COMPUTATIONAL LEARNING MODELS AND METHODS DRIVEN BY OMICS FOR PRECISION MEDICINE

EDITED BY: Hongmin Cai, Huiru Zheng, Fa Zhang, Quan Zou and Yanjie Wei

PUBLISHED IN: Frontiers in Genetics and
Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-460-3
DOI 10.3389/978-2-88966-460-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL LEARNING MODELS AND METHODS DRIVEN BY OMICS FOR PRECISION MEDICINE

Topic Editors:
Hongmin Cai, South China University of Technology, China
Huiru Zheng, Ulster University, United Kingdom
Fa Zhang, Institute of Computing Technology (CAS), China
Quan Zou, University of Electronic Science and Technology of China, China
Yanjie Wei, Shenzhen Institutes of Advanced Technology (CAS), China

Citation: Cai, H., Zheng, H., Zhang, F., Zou, Q., Wei, Y., eds. (2021). Computational Learning Models and Methods Driven by Omics for Precision Medicine. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-460-3

Table of Contents

frontiers in Genetics

EDITORIAL
published: 23 December 2020
doi: 10.3389/fgene.2020.620976

Editorial: Computational Learning Models and Methods Driven by Omics for Precision Medicine

Lei Zhu[1], *Hongmin Cai*[1]*, *Fa Zhang*[2], *Quan Zou*[3], *Yanjie Wei*[4] *and Huiru Zheng*[5]

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, [2] Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China, [3] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [4] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences (CAS), Shenzhen, China, [5] Faculty of Computing, Engineering and the Built Environment, School of Computing, Engineering and Intelligent Systems, Ulster University, Coleraine, United Kingdom

Keywords: omics, machine learning, drug, RNA, disease, biomarker, network analysis, deep learning

Editorial on the Research Topic

Computational Learning Models and Methods Driven by Omics for Precision Medicine

Due to the high experimental cost and the exponential decline in the cost of high-throughput sequencing, computational models, and methods are preferred by scholars. The curse of dimensionality is the primary obstacle to dealing with the explosive growth of omics data. Machine learning methods are applied to reduce dimensionality and perform feature selection from massive data. Researchers meet the requirements of data sparsity by increasing the sparsity constraints of the computational models. The models combined with the deep learning method help to discover potential non-linear associations. Improving data representation or adding embedding layers could provide better performance of the models. Computational methods for biomarker discovery, sample classification, and disease process interpretation pave the way for precision medicine.

This topic includes 34 papers and a corrigendum. These papers introduce latest researches in the area of computational biology, catering for precision medicine and complex diseases. They include sequencing alignment, correlation detection between omics data and biological traits, prediction of biological functionality, computational methods for cancer subtyping, finding of pathogenic causes, repositions and targeting, and computational methods specially designed for biological knowledge mining.

SEQUENCE ALIGNMENT

The raw sequencing data is unstructured short sequences. The structured data can be generated from downstream analysis through filtering, quality control, and assembly of these unstructured data. Assembly reconciliation can generate high-quality assembly results. In Tang et al., using the consensus blocks between contigs to construct adjacency graphs to avoid varying sequencing depth and sequencing errors, the authors propose a scoring function to rank the input assembly sets. They use an adjacency algebra model for accurate fusion, which performs well on *M. abscessus*, *B. fragilis*, *R. sphaeroides,* and *V. cholerae*. Shi and Zhang apply the partition and recur platform to generate a high-level abstraction of the sequence alignments. The algorithm component library is verified

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Hongmin Cai
hmcai@scut.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *24 October 2020*
Accepted: *01 December 2020*
Published: *23 December 2020*

Citation:
Zhu L, Cai H, Zhang F, Zou Q, Wei Y and Zheng H (2020) Editorial: Computational Learning Models and Methods Driven by Omics for Precision Medicine. Front. Genet. 11:620976. doi: 10.3389/fgene.2020.620976

by Apla language. The advantage of implementing the sequence assembly process through abstract components is that it can effectively improve stability and reduce the possibility of errors caused by manual selection.

ESTABLISHING OMICS—DISEASE ASSOCIATIONS

Four groups present research on RNA association prediction, including Long non-coding RNA(lncRNA)–protein interactions (LPI), LncRNA-Disease, microRNA(miRNA)-Disease, and Circular RNA(circRNA)-Disease. Peng et al. give us an overview of how to identify lncRNA–protein interactions(LPI), and they introduced 16 related repositories and methods. Among these network-based and deep learning-based methods for predicting LPI, the proposed SFPEL-LPI used assembly learning and achieved the best Area Under Curve(AUC) performance. Hu et al. combined the two methods of neural network and matrix factorization (MF) to predict lncRNA-disease associations. They achieved this combination by concatenating outputs and sharing inputs between the two methods. Both the MF and the neural network are trained simultaneously under the framework of TensorFlow. In Yu, Shen et al., prior information (lncRNA-miRNA and lncRNA-disease associations) and known miRNA-disease associations are integrated to construct a three-layer heterogeneous network of LncRNA, miRNA, and disease. In this three-layer network, the edges between the layers are filled with prior information. Random walk is applied to predict miRNA-disease associations. The proposed methods are evaluated using cancer data. Their results show that most potential miRNAs can be confirmed by databases. In Lei X. et al., the cirRNA similarity network and the disease similarity network are used as the input of the collaboration filtering recommendation system. Their experiments on predicting potential circRNA–disease associations indicate the effectiveness of the recommendation system algorithm.

Like RNA, microbes and pathogens are also the causes of diseases. In Li, Wang, Chen et al., a bipartite network is applied to avoid the omission of neighbor information for predicting Pathogen–Host associations. Among the top 20 pathogen-host pairs discovered, 16 pairs can be verified by biological experiments. In Ma et al., to explore the pathogenesis of complex diseases from the modular perspective, the similarity matrix is decomposed to generate microbe-disease co-modules by non-negative matrix tri-factorization. Their method achieves nice performance in the enrichment index and the number of significantly enriched taxon sets. In Li S. et al., on the strength of a matrix containing microbes similarity, disease similarity and a bipartite graph network of the two interactions, the potential microbe-disease associations are calculated by Katz centrality. The prediction performance was evaluated by the leave-one-out cross validation and reached an AUC of 0.9098. Zhu et al. use a deep feedforward network to identify microbial markers and realize graph embedding by replacing the first two layers of the network with a sparse graph. Experiments show that this Graph Embedding Deep Feedforward Network has the best performance, comparing deep forest, random forest and Support Vector Machine(SVM).

PREDICTION OF BIOLOGICAL FUNCTIONALITY

Identifying acetylation proteins is conducive to understanding the post-translational modification process. In Qiu et al., the authors first generate a k-nearest neighbors (KNN) score, and then use random forest to classify the acetylation proteins. The formation of KNN scores is based on domain annotation and subcellular localization. Five-fold cross-validation on the three data sets was performed, and finally, an average AUC of 0.8389 was obtained. In Miao et al., the authors aim to identify which proteins are endoplasmic reticulum-resident proteins, and they achieved accuracy over 86%. Such work allows us to understand the functionality of proteins, which may be potential points of drug design. The promoter drives the flow of genetic information from DNA to RNA, and its sequence information determines the strength of the promoter. In Le et al., the promoter sequence is divided into 10-gram levels and is used to form a 1,000-dimensional vector. The vector is input into a deep neural networks model to classify the promoter strength. Compared with other latest methods in the same test set, this method improves 1–4% on all indicators.

COMPUTATIONAL APPROACH FOR CANCER SUBTYPING

Cancer subtyping is fundamental for precision therapy. Accurately identifying cancer subtypes enables us to understand cancer evolution. In Lu et al., Laplacian score and low-rank representation methods are integrated to obtain a low-rank expression of cancer gene expression data. This low rank matrix is hoping to preserve subtype information. By sorting the obtained matrix, the feature genes are heuristically selected to comprise of a gene subset for accurate cancer subtyping. The method is tested on five cancer dataset and is shown to achieve superior performance over k-means, non-negative matrix factorization (NMF) and several other baseline methods. Aouiche et al. obtained the cancer stages on copy number variation(CNV) data. The positive significance of distinct stage division is dependent on not only a high cure rate after cancer been detected, but also on critical markers, which are potential therapeutic targets. Li, Wang, Wang et al. identify differentially expressed genes(DEGs) in tumor by analyzing the residues of each gene via a regression model and found potential biomarkers of the individual sample from DEGs. Survival analysis is performed on samples collected from human and mouse cancer data, and is shown to be statistically differently.

QUANTITATIVE UNDERSTANDING OF PATHOGENIC CAUSES

The goal of developing computational disease models is to find a therapeutic target. As the first step, computational tools are

required to explain the cause of the disease. Regarding the identification of Schizophrenia (SZ), Xiang et al. construct a Brainnetome atlas based on resting-state functional magnetic resonance imaging. Brainnetome atlas is a weighted undirected graph constructed with brain regions as nodes and correlations as edges. The authors calculate the features from the altas and, then use least absolute shrinkage and selection operator(lasso) learning to prune the features. The classification is SZ is achieved by using SVM with an accuracy of 93.10%. In Li X. et al., each single sample is classified by a pathway-based approach, into Ulcerative colitis (UC) and Crohn's disease (CD). Even though UC and CD have common clinical characteristics, they have different responses to drugs. According to the gene expression data of the sample, the author scores each pathway to form a pathway activation for single sample matrix, which is classified by a random forest classifier. In Zhang S. et al., the authors aim to select CNV markers to distinguish between three different states of mono-ADP-ribosylhydrolase 2 (MACROD2). The frequent deletions of MACROD2 locus may lead to chromosomal instability of human colorectal cancer. The authors firstly select 17 important single nucleotide polymorphism(SNP) site via mutual information, and then uses bootstrapping scheme to train multiple classifiers. The trained classifiers are finally ensembled to effectively distinguish three types of MACROD2. In Lei W. et al., the effectiveness of lipoprotein 2 on Subarachnoid hemorrhage (SAH) intervention is revealed from the perspective of the cell signaling pathway. The authors discover five biomarkers, three of which have been verified by previous experimental evidence. Finally, the early SAH prediction is performed based on the assembly learning of logistic regression, SVM and Naive-Bayes, achieving an accuracy of 79%. Zhang P. et al. clarify a pathway of polycistronic mRNA ORF73 involved in host apoptosis through protein p53, supplementing the pathogenic process of Kaposi sarcoma-associated herpes virus. This work is mainly done through protein-protein interactions (PPI) analysis, Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway analyses. In Shao et al., 108 whole-non-structural protein 5 sequences are analyzed in Zika virus, and 35 potential glycosylation and phosphorylation sites have been discussed. Mutations in amino acid sites are found to be correlated with their pathogenicity and transmission efficiency. The relatively stable nucleic acid sequence is shown to be helpful for detection and vaccine development.

A meta-analysis can combine multiple studies, and the two groups apply meta-analysis methods. In Fukutani et al., after the analysis of Human T-lymphotropic virus 1 (HTLV-1)-infected patients, the authors find that gene CD40LG and gene GBP2 can be used as two phenotypic classifications of HTLV-1 infection, with accuracy rates of 0.88 and 1. In Jin and Shi, a meta-analysis is performed to test SNP-environment interaction. Based on meta-regression (MR), the author proposes overlapping MR combined with the method of processing overlapping data. This method can reduce type I error and is more robust than MR in dealing with the non-linear interaction effect.

Gao et al. screen 107 methylomic features in whole blood methylation samples and use Support Vector Regressor to predict age. What is interesting is that only gene CALB1 and gene KLF14 are both found in the male and female age prediction models.

DRUG REPOSITIONS AND TARGETING

Four works focus on drug repositions. In Manibalan et al., the authors focus on the S100A8 protein, which has a strong interaction with the prevalence of polycystic ovary syndrome biomarkers. Therefore, they design a series of RNA aptamers targeting the S100A8, and select the one with minimal binding energy as the targeted drug. Wound Scratch experiments confirm that the synthesized 18-mer oligo has a significant inhibition effect on tumor cell migration. Wu et al. hope to level the differences in chemotherapy prognosis through cisplatin resistance analysis of oral squamous cell carcinoma. Through the analysis of differentially expressed genes, PPI network and miRNA-mRNA targeted regulatory network, they find that five hub genes and the miR-200 family members that regulate hub genes may be potential drug targets. In Yu, Xu et al., new targeted drugs for hepatocellular carcinoma (HCC) are found by the drug repositioning bioinformatics method. Finding HCC's kernel genes is the first step in work. The next step is to combine the relationship between the drug and gene expression in the Connectivity Map database to score the relationship between the drug and HCC. Among the top ten drugs screened by this method, eight drugs have been supported by publications. In Emdadi and Eslahchi, cell line similarity, drug similarity and half maximal inhibitory concentration are combined to predict the drug sensitivity of cells, and logistic matrix factorization is applied to obtain latent vectors. For the drug sensitivity prediction of the new cell line, the k-nearest neighbors of the cell line are estimated through the decision tree to obtain the latent vectors of the cell line. Finally, a threshold based on the probability of the latent vector is used to predict whether the cell line is sensitive to drugs. The genomics of drug sensitivity on haematopoietic cell lines in cancer was tested for model performance, with an accuracy of 0.721.

BIOLOGY-ORIENTED LEARNING METHODS

Traditional learning methods have achieved tremendous success and have provided solutions to even some difficult biological problems. In Wang et al., Huber loss is applied to alleviate non-Gaussian noise contaminations. A sparsity penalty item is used to encourage the sparsity of representation of The Cancer Genome Atlas data, and a graph regularization is used to preserve the manifold structure. The clustering accuracy is improved by 5% compared with non-negative matrix factorization. Che et al. improve the traditional methods on the basis of Sparse Group Lasso (SGL) and proposed a weighted sparse group lasso (WSGL) by introducing prior constraint on the sparse term. Compared with lasso and SGL, the performance is significantly improved, indicating that prior biological knowledge carries on valuable message. Comparing the lasso and SGL methods, WSGL can

screen less genes, and the ratio of candidate genes is higher using *Arabidopsis* flowering time data. In Lemaçon et al., a visualization method is proposed based on a scoring system for rating susceptibility loci. In general, this is a visualization method for searching for the best potential variants through aggregating prediction approaches. In Guo, Kullback-Leibler divergence is used to measure the distance between two SNPs, and these distances are used as k-means clustering. Then, statistical testing methods are applied to find epistatic interactions, and the time cost of this method is about one-tenth that of Bayesian inference-based method. Zheng et al. use sparse subspace clustering to perform single-cell clustering. This method assumes that the feature vector of a sample can be expressed as a linear combination of other samples in the same subspace. In the test of 10 single-cell datasets, this method maintains the leading position in normalized mutual information and adjusted rand index.

These teams work together to continuously improve model accuracy. Most articles related to computational methods are tailored from early established models for biology knowledge learning.

AUTHOR CONTRIBUTIONS

The article was written by LZ, HC, FZ, QZ, YW, and HZ have provided guidance to the manuscript preparation, have also reviewed and edited the paper. All authors have approved the final version of the editorial.

ACKNOWLEDGMENTS

We thank all the authors who contributed to this topic.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 September 2019
doi: 10.3389/fgene.2019.00897

Predicting circRNA–Disease Associations Based on Improved Collaboration Filtering Recommendation System With Multiple Data

Xiujuan Lei[1]*, *Zengqiang Fang*[1] and *Ling Guo*[2]

[1] School of Computer Science, Shaanxi Normal University, Xi'an, China, [2] College of Life Sciences, Shaanxi Normal University, Xi'an, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Zhipeng Cai,
Georgia State University, United States
Chaoyang Zhang,
University of Southern Mississippi, United States

***Correspondence:**
Xiujuan Lei
xjlei@snnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 12 June 2019
Accepted: 23 August 2019
Published: 25 September 2019

Citation:
Lei X, Fang Z and Guo L (2019) Predicting circRNA–Disease Associations Based on Improved Collaboration Filtering Recommendation System With Multiple Data. Front. Genet. 10:897. doi: 10.3389/fgene.2019.00897

With the development of high-throughput techniques, various biological molecules are discovered, which includes the circular RNAs (circRNAs). Circular RNA is a novel endogenous noncoding RNA that plays significant roles in regulating gene expression, moderating the microRNAs transcription as sponges, diagnosing diseases, and so on. Based on the circRNA particular molecular structures that are closed-loop structures with neither 5′-3′ polarities nor polyadenylated tails, circRNAs are more stable and conservative than the normal linear coding or noncoding RNAs, which makes circRNAs a biomarker of various diseases. Although some conventional experiments are used to identify the associations between circRNAs and diseases, almost the techniques and experiments are time-consuming and expensive. In this study, we propose a collaboration filtering recommendation system–based computational method, which handles the "cold start" problem to predict the potential circRNA–disease associations, which is named ICFCDA. All the known circRNA–disease associations data are downloaded from circR2Disease database (http://bioinfo.snnu.edu.cn/CircR2Disease/). Based on these data, multiple data are extracted from different databases to calculate the circRNA similarity networks and the disease similarity networks. The collaboration filtering recommendation system algorithm is first employed to predict circRNA–disease associations. Then, the leave-one-out cross validation mechanism is adopted to measure the performance of our proposed computational method. ICFCDA achieves the areas under the curve of 0.946, which is better than other existing methods. In order to further illustrate the performance of ICFCDA, case studies of some common diseases are made, and the results are confirmed by other databases. The experimental results show that ICFCDA is competent in predicting the circRNA–disease associations.

Keywords: circRNA–disease association, collaboration filtering, multiple biological data, recommendation system, neighbor information

INTRODUCTION

Circular RNA (circRNA) is a relatively novel biological molecule compared with the usual linear RNAs. Circular RNAs were first discovered in the RNA viruses before 1970 (Sanger et al., 1976). It is said that circRNAs lack covalently closed-loop structures with neither 5′-3′ polarities nor polyadenylated tails (Chen and Yang, 2015), which causes that it is not easy to find circRNAs compared with linear RNAs. Because of circRNAs closed-loop structure, however, more and more circRNAs (Hsu and Coca-Prados, 1979; Arnberg et al., 1980; Pasman et al., 1996) are revealed based on the development of the RNA base sequence high-throughput techniques. In terms of recent researches, there are various kinds of circRNAs in the creatures, which include as follows: exonic circRNAs, which are mainly produced by back-spliced exons (Wilusz and Sharp, 2013), introns circRNAs extracted from introns (Lasda and Parker, 2014), exon–intron circRNAs that are analogous to ecircRNAs (Li et al., 2015), and integrated circRNAs discovered by a biological identifier, CIRI (Gao et al., 2015). Many recent evidences (Danan et al., 2011) show that circRNAs play significant roles in different biological processes and have multiple biological functions (Jeck and Sharpless, 2014; Qu et al., 2015). First, circRNA can be regarded as miRNA sponges (Hansen et al., 2013; Kulcheski et al., 2016), which could be adopted to be an identifier for diseases. Second, some evidences illustrate that circRNAs also can regulate some transcriptional processes (Chao et al., 1998). Simultaneously, circRNAs also have associations with RNA-binding proteins (RBPs) (Panda et al., 2017) bases on their stable circular structures. Circular RNA has different ways to bind with the RBPs compared with the linear RNA (Memczak et al., 2013), which indicates that circRNAs have potential to be disease biomarkers. Moreover, circRNAs have some translational functions (Chen and Sarnow, 1995) like common RNAs.

With the further study of circRNAs' functions, increasing numbers of evidences point out that circRNAs have associations with complicated diseases (Xu et al., 2017) or have effects on the translation of some proteins (Bartsch et al., 2018). There are many previous searches revealing the associations between circRNAs and some cancers. Circular RNA circ-PVT1 has been discovered to upregulate the gene expression in the gastric cancer (GC) tissues and promotes the GC cells reproduction (Chen et al., 2017a). In contrast circRNA hsa_circ_0000190, it regulates the gene expression in GC tissues by downregulation (Chen et al., 2017b). CircRNA circTCF25 can upregulate the gene expression or cell proliferation of 13 target locus of miRNA miR-103a-3p/miR-107, which can be regarded as a biomarker of bladder cancer (BC) (Zhong et al., 2016). Circular RNA hsa_circRNA_105055 and hsa_circRNA_086376 are the potential biomarkers of colorectal cancer by working as sponges for miR-7 (Zeng et al., 2017). Moreover, circRNA hsa_circ_0054633 also has association with diabetes, especially for prediabetes and type 2 diabetes mellitus (Zhao et al., 2017).

Because of the development of RNA base sequence techniques, more and more circRNA-related information is excavated. Thus, many different kinds of circRNA-related databases are established for further researches of various diseases, biological molecules and pathways, etc. To create more convenience to the researchers, circBase database (Glazar et al., 2014) was developed to provide the evidence supporting their expression, and all the data can be accessed, downloaded, and browsed within the genomic context. Circular RNADb (Chen et al., 2016a) is a comprehensive circRNA database that collects human protein-coding annotations of circRNAs and includes some important information about exonic circRNAs such as genomic information, exon splicing, genome sequence, internal ribosome entry site, open reading frame, and cricRNA-related references. Furthermore, ExoRBase (Li et al., 2017) is an online accessible database that extracts data from RNA-seq data analyses of human blood exosomes. circNet (Lin et al., 2015) is also a circRNA-related database from which tissue-specific circRNA expression profiles and circRNA-miRNA-gene regulatory networks can be downloaded. Moreover, circ2Traits (Ghosal et al., 2013) is an overall circRNA–disease associations database, which obtains the associations as follows: one is identifying the interactions of circRNAs with disease-related miRNAs; the other is matching the diseases associated SNPs on circRNA loci. To obtain more reliable circRNA–disease associations, circR2Disaese (Fan et al., 2018) database (http://bioinfo.snnu.edu.cn/CircR2Disease/) was developed. The whole circRNA–disease associations are collected manually from relevant references and reviews, which provides more convenience and basics to infer novel circRNA–disease associations.

Although, there are many circRNA–disease associations discovered by biological experiments, whose experimental processes are extremely expensive and time-consuming. On the one hand, there are a limited number of computational methods existing to predict potential circRNA–disease associations. On the other hand, we lack comprehensive circRNA-related diseases databases, which are our main motivation to propose a new computational method based on circR2Disease database. In this study, we develop an improved collaboration filtering recommendation system (Pan et al., 2008) method to predict circRNA–disease associations, which is named ICFCDA. First, circRNAs target gene–related gene ontology (GO) terms, circRNAs base corresponding sequences data, and circRNA–disease associations are adopted to calculate the circRNA functional annotation semantic similarity, sequence similarity, and Gaussian interaction profile (GIP) kernel similarity. Second, disease-related genes and circRNA–disease associations are used to calculate the disease functional similarity and disease GIP kernel similarity. Furthermore, we also replace the disease names into disease ontology (DO) IDs to calculate the disease semantic similarity based on the DOSE (Yu et al., 2015) tool. Third, multiple disease similarities and circRNA similarities are combined with the final disease similarity matrix and circRNA similarity matrix, respectively. Finally, collaboration filtering method is adopted to calculate the score of each circRNA–disease pair. For the sake of evaluating the performance of method we proposed, leave-one-out cross validation (LOOCV) is used to calculate the area under receiver operating characteristic (ROC) curve (AUC) value. Moreover, several common diseases also are tested by the LOOCV mechanism. In addition, case studies of two common diseases are implemented to further illustrate the performance of ICFCDA.

MATERIALS AND METHODS

Human circRNA–Disease Associations

To extract circRNA–disease associations, the initial circRNA–disease associations datasets are downloaded from circR2Disease database (Fan et al., 2018) (http://bioinfo.snnu.edu.cn/CircR2Disease/). In the original dataset, there are 725 circRNA–disease associations that have been verified by biological experiments. These 725 circRNA–disease associations contain 661 circRNA individuals and 100 disease individuals. In term of the initial dataset, 212 circRNA–disease associations are picked out as the known associations in this study, which are composed of 42 disease entities and 200 circRNA entities. The adjacency circRNA–disease association matrix is deciphered by matrix A. If there is an association between the disease i and circRNA j, $A(i, j)$ is equal to 1 or $A(i, j)$ is equal to 0.

circRNA Similarity

circRNA Functional Annotation Semantic Similarity

On the basis of the original circRNA–disease associations, 200 circRNA entities are screened out. Then human GO terms data are downloaded from human protein reference database (HPRD) database (Keshava Prasad et al., 2009). The initial circRNA–disease associations provide the circRNAs-related genes. Thus, the circRNA-related genes are utilized to match GO data extracted from HPRD database. In this study, an information content algorithm (Lin, 1998) is adopted to calculate the circRNA functional annotation semantic similarity. *CFS* is used to describe the circRNA functional annotation semantic similarity network. Moreover, the following equation is used to calculate the circRNA functional annotation semantic similarity:

$$CFS(i,j)=\frac{2\times\log P(C_i\cup C_j)}{\log P(C_i)+\log P(C_j)} \tag{1}$$

where *CFS(i, j)* denotes the functional annotation semantic similarity between circRNA *Ci* and *Cj*; *P*(*Ci*) and *P*(*Cj*) represent the probability between the number of *Ci* and *Cj* target gene–related GO terms and the number of the entire GO terms. $P(C_i\cup C_j)$ is the ratio of between the union of the number of circRNA *Ci* and *Cj* target gene–related GO terms and the number of the entire GO terms.

circRNA Sequence Similarity

For the sake of calculating the circRNA sequence similarity, the circRNA corresponding RNA base sequence data are downloaded from circBase database (Glazar et al., 2014) (http://www.circbase.org/). In our computational model, there are 200 circRNAs needing matching their related RNA base sequences. A base pairing algorithm named the Needleman-Wunsch pairwise alignment algorithm is used to calculate the circRNA sequence similarity, which is integrated into a python toolkit called Biopython (Cock et al., 2009). Therefore, there are some parameters needing setting up for obtaining a better result. The gap-open penalty is set as 2, and the gap-open extending penalty is set as −0.5 to −0.1. *CSS* is adopted to represent the circRNA sequence similarity matrix, and *CSS*(*i*, *j*) represents the similarity value between the circRNA C_i and C_j. Then, the Needleman-Wunsch score of each circRNA pair is normalized as follows:

$$CSS(i,j)=\frac{NW(i,j)}{\sqrt{NW(i,i)}\sqrt{NW(j,j)}} \tag{2}$$

where *NW(i, j)* is the score of the Needleman-Wunsch algorithm between circRNA *i* and *j*.

circRNA GIP Kernel Similarity

Known circRNA–disease associations are adopted to calculate circRNA GIP kernel (Van Laarhoven et al., 2011) similarity marked as *CGS*. According to an assumption (Van Laarhoven et al., 2011) that the more similar the two circRNAs are, the more likely the disease associated with one of them is to be associated with another. Therefore, V_{Ci} is used to represent the interaction profile of circRNA *C(i)* with each disease, which means the ith row in the circRNA–disease association network. The GIP kernel similarity between circRNA *C(i)* and *C(j)* is calculated as follows:

$$CGS(i,j)=\exp(-\gamma_c\left\|V_{C_i}-V_{C_j}\right\|^2) \tag{3}$$

where *CGS*(*i*, *j*) is the GIP kernel similarity of circRNA *i* and *j*. γ_c is an adjusting parameter, which controls the bandwidth of each kernel, which can be initialized as follows:

$$\gamma_c=\hat{\gamma_c}\Big/\left(\frac{1}{N_c}\sum_{i=1}^{N_c}\left\|V_{C_i}\right\|\right) \tag{4}$$

Where $\hat{\gamma_c}$ is the initial value, which is set as 1 based on the previous study (Van Laarhoven et al., 2011). N_c is total number of circRNAs.

circRNA Similarity Integration

Finally, we obtain the circRNA functional annotation semantic similarity, sequence similarity, and GIP kernel similarity. In order to make full use of these three circRNA similarities, the following equation is adopted to integrate the circRNA similarities:

$$CS(i,j)=\begin{cases} CGS(i,j), \text{ if } CGS(i,j)\neq 0 \\ \alpha CFS(i,j)+(1-\alpha)CSS(i,j), \text{otherwise} \end{cases} \tag{5}$$

where *CS* denotes the integrated circRNA similarity network; α is a harmonic mean factor to integrate the circRNA functional annotation semantic similarity *CFS*, and the circRNA sequences similarity *CSS*.

Disease Similarity

Disease Functional Similarity

Furthermore, disease-related genes are downloaded from DisGeNET (Pinero et al., 2017) database, which gathers more than 3,815,056 gene–disease associations between 16,666 gene individuals and 13,172 disease individuals. In order to obtain more reliable disease similarity, we also extract disease-related genes from Online Mendelian Inheritance in Man (OMIM)

(Hamosh et al., 2005) database. Based on the initial circRNA–disease associations, 42 independent disease entities are picked out as the experimental data. Then, those above disease entities are used to match the disease phenotype corresponding genes in the OMIM dataset manually. In this study, JACCARD algorithm, a statistic method, is used to calculate the disease functional similarity as follows:

$$DS1(i,j)=\frac{\left|DG(i)\cap DG(j)\right|}{\left|DG(i)\cup DG(j)\right|} \tag{6}$$

where *DG(i)* and *DG(j)* denote the subsets of the disease i and j related genes.

Disease GIP Kernel Similarity

GIP kernel similarity algorithm is also adopted to calculate the disease GIP kernel similarity between *D*(*i*) and *D*(*j*), which is similar to calculate circRNA GIP kernel similarities. The computing process is as follows:

$$DGS(i,j)=\exp(-\gamma_d\left\|V_{D_i}-V_{D_j}\right\|^2) \tag{7}$$

where *DGS* is the disease GIP kernel similarity network, and the *DGS*(*i*, *j*) is GIP kernel similarity score between disease *i* and *j*. γ_d is also a bandwidth adjustment parameter, which is defined as follows:

$$\gamma_d=\hat{\gamma_d}\Big/\left(\frac{1}{N_d}\sum_{i=1}^{N_d}\left\|V_{D_i}\right\|\right) \tag{8}$$

where $\hat{\gamma_d}$ is the initial value, which is set as 1 based on the previous study (Van Laarhoven et al., 2011). N_d is total number of diseases.

Disease Semantic Similarity

In order to calculate the semantic similarity between these 42 diseases, the disease-relevant DO IDs are extracted from the DO (Kibbe et al., 2015) database. Then all the 42 diseases' names are replaced into the corresponding DO IDs, which are adopted to input into a R package named DOSE (Yu et al., 2015) to calculate the disease semantic similarity. After the semantic similarity score of each disease pair is obtained, *DS*2 can be used to represent the diseases semantic similarity matrix.

Disease Similarity Integration

Thus, the integrated disease similarity thereby can be accessed by combining the disease functional similarity, GIP kernel similarity, and semantic similarity. In this study, when we fuse different disease similarities, different weights are allocated to the disease functional similarity matrix, GIP kernel similarity matrix, and semantic similarity matrix based on the following formula:

$$DS(i,j)=\begin{cases} DGS(i,j), \text{ if } DGS(i,j)\neq 0 \\ \beta DS1(i,j)+(1-\beta)DS2(i,j), \text{otherwise} \end{cases} \tag{9}$$

where *DS* denotes the integrated disease similarity network.

ICFCDA

With the increasing numbers of data in all aspects, it is important to predict or recommend some associations between the two different things. It is in this case that the recommendation system algorithm has attracted the attention of many experts. Collaborative filtering algorithm (Schafer et al., 2007; Zhou et al., 2015) is one of the recommendation system algorithms, which is applied to recommend movies (Zhou et al., 2008) or news (Das et al., 2007) for users. In this study, we first adopt the collaborative filtering recommendation system algorithms to predict the circRNA–disease associations, which is named as ICFCDA, and its flowchart is illustrated in **Figure 1**.

For scoring each circRNA–disease association, there are five steps in our computational method as follows:

Step 1: Obtaining the top *k* similar neighbors of each circRNA based on circRNA similarity network *CS*.
Step 2: Obtaining the top *k* similar neighbors of each disease based on disease similarity network *DS*.
Step 3: Calculating the scores of circRNA–disease association by the collaborative filtering recommending based on circRNAs.
Step 4: Calculating the scores of circRNA–disease association by the collaborative filtering recommending based on diseases.
Step 5: Integrating the final recommendation scores based on Steps 3 and 4.

First, the similarity scores between circRNA *j* and other circRNAs in the circRNAs dataset are listed in descending order. Then, the most similar top *k* neighbors of each circRNA are picked out based on the final integrated circRNA similarity network *CS*. We conduct the same above processes for each circRNA. Therefore, we obtain the most similar top *k* neighbors of each circRNA. Furthermore, the value of *k* is set as the 4% of the number of the whole circRNAs, which can be described as nc*0.04.

Second, in terms of the most similar top *k* neighbors of cirRNA *j* and the associations between the disease *i* and the neighbors of the circRNA *j*, the most similar top *k* neighbors of the circRNA-based recommendation score between the disease *i* and the circRNA *j* can be calculated as follows:

$$CRS(i,j)=\frac{1}{k}\left(\sum_{n=1}^{k}A(i,n)\times CS(n,j)\right) \tag{10}$$

where *CRS*(*i*, *j*) is the recommendation score between the disease *i* and the circRNA *j* based on the top *k* most similar neighbors of circRNA *j*. *A*(*i*, *n*) is the association information of the *n*th most similar neighbor of circRNA *j* and the disease *i*. *CS*(*n*, *j*) is the similarity score of the *n*th most similar neighbor circRNA and circRNA *j*.

Third, the similarity scores between disease *i* and other diseases in the disease dataset are listed in descending order. Then, the most similar top k neighbors of each disease are screened out based on the final integrated disease similarity network *DS*. We also carry out the same processes for each disease. Therefore,

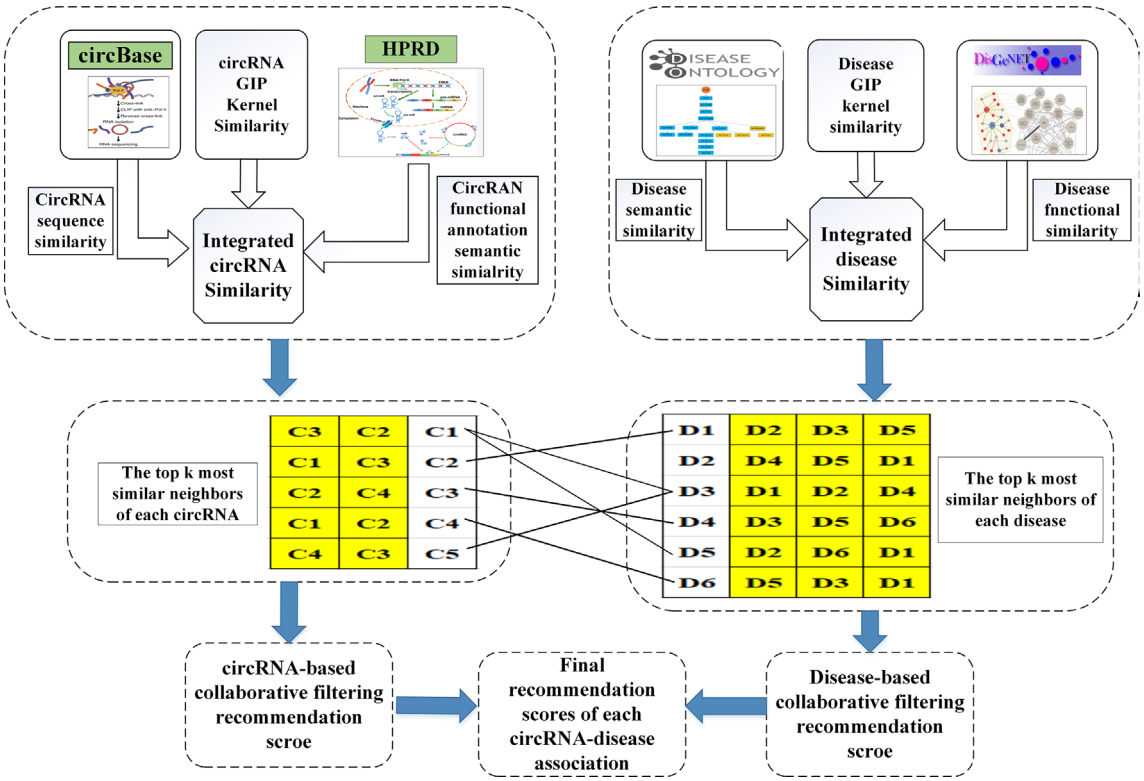


FIGURE 1 | The flowchart of the computational method ICFCDA.

the most similar top k neighbors of each disease are selected. Moreover, the value of k is set as the 4% of the number of the whole diseases, which can be described as $nd * 0.04$.

Fourth, based on the most similar top k neighbors of disease i and the associations between the neighbors of the disease i and the circRNA j, the most similar top k neighbors of the disease-based recommendation score between the disease i and the circRNA j can be calculated as follows:

$$DRS(i,j)=\frac{1}{k}\left(\sum_{m=1}^{k}DS(i,m)\times A(m,j)\right) \tag{11}$$

where $DRS(i, j)$ is the recommendation score between the disease i and the circRNA j based on the top k most similar neighbors of disease i. $A(m, j)$ is the association information of the mth most similar neighbor of disease i and the circRNA j. $DS(i, m)$ is the similarity score of the mth most similar neighbor disease and disease i.

Finally, the circRNA-based recommendation scores and the disease-based recommendation scores are combined with the final recommendation scores as follows:

$$IRS(i,j)=\gamma DRS(i,j)+(1-\gamma)CRS(i,j) \tag{12}$$

where $IRS(i, j)$ is the integrated recommendation scores between the disease i and the circRNA j. The parameter $\gamma\in[0, 1.0]$ is a balance factor that can control the significance of the circRNA-based recommendation scores and the disease-based recommendation scores.

In order to solve the "cold start" problem in the collaborative filtering recommendation system, the importance of neighbors is taken into consideration. The more diseases/circRNAs are shared by two cicRNAs/diseases, the more significant it is. The importance of two diseases/circRNAs can be defined as follows:

$$IMP(C(i),C(j))=f_{exp}(C(i))*f_{ns}(C(j))*\sum_{C(c(k))}f_{cod}(c(k)) \tag{13}$$

where $IMP(C(i), C(j))$ is the significance coefficient between circRNA i and j. IMP is divided into three parts, which include the circRNA $C(i)$ related diseases $f_{exp}(C(i))$, which can be calculated as the following equation:

$$f_{exp}(C(j))=\frac{1}{D(C(i))} \tag{14}$$

where $D(C(i))$ is circRNA i–related diseases, which means that circRNA i would provide more useful suggestion, if it is associated with fewer diseases. $f_{ns}(C(j))$ is the similarity if disease j based on the disease i, which is defined as follows:

$$f_{ns}(C(j))=\frac{1}{D(C(j))-I(C(i),C(j))+1} \tag{15}$$

where $I(C(i), C(j))$ is intersection of the circRNA i and j–related disease dataset. $f_{cod}(C(k))$ is the disease that is merely associated

with circRNA *i* and *j*. Therefore, for those circRNAs that have only one relevant disease, the following equation is adopted to calculate the recommendation score:

$$Score_{cold\ start} = \sum_{i=1}^{Nc} IMP(C(t),C(i)) * CS(C(t),C(i)) \quad (16)$$

Performance Metric

In order to evaluate the performance of our proposed computational method, the AUC value that is the area of the ROC curve and the *f*-measure, which is a comprehensive metric using the *precision* and the *recall*, are the two main evaluation metrics in this study. The ROC curve consists of the true-positive rate (TPR) and the false-positive rate (FPR), which are calculated by the following equations:

$$TPR = \frac{TP}{TP + FN} \quad (17)$$

$$FPR = \frac{FP}{FP + TN} \quad (18)$$

where TP is the number of the positive samples that is the known circRNA–disease associations, which are predicted as the true circRNA–disease associations, and FN is the number of the negative samples predicted as the false circRNA–disease associations. FP is the number of the incorrectly predicted positive samples, and the TN is the number of the truly predicted negative samples. In addition, the *precision* is the true-positive samples in the dataset, which are predicted as positive samples dataset. The *recall* is the ratio between the samples that are predicted as positive samples and the whole positive samples. Thus, *f*-measure is illustrated as follows:

$$precision = \frac{TP}{TP + FP} \quad (19)$$

$$recall = \frac{TP}{TP + FN} \quad (20)$$

$$f - measure = \frac{2 \times precision \times recall}{precision + recall} \quad (21)$$

RESULTS

Leave-One-Out Cross Validation

In this study, a cross validation mechanism, LOOCV, is adopted to test the performance of our proposed computational method, ICFCDA. For each given disease in the circRNA–disease association network, there could be one or several relevant circRNAs with each specific disease. First, for each given disease *i*, some circRNAs are confirmed that they are associated with the disease *i*, which are the known circRNA–disease associations. Each association between the disease *i* and one particular circRNA could be regarded as test data, and all the left circRNA–disease associations are seen as training dataset. During each LOOCV procedure, one circRNA–disease association potential score is generated. When all the scores of the test dataset are obtained, the remaining unknown circRNA–disease associations are treated as the test dataset. Finally, the predictive score of each circRNA–disease pair is obtained. Each circRNA–disease association score is a threshold after the final potential scores of the circRNA–disease associations are sorted in descending order. With the changing threshold, we can calculate the TPRs and the FPRs, which are adopted to draw the ROC curve and calculate the AUC value. In order to evaluate the performance of ICFCDA, the AUC value is compared with other seven state-of-the-art methods such as heterogeneous graph inference (HGI) method (Chen et al., 2016b), KATZ (Ganegoda et al., 2014), random walk restart (RWR) (Chen et al., 2012), and graph regularized nonnegative matrix factorization (NMF) (Liu et al., 2018), respectively. The result is shown in **Figure 2**, which illustrates that the performance of ICFCDA is better than others. According to **Figure 2**, we can find that ICFCDA achieves greater AUC value of 0.946 compared with HGI (0.821), KATZ (0.841), RWR (0.774), NMF (0.776), K-nearest neighbor regression (0.559), support vector regression with rbf kernel (0.497), and support vector regression with poly kernel (0.451), respectively. Moreover, the experiment of collaborative filtering without solving the "cold start" problem is supplemented to evaluate the performance of ICFCDA, which is presented in **Figure 3**. We also make the prediction of other nine common diseases including BC, breast cancer, colorectal cancer, and so on, which are represented in **Figure 4**. In order to illustrate the stability of our proposed computational method, the average AUC values based on the 42 diseases of other methods are shown in **Table 1**. Based on **Figure 2** and **Table 1**, ICFCDA can obtain better and more stable performance than other computational methods. Furthermore, for the sake of obtaining more comprehensive and reliable results, *f*-measure is also treated as one of our evaluating metric, which is described in **Figure 5**. In addition, we also show the first *k* correct circRNA–disease relationships in the predicting results, which is described in **Figure 6**.

Parameters Analysis

In this study, there are three main parameters that are the most similar top *k* neighbors of each circRNA/disease, the circRNA similarity integration adjustment factor α and the disease similarity integration adjustment factor β, respectively. Parameter *k* controls the selecting neighbors' number of each circRNA/disease, which provides the recommendation information from neighbors. The parameter α determines the importance between the circRNA functional annotation semantic similarity and the circRNA sequence similarity, and its value is changed from 0.1 to 0.9. The third parameter β is a tradeoff between the disease functional similarity and the disease semantic similarity, whose value ranges from 0.1 to 0.9. At first, to avoid causing the bias between the circRNA and the disease recommendation scores, the recommendation integration factor γ is set as $N_c/(N_d+N_c)$, where N_c is the number the circRNA entries, and the N_d is the number of the disease entries. At first, for testing the suitable value of the parameter *k*, the parameter α and the parameter β and γ are set up as 0.5, 0.5, and $N_c/(N_d+N_c)$, which means that different disease

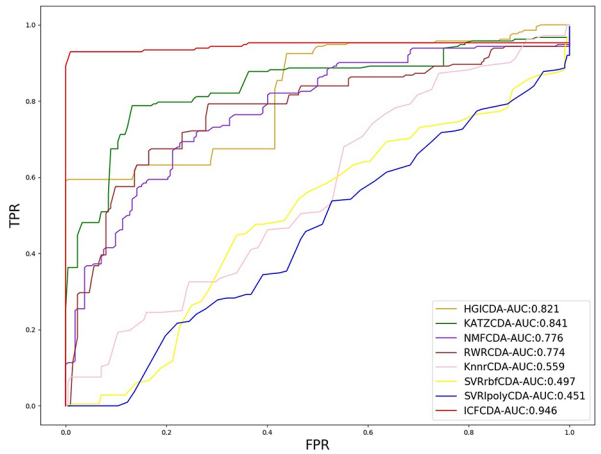


FIGURE 2 | The AUC value of ICFCDA compared with other computational methods.

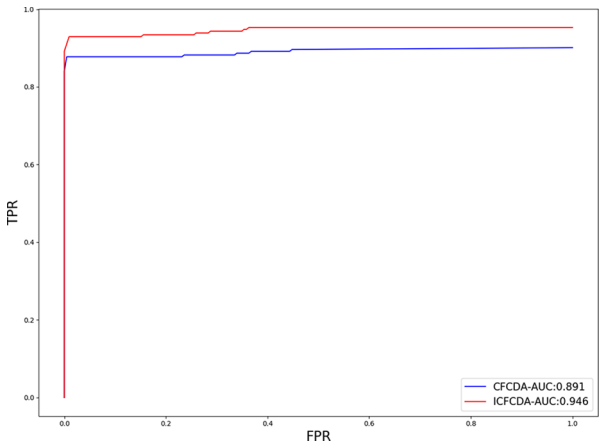


FIGURE 3 | The AUC value of ICFCDA compared with CFCDA without solving the "cold start" problem.

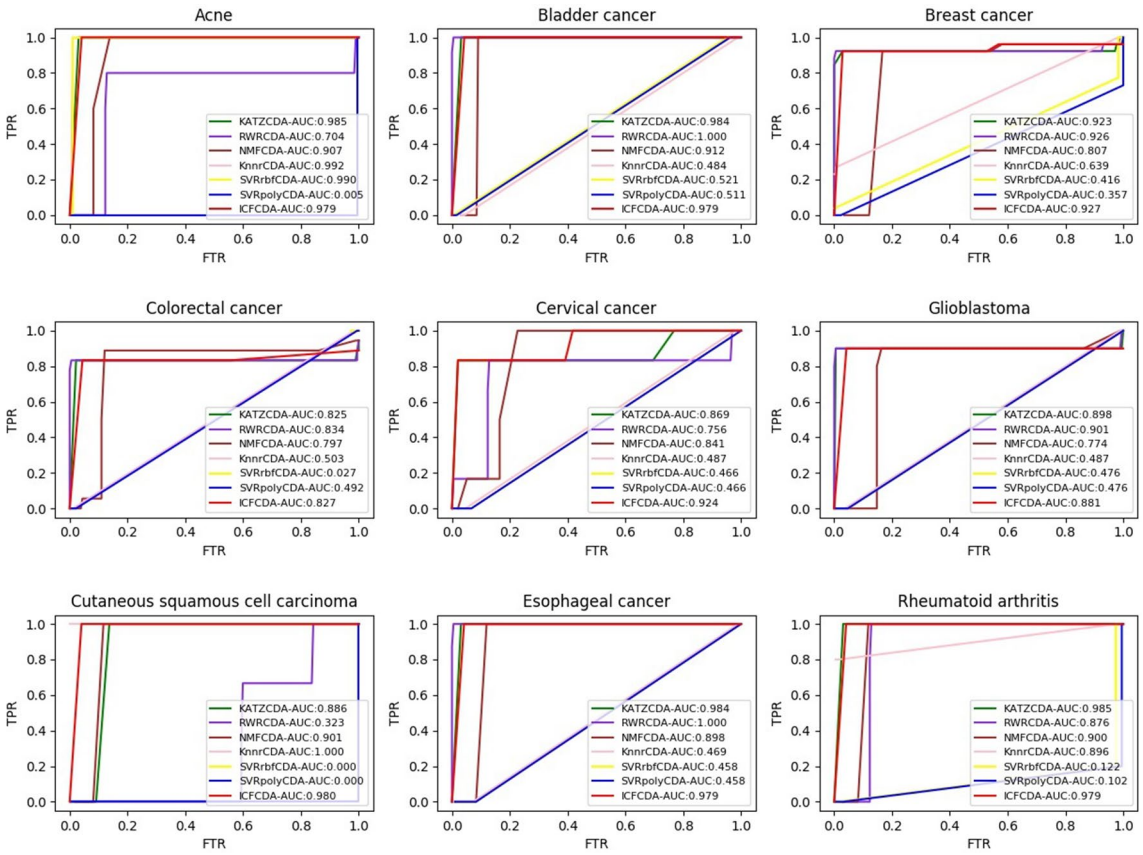


FIGURE 4 | The AUC values of nine kinds of specific diseases.

TABLE 1 | The average AUC values of 42 diseases.

	KATZCDA	RWRCDA	NMFCDA	KNNR	SVRrbf	SVRpoly	ICFCDA
Average AUC	0.719	0.478	0.616	0.536	0.441	0.415	0.885

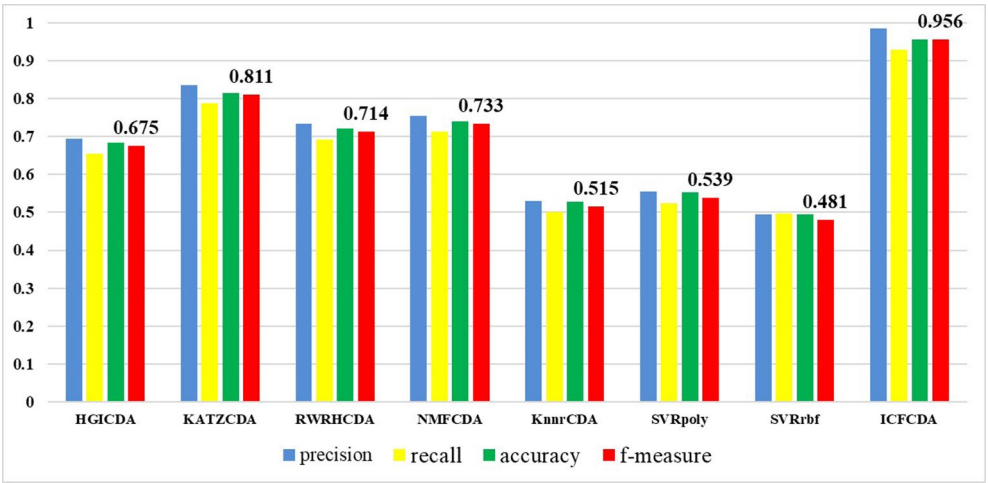


FIGURE 5 | Comparison of the precision, recall, accuracy, and *f*-measure with different methods.

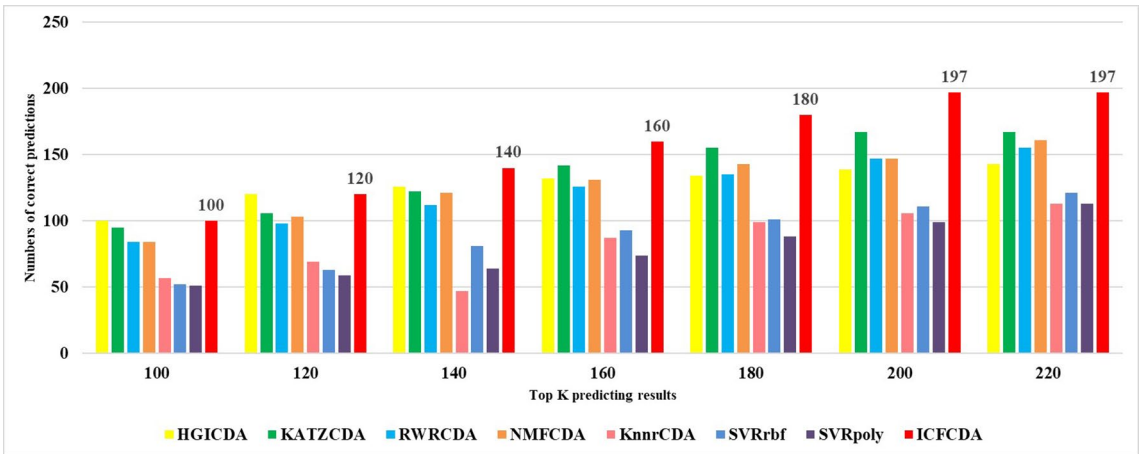


FIGURE 6 | The number of correct circRNA–disease association in top *k* predicting results.

similarity scores are treated equally. According to the above experiments, the parameter α, β, and γ are fixed. When *k* is set as 4%, ICFCDA can obtain the best AUC value (0.946), which is shown in **Table 2**. After that, we can find that the parameter α and β are not sensitive in our computational method according to **Figure 7**. Therefore, both the parameter α and β are set as 0.5.

Case Study

In order to further evaluate the performance of our proposed computational method ICFCDA, we also conduct case studies of two common diseases in the world, which are BC (Kaufman et al., 2009) and breast cancer (Veronesi et al., 2005). Bladder cancer is one of the most common genitourinary malignant diseases, which has caused hundreds of thousands of people's death since it was discovered clinically. What's worse, the risk of BC increases with the increasing age. Another case study is about the breast cancer, which is an important public healthy disease worldwide and is also hard to prevent. Breast cancer has a very high mortality rate. Therefore, some computational methods should be put forward to identify the potential biomarkers of these above two diseases. In this study, the prediction results of ICFCDA are validated by the other three circRNA–disease association–related databases, which are the circ2Disease (Yao et al., 2018), circRNADisease (Zhao et al., 2018), and LncRNADisease v2.0 (Bao et al., 2019),

TABLE 2 | AUC with different values for parameter *k*.

k	1	2	3	4	5	6	7	8	9	10
AUC	0.930	0.932	0.940	**0.946**	0.923	0.921	0.921	0.906	0.906	0.902

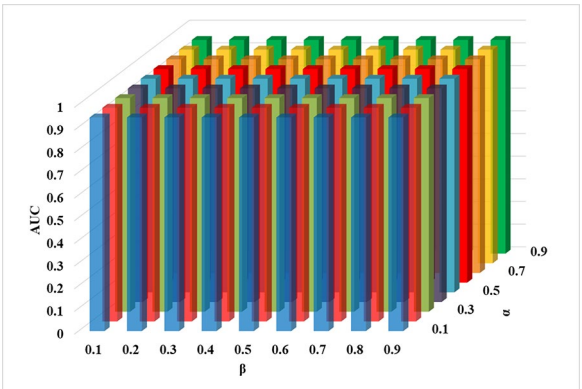


FIGURE 7 | The AUC of the parameter α and β based on the fixed parameter γ and *k*.

which are shown in **Tables 3** and **4**. Both **Tables 3** and **4** are the predicting results of the top 10 BC- and breast cancer–relevant circRNAs. Circ2Disease, circRNADiseaes, and LncRNADisease are represented by *, #, and +, respectively.

CONCLUSION

With the discovery of an increasing numbers of disease-related circRNAs, more and more attention is paid by biologists. People might have lots of interests to explore the complicated associations between the various kinds of diseases and circRNAs. Simultaneously, because of the development of the RNA high-throughput techniques, it makes more convenience to find the potential associations of circRNAs and diseases. While the RNA high-throughput techniques can make this procedure easier than before, it is not only time consuming but also expensive, which becomes the main motivation to develop a computational method to predict the circRNA–disease associations. In this study, we propose a collaborative filtering recommendation system solving the "cold start" problem-based method to predict the circRNA–disease associations, which is named ICFCDA. For evaluating the performance of ICFCDA, LOOCV and *f*-measure show that ICFCDA can obtain better results than other novel computational methods. Moreover, case studies of BC and breast cancer also are adopted to test the performance of the ICFCDA. In terms of the different evaluations, we believe that our proposed computational method is a useful method to predict the associations of the circRNAs and the diseases.

ICFCDA can obtain better performance because of some following nonnegligible reasons. First, our proposed computational method is based on the recommendation system algorithm, collaborative filtering, which is suitable to be used to predict the circRNA–disease associations. Because circRNAs can be treated as the items, and the diseases can be regarded as the users, the different items (circRNAs) can be recommended to different users (diseases). Second, in order to solve the "cold start" problem, the circRNA similarity and the disease similarity are involved to figure out this problem. For obtaining more reliable recommendation information, various kinds of biological data are adopted to measure the circRNA and disease similarity. We download the circRNA-related gene annotation terms to calculate the circRNA functional annotation semantic similarity and the RNA base sequences to calculate the circRNA sequence similarity. Disease-related genes and phenotypes (DO ID) are used to calculate the disease functional and semantic similarity, respectively. Third, in order to screen out more informative information from the noise,

TABLE 3 | The top 10 bladder cancer related candidates' circRNAs.

Rank	CirRNA name/id	Evidences	Rank	CircRNA name/id	Evidences
1	hsa_circ_0000172	+	**6**	hsa_circ_0002024	+
2	hsa_circ_0002495	+	**7**	circMylk/ circRNAMYLK/ hsa_circ_0002768	*, #
3	circRNABCRC4/ hsa_circ_001598/ hsa_circ_0001577	PMID: 29270748	**8**	circTCF25/ hsa_circ_0041103	#
4	hsa_circ_0003221/ circPTK2	#, +	**9**	circFAM169A/ hsa_circ_0007158	#
5	hsa_circ_0091017	#, +	**10**	circTRIM24/ hsa_circ_0082582	#

TABLE 4 | The top 10 breast cancer–related candidates' circRNAs.

Rank	CirRNA name/id	Evidences	Rank	CircRNA name/id	Evidences
1	hsa_circ_0011946	+	**6**	circAmotl1/ hsa_circ_0004214	*, #
2	hsa_circ_0093859	+	**7**	hsa_circ_0006528	*, #, +
3	hsa_circ_0001982	#, +	**8**	hsa_circ_0002874	#, +
4	hsa_circ_0001785	#, +	**9**	hsa_circ_0085495	#, +
5	hsa_circ_0108942	#, +	**10**	hsa_circ_0086241	#, +

we merely use the top 4% most similar neighbors of each circRNA and disease to obtain more reliable recommendation score.

For the future work, more biological data will be added to calculate the disease and the circRNA similarity for reducing the useless noisy information. Adding multiple data can enrich the information of the different biological network, such as circRNA-lncRNA, circRNA-miRNA, and so on.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://bioinfo.snnu.edu.cn/CircR2Disease/article/DownLoad.aspx, http://www.circbase.org/cgi-bin/downloads.cgi, http://www.disgenet.org/downloads, http://www.disease-ontology.org/, http://hprd.org/, https://www.omim.org/.

AUTHOR CONTRIBUTIONS

XL conceptualized the algorithm, designed the method, and drafted the manuscript. ZF designed the method and drafted the manuscript. ZF and LG analyzed the data and carried out the experiments. XL modified the manuscript and polished the English expression.

FUNDING

This work was supported by the funding from National Natural Science Foundation of China (61972451, 61672334, 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

REFERENCES

Arnberg, A. C., Van Ommen, G. J., Grivell, L. A., Van Bruggen, E. F., and Borst, P. (1980). Some yeast mitochondrial RNAs are circular. *Cell* 19, 313–319. doi: 10.1016/0092-8674(80)90505-X

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–d1037. doi: 10.1093/nar/gky905

Bartsch, D., Zirkel, A., and Kurian, L. (2018). Characterization of circular RNAs (circRNA) associated with the translation machinery. *Methods Mol. Biol. (Clifton, N.J.)* 1724, 159–166. doi: 10.1007/978-1-4939-7562-4_13

Chao, C. W., Chan, D. C., Kuo, A., and Leder, P. (1998). The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. *Mol. Med.* 4, 614–628. doi: 10.1007/BF03401761

Chen, C. Y., and Sarnow, P. (1995). Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. *Science* 268, 415–417. doi: 10.1126/science.7536344

Chen, L. L., and Yang, L. (2015). Regulation of circRNA biogenesis. *RNA Biol.* 12, 381–388. doi: 10.1080/15476286.2015.1020271

Chen, X., Han, P., Zhou, T., Guo, X., Song, X., and Li, Y. (2016a). circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Sci. Rep.* 6, 34985. doi: 10.1038/srep34985

Chen, S., Li, T., Zhao, Q., Xiao, B., and Guo, J. (2017b). Using circular RNA hsa_circ_0000190 as a new biomarker in the diagnosis of gastric cancer. *Clin. Chim. Acta* 466, 167–171. doi: 10.1016/j.cca.2017.01.025

Chen, J., Li, Y., Zheng, Q., Bao, C., He, J., Chen, B., et al. (2017a). Circular RNA profile identifies circPVT1 as a proliferative factor and prognostic marker in gastric cancer. *Cancer Lett.* 388, 208–219. doi: 10.1016/j.canlet.2016.12.006

Chen, X., Liu, M. X., and Yan, G. Y. (2012). RWRMDA: predicting novel human microRNA-disease associations. *Mol. Biosyst.* 8, 2792–2798. doi: 10.1039/c2mb25180a

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Huang, Y. A., and Yan, G. Y. (2016b). HGIMDA: heterogeneous graph inference for miRNA-disease association prediction. *Oncotarget* 7, 65257–65269. doi: 10.18632/oncotarget.11251

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2011). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40, 3131–3142. doi: 10.1093/nar/gkr1009

Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web* (Banff, Alberta, Canada: ACM). doi: 10.1145/1242572.1242610

Fan, C., Lei, X., Fang, Z., Jiang, Q., and Wu, F.-X. (2018). CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database* 2018, bay044. doi: 10.1093/database/bay044

Ganegoda, G., Wang, J., Wu, F. X., and Li, M. (2014). Prediction of disease genes using tissue-specified gene-gene network. *BMC Syst. Biol.* 8 Suppl 3, S3. doi: 10.1186/1752-0509-8-S3-S3

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for *de novo* circular RNA identification. *Genome Biol.* 16, 4. doi: 10.1186/s13059-014-0571-3

Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013). Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* 4, 283. doi: 10.3389/fgene.2013.00283

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *RNA* 20, 1666–1670. doi: 10.1261/rna.043687.113

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and Mckusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993

Hsu, M. T., and Coca-Prados, M. (1979). Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* 280, 339–340. doi: 10.1038/280339a0

Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nat. Biotechnol.* 32, 453–461. doi: 10.1038/nbt.2890

Kaufman, D. S., Shipley, W. U., and Feldman, A. S. (2009). Bladder cancer. *Lancet* 374, 239–249. doi: 10.1016/S0140-6736(09)60491-8

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human Protein Reference Database—2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Kibbe, W. A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., et al. (2015). Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* 43, D1071–D1078. doi: 10.1093/nar/gku1011

Kulcheski, F. R., Christoff, A. P., and Margis, R. (2016). Circular RNAs are miRNA sponges and can be used as a new class of biomarker. *J. Biotechnol.* 238, 42–51. doi: 10.1016/j.jbiotec.2016.09.011

Lasda, E., and Parker, R. (2014). Circular RNAs: diversity of form and function. *RNA* 20, 1829–1842. doi: 10.1261/rna.047126.114

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., et al. (2015). Exon–intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 22, 256–264. doi: 10.1038/nsmb.2959

Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., et al. (2017). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.* 46, D106–D112. doi: 10.1093/nar/gkx891

Lin, D. (1998). "An information-theoretic definition of similarity," in *Proceedings of the fifteenth international conference on machine learning* (San Francisco CA, USA: Morgan Kaufmann Publishers Inc.).

Lin, F.-M., Hsu, S.-D., Liu, Y.-C., Huang, H.-D., Sun, C.-H., Li, J.-R., et al. (2015). CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.* 44, D209–D215. doi: 10.1093/nar/gkv940

Liu, Y., Wang, S. L., and Zhang, J. F. (2018). Prediction of microbe–disease associations by graph regularized non-negative matrix factorization. *J. Comput. Biol.* 25. doi: 10.1089/cmb.2018.0072

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Pan, R., Zhou, Y., Cao, B., Liu, N. N., Lukose, R., Scholz, M., et al. (2008). "One-class collaborative filtering," in *Proceedings of the 2008 Eighth IEEE international conference on data mining* (Washington DC, USA: IEEE Computer Society). doi: 10.1109/ICDM.2008.16

Panda, A. C., De, S., Grammatikakis, I., Munk, R., Yang, X., Piao, Y., et al. (2017). High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs. *Nucleic Acids Res.* 45, e116. doi: 10.1093/nar/gkx297

Pasman, Z., Been, M. D., and Garcia-Blanco, M. A. (1996). Exon circularization in mammalian nuclear extracts. *RNA* (Berlin Heidelberg: Springer), 2, 603–610.

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–d839. doi: 10.1093/nar/gkw943

Qu, S., Yang, X., Li, X., Wang, J., Gao, Y., Shang, R., et al. (2015). Circular RNA: a new star of noncoding RNAs. *Cancer Lett.* 365, 141–148. doi: 10.1016/j.canlet.2015.06.003

Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc. Natl. Acad. Sci. U.S.A.* 73, 3852–3856. doi: 10.1073/pnas.73.11.3852

Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S. (2007). "Collaborative filtering recommender systems," in *The adaptive web*. Eds. B. Peter, K. Alfred, and N. Wolfgang (Springer Berlin Heidelberg: Springer-Verlag), 291–324. doi: 10.1007/978-3-540-72079-9_9

Van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

Veronesi, U., Boyle, P., Goldhirsch, A., Orecchia, R., and Viale, G. (2005). Breast cancer. *Lancet* 365, 1727–1741. doi: 10.1016/S0140-6736(05)66546-4

Wilusz, J. E., and Sharp, P. A. (2013). Molecular biology. A circuitous route to noncoding RNA. *Science (New York, N.Y.)* 340, 440–441. doi: 10.1126/science.1238522

Xu, Z., Yan, Y., Zeng, S., Dai, S., Chen, X., Wei, J., et al. (2017). Circular RNAs: clinical relevance in cancer. *Oncotarget* 9, 1444–1460. doi: 10.18632/oncotarget.22846

Yao, D., Zhang, L., Zheng, M., Sun, X., Lu, Y., and Liu, P. (2018). Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci. Rep.* 8, 11018–11018. doi: 10.1038/s41598-018-29360-3

Yu, G., Wang, L. G., Yan, G. R., and He, Q. Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Zeng, Y., Xu, Y., Shu, R., Sun, L., Tian, Y., Shi, C., et al. (2017). Altered expression profiles of circular RNA in colorectal cancer tissues from patients with lung metastasis. *Int. J. Mol. Med.* 40, 1818–1828. doi: 10.3892/ijmm.2017.3189

Zhao, Z., Li, X., Jian, D., Hao, P., Rao, L., and Li, M. (2017). Hsa_circ_0054633 in peripheral blood can be used as a diagnostic biomarker of pre-diabetes and type 2 diabetes mellitus. *Acta Diabetol.* 54, 237–245. doi: 10.1007/s00592-016-0943-0

Zhao, Z., Wang, K., Wu, F., Wang, W., Zhang, K., Hu, H., et al. (2018). circRNA disease: a manually curated database of experimentally supported circRNA–disease associations. *Cell Death Dis.* 9, 475–475. doi: 10.1038/s41419-018-0503-3

Zhong, Z., Lv, M., and Chen, J. (2016). Screening differential circular RNA expression profiles reveals the regulatory role of circTCF25-miR-103a-3p/miR-107-CDK6 pathway in bladder carcinoma. *Sci. Rep.* 6, 30919. doi: 10.1038/srep30919

Zhou, Y., Song, B., and Zheng, H.-T. (2015). *Exploiting latent relations between users and items for collaborative filtering*. Springer Nature Switzerland AG: Springer International Publishing, 365–374. doi: 10.1007/978-3-319-26555-1_41

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). *Large-scale parallel collaborative filtering for the Netflix prize*. Springer Berlin Heidelberg, 337–348. doi: 10.1007/978-3-540-68880-8_32

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 09 October 2019
doi: 10.3389/fgene.2019.00937

Deep Learning Enables Accurate Prediction of Interplay Between lncRNA and Disease

Jialu Hu[1,2,3], *Yiqun Gao*[1], *Jing Li*[4] *and Xuequn Shang*[1,3]*

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [2] Centre for Multidisciplinary Convergence Computing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [3] Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Xi'an, China, [4] Ming De College, Northwestern Polytechnical University, Xi'an, China

Many studies have suggested that lncRNAs are involved in distinct and diverse biological processes. The mutation of lncRNAs plays a major role in a wide range of diseases. A comprehensive information of lncRNA-disease associations would improve our understanding of the underlying molecular mechanism that can explain the development of disease. However, the discovery of the relationship between lncRNA and disease in biological experiment is costly and time-consuming. Although many computational algorithms have been proposed in the last decade, there still exists much room to improve because of diverse computational limitations. In this paper, we proposed a deep-learning framework, NNLDA, to predict potential lncRNA-disease associations. We compared it with other two widely-used algorithms on a network with 205,959 interactions between 19,166 lncRNAs and 529 diseases. Results show that NNLDA outperforms other existing algorithm in the prediction of lncRNA-disease association. Additionally, NNLDA can be easily applied to large-scale datasets using the technique of mini-batch stochastic gradient descent. To our best knowledge, NNLDA is the first algorithm that uses deep neural networks to predict lncRNA-disease association. The source code of NNLDA can be freely accessed at https://github.com/gao793583308/NNLDA.

Keywords: lncRNA, neural network, large dataset, non-linear, computational model

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Lei Deng,
Central South University, China
Liang Cheng,
Harbin Medical University, China

***Correspondence:**
Xuequn Shang
shang@nwpu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 July 2019*
Accepted: *05 September 2019*
Published: *09 October 2019*

Citation:
Hu J, Gao Y, Li J and Shang X (2019) Deep Learning Enables Accurate Prediction of Interplay Between lncRNA and Disease. Front. Genet. 10:937. doi: 10.3389/fgene.2019.00937

INTRODUCTION

There are about 30,000–40,000 protein-coding genes in the human genome, which are only about twice as many as in worm or fly (Lander et al., 2001). But the majority of the human genome transcripts are non-coding RNAs, in particular, long non-coding RNAs (lncRNAs) (Geng et al., 2013). Protein-coding genes account for only 1.5% of the human genome. However, researchers observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts respectively (Djebali et al., 2012). This suggests that lncRNA also plays an important role in biological processes. Recent studies revealed that numerous sets of non-coding RNA involved in distinct and diverse biological processes, such as cell proliferation, RNA binding complexes, immune surveillance, ESC pluripotency, neuronal processes, morphogenesis, gametogenesis, and muscle development (Mitchell et al., 2009). Furthermore, some important lncRNA biomarkers were found in a wide range of human diseases. For example, the expression of HOTAIR would induce androgen-independent (AR) activation, which plays a central role in

establishing an oncogenic cascade that drives prostate cancer progression. It can also drive AR-mediated transcriptional programs in the absence of androgen (Zhang et al., 2015). So finding the relationship between lncRNA and disease can not only help us understand the mechanism of disease, but also accelerate the discovery of biomarker. However, discovering the potential relationship between lncRNA and disease by experimental ways are costly and time-consuming. Thus, many computational models have been proposed to predict potential connection patterns by utilizing existing data such as LncRNADisease (Geng et al., 2013), LncRNAdb (Cheng et al., 2015), and NONE-CODE (Cheng et al., 2015).

The existing computational models can be divided into two categories. The first class of methods make predictions based on the similarity of artificial definitions. It assumed that similar diseases or lncRNA have similar connection patterns. Take a simple example, if we know that disease(i) is related to lncRNA(i) and disease(i) and disease(j) are very similar. It's obvious that we can infer that disease(j) and lncRNA(i) are also related. This algorithm needs to collect a lot of additional data to accurately define similarity. If the definition of similarity is accurate, the algorithm can achieve high performance. For example, LncRDNetFlow utilizes a flow propagation algorithm to integrate multiple networks based on a variety of biological information including lncRNA similarity, protein-protein interactions, disease similarity, and the associations between them to infer lncRNA-disease associations (Zhang et al., 2017a). IRWRLDA construct lncRNA expression similarity and lncRNA functional similarity to make prediction (Chen et al., 2016). RWRlncD infer potential human lncRNAdisease associations by implementing the random walk with restart method on a lncRNA functional similarity network (Sun et al., 2014). BiWalkLDA integrating interaction profiles and gene ontology information to construct similarity network. Such an algorithm also has KATZLGO (Zhang et al., 2017b) and IDHI-MIRW (Fan et al., 2019). It can be seen that this algorithm first constructs the similarity network based on the relevant data and then making prediction according to the constructed similarity. The second class of methods make predictions based on matrix factorization (MF). Their core idea is to learn a similarity rather than artificial definition similarity. This actually turns the prediction process into a classification question. For each lncRNA and disease, the aim of MF is to learn a latent factor to represent them and then make prediction based on learned latent factors. In this way, no additional knowledge is needed to define similarity. This method is widely used in prediction lncRNA-disease association. For example, the algorithm of MFLDA decomposes data matrices of heterogeneous data sources into low-rank matrices *via* matrix tri-factorization to explore and exploit their intrinsic and shared structure (Fu et al., 2017). SIMCLDA models the lncRNA-disease association prediction problem as a recommendation task and solves it with inductive matrix completion (IMC) (Lu et al., 2018).

The known lncRNA-disease association data used by current algorithms is derived from LncRNADisease (Geng et al., 2013). This database was proposed in 2013 and does not contain much lncRNA and disease (almost 300 lncRNA and 700 diseases). Because the data is relatively small, even though the existing prediction algorithms can achieve high accuracy, many results are repetitive and therefore cannot provide more valuable results. Fortunately, recently, a larger dataset LncRNADisease 2.0 can be used (Bao et al., 2019). LncRNADisease 2.0 curated 19,166 lncRNAs, 823 circRNAs, and 529 diseases from 3878 literatures. Although the form of data remains unchanged, only the increase in the amount of data makes previous algorithms not applicable to LncRNADisease 2.0. For methods that need to artificially define similarity, it is difficult to collect the additional information needed comprehensively in the face of such large data. So, it is difficult to define an appropriate similarity for prediction. For the method based on MF, the time cost of the algorithm is unacceptable with the increase of data. Besides, MF is actually a linear model of latent factors, so it cannot describe more complex relational patterns well (He et al., 2017). As we all know, deep learning can be applied to large-scale data and learn complex non-linear relationships by means of mini-batch stochastic gradient descentand and nonlinear activation function. In recent years, deep neural networks have yielded immense success on object detection (Ren et al., 2017), recommendation System (Zhou et al., 2017), single cell denoising (Eraslan and Simon, 2019; Peng et al., 2019), and many other fields. However, no deep learning-based algorithm has been proposed to predict potential lncRNA-disease association. In this article, we will introduce our proposed framework NNLDA which uses neural networks to predict lncRNA-disease association. To our best knowledge, NNLDA is the first algorithm that uses deep neural networks to predict lncRNA-disease association. Experiments show that NNLDA can be well applied to large data and to learn more complex non-linear relationships.

METHOD

Our prediction framework NNLDA is improved based on the MF method. In this section, I will first introduce the method of MF and point out its shortcomings. Then, we will explain how we solve these shortcomings and introduce the procedure of NNLDA in detail.

Matrix Factorization (MF)

MF is a frequently used method in the problem of predicting lncRNA-disease association (Fu et al., 2017; Lu et al., 2018). Its core idea is to learn a corresponding latent factor for each lncRNA and disease. The dot product of the latent factor was used to represent the possible score of corresponding lncRNA and disease. Take the prediction of lncRNA-disease association, for example. First, we should construct an adjacency matrix $A_{nl \times nd}$, where n_l is the number of lncRNA and n_d is the number of diseases. $A_{ij} = 1$ represents that the i^{th} lncRNA is associated with d_j, otherwise, $A_{ij} = 0$. Then, we assign a k-dimensional latent factor L(i) for each lncRNA(i) and a k-dimensional latent factor D(i) for each disease(i). These latent factors are usually randomly initialized at the beginning and then be adjusted by some optimization algorithm such as stochastic gradient descent. Now, we can use the dot product of the latent factor to re-estimate A.

For each pair of lncRNA(i) and disease(j), we predict its association using $\hat{A}_{ij} = \sum_{n=1}^{k} L_{in} D_{nj}$. Our objective funcition is to minimize the following loss function:

$$Loss = \sum_{i=1}^{n_l} \sum_{j=1}^{n_d} (A_{ij} - \hat{A}_{ij})$$

A new L and D can be learned by minimizing loss. This loss is actually equivalent to $Loss = || A - LD ||_F^2$, which was frequently used in other literatures because the dot product of vectors can be seen as the angle of vectors in space ($a \cdot b = |a||b|cos(a,b)$). So, matrix factorization actually maps each lncRNA and disease into k-dimensional space and then defines the relationship between lncRNA and disease by using the length and angle of the latent factor. However, there are several shortcomings in doing so: (1) There are limitations in utilizing the angle between latent factor to define the relationship between lncRNA and disease. Take two-dimensional space as an example, suppose we now learn three latent factors: $a_1(1,0)$, $a_2(0,1)$, $a_3(1,1)$, if we also have latent factor a_4 and we want the angle between a_4 and a_1, a_2 to be as small as possible, but the angle between a_2 and a_3 to be as large as possible. Obviously, no matter where a_4 is, it can't be satisfied. Of course, we can describe this relationship by adding spatial dimensions, but the increase of k actually increases the risk of over-fitting. It can be concluded that angle can't actually describe some complex relationship patterns perfectly. (2) The time complexity of matrix decomposition is too high. When calculating the loss, it needs to calculate all possible connections between lncRNA and disease. As the amount of data increases, the time required is unacceptable. Besides directly optimizing, global loss is easy to fall into local minima.

Making Matrix Factorization Applicable to Large Data

In order to make the matrix factorization method suitable for large-scale data, we made some improvements to the original method and implemented the method with tensorflow. We named this method NNMF, which is different from the traditional MF method in two aspects:

(1) Unlike previous MF, full data is used to minimize loss. We adopt mini-batch stochastic gradient descent to train model. This means that we use only one batch data per round to minimize loss, which makes our algorithm suitable for large-scale data.
(2) The traditional matrix factorization uses mean square error or absolute value error to measure loss. Its goal is to $min || A - LD ||_F^2$. In NNMF, we use cross-entropy as our loss function, which is proved to be more applicable to classification problems and easier to optimize.

With above two improvements, NNMF can be adapted to large-scale data. The structure of the network and an example of computational processes are shown in **Figure 1**. NNMF takes lncRNA(i) and disease(j) as its input and outputs the probability of the relationship between lncRNA(i) and disease(j). First, the network generates a dense latent factor for corresponding lncRNA(i) and disease(j). This operation is done by embedding lookup function in tensorflow. Then, the corresponding position elements of the two vectors are multiplied and summed. Sigmoid activation functions are added to limit output to between 0 and 1. With the predicted results, we can calculate the cross-entropy loss to adjust the corresponding latent factor. To avoid storing the whole data set into memory each time we take a batch data to train, the batch size is set to 1,024. This process is repeated until the loss is no longer reduced. NNMF changes the way of training and the loss function compared with the traditional matrix decomposition algorithm. With these small changes, NNMF can be adapted to large-scale data easily.

Learning More Complex Relationships by Using Full Connectivity Layer

Matrix factorization actually maps lncRNA and disease into k-dimensional space, and then measures their relationship by using dot product of latent factors. This approach undoubtedly has its limitations. In order to learn more complex non-linear features, a natural idea is to use the full connection layer of the

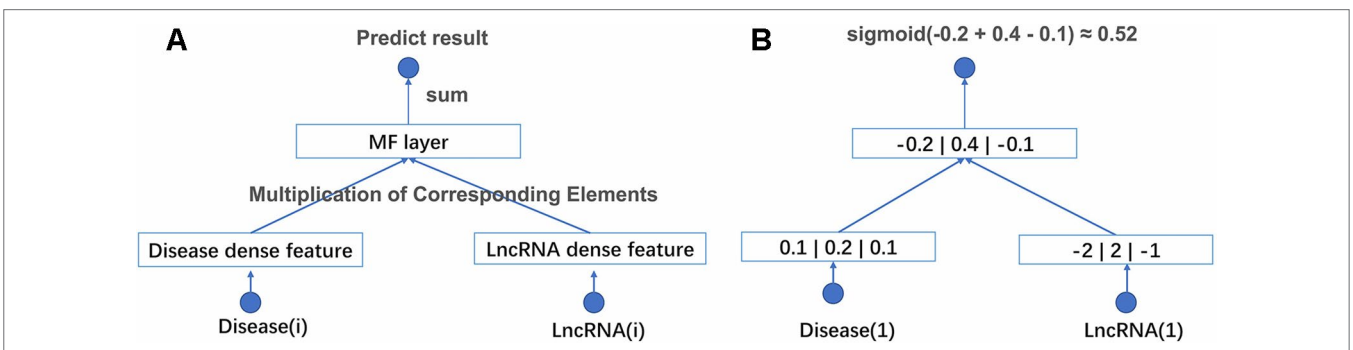


FIGURE 1 | (A) The structure of NNMF. Each lncRNA and disease is projected into a k-dimensional space. It means each lncRNA and disease would be represented by a corresponding k*1 eigenvector. The relationship between lncRNA and disease is measured by the dot product of their corresponding eigenvector. The activation function is sigmoid. **(B)** A toy example of the NNMF, where k is set to 3.

neural network to improve it. Similar to the NNMF process, we initialize a latent factor for each lncRNA and disease at the beginning. Then, we concatenated the latent factors and add full connection layers to learn more complex relationships. RelU activation function is used on each full connection layer to increase the non-linear description ability of the network. Sigmoid activation functions are added to limit output to between 0 and 1. Considering that using full connection layer alone may increase the risk of over-fitting. We adopt the following two strategies to prevent over-fitting:

(1) Add L2 regularization to latent factors and full connection layer to limit models from learning too complex features.
(2) The deep part is trained together with NNMF. In this way, we cannot only learn more diverse connection relationships, but also improve the generalization ability of the model.

We name this new model NNLDA. It means predicting lncRNA-disease association by means of neural networks. The overall structure of NNLDA is shown in **Figure 2**. First, for each lncRNA and disease, we will find their corresponding latent factors. MF part multiplies the corresponding elements of latent factors and deep part use several full-connection layers to learn the complex relationship between lncRNA and disease. Their results are concatenated together and connected to a full connection layer for final prediction. Sigmoid activation function is added to limit output to between 0 and 1. NNLDA learns more complex relational patterns by combining dot product of latent factors and full connectivity layer. Because NNLDA uses mini-batch stochastic gradient descent to minimize loss, it can also be well applied to large-scale data. We believe that NNLDA can perfectly solve the shortcomings of traditional MF methods.

Implementation

NNLDA is implemented in Python 3.5 and uses TensorFlow1.12.0. Length of latent factor is set to 32. Three full-connection layers with lengths of 32, 16 and 8 are added in deep part. L2 regulation is added in all full-connection layers and latent factors to prevent over-fitting and regulation rate is set to 0.01. We use adam for optimization with learning rate 0.01. Epoch is set to 100 and batch size is set to 1024.

EXPERIMENT

Dataset

Unlike previous algorithms which usually perform on small data sets such as LncRNADisease database, we use LncRNADisease 2.0 to measure the results of the algorithm. LncRNADisease 2.0 shows that there exists 205,959 interactions between 19,166 lncRNAs and 529 diseases. We believe that more valuable results can be found by using larger data. Such large-scale data also challenges previous algorithms. The experimental data can be downloaded from http://www.rnanut.net/lncrnadisease/. We remove all repeating records with the same lncRNA and disease, and all these non-human associations. Finally, we retained 187,55 lncRNA and 463 disease with 177,899 associations.

10-Fold Cross Validation

To test the algorithm performance, we employed a widely-used strategy, 10-fold cross validation. Known lncRNA and disease associations are divided into 10 copies. In each round, nine of them are used to train algorithms and the remaining one is used as a test set. Notice that we need negative samples to train the algorithm, but in fact we don't know which lncRNAs are not associated with diseases. So, for each known LncRNA-disease,

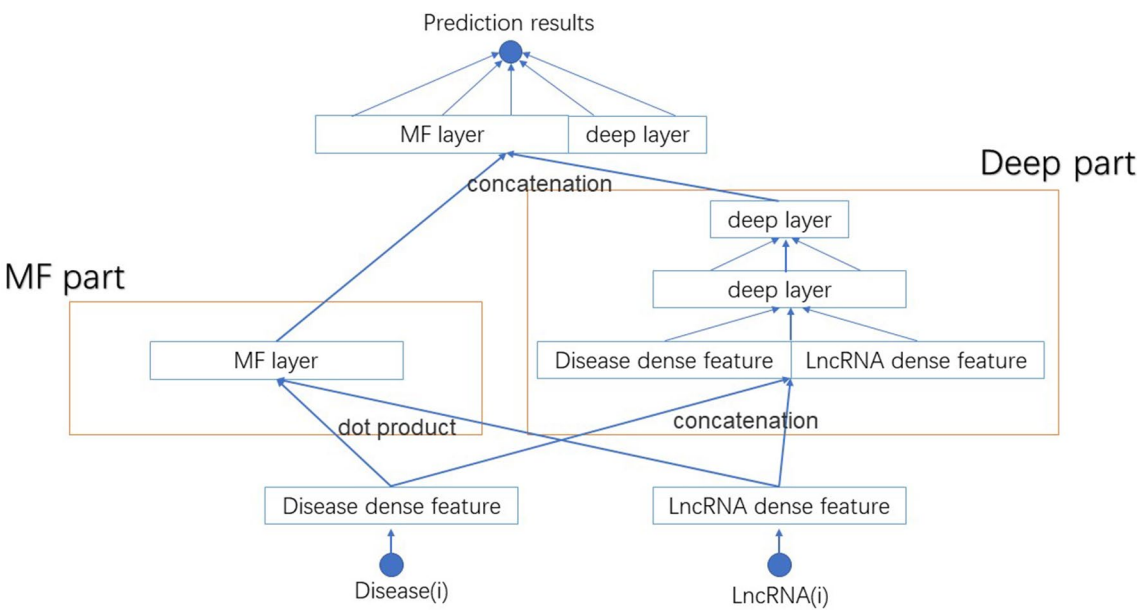


FIGURE 2 | The structure of NNLDA. MF part is same as NNMF. Deep part use several full connection layers to learn complex association relationships. Their results are concatenated together to make final predictions.

we will randomly sample four lncRNA that do not interact by this disease as negative samples. When predicting test sets, we no longer use AUC as an evaluation criterion. This is because AUC needs to compute all possible associations. This means that if there are n lncRNA and m disease, we need to calculate n*m possible cases and then generate a rank list. It's obvious that it's unrealistic when the data set is large. So we adopt a new evaluation strategy. For each test sample, we will sample 99 random lncRNA that not interact by this disease. The model scores 99 negative samples and one positive sample to generate the corresponding rank list. Then, we use Hit Ratio (HR) to assessment results. The HR intuitively measures whether the test item is present on the top-k rank list and we can interpret HR (K) as the probability of positive samples appearing in top-k rank list. If the test sample is in the first k of rank list, its value is plus one. The hit rate value can be obtained by dividing the final hit value by the number of test samples. The higher the hit rate, the higher the likelihood that true sample will appear in the top-k rank list.

The Effects of Parameters

Length of Latent Factors

In the first step of NNMF and NNLDA, both lncRNA and disease need to be mapped into a k-dimensional vector. This vector is called latent factors. Here, k is an artificially defined parameter and represents the dimension of feature space. If the value of k is very small, the model cannot learn complex relationships. If the value of k is big, the risk of over-fitting of the model increases. In order to test possible effects on the performance of the algorithm under different value of k, we changed the value of k in 8, 16, 32, 64, and 128 each time, and then calculated the HR 10. Because KNN does not use latent factors, we only compared NNMF and NNLDA here. The experimental results show in **Figure 3**. The result shows that the length of latent factors don't actually have much impact on the hit ration. This is because we added L2 regularization to latent factors. Even if the length of latent factors increases, it will not be over-fitting data. If no regularization is added, the loss of the model decreases rapidly and over-fitting will occur soon.

Number of Layers

We used several full-connection layers in deep part to learn more complex relationships. More layers can theoretically learn more complex models, which also increases the risk of over-fitting. In order to test the possible effect of number of layers on the performance of the algorithm. We changed the number of layers in 1-layer (32), two-layer (32 and 16), three-layer (32, 16, and 8) and four-layer (32, 16, 8, and 4), and calculate the hit ration value. The experimental results are shown in **Figure 4**. It can be seen that increasing the number of layers of the network will not greatly improve the effectiveness of the algorithm. Algorithm performance is poor when the number of layers is 4. This shows that even if we use L2 regularization to prevent over-fitting, the number of layers of the network should not be too big.

Comparison With Other Algorithms

Because we use LncRNADisease 2.0 to compare the performance of our algorithm. Traditional algorithms cannot be applied to such large dataset. So, although many computational models have been proposed, they cannot be used for comparison. We have made some changes to the traditional algorithm. NNMF can be seen as a matrix factorization algorithm suitable for large-scale data. For algorithms that need to define similarity artificially, we implement an algorithm manually based on the idea of KNN. The specific process is as follows: First, we calculate the gauss similarity between diseases which is widely used in other papers. Then for each disease, we will find 40 diseases that are most similar to it and use their average interaction profile to make predictions.

We compare NNLDA with other two computational methods (NNMF and KNN) of lncRNA-disease association prediction in terms of HR. All algorithms use the same data to make predictions. The experimental results are shown in **Figure 5**. It can be seen that the performance of KNN is very poor. This is

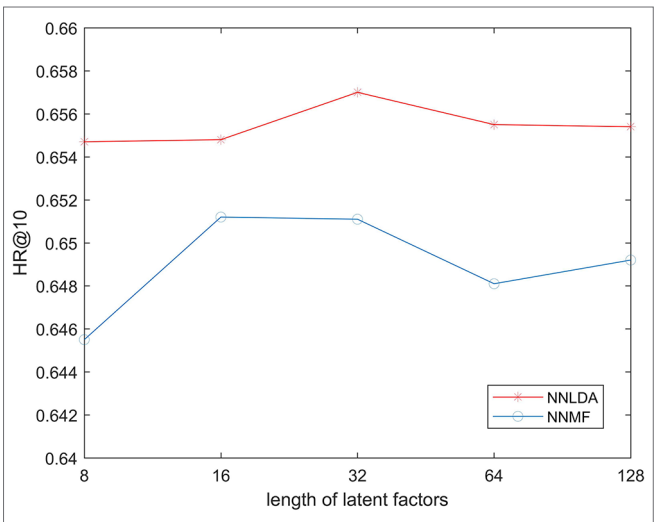


FIGURE 3 | HR @ k Three Algorithms under Different value of k.

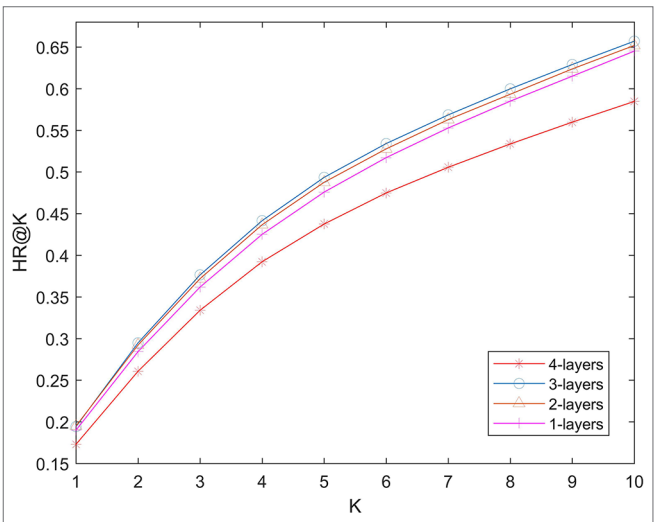


FIGURE 4 | Effects of lengths of latent factors.

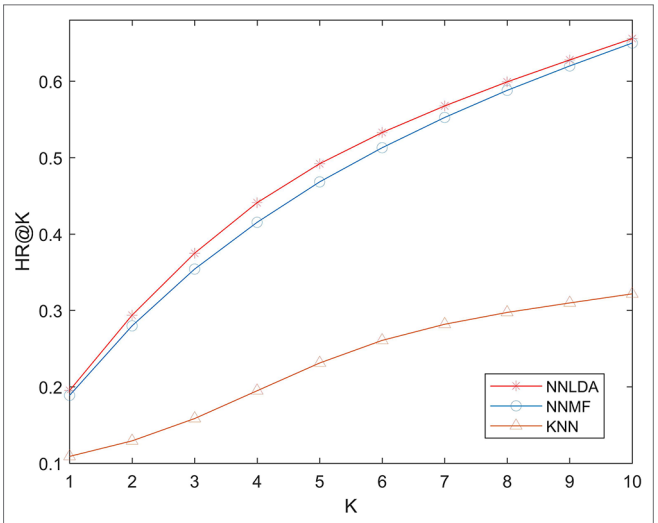


FIGURE 5 | Effects of lengths of latent factors.

because similarity-based algorithms need to artificially define the similarity between diseases and then make predictions based on similarity. As the amount of data increases, additional data becomes more and more difficult to obtain. Because of this, it is difficult to define an accurate and reasonable similarity. So, the performance of this algorithm is limited by similarity. Comparing NNLDA and NNMF, we can find that NNLDA outperforms NNMF in all k values. In fact, NNLDA can be seen as model fusion of NNMF and full connectivity layer. This shows that more complex connection relationships can be learned by using the full-connection layer.

CONCLUSION

Many recent studies suggest that lncRNAs are strongly associated with various complex human diseases. Therefore, the discovery of the potential association between lncRNA and diseases helps to understand the biological processes and underlying mechanisms of diseases. Many prediction algorithms have been proposed to predict lncRNA-disease association. Although the algorithm can achieve high accuracy, traditional prediction algorithms can no longer be applied to more and more large-scale data. In this paper, we propose NNLDA to predict lncRNA-disease association. NNLDA uses mini-batch stochastic gradient descent and cross-entropy loss to enable the algorithm to be applied to large-data sets and use full-connection layer to make up for the deficiency of MF expression ability. Our contributions can be summarized as follows: 1) NNLDA is the first algorithm can predict lncRNA-disease association on large datasets. 2) NNLDA is the first algorithm to use neural network to predict potential lncRNA-disease association. Compared with traditional MF algorithm, NNMF can better describe their relationship by using full-connection layer. In the experimental part, we compare NNLDA, KNN, and NNMF. The experimental results show that NNLDA performs better in terms of hit rate on LncRNADisease 2.0 database. The experiment of parameter influence shows that NNLDA is robust to different parameter setting.

DATA AVAILABILITY STATEMENT

All datasets generated/analyzed for this study are included in the manuscript/supplementary files.

AUTHOR CONTRIBUTIONS

JH designed the computational framework, JH and YG implemented the program. YG and JL performed all the analyses of the data and wrote the manuscript. XS is the major coordinator, who contributed a lot of time and efforts in the discussion of this project. All authors read and approved the final manuscript.

FUNDING

Publication costs were funded by the National Natural Science Foundation of China (Grant No. 61702420). This project has also been funded by the National Natural Science Foundation of China (Grant No. 61332014, 61702420 and 61772426); the China Postdoctoral Science Foundation (Grant No. 2017M613203); the Natural Science Foundation of Shaanxi Province (Grant No. 2017JQ6037); the Fundamental Research Funds for the Central Universities (Grant No. 3102018zy032); and the Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

REFERENCES

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNAdisease 2.0: an updated data base of long non-coding rna-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi: 10.1093/nar/gky905

Chen, X., You, Z. H., Yan, G. Y., and Gong, D. W. (2016). Irwrlda: improved random walk with restart for lncRNA-disease association prediction. *Oncotarget* 7, 57919–57931. doi: 10.18632/oncotarget.11141

Cheng, Q. X., Thomson, D. W., Maag, J. L. V., Nenad, B., Bethany, S., Clark, M. B., et al. (2015). LncRNAdb v2.0: expanding the reference database for functional long noncoding rnas. *Nucleic Acids Res.* 43, 168–173. doi: 10.1093/nar/gku988

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A. M., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233

Eraslan, G., and Simon, L. M. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nat. Commun.* 10, 390. doi: 10.1038/s41467-018-07931-2

Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., Zhu, K., and Lu, S. (2019). Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with rwr algorithm and positive pointwise mutual information. *BMC Bioinf.* 20. doi: 10.1186/s12859-019-2675-y

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2017). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Geng, C., Wang, Z., Wang, D., Qiu, C., Liu, M., Xing, C., et al. (2013). Lncrnadisease: a database for long-non-coding rna-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

He, X., Liao, L., Zhang, H., Nie, L., Xia, H., and Chua, T. S. (2017). Neural collaborative filtering. in *Proceedings of the 26th International Conference on World Wide Web*. (Perth, Australia), 173–182. https://dl.acm.org/citation.cfm?doid=3038912.3052569

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lu, C., Yang, M., Luo, F., Wu, F. X., Li, M., Pan, Y., et al. (2018). Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics*. 34, 3357–3364. doi: 10.1093/bioinformatics/bty327

Mitchell, G., Ido, A., Manuel, G., Courtney, F., Lin, M. F., David, F., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature* 458, 223. doi: 10.1038/nature07672

Peng, J., Wang, X., and Shang, X. (2019). Combining gene ontology with deep neural networks to enhance the clustering of single cell rna-seq data. *BMC Bioinf.* 20, 284. doi: 10.1186/s12859-019-2769-6

Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* 10, 2074–2081. doi: 10.1039/C3MB70608G

Zhang, A., Zhao, J. C., Kim, J., Fong, K. W., Yang, Y. A., Chakravarti, D., et al. (2015). LncRNA hotair enhances the androgen-receptor-mediated transcriptional program and drives castration-resistant prostate cancer. *Cell Rep.* 13, 209–221. doi: 10.1016/j.celrep.2015.08.069

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2017a). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE/ACM Trans. Comp. Biol. Bioinf.* PP, 1–1. doi: 10.1109/TCBB.2017.2701379

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2017b). Katzlgo: large-scale prediction of lncRNA functions by using the katz measure based on multiple networks. *IEEE/ACM Trans. Comp. Biol. Bioinf.* PP, 1–1. doi: 10.1109/TCBB.2017.2704587

Zhou, G., Song, C., Zhu, X., Fan, Y., Zhu, H., Ma, X. et al. (2017). Deep interest network for click-through rate prediction. in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. (London, United Kingdom) 1059–1068. https://dl.acm.org/citation.cfm?doid=3219819.3219823

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fbioe.2019.00305

Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams

Nguyen Quoc Khanh Le[1]*, *Edward Kien Yee Yapp*[2], *N. Nagasundaram*[3] and *Hui-Yuan Yeh*[3]*

[1] Professional Master Program in Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan, [2] Singapore Institute of Manufacturing Technology, Innovis, Singapore, Singapore, [3] Medical Humanities Research Cluster, School of Humanities, Nanyang Technological University, Singapore, Singapore

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Xue Xu,
Harvard Medical School, United States
Mahashweta Basu,
United States Food and Drug Administration, United States

***Correspondence:**
Nguyen Quoc Khanh Le
khanhlee@tmu.edu.tw
Hui-Yuan Yeh
hyyeh@ntu.edu.sg

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *13 August 2019*
Accepted: *17 October 2019*
Published: *05 November 2019*

Citation:
Le NQK, Yapp EKY, Nagasundaram N and Yeh H-Y (2019) Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. Front. Bioeng. Biotechnol. 7:305. doi: 10.3389/fbioe.2019.00305

A promoter is a short region of DNA (100–1,000 bp) where transcription of a gene by RNA polymerase begins. It is typically located directly upstream or at the 5′ end of the transcription initiation site. DNA promoter has been proven to be the primary cause of many human diseases, especially diabetes, cancer, or Huntington's disease. Therefore, classifying promoters has become an interesting problem and it has attracted the attention of a lot of researchers in the bioinformatics field. There were a variety of studies conducted to resolve this problem, however, their performance results still require further improvement. In this study, we will present an innovative approach by interpreting DNA sequences as a combination of continuous FastText N-grams, which are then fed into a deep neural network in order to classify them. Our approach is able to attain a cross-validation accuracy of 85.41 and 73.1% in the two layers, respectively. Our results outperformed the state-of-the-art methods on the same dataset, especially in the second layer (strength classification). Throughout this study, promoter regions could be identified with high accuracy and it provides analysis for further biological research as well as precision medicine. In addition, this study opens new paths for the natural language processing application in omics data in general and DNA sequences in particular.

Keywords: DNA promoter, transcription factor, word embedding, convolutional neural network, natural language processing, precision medicine

INTRODUCTION

A promoter is a region of DNA where RNA polymerase begins to transcribe a gene. Normally, promoter sequences are typically located directly upstream or at the 5′ end of the transcription initiation site (Lin et al., 2018). Both promoters and transcription initiation sites are bound by RNA polymerase and the necessary transcription factors. Promoter sequences describe the direction of transcription and point out which DNA strand will be transcribed (known as sense strand). The transcription process is shown in **Figure 1**, which contains two steps: turning on and turning off genes. In these two stages, promoters receive information from RNA polymerase to decide the

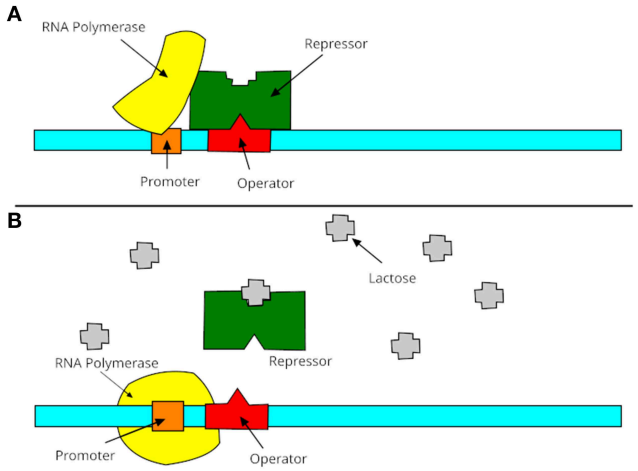


FIGURE 1 | Process of promoters in transcription. **(A)** The gene is essentially turned off. The repressor is not inhibited by lactose and binds to operator, then promoter is bound to make lactase; **(B)** the gene is turned on. The repressor is inhibited by lactose, then the promoter is bound by the RNA polymerase and express the genes to synthesize lactase. Finally, the lactase will digest all of the lactose, until nothing binds to the repressor. The repressor will then bind to the operator, stopping the manufacture of lactase.

manufacture of lactase. Promoters can be about 100–1,000 base pairs long. There are three elements of promoters in eukaryotic cells, such as core promoter, proximal promoter, and distal promoter. Each of them plays a different role in DNA transcription and RNA polymerase. Many recent studies suggested that DNA promoters may be the primary cause of many human diseases, especially diabetes (Döhr et al., 2005; Ionescu-Tîrgovişte et al., 2015) or Huntington's disease (Coles et al., 1998).

Owing to the huge importance of promoters in genetics and human diseases, the detection of them is an essential problem in genome research. A lot of efforts had been made to address this issue, from researchers with wet-lab, experimental, and computational techniques. One of the most important techniques is to detect the promoters based on TATA box, which is a motif that contains 24% of promoter genes in eukaryotes. Examples of this approach include: Promoter Scan (Prestridge, 1995) built a scoring profile by combining a weighted matrix for scoring a TATA box; Promoter2.0 (Knudsen, 1999) combined genetic algorithms and elements similar to neural networks to recognize promoter regions; Reese (2001) annotated promoters in the *Drosophila* melanogaster genome using a time-delay neural network; and (Down and Hubbard, 2002) combined TATA box with flanking regions of C-G enrichment. Later, some approaches focused on addressing this problem with spatial information of the base pairs in the sequences. There are some examples in this case: PromoterInspector identified promoters, based on the genetic context of promoters rather than their exact location; MCPromoter1.1 (Ohler et al., 1999) identified promoters based on three interpolated Markov chains (IMCs) of a different order. Moreover, the location of GpG islands had been used to predict the promoters region, as shown in Ioshikhes and Zhang (2000), Davuluri et al. (2001), and Ponger (2002).

Over the past decade, with the development of NGS technology, a large number of sequences was transcribed, which motivates researchers to build their predictors on sequence information. Similarly for promoters, it is necessary and urgent to develop highly efficient prediction techniques on it. Some notable research have been reported in the identification of promoters using sequence information. For instance (Li and Lin, 2006) recognized and predicted σ70 promoters in *Escherichia coli* K-12 by using position-correlation scoring matrix (PCSM) algorithm. This problem has been improved upon using variable-window Z-curve composition (Song, 2011) and six local DNA structural properties (Lin et al., 2018). Yang et al. (2017) exploited sex cell types and word embedding to identify enhancer–promoter interaction. Two types of promoters (σ54 and σ28) were identified by integrating DNA duplex stability into neural networks (de Avila e Silva et al., 2014). Later, (Lin et al., 2014) identified σ54 promoters using PseKNC, which is an advanced feature in bioinformatics fields. PseKNC had been used in the latter applications to classify promoter's types (Liu et al., 2017) and promoter's strength (Xiao et al., 2018). The promoter strength of *Escherichia coli* σ70 has been also predicted in Bharanikumar et al. (2018) with use of respective position weight matrices (PWM). Deep convolutional neural networks have been used to identify promoters using sequence information, such as recognition of prokaryotic and eukaryotic promoters (Umarov and Solovyev, 2017).

Identifying promoters, especially their strength, is an important problem in this aspect and latest research (Xiao et al., 2018) has achieved an accuracy of 83.13 and 71.20% for two layers, respectively. However, the performance results are not satisfactory and requires a lot of efforts from bioinformatics researchers to enhance the accuracy. A novel approach, proposed in this study, aims to address this problem. Our idea is based upon the natural language processing (NLP) field which classifies the text/sentence into its appropriate scenario. Therefore, we would like to apply it to bioinformatics to interpret the hidden information of DNA sequences (represented by promoters). Over the past decade, some researchers have successfully applied NLP techniques into biological sequences. One of the pioneering studies is from Asgari and Mofrad (2015) and it had been applied successfully in many later bioinformatics applications (Habibi et al., 2017; Hamid and Friedberg, 2018; Öztürk et al., 2018). However, most studies used the Word2Vec model or FastText model with a single level of N-gram. Here, a novel approach is presented, in which we used a combination of FastText N-grams to represent the DNA sequences. With this idea, we are able to take into account the sub-word information of DNA sequences as well as many N-gram levels in order to aid the increase in the predictive performance. Another point is the use of deep learning to take advantage of the numerous promoter sequences in this problem.

We listed some key contributions of this study which are as follows: (1) a computational model for classifying promoters which achieved better performance than the previous methods; (2) a novel method for generating hidden information of

DNA sequences by incorporating a combination of FastText N-grams and deep learning; (3) a study that provides significant information for researchers and biologists to better understand the promoter's functions; and (4) a basis for further study that would apply the FastText model and deep learning architecture in solving the bioinformatics problem. Here we deal with these contributions clearly in the following sections.

METHODS

Under the operation of a specifically designed pipeline, an overall flowchart of our approach is presented in **Figure 2**. Each of the experimental steps of this proposed pipeline will be sequentially addressed in the following subsections.

Benchmark Dataset

Collecting a high-quality dataset is one of the most important steps to address a bioinformatics problem. In this study, we re-used the benchmark dataset from Xiao et al. (2018) to objectively assess the difference in performance between our model and other existing ones. In this dataset, they collected all experimentally—confirmed promoter sequences from RegulonDB (Gama-Castro et al., 2015), which is a huge database of the regulatory network of gene expression. These sequences were categorized into two groups: strong and weak promoters based on their levels in transcription activation and expression. They also extracted non-promoter sequences by considering intron, exon, and intergenic sequences excluding the positive sequences. After that, the CD-HIT [26] was also used to exclude the pairwise sequences whose similarities were calculated to be more than 85%.

The benchmark dataset encompasses 3,382 promoter samples and 3,382 non-promoter samples. In 3,382 promoter samples, there are 1,591 strong promoter samples and 1,792 weak promoter samples for construction of second layer classification. It can be freely downloaded at http://www.jci-bioinfo.cn/iPSW(2L)-PseKNC/images/Supp.pdf. The whole dataset was randomly divided into five subsets to perform a 5-fold cross-validation. The training process was performed using a fixed ratio of the training set over the validation set of 4:1 with alternation.

DNA Representation With Language Model

A DNA sequence consists of four nucleotides: adenosine (A), cytidine (C), guanosine (G), and thymine (T). These nucleotides will combine together to form a definite sequence in the DNA

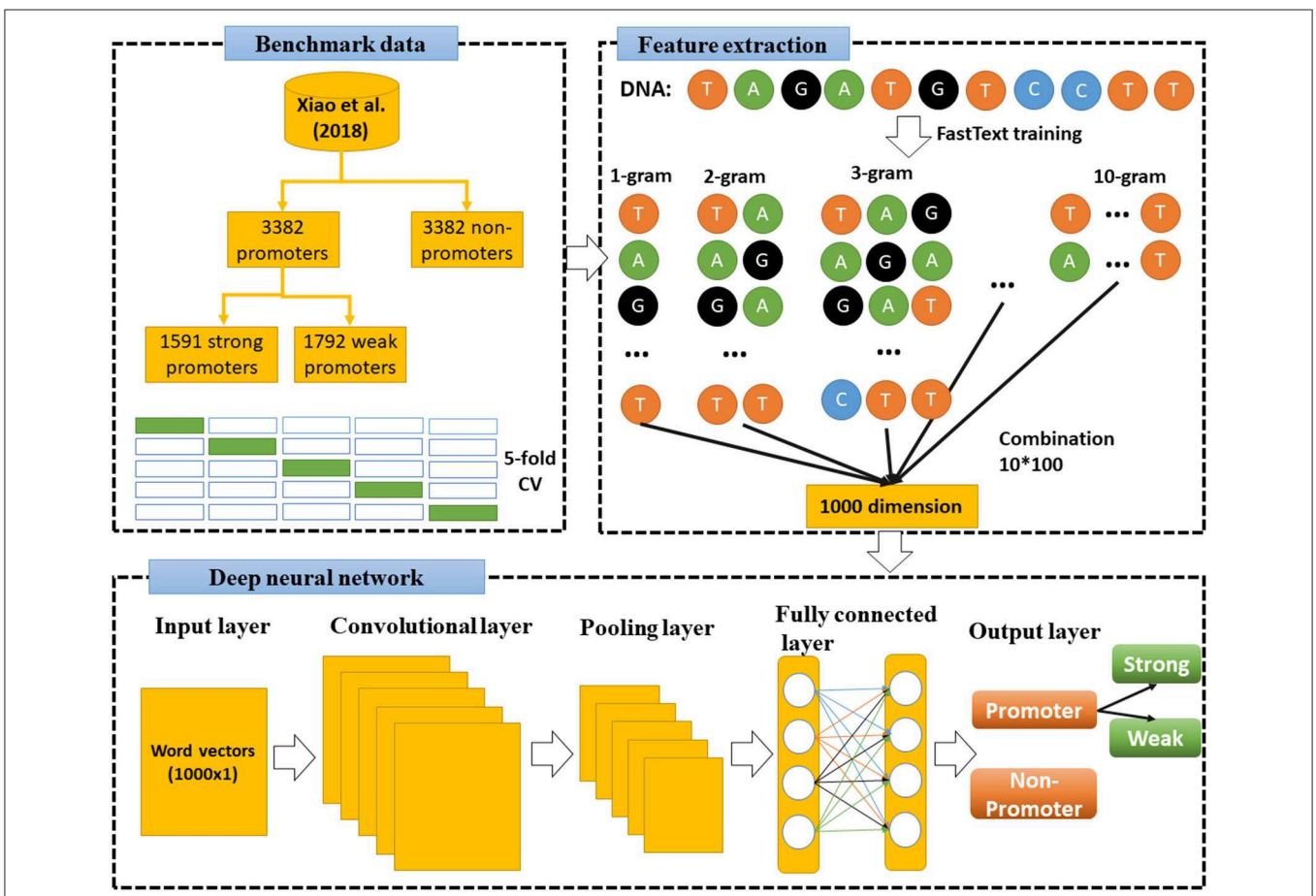


FIGURE 2 | Flowchart of this study. First, we used FastText to train model and extract features from benchmark dataset (Xiao et al., 2018), then combined 10-gram levels to a combination sets of vectors (1,000 dimensions). Deep neural network was then constructed to learn these vectors and classify the DNA sequences.

sequence. Feature extraction is an important step in most of the bioinformatics problems, whereby the main features will help in discriminating DNA sequences. One of the most common methods is the use of k-mer. K-mers are described as all the possible subsequences (of length k) from a read accessed through DNA sequencing. The number of k-mers possible given a string of length L is L-k+1, whilst the number of possible k-mers given n possibilities (four in the case of DNA e.g., ATGC) is nk. K-mer has been used in a lot of bioinformatics problems and has achieved promising results. Next, Chou highlighted PseDNC which has extracted DNA sequences via different ways. PseDNC has helped to rectify numerous problems relating to bioinformatics, as compared to using k-mer. Another approach is the use of language model to represent the information of DNA sequences. In this approach, DNA sequence will be treated as a language sentence and then fed into supervised learning for classification. We can easily list the methods using this approach, from Word2Vector to FastText. In these approaches, FastText has been proven to achieve better performance as compared to Word2Vector or Glove.

FastText Implementation

In order to generate continuous N-grams, we made use of FastText (Bojanowski et al., 2017), which is a library from Facebook for representation and classification of text. In FastText, we can train different language models such as skip-gram or CBOW and apply a variety of parameters such as sampling or loss functions. There are a lot of improvements from Word2Vector to FastText as described in Bojanowski et al. (2017) and Le et al. (2019a). In this study, each DNA sequence was treated as a sentence with a lot of words. Moreover, each word contains a bag of character n-gram. As mentioned in FastText's document, they modified the algorithm of Word2Vector whereby special symbols "and" are added at the boundary of words, which helps to differentiate prefixes and suffixes from other character sequences. Moreover, the word itself has been also included in the n-gram set to learn a representation for each word (together with character n-grams). To explain the idea, we used our DNA word "ATGAC" as an example. If we would like to generate the representation of this word with 3-gram, they will be consequently: <AT, ATG, TGA, GAC, AC> and the special sequence <ATGAC>. Here, it is noteworthy that the representation <TGA>, corresponding to the word "TGA," is different from the tri-gram "TGA," derived from the word "ATGAC." The reason is because of the potential of extracting sub-word information in word "TGA" of FastText and it could help generate more information for each word. The word generated by FastText could be considered as a continuous bag of words. In this study, we extracted all the n-grams from 1 to 10 to consider the optimal levels of them.

What makes FastText different from Word2Vector is the sub-word information, and it is proposed via a scoring function *s* as follows:

$$s(w, c) = \sum_{g \in G_w} z_g^T v_c \quad (1)$$

where G is the size of n-grams, G_w ranges from 1 to G, w is a given word, z_g is a vector representation to each n-gram g, v_c is context vector. This simple modification allows objective representation of words, thus helping the model learn reliable representation for rare words.

Based on the recent successful applications of FastText model in representing biological sequence (Le, 2019; Le et al., 2019a), we introduced a more in-depth benchmark method using FastText to improve this representation. Here we take into account the combination of continuous N-gram levels, which was not considered by the previous studies. It means that instead of using only one level of N-gram and sub-word information, we used a lot of N-gram combinations and considered which was the best combination for this problem. A huge advantage of this approach is that we can have many features for learning. In addition, we can easily implement feature selection techniques and improve the performance results in the model.

1D Convolutional Neural Network

In general, CNN is a class of deep neural networks that has been demonstrated to be exceptionally successful in territories, such as picture acknowledgment and order. CNN has been fruitful in computer vision related issues such as face recognition, object detection, or self-driving cars. CNN appears ready to reproduce and upgrade these key strides in a bound together structure and learn various leveled portrayals specifically from crude images. If we take a convolutional neural organization that has been prepared to perceive protests inside pictures, then that system will have built up some inward autonomous portrayals of the substance and style contained inside a given picture. Since the input of this problem was a vector, therefore, we used 1D CNN. Similar to 2D CNN approaches which has been used in bioinformatics (Le and Nguyen, 2019; Le et al., 2019b; Nguyen et al., 2019), it consisted of the following layers:

(1) Input layer: The input of our model is a 1D vector, which is a vector of size 1×100 (created by FastText model).
(2) Convolutional layer: A 1D convolutional layer (e.g., temporal convolution) is used to construct a convolution kernel and then derive features encoded in the 1D input vector. The convolutional layer moves in stride over the input, transforming the values into representative values via a sliding window. This process helps conserve the dimensional relationship between numeric values in the vectors, by gaining beneficial features using small parts of input data. Since our input size was not big, a kernel size of 3 was applied to figure out more information.
(3) Rectified Linear Unit (ReLU): an additional non-linear operation is presented after every convolution operation. It aims to perform non-linear function in our CNN and help our model understand data better. The output function of ReLU is as follows:

$$f(x) = \max(0, x) \quad (2)$$

where x is the number of inputs in a neural network.

(4) Pooling layer: It is normally added inside the convolutional layers to reduce the calculation of the next layers. Max pooling was selected in this step with stride of 2.
(5) Dropout layer: A technique which aims to prevent overfitting and also help to increase the model's performance (Srivastava et al., 2014).
(6) Flatten layer: a layer helps to transform the input matrix into a vector.
(7) Fully connected layer: is normally inserted by the last stage of the deep networks. The layer is fully-connected if each node is connected with all of the previous nodes in the network. Our problem is to identify between promoter and non-promoter (or classify strong and weak promoter), thus it was a binary classification. Therefore, the final number of nodes in our output is 2.
(8) Softmax is a logistic function defined by the formula:

$$\sigma\ (z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \tag{3}$$

where z is the input vector with K-dimensional vector, $\sigma(z)_i$ is real values in the range (0, 1) and ith class is the predicted probability from sample vector x. It was compulsory to insert Softmax, in order to determine the probability of each possible output.

Assessment of Predictive Ability

To evaluate the performance of the classifiers that were constructed by the aforementioned deep learning architecture, the 5-fold cross-validation technique was implemented. The average metrics among the five testing sets were determined in order to compare the performance when constructing the classifier. We follow Chou's evaluation criteria which is widely used in many bioinformatics studies (Chou, 2001; Xiao et al., 2018; Le et al., 2019a). The criteria includes sensitivity (Sens), specificity (Spec), accuracy (Acc), and Matthews Correlation Coefficient (MCC) which are defined as:

$$Sensitivity = 1 - \frac{N_-^+}{N^+},\ 0 \le Sen \le 1 \tag{4}$$

$$Specificity = 1 - \frac{N_+^-}{N^-},\ 0 \le Spec \le 1 \tag{5}$$

$$Accuracy = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-},\ 0 \le Acc \le 1 \tag{6}$$

$$MCC = \frac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}},\ -1 \le MCC \le 1 \tag{7}$$

The relations between these symbols and the symbols in Equations (4, 5, 6, and 7) are given by:

$$\begin{cases} N_+^- = FP \\ N_-^+ = FN \\ N^+ = TP + N_-^+ \\ N^- = TN + N_+^- \end{cases} \tag{8}$$

True positive (TP) and true negative (TN) are the respective numbers of correctly predicted promoter and non-promoter, whereas false positive (FP) and false negative (FN) are the respective numbers of misclassified promoter and non-promoter.

Likewise, we also used Receiver Operating Characteristics (ROC) curve and Area Under Curve (AUC) (Bradley, 1997) as the additional metrics for performance evaluation. The AUC is a probability value ranging from 0 to 1 in which the greater AUC shows the better predictive performance.

RESULTS

Optimal Experimental Setup

In this analysis, we attempted to observe the optimal hyperparameters that were used in this study. Because we integrated FastText and deep learning model, we chose the best parameters for both methods. FastText has a lot of different parameters for training purpose. Many prior research on it determined that changing these parameters will help to change the model's accuracy drastically. Therefore, we would like to perform a one-by-one strategy to tune up the optimal parameters in FastText. There are a lot of parameters that may affect the performance results and we decided to adapt these parameters such as wordNgrams (max length of word n-gram), lr (learning rate), dim (size of word vectors), ws (size of context window), epoch (number of iterations), and loss (loss function). We used a basic setting on FastText classifier to perform supervised learning for text classification. The dataset used in this section helped distinguish between promoters and non-promoters. In the first experiment, we would like to examine the effect of different levels of N-grams (from 1 to 10) on the performance results. The important measurement metric used in this evaluation is ROC AUC value. As shown in **Figure 3**, our classifier could classify promoters with high performance (AUC $\sim$ 0.9), especially in two levels: 4-gram and 5-gram. However, the differences were not significant and it indicates that we can select any level of N-gram to create a good model for promoter classification. **Table 1** shows the hyperparameters used for tuning the model. After the tuning process, we also presented the best set of hyperparameters found: learning rate of 0.1, vector dimension of 100, context window size of 5, epoch of 100, and softmax loss function.

The next tuning is from deep learning architecture, in which we performed a grid search CV on a set of potential hyperparameters. All of the parameters selected for tuning in CNN include the number of layers, epochs, batch sizes, dropout values, weight constant as well as the optimizer and activation function. After this step, we identified a set of optimal hyperparameters in CNN as follows: 64 filter layers, batch size of 100, epoch of 100, dropout of 0.3, weight constraint of 4, adadelta optimizer, and linear activation. We then used all of the optimal parameters in the next experiments as well as the later comparisons.

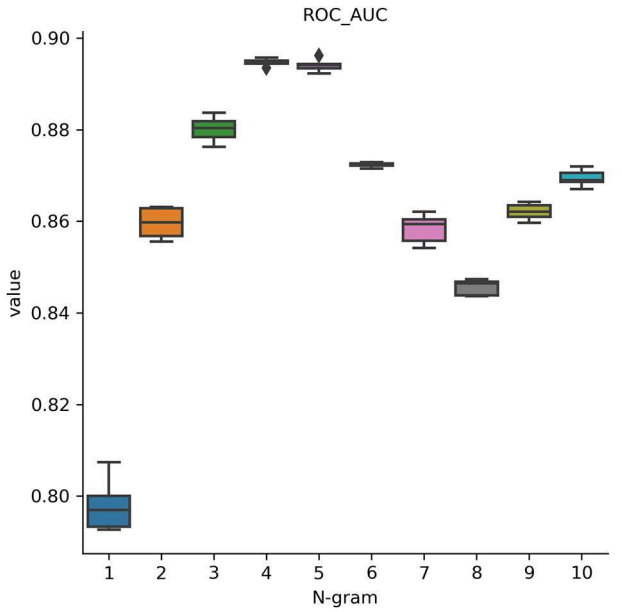


FIGURE 3 | Performance results on identifying promoters using different levels of N-gram. Our classifier could classify promoters with high performance (AUC ~ 0.9), especially at 4-gram and 5-gram levels.

TABLE 1 | Hyperparameters chosen for tuning FastText model.

Parameters	Range	Stepsize	Optimal
lr	0.05–0.25	0.05	0.1
Dim	50–500	25	100
Ws	1–10	1	5
Epoch	25–500	25	100
Loss	[ns, hs, softmax]	-	softmax

Lr, learning rate; dim, dimension; ws, size of context window; epoch, number of iterations; loss, loss function.

Effects of Different Levels of N-Gram and Combination of Continuous N-Grams in Classifying Promoters

According to the previous section, changing the number of N-grams did not make significant effect on promoter classification. It has been also proven in some of the previous works which used the FastText model (Le, 2019; Le et al., 2019a). However, one novel idea implemented in this study was to increase the performance results by using a combination of N-grams. The idea was to combine all of the N-gram levels into a big set of features, which will then be fed into classifiers. As such, our classifier will take full advantage of important features for each specific N-gram level and remove some less important features inside all of the levels. The performance results were shown in detail in **Table 2**. It is noted that the 5-fold cross-validation has been performed for several independent iterations to give a confidence interval for the results. In these results, we fed all 1,000 features from 10 levels of N-gram into our CNN architecture. It is easy to say that the combination of N-grams outperforms the single level of N-gram. This method achieved a sensitivity of 82.76%, specificity of 88.05%, accuracy of 85.41% and MCC of 0.709, which is improved ~1–4% from single N-gram in term of specificity, accuracy, and MCC. To statistically compare between N-gram combination and N-gram single levels, we performed 10 times of one-sided Wilcoxon tests of the ROC AUC values between the combination model and each of the 1–10-gram model. After that, all of Wilcoxon tests showed a p-value of 0.0005 (less than significance level $\alpha = 0.05$) which could strongly conclude that the performance results of combination features were significantly better than the single ones at high confidence level.

TABLE 2 | Comparison between single N-gram and combination of continuous N-grams.

Methods	Sens	Spec	Acc	MCC
Single N-gram	82.43	83.34	82.88	0.658
Combination of N-grams	82.76	88.05	85.41	0.709

Single N-gram, representative by 4-gram; Combination of N-grams, combine 10 levels of N-gram together.

TABLE 3 | Top-ranked features using MRMD feature selection technique.

No.	Feature number	Score
1	feature_97	1.0
2	feature_21	0.9170726107858075
3	feature_34	0.9096134637807235
4	feature_92	0.8914645287023287
5	feature_54	0.8463944338892277
6	feature_9	0.8368290059895386
7	feature_41	0.824726606348234
8	feature_8	0.8020998165541897
9	feature_77	0.7714372077391476
10	feature_3	0.7598084153408637

Since deep learning is a black-box manner, it automatically generated the hidden information from our feature sets. Therefore, it is challenging to understand which features have most contribution or play critical role for promoter distinction in our model. As a reference, we used a common technique namely Maximum-Relevance-Maximum-Distance (MRMD) (Zou et al., 2016a) to evaluate and extract the important features of our datasets. MRMD has been used a lot of works in bioinformatics with promising results (Zou et al., 2016b; Wei et al., 2017). According to the results, MRMD suggested that our model will reach the highest accuracy when we selected 835 top-ranked features (out of 1,000) to insert into our neural network. To detail, 10 features had the highest scores were shown in **Table 3**. These features, therefore, play an essential role in classifying promoter sequences using our model.

Next, we would like to compare our performance results with a baseline machine learning technique to check whether the deep CNN has generated more hidden information and given

a significant performance. Since nearest neighbor (kNN) (Keller et al., 1985) has been used to represent for traditional machine learning classifiers in different problems, we implemented it in our study for comparison. We used hyperparameter optimization process and found that the model performed consistently at 10 neighbor trees. The optimal performance reached 78.8%, 86.8%, 82.8%, 0.66, and 0.885 for sensitivity, specificity, accuracy, MCC, and AUC, respectively. Compared with the performance from CNN, kNN was lower in term of sensitivity, accuracy, MCC, and AUC. It is enough evidence to say that the deep neural network could learn more features and produce a better performance than traditional neural networks.

Classifying Promoters' Strength

Since the combination of N-grams performed well in the first layer classification, we aimed to use the same experimental setups for the second layer (classifying promoter's strength). Our dataset includes 1591 strong promoters and 1792 weak promoters as collected from Xiao et al. (2018) and has been mentioned in the dataset section. The experiments show that our method, which used a combination of N-grams, could classify the promoter's strength with an accuracy of 73.1%, sensitivity of 69.4%, specificity of 76.4%, and MCC of 0.46. The performance was also better than the baseline models with single levels of N-grams. It means that we can use this setup for both layers with promising results.

Comparison the Performance Results Between Proposed Method and the Existing Methods

Our best model as mentioned in the previous sections is the combination of different N-gram levels and deep convolutional neural networks. To be fair, we have to compare our proposed method with the other previous works that regarding promoter classification. Also it is noted that we surely chose the previous works that used the same benchmark dataset. For the first layer, numerous studies had been done, including PCSF (Li and Lin, 2006), vw Z-curve (Song, 2011), Stability (de Avila e Silva et al., 2014), iPro54 (Lin et al., 2014), iPromoter-2L (Liu et al., 2017), and iPSW(2L)-PseKNC (Xiao et al., 2018). Among these studies, only the last one performed the classification of promoter's strength, thus we also compared with this predictor in our second layer. The results are shown in **Table 4**, and we highlighted the highest values to highlight the significance of each metrics. We then observed that our method outperforms other predictors in all metrics (sensitivity, specificity, accuracy, and MCC) in both layer classifications. Another improvement is that our approach could be applied to actual genome sequences (long fragments of bacterial genomes) rather only short sequences. All sequences with different length will be trained to become a vector with a fix-length. It helps to input any form of sequences flexibly.

TABLE 4 | Comparison with previous predictors on the same benchmark dataset.

Predictors	Sens	Spec	Acc	MCC
1st layer				
Ours	**82.76**	**88.05**	**85.41**	**0.709**
iPSW(2L)-PseKNC	81.37	84.89	83.13	0.663
iPromoter-2L	79.2	84.16	81.68	0.6343
iPro54	77.76	83.15	80.45	0.61
Stability	76.61	79.48	78.04	0.5615
vw Z-curve	77.76	82.8	80.28	0.6098
PCSF	78.92	70.7	74.81	0.498
2nd layer				
Ours	**69.4**	76.4	**73.1**	**0.46**
iPSW(2L)-PseKNC	62.23	**79.17**	71.2	0.4213

Highlighted values are the significant values for each metric.

DISCUSSIONS

Promoters play an important role in the transcription of genes affect numerous human diseases. Therefore, identification of promoters using their sequence information is one of the most important tasks in bioinformatics. Although few computational tools had already been presented, the performance results require improvements. This study presents a new hybrid system, from deep learning and a combination of FastText N-grams, to identify promoters and their respective strengths. To our knowledge, this is the first bioinformatics study which has applied this hybrid into biological sequences. By using this method, we are able to generate the hidden information of DNA sequences unlike other methods. Our performance results were evaluated via a 5-fold cross-validation test on a benchmark dataset. It was found that the proposed method could identify promoters and their strength, with an accuracy of 85.41 and 73.1%, respectively. The rest of the measurement metrics, such as sensitivity, specificity, and MCC, also attained superior performances. When compared to the other state-of-the-art predictors regarding the same problem and dataset, our proposed method has improved at about 1–4% in all of the metrics. Therefore, our model can be considered as a reliable method for identifying promoters and their strength, with use of sequence information. It can also act a basis for further study that aims to interpret the language context of DNA sequences.

Last but not least, scientists can use our approach to solve further bioinformatics problems on sequencing. Since most bioinformatics problems focused on sequencing data, their features could be extracted by using our combination (different levels of FastText N-grams). They then be fed into a supervised learning to perform the prediction or classification (e.g., using deep neural network as proposed in this work). It could also provide a new approach for the previous works that only used one level of FastText (Le, 2019; Le et al., 2019a). A combination of more levels could be a solution for boosting their predictive performances. We also provided our source codes at https://github.com/khanhlee/deepPromoter to help reproducing our method. Furthermore, since a lot of previous works on promoter classification extracted features by using PseKNC [such as (Liu et al., 2017; Lin et al., 2018; Xiao et al., 2018)], a hybrid of this feature and our features

could be considered in the future works for the purpose of performance improvement.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.jci-bioinfo.cn/iPSW(2L)-PseKNC/images/Supp.pdf.

AUTHOR CONTRIBUTIONS

NL and EY conceived the ideas and designed study. NL conducted the experiments and analyzed the results. NL, EY, NN, and H-YY participated in the discussion of the results and writing of the article. All authors read and approved the final version of the manuscript.

FUNDING

This work has been supported by the Nanyang Technological University Start-Up Grant.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

Asgari, E., and Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* 10:e0141287. doi: 10.1371/journal.pone.0141287

Bharanikumar, R., Premkumar, K. A. R., and Palaniappan, A. (2018). PromoterPredict: sequence-based modelling of *Escherichia coli* σ70 promoter strength yields logarithmic dependence between promoter strength and sequence. *PeerJ* 6:e5862. doi: 10.7717/peerj.5862

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Trans. Assoc. Comp. Lingu.* 5, 135–146. doi: 10.1162/tacl_a_00051

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* 30, 1145–1159. doi: 10.1016/S0031-3203(96)00142-2

Chou, K. C. (2001). Prediction of protein signal sequences and their cleavage sites. *Proteins* 42, 136–139. doi: 10.1002/1097-0134(20010101)42:1<136::AID-PROT130>3.0.CO;2-F

Coles, R., Caswell, R., and Rubinsztein, D. C. (1998). Functional analysis of the huntington's disease (HD) gene promoter. *Hum. Mol. Genet.* 7, 791–800. doi: 10.1093/hmg/7.5.791

Davuluri, R. V., Grosse, I., and Zhang, M. Q. (2001). Computational identification of promoters and first exons in the human genome. *Nat. Genet.* 29, 412–417. doi: 10.1038/ng780

de Avila e Silva, S., Forte, F., Sartor, T. S. I., Andrighetti, T., Gerhardt, G. J. L., Longaray Delamare, A. P., et al. (2014). DNA duplex stability as discriminative characteristic for *Escherichia coli* σ54- and σ28- dependent promoter sequences. *Biologicals* 42, 22–8. doi: 10.1016/j.biologicals.2013.10.001

Döhr, S., Klingenhoff, A., Maier, H., de Angelis, M. H., Werner, T., and Schneider, R. (2005). Linking disease-associated genes to regulatory networks via promoter organization. *Nucleic Acids Res.* 33, 864–872. doi: 10.1093/nar/gki230

Down, T. A., and Hubbard, T. J. P. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* 12, 458–461. doi: 10.1101/gr.216102

Gama-Castro, S., Salgado, H., Santos-Zavaleta, A., Ledezma-Tejeida, D., Muñiz-Rascado, L., García-Sotelo, J. S., et al. (2015). RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 44, D133–D43. doi: 10.1093/nar/gkv1156

Habibi, M., Weber, L., Neves, M., Wiegandt, D. L., and Leser, U. (2017). Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 33, i37–i48. doi: 10.1093/bioinformatics/btx228

Hamid, M.-N., and Friedberg, I. (2018). Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformatics* 35, 2009–2016. doi: 10.1101/255505

Ionescu-Tîrgovişte, C., Gagniuc, P. A., and Guja, C. (2015). Structural properties of gene promoters highlight more than two phenotypes of diabetes. *PLoS ONE* 10:e0137950. doi: 10.1371/journal.pone.0137950

Ioshikhes, I. P., and Zhang, M. Q. (2000). Large-scale human promoter mapping using CpG islands. *Nat. Genet.* 26, 61–63. doi: 10.1038/79189

Keller, J. M., Gray, M. R., and Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst. Man Cybern.* 580–5. doi: 10.1109/TSMC.1985.6313426

Knudsen, S. (1999). Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* 15, 356–361. doi: 10.1093/bioinformatics/15.5.356

Le, N. Q. K. (2019). iN6-methylat (5-step): identifying DNA N6-methyladenine sites in rice genome using continuous bag of nucleobases via Chou's 5-step rule. *Mol. Genet. Genomics*. 294, 1173–1182. doi: 10.1007/s00438-019-01570-y

Le, N. Q. K., Huynh, T.-T., Yapp, E. K. Y., and Yeh, H.-Y. (2019b). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. *Comput. Methods Programs Biomed.* 177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Le, N. Q. K., and Nguyen, V.-N. (2019). SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data. *PeerJ Comp. Sci.* 5:e177. doi: 10.7717/peerj-cs.177

Le, N. Q. K., Yapp, E. K. Y., Ho, Q.-T., Nagasundaram, N., Ou, Y.-Y., and Yeh, H.-Y. (2019a). iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* 571, 53–61. doi: 10.1016/j.ab.2019.02.017

Li, Q.-Z., and Lin, H. (2006). The recognition and prediction of σ70 promoters in *Escherichia coli* K-12. *J. Theor. Biol.* 242, 135–141. doi: 10.1016/j.jtbi.2006.02.007

Lin, H., Deng, E.-Z., Ding, H., Chen, W., and Chou, K.-C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.* 42, 12961–12972. doi: 10.1093/nar/gku1019

Lin, H., Liang, Z., Tang, H., and Chen, W. (2018). Identifying sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comp. Biol. Bioinform.* 16, 1316–1321. doi: 10.1109/TCBB.2017.2666141

Liu, B., Yang, F., Huang, D.-S., and Chou, K.-C. (2017). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Nguyen, T.-T.-D., Le, N.-Q.-K., Kusuma, R. M. I., and Ou, Y.-Y. (2019). Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network. *J. Mol. Graph. Model.* 92:86–93. doi: 10.1016/j.jmgm.2019.07.003

Ohler, U., Harbeck, S., Niemann, H., N$\sqrt{}$∂th, E., and Reese, M. G. (1999). Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics* 15, 362–369. doi: 10.1093/bioinformatics/15.5.362

Öztürk, H., Ozkirimli, E., and Özgür, A. (2018). A novel methodology on distributed representations of proteins using their interacting ligands. *Bioinformatics* 34, i295–i303. doi: 10.1093/bioinformatics/bty287

Ponger, L. C, and Mouchiroud, D. (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18, 631–633. doi: 10.1093/bioinformatics/18.4.631

Prestridge, D. S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* 249, 923–932. doi: 10.1006/jmbi.1995.0349

Reese, M. G. (2001). Application of a time-delay neural network to promoter annotation in the *Drosophila* melanogaster genome. *Comput. Chem.* 26, 51–56. doi: 10.1016/S0097-8485(01)00099-7

Song, K. (2011). Recognition of prokaryotic promoters based on a novel variable-window Z-curve method. *Nucleic Acids Res.* 40, 963–971. doi: 10.1093/nar/gkr795

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Umarov, R. K., and Solovyev, V. V. (2017). Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* 12:e0171410. doi: 10.1371/journal.pone.0171410

Wei, L., Xing, P., Su, R., Shi, G., Ma, Z. S., and Zou, Q. (2017). CPPred-RF: a sequence-based predictor for identifying cell-penetrating peptides and their uptake efficiency. *J. Proteome Res.* 16, 2044–2053. doi: 10.1021/acs.jproteome.7b00019

Xiao, X., Xu, Z.-C., Qiu, W.-R., Wang, P., Ge, H.-T., and Chou, K.-C. (2018). iPSW(2L)-PseKNC: a two-layer predictor for identifying promoters and their strength by hybrid features via pseudo K-tuple nucleotide composition. *Genomics* (2018). doi: 10.1016/j.ygeno.2018.12.001. [Epub ahead of print]

Yang, Y., Zhang, R., Singh, S., and Ma, J. (2017). Exploiting sequence-based features for predicting enhancer-promoter interactions. *Bioinformatics* 33, i252–i60. doi: 10.1093/bioinformatics/btx257

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016b). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016a). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 November 2019
doi: 10.3389/fgene.2019.01056

Meta-Analysis of HTLV-1-Infected Patients Identifies CD40LG and GBP2 as Markers of ATLL and HAM/TSP Clinical Status: Two Genes Beat as One

Eduardo Rocha Fukutani[1], Pablo Ivan Pereira Ramos[1], José Irahe Kasprzykowski[1], Lucas Gentil Azevedo[1], Moreno Magalhães de Souza Rodrigues[2], João Victor de Oliveira Pimenta Lima[1], Helton Fábio Santos de Araújo Junior[1], Kiyoshi Ferreira Fukutani[1,3,4] and Artur Trancoso Lopo de Queiroz[1]*

[1] Center of Data and Knowledge Integration for Health (CIDACS), Instituto Gonçalo Moniz, FIOCRUZ, Salvador, Brazil, [2] Laboratório de Análise e Visualização de Dados, FIOCRUZ-RO, Salvador, Brazil, [3] Fundação José Silveira, Multinational Organization Network Sponsoring Translational and Epidemiological Research, FJS, Salvador, Brazil, [4] Faculdade de Medicina, Faculdade de Tecnologia e Ciências, Salvador, Brazil

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Andrew Dellinger, Elon University, United States
Lei Hou, Massachusetts Institute of Technology, United States

***Correspondence:**
Eduardo Rocha Fukutani
eduardofukutani@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 11 June 2019
Accepted: 02 October 2019
Published: 08 November 2019

Citation:
Fukutani ER, Ramos PIP, Kasprzykowski JI, Azevedo LG, Rodrigues MMdS, Lima JVdOP, Araújo Junior HFSd, Fukutani KF and Queiroz ATLd (2019) Meta-Analysis of HTLV-1-Infected Patients Identifies CD40LG and GBP2 as Markers of ATLL and HAM/TSP Clinical Status: Two Genes Beat as One. Front. Genet. 10:1056. doi: 10.3389/fgene.2019.01056

Human T-lymphotropic virus 1 (HTLV-1) was the first recognized human retrovirus. Infection can lead to two main symptomatologies: adult T-cell lymphoma/leukemia (ATLL) and HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP). Each manifestation is associated with distinct characteristics, as ATLL presents as a leukemia-like disease, while HAM/TSP presents as severe inflammation in the central nervous system, leading to paraparesis. Previous studies have identified molecules associated with disease development, e.g., the downregulation of Foxp3 in Treg cells was associated with increased risk of HAM/TSP. In addition, elevated levels of CXCL10, CXCL9, and Neopterin in cerebrospinal fluid also present increased risk. However, these molecules were only associated with specific patient groups or viral strains. Furthermore, the majority of studies did not jointly compare all clinical manifestations, and robust analysis entails the inclusion of both ATLL and HAM/TSP. The low numbers of samples also pose difficulties in conducting gene expression analysis to identify specific molecular relationships. To address these limitations and increase the power of manifestation-specific gene associations, meta-analysis was performed using publicly available gene expression data. The application of supervised learning techniques identified alterations in two genes observed to act in tandem as potential biomarkers: *GBP2* was associated with HAM/TSP, and *CD40LG* with ATLL. Together, both molecules demonstrated high sample-classification accuracy (AUC values: 0.88 and 1.0, respectively). Next, other genes with expression correlated to these genes were identified, and we attempted to relate the enriched pathways identified with the characteristic of each clinical manifestation. The present findings contribute to knowledge surrounding viral progression and suggest a potentially powerful new tool for the molecular classification of HTLV-associated diseases.

Keywords: human T-lymphotropic virus 1, bioinformatics, biomarkers, adult T-cell lymphoma/leukemia, HTLV-1 associated myelopathy/tropical spastic paraparesis, meta-analysis

INTRODUCTION

Human T-lymphotropic virus 1 (HTLV-1) belongs to the *Retroviridae* family and Deltaretrovirus genus, and presents tropism in the infection of T lymphocyte cells (Mirvish et al., 2011). Two diseases are mainly associated with this infection: adult T-cell lymphoma/leukemia (ATLL) and HTLV-associated myelopathy/tropical spastic paraparesis (HAM/TSP) (Gessain and Mahieux, 2012). Around 2–5% of HTLV-infected subjects develop ATLL (Uchiyama et al., 1977) and 0.25–3.8% develop HAM/TSP (Osame et al., 1986), while the majority of HTLV-infected subjects remain asymptomatic (Galvão-Castro et al., 1997). ATLL is a lymphoma-like disease classified into four subtypes: acute, chronic, smoldering, and lymphoma (Shimoyama and members of The Lymphoma Study Group (1984–87)*, 1991). Developing this symptomatology results in a life expectancy less than 1 year in around 65% of affected individuals (Matutes, 2007), in addition to low documented chemotherapeutic response (Yamada et al., 2001). HAM/TSP is characterized as an inflammatory disease of the central nervous system (CNS), can progressively evolve to spastic paraparesis, and results in sensory disturbance in the lower extremities and bladder/bowel dysfunction (Nakagawa et al., 1995).

Currently, ATLL can be diagnosed by integrating cytology and lymphocyte immunophenotyping with HTLV-1 serology (Matutes, 2007). The diagnosis of HAM/TSP is based on clinical evaluation and the exclusion of other disorders and molecular and serological diagnosis, including HTLV-1 serology, Western blotting, and PCR analysis (Yamano and Sato, 2012). In this complex scenario, the identification of biomarkers of this disease is crucial for improving patient care and treatment. With the goal of furthering the understanding surrounding the mechanisms related to disease manifestation, some studies employing gene expression have been conducted. For instance, the downregulation of the FOXP3 gene in T-reg cells was reported to be induced by the HBZ viral protein from HTLV-1. Accordingly, the stimulated proinflammatory response was found to be associated with HAM/TSP development (Yamamoto-Taguchi et al., 2013). Furthermore, other molecules in cerebrospinal fluid, such as CXCL10, CXCL9, and neopterin, have been proposed as promising candidates for prognostic biomarkers of HAM/TSP, offering improved predictive values in comparison to proviral load (Sato et al., 2013).

On the other hand, CAN2 and SPTA2 proteins have been proposed as biomarkers capable of classifying ATLL patients. CAN2 activity was found to induce ATLL cell death and the corresponding gene was downregulated in these cells. In addition, 17 proteins were proposed as capable of classifying healthy controls from asymptomatic carriers (ACs), HAM/TSP, and ATLL patients (Ishihara et al., 2013). Several alterations in anti-inflammatory cytokine levels in infected T cells, e.g., increased IL-10 and suppressed pro-inflammatory cytokines, were also associated with this disease (Kagdi et al., 2018). Another study suggested diagnosing patients by measuring antibody responses to HTLV-1 gag, Env, and Tax proteins (Enose-Akahata et al., 2012); however, this is akin to an immunological diagnosis. Despite the identification of biomarker candidates, various limitations have prevented adoption, as some markers were only identified in specific populations (Yasuma et al., 2016), small sample sizes were used (Ishihara et al., 2013), and the identification was performed only in specific clinical manifestations without appropriate confirmation for use as a general biomarker (Sato et al., 2013; Yamamoto-Taguchi et al., 2013).

To mitigate the impact of low sample sizes, which have limited the interpretation of individual studies, meta-analysis approaches have been employed in the field of gene/marker identification. This approach was used to highlight important genes and molecular pathways in endometrioid endometrial cancer (O'Mara et al., 2016), for the identification of programmed death-ligand 1 as a potential biomarker in glioblastoma (Xue et al., 2017), to identify a set of candidate genes, pathways, and transcription factors not previously associated with the pathogenesis of sickle cell disease (Hounkpe et al., 2015), and to disclose a novel set of candidate genetic markers, pathways, and transcription factors common to both thrombosis and myeloproliferative disorders (Jha et al., 2016). Meta-analysis, in combination with classical approaches and machine learning, has also been applied to identify biomarkers of viral infection in the *Aedes aegypti* mosquito (Fukutani et al., 2017). This methodology has proven powerful in discriminatory classification using gene expression data and was recently highlighted as a potentially useful method for discovering new evidences (Debray et al., 2017); Sweeney et al., 2017). Given the need to identify biomarkers associated with HTLV-1 infection, and considering the abundance of individual studies that resulted in the generation of gene expression datasets, we performed meta-analysis in an attempt to identify candidate transcriptional biomarkers that could offer improved predictive power in the classification of clinical manifestations in HTLV-1, a novelty in this field that has never been done before.

METHODOLOGY

Description of Datasets Comprising the Discovery Dataset

To identify published datasets relevant to HTLV infection, the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) was searched filtering *Homo sapiens* as the organism of interest and "HTLV" as the keyword. This query returned a total of 41 datasets (search performed in September 2017). After manual evaluation, 32 datasets were excluded due to methodological incompatibility (non-blood cell tissues and absence of symptomatologic information). Of the remaining datasets, three with detailed gene expression by peripheral blood mononuclear cells (PBMCs) were selected to build the Discovery dataset: GSE55851 (Kobayashi et al., 2014), GSE29312, and GSE29332 (Tattermusch et al., 2012). All of the studies that produced these datasets were performed in PBMCs and included at least two different clinical forms of infection, as

Abbreviations: ATLL, Adult T-Cell Lymphoma/Leukemia; HAM/TSP, HTLV-associated myelopathy/tropical spastic paraparesis; AC, Asymptomatic Carriers.

TABLE 1 | Description of the datasets used as the Discovery set.

Accession number	Reference	Symptomatology	Sample number	Tissue
GSE55851	Kobayashi et al. (2014)	Control	3	PBMCs
		Asymptomatic	6	PBMCs
		ATLL	12	PBMCs
GSE29312	Tattermusch et al. (2012)	Control	9	PBMCs
		Asymptomatic	20	PBMCs
		HAM/TSP	10	PBMCs
GSE29332	Tattermusch et al. (2012)	Control	8	PBMCs
		Asymptomatic	17	PBMCs
		HAM/TSP	10	PBMCs
Total		Control	20	PBMCs
		Asymptomatic	43	PBMCs
		ATLL	12	PBMCs
		HAM/TSP	20	PBMCs

well as controls (healthy individuals). When combined, the three datasets included 20 controls, 43 AC, 12 ATLL, and 20 HAM/TSP samples (**Table 1**). For our analysis, the AC samples were discarded to avoid possible classification bias, since this form can evolve to another clinical manifestation at some point during the patient's life, and no information regarding disease progression was provided. The remaining six datasets performed in other tissue types were used for *in silico* validation.

Data Retrieval, Pre-Processing, and Batch Correction

Raw expression data were downloaded from GEO/NCBI using the *GEOquery* package (Davis and Meltzer, 2007). Next, the *collapseRows* R function in the *WGCNA* package (Miller et al., 2011) was used to collapse the data, and only probes mapping to genes common to all datasets were maintained. Log transformation was applied to the expression data using the *preProcessCore* package (Bolstad, 2018), and outlier samples were identified and removed by the *ArrayQualityMetrics* package for R (Kauffmann et al., 2008). The *plyr* package was subsequently used to merge all data (Wickham, 2011). Following pre-processing, the combined dataset was submitted to a batch correction procedure using an empirical Bayes framework implemented in the *ComBat* function of the *sva* package (Leek et al., 2013), with clinical manifestations and original datasets as covariates. This allowed us to account for known or unknown sources of variation in the datasets, enabling the use of samples from different datasets in the integrated dataset (i.e., Discovery dataset). This method allowed for the inclusion of the maximum number of samples for analysis, in addition to more robust data interpretation, leading to the identification of consistent insights regarding biological phenomena. *ComBat* has been used in other studies and was shown to outperform other similar tools designed for this purpose (Chen et al., 2011). The final dataset consisted of 94 samples, with expression data pertaining to 10,533 genes in total.

Classification of HTLV Patient Clinical Manifestation *via* Decision Tree

A decision tree classification procedure was performed in the Discovery dataset to identify the key genes related to HTLV patient clinical manifestation (ATLL or HAM/TSP). Decision trees were constructed using the *rpart* package Therneau et al. (2015), which screens for the key factors that allow for the separation of the groups with maximum accuracy. To measure the performance of the classification model, areas under receiver operating characteristic (ROC) curves were calculated to determine a given model's sensitivity and specificity. The overall accuracy of a model is calculated by estimating the area under the curve (AUC), permitting measurements of the degree of class separability in a given model. Values approximating 1.0 indicate that the model is suitably capable of distinguishing among different classes. Finally, scatterplots were generated to visualize the dispersion of samples according to the model threshold in order to verify the accuracy estimated by ROC curve analysis.

Co-Expression and Enrichment Analysis of Genes Related to *CD40LG* and *GBP2*

A correlation matrix between the genes *CD40LG* and *GBP2* (identified as best classifiers) and all the genes within the Discovery dataset was constructed. Correlation was calculated separately for each group (control, ATLL, and HAM/TSP) using gene expression values measured as biweight midcorrelation coefficients, which function similarly to Pearson's r, except this technique is more robust with regard to data outliers (Langfelder and Horvath, 2012). Correlations were considered significant using a threshold of $|r| \geq 0.7$ and p-value ≤ 0.05. Next, correlated genes were clustered according to the functional terms of the REACTOME pathway database (https://reactome.org/). This enrichment analysis was performed using *clusterProfiler* Yu et al. (2012) with the following parameters: p-value threshold = 0.05, Q-value threshold = 0.05, minimum number of genes to cluster = 20, maximum number of genes to cluster = 500.

Description of Datasets Used for Validation

Six microarray expression datasets were retrieved from GEO: GSE17718 (Kress et al., 2010), GSE6034 (Hamamura et al., 2007), GSE38537 (Pinto et al., 2014), GSE33615 (Fujikawa et al., 2016), GSE57259 (Araya et al., 2014), and GSE19080 (no citation available at GEO/NCBI). To confirm the gene signature performance, we performed the gene model comparison in the validation dataset independently, without using the thresholds yielded by the decision tree model estimated during the discovery phase. The model comparison in each different dataset was obtained by applying a logistic regression fitting, which estimated the variable accuracy (CD40LG and GBP2), according to the response variable [determined by dataset metadata (HTLV status)]. Then, the ROC curve and the AUC were measured, which allows the comparison of the gene signature classification power across the validation datasets. A full description of the selected datasets is available in **Table S1**.

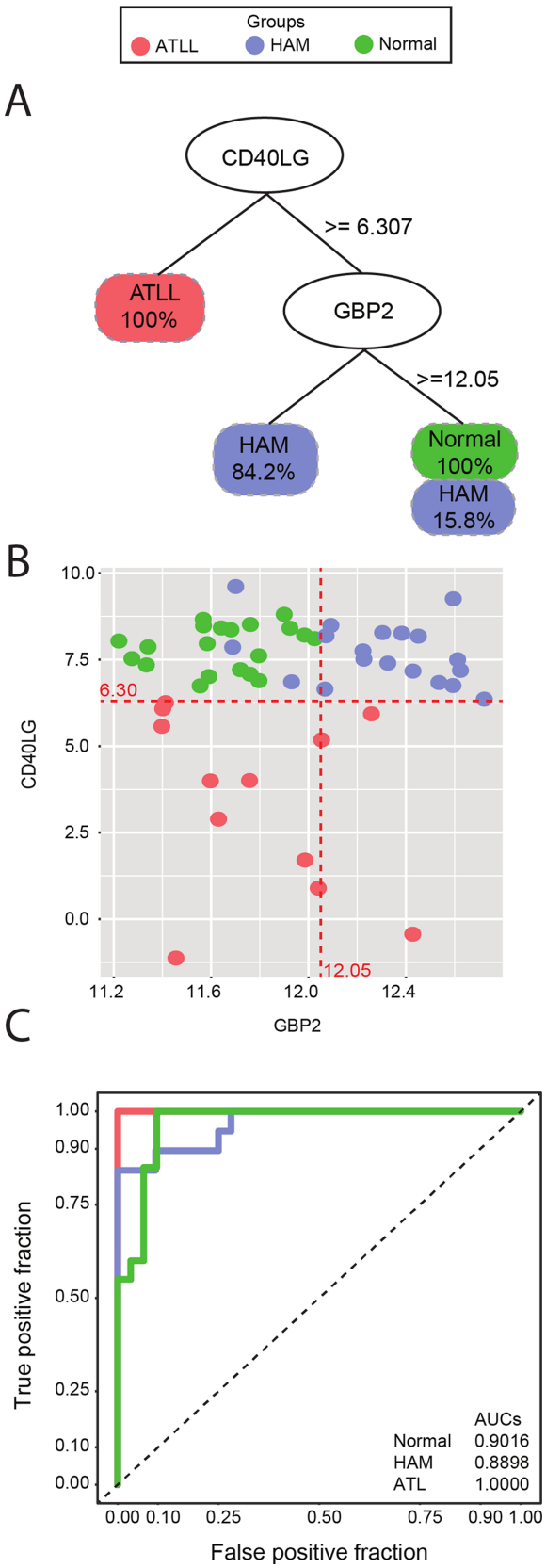


FIGURE 1 | Continued

FIGURE 1 | **(A)** Decision tree classification of three different symptomatologies using *CD40LG* to separate all ATLL samples from the others, and *GBP2* to separate 84.2% of the HAM/TSP samples from controls. **(B)** Scatterplot of *CD40LG* (Y axis) and *GBP2* (X axis) gene expression detailing the dispersion of the analyzed samples. Red lines represent the thresholds suggested by decision tree analysis. **(C)** ROC curve representing accuracy. An AUC of 0.9016 was found for the control group, 0.8898 for the HAM/TSP group, and 1.000 for ATLL. The red line represents the ATLL group, blue indicates HAM/TSP, and green is indicative of controls.

RESULTS

Gene Expression of *CD40LG* and *GBP2* Permits Accurate Discrimination of ATLL and HAM/TSP Patients

The decision tree algorithm identified two genes, *CD40LG* and *GBP2*, as the most informative in differentiating between the clinical manifestations of HTLV-infected samples and controls. The expression of *CD40LG* allowed for the discrimination of individuals with ATLL with 100% accuracy. To correctly classify the remaining samples (HAM/TSP and controls), a second gene (*GBP2*) was required. Expression levels of *GBP2* were able to discriminate HAM/TSP samples with 84.2% classification accuracy, and controls with 100% accuracy, with a 15.8% misclassification rate occurring between HAM/TSP and controls (**Figure 1A**). In addition, sample dispersion was visually checked by scatterplot using the log expression cutoffs returned by the decision tree algorithm: 6.30 for *CD40LG* and 12.05 for *GBP2* (**Figure 1B**). Finally, sensitivity and specificity were measured using ROC curve analysis, revealing high accuracy in discriminating among samples using genes *CD40LG* and *GBP2*: AUC of 0.90 for controls, 0.88 for HAM/TSP, and 1.00 for ATLL (**Figure 1C**).

Gene Expression of *CD40LG* and *GBP2* Correlate With Various Immune and Metabolic Pathways That Could Impact the Course of HTLV Infection

After evaluating the high predictive power of *CD40LG* and *GBP2* in discriminating HTLV clinical status, the roles played by these genes were investigated. Correlation analysis was performed considering global expression for each clinical manifestation (HAM/TSP or ATLL) and controls. Our results showed that 208 genes were significantly positively ($r > 0.7$ and p-value < 0.05) and 13 genes were significantly negatively ($r > 0.7$ and p-value < 0.05) correlated with *CD40LG*. Also, 84 genes were significantly positively and 1 gene was significantly negatively correlated with *GBP2*. In contrast, in the ATLL samples, 399 genes were significantly negatively correlated with *CD40LG* and 743 genes were significantly positively correlated with *GBP2*. A total of 12 genes were found to be correlated with both *CD40LG* and *GBP2* (*OAZ1, SLC39A11, NADK, TMED2, SLC38A5, P4HA1, HM13, MGAT2, HIST1H2BG, UQCRFS1, PTDSS1,* and *TAP1B*) (**Figure S1A**). In addition, the HAM/TSP samples presented 394 positive and 420 negative correlations, with three being associated with both *CD40LG* and *GBP2*

(*PWP1, H3F3A*, and *GNE*). In these samples, correlations with *CD40LG* were mostly positive, with 367 positive correlations, while those with *GBP2* were mostly negative, with 230 negative correlations (**Figure S1B**). More comprehensive information regarding this correlation analysis and the commonly observed genes is available as supplementary material (**Tables S2–S4**). The gene set previously identified correlated with the biomarkers (*CD40LG* and *GBP2*) was analyzed in order to identify their enriched pathways. Thus, the top four pathways identified from being negatively correlated with the *CD40LG* gene set in the HAM/TSP were "Neutrophil degranulation," "Signaling by interleukins," "TRAF6-mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation," and "Toll Like Receptor 7/8 (TLR7/8) Cascade." The main pathways identified from the gene set that negatively correlated with GBP2 in the HAM/TSP were "SUMO E3 ligases SUMOylate target proteins," "SUMOylation," "rRNA processing," and "tRNA processing" (**Figure 2B**). Only one pathway was identified from the gene set that positively correlated with *CD40LG* in HAM/TSP: "SUMOylation of DNA replication proteins." Several pathways were identified from the genes that were positively correlated with *GBP2* in HAM/TSP: "Interferon Signaling," "Interferon alpha/beta signaling," "Activation of G protein gated Potassium channels," "G protein gated Potassium channels," and "Interleukin-20 family signaling" (**Figure 2A**).

The top 5 pathways identified from the gene set that negatively correlated with *CD40LG* in the ATLL were "MAPK family signaling cascades," "MAPK1/MAPK3 signaling," "RAF/MAP kinase cascade," "Mitotic G1–G1/S phases," and "G1/S Transition" (**Figure 2B**). Moreover, the associated pathways from the gene set that positively correlated with *GBP2* in ATLL patients were "tRNA processing in the nucleus," "tRNA processing," "Viral Messenger RNA synthesis," "Late Phase of HIV Life Cycle," and "HIV Life Cycle" (**Figure 2A**).

By contrast, in the control group, the pathways identified from the gene set that correlated with *CD40LG* were "Processing of Capped Intron-Containing Pre-mRNA," "tRNA processing in the nucleus," "tRNA processing," "Viral Messenger RNA Synthesis," "Dual incision in TC-NER," "Transcription-Coupled Nucleotide Excision Repair (TC-NER)," "Late Phase of HIV Life Cycle," "mRNA Splicing—Major Pathway," "HIV Life Cycle," "Synthesis of DNA," "SUMOylation of DNA replication proteins," and "HIV infection." With regard to GBP2's positively correlated genes, the following pathways were found in the control group: "Neutrophil degranulation," "Metabolism of water-soluble

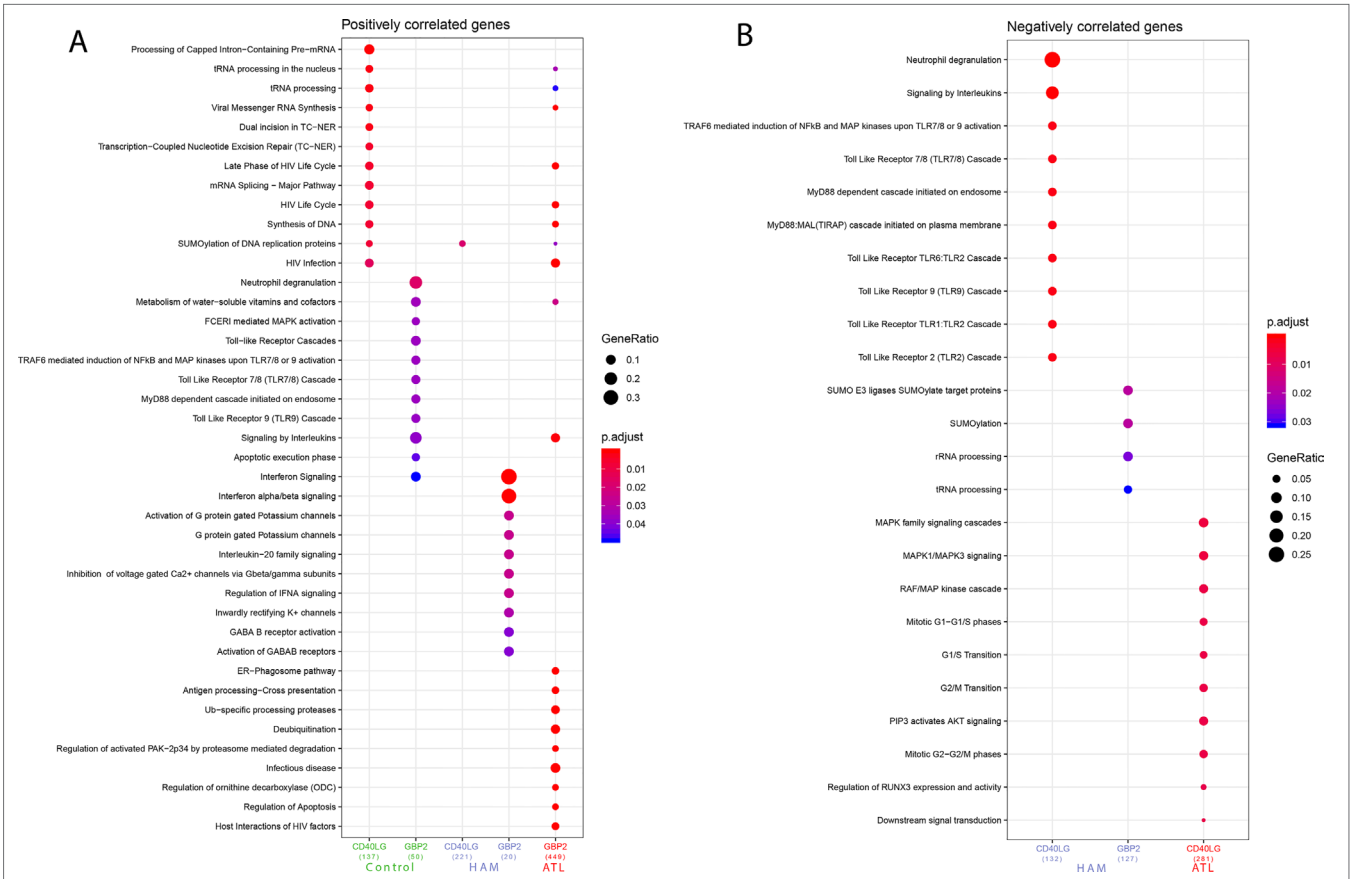


FIGURE 2 | (A) Pathways associated with genes found to be positively correlated with CD40LG and GBP2, grouped according to symptomatology. **(B)** Pathways associated with genes found to be negatively correlated with CD40LG and GBP2, grouped according to symptomatology. Analysis performed using the following parameters: p-value = 0.05, q-value = 0.2, minimum number of genes to cluster = 20, maximum number of genes to cluster = 500.

vitamins and cofactors," "FCERI mediated MAPK activation," "Toll-Like Receptors Cascades," "TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation," "Toll Like Receptor 7/8 (TLR7/8) Cascade," "MyD88 dependent cascade initiated on endosome," "Toll Like Receptor 9 (TLR9) Cascade," "Signaling by Interleukins," "Apoptotic execution phase," and "Interferon signaling" (**Figure 2A**). Further information regarding the pathways associated with these genes (ENTREZ ID) is available as supplementary material, separated into negatively correlated (**Table S5**) and positively correlated categories (**Table S6**).

Validation of *CD40LG* and *GBP2* in Independent Datasets Reveals Classification Robustness in Different Tissue Types

To validate the accuracy of our two-gene model in the discrimination of ATLL, HAM/TSP, and control samples, this model was applied to the other datasets not used in the discovery set: (Kress et al., 2010) (GSE17718), (Hamamura et al., 2007) (GSE6034), (Pinto et al., 2014) (GSE38537), (Yamagishi et al., 2012) (GSE33615), (Olière et al., 2010) (GSE57259), and GSE19080. After downloading and pre-processing these datasets, ROC curve analysis was applied to measure the discriminant power of *CD40LG* and *GBP2* in classifying HLTV-1 clinical manifestations. The discriminant power of this two-gene signature was found to be very high, allowing for the discrimination of the HTLV-1 clinical status in five of the datasets with an AUC value of 1 (GSE17718, GSE6034, GSE38537, GSE33615, and GSE57259). The need to include both genes for accurate classification was evidenced in the GSE19080 dataset (in which the *CD40LG* gene is absent), yielding a much lower AUC (0.875) in the discrimination of control samples, compared to 0.666 for HAM/TSP samples and 0.5 when discriminating ATLL samples. These validation datasets were derived from a variety of tissues, such as cell lines (StEd, MT-2, Tay and MT-4), CD4 lymphocytes, and PBMCs. The overall accuracy of this two-gene signature model is delineated in **Table 2**. Also, the sample distribution using the two-gene expression in all validation dataset is summarized in **Figure S2**.

DISCUSSION

To date, few studies have attempted to identify biomarkers capable of discriminating between ATLL and HAM/TSP in HTLV-1 infection. A previous report (Sato et al., 2013) suggested three potential prognostic biomarkers in cerebrospinal fluid for HAM/TSP disease progression: CXCL10, CXCL9, and neopterin. Another study (Baratella et al., 2017) stated that the HBZ protein, exclusively localized in the cytoplasm, could be a biomarker of HAM/TSP. In addition, CAN-2 and SPTA-2 were identified as biomarkers capable of discriminating ATLL (Ishihara et al., 2013). However, these biomarkers were found in a specific population and, to the best of our knowledge, the literature contains no sets of biomarkers offering sufficient accuracy to reliably identify both the ATLL and HAM/TSP phenotypes. With the objective of achieving accurate discrimination, we employed a robust bioinformatic approach to consolidate the available expression data using three different datasets combined into a single Discovery dataset. Three studies were selected for this analysis, one submitted by Kobayashi et al. (acc number: GSE55851) and two submitted by Tattermusch et al. (acc number: GSE29332 and GSE29312). The study by Kobayashi et al. compares gene expression levels in PBMCs from ATLL, asymptomatic, and control patients. The other studies submitted by Tattermusch et al. compared gene expression levels in PBMCs from HAM/TSP, asymptomatic, and control individuals. Next, a data mining technique was applied to the merged, batch-corrected Discovery dataset to identify which variables (genes) could effectively discriminate clinical status among the samples. Decision tree analysis revealed genes *CD40LG* and *GBP2* as discriminators of ATLL and HAM/TSP, offering accuracy rates of 100% and 84.2%, respectively. A previous report identified lower *CD40LG* expression in cells expressing PTHrP and MIP-1α, two proteins associated with ATLL progression (Shu et al., 2012). The second marker identified herein, *GBP2*, was previously associated with tax protein activity in HTLV-1 (Arainga et al., 2012). Despite identifying these associations, no previous studies proposed either of these genes as biomarkers of ATLL or HAM/TSP symptomatology.

The *CD40LG* gene encodes a protein located on the surface of T cells and exerts the role of regulating B cell functions (Stelzer et al., 2016). *GBP2* is a guanylate binding protein induced

TABLE 2 | Performance of the two-gene signature classifying the samples from validation datasets.

Accession number	Symptomatology	Tissue	Biomarkers	AUC
GSE17718	Control	CD4+ Lymphocyte	CD40LG and GBP2	1.00
	ATLL	Cell lines StEd and MT-2	CD40LG and GBP2	1.00
GSE6034	Control	CD4+ Lymphocyte	CD40LG and GBP2	1.00
	ATLL	Cell lines TaY, MT-2 and MT-4	CD40LG and GBP2	1.00
GSE38537	Control	CD4+ Lymphocyte	CD40LG and GBP2	1.00
	HAM/TSP	CD4+ Lymphocyte	CD40LG and GBP2	1.00
GSE33615	Control	CD4+ Lymphocyte	CD40LG and GBP2	1.00
	ATLL	PBMCs (Mostly CD4+ Lymphocytes)	CD40LG and GBP2	1.00
GSE19080	Control	CD4+ Lymphocyte	GBP2	0.87
	ATLL	CD4+ Lymphocyte	GBP2	0.50
	HAM/TSP	CD4+ Lymphocyte	GBP2	0.66
GSE57259	Control	CD4+ CD25+ CCR4+ Lymphocytes	CD40LG and GBP2	1.00
	HAM/TSP	CD4+ CD25+ CCR4+ Lymphocytes	CD40LG and GBP2	1.00
	ATLL	CD4+ CD25+ CCR4+ Lymphocytes	CD40LG and GBP2	1.00

by IFN-γ and is considered as a control factor for tumor cell proliferation and spreading (Messmer-Blust et al., 2010). Our functional approach entailed the correlation of these biomarkers with the global expression of other genes, followed by enrichment analysis using the REACTOME database (Fabregat et al., 2018). This analysis showed that the genes positively correlated with *CD40LG* are associated with pathways mainly related to tRNA processing, viral replication, and mRNA splicing in the control group. However, in the HAM/TSP group, these genes were only found to be associated with the SUMOylation of DNA replication pathway, which is specifically associated with transcription and replication pathways. In addition, the genes negatively correlated with *CD40LG* were found to be associated primarily with neutrophil degranulation, signaling for interleukins and several cascades of Toll Like Receptors in HAM/TSP patients. These pathways may be associated with immune responses involving inflammation (Faurschou and Borregaard, 2003; Lacagnina et al., 2018; Weitzman, 2003), which is frequently observed in HAM/TSP patients (Nakagawa et al., 1995).

On the other hand, the genes negatively correlated with *CD40LG* were found to be associated with MAPK cascade-associated pathways and cell cycle-related pathways. MAPK cascade-related pathways are associated with a wide spectrum of metabolic pathways related to cell proliferation, differentiation, and apoptosis (Shaul and Seger, 2007). Cell cycle-related pathways, such as Mitotic G1-G1/S phases, G1/S Transition, G2/M Transition, and Mitotic G2-G2/M phases, are related to cell proliferation (Matson and Cook, 2017). These pathways are all related to cell proliferation, which is consistent with ATLL symptomatology and the uncontrolled proliferation of T cells (Shimoyama and members of The Lymphoma Study Group (1984–87)*, 1991).

The top pathways that positively correlated with *GBP2* were mainly related to HIV infection, tRNA, and viral mRNA processing and synthesis, signaling by interleukins, and apoptosis regulation. The pathways observed to be related to HIV infection may be due to similarities between HTLV-1 and HIV, as both these retroviruses mainly infect T CD4+ lymphocytes. The tRNA and viral mRNA pathways are associated with the highly active processing of RNAs that occurs in ATLL cells. Furthermore, the regulation of apoptosis could be associated with the immortalization of T CD4+ cells that characterizes the leukemic aspect of ATLL (Bellon et al., 2010).

In order to evaluate the predictive power of the *CD40LG/GBP2* two-gene signature in the accurate classification of HAM/TSP and ATLL samples, we conducted a validation step using independent datasets, which revealed excellent predictive values. The majority of datasets returned an AUC of 1.0, corresponding to an accuracy rate of 100% when classifying samples as ATLL, HAM/TSP, or controls. In one of six validation datasets (GSE19080), a poorer classification accuracy was found, which is likely due to the absence of the *CD40LG* in the array, indicating the requirement of both genes in order to maintain reliably consistent classification. Additionally, the selected validation datasets sampled not only PBMCs but also several transformed cell lines, including MT-2, MT-4, StEd, and TaY, as well as isolated CD4+ cells. These high rates of accuracy seen in a diverse range of tissue types serve to confirm the robustness of the two-gene signature identified herein, suggesting a conserved mechanism in the regulation of genes associated with each symptomatology. Despite some limitations such as the absence of available datasets studying HTLV-1 biomarkers in a transcriptional approach and the reduced sample numbers, our findings provide useful biomarkers to independently identify populations affected by HTLV-1.

CONCLUSION

Our meta-analysis of gene expression datasets in HTLV-1-infected patients with specific disease manifestations identified a two-gene signature (*CD40LG/GBP2*) allowing for excellent classification of the HAM/TSP and ATLL phenotypes. This signature was subsequently validated in six independent datasets. An exploratory functional enrichment analysis of the genes found to be positively and negatively correlated with this signature revealed diverse activation and repression of pathways relevant to this viral disease. Our findings add to the accumulation of knowledge surrounding HTLV-1 infection and may contribute to early diagnosis, as well as the treatment of related symptomatologies.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE55851, GSE29312,GSE29332,GSE17718, GSE6034,GSE38537,GSE33615,GSE57259,GSE19080.

AUTHOR CONTRIBUTIONS

EF, AQ, KF, MR and PR participated in the data analysis. EF, AQ, KF and PR participated in the manuscript writing. JK, LA, JL and HJ participated in the idea generation for this work.

FUNDING

AQ acknowledges financial support from the program Inova Fiocruz (Project number VPPIS-001-FIO18).

ACKNOWLEDGMENTS

We thank Mr. Olival Rocha for his assistance. The authors would also like to thank Andris K. Walter for English language revision and manuscript copyediting assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01056/full#supplementary-material

TABLE S1 | Datasets used on the validation step's detailed information

TABLE S2 | Correlations observed between CD40LG and GBP2 in the HAM/TSP group.

TABLE S3 | Correlations observed between CD40LG and GBP2 in the ATLL group.

TABLE S4 | Correlations observed between CD40LG and GBP2 in the control group.

TABLE S5 | Detailed information regarding the negatively correlated pathways and associated genes (ENTREZ ID).

TABLE S6 | Detailed information regarding the positively correlated pathways and associated genes (ENTREZ ID).

FIGURE S1 | Correlation network based on gene expression values in ATLL samples. Highlighted genes were found to correlate with both CD40LG and GBP2. B - Correlation network based on the gene expression values in the HAM/TSP group. Highlighted genes were found to correlate with both CD40LG and GBP2. C - Correlation network based on the gene expression values in the control group.

FIGURE S2 | Scatterplot of validation datasets sample distribution using the CD40LG and GBP2's log transformed expression values. The samples can be separated by symptomatology [ATLL (green), HAM/TSP (red) and control (blue)], this separation is shown by the collored ellipses. The GSE19080's scatterplot has only GBP2 within the dataset, the values of X and Y axis are both representing GBP2's log transformed expression value.

REFERENCES

Arainga, M., Murakami, H., and Aida, Y. (2012). Visualizing spatiotemporal dynamics of apoptosis after G1 arrest by human T cell leukemia virus type 1 Tax and insights into gene expression changes using microarray-based gene expression analysis. *BMC Genomics* 13, 275. doi: 10.1186/1471-2164-13-275

Araya, N., Sato, T., Ando, H., Tomaru, U., Yoshida, M., Coler-Reilly, A., et al. (2014). HTLV-1 induces a Th1-like state in CD4+CCR4+ T cells. *J. Clin. Invest.* 124, 3431–3442. doi: 10.1172/JCI75250

Baratella, M., Forlani, G., Raval, G. U., Tedeschi, A., Gout, O., Gessain, A., et al. (2017). Cytoplasmic localization of HTLV-1 HBZ protein: a biomarker of HTLV-1-Associated myelopathy/tropical spastic paraparesis (HAM/TSP). *PLoS Negl. Trop. Dis.* 11, e0005285. doi: 10.1371/journal.pntd.0005285

Bellon, M., Baydoun, H. H., Yao, Y., and Nicot, C. (2010). HTLV-I Tax-dependent and -independent events associated with immortalization of human primary T lymphocytes. *Blood* 115, 2441–2448. doi: 10.1182/blood-2009-08-241117

Bolstad, B. (2018). preprocessCore: A collection of pre-processing functions. *R package version 1.44.0*. Available at: https://github.com/bmbolstad/preprocessCore.

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., et al. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 6, e17238. doi: 10.1371/journal.pone.0017238

Davis, S., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Debray, T. P. A., Damen, J. A. A. G., Snell, K. I. E., Ensor, J., Hooft, L., Reitsma, J. B., et al. (2017). A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 356, i6460. doi: 10.1136/bmj.i6460

Enose-Akahata, Y., Abrams, A., Johnson, K. R., Maloney, E. M., and Jacobson, S. (2012). Quantitative differences in HTLV-I antibody responses: classification and relative risk assessment for asymptomatic carriers and ATL and HAM/TSP patients from Jamaica. *Blood* 119, 2829–2836. doi: 10.1182/blood-2011-11-390807

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., et al. (2018). The reactome pathway knowledgebase. *Nucleic Acids Res.* 46, D649–D655. doi: 10.1093/nar/gkx1132

Faurschou, M., and Borregaard, N. (2003). Neutrophil granules and secretory vesicles in inflammation. *Microbes Infect.* 5, 1317–1327. doi: 10.1016/j.micinf.2003.09.008

Fujikawa, D., Nakagawa, S., Hori, M., Kurokawa, N., Soejima, A., Nakano, K., et al. (2016). Polycomb-dependent epigenetic landscape in adult T-cell leukemia. *Blood* 127, 1790–1802. doi: 10.1182/blood-2015-08-662593

Fukutani, K. F., Kasprzykowski, J. I., Paschoal, A. R., Gomes, M., de, S., Barral, A., et al. (2017). Meta-analysis of expression datasets: comparing virus infection and blood-fed transcriptomes to identify markers of virus presence. *Front. Bioeng. Biotechnol.* 5, 84. doi: 10.3389/fbioe.2017.00084

Galvão-Castro, B., Loures, L., Rodriques, L. G., Sereno, A., Ferreira Júnior, O. C., Franco, L. G., et al. (1997). Distribution of human T-lymphotropic virus type I among blood donors: a nationwide Brazilian study. *Transfusion* 37, 242–243. doi: 10.1046/j.1537-2995.1997.37297203532.x

Gessain, A., and Mahieux, R. (2012). Tropical spastic paraparesis and HTLV-1 associated myelopathy: clinical, epidemiological, virological and therapeutic aspects. *Rev. Neurol.* 168, 257–269. doi: 10.1016/j.neurol.2011.12.006

GSE19080, Hernandez, E., and Oliere, S. (2010). Gene expression profiling in patients infected with HTLV-1: Identification of ATL and HAM/TSP-specific genetic profiles. *Gene Expression Omnibus*. GSE19080.

Hamamura, R. S., Ohyashiki, J. H., Kurashina, R., Kobayashi, C., Zhang, Y., Takaku, T., et al. (2007). Induction of heme oxygenase-1 by cobalt protoporphyrin enhances the antitumour effect of bortezomib in adult T-cell leukaemia cells. *Br. J. Cancer* 97, 1099–1105. doi: 10.1038/sj.bjc.6604003

Hounkpe, B. W., Fiusa, M. M. L., Colella, M. P., da Costa, L. N. G., Benatti, R., de, O., et al. (2015). Role of innate immunity-triggered pathways in the pathogenesis of Sickle Cell Disease: a meta-analysis of gene expression studies. *Sci. Rep.* 5, 17822. doi: 10.1038/srep17822

Ishihara, M., Araya, N., Sato, T., Tatsuguchi, A., Saichi, N., Utsunomiya, A., et al. (2013). Preapoptotic protease calpain-2 is frequently suppressed in adult T-cell leukemia. *Blood* 121, 4340–4347. doi: 10.1182/blood-2012-08-446922

Jha, P. K., Vijay, A., Sahu, A., and Ashraf, M. Z. (2016). Comprehensive Gene expression meta-analysis and integrated bioinformatic approaches reveal shared signatures between thrombosis and myeloproliferative disorders. *Sci. Rep.* 6, 37099. doi: 10.1038/srep37099

Kagdi, H., Demontis, M. A., Ramos, J. C., and Taylor, G. P. (2018). Switching and loss of cellular cytokine producing capacity characterize *in vivo* viral infection and malignant transformation in human T- lymphotropic virus type 1 infection. *PLoS Pathog.* 14, e1006861. doi: 10.1371/journal.ppat.1006861

Kauffmann, A., Gentleman, R., and Huber, W. (2008). arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* 25, 415–416. doi: 10.1093/bioinformatics/btn647

Kobayashi, S., Nakano, K., Watanabe, E., Ishigaki, T., Ohno, N., Yuji, K., et al. (2014). CADM1 expression and stepwise downregulation of CD7 are closely associated with clonal expansion of HTLV-I-infected cells in adult T-cell leukemia/lymphoma. *Clin. Cancer Res.* 20, 2851–2861. doi: 10.1158/1078-0432.CCR-13-3169

Kress, A. K., Schneider, G., Pichler, K., Kalmer, M., Fleckenstein, B., and Grassmann, R. (2010). Elevated cyclic AMP levels in T lymphocytes transformed by human T-cell lymphotropic virus type 1. *J. Virol.* 84, 8732–8742. doi: 10.1128/JVI.00487-10

Lacagnina, M. J., Watkins, L. R., and Grace, P. M. (2018). Toll-like receptors and their role in persistent pain. *Pharmacol. Ther.* 184, 145–158. doi: 10.1016/j.pharmthera.2017.10.006

Langfelder, P., and Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* 46. doi: 10.18637/jss.v046.i11

Leek, J. T., Johnson, W. E., Parker, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., et al. (2013). sva: Surrogate variable analysis. *R Package Version* 3.

Matson, J. P., and Cook, J. G. (2017). Cell cycle proliferation decisions: the impact of single cell analyses. *FEBS J.* 284, 362–375. doi: 10.1111/febs.13898

Matutes, E. (2007). Adult T-cell leukaemia/lymphoma. *J. Clin. Pathol.* 60, 1373–1377. doi: 10.1136/jcp.2007.052456

Messmer-Blust, A. F., Balasubramanian, S., Gorbacheva, V. Y., Jeyaratnam, J. A., and Vestal, D. J. (2010). The interferon-gamma-induced murine guanylate-binding protein-2 inhibits rac activation during cell spreading on fibronectin and after platelet-derived growth factor treatment: role for phosphatidylinositol 3-kinase. *Mol. Biol. Cell* 21, 2514–2528. doi: 10.1091/mbc.e09-04-0344

Miller, J. A., Cai, C., Langfelder, P., Geschwind, D. H., Kurian, S. M., Salomon, D. R., et al. (2011). Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinf.* 12, 322. doi: 10.1186/1471-2105-12-322

Mirvish, E. D., Pomerantz, R. G., and Geskin, L. J. (2011). Infectious agents in cutaneous T-cell lymphoma. *J. Am. Acad. Dermatol.* 64, 423–431. doi: 10.1016/j.jaad.2009.11.692

Nakagawa, M., Izumo, S., Ijichi, S., Kubota, H., Arimura, K., Kawabata, M., et al. (1995). HTLV-I-associated myelopathy: analysis of 213 patients based on clinical features and laboratory findings. *J. Neurovirol.* 1, 50–61. doi: 10.3109/13550289509111010

Olière, S., Hernandez, E., Lézin, A., Arguello, M., Douville, R., Nguyen, T. L.-A., et al. (2010). HTLV-1 evades type I interferon antiviral signaling by inducing the suppressor of cytokine signaling 1 (SOCS1). *PLoS Pathog.* 6, e1001177. doi: 10.1371/journal.ppat.1001177

O'Mara, T. A., Zhao, M., and Spurdle, A. B. (2016). Meta-analysis of gene expression studies in endometrial cancer identifies gene expression profiles associated with aggressive disease and patient outcome. *Sci. Rep.* 6, 36677. doi: 10.1038/srep36677

Osame, M., Usuku, K., Izumo, S., Ijichi, N., Amitani, H., Igata, A., et al. (1986). HTLV-I associated myelopathy, a new clinical entity. *Lancet* 1, 1031–1032. doi: 10.1016/S0140-6736(86)91298-5

Pinto, M. T., Malta, T. M., Rodrigues, E. S., Pinheiro, D. G., Panepucci, R. A., Malmegrim de Farias, K. C. R., et al. (2014). Genes related to antiviral activity, cell migration, and lysis are differentially expressed in CD4(+) T cells in human t cell leukemia virus type 1-associated myelopathy/tropical spastic paraparesis patients. *AIDS Res. Hum. Retroviruses* 30, 610–622. doi: 10.1089/aid.2013.0109

Sato, T., Coler-Reilly, A., Utsunomiya, A., Araya, N., Yagishita, N., Ando, H., et al. (2013). CSF CXCL10, CXCL9, and neopterin as candidate prognostic biomarkers for HTLV-1-associated myelopathy/tropical spastic paraparesis. *PLoS Negl. Trop. Dis.* 7, e2479. doi: 10.1371/journal.pntd.0002479

Shaul, Y. D., and Seger, R. (2007). The MEK/ERK cascade: from signaling specificity to diverse functions. *Biochim. Biophys. Acta* 1773, 1213–1226. doi: 10.1016/j.bbamcr.2006.10.005

Shimoyama, M., and members of The Lymphoma Study Group (1984–87)* (1991). Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. *Br. J. Haematol.* 79, 428–437. doi: 10.1111/j.1365-2141.1991.tb08051.x

Shu, S. T., Dirksen, W. P., Lanigan, L. G., Martin, C. K., Thudi, N. K., Werbeck, J. L., et al. (2012). Effects of parathyroid hormone-related protein and macrophage inflammatory protein-1α in Jurkat T-cells on tumor formation *in vivo* and expression of apoptosis regulatory genes *in vitro*. *Leuk. Lymphoma* 53, 688–698. doi: 10.3109/10428194.2011.626883

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards Suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinf.* 54, 1.30.1–1.30.33. doi: 10.1002/cpbi.5

Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., and Khatri, P. (2017). Methods to increase reproducibility in differential gene expression *via* meta-analysis. *Nucleic Acids Res.* 45, e1. doi: 10.1093/nar/gkw797

Tattermusch, S., Skinner, J. A., Chaussabel, D., Banchereau, J., Berry, M. P., McNab, F. W., et al. (2012). Systems biology approaches reveal a specific interferon-inducible signature in HTLV-1 associated myelopathy. *PLoS Pathog.* 8, e1002480. doi: 10.1371/journal.ppat.1002480

Therneau, T., Atkinson, B., and Ripley, B. (2015). *rpart: Recursive Partitioning and Regression Trees. R package version 4*. pp. 1–10.

Uchiyama, T., Yodoi, J., Sagawa, K., Takatsuki, K., and Uchino, H. (1977). Adult T-cell leukemia: clinical and hematologic features of 16 cases. *Blood* 50, 481–492. doi: 10.1182/blood.V50.3.481.bloodjournal503481

Weitzman, J. (2003). Interleukins in inflammation. *Genome Biol.* 4, spotlight–20030217. doi: 10.1186/gb-spotlight-20030217-01

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *J. Stat. Softw.* 40. doi: 10.18637/jss.v040.i01

Xue, S., Song, G., and Yu, J. (2017). The prognostic significance of PD-L1 expression in patients with glioma: A meta-analysis. *Sci. Rep.* 7, 4231. doi: 10.1038/s41598-017-04023-x

Yamada, Y., Tomonaga, M., Fukuda, H., Hanada, S., Utsunomiya, A., Tara, M., et al. (2001). A new G-CSF-supported combination chemotherapy, LSG15, for adult T-cell leukaemia-lymphoma: Japan Clinical Oncology Group Study 9303. *Br. J. Haematol.* 113, 375–382. doi: 10.1046/j.1365-2141.2001.02737.x

Yamagishi, M., Nakano, K., Miyake, A., Yamochi, T., Kagami, Y., Tsutsumi, A., et al. (2012). Polycomb-mediated loss of miR-31 activates NIK-dependent NF-κB pathway in adult T cell leukemia and other cancers. *Cancer Cell* 21, 121–135. doi: 10.1016/j.ccr.2011.12.015

Yamamoto-Taguchi, N., Satou, Y., Miyazato, P., Ohshima, K., Nakagawa, M., Katagiri, K., et al. (2013). HTLV-1 bZIP factor induces inflammation through labile Foxp3 expression. *PLoS Pathog.* 9, e1003630. doi: 10.1371/journal.ppat.1003630

Yamano, Y., and Sato, T. (2012). Clinical pathophysiology of human T-lymphotropic virus-type 1-associated myelopathy/tropical spastic paraparesis. *Front. Microbiol.* 3, 389. doi: 10.3389/fmicb.2012.00389

Yasuma, K., Matsuzaki, T., Yamano, Y., Takashima, H., Matsuoka, M., and Saito, M. (2016). HTLV-1 subgroups associated with the risk of HAM/TSP are related to viral and host gene expression in peripheral blood mononuclear cells, independent of the transactivation functions of the viral factors. *J. Neurovirol.* 22, 416–430. doi: 10.1007/s13365-015-0407-2

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* 16, 284–287 doi: 10.1089/omi.2011.0118

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 November 2019
doi: 10.3389/fgene.2019.01147

A Novel Approach Based on Bipartite Network Recommendation and KATZ Model to Predict Potential Micro-Disease Associations

Shiru Li[1], *Minzhu Xie*[1]* *and Xinqiu Liu*[2]

[1] College of Information Science and Engineering, Hunan Normal University, Changsha, China, [2] Hunan Vocational College of Engineering, Changsha, China

Accumulating evidence indicates that the microbes colonizing human bodies have crucial effects on human health and the discovery of disease-related microbes will promote the discovery of biomarkers and drugs for the prevention, diagnosis, treatment, and prognosis of diseases. However clinical experiments of disease-microbe associations are time-consuming, laborious and expensive, and there are few methods for predicting potential microbe-disease association. Therefore, developing effective computational models utilizing the accumulated public data of clinically validated microbe-disease associations to identify novel disease-microbe associations is of practical importance. We propose a novel method based on the KATZ model and Bipartite Network Recommendation Algorithm (KATZBNRA) to discover potential associations between microbes and diseases. We calculate the Gaussian interaction profile kernel similarity of diseases and microbes based on validated disease-microbe associations. Then, we construct a bipartite graph and execute a bipartite network recommendation algorithm. Finally, we integrate the disease similarity, microbe similarity and bipartite network recommendation score to obtain the final score, which is used to infer whether there are some novel disease-microbe interactions. To evaluate the predictive power of KATZBNRA, we tested it with the walk length 2 using global leave-one-out cross validation (LOOV), two-fold and five-fold cross validations, with AUCs of 0.9098, 0.8463 and 0.8969, respectively. The test results also show that KATZBNRA is more accurate than two recent similar methods KATZHMDA and BNPMDA.

Keywords: microbe, disease, KATZ model, bipartite network recommendation, Gaussian interaction profile kernel similarity

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Xiangxiang Zeng,
Xiamen University, China
Wen Zhang,
Huazhong Agricultural University, China

***Correspondence:**
Minzhu Xie
xieminzhu@hunnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *19 July 2019*
Accepted: *21 October 2019*
Published: *15 November 2019*

Citation:
Li S, Xie M and Liu X (2019) A Novel Approach Based on Bipartite Network Recommendation and KATZ Model to Predict Potential Micro-Disease Associations. Front. Genet. 10:1147. doi: 10.3389/fgene.2019.01147

INTRODUCTION

A microbe is a microscopic organism, including bacteria, eukaryotes, archaea, and viruses (Wu et al., 2018). Various types of microbes live on or in different parts of a human body such as the skin, mouth, hair, stomach, and gastrointestinal tract. An adult human body contains a large number of bacterial cells, which is estimated to reach 10^{14} and much more than the total number of human cells, with more than 5 million microbe genes, outnumbering the human genes by more than 100 fold (Sommer and Backhed, 2013). Most microbes are harmless and some are beneficial to humans

(Grice and Segre, 2011). Recently, accumulated experimental evidence shows that microbes have an important impact on human health, nutrient absorption, immune response, cancer control, and the prevention of pathogen colonization (Wu et al., 2018). For example, the gut microbiota could significantly contribute nutrition absorption by producing indispensable vitamins and decomposing indigestible polysaccharides, and it also has an important impact on the mucus layer, the balance of antimicrobial peptides, and immunoglobulin A, and the differentiation and activation of some lymphocyte populations (Sommer and Backhed, 2013). Therefore, the gut microbiota is thought to be an extra 'organ' of humans (Gill et al., 2006). But some microbes may contribute to disease, such as psoriasis and inflammatory bowel disease (IBD). There have been reports that psoriasis occurs after strep throat and could worsen due to the colonization of Candida albicans, Malassezia, and Staphylococcus aureus on the skin or in the gut (Fry and Baker, 2007). Aroniadis et. al. (Aroniadis and Brandt, 2013) indicated that the biodiversity of bacteria, such as Bacteroidetes and Firmicutes, colonizing in individuals affected by IBD has been found to be reduced by 30 to 50%. Wang et. al. (Wang and Jia, 2016) showed that the gut microbiota's dysbiosis might be a key environmental risk factor of many human diseases, though it's difficult to reveal the true causality.

To explore the relationship between microbes and their human hosts, scientists from many countries collaborated and launched the Human Microbiome Project (HMP) (Human Microbiome Project, 2012a). Recently, high-throughput sequencing techniques and corresponding software packages have been developed rapidly, and a growing number of research analyses have been carried out on the microbiome, such as whole-genome shotgun (WGS), 16S, and the taxonomic profiling (Human Microbiome Project, 2012b), and have demonstrated significant associations between microbes and complex human diseases such as rheumatoid arthritis, colorectal cancer, obesity, and type 2 diabetes (Wang and Jia, 2016). However, these studies involve time-consuming and expensive biological experiments. Therefore, it is necessary to utilize the known information to predict the unknown microbe-disease interactions. Identifying microbe-disease interactions could promote discovering biomarkers and drugs for the prevention, diagnosis, treatment, and prognosis of diseases. Now, more and more computer algorithms (Chen and Zhang, 2013; Yang et al., 2014; Zhang et al., 2017; Zeng et al., 2018; Zhang et al., 2018a; Zhang et al., 2018b; Zhang et al., 2018c; Zhang et al., 2018d; Zeng et al., 2019) have been proposed for interaction prediction of miRNA-disease, lncRNA-disease, and drug-drug, and it is feasible to apply these methods to the microbe-disease association prediction field.

Recently, Ma et al. (2017) collected microbe-disease association data from previous published studies and constructed the Human Microbe-Disease Association Database (HMDAD). Based on the data from HMDAD, some microbe-disease association prediction methods have been proposed. Chen et al. (2017) used a KATZ measure to predict human microbe-disease association, and proposed an algorithm named KATZHMDA. KATZHMDA can predict new microbe-disease associations at a large scale. Bao et al. (2017) used network consistency projection and introduced an algorithm NCPHMD to predict human microbe-disease association. NCPHMD deals with unknown diseases or microbes that are not present in the disease-microbe databases. He et al. (He et al., 2018) presented an algorithm GRNMFHMDA. GRNMFHMDA assigns likelihood scores to unknown microbe-disease pairs by calculating weighted K nearest neighbor profiles of microbes and diseases, and then adapts the standard non-negative matrix factorization by integrating graph Laplacian and Tikhonov (L2) regularization to obtain a microbe-disease association prediction score matrix. Zou et al. (2017) designed an approach BiRWHMDA. BiRWHMDA constructs a heterogeneous network by connecting the microbe similarity network and the disease similarity network based on known microbe-disease associations, and then uses a bi-random walk to predict microbe-disease association.

In the paper, we propose a novel approach to predict potential micro-disease association based on the KATZ measure and bipartite network recommendation algorithm (KATZBNRA), which is an improvement on KATZHMDA (Chen et al., 2017). Similar to KATZHMDA, KATZBNRA uses the KATZ measure and the similarity of diseases and microbes according to the Gaussian interaction profile kernel to predict novel microbe-disease associations based on the known microbe-disease associations. Furthermore, in order to improve the predicting accuracy, KATZBNRA uses a bipartite network recommendation algorithm.

MATERIALS AND METHODS

Known Disease-Microbe Associations

HMDAD (Human Microbe-Disease Association Database, http://www.cuilab.cn/hmdad) collected the curated human microbe-disease association data from microbiota studies where the microbes were determined by 16s RNA sequencing on the genus level (Ma et al., 2017). HMDAD provides public access to the data, and our known microbe-disease association data were downloaded from HMDAD. The data contains 450 distinct confirmed associations between 39 diseases and 292 microbes and is coded in an adjacency matrix $A \in R^{n_d \times n_m}$, where n_d (or n_m) is the number of diseases (or microbes). If there has been an experiment confirming that microbe m_j relates to disease d_i,$A(i,j)$ is set to 1, otherwise $A(i,j)$ is set to 0.

Disease Gaussian Interaction Profile Kernel Similarity

According to (Chen et al., 2017), there is a generally accepted assumption that similar diseases show an interaction tendency to similar microbes. Similar to (Chen and Yan, 2013) and (Chen et al., 2017), we compute the disease network topologic similarity based on the Gaussian interaction profile kernel. For a disease-microbe association adjacent matrix A, the binary element $A(i,j)$ at row i and column j encodes whether there is a confirmed association between disease $d(i)$ and microbe $m(j)$. The ith row of A is denoted by $IP(d(i))$. $IP(d(i))$ can be regarded a binary vector and is called the interaction profile of $d(i)$ since it provides

the association information of disease $d(i)$ with all microbes. For two diseases, their similarity $KD(d_i,d_j)$, based on the Gaussian interaction profile kernel, is calculated from their interaction profiles according to the following equations.

$$KD(d_i,d_j)=\exp(-\gamma_d \,||\, IP(d_i)-IP(d_j)\,||^2) \tag{1}$$

$$\gamma_d=\frac{\gamma_d^{'}}{\left(\frac{1}{n_d}\sum_{k=1}^{n_d}||\, IP(d_k)\,||^2\right)} \tag{2}$$

$KD(d_i,d_j)$ is adjusted by the norm kernel bandwidth γ_d, which is controlled by the bandwidth parameter γ_d'. It is obvious that $KD(d_i,d_i) = 1$ and $0 < KD(d_i,d_j) \leq 1$. According to (Vanunu et al., 2010), KD values in (0, 0.3] may be not informative, while KD values in [0.6, 1] may show significant similarity. Therefore, a logistic function transformation from $KD(x, y)$ to $KD'(x, y)$ in Equation (3) is utilized in order to measure the similarity of diseases x and y more appropriately.

$$KD'(d_i,d_j)=\frac{1}{1+e^{c*KD(d_i,d_j)+d}} \tag{3}$$

The parameters γ_d' and c could be set with cross-validation, but to simplify the calculation, we set $\gamma_d' = 1$ as in van Laarhoven et al., 2011, c = -15 as in (Vanunu et al., 2010). According to (Vanunu et al., 2010), we set $d = log(9999)$ such that $KD'(d_i,d_j) = 0.0001$ when $KD(d_i,d_j) = 0$.

Microbe Gaussian Interaction Profile Kernel Similarity

As mentioned before, similar diseases show an association tendency with similar microbes. To measure the similarity of microbes, we also used the Gaussian interaction profile kernel as before. It could be calculated in a similar way as follows.

$$KM(m_i,m_j)=\exp(-\gamma_m \,||\, IP(m_i)-IP(m_j)\,||^2) \tag{4}$$

$$\gamma_m=\frac{\gamma_m'}{\left(\frac{1}{n_m}\sum_{k=1}^{n_m}||\, IP(m_k)\,||^2\right)} \tag{5}$$

where γ_m' is also set to 1, and $IP(m_i)$ is the ith column of matrix A. Similarly, $KM'(m_i, m_j)$ could be calculated as Equation (3). It should be noted that in each cross-validation experiment, the similarities of diseases and microbes will be recalculated (Sun et al., 2018).

Bipartite Network Recommendation

The bipartite network recommendation is a two-step resource allocation process (Chen et al., 2018b), which is based on a bipartite graph $G(D, M, E)$, where D represents disease nodes, M microbe nodes, E the edges corresponding to the known microbe-disease associations. Let $f_{0,i}(m_j)$ denote the initial resource allocated to a microbe node m_j when considering disease d_i, $k(m_j)$ be the number of adjacent disease nodes of microbe m_j, and let $k(d_i)$ be the number of adjacent microbe nodes of disease d_i in graph G.

When focusing on disease d_i, each disease d_i related microbe node is initially allocated with a resource value of 1, i.e. if there is an edge between the disease node d_i and a microbe node m_j in G, allocate an initial resource of 1 to m_j. The first step of the bipartite network recommendation is to transfer the resource from microbe nodes to disease nodes according to Equation (6), and the second step is to transfer the resource of the disease nodes back to microbe nodes according to Equation (8).

$$f_{1,i}(d_l)=\sum_{i=1}^{n_m}\frac{a_{lj}f_{0,i}(m_j)}{k(m_j)} \tag{6}$$

where a_{lj} is an element of matrix A, i.e. $a_{lj} = A(l, j)$ and

$$a_{lj}=\begin{cases}1, if\ disease\ d_i\ is\ related\ to\ microbe\ m_j\\ 0, otherwise\end{cases} \tag{7}$$

In fact, $f_{0,i}(m_j)$ is also equal to $A(i, j)$.

$$f_{2,i}(m_j)=\sum_{l=1}^{n_d}\frac{a_{lj}f_{1,i}(d_l)}{k(d_l)} \tag{8}$$

Equations (6) and (8) are integrated into Equation (9).

$$f_{2,i}(m_l)=\sum_{j=1}^{n_m}w_{lj}f_{0,i}(m_j) \tag{9}$$

$$w_{lj}=\frac{1}{k(m_j)}\sum_{q=1}^{n_d}\frac{a_{qi}a_{qj}}{k(d_q)} \tag{10}$$

Please see an example of the process of the bipartite network recommendation focusing on disease d_1 in **Figure 1**. After the process, we obtain the recommendation scores (1, 0, 1/4, 3/4, 0) of the microbes for disease d_1, which suggests that besides m_1 and m_4, m_3 may also be related to the disease.

The matrix form of Equation (9) is as follows.

$$B=W\times A^T \tag{11}$$

where $W=\{w_{ij}\}_{n_m\times n_m}$, and B is a matrix with n_m rows and n_d columns. The ith column of B is the recommend scores of bipartite network recommendation regarding disease d_i

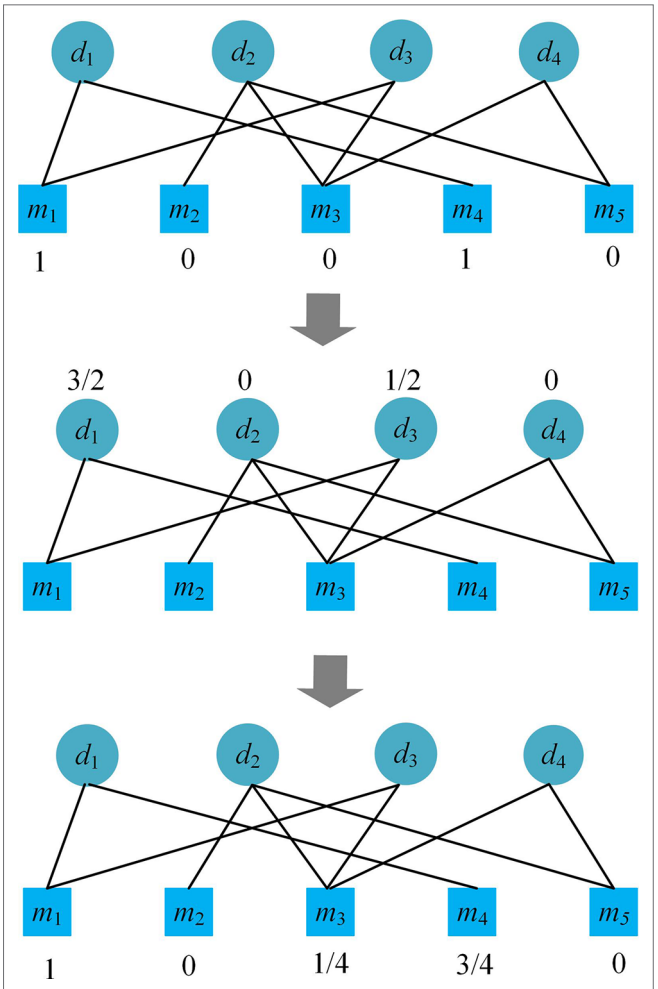


FIGURE 1 | Illustration of the two-step resource-allocation process in a bipartite graph.

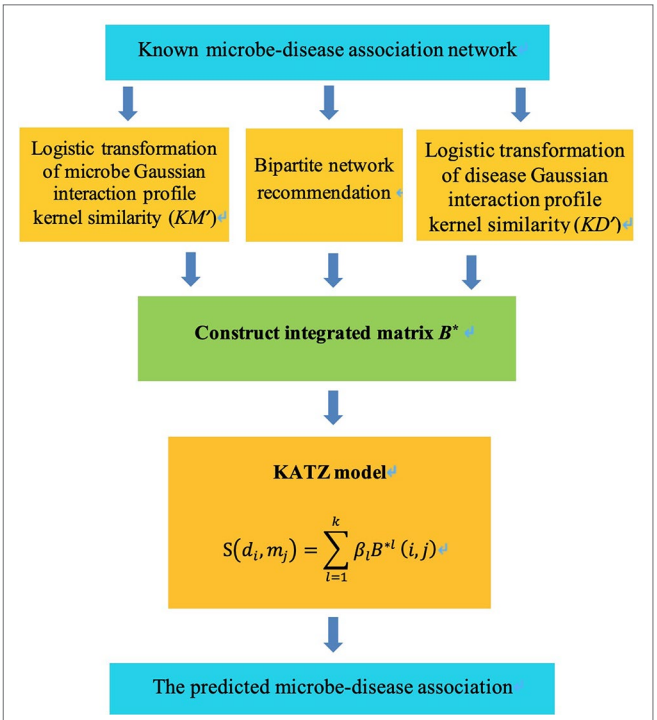


FIGURE 2 | The diagram of KATZBNRA.

KATZBNRA

KATZBNRA uses the KATZ model to compute the associations between diseases and microbes and is illustrated in **Figure 2**. As a network-based computation method, the KATZ model (Chen, 2015) had been used in the problem of link prediction in the heterogenous network to calculate the similarity of nodes. There are two factors that have been regarded as effective similarity metrics in the KATZ model, the walk steps (length, i.e. the number of edges of the walk) and the number of walks from one node to another. We use the KATZ model to calculate similarities between the nodes of the microbe and disease by counting the number of walks between them. Here $A^l(i,j)$, the element of the l-th power of A, is the number of l-length walks between disease node d_i and microbe node m_j. Due to the limited data from HMDAD, matrix A is sparse. In order to use more information, we integrated the matrices KM, KD, B into a matrix B^* as Equation (12) and replace A by B^* in the KATZ model to calculate similarities between microbes and diseases.

$$B^* = \begin{bmatrix} KD' & B \\ B^T & KM' \end{bmatrix} \tag{12}$$

Since walks between nodes of microbe and disease with different lengths have different contributions to similarities of node pairs, in order to dampen longer walks' contribution, we introduced a parameter β_l which is no smaller than 0, and if $l_1 > l_2$, then $\beta_{l_1} < \beta_{l_2}$. The potential association between diseases d_i and microbe m_j can be calculated as follows.

$$S(d_i, m_j) = \sum_{l=1}^{k} \beta_l B^{*l}(i,j) \tag{13}$$

If $k \to \infty$, replace β_l with β^l ($0<\beta<1$) (Qu et al., 2018) and the matrix form of Equation (13) is as follows.

$$S = \sum_{l \geq 1} \beta^l B^{*l} = (I - \beta B^*)^{-1} - I \tag{14}$$

S is a matrix of size $(n_d + n_m) \times (n_d + n_m)$, and could be partitioned into four sub-matrices as shown in Equation (12).

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{bmatrix} \tag{15}$$

where the rows of $S_{1,1}$ and $S_{1,2}$ are n_d, the rows of $S_{2,1}$ and $S_{2,2}$ are n_m, the columns of $S_{1,1}$ and $S_{2,1}$ are n_d, and the columns of $S_{1,2}$ and $S_{2,2}$ are n_m. The element $S_{1,2}(i, j)$ of $S_{1,2}$ provides the possibility that an association between disease d_i and the microbe m_j exists, and our prediction result can be obtained from $S_{1,2}$.

Considering that the walks of long lengths may be meaningless, we limit k in Equation (13) to be 2, 3 and 4, and the expression can be as follows.

$$S_{k=2} = \beta \cdot B + \beta^2 \cdot (KM' \cdot B + B \cdot KD') \tag{16}$$

$$S_{k=3} = S_{k=2} + \beta^3 \cdot (B \cdot B^T \cdot B + KM'^2 \cdot B + KM' \cdot B \cdot KD' + B \cdot KD'^2) \tag{17}$$

$$\begin{aligned} S_{k=4} &= S_{k=3} \\ &+ \beta^4 \cdot (KM'^3 \cdot B + B \cdot B^T \cdot KM' \cdot B + KM' \cdot B \cdot B^T \cdot B + B \cdot KD' \cdot B^T \cdot B) \\ &+ \beta^4 \cdot (B \cdot B^T \cdot B \cdot KD' + KM'^2 \cdot B \cdot KD' + KM' \cdot B \cdot KD'^2 + B \cdot KD'^3) \end{aligned} \tag{18}$$

RESULTS

Performance Evaluation

The test dataset of microbe-disease association was downloaded from HMDAD. We used LOOCV (leave-one-out cross validation), two-fold cross validation and five-fold cross validation to test the prediction performance of KATZBNRA on the HMDAD data.

In LOOCV, each known microbe-disease association takes turns to be picked out as the testing case and the other known associations are regarded as training data. We then obtained the prediction score of the test case output by KATZBNRA and ranked of the test case in the sorted list of all predicted microbe-disease associations in descending order of their scores. We used different thresholds to determine the correct predictions and wrong predictions and calculated corresponding FPR (false positive rate) and TPR (true positive rate) according to Equation (19). Finally, the results were presented in the ROC (receiver operating characteristics) curve plot of TPR against FPR.

$$TPR = \frac{TP}{FN + TP}, \quad FPR = \frac{FP}{TN + FP} \tag{19}$$

where FN is the number of false negative predictions (i.e. the cases whose prediction scores below the threshold), and TP is the number of true positive predictions (i.e. the cases whose prediction scores are not smaller than the threshold). FP is the number of the predicted associations that are not in the HMDAD dataset with scores not smaller than the threshold, and TN is the number of predicted associations that are not in the HMDAD dataset with scores smaller than the threshold. The area under a ROC curve is called AUC, and AUC is generally utilized to compare the power of predictive models. AUC of 0.5 indicates an entirely random prediction while AUC = 1 means a completely correct prediction.

In order to further test the prediction power of KATZBNRA, we also adopted 5-fold cross validation and 2-fold cross validation besides LOOCV. 5-fold (or 2-fold) cross validation randomly divides the microbe-disease associations equally into five (or two) parts and one of the five (or two) parts is reserved as the verification data while the remaining is used as training data. Considering the potential random sampling bias, we repeated each LOOCV, 2-fold and 5-fold cross validation test 100 times, and all ROC curves and AUCs are the average results of the 100 repeated tests. Meanwhile, we compared KATZBNRA with several state-of-the-art predictive methods using these validations.

For our method KATZBNRA, the walk length k plays a critical role. To test the effect of k, we changed the value of k, and carried out a series of LOOCV experiments. As shown in **Figure 3**, when k is set to 2, 3 and 4, the AUCs of each walk lengths are 0.9098, 0.8968, and 0.8827, respectively. Obviously, when parameter k = 2, KATZBNRA achieved the best prediction performance and walks of longer lengths may make the association prediction worse. Therefore, in the following experiments, we set k = 2. KATZBNRA has two more parameters, γ′ and β. The test in a previous work (Chen et al., 2016) showed AUC tended to decrease when γ′ was increased from 1.0 to 1.5, 2.0 and 2.5, and β was increased from 0.01 to 0.05 and 0.1. We also evaluated the AUC of KATZBNRA with different values of parameter γ′ and c in Equation (2) and Equation (3), and the test results are shown in **Tables 1** and **2**, showing similar results as Chen et al., 2016. Therefore, we set γ′=1.0 and β=0.01.

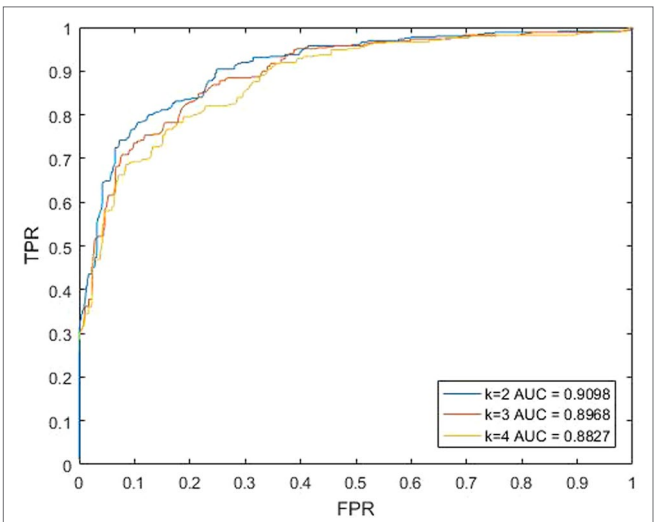


FIGURE 3 | The predictive performances of KATZBNRA with different ks.

TABLE 1 | The AUC of KATZBNRA with γ'set different values.

γ'	AUC
1	0.9098
1.5	0.9083
2	0.9033

TABLE 2 | The AUC of KATZBNRA with c set different values.

c	AUC
-15	0.9098
-10	0.9038
-5	0.8935

We compared KATZBNRA with another three prediction methods, the native bipartite network recommendation (BNR) (Zhou et al., 2007), KATZHMDA (Chen et al., 2017), and IMCMDA (Chen et al., 2018a) using LOOCV, 5-fold cross validation and 2-fold cross validation. The global LOOCV showed that the AUCs of KATZBNRA, KATZHMDA, IMCMDA and BNR were 0.9098, 0.8382, 0.7786, and 0.4113, respectively, as shown in **Figure 4**–**6** show the 5-fold cross validation experimental results and the 2-fold cross validation experimental results, respectively. In 5-fold cross validation KATZBNRA, KATZHMDA, IMCMDA, and BNR obtained AUCs of 0.8972, 0.8330, 0.8041, and 0.5645, respectively, and in 2-fold cross validation, their AUCs were 0.8463, 0.8190, 0.7988 and 0.5434, respectively. In all the above experiments, the curves of KATZBNRA are above those of the other methods, which means that among the four methods, KATZBNRA achieved the best prediction performance.

Case Studies

We studied asthma and inflammatory bowel disease (IBD) of microbe-related diseases of human beings based on recent published clinical and biological reports to further evaluate the ability of our method. The predicted disease-microbe associations which are contained in the HMDAD dataset are sorted according to their prediction scores in descending order. For asthma and IBD, we observed the microbes in the top 10 associations of the lists. This guarantees absolute independence between the verification candidate and the known association for model training.

As a common chronic lung inflammatory disease, asthma causes difficulty in breathing (Martinez, 2007). It is believed that asthma is caused by the environment and a combination of genes. For severe asthma, one of the leading causes is a microbe (Huang et al., 2011). All of top predicted 10 candidate microbes of KATZBNRA (**Table 3**) have been verified by recent studies.

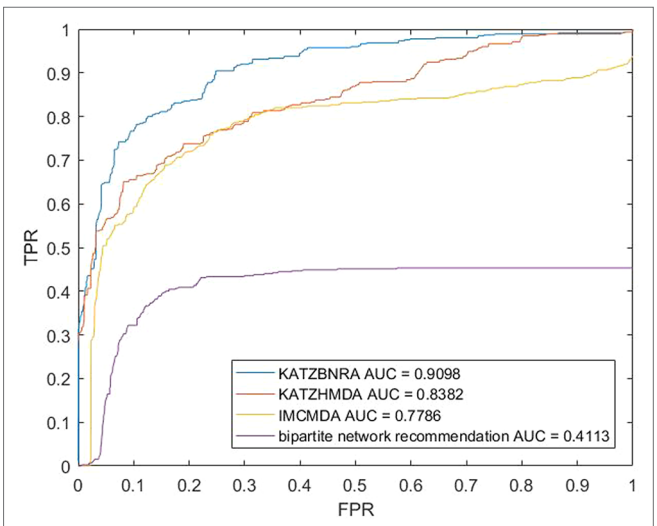


FIGURE 4 | The LOOCV experimental results of KAZTBNRA, KATZHMDA, IMCMDA, and the native bipartite network recommendation.

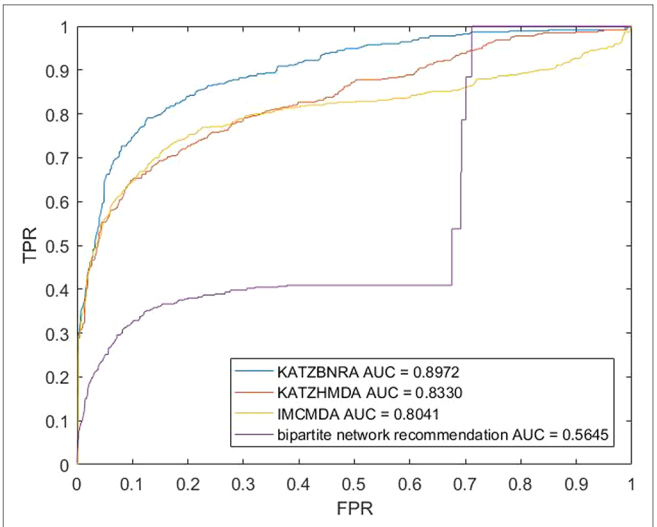


FIGURE 5 | The 5-fold cross validation experimental results of KAZTBNRA, KATZHMDA, IMCMDA, and the native bipartite network recommendation.

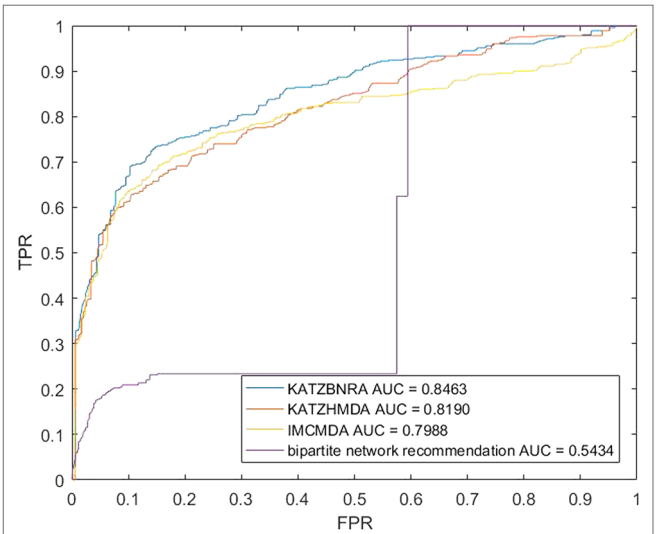


FIGURE 6 | The 2-fold cross validation experimental results of KAZTBNRA, KATZHMDA, IMCMDA, and the native bipartite network recommendation.

TABLE 3 | The Asthma-related microbe prediction of KATZBNRA. All of top 10 microbes were confirmed by recent studies.

Rank	Microbe	Evidence
1	Firmicutes	PMID: 23265859(Marri et al., 2013)
2	Actinobacteria	PMID: 23265859(Marri et al., 2013)
3	Clostridium coccoides	PMID:21477358(Vael et al., 2011)
4	Streptococcus	PMID: 17950502(Preston et al., 2007)
5	Lactobacillus	PMID: 20592920(Yu et al., 2010)
6	Lachnospiraceae	PMID:17433177(Rados et al., 2007)
7	Pseudomonas	PMID:13268970(Fein, 1955)
8	Burkholderia	PMID:24451910(Beigelman et al., 2014)
9	Fusobacterium	Dang et al., 2013(Dang et al., 2013)
10	Propionibacterium	PMID:27433177(Jung et al., 2016)

TABLE 4 | Top 10 potential IBD-related microbes predicted by KATZBNRA

Rank	Microbe	Evidence
1	Clostridium coccoides	PMID:19235886(Sokol et al., 2009)
2	Firmicutes	PMID:25307765(Walters et al., 2014)
3	Bacteroidetes	PMID:25307765(Walters et al., 2014)
4	Staphylococcus	PMID:28174737(Pedamallu et al., 2016)
5	Prevotella	PMID:25307765(Walters et al., 2014)
6	Streptococcus	PMID:23679203(Kojima et al., 2014)
7	Propionibacterium	unconfirmed
8	Propionibacterium acnes	unconfirmed
9	Bacteroidaceae	PMID:17897884(Takaishi et al., 2008)
10	Haemophilus	PMID:24013298(Said et al., 2014)

As a typical chronic GI (gastrointestinal) tract inflammatory bowel disease, IBD includes ulcerative colitis and Crohn's disease (Lomax et al., 2006). We listed the top 10 IBD-related candidate microbes predicted by KATZBNRA in **Table 4**, among which eight microbes have been previously validated.

DISCUSSION

Based on the bipartite network recommendation and the KATZ model, the paper introduced a novel disease-microbe association prediction method called KATZBNRA. KATZBNRA uses the Gaussian interaction profile kernel to calculate the similarity of diseases and microbes in the bipartite network containing the known microbe-disease associations from the HMDAD database, and the bipartite network recommendation score on the KATZ model enables KATZBNRA to predict potential disease-microbe associations with high accuracy. The experimental results of LOOCV, 5-fold cross validation, 2-fold cross validation and the IBD and asthma case studies have demonstrated the excellent and reliable prediction ability of KATZBNRA. With regard to similar prediction problems such as predicting lncRNA-disease, drug-target, gene-disease, miRNA-disease, and other biological associations, this model can be applied with small modifications.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.cuilab.cn/hmdad.

AUTHOR CONTRIBUTIONS

SL and MX conceived the study and the conceptual design of the work. SL and XL collected the data and implemented the KATZBNRA algorithm. SL tested the algorithms and drafted the manuscript. MX and XL polished the manuscript. All authors have read and approved the manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China (Grant No. 61772197).

ACKNOWLEDGMENTS

We sincerely thank all reviewers for their valuable comments that we have used to improve the quality of our manuscript.The paper is an extended version of the abstract presented orally at CBC 2019 and has been recommended by CBC 2019.

REFERENCES

Aroniadis, O. C., and Brandt, L. J. (2013). Fecal microbiota transplantation: past, present and future. *Curr. Opin. Gastroenterol.* 29 (1), 79–84. doi: 10.1097/MOG.0b013e32835a4b3e

Bao, W., Jiang, Z., and Huang, D. S. (2017). Novel human microbe-disease association prediction using network consistency projection. *BMC Bioinf.* 18 (Suppl 16), 543. doi: 10.1186/s12859-017-1968-2

Beigelman, A., Weinstock, G. M., and Bacharier, L. B. (2014). The relationships between environmental bacterial exposure, airway bacterial colonization, and asthma. *Curr. Opin. Allergy Clin. Immunol.* 14 (2), 137–142. doi: 10.1097/ACI.0000000000000036

Chen, H., and Zhang, Z. (2013). Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med. Genomics* 6, 12. doi: 10.1186/1755-8794-6-12

Chen, X. (2015). KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5, 16840. doi: 10.1038/srep16840

Chen, X., Huang, Y. A., Wang, X. S., You, Z. H., and Chan, K. C. (2016). FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. *Oncotarget* 7 (29), 45948–45958. doi: 10.18632/oncotarget.10008

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017). A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 33 (5), 733–739. doi: 10.1093/bioinformatics/btw715

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018a). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34 (24), 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018b). BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. *Bioinformatics* 34 (18), 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29 (20), 2617–2624. doi: 10.1093/bioinformatics/btt426

Dang, H. T., Park, H. K., Shin, J. W., Park, S.-G., and Kim, W. (2013). Analysis of oropharyngeal microbiota between the patients with bronchial asthma and the non-asthmatic persons. *J. Bacteriology Virol.* 43 (4), 270–278. doi: 10.4167/jbv.2013.43.4.270

Fein, B. T. (1955). Bronchial asthma caused by Pseudomonas aeruginosa diagnosed by bronchoscopic examination. *Ann. Allergy* 13 (6), 639–641.

Fry, L., and Baker, B. S. (2007). Triggering psoriasis: the role of infections and medications. *Clin. Dermatol.* 25 (6), 606–615. doi: 10.1016/j.clindermatol.2007.08.015

Gill, S. R., Pop, M., Deboy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., et al. (2006). Metagenomic analysis of the human distal gut microbiome. *Science* 312 (5778), 1355–1359. doi: 10.1126/science.1124234

Grice, E. A., and Segre, J. A. (2011). The skin microbiome. *Nat. Rev. Microbiol* 9 (4), 244–253. doi: 10.1038/nrmicro2537

He, B. S., Peng, L. H., and Li, Z. (2018). Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front. Microbiol* 9, 2560. doi: 10.3389/fmicb.2018.02560

Huang, Y. J., Nelson, C. E., Brodie, E. L., Desantis, T. Z., Baek, M. S., Liu, J., et al. (2011). Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J. Allergy Clin. Immunol.* 127372-381 (2), e371–e373. doi: 10.1016/j.jaci.2010.10.048

Human Microbiome Project, C. (2012a). A framework for human microbiome research. *Nature* 486 (7402), 215–221. doi: 10.1038/nature11209

Human Microbiome Project, C. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature* 486 (7402), 207–214. doi: 10.1038/nature11234

Jung, J. W., Choi, J. C., Shin, J. W., Kim, J. Y., Park, I. W., Choi, B. W., et al. (2016). Lung microbiome analysis in steroid-nasmall yi, ukrainianve asthma patients by using whole sputum. *Tuberc Respir. Dis. (Seoul)* 79 (3), 165–178. doi: 10.4046/trd.2016.79.3.165

Kojima, A., Nomura, R., Naka, S., Okawa, R., Ooshima, T., and Nakano, K. (2014). Aggravation of inflammatory bowel diseases by oral streptococci. *Dis.* 20 (4), 359–366. doi: 10.1111/odi.12125

Lomax, A. E., Linden, D. R., Mawe, G. M., and Sharkey, K. A. (2006). Effects of gastrointestinal inflammation on enteroendocrine cells and enteric neural reflex circuits. *Auton Neurosci.*, 126–127. doi: 10.1016/j.autneu.2006.02.015

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2017). An analysis of human microbe-disease associations. *Brief Bioinform.* 18 (1), 85–97. doi: 10.1093/bib/bbw005

Marri, P. R., Stern, D. A., Wright, A. L., Billheimer, D., and Martinez, F. D. (2013). Asthma-associated differences in microbial composition of induced sputum. *J. Allergy Clin. Immunol.* 131346-352 (2), e341–e343. doi: 10.1016/j.jaci.2012.11.013

Martinez, F. D. (2007). Genes, environments, development and asthma: a reappraisal. *Eur. Respir. J.* 29 (1), 179–184. doi: 10.1183/09031936.00087906

Pedamallu, C. S., Bhatt, A. S., Bullman, S., Fowler, S., Freeman, S. S., Durand, J., et al. (2016). Metagenomic characterization of microbial communities in situ within the deeper layers of the ileum in crohn's disease. *Cell Mol. Gastroenterol. Hepatol* 2563-566 (5), e565. doi: 10.1016/j.jcmgh.2016.05.011

Preston, J. A., Essilfie, A. T., Horvat, J. C., Wade, M. A., Beagley, K. W., Gibson, P. G., et al. (2007). Inhibition of allergic airways disease by immunomodulatory therapy with whole killed Streptococcus pneumoniae. *Vaccine* 25 (48), 8154–8162. doi: 10.1016/j.vaccine.2007.09.034

Qu, Y., Zhang, H., Liang, C., and Dong, X. (2018). KATZMDA: prediction of miRNA-disease associations based on KATZ model. *IEEE Access* 6, 3943–3950. doi: 10.1109/access.2017.2754409

Rados, J., Dobric, I., Pasic, A., Lipozencic, J., Ledic-Drvar, D., and Stajminger, G. (2007). Normalization in the appearance of severly damaged psoriatic nails using soft x-rays. A case report. *Acta Dermatovenerol Croat* 15 (1), 27–32.

Said, H. S., Suda, W., Nakagome, S., Chinen, H., Oshima, K., Kim, S., et al. (2014). Dysbiosis of salivary microbiota in inflammatory bowel disease and its association with oral immunological biomarkers. *DNA Res.* 21 (1), 15–25. doi: 10.1093/dnares/dst037

Sokol, H., Seksik, P., Furet, J. P., Firmesse, O., Nion-Larmurier, I., Beaugerie, L., et al. (2009). Low counts of Faecalibacterium prausnitzii in colitis microbiota. *Inflammation Bowel Dis.* 15 (8), 1183–1189. doi: 10.1002/ibd.20903

Sommer, F., and Backhed, F. (2013). The gut microbiota–masters of host development and physiology. *Nat. Rev. Microbiol* 11 (4), 227–238. doi: 10.1038/nrmicro2974

Sun, Y., Zhu, Z., You, Z. H., Zeng, Z., Huang, Z. A., and Huang, Y. A. (2018). FMSM: a novel computational model for predicting potential miRNA biomarkers for various human diseases. *BMC Syst. Biol.* 12 (Suppl 9), 121. doi: 10.1186/s12918-018-0664-9

Takaishi, H., Matsuki, T., Nakazawa, A., Takada, T., Kado, S., Asahara, T., et al. (2008). Imbalance in intestinal microflora constitution could be involved in the pathogenesis of inflammatory bowel disease. *Int. J. Med. Microbiol* 298 (5-6), 463–472. doi: 10.1016/j.ijmm.2007.07.016

Vael, C., Vanheirstraeten, L., Desager, K. N., and Goossens, H. (2011). Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. *BMC Microbiol* 11, 68. doi: 10.1186/1471-2180-11-68

van Laarhoven, T., Nabuurs, S. B., and Marchiori E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27(21), 3036–3043.

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PloS Comput. Biol.* 6 (1), e1000641. doi: 10.1371/journal.pcbi.1000641

Walters, W. A., Xu, Z., and Knight, R. (2014). Meta-analyses of human gut microbes associated with obesity and IBD. *FEBS Lett.* 588 (22), 4223–4233. doi: 10.1016/j.febslet.2014.09.039

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol* 14 (8), 508–522. doi: 10.1038/nrmicro.2016.83

Wu, C., Gao, R., Zhang, D., Han, S., and Zhang, Y. (2018). PRWHMDA: human microbe-disease association prediction by random walk on the heterogeneous network with PSO. *Int. J. Biol. Sci.* 14 (8), 849–857. doi: 10.7150/ijbs.24539

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PloS One* 9 (1), e87797. doi: 10.1371/journal.pone.0087797

Yu, J., Jang, S. O., Kim, B. J., Song, Y. H., Kwon, J. W., Kang, M. J., et al. (2010). The Effects of Lactobacillus rhamnosus on the Prevention of Asthma in a Murine Model. *Allergy Asthma Immunol. Res.* 2 (3), 199–205. doi: 10.4168/aair.2010.2.3.199

Zeng, X., Liu, L., Lu, L., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34 (14), 2425–2432. doi: 10.1093/bioinformatics/bty112

Zeng, X., Wang, W., Deng, G., Bing, J., and Zou, Q. (2019). Prediction of potential disease-associated MicroRNAs by using neural networks. *Mol. Ther. Nucleic Acids* 16, 566–575. doi: 10.1016/j.omtn.2019.04.010

Zhang, W., Chen, Y., Liu, F., Luo, F., Tian, G., and Li, X. (2017). Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinf.* 18 (1), 18. doi: 10.1186/s12859-016-1415-9

Zhang, W., Lu, X., Yang, W., Huang, F., Wang, B., Wang, A., et al. (2018a)."HNGRNMF: Heterogeneous Network-based Graph Regularized Nonnegative Matrix Factorization for predicting events of microbe-disease associations", in: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018),*. (Madrid, Spain).

Zhang, W., Yang, W., Lu, X., Huang, F., and Luo, F. (2018b). The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018c). Predicting drug-disease associations and their therapeutic function based on the

drug-disease association bipartite network. *Methods* 145, 51–59. doi: 10.1016/j.ymeth.2018.06.001

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018d). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinf.* 19 (1), 233. doi: 10.1186/s12859-018-2220-4

Zhou, T., Ren, J., Medo, M., and Zhang, Y. C. (2007). Bipartite network projection and personal recommendation. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 76 (4 Pt 2), 046115. doi: 10.1103/PhysRevE.76.046115

Zou, S., Zhang, J., and Zhang, Z. (2017). A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. *PloS One* 12 (9), e0184394. doi: 10.1371/journal.pone.0184394

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2019
doi: 10.3389/fgene.2019.01054

Sparse Graph Regularization Non-Negative Matrix Factorization Based on Huber Loss Model for Cancer Data Analysis

Chuan-Yuan Wang[1], Jin-Xing Liu[1], Na Yu[1] and Chun-Hou Zheng[2]*

[1] School of Information Science and Engineering, Qufu Normal University, Rizhao, China, [2] School of Software Engineering, Qufu Normal University, Qufu, China

Non-negative matrix factorization (NMF) is a matrix decomposition method based on the square loss function. To exploit cancer information, cancer gene expression data often uses the NMF method to reduce dimensionality. Gene expression data usually have some noise and outliers, while the original NMF loss function is very sensitive to non-Gaussian noise. To improve the robustness and clustering performance of the algorithm, we propose a sparse graph regularization NMF based on Huber loss model for cancer data analysis (Huber-SGNMF). Huber loss is a function between L_1-norm and L_2-norm that can effectively handle non-Gaussian noise and outliers. Taking into account the sparsity matrix and data geometry information, sparse penalty and graph regularization terms are introduced into the model to enhance matrix sparsity and capture data manifold structure. Before the experiment, we first analyzed the robustness of Huber-SGNMF and other models. Experiments on The Cancer Genome Atlas (TCGA) data have shown that Huber-SGNMF performs better than other most advanced methods in sample clustering and differentially expressed gene selection.

Keywords: non-negative matrix factorization, Huber loss, sample clustering, graph regularization, robustness

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Wen-Sheng Chen,
Shenzhen University, China
Xiangxiang Zeng,
Xiamen University, China

****Correspondence:***
Jin-Xing Liu
sdcavell@126.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *20 July 2019*
Accepted: *01 October 2019*
Published: *20 November 2019*

Citation:
Wang C-Y, Liu J-X, Yu N and Zheng C-H (2019) Sparse Graph Regularization Non-Negative Matrix Factorization Based on Huber Loss Model for Cancer Data Analysis. Front. Genet. 10:1054. doi: 10.3389/fgene.2019.01054

INTRODUCTION

Cancer is considered to be the number one killer of human health. The development of high-throughput sequencing technology has enabled researchers to obtain more comprehensive information about cancer patients (Chen et al., 2019). The gene expression data of cancer patients can be more used for effective data mining through computational methods (Chen et al., 2018). In general, cancer gene expression data are characterized by high dimensionality, which is extremely difficult for data analysis. How to effectively reduce the dimensionality of data is the key to analyzing cancer data. Principal component analysis (PCA) (Feng et al., 2019), locally linear embedding (LLE) (Roweis and Saul, 2000), and non-negative matrix factorization (NMF) (Yu et al., 2017) are common methods for reducing the data dimensionality. Unlike several other methods, NMF can find two non-negative matrices and its product can effectively restore the original matrix. The non-negative constraint guarantees additive combinations between different elements. NMF demonstrates its advantages in facial recognition, speech processing, document clustering, and recommendation systems (Guillamet and Vitrià, 2002; Xu et al., 2003; Schmidt and Olsson, 2006; Luo et al., 2014).

NMF has developed rapidly in recent years, and several variants of NMF have been proposed to improve the effectiveness of the decomposition. Cai et al. proposed graph regularized NMF (GNMF) for data representation (Cai et al., 2011). GNMF considers the association between points to preserve the internal structure of the data. Kim et al. applied the L^1-norm constraint on the coefficient matrix to introduce sparse NMF for clustering (SNMF) (Kim and Park, 2007). Sparseness is more likely to remove redundant features of data. The most of cancer data have noise and outliers, and the original NMF cannot solve this. Wang et al. introduced Characteristic Gene Selection Based on Robust GNMF (RGNMF) (Wang et al., 2016a) to improve the robustness of the algorithm. RGNMF assumes that the loss follows Laplacian distribution and uses the loss function of the $L^{2,1}$-norm (Kong et al., 2011) constraint. The $L^{2,1}$-norm combines the advantages of the L^2-norm and the L^1-norm, which impose an L^2-norm constraint on the entire data space and an L^1-norm constraint on the sum of the different data points (Ding et al., 2006).

The original NMF model is simple to understand and computationally fast, but the squared loss function is too sensitive to outliers and noise. Mao et al. proposed the correntropy induced metric based GNMF (CGNMF) (Mao et al., 2014) that changed the original loss function. The correntropy uses L^0-norm approximation for large outliers and noise through kernel function filtering, and the normal data is constrained by the L^2-norm (Liu et al., 2007), so it is not sensitive to outliers and noise. Du et al. proposed Huber-NMF (Du et al., 2012), which is also a loss function that is insensitive to outliers and noise. It uses approximate L^1-norm processing for outliers and noise, and L^2-norm for valuable data. Correntropy uses kernel functions to control weights, and Huber loss uses a different function approximation for different data through threshold adjustment. The robustness analysis of these several non-square loss models is given in the experimental part. To compare the performance of the NMF algorithm, the robust PCA (RPCA) based method for discovering differentially expressed genes proposed by Liu et al. (2013) is added to the experiment.

In this paper, we propose a model called sparse graph regularization NMF based on Huber Loss Model for Cancer Data Analysis (Huber-SGNMF). It effectively combines Huber loss, manifold structure, and sparse constraint. Huber loss is based on the relationship between L^1-norm and L^2-norm to approximate different data. In detail, Huber loss adjusts the square loss or linear loss to the data according to the threshold to enhance the robustness of the model to outliers. Geometric information in high-dimensional data should remain locally constant in low-dimensional representations (Cai et al., 2011), so graph regularization is added to preserve the manifold structure of the data. Sparse constraints in the model can remove redundant features contained in the data to reduce the amount of model calculations and improve clustering performance (Kim and Park, 2007).

The contributions of this article are as follows:

1. The squared loss of the original NMF is too sensitive to outliers and noise; so, we use a more robust Huber loss combined with NMF. The Huber loss considers the relationship between the L^1-norm and the L^2-norm to effectively handle non-Gaussian noise and large outliers. For the update rules of Huber loss, we use the multiplicative iterative algorithm based on semi-quadratic optimization to find the optimal solution.
2. The NMF model fits the data in Euclidean space but does not consider the intrinsic geometry of the data space. If the data is related in high-dimensional space, then we believe that the data represented by the low-dimensional should also be closely related. Considering the manifolds embedded in the high-dimensional environment space, we add graph Laplacian as a regularization term to the model. Graph regularization takes into account the impact of recent neighbors on data, and retaining graph structure can make NMF more distinguishable.
3. Sparse matrices can remove redundant data, reducing data complexity and model computational difficulty. In data analysis, sparsity can improve clustering performance by reducing the difficulty of feature selection. The $L^{2,1}$-norm as a sparse constraint is added to the model because the $L^{2,1}$-norm is robust and can achieve row sparse effect.

The remainder of this paper is organized as follows. The introduction of related work is shown in Section 2. Models and solution optimization are presented in Section 3. The experiment and analysis are arranged in Section 4. Section 5 summarizes the entire paper.

RELATED WORK

Non-Negative Matrix Factorization

NMF is a dimensionality reduction method based on partial representation. For a given dataset $\mathbf{X}=\left[x_1,x_2\ldots,x_n\right]\in\mathbb{R}^{m\times n}$, NMF can decompose it into the basic matrix $\mathbf{U}\in\mathbb{R}^{m\times k}$ and the coefficient matrix $\mathbf{V}\in\mathbb{R}^{k\times n}$, with the purpose of approximating the original matrix by two matrix products. In general, the rank of matrix factorization k is selected by the number of larger singular values.

For gene expression data matrix $\mathbf{X}\in\mathbb{R}^{m\times n}$, each row represents a gene corresponding to n samples, and each column represents a sample composed of m genes. Moreover, $\mathbf{U}$ contains m rows of metagene and $\mathbf{V}$ contains n rows of metapattern (Liu et al., 2018). Each column of $\mathbf{V}$ is a projection of a corresponding sample vector in $\mathbf{X}$ according to the basic matrix $\mathbf{U}$ (Li et al., 2017). NMF is visualized on gene expression data as shown in **Figure 1**.

The NMF loss function is minimized as follows:

$$\min \quad \left\|\mathbf{X}-\mathbf{UV}\right\|^2, \quad s.t.\ \mathbf{U}\geq 0,\mathbf{V}\geq 0, \tag{1}$$

where $\|\cdot\|$ represents the application of the Frobenius norm to the matrix.

Lee and Seung proposed the use of multiplicative iterative update rules to solve the optimal solution of NMF (Lee and Seung, 1999). Its update formula is as follows:

$$u_{ik}=u_{ik}\frac{(\mathbf{XV})_{ik}}{(\mathbf{UVV}^T)_{ik}}, \tag{2}$$

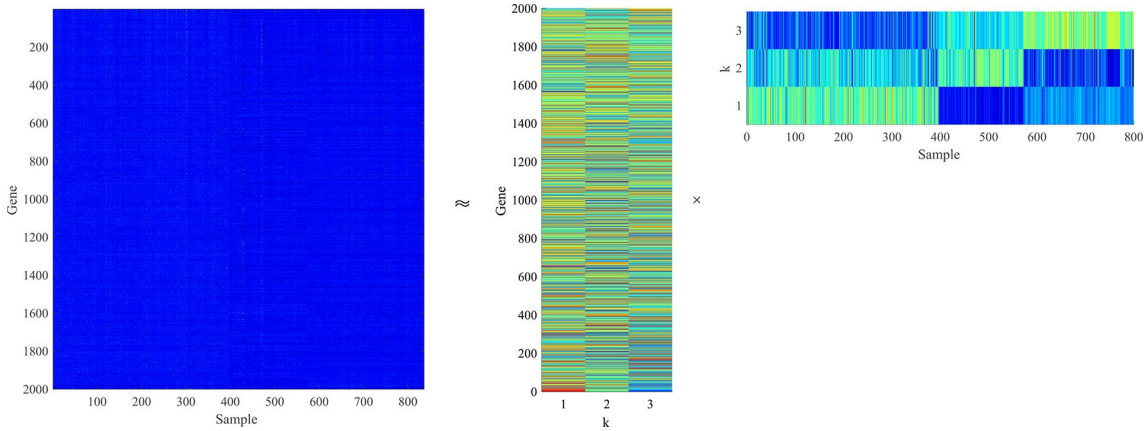


FIGURE 1 | The gene expression data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ is decomposed into a low-dimensional basic matrix $\mathbf{U} \in \mathbb{R}^{m \times k}$ and a low-dimensional coefficient matrix $\mathbf{V} \in \mathbb{R}^{k \times n}$. The product of two low-dimensional matrices can approximate the original matrix.

$$v_{kj} = v_{kj} \frac{\left(\mathbf{U}^T \mathbf{X}\right)_{kj}}{\left(\mathbf{U}^T \mathbf{U} \mathbf{V}\right)_{kj}}, \tag{3}$$

where u_{ik} and v_{kj} are elements belonging to $\mathbf{U}$ and $\mathbf{V}$, respectively. The non-negative constraints of $\mathbf{U}$ and $\mathbf{V}$ only allow additive combinations between different elements, so NMF can learn part-based representations (Cai et al., 2011).

Huber Loss

Data usually contain a small amount of outliers and noise, which can have a worse effect on model reconstruction. For noise and outliers in the dataset, Huber loss uses weighted L_1-norm processing because the L_1-norm is robust and can effectively handle outliers and noise (Guofa et al., 2011; Yu et al., 2016). For other valuable data in the dataset, Huber losses still use L_2-norm loss to fit the data. Huber loss function $\delta(\cdot)$ is defined as follows:

$$\delta\left(e\right) = \begin{cases} e^2 & if \ \left|e\right| < c, \\ 2c\left|e\right| - c^2 & if \ \left|e\right| \geq c, \end{cases} \tag{4}$$

where c represents the threshold parameter of the data using the L_1-norm or the L_2-norm. This function is a bounded and convex function that minimizes the effects of a single anomaly point (Chreiky et al., 2016). Huber losses often apply to the insensitive outliers and noise contained in the data, which are often difficult to find using the squared loss function (Du et al., 2012).

Manifold Regularization

The manifold learning theory (Belkin and Niyogi, 2001) shows that the internal manifold structure of the data can be effectively simulated by the nearest neighbor of the data points. Each data point finds its nearest p neighbors and connects the data points to the neighbors with edges. There are many ways to define the weight of an edge, most commonly 0–1 weighted: $\mathbf{W}_{ij}$=1, if and only if nodes i and j are connected by edges. The advantage of this weighting method is that it is easy to calculate.

Weight matrix $\mathbf{W}_{ij}$ is only used to measure the intimacy between data points. For the low-dimensional representation $\mathbf{s}_j$ of the high dimensional data $\mathbf{x}_j$, the Euclidean distance $\mathbf{O}(\mathbf{s}_j, \mathbf{s}_l) = \left\|\mathbf{s}_j - \mathbf{s}_l\right\|^2$ is typically used to measure the similarity between two low-dimensional data points. According to the intimacy weight W, the smoothness of the two low-dimensional vectors can be measured as follows:

$$\begin{aligned} R &= \frac{1}{2} \sum_{j,l}^{N} \left\|\mathbf{s}_j - \mathbf{s}_l\right\|^2 \mathbf{W}_{jl} \\ &= \sum_{j=1}^{N} \mathbf{s}_j^T \mathbf{s}_j \mathbf{D}_{jj} - \sum_{j,l=1}^{N} \mathbf{s}_j^T \mathbf{s}_l \mathbf{W}_{jl} \\ &= \mathrm{tr}\left(\mathbf{V}\mathbf{D}\mathbf{V}^T\right) - \mathrm{tr}\left(\mathbf{V}\mathbf{W}\mathbf{V}^T\right) \\ &= \mathrm{tr}\left(\mathbf{V}\mathbf{L}\mathbf{V}^T\right), \end{aligned} \tag{5}$$

where tr(·) denotes the trace of a matrix. The matrix $\mathbf{D}$ is defined as a diagonal matrix with diagonal elements $\mathbf{D}_{ii} = \sum_{jj} \mathbf{W}_{jl}$ The graph Laplacian (Liu et al., 2014) matrix $\mathbf{L}$ is defined as $\mathbf{L}$=$\mathbf{D}$-$\mathbf{W}$.

We hope that if the high-dimensional data $\mathbf{x}_j$ and $\mathbf{x}_l$ are very intimate, then $\mathbf{s}_j$ and $\mathbf{s}_l$ should be close enough in low-dimensional representations (Cai et al., 2011). Therefore, minimizing R is added to our model to encode the internal geometry of the data.

METHOD

The Huber-Sgnmf Model

Based on the Huber loss function, we proposed a novel model that preserves the manifold structure and sparsity simultaneously. The Huber loss is combined with NMF to enhance NMF robustness. To further optimize the model, the graph regularization term and the $L_{2,1}$-norm are added to the loss function as constraints. $L_{2,1}$-norm mathematical expression is as follows:

$$\left\|\mathbf{X}\right\|_{2,1} = \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} x_{ij}^2} = \sum_{i=1}^{m} \left\|x_{i,*}\right\|_2. \tag{6}$$

The Huber-SGNMF final model $O_{Huber-SGNMF}$ is as follows:

$$\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} \delta(\mathbf{X}\text{-}\mathbf{UV})+\alpha \mathrm{tr}\left(\mathbf{VLV}^T\right)+\beta\left\|\mathbf{V}\right\|_{2,1}, \tag{7}$$

where tr(·), α, and β represent the trace of the matrix, the regularization term parameters, and the sparse constraint parameters, respectively. In the experiment, the basic matrix **U** and the coefficient matrix **V** are used for differential gene selection and cluster analysis, respectively.

Optimization

Obviously, the loss function is a non-quadratic optimization problem, and finding the optimal solution is not simple. Fortunately, the semi-quadratic optimization technique that has been proposed can effectively optimize the loss function. The loss function can be reconstructed to find the optimal solution by introducing auxiliary variables. According to the conjugate function and the semi-quadratic technique (Nikolova and Chan, 2007), the fixed loss function $\sigma(\mathbf{Z})$ can be constructed as follows:

$$\sigma\left(\mathbf{Z}_{ij}\right)=\min_{\mathbf{W}\in\mathbb{R}} \tau\left(\mathbf{Z}_{ij},\mathbf{W}_{ij}\right)+\phi\left(\mathbf{W}_{ij}\right), \tag{8}$$

where $\mathbf{Z}_{ij}=\mathbf{X}_{ij}\text{-}\sum_{k=1}^{K}\mathbf{U}_{ik}\mathbf{V}_{kj}$ represents the difference between the NMF predicted value and the actual value. $\sigma(\cdot)$ indicates the noise or normal data, which is processed using different loss functions. $\mathbf{W}\in\mathbb{R}^{m\times n}$ is the introduced auxiliary variable. $\phi(\mathbf{W}_{ij})$ is the conjugate function of $\mathbf{Z}_{ij}$. $\tau(\cdot,\cdot)$ is a quadratic term for $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$. The NMF model only needs to consider the quadratic term of the multiplication form:

$$\tau\left(\mathbf{Z}_{ij},\mathbf{W}_{ij}\right)=\frac{1}{2}\mathbf{W}_{ij}\mathbf{Z}_{ij}^2. \tag{9}$$

Combine Equation (8) and Equation (9) with the loss function (7):

$$\begin{aligned}&\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} \delta\left(\mathbf{X}\text{-}\mathbf{UV}\right)+\alpha\mathrm{tr}\left(\mathbf{VLV}^T\right)+\beta\left\|\mathbf{V}\right\|_{2,1}\\&=\min_{\mathbf{U}\geq 0,\mathbf{V}\geq 0} \mathbf{W}\otimes\left(\mathbf{X}\text{-}\mathbf{UV}\right)^2+\phi\left(\mathbf{W}\right)+\alpha\mathbf{tr}\left(\mathbf{VLV}^T\right)+\beta\left\|\mathbf{V}\right\|_{2,1},\end{aligned} \tag{10}$$

where ⊗ represents the Hadamard product, which is the product between two matrices' elements. Operator ⊗ takes precedence over other matrices operators. Its Lagrangian function expansion is expressed as follows:

$$\mathcal{L}_{Huber-SGNMF}\left(\mathbf{U}\right)=\sum_{i=1}^{m}\left(\mathbf{X}_{i*}-\mathbf{U}_{i*}\mathbf{V}_{i*}\right)\mathbf{Q}_i\left(\mathbf{X}_{i*}-\mathbf{U}_{i*}\mathbf{V}_{i*}\right)^T+\mathrm{tr}\left(\boldsymbol{\psi}\mathbf{U}^T\right), \tag{11}$$

and

$$\begin{aligned}\mathcal{L}_{Huber-SGNMF}\left(\mathbf{V}\right)=&\sum_{j=1}^{n}\left(\mathbf{X}_{*j}-\mathbf{U}_{*j}\mathbf{V}_{*j}\right)\mathbf{R}_j\left(\mathbf{X}_{*j}-\mathbf{U}_{*j}\mathbf{V}_{*j}\right)^T\\&+\alpha\mathrm{tr}\left(\mathbf{VLV}^T\right)+\beta\mathrm{tr}\left(\mathbf{VGV}^T\right)+\mathrm{tr}\left(\boldsymbol{\psi}\mathbf{U}^T\right)+\mathrm{tr}\left(\boldsymbol{\varphi}\mathbf{V}^T\right),\end{aligned} \tag{12}$$

where $\mathbf{Q}_i$ and $\mathbf{R}_j$ are defined as $\mathbf{Q}_i=diag(\mathbf{W}_{i*})$ and $\mathbf{R}=diag(\mathbf{W}_{*j})$, respectively. $\boldsymbol{\psi}=\left[\psi_{ik}\right]$ and $\boldsymbol{\varphi}=\left[\varphi_{kj}\right]$ are Lagrangian multipliers of non-negative constraints $\mathbf{U}\ 0$ and $\mathbf{V}\ 0$, respectively. **G** is a diagonal matrix with diagonal elements, which is given by:

$$G_{jj}=1/\sqrt{\sum_{m=1}^{k}v_{mj}^2+\omega} \tag{13}$$

where ω is a number that is very close but not equal to zero.

Let $\boldsymbol{\psi}\mathbf{U}=0$ and $\boldsymbol{\varphi}\mathbf{V}=0$ by using Karush–Kuhn–Tucher (KKT) (Qi and Jiang, 1997) conditions. The loss function (10) can be iteratively optimized by the following schemes:

Update **W** when **U** and **V** are fixed. The weight matrix **W** according to equation (8) is defined as follows:

$$w_{ij}=\frac{\sigma'\left(\mathbf{Z}_{ij}\right)}{\mathbf{Z}_{ij}}, \tag{14}$$

where the elements of weight matrix is $w_{ij}\in\mathbf{W}$ Combine the loss function (7) with the equation (14) are as follows:

$$w_{ij}=\begin{cases}1 & if\ \left|x_{ij}\text{-}u_{ik}v_{kj}\right|\leq c,\\ \dfrac{c}{\left|\mathbf{X}-\mathbf{UV}\right|_{ij}} & otherwise,\end{cases} \tag{15}$$

Update **U** and **V** when **W** is fixed. The update rules for **U** and **V**are as follows:

$$\begin{aligned}u_{ik}&=u_{ik}\frac{\left(\mathbf{X}_{i*}\mathbf{Q}_i\mathbf{V}^T\right)_{ik}}{\left(\mathbf{U}_{i*}\mathbf{VQ}_i\mathbf{V}^T\right)_{ik}}\\&=u_{ik}\frac{\left(\mathbf{W}\otimes\mathbf{XV}^T\right)_{ik}}{\left(\mathbf{W}\otimes\left(\mathbf{UV}\right)\mathbf{V}^T\right)_{ik}},\end{aligned} \tag{16}$$

$$\begin{aligned}v_{kj}&=v_{kj}\frac{\left(\mathbf{U}^T\mathbf{X}_{*j}\mathbf{R}_j\right)_{kj}}{\left(\mathbf{U}^T\mathbf{R}_j\mathbf{UV}_{*j}+\alpha\mathbf{VL}+\beta\mathbf{VG}\right)_{kj}}\\&=v_{kj}\frac{\left(\mathbf{U}^T\left(\mathbf{W}\otimes\mathbf{X}\right)\right)_{kj}}{\left(\mathbf{U}^T\left(\mathbf{W}\otimes\mathbf{UV}\right)+\alpha\mathbf{VL}+\beta\mathbf{VG}\right)_{kj}},\end{aligned} \tag{17}$$

The threshold parameter c is set to the median of the reconstruction error,

$$c = \ median\ \left(\left|\mathbf{X}-\mathbf{UV}\right|_{ij}\right). \tag{18}$$

The corresponding algorithm is shown in Algorithm 1.

ALGORITHM 1 | Huber-SGNMF.

Data input: $\mathbf{X}\in\mathbb{R}^{m\times n}$, $\mathbf{L}\in\mathbb{R}^{n\times n}$
Data output: $\mathbf{U}\in\mathbb{R}^{m\times k}$, $\mathbf{V}\in\mathbb{R}^{k\times n}$ and weight matrix $\mathbf{W}\in\mathbb{R}^{m\times n}$
Parameters: α,β
Data initialize: $\mathbf{U}\geq 0$, $\mathbf{V}\geq 0$
Repeat
Update $\mathbf{G}$ by (13);
Update $\mathbf{W}$ by (15);
Update $\mathbf{u}_{ik}$ by (16);
Update $\mathbf{v}_{jk}$ by (17);
Update c by (18);
End convergence

Convergence Analysis

According to the update rules of Huber-SGNMF, the loss function $O_{Huber\text{-}SGNMF}$ can converge to the local optimum through theorem 1.

Theorem 1. The loss function (7) is guaranteed to be non-increasing under the update rules (16) and (17). The loss function is constant when the elements u_{ik} and v_{kj} have fixed values.

To prove theorem 1, we introduce the auxiliary function $\mathbf{H}$ in Algorithm.

Lemma 1. Suppose $\mathbf{H}$ (r, r') is an auxiliary function of F (r). If the conditions $\mathbf{H}$ (r,r') F(r) and $\mathbf{H}$ (r,r)=F(r) are satisfied, then it can be concluded that F(r) is non-increasing from iteration t to t+1:

$$r^{t+1} = \arg \min_{r} \mathbf{H}(r,r') \tag{19}$$

Proof:

$$\mathrm{F}(r^{t+1}) \leq \mathbf{H}(r^{t+1},r^{t}) \leq \mathbf{H}(r^{t},r^{t}) = \mathrm{F}(r^{t}). \tag{20}$$

Suppose loss function $O_{Huber\text{-}SGNMF}$ has a suitable auxiliary function $\mathbf{H}_{Huber}$ If the minimum updates rule for $\mathbf{H}_{Huber}$ is equal to (16) and (17), then the convergence of $O_{Huber\text{-}SGNMF}$ can be proved. Furthermore, the parts of the loss function $O_{Huber\text{-}SGNMF}$ associated with the elements $u_{ik} \in \mathbf{U}$ and $v_{kj} \in \mathbf{V}$ are represented by F_{ik} and F_{kj}, respectively. The partial derivative equation of $O_{Huber\text{-}SGNMF}$ can be derived as follows:

$$\mathrm{F}'_{ik} = \left(\frac{\partial O_{Huber-SGNMF}}{\partial \boldsymbol{U}}\right)_{ik} = \left(-2\mathbf{X}_{i*}\mathbf{Q}_i\mathbf{V}^T + 2\mathbf{U}_{i*}\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{ik}, \tag{21}$$

$$\mathrm{F}''_{ik} = \left(\frac{\partial^2 O_{Huber-SGNMF}}{\partial \mathbf{U}^2}\right)_{ik} = 2\left(\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{kk}, \tag{22}$$

$$\begin{aligned}\mathrm{F}'_{kj} = \left(\frac{\partial O_{Huber-SGNMF}}{\partial \mathbf{V}}\right)_{kj} &= -2\left(\mathbf{U}^T\mathbf{X}_{*j}\mathbf{R}_j\right)_k \\ &+2\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\mathbf{V}_{*j}\right)_k + 2\alpha\mathbf{V}\mathbf{L} + 2\beta\mathbf{V}\mathbf{G},\end{aligned} \tag{23}$$

$$\mathrm{F}''_{kj} = \left(\frac{\partial^2 O_{Huber-SGNMF}}{\partial \mathbf{V}^2}\right)_{kj} = 2\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\right)_{kk} + \left(2\alpha\mathbf{L}\right)_{jj} + \left(2\beta\mathbf{G}\right)_{jj}. \tag{24}$$

Essentially, the algorithm updates each element, which means that if the elements F_{ik} and F_{kj} are non-increasing, then $O_{Huber\text{-}SGNMF}$ is also non-increasing.

Lemma 2. Define $\mathbf{H}_{Huber}\left(u,u_{ik}^t\right)$ and $\mathrm{H}_{Huber}\left(v,v_{kj}^t\right)$ as auxiliary functions for $\mathbf{u}_{ik}$ and $\mathbf{v}_{kj}$, respectively. The expansion items are as follows:

$$\begin{aligned}\mathbf{H}_{Huber}\left(u,u_{ik}^t\right) &= \mathrm{F}_{ik}\left(u_{ik}^t\right) + \mathrm{F}'_{ik}\left(u_{ik}^t\right)\left(u-u_{ik}^t\right) \\ &+\frac{\left(\mathbf{U}_{i*}\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{ik}}{u_{ik}^t}\left(u-u_{ik}^t\right)^2,\end{aligned} \tag{25}$$

$$\begin{aligned}\mathbf{H}_{Huber}\left(v,v_{kj}^t\right) &= \mathrm{F}_{kj}\left(v_{kj}^t\right) + \mathrm{F}'_{kj}\left(v_{kj}^t\right)\left(v-v_{kj}^t\right) \\ &+\frac{\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\mathbf{V}_{*j} + \alpha\mathbf{V}\mathbf{L} + \beta\mathbf{V}\mathbf{G}\right)_{kj}}{v_{kj}^t}\left(v-v_{kj}^t\right)^2.\end{aligned} \tag{26}$$

Proof:

According to the lemma 1, $\mathbf{H}_{Huber}$ (u,u)=$\mathrm{F}_{ik}(u)$ and $\mathbf{H}_{Huber}$ (v,v)=$\mathrm{F}_{kj}(v)$ can be obtained. We have the following formulas through the Taylor series expansion of the auxiliary function.

$$\begin{aligned}\mathrm{F}_{ik}(u) &\approx \mathrm{F}_{ik}\left(u_{ik}^t\right) + \mathrm{F}'_{ik}\left(u_{ik}^t\right)\left(u-u_{ik}^t\right) \\ &+\frac{1}{2}\mathrm{F}''_{ik}\left(u_{ik}^t\right)\left(u-u_{ik}^t\right)^2,\end{aligned} \tag{27}$$

$$\begin{aligned}\mathrm{F}_{kj}(v) &\approx \mathrm{F}_{kj}\left(v_{kj}^t\right) + \mathrm{F}'_{kj}\left(v_{kj}^t\right)\left(v-v_{kj}^t\right) \\ &+\frac{1}{2}\mathrm{F}''_{kj}\left(v_{kj}^t\right)\left(v-v_{kj}^t\right)^2.\end{aligned} \tag{28}$$

Next, $\mathbf{H}_{Huber}\left(u,u_{ik}^t\right) \geq \mathrm{F}_{ik}(u)$ and $\mathbf{H}_{Huber}\left(v,v_{kj}^t\right) \geq \mathrm{F}_{kj}(v)$ need to be guaranteed.

According to (25) and (27), expand $\mathbf{H}_{Huber}\left(u,u_{ik}^t\right) \geq \mathrm{F}_{ik}(u)$ is as follows:

$$\frac{\left(\mathbf{U}_{i*}\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{ik}}{u_{ik}^t} \geq \mathbf{V}\mathbf{Q}_i\mathbf{V}^T, \tag{29}$$

since

$$\left(\mathbf{U}_{i*}\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{ik} = \sum_{a=1}^{K} u_{ia}\left(\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{ka} \geq u_{ik}\left(\mathbf{V}\mathbf{Q}_i\mathbf{V}^T\right)_{kk}. \quad (30)$$

According to (26) and (28), expand $\mathbf{H}_{Huber}\left(v, v_{kj}^t\right) \geq \mathrm{F}_{kj}\left(v\right)$ is as follows:

$$\frac{\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\mathbf{V}_{*j} + \alpha\mathbf{V}\mathbf{L} + \beta\mathbf{V}\mathbf{G}\right)_{kj}}{v_{kj}^t} \geq \left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\right)_{kk} + \left(\alpha\mathbf{L}\right)_{jj} + \left(\beta\mathbf{G}\right)_{jj}, \quad (31)$$

since

$$\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\mathbf{V}_{*j}\right)_{kj} = \sum_{b=1}^{K}\left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\right)_{bk} v_{bj} \geq \left(\mathbf{U}^T\mathbf{R}_j\mathbf{U}\right)_{kk} v_{kj}, \quad (32)$$

$$\left(\beta\mathbf{V}\mathbf{G}\right)_{kj} = \beta\sum_{b=1}^{N} v_{kb}\mathbf{G}_{bb} \geq \beta v_{kj}\mathbf{G}_{jj}, \quad (33)$$

and

$$\begin{aligned}\left(\alpha\mathbf{V}\mathbf{D}\right)_{kj} &= \alpha\sum_{c=1}^{N} v_{kc}\mathbf{D}_{cc} \geq \alpha v_{kj}\mathbf{D}_{jj} \\ &\geq \alpha v_{kj}\left(\mathbf{D}-\mathbf{W}\right)_{jj} = \alpha v_{kj}\mathbf{L}_{jj}.\end{aligned} \quad (34)$$

So, $\mathbf{H}_{Huber}\left(u, u_{ik}^t\right) \geq \mathrm{F}_{ik}\left(u\right)$ and $\mathbf{H}_{Huber}\left(v, v_{kj}^t\right) \geq \mathrm{F}_{kj}\left(v\right)$ can be obtained. In other words, the auxiliary functions $\mathrm{F}_{ik}(u)$ and $\mathrm{F}_{kj}(v)$ of the updated rules (16) and (17) are non-increasing, and the derivation of theorem 1 is completed. Finally, the convergence of the loss function $O_{Huber\text{-}SGNMF}$ is proved.

The corresponding convergence analysis curve is shown in **Figure 2**.

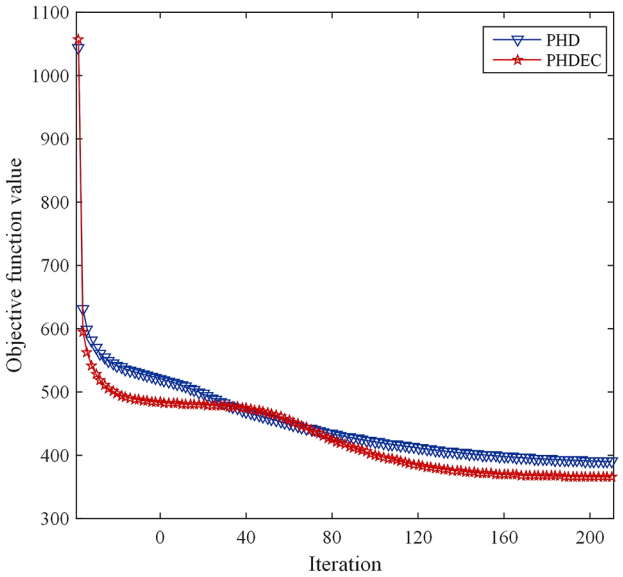


FIGURE 2 | Convergence analysis curve of Huber-SGNMF model. Each curve represents a dataset. PHD and PHDEC are the datasets used in the experiment.

RESULTS AND DISCUSSION

Datasets

Five gene expression datasets downloaded from TCGA are used in the experiment. TCGA is a gene data sharing system that contains information on thousands of cancer patients and has made great contributions to the path of human exploration of cancer genomics. The experiment used five datasets including cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), pancreatic cancer (PAAD), and esophageal cancer (ESCA).

To explore the association between genes and multiple cancers, diseased samples from multiple datasets are integrated into one dataset. In detail, the detesteds PAAD, HNSC, and COAD are integrated into one dataset, which is represented as PHD. The detesteds PAAD, HNSC, and COAD are integrated into one dataset, which is represented as PHD. These two integrated datasets contain only diseased samples of different diseases. Datasets are standardized before using, and the data normalization scales data to specific time intervals. Pre-processing data speeds up searching for the best solution and optimizes convergence speed. Since high-dimensional gene expression data contains a large amount of redundant information, PCA (Wu et al., 2017) is used to reduce the dimensions to 2,000 genes in the pre-processing.

Model Robustness

To analyze the robustness of RGNMF, CGNMF, and Huber-SGNMF, we apply these methods to a composite dataset consisting of 200 two-dimensional data points (**Figure 3A**). All data points are distributed in one dimensional space. In **Figure 3A**, there is only one contaminated point, and each model can restore the original data normally. The contaminated points in **Figures 3B–D** are 50 points, 100 points, and 150 points, respectively. In the case where a part of the data is contaminated, only Huber-SGNMF successfully restores the original data. CGNMF and RGNMF are affected by some noise or outliers when restoring data, while NMF is most affected by noise or outliers.

Parameter Selection

In the experiment, we consider the effect of each parameter on the solution model. A grid search method is used to find the optimal parameters of the model. The grid search range is $[10^{-2}\sim10^{2}]$. As shown in **Figure 4**, the PHD dataset is used as an example to find the optimal parameters of the Huber-SGNMF model. Specifically, the two datasets are set to the same parameters α = 100 and β = 0.01 Other methods in the experiment are set up with prior parameters or grid searches to find the optimal parameters.

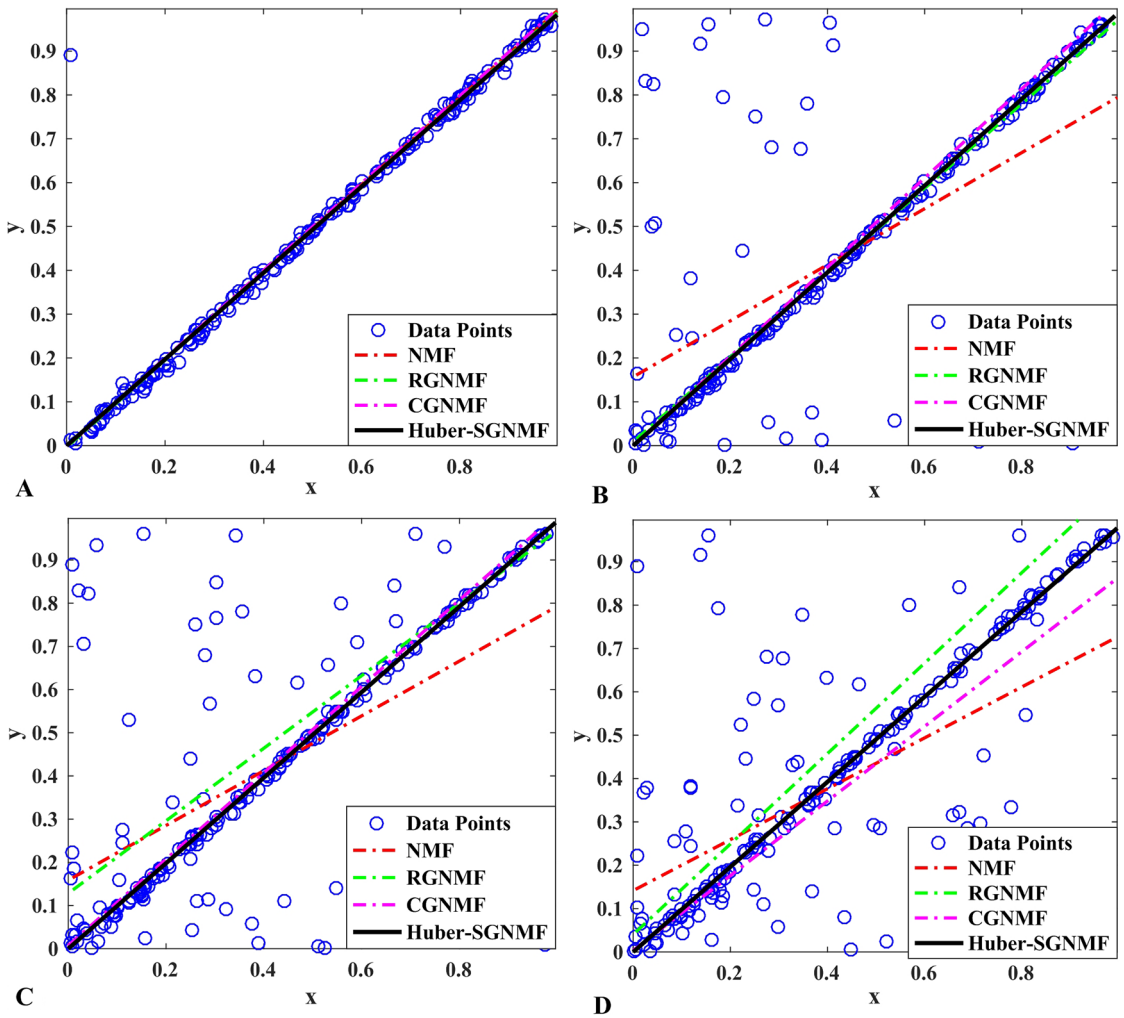


FIGURE 3 | In the case of different data points are contaminated, NMF, RGNMF, CGNMF, and Huber-SGNMF restore 200 synthetic two-dimensional data points: **(A)** the data contains 1 noise or outlier, **(B)** the data contains 50 noise or outliers, **(C)** the data contains 100 noise or outliers, **(D)** the data contains 150 noise or outliers.

Performance Evaluation and Comparisons

To prove the validity of the performance of the model, six states of the art methods including RPCA (Liu et al., 2013), NMF (Lee and Seung, 1999), SNMF (Kim and Park, 2007), GNMF (Cai et al., 2011), RGNMF (Wang et al., 2016a), CGNMF (Mao et al., 2014), and Huber-NMF (Du et al., 2012) are compared with Huber-SGNMF. In the experiment, the basic matrix and the coefficient matrix are used to differentially gene selection and cluster analysis, respectively.

Feature Selection Results and Analysis

Feature selection is the selection of representative features from multiple feature values (Yu and Liu, 2003). In the analysis of cancer data, the feature selection is to find differentially expressed genes for cancer (that is, pathogenic genes). This is of great significance in exploring the link between cancer and genes (Chen et al., 2017). For each method, the top 500 genes with the greatest differential expression are analyzed.

The GeneCards (https://www.genecards.org/) system is used to download all gene libraries associated with the disease. The selected genes are compared with the gene bank to select overlapping genes and obtain a corresponding relevance score. The relevance score is the indicator that GeneCards assesses the association between the gene and the disease. The higher the relevance score is, the greater the intimacy of the gene and the disease. The average relevance score (ARS) and the maximum relevance score (MRS) are used to evaluate the performance of the model.

The specific experimental results of the seven methods are listed in **Table 1**. The results show that the genes selected by Huber-SGNMF model have higher ARS and MRS. This means that the model can effectively find the genes associated with cancer. **Table 2** lists the genes for the top 10 largest relevance scores selected by the Huber-SGNMF model on the PHD dataset. The detailed genetic analysis is as follows.

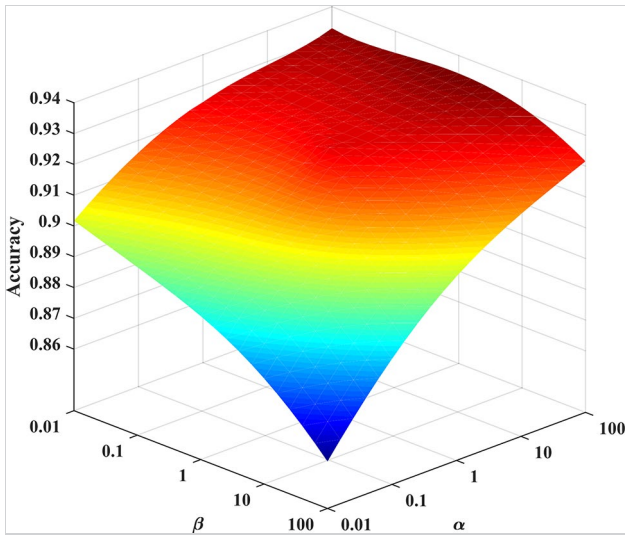


FIGURE 4 | Optimal parameter selection for the Huber-SGNMF model on the PHD dataset. Huber-SGNMF is set with parameters α = 100 and β = 0.01.

CTNNB1 is a protein-coding gene from which the protein encoded by the gene forms part of an adhesion-linked protein complex. Mutations in the CTNNB1 proto-oncogene are associated with most human colorectal epithelial tumors, and a significant increase in expression in the same tumor may indirectly or directly lead to intestinal adenocarcinoma (Wang et al., 2011). Moreover, deep sequencing of patients with pancreatic ductal adenocarcinoma also found CTNNB1 mutations (Honda et al., 2013; Javadinia et al., 2019). Multiple studies have shown that CTNNB1 mutation analysis is important for PAAD and COAD (Kubota et al., 2015).

ERBB2, commonly referred to as HER2, may be critical for enhancing the synergistic effect of PI3K inhibitors in HNSC patients (Michmerhuizen et al., 2019). It is generally believed that dysregulated ERBB2 signaling plays a key role in the development of pancreatic cancer (Lin et al., 2019). For the treatment of intestinal adenocarcinoma, ERBB2 mutations and amplification in small intestinal adenocarcinoma patients also make a great contribution (Adam et al., 2019). Recent studies have shown that HER2 targeted therapy has significantly improved outcomes in patients with breast and stomach problems with ERBB2 mutation/amplification (Meric-Bernstam et al., 2019).

The CDH1 gene plays a regulatory role in cell growth (Nagai et al., 2018), and the CDH1 gene located on chromosome 16q22.1 is considered to be a tumor suppressor of diffuse gastric cancer. By measuring the methylation profile of gastric cancer and breast cancer patients, it is found that CDH1 is closely related to low protein expression (Wang et al., 2016b; Wang et al., 2016c). Studies have shown that abnormal expression of CDH1 gene leads to uncontrolled growth of tumor cells (Dial et al., 2007; Chen et al., 2012).

The above experimental results show that Huber-SGNMF model can find pathogenic genes more effectively. Although some genes have not been confirmed, they may be a key part of solving cancer problems in the future.

Clustering Results and Analysis

After the Huber-SGNMF model reduces the dimensions of the data, the coefficient matrix is used for k-means clustering. Sample clustering is a common analytical method for cancer diagnosis and molecular subtype discrimination (Xu et al., 2019). Moreover, multiple evaluation criteria accuracy (ACC), recall (R), precision (P), and F-measure (F) are adopted to judge the model to be feasible and effective. ACC is an evaluation standard

TABLE 1 | Relevance scores for seven methods.

		NMF	SNMF	GNMF	RGNMF	RPCA	CGNMF	Huber-NMF	Huber-SGNMF	
PHD	MRS	116.4	113.99	116.4	164.03	**194.01**	113.99	164.03	164.03
	ARS	22.64	21.75	22.03	22.16	26.03	21.75	25.56	**27.19**
PHDEC	MRS	92.51	96.36	153.66	124.37	**172.9**	164.91	145	**172.9**
	ARS	29.18	30.05	36.58	27.87	37.83	35.07	35.93	**44.97**

Bolded texts denoted best experimental results.

TABLE 2 | Detailed analysis of the differentially expressed genes in PHD dataset.

Gene name	Relevance score	Gene official name	Related diseases
CTNNB1	164.03	Catenin beta 1	Colorectal cancer and pilomatrixoma
ERBB2	152.33	Erb-B2 receptor tyrosine kinase 2	Lung cancer and ovary adenocarcinoma
CDH1	149.92	Cadherin 1	Gastric cancer and breast cancer
TGFBR2	102.74	Transforming growth factor beta receptor 2	Colorectal cancer and esophageal cancer
CDK4	93.35	Cyclin dependent kinase 4	Myeloma and melanoma
EPCAM	86.79	Epithelial cell adhesion molecule	Pancreatic cancer and gastrointestinal carcinoma
GNAS	76.17	GNAS complex locus	Osseous heteroplasia
ERBB3	74.35	Erb-B2 receptor tyrosine kinase 3	Transitional cell carcinoma
CEACAM5	59.9	Carcinoembryonic antigen related cell Adhesion molecule 5	Colorectal cancer and lung cancer
MAP2K2	51.51	Mitogen-activated protein kinase kinase 2	Head and neck squamous cell carcinoma

that can visually reflect the clustering of samples. It is defined as follows:

$$ACC = \frac{\sum_{i=1}^{n} \delta\left(a_i, \text{map}\left(b_i\right)\right)}{n}, \tag{35}$$

Where δ (•) and map (•) represent function permutation and delta mapping function, respectively. The actual sample label, the predicted sample label, and the total number of samples are denoted by *a, b* and *n*, respectively.

Considering clustering accuracy alone does not fully demonstrate clustering performance, and more evaluation criteria need to be introduced. The clustering results can be divided into true positive (TP), true negative (TN), false positive (FP) cases, and false negative (FN) according to real and predictive labels. These four measures are listed in **Table 3**. The detailed evaluation criteria are as follows.

$$R = \frac{TP}{TP + FN}, \tag{36}$$

$$P = \frac{TP}{TP + FP}, \tag{37}$$

$$F = \frac{2 \times R \times P}{R + P}. \tag{38}$$

Since *R, P,* and *F* can only reflect the clustering performance of a certain sample categories, for multi-category problems, the average of each category of indicators is usually used as the evaluation criterions:

$$Macro - R = \frac{1}{n} \sum_{i=1}^{n} R_i, \tag{39}$$

$$Macro - P = \frac{1}{n} \sum_{i=1}^{n} P_i, \tag{40}$$

$$Macro - F = \frac{2 \times Macro - R \times Macro - P}{Macro - R + Macro - P}, \tag{41}$$

where *n* represents the number of sample categories.

According to the above evaluation criterions, each algorithm is performed 50 times to get an average result. Since the initialization matrix is random, the average value can reduce the chance of the algorithm. **Table 4** lists the comparative experiments of seven methods based on four evaluation criterions. Compared with the other six methods, our proposed model has the excellent clustering performance under the four evaluation criterions. The specific analysis of the clustering results is as follows:

1. Since the squared loss of the original NMF is sensitive to noise and outliers, the squared loss is replaced by Huber loss to improve the robustness of the algorithm. The experimental results show that the clustering performance of RPCA, CGNMF, RGNMF, Huber-NMF, and Huber-SGNMF is higher than standard NMF and GNMF. The reason is that both NMF and GNMF use square loss while other methods use more robust loss function. Moreover, the experimental results show that the robustness of the Huber loss model is higher than the $L_{2,1}$ -norm loss and correntropy loss. The RPCA model has higher performance as a state-of-the-art algorithm and is still lower than Huber-SGNMF. The Huber loss use L_1 -norm or L_2 -norm to different data, which can effectively reduce the influence of noise and outliers and enhance the robustness of the algorithm. Compared with NMF, the clustering accuracy of Huber-SGNMF model on the two datasets increased by 4.90 and 5.68%, respectively.
2. Assuming that data points are related in a high-dimensional state, they should also be relevant in low-dimensional representations. However, the association between data points is difficult to preserve when the data is mapped to low-dimensions. The manifold structure preserves the spatial structure of high-dimensional data in low-dimensional representations, enhancing the correlation between data points. Constructing a sample association graph of gene expression data to preserve the relationship between the samples. The experimental results of several models (NMF

TABLE 3 | Clustering result confusion matrix.

The true situation	Clustering result	
	Positive	Negative
Positive	TP (true positive)	FN (false negative)
Negative	FP (false positive)	TN (true negative)

TABLE 4 | Clustering effect for seven methods.

Dataset	Evaluation	NMF	SNMF	GNMF	RGNMF	RPCA	CGNMF	Huber-NMF	Huber-SGNMF
PHD	ACC (%)	85.38 ± 1.24	88.93 ± 0.58	86.05 ± 1.97	86.50 ± 1.84	86.37 ± 2.04	87.18 ± 1.43	88.55 ± 0.98	**90.36 ± 0.91**
	Macro-R (%)	82.99 ± 1.57	86.86 ± 0.82	81.02 ± 1.09	84.28 ± 2.40	84.10 ± 2.79	85.00 ± 1.79	86.41 ± 1.27	**88.50 ± 1.19**
	Macro-P (%)	84.88 ± 1.74	89.08 ± 0.86	83.55 ± 3.76	85.68 ± 2.74	85.58 ± 2.96	86.77 ± 2.02	88.32 ± 1.36	**90.18 ± 1.28**
	Macro-F (%)	83.92 ± 1.65	87.95 ± 0.84	82.25 ± 3.51	84.92 ± 2.60	84.83 ± 2.88	85.87 ± 1.90	87.35 ± 1.31	**89.33 ± 1.23**
PHDEC	ACC (%)	69.84 ± 0.26	71.51 ± 0.31	70.15 ± 0.08	71.86 ± 0.69	75.02 ± 0.32	73.81 ± 0.27	72.53 ± 0.21	**75.52 ± 0.20**
	Macro-R (%)	63.95 ± 0.18	65.33 ± 0.14	61.98 ± 0.38	64.45 ± 0.87	68.37 ± 0.28	66.74 ± 0.15	67.09 ± 0.07	**69.02 ± 0.07**
	Macro-P (%)	61.34 ± 0.26	62.45 ± 0.19	58.77 ± 0.10	62.80 ± 0.97	**65.81 ± 0.50**	64.47 ± 0.27	63.92 ± 0.25	65.56 ± 0.25
	Macro-F (%)	64.17 ± 0.20	63.79 ± 0.21	60.24 ± 0.27	63.49 ± 0.87	66.92 ± 0.29	65.51 ± 0.17	65.34 ± 0.12	**67.17 ± 0.10**

Bolded texts denoted best experimental results.

and GNMF, Huber-NMF, and Huber-SGNMF) show that the clustering performance of the model with the addition of graph regularity constraints is improved. Compared with Huber-NMF, Huber-SGNMF has improved clustering accuracy by 1.73 and 2.99% in the two datasets, respectively.

3. Matrix sparseness removes redundant data and simplifies model calculations. The sparsity constraint of the coefficient matrix removes redundant features and improves clustering performance. The experimental results of SNMF and Huber-SGNMF prove this. Compared with SNMF, since Huber-SGNMF improves the loss function and manifold structure, the clustering accuracy in the two datasets is increased by 1.35 and 4.02%, respectively.

In summary, the experimental results based on the four evaluation indicators demonstrate the excellent clustering performance of the Huber-SGNMF model. Compared with NMF, the clustering performance of Huber-SGNMF has improved 5.30 and 4.49% on average in PHD dataset and PHDEC dataset, respectively. Huber-SGNMF clustering performance improves 1.93 and 2.07% on average compared to Huber-NMF. The above experimental results strongly prove the effectiveness of Huber-SGNMF in clustering performance.

CONCLUSION

In this paper, we propose a novel model based on Huber loss: Huber-SGNMF, which is dedicated to samples clustering and differentially expressed gene selection. On the one hand, the squared loss is replaced by Huber loss to enhance algorithm robustness. On the other hand, sparse penalty and graph regularization terms are added to the model to enhance the sparsity of the matrix and preserve data geometry information. Numerous experimental results confirm that the Huber-SGNMF method is more effective. In the future work, we will actively explore more effective constraints based on the traditional NMF method to improve the robustness and sparsity of the method.

DATA AVAILABILITY STATEMENT

The datasets for this study can be downloaded in the The Cancer Genome Atlas [https://cancergenome.nih.gov/].

AUTHOR CONTRIBUTIONS

C-YW and NY proposed and designed the algorithm. J-XL demonstrated the robustness of the algorithm and analyzed the experimental data. C-YW and C-HZ drafted the manuscript.

ACKNOWLEDGMENTS

This work was supported in part by the grants provided by the National Science Foundation of China, Nos. 61872220, 61873001, and 61572284.

REFERENCES

Adam, L., San Lucas, F. A., Fowler, R., Yu, Y., Wu, W., Liu, Y., et al. (2019). DNA sequencing of small bowel adenocarcinomas identifies targetable recurrent mutations in the ERBB2 signaling pathway. *Clin. Cancer Res.* 25 (2), 641–651. doi: 10.1158/1078-0432.CCR-18-1480

Belkin, M., and Niyogi, P. (2001). "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. (MIT Press).

Cai, D., He, X., Han, J., and Huang, T. S. (2011). Graph regularized non-negative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8), 1548–1560. doi: 10.1109/TPAMI.2010.231

Chen, J., Han, G., Xu, A., and Cai, H. (2019). Identification of multidimensional regulatory modules through multi-graph matching with network constraints. *IEEE Trans. Biomed. Eng*. doi: 10.1109/TBME.2019.2927157

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2018). HOGMMNC: a higher order graph matching with multiple network constraints model for gene-drug regulatory modules identification. *Bioinformatics* 35 (4), 602–610. doi: 10.1093/bioinformatics/bty662

Chen, M., Gutierrez, G. J., and Ronai, Z. A. (2012). The anaphase-promoting complex or cyclosome supports cell survival in response to endoplasmic reticulum stress. *PloS One* 7 (4), e35520. doi: 10.1371/journal.pone.0035520

Chen, X., Feng, M., Shen, C., He, B., Du, X., Yu, Y., et al. (2017). A novel approach to select differential pathways associated with hypertrophic cardiomyopathy based on gene coexpression analysis. *Mol. Med. Rep.* 16 (1), 773–777. doi: 10.3892/mmr.2017.6667

Chreiky, R., Delmaire, G., Puigt, M., Roussel, G., and Abche, A. (2016). "Informed split gradient non-negative matrix factorization using huber cost function for source apportionment," in: 2016 *IEEE International Symposium on Signal Processing and Information Technology* (*ISSPIT*): *IEEE*), 69–74. doi: 10.1109/ISSPIT.2016.7886011

Dial, J. M., Petrotchenko, E. V., and Borchers, C. H. (2007). Inhibition of APCCdh1 activity by Cdh1/Acm1/Bmh1 ternary complex formation. *J. Boil. Chem.* 282 (8), 5237–5248. doi: 10.1074/jbc.M606589200

Ding, C., Zhou, D., He, X., and Zha, H. (2006). "R 1-PCA: rotational invariant L 1-norm principal component analysis for robust subspace factorization," in: *Proceedings of the 23rd international conference on Machine learning*: ACM), 281–288.

Du, L., Li, X., and Shen, Y.-D. (2012). "Robust nonnegative matrix factorization via half-quadratic minimization," in 2012 *IEEE 12th International Conference on Data Mining*. (IEEE), 201–210. doi: 10.1109/ICDM.2012.39

Feng, C.-M., Xu, Y., Liu, J.-X., Gao, Y.-L., and Zheng, C.-H. (2019). Supervised discriminative sparse PCA for com-characteristic gene selection and tumor classification on multiview biological data. *IEEE Trans. Neural Netw. Learn. Syst*. doi: 10.1109/TNNLS.2019.2893190

Guillamet, D., and Vitrià, J. (2002). "Non-negative matrix factorization for face recognition," in *Catalonian Conference on Artificial Intelligence*. (Springer), 336–344. doi: 10.1007/3-540-36079-4_29

Guofa, S., Ling, X., Feng, L., Mingfeng, J., and Stuart, C. (2011). On epicardial potential reconstruction using regularization schemes with the L1-norm data term. *Phys. Med. Biol.* 56 (1), 57–72. doi: 10.1088/0031-9155/56/1/004

Honda, S., Okada, T., Miyagi, H., Minato, M., Suzuki, H., and Taketomi, A. (2013). Spontaneous rupture of an advanced pancreatoblastoma: Aberrant RASSF1A methylation and CTNNB1 mutation as molecular genetic markers. *J. Pediatr. Surg.* 48 (4), e29–e32. doi: 10.1016/j.jpedsurg.2013.02.038

Javadinia, S. A., Shahidsales, S., Fanipakdel, A., Joudi-Mashhad, M., Mehramiz, M., Talebian, S., et al. (2019). Therapeutic potential of targeting the Wnt/β-catenin pathway in the treatment of pancreatic cancer. *J. Cell. Biochem.* 120 (5), 6833–6840. doi: 10.1002/jcb.27835

Kim, H., and Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23 (12), 1495–1502. doi: 10.1093/bioinformatics/btm134

Kong, D., Ding, C. H. Q., and Huang, H. (2011). "Robust nonnegative matrix factorization using L21-norm," in *Proceedings of the 20th ACM international conference on Information and knowledge management* (ACM) 673–682. doi: 10.1145/2063576.2063676

Kubota, Y., Kawakami, H., Natsuizaka, M., Kawakubo, K., Marukawa, K., Kudo, T., et al. (2015). CTNNB1 mutational analysis of solid-pseudopapillary neoplasms of the pancreas using endoscopic ultrasound-guided fine-needle aspiration and next-generation deep sequencing. *J. Gastroenterol.* 50 (2), 203–210. doi: 10.1007/s00535-014-0954-y

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788. doi: 10.1038/44565

Li, X., Cui, G., and Dong, Y. (2017). Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE Trans. Cybern.* 47 (11), 3840–3853. doi: 10.1109/TCYB.2016.2585355

Lin, T., Ren, Q., Zuo, W., Jia, R., Xie, L., Lin, R., et al. (2019). Valproic acid exhibits anti-tumor activity selectively against EGFR/ErbB2/ErbB3-coexpressing pancreatic cancer *via* induction of ErbB family members-targeting microRNAs. *J. Exp. Clin. Cancer Res.* 38 (1), 150. doi: 10.1186/s13046-019-1160-9

Liu, J.-X., Wang, D., Gao, Y.-L., Zheng, C.-H., Xu, Y., and Yu, J. (2018). Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: a survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (3), 974–987. doi: 10.1109/TCBB.2017.2665557

Liu, J.-X., Wang, Y.-T., Zheng, C.-H., Sha, W., Mi, J.-X., and Xu, Y. (2013). "Robust PCA based method for discovering differentially expressed genes," in *BMC bioinformatics.* (BioMed Central), 14(8) S3. doi: 10.1186/1471-2105-14-S8-S3

Liu, W., Pokharel, P. P., and Principe, J. C. (2007). Correntropy: properties and applications in non-gaussian signal processing. *IEEE Trans. Signal Process.* 55 (11), 5286–5298. doi: 10.1109/TSP.2007.896065

Liu, X., Zhai, D., Zhao, D., Zhai, G., and Gao, W. (2014). Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. *IEEE Trans. Image Process.* 23 (4), 1491–1503. doi: 10.1109/TIP.2014.2303638

Luo, X., Zhou, M., Xia, Y., and Zhu, Q. (2014). An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Trans. Ind. Inform.* 10 (2), 1273–1284. doi: 10.1109/TII.2014.2308433

Mao, B., Guan, N., Tao, D., Huang, X., and Luo, Z. (2014). "Correntropy induced metric based graph regularized non-negative matrix factorization," *Proceedings 2014 IEEE international conference on Security, Pattern Analysis, and Cybernetics (SPAC).* (IEEE), 163–168. doi: 10.1109/SPAC.2014.6982679

Meric-Bernstam, F., Johnson, A. M., Dumbrava, E. E. I., Raghav, K., Balaji, K., Bhatt, M., et al. (2019). Advances in HER2-targeted therapy: novel agents and opportunities beyond breast and gastric cancer. *Clin. Cancer Res.* 25 (7), 2033–2041. doi: 10.1158/1078-0432.CCR-18-2275

Michmerhuizen, N. L., Leonard, E., Matovina, C., Harris, M., Herbst, G., Kulkarni, A., et al. (2019). Rationale for using irreversible epidermal growth factor receptor inhibitors in combination with phosphatidylinositol 3-kinase inhibitors for advanced head and neck squamous cell carcinoma. *Mol. Pharmacol.* 95 (5), 528–536. doi: 10.1124/mol.118.115162

Nagai, M., Shibata, A., and Ushimaru, T. (2018). Cdh1 degradation is mediated by APC/C-Cdh1 and SCF-Cdc4 in budding yeast. *Biochem. Biophys. Res. Commun.* 506 (4), 932–938. doi: 10.1016/j.bbrc.2018.10.179

Nikolova, M., and Chan, R. H. (2007). The equivalence of half-quadratic minimization and the gradient linearization iteration. *IEEE Trans. Image Process.* 16 (6), 1623–1627. doi: 10.1109/TIP.2007.896622

Qi, L., and Jiang, H. (1997). Semismooth Karush-Kuhn-Tucker equations and convergence analysis of newton and quasi-newton methods for solving these equations. *Math. Oper. Res.* 22 (2), 301–325. doi: 10.1287/moor.22.2.301

Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290 (5500), 2323–2326. doi: 10.1126/science.290.5500.2323

Schmidt, M. N., and Olsson, R. K. (2006). "Single-channel speech separation using sparse non-negative matrix factorization," in *Ninth International Conference on Spoken Language Processing* (IEEE).

Wang, D., Liu, J.-X., Gao, Y.-L., Zheng, C.-H., and Xu, Y. (2016a). Characteristic gene selection based on robust graph regularized non-negative matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13 (6), 1059–1067. doi: 10.1109/TCBB.2015.2505294

Wang, H. L., Hart, J., Fan, L., Mustafi, R., and Bissonnette, M. (2011). Upregulation of glycogen synthase kinase 3β in human colorectal adenocarcinomas correlates with accumulation of CTNNB1. *Clin. Colorectal Cancer* 10 (1), 30–36. doi: 10.3816/CCC.2011.n.004

Wang, Q., Wang, B., Zhang, Y.-M., and Wang, W. (2016b). The association between CDH1 promoter methylation and patients with ovarian cancer: a systematic meta-analysis. *J. Ovarian Res.* 9, 23. doi: 10.1186/s13048-016-0231-1

Wang, Y. Q., Yuan, Y., Jiang, S., and Jiang, H. (2016c). Promoter methylation and expression of CDH1 and susceptibility and prognosis of eyelid squamous cell carcinoma. *Tumor Biol.* 37 (7), 9521–9526. doi: 10.1007/s13277-016-4851-2

Wu, M.-J., Liu, J.-X., Gao, Y.-L., Kong, X.-Z., and Feng, C.-M. (2017). "Feature selection and clustering *via* robust graph-laplacian PCA based on capped L 1-norm," in 2017 *IEEE international conference on Bioinformatics and Biomedicine (BIBM)* (IEEE), 1741–1745. doi: 10.1109/BIBM.2017.8217923

Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. *Front. Genet.* 10, 236. doi: 10.3389/fgene.2019.00236

Xu, W., Liu, X., and Gong, Y. (2003). "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (ACM), 267–273.

Yu, L., and Liu, H. (2003). "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03).* (AAAI), 856–863.

Yu, N., Liu, J.-X., Gao, Y.-L., Zheng, C.-H., Wang, J., and Wu, M.-J. (2017). "Graph regularized robust non-negative matrix factorization for clustering and selecting differentially expressed genes," in *2017 IEEE international conference on Bioinformatics and Biomedicine (BIBM)* (IEEE), 1752–1756. doi: 10.1109/BIBM.2017.8217925

Yu, T., Zhao, Z., Yan, Z., and Li, F. (2016). "Robust L1-norm matrixed locality preserving projection for discriminative subspace learning," in *2016 International Joint Conference on Neural Networks (IJCNN)* (IEEE), 4199–4204.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fgene.2019.01182

Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification

Qiang Zhu[1,2], *Xingpeng Jiang*[2,3]*, *Qing Zhu*[2,3], *Min Pan*[2,3] *and Tingting He*[2,3]

[1] School of Information Management, Central China Normal University, Wuhan, China, [2] School of Computer, Central China Normal University, Wuhan, China, [3] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China

The microbiome-wide association studies are to figure out the relationship between microorganisms and humans, with the goal of discovering relevant biomarkers to guide disease diagnosis. However, the microbiome data is complex, with high noise and dimensions. Traditional machine learning methods are limited by the models' representation ability and cannot learn complex patterns from the data. Recently, deep learning has been widely applied to fields ranging from text processing to image recognition due to its efficient flexibility and high capacity. But the deep learning models must be trained with enough data in order to achieve good performance, which is impractical in reality. In addition, deep learning is considered as black box and hard to interpret. These factors make deep learning not widely used in microbiome-wide association studies. In this work, we construct a sparse microbial interaction network and embed this graph into deep model to alleviate the risk of overfitting and improve the performance. Further, we explore a Graph Embedding Deep Feedforward Network (GEDFN) to conduct feature selection and guide meaningful microbial markers' identification. Based on the experimental results, we verify the feasibility of combining the microbial graph model with the deep learning model, and demonstrate the feasibility of applying deep learning and feature selection on microbial data. Our main contributions are: firstly, we utilize different methods to construct a variety of microbial interaction networks and combine the network *via* graph embedding deep learning. Secondly, we introduce a feature selection method based on graph embedding and validate the biological meaning of microbial markers. The code is available at https://github.com/MicroAVA/GEDFN.git.

Keywords: graph embedding, deep learning, feature selection, biomarkers, microbiome

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Lingling Jin,
Thompson Rivers University, Canada
Xishuang Dong,
Prairie View A&M University, United States
Xiangrong Liu,
Xiamen University, China

***Correspondence:**
Xingpeng Jiang
xpjiang@mail.ccnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *12 August 2019*
Accepted: *24 October 2019*
Published: *22 November 2019*

Citation:
Zhu Q, Jiang X, Zhu Q, Pan M and He T (2019) Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. Front. Genet. 10:1182. doi: 10.3389/fgene.2019.01182

INTRODUCTION

A large number of microorganisms are parasite on various parts of the human body, mainly concentrated in the intestine, oral cavity, reproductive tract, epidermis and skin. The microbial communities existing in different parts of the body or in different host environments are very different (Turnbaugh et al., 2007; Lloyd-Price et al., 2017). These microorganisms include bacteria, fungi, viruses and protozoa. All genetic material in the particular microbial community is called the microbiome. Recent studies have shown that microorganisms are directly or indirectly related to many diseases. For example, the gut microbiome may be closely related to irritable

bowel syndrome and its imbalance may lead to chronic kidney diseases. Microorganisms may also be closely related to digestive tract diseases, endocrine diseases, circulatory diseases, reproductive system diseases, respiratory and psychiatric diseases (Kho and Lal, 2018). Since the microbiome plays a central role in the hosts' health, understanding the distribution and composition of microbial communities in humans, especially under different diseases or physiological conditions, is of great significance for disease diagnosis, prevention and treatment. The microbiome-wide association studies are to find disease-associated microbial markers to guide disease diagnosis and treatment (Gilbert et al., 2016; Wang and Jia, 2016). Compared with the human genome, the microbiome is an ideal target and more convenient to regulate. Therefore, the microbiome is often named "the second human genome" (Brüls and Weissenbach, 2011). However, there are many types of microorganisms and most of them cannot be cultured. Therefore, a high-throughput sequencing method is a feasible means of understanding microbial communities. Through high-throughput sequencing, we can understand the types of microorganisms and even their functions in the community (Ranjan et al., 2016).

The microbiome data is from high-throughput sequencing methods such as 16s or shotgun sequencing, which is often with high dimensions with noise. As a result, it is difficult to mine microbial signatures from these data. Traditionally, statistical-based methods identify markers mainly through microbial abundance differential expression (Paulson et al., 2013). However, the statistical approaches often have strong assumptions and the real data often do not satisfy these assumptions (Hawinkel et al., 2017; Weiss et al., 2017). Other machine learning methods are widely explored (Pasolli et al., 2016). Recently, deep learning has received great attention, especially its end-to-end automatic learning ability. At present, deep learning is widely used in automatic driving, image recognition and text processing, which has received exciting results (LeCun et al., 2015). The deep models can learn specific patterns directly from the data, thus avoiding the artificial feature engineering (Goodfellow et al., 2016; Kong and Yu, 2018). In the analysis of biomedical data, especially the analysis of various omics data, deep learning has achieved good improvement, but still faces many problems and challenges (Angermueller et al., 2016; Camacho et al., 2018; Eraslan et al., 2019). First, deep learning requires a large amount of training data to learn useful information while the biological sample size is often limited and cannot fully utilize its capabilities. Second, the training process is often considered a black box and people can only control the input and models' parameters. More specifically, deep learning involves complex network structures and nonlinear transformations, as well as a large number of hyperparameters, which hinder people from understanding how deep neural networks are making predictions. Although deep neural networks perform well on some classification tasks, biological problems should be paid more attention to which features lead to better classification (Ching et al., 2018).

In this paper, we propose a feature selection method based on Graph Embedding Deep Feedforward Network (GEDFN) to conduct microbiome-wide association studies. Firstly, we construct three different microbial co-occurrence interaction networks. We utilize a graph embedding method to embed the network as *a priori* knowledge into Deep Feedforward Neural Network to reduce parameters, alleviate the overfitting problem and improve the models' performance. Secondly, we propose a feature selection approach based on GEDFN. Experiments show the microbial feature markers obtained *via* this method have biological significance. In other words, our results demonstrate graph embedding deep learning could guide feature selection.

RELATED WORK

Microbial Interaction Network

Because of the various relationships between microorganisms, such as symbiosis, competition and so on, as well as the complex structure and function of microorganisms due to their dynamic properties, the network is a good way to represent complex relationships. Understanding microbial interaction can help us understand microbial functions. System-oriented graph theory can facilitate microbial analysis and enhance our understanding of complex ecosystems and evolutionary processes (Faust et al., 2012; Layeghifard et al., 2017). However, most microorganisms are uncultured, we can only construct microbial interaction networks from high-throughput sequencing data. At present, there are many computational methods to construct microbial interaction networks. In theory, any method of calculating features' relationships can be used. For example, Bray–Curtis can be used to measure species abundance similarity (Bray and Curtis, 1957). The Pearson correlation coefficient is used to evaluate the linear relationship and the Spearman correlation coefficient can measure the rank relationship (Mukaka, 2012). CoNet uses an ensemble approach and combines with different comparison metrics to detect different relationships (Faust and Raes, 2016). Maximum mutual information is designed to capture broader relationships, not limited to specific function families (Reshef et al., 2011). MENA applies random matrix theory to conduct microbial analysis and experiments show it is robust to the noise and threshold (Deng et al., 2012). Sparse Correlations for Compositional data (SparCC) is a tool based on Aitchison's log ratio transformation to conduct microbial composition analysis (Friedman and Alm, 2012). SParse InversE Covariance Estimation for Ecological Association Inference (SPIEC-EASI) combines data logarithmic transformation with graph model inference framework to build a correlation network (Kurtz et al., 2015).

Feature Selection

Real biomedical data, especially various omics data with high dimensions and noise, often has feature redundancy problem. Feature selection is a step of data preprocessing, which involves selecting related features from a large number of features to improve subsequent learning tasks (Li et al., 2017).

There are mainly three kinds of feature selection methods, including filter, wrapper and embedded method. The filter approach selects subset features and then trains the learner. The feature selection process is independent of the subsequent learner. This is equivalent to filter the initial feature with the feature selection process and train the model with the filtered features. However, filter methods often ignore some features that are helpful for classification. At the same time, many filter methods are based on a single-featured greedy algorithm. The assumption is that each feature is independent while this is often not the case in microbiological data. The wrapper feature selection directly takes the performance of the learner to be used as the evaluation criterion of the feature subset. In other words, the purpose of the wrapper feature selection is to select a feature subset that is most efficient in its performance for a given learner. Compared to the filter method, the wrapper method can evaluate the result of feature selection to improve the classification performance; however, the feature selection process requires to train the learner iteratively and the calculation is huge (Li et al., 2017). The embedded feature selection combines the feature selection in the learning and training process, both of which are completed in the same optimization. In other words, the feature selection is automatically performed during the training.

Feature selection is a traditional machine learning research field with many methods. For more information, please refer to the literature (Li et al., 2017). The previous work proposed a feature selection method based on Deep Forest (Zhu et al., 2018); however, there is less work on microbiome-wide association studies *via* Deep Neural Network and less research is done from the perspective of embedding approach for feature selection. The challenge of feature selection based on microbial network is that there is no microbial network available at present. The commonly used statistical-based interaction network methods may lead to high false positive rate due to the compositional bias (Gloor et al., 2017).

MATERIALS AND METHODS

We mainly explain the feature selection method based on GEDFN from the following three aspects (**Figure 1**). First, we will introduce the construction method of microbial interaction network, including sparcc, SPIEC-EASI and Maximal Information Coefficient (MIC) then, we will introduce a deep embedding structure to embed the graph into Deep Feedforward Network. Finally, we will propose a feature selection approach for GEDFN.

Microbial Correlation Network

The total amount of genetic material extracted from the microbial community and the sequencing depth will affect the whole reads. It is often necessary to normalize the reads in the sample. As a result, the microbial abundance obtained by 16s sequencing is relative rather than absolute, which is not independent. The traditional statistical measures for detecting microbial interactions, for example, Pearson correlation, will lead to false positives (Gloor et al., 2017).

Sparcc

Assuming that the network is sparse, sparcc constructs the association network by using standard logarithmic ratio transformation and iteratively calculates the variance matrix of

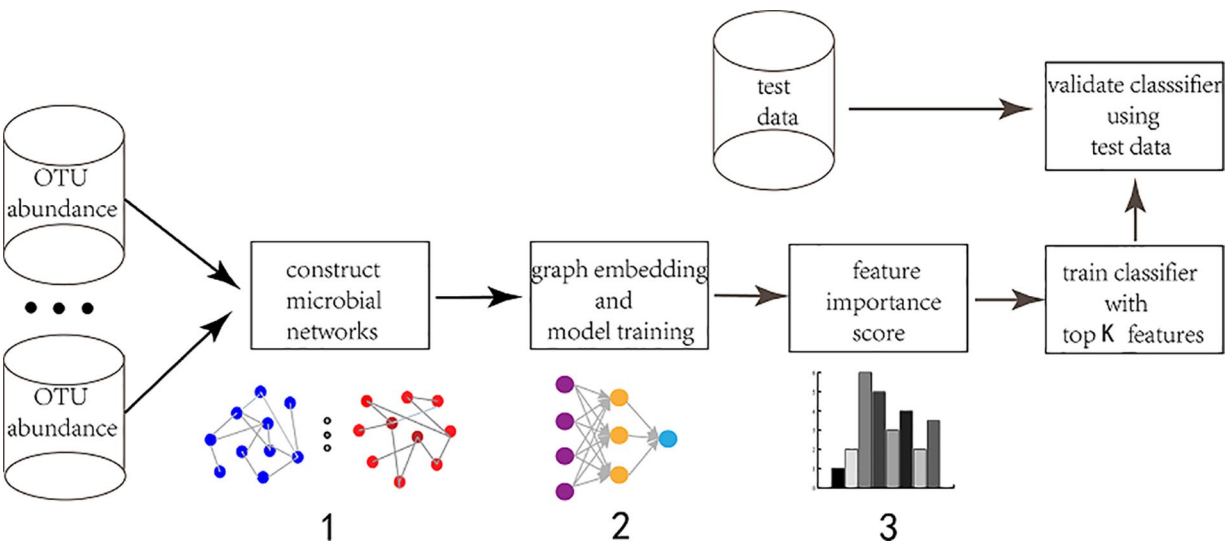


FIGURE 1 | The workflow of graph embedding deep network to conduct feature selection. 1. Construct microbial interaction network. The input is Operational Taxonomic Unit (OTU) abundance. Different approaches are adopted to obtain different interaction networks. The vertexes are species and the edges are correlation coefficient. 2. Graph embedding and model training. The feature graph is embedded into the first hidden layer in order to achieve sparse connection instead of fully-connected between the input layer and the first hidden layer. The first hidden layer (graph embedding layer) has the same neurons as the input layer. 3. Feature selection. The neurons (features) are ranked according to their importance score which is calculated *via* each neuron's connection weights.

compositional dependence. For details of the algorithm, please refer to the literature (Friedman and Alm, 2012).

SPIEC-EASI

SPIEC-EASI assumes the network is sparse and combines logarithmic transformation of compositional data with graph inference framework to construct the network. It consists of two steps: first, logarithmic ratio transforms the data; then, SPIEC-EASI uses the neighborhood selection and sparse inverse covariance selection to infer the interaction graph from the transformed data (Kurtz et al., 2015).

Maximal Information Coefficient

The maximal information coefficient (MIC) is used to measure the degree of linear and nonlinear correlation between two variables (Reshef et al., 2011). The main idea of the MIC method is based on the recognition that if there is some correlation between two variables, the distribution of the data in the grid can be reflected after meshing the scatter plots formed by the two variables. The MIC divides the scatter plot of the variable pair (x, y) and uses dynamic programming to calculate and search for the maximum mutual information value that can be achieved under different split modes. Finally, the maximum mutual information value is normalized and the result is MIC.

The Framework of Graph Embedding Deep Feedforward Network

Deep Feedforward Neural Network

Deep Feedforward Network, also known as feedforward neural network or multilayer perceptron, is a typical deep learning model. In this model, the information moves only in one direction from the input nodes to the output nodes through the hidden nodes. There is no loop in the network. A feedforward neural network structure with l hidden layers is:

$$\mathrm{P}(\mathrm{y}|\mathrm{X},\theta) = f\left(Z_{out}W_{out} + b_{out}\right) \tag{1}$$

$$Z_{out} = \sigma\left(Z_l W_l + b_l\right) \tag{2}$$

… …

$$Z_{k+1} = \sigma\left(Z_k W_k + b_k\right) \tag{3}$$

… …

$$Z_1 = \sigma\left(XW_{in} + b_{in}\right) \tag{4}$$

where $X \in R^{nxp}$ is an input matrix with n samples and p features, $y \in R^n$ is the output label for the classification task. In this work, it is a binary classification. The label for each sample is normal or disease. Z_{out} and Z_k,(k=1,…,l-1) are the neurons in the hidden layer. W_k is the weight matrix. b_k is the bias. θ is the parameters. $\sigma(\cdot)$is the activation function(such as, sigmoid, tanh, rectifiers). $F(\cdot)$ is a softmax function which is used to convert the output layer value into the predicted probability.

The model uses a stochastic gradient descent (SGD) algorithm to minimize the cross entropy loss function to update the parameter θ. When a feedforward neural network is used to receive input x and produce an output $\hat{y}$. During training, forward propagation can continue until it produces a scalar cost function $J(\theta)$. The backpropagation algorithm runs information from the cost function and flow backward through the network to calculate the gradient in order to update the weight parameters (Goodfellow et al., 2016).

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i log\hat{p}_i + (1-y_i)\log\left(1-\hat{p}_i\right)\right) \tag{5}$$

Graph Embedding Deep Feedforward Network

The fully connected deep feedforward neural network has many parameters and requires a large number of training data, but often the biological sample size is limited, which often leads to overfitting. Therefore, we construct a microbial sparse network and embed this graph network into the model. There are two main advantages. First, the sparse graph embedding will greatly reduce the parameters of deep feedforward network and mitigate the overfitting risk. Second, the sparse graph structure is derived from existing prior information and combining the priori information into the network can improve the reliability of the model. The main idea of graph embedding is to replace the full connections between the input layer and the first hidden layer with a sparse graph (**Figure 2**).

Consider a graph G=(V,E), V is the vertical set with p features. E is a collection of all edges. A common way of representing a graph is to use an adjacency matrix. Given a graph G with p vertices, a pxp adjacency matrix A is:

$$A_{ij} = \begin{cases} 1, \textit{ if } \mathrm{V_i} \textit{ and } \mathrm{V_j} \text{ connected}, & \forall i,j = 1,\ldots,p \\ 0, & \text{otherwise.} \end{cases}$$

G is an undirected graph and A is a symmetric matrix. At the same time, we consider A_{ii}=1 which indicates that the vertex itself is connected. We construct a feedforward neural network in which the first hidden layer has the same dimensions as the input layer, h_{in}=p, similarly,W_{in} is a pxp matrix. The input X is sparsely connected with Z_1 (**Figure 2**). In other words, the original fully connected layer:

$$Z_1 = \sigma\left(XW_{in} + b_{in}\right) \tag{6}$$

is changed to:

$$Z_1 = \sigma\left(X\left(W_{in} \odot A\right) + b_{in}\right) \tag{7}$$

Where $\odot$ is element-wise product. Therefore, the connection between the input and first hidden layer of the feedforward network is filtered by the graph adjacency matrix. Each feature is corresponding to a hidden neuron. All features have corresponding

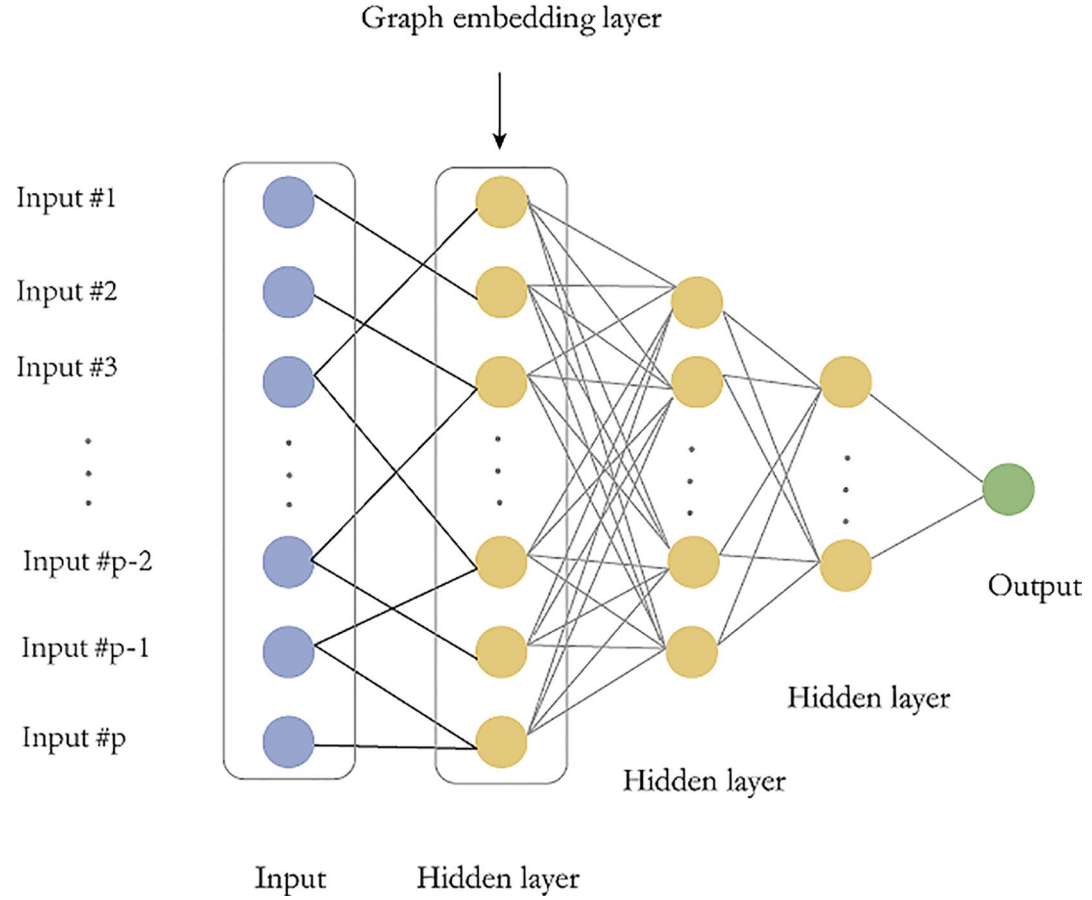


FIGURE 2 | Graph embedding deep feedforward network (GEDFN). The graph embedding layer (first hidden layer) has same neurons with the input layer. The sparse connect between the input layer and the first hidden layer is marked as black. Other hidden layers are fully-connected.

hidden neurons in the first hidden layer. The feature can only provide information to the connected graph. In this way, the graph helps to achieve the sparsity of the connection between the input layer and the first hidden layer (Kong and Yu, 2018).

Feature Selection Based on GEDFN

In addition to improving classification, it is also meaningful to find features that contribute significantly to classification because they reveal potential biological mechanisms. However, Deep neural network is a "black box", the interpretability of deep learning hasn't been well-defined (Guidotti et al., 2019). In our experiment, we focus on how the input features influence the prediction and we borrow the idea from Olden and Jackson (2002) and Kong and Yu (2018). The feature importance score is the quantification values of the contributions of features to a model prediction, which links the input features and output prediction. They highlight the parts of a given input that are most influential for the model prediction and thereby help to explain why such a prediction was made. The feature selection is based on feature score, which means the score is high if the feature is important. As a result, we develop a feature ranking method based on the feature relative importance score, similar to the connection weights method introduced by Olden and Jackson (2002) and Kong and Yu (2018). What is learned by neural networks is contained in the connection weights. Based on idea of connection weight, we propose a graphical connect weight method that emphasizes the importance of the features of our proposed neural network architecture.

The main idea of a graphical connect weight is: the contribution of a particular variable directly reflects the magnitude of the connection weights associated with the corresponding hidden neurons in the graph embedding layer. The sum of the absolute values of the directly related weights for a neuron (or feature) gives its relative importance:

$$s_j = \gamma_j \sum_{k=1}^{p} \left| w_{kj}^{(in)} I(A_{kj} = 1) \right| + \Sigma_{m=1}^{h_1} \left| w_{jm}^{(1)} \right|, \tag{8}$$

$$\gamma_j = \min\left(c / \sum_{k=1}^{p} \left(A_{kj} = 1 \right), 1 \right), \; j = 1, \ldots, p. \tag{9}$$

Where s_j is importance score of the feature j $w^{(in)}$ indicates the weights between the input layer and the first hidden layer, while $w^{(1)}$ indicates the weights between the first and second

hidden layer. The constant c is to penalize vertices with too many connections so that they don't over impact the result. In the following experiments, we set the parameter $c = 50$.

EXPERIMENTS AND RESULTS

Data Set

Inflammatory bowel diseases (IBD) are a group of specific chronic intestinal diseases, mainly including Crohn's disease and ulcerative colitis. The occurrence and development of IBD are closely related to intestinal microorganisms (Gevers et al., 2014). In our experiment, OTU BIOM files and metadata were downloaded from the QIITA (https://qiita.ucsd.edu/) database (study id: 1939). The detailed experiment was described in Gevers et al., 2014. The IBD data set consists of 1,359 metagenomic samples, including rectal, ileal biopsy and fecal samples (Gevers et al., 2014). We retained samples of mucosal tissue biopsies (terminal ileum and rectum) samples under the age of 18. The control group were without inflammatory conditions, such as abdominal pain and diarrhea. The final data set consisted of 657 IBD samples and 316 normal samples, respectively. We used QIIME's taxa collapse to filter the strain's species, limiting features at genus level.

Results

The Hyperparameters of Graph Embedding Deep Feedforward Neural Network

The structure of the graph embedding deep feedforward neural network (GEDFN) is shown in **Figure 2**. The most important part of GEDFN is that the number of neurons in the first hidden layer is the same as the number of neurons in the input layer and they are sparsely connected, which is different with normal fully connected feed forward neural network. The second layer, third and fourth hidden layers are consisting of 128, 64 and 16 neurons respectively and they are fully connected.

We use three different methods to construct a microbial co-occurrence interaction network from microbial abundance data. When the sparcc method is used to build the network, we reserve the vertexes if the correlation of two vertexes is larger than 0.3. We get an adjacency network with 63 vertexes and 315 edges. We adopt the mictools (Albanese et al., 2018) to build the MIC relevant network and we get 279 vertexes and 3230 edges when the correlation threshold is 0.2. The network constructed by sparcc and SPEC-EASI methods is sparse while MIC gets relatively a dense network. Different methods get different interaction networks. We find the higher the threshold, the more reliable is the network. However, the high threshold will make the network too sparse. As a result, we combine three kinds of networks to get a larger network with 736 vertexes and 18,034 edges. In this way, the connections between the input layer and the first hidden layer are more reliable and less dense than the fully connected approach.

Other hyperparameters of GEDFN are as follows: the learning rate is 0.0001, the activation function is Rectified Linear Unit (ReLU) and the weight initializer is he_uniform, the drop out is 0.2. the code is implemented in keras and available at https://github.com/MicroAVA/GEDFN.git.

The Evaluation of Classification

Traditional classification methods such as Random Forest has been shown to be the best performers in omics data classification tasks and the results show that Random Forest has achieved the best performance on microbial classification (Pasolli et al., 2016). Therefore, we compare GEDFN with Deep Forest (DF), Random Forest (RF) and Support Vector Machines (SVM). For the binary classification, we calculate the Area Under the Receiver Operating Characteristics (AUROC) and classification accuracy for each method (**Figure 3**).

AUROC curve is a performance measurement for classification problem at various thresholds settings, which can evaluate classifiers considering all true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Receiver Operating Characteristics (ROC) is a probability curve and Area Under the Curve (AUC) represents degree or measure of separability. It tells how much a model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. By analogy, the higher the AUC, the better the model is at distinguishing between patients with disease and no disease. The ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where TPR is on the y-axis and FPR is on the x-axis.

$$\mathrm{TPR} = \frac{TP}{TP + FN}, \; FPR = \frac{FP}{TN + FP}$$

The classification accuracy means the percentage of correct predictions from the total number of predictions made.

$$\mathrm{ACC} = \frac{1}{m}\sum_{i=1}^{m} I\left(\hat{y} = y\right)$$

Where $\hat{y}$ is the predicted label and y_i is the true label for the sample i. The m means the sample size and $I(\cdot)$is the indicator function.

In this experiment, we adopt a five-fold cross-validation. We use the implementation of Random Forest in python's scikit-learn package. We set the estimator parameter to 300. The Deep Forest is based on the work (Zhu et al., 2018). From the AUC value, we find that the Graph Embedding Deep Feedforward Network (GEDFN) is much better than SVM (AUC = 0.663). Compared with Deep Forest and Random Forest, GEDFN is also very competitive. GEDFN achieves an AUC value of 0.843, which is slightly better than Deep Forest (AUC = 0.834) and Random Forest (AUC = 0.823). In terms of classification accuracy, GEDFN achieves an average accuracy of 79.52%, Deep Forest achieves 76.6% and Random Forest achieves 75.16%. GEDFN outperforms 2–4% than Deep Forest and Random Forest. These methods are much better than SVM (67.5%).

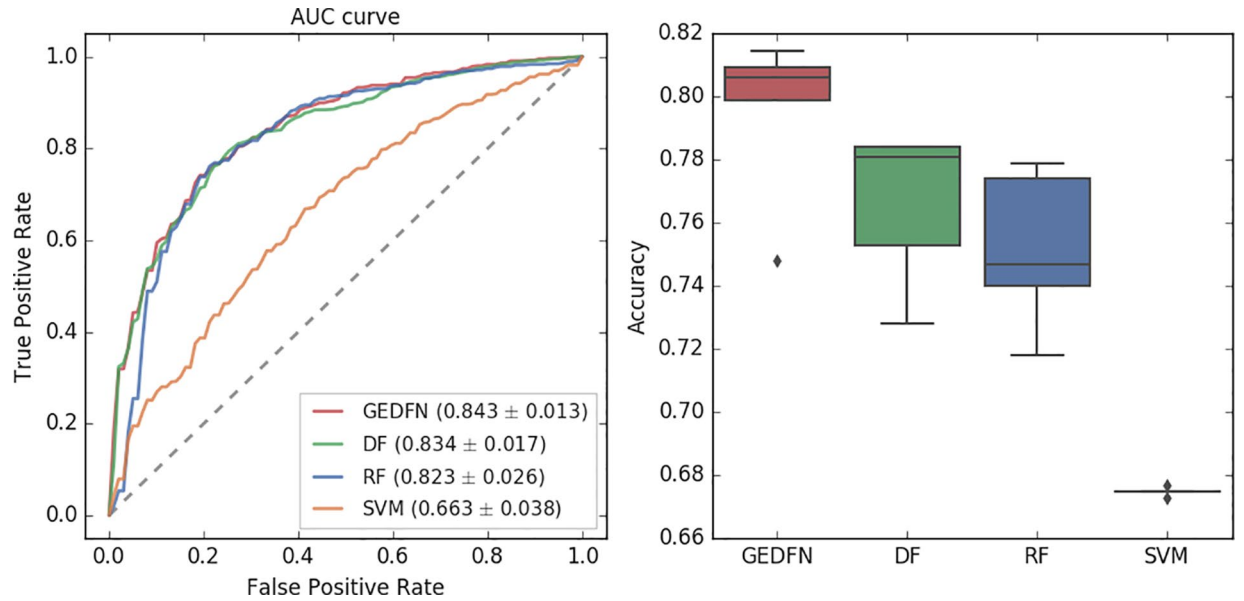


FIGURE 3 | The Area Under Receiver Operating Characteristic curve (left) and accuracy of classification (right) for GEDFN, Deep Forest (DF), Random Forest (RF) and Support Vector Machines (SVM). Left: the grey dash line is the chance discrimination that located on diagonal line (AUC = 0.5). The maximum AUC = 1 means the classifier could discriminate the diseased and non-diseased perfectly while AUC = 0 means the classifier incorrectly classified all subjects with diseased as negative and all subjects with non-diseased as positive. The AUC is averaged through a five-fold cross validation. Right: the boxplot for classifiers' classification accuracy.

The Evaluation of Feature Selection

In our experiment, we compare GEDFN with traditional feature selection methods, such as minimum redundancy and maximum Relevance (mRMR) (Ding and Peng, 2005), Random Forest and Deep Forest respectively. Each method selects 50 features. We want to know if the features obtained by the traditional machine learning feature selection method can also be selected by GEDFN. As can be seen from the Venn diagram (**Figure 4**), most of the features selected by the mRMR are different from those selected by the other three methods. Among these 50 features selected by

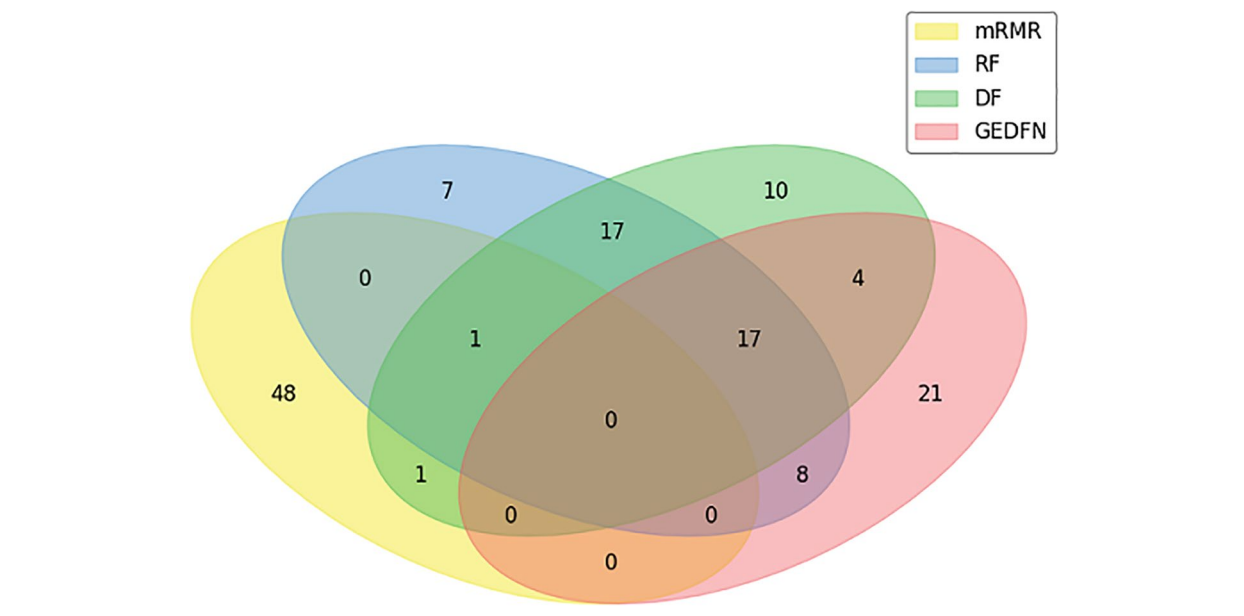


FIGURE 4 | The feature selection based on Graph Embedding Deep Feedforward Network (GEDFN). The Venn diagram for top the 50 features selected *via* minimum Redundancy and Maximum Relevance (mRMR), Random Forest (RF), Deep Forest (DF) and GEDFN.

GEDFN, there are 25 and 21 features which are consistent with the Random Forest and Deep Forest respectively.

In addition, we compare the performance of GEDFN + SVM, RF + SVM, RF + SVM and RF + DF. Our approach is to select top 10, top 15, top 20,…, top 50 feature subsets from GEDFN and RF respectively, and test them on SVM and Deep Forest (DF) classifiers with five-fold cross-validation (**Table 1**). GEDFN + SVM, means GEDFN is utilized to conduct feature selection and SVM is the classifier. RF + SVM, means RF is utilized to conduct feature selection and SVM is the classifier. GEDFN + DF, means GEDFN is utilized to conduct feature selection and DF is the classifier. RF + DF, means RF is utilized to conduct feature selection and DF is the classifier.

From **Table 1**, the combination of GEDFN and SVM achieves the best f1 score, while RF + SVM gets the worst performance. Meanwhile, GEDFN + SVM and GEDFN + DF have consistent performance. We find GEDFN prefers the sparse features while RF prefers the dense features. In other words, RF has a bias in the feature selection process where multivalued features are favored (Nguyen et al., 2015). In addition, RF is biased in the presence of correlation and often identifies non-predictive features that are independent from each other (Nicodemus and Malley, 2009). Actually, the microbial data is sparse and the features are dependent, which makes RF not the best choice to conduct feature selection in microbiome. However, GEDFN is to embed the *priori* sparse correlation network and find biomarkers as a whole, which makes it more suitable for microbiome-wide association studies than RF-based models.

The cophenetic similarity or cophenetic distance of two objects is a measure of how similar those two objects have to be in order to be grouped into the same cluster (Sokal and Rohlf, 1962; Saraçli et al., 2013). We calculate the cophenetic distance of the feature subsets. The specific process is as follows: we select different feature subsets obtained by Random Forest, Deep Forest and GEDFN, such as top 10–50 features, and then calculate node-node pairwise distance. The distance is characterized by the leaf nodes of the phylogenetic tree. We use the cophenetic method of the ape package in R to calculate the node-node pairwise cophenetic distance. The value in the matrix is the sum of the branch lengths separating each pair of species. We compare the top 50 features of Random Forest, Deep Forest and GEDFN respectively. We find the feature subsets of GEDFN has smallest cophenetic distances among these methods, which means that the subset of these features is better cohesive and we speculate that this cohesion may be functional meaningful (**Figure 5**). Deep Forest and Random Forest have similar cophenetic distance because Deep Forest is a cascade structure based on Random Forest.

In addition, we utilize interactive Tree Of Life (iTOL) (Letunic and Bork, 2016) to visualize the top 20 features selected by GEDFN (**Figure 6**). The features are ranked according to their importance score. We average each species' relative abundance for diseased and normal groups respectively. We find that *Neisseria*, *Pasteurellaceae*, *Bamesiellaceae*, *S24-7*, *Fusobacterium*, *Anaeroplasma* and *Gemellaceae* had high abundance compared to the normal group, while other microorganisms are lowly expressed in the disease group. The *Neisseria*, *Pasteurellaceae*, *Fusobacterium* and *Gemellaceae* increased in Crohn's disease, which was reported in the research (Gevers et al., 2014). The *Clostridiales*, *Eubacterium*, *Erysipelotrichaceae* and *Peptostreptococcaceae*, *Christensenellaceae* were found in lower relative abundance in Crohn's disease (Gevers et al., 2014; Matsuoka and Kanai, 2015; Pascal et al., 2017). However, there is no unified option on the Crohn's disease-related microbial biomarkers. As a result, our findings must need further experiments to explore and verify.

CONCLUSIONS

In this work, we propose a method of embedding a microbial graph into a Deep Feedforward Network to achieve feature selection purpose. We have verified the feasibility of this method through experiments. The main contributions of our work are as follows: Firstly, the feasibility of this method is verified through combining microbial interaction structure and deep learning, and a sparse network structure is proposed. Secondly, the feature selection method is introduced into the microbial sparse network and the reliability of the feature selection results is verified, indicating that deep neural networks can also conduct feature

TABLE 1 | The performance among GEDFN + SVM, RF + SVM, GEDFN + DF and RF + DF.

#	GEDFN + SVM			RF + SVM			GEDGN+DF			RF+DF		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
10	0.733	1	**0.846**	0.675	1	0.806	0.733	1	**0.846**	0.785	0.871	0.825
15	0.745	1	**0.854**	0.675	1	0.806	0.745	1	**0.854**	0.722	0.909	0.800
20	0.752	1	**0.858**	0.675	1	0.806	0.750	0.991	0.854	0.717	0.927	0.805
25	0.706	1	0.828	0.675	1	0.806	0.705	0.991	0.824	0.765	0.907	**0.829**
30	0.707	1	**0.828**	0.675	1	0.806	0.707	0.983	0.823	0.718	0.957	0.821
35	0.698	1	**0.822**	0.675	1	0.806	0.698	1	**0.822**	0.692	0.977	0.810
40	0.704	1	**0.826**	0.675	1	0.806	0.709	0.985	0.824	0.706	0.962	0.813
45	0.707	1	**0.828**	0.675	1	0.806	0.707	1	**0.828**	0.687	0.991	0.811
50	0.697	1	**0.822**	0.675	1	0.806	0.697	1	**0.822**	0.695	0.974	0.810

#, number of top features; P, precision; R, recall; F1= $\frac{2 \times P \times R}{P + R}$*. The best F1 scores are marked as bold.*

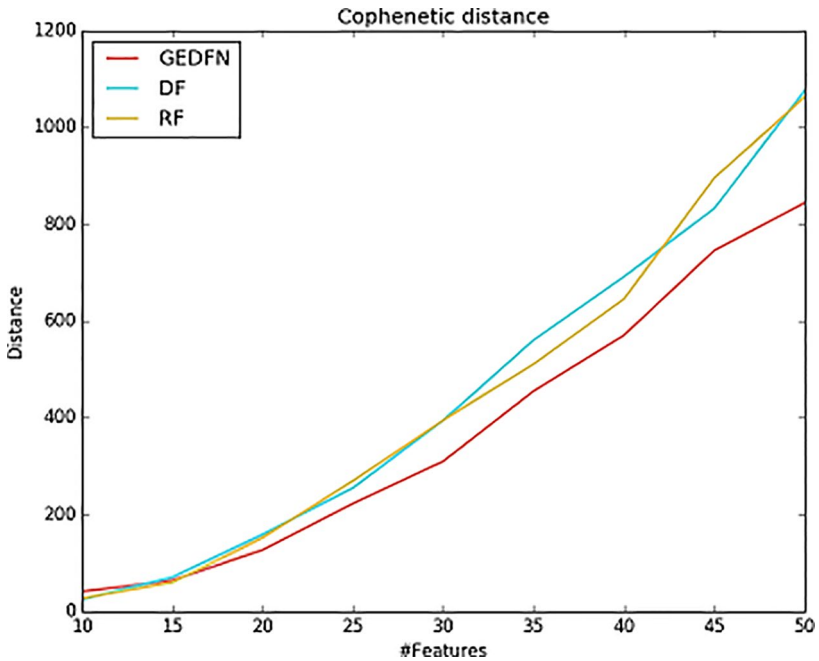


FIGURE 5 | The cophenetic distance for top 50 features selected *via* Random Forest (RF), Deep Forest (DF) and Graph Embedding Deep Feedforward Network (GEDFN) respectively (The cophenetic distance is the sum of the features' pair-wise distance.). The cophenetic distance of two objects is a measure of how similar those two objects have to be in order to be grouped into the same cluster.

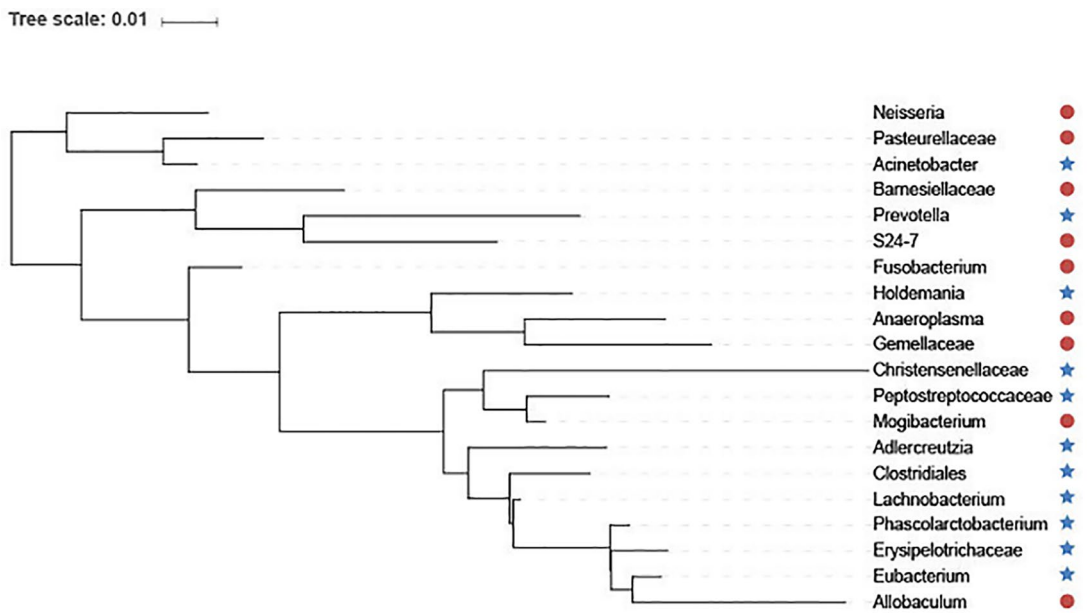


FIGURE 6 | The top 20 species selected *via* Graph Embedding Deep Feedforward Network (GEDFN). The species in red circle are higher relative abundance while species in blue star are lower relative abundance in diseased group. These species are visualized on the phylogenetic tree.

selection. We hope our work will bring another perspective to the interpretability of deep learning.

The problems still exist in the research work. First of all, our work does not compare the influence of various methods of constructing microbial networks on feature selection (Weiss et al., 2016). The networks constructed by various methods are varying. We found that the reliability of the microbial network directly affected the subsequent results. Secondly, the threshold of association network was traded off and there was no relevant guidance suggestion. In general, the higher the threshold, the

more reliable the network, but it would make the network too sparse. It would be required to balance the threshold and the network's sparseness. Finally, we only consider the influence of the weight parameters of the Deep Neural Network on the feature selection without considering the threshold of the neuron. Because it would involve the nonlinear transformation which could make the problem complicated and difficult. Therefore, our future work will focus on how to build a more reliable microbial interaction network and get more meaningful microbial markers.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

Qiang Z, XJ and TH conceived the concept of the work and designed the experiments. Qing Z and MP performed literature search. Qing Z, XJ, MP and TH collected and analyzed the data. Qiang Z, XJ and MP wrote the paper. All authors have approved the final manuscript.

FUNDING

This research is supported by the National Key Research and Development Program of China (2017YFC0909502) and the National Natural Science Foundation of China (No. 61532008 and 61872157).

REFERENCES

Albanese, D., Riccadonna, S., Donati, C., and Franceschi, P. (2018). A practical tool for maximal information coefficient analysis. *GigaScience* 7 (4), giy032. doi: 10.1093/gigascience/giy032

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12 (7), 878. doi: 10.15252/msb.20156651

Bray, J. R., and Curtis, J. T. (1957). An ordination of the upland forest communities of southern wisconsin. *Ecol. Monographs* 27 (4), 325–349. doi: 10.2307/1942268

Brüls, T., and Weissenbach, J. (2011). The human metagenome: our other genome. *Hum. Mol. Genet.* 20 (R2), R142–R148. doi: 10.1093/hmg/ddr353

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-generation machine learning for biological networks. *Cell* 173 (7), 1581–1592. doi: 10.1016/j.cell.2018.05.015

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. Royal Soc. Inter.* 15 (141), 20170387. doi: 10.1098/rsif.2017.0387

Deng, Y., Jiang, Y. H., Yang, Y., He, Z., Luo, F., and Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics* 13 (1), 113. doi: 10.1186/1471-2105-13-113

Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3 (02), 185–205. doi: 10.1142/S0219720005001004

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20 (7), 389–403. doi: 10.1038/s41576-019-0122-6

Faust, K., and Raes, J. (2016). CoNet app: inference of biological association networks using cytoscape. *F1000 Research* 5, 1519. doi: 10.12688/f1000research.9050.2 F1000.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8 (7), e1002606. doi: 10.1371/journal.pcbi.1002606

Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8 (9), e1002687. doi: 10.1371/journal.pcbi.1002687

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Treuren, W. V., Ren, B., et al. (2014). The treatment-naive microbiome in new-onset crohn's disease. *Cell Host Microbe* 15 (3), 382–392. doi: 10.1016/j.chom.2014.02.005

Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., et al. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* 535 (7610), 94. doi: 10.1038/nature18850

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8, 2224. doi: 10.3389/fmicb.2017.02224

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge: MIT press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A Survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51 (5), 93. doi: 10.1145/3236009

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2017). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinform.* 20 (1), 210–221. doi: 10.1093/bib/bbx104

Kho, Z. Y., and Lal, Sunil K. (2018). The human gut microbiome-a potential controller of wellness and disease. *Front. Microbiol.* 9, 1835. doi: 10.3389/fmicb.2018.01835

Kong, Y., and Yu, T. (2018). A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics* 34 (21), 3727–3737. doi: 10.1093/bioinformatics/bty429

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological Networks. *PLoS Comput. Biol.* 11 (5), e1004226. doi: 10.1371/journal.pcbi.1004226

Layeghifard, M., Hwang, D. M., and Guttman, D. S. (2017). Disentangling interactions in the microbiome: a network perspective. *Trends Microbiol.* 25 (3), 217–228. doi: 10.1016/j.tim.2016.11.008

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436. doi: 10.1038/nature14539

Letunic, I., and Bork, P. (2016). Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* 44 (W1), W242–W245. doi: 10.1093/nar/gkw290

Li, J., Cheng, K., Wang, S., F Morstatter, R. P. T., Tang, J., and Liu, H. (2017). Feature selection: a data perspective. *ACM Comput. Surveys* 50 (6), 1–45. doi: 10.1145/3136625

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., et al. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature* 550 (7674), 61. doi: 10.1038/nature23889

Matsuoka, K., and Kanai, T. (2015). "The Gut Microbiota and Inflammatory Bowel Disease," in *Seminars in immunopathology*, vol. 37. (Verlag GmbH Germany: Springer), 47–55. doi: 10.1007/s00281-014-0454-4

Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24 (3), 69–71.

Nguyen, T. T., Huang, J. Z., and Nguyen, T. T. (2015). Unbiased feature selection in learning random forests for high-dimensional data. *Sci. World J.* 471371. doi: 10.1155/2015/471371 2015.

Nicodemus, K. K., and Malley, J. D. (2009). Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* 25 (15), 1884–1890. doi: 10.1093/bioinformatics/btp331

Olden, J. D., and Jackson, D. A. (2002). Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* 154 (1–2), 135–150. doi: 10.1016/S0304-3800(02)00064-9

Pascal, V., Pozuelo, M., Borruel, N., Casellas, F., Campos, D., Santiago, A., et al. (2017). A microbial signature for crohn's disease. *Gut* 66 (5), 813–822. doi: 10.1136/gutjnl-2016-313235

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12 (7), e1004977. doi: 10.1371/journal.pcbi.1004977

Paulson, J. N., Stine, O. C., Bravo, H. C., and Pop, M. (2013). Differential abundance Analysis for microbial marker-gene surveys. *Nat. Methods* 10 (12), 1200. doi: 10.1038/nmeth.2658

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016). Analysis of the microbiome: advantages of whole genome shotgun versus 16s amplicon sequencing. *Biochem. Biophys. Res. Commun.* 469 (4), 967–977. doi: 10.1016/j.bbrc.2015.12.083

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting novel associations in large data sets. *Science* 334 (6062), 1518–1524. doi: 10.1126/science.1205438

Saraçli, S., Doğan, N., and Doğan, İ (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J. Inequal. Appl.* 2013 (1), 203. doi: 10.1186/1029-242X-2013-203

Sokal, R. R., and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon.* 11 (2), 33–40. doi: 10.2307/1217208

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature* 449 (7164), 804. doi: 10.1038/nature06244

Wang, J., and Jia, H. (2016). Metagenome-wide association studies: fine-mining the microbiome. *Nat. Rev. Microbiol.* 14 (8), 508. doi: 10.1038/nrmicro.2016.83

Weiss, S., Treuren, W. V., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10 (7), 1669. doi: 10.1038/ismej.2015.235

Weiss, S., Xu, Z Zech, Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5 (1), 27. doi: 10.1186/s40168-017-0237-y

Zhu, Q., Pan, M., Liu, L., Li, B., He, T., Jiang, X., et al. (2018). "An ensemble feature selection method based on deep forest for microbiome-wide association studies," in *2018 IEEE international conference on Bioinformatics and Biomedicine (BIBM)*, vol. 248–253. (Washington, D.C.: IEEE Computer Society), 248–253. doi: 10.1109/BIBM.2018.8621461

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

CORRECTION
published: 15 May 2020
doi: 10.3389/fgene.2020.00487

Corrigendum: Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification

Qiang Zhu[1,2], *Xingpeng Jiang*[2,3]*, *Qing Zhu*[2,3], *Min Pan*[2,3] *and Tingting He*[2,3]

[1] School of Information Management, Central China Normal University, Wuhan, China, [2] School of Computer, Central China Normal University, Wuhan, China, [3] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China

Keywords: graph embedding, deep learning, feature selection, biomarkers, microbiome

A Corrigendum on

Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification
by Zhu, Q., Jiang, X., Zhu, Q., Pan, M., and He, T. (2019). Front. Genet. 10:1182. doi: 10.3389/fgene.2019.01182

Although in the original article although we have cited the work (Kong and Yu, 2018) in the Introduction section, we did not cite the work in the Materials and Methods section. Our approach to embedding deep learning for identifying microbial biomarkers is based on their methods and thus contributed a lot to our article. Therefore, this citation has been added to the following sections.

In order to avoid misinterpretation, we would like to add the reference in the following places which were highlighted in RED:

The **Materials and Methods** section, subsection **The Framework of Graph Embedding Deep Feedforward Network**, sub-subsection **Graph Embedding Deep Feedforward Network**, paragraph 4:

"Where $\odot$ is element-wise product. Therefore, the connection between the input and first hidden layer of the feedforward network is filtered by the graph adjacency matrix. Each feature is corresponding to a hidden neuron. All features have corresponding hidden neurons in the first hidden layer. The feature can only provide information to the connected graph. In this way, the graph helps to achieve the sparsity of the connection between the input layer and the first hidden layer (Kong and Yu, 2018)."

The **Materials and Methods** section, subsection **Feature Selection Based on GEDFN**, paragraph 1:

"In addition to improving classification, it is also meaningful to find features that contribute significantly to classification because they reveal potential biological mechanisms. However, Deep neural network is a "black box", the interpretability of deep learning hasn't been well-defined (Guidotti et al., 2019). In our experiment, we focus on how the input features influence the prediction and we borrow the idea from Olden and Jackson (2002) and Kong and Yu (2018). The feature importance score is the quantification values of the contributions of features to a model prediction, which links the input features and output prediction. They highlight the parts of a given input that are most influential for the model prediction and thereby help to explain why such a prediction was made. The feature selection is based on feature score, which means the score is high if the feature is important. As a result, we develop a feature ranking method based on the feature relative importance score, similar to the connection weights method introduced by Olden and Jackson (2002) and Kong and Yu (2018). What is learned by neural networks is

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Xingpeng Jiang
xpjiang@mail.ccnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *27 March 2020*
Accepted: *20 April 2020*
Published: *15 May 2020*

Citation:
Zhu Q, Jiang X, Zhu Q, Pan M and He T (2020) Corrigendum: Graph Embedding Deep Learning Guides Microbial Biomarkers' Identification. Front. Genet. 11:487. doi: 10.3389/fgene.2020.00487

contained in the connection weights. Based on idea of connection weight, we propose a graphical connect weight method that emphasizes the importance of the features of our proposed neural network architecture."

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

REFERENCES

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2019). A Survey of methods for explaining black box models. *ACM Comput. Surv.* 51:93. doi: 10.1145/3236009

Kong, Y., and Yu, T. (2018). A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. *Bioinformatics* 34, 3727–3737. doi: 10.1093/bioinformatics/bty429

Olden, J. D., and Jackson, D. A. (2002). Illuminating the 'black box': a randomization approach for understanding variable contributions in artificial neural networks. *Ecol. Model.* 154, 135–150. doi: 10.1016/S0304-3800(02)00064-9

frontiers in Genetics

METHODS
published: 26 November 2019
doi: 10.3389/fgene.2019.01236

IBI: Identification of Biomarker Genes in Individual Tumor Samples

Jie Li, Dong Wang and Yadong Wang*

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Individual patient biomarkers have an important role in personalized treatment. Although various high-throughput sequencing technologies are widely used in biological experiments, these are usually conducted only once or a few times for each patient, which makes it a challenging problem to identify biomarkers in individual patients. At present, there is a lack of effective methods to identify biomarkers in individual sample data. Here, we propose a novel method, IBI, to identify biomarkers in individual tumor samples. Experimental results from several tumor data sets showed that the proposed method could effectively find biomarker genes for individual patients, including common biomarkers related to the mechanisms of the development of cancer, which can be used to predict survival and drug response in patients. In summary, these results demonstrate that the proposed method offers a new perspective for analyzing individual samples.

Keywords: biomarker, individual sample, tumor, regression model, gene expression data

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Weijia Zhang,
Icahn School of Medicine at Mount Sinai, United States
Jingyang Gao,
Beijing University of Chemical Technology, China

****Correspondence:***
Jie Li
jieli@hit.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *17 October 2019*
Accepted: *07 November 2019*
Published: *26 November 2019*

Citation:
Li J, Wang D and Wang Y (2019) IBI: Identification of Biomarker Genes in Individual Tumor Samples. Front. Genet. 10:1236. doi: 10.3389/fgene.2019.01236

INTRODUCTION

Biomarker discovery is critical for cancer diagnostics, prognosis, and monitoring of therapy in clinical trials. With the development of high-throughput biochip technologies such as next-generation sequencing, massive quantities of cancer genomic data are being generated in the healthcare field, which offers an opportunity to identify high-quality cancer biomarkers for use in personalized medicine. Therefore, various computational methods have been proposed to identify cancer biomarkers. At present, the most commonly used methods are statistical tests, such as t-test, KS-test, and Wilcoxon's rank sum test (Li et al., 2007; Dembélé and Kastner, 2014; Love et al., 2014; Moore et al., 2016; Wang et al., 2018), which identify differentially expressed genes (DEGs) from two types of samples and choose the group of genes with the lower p-value as potential biomarkers. However, the method often ignores and misses information between genes (Lewis-Wambi et al., 2008). Machine learning algorithms and statistical models also are widely used to identify cancer biomarkers. For example, the 70-gene biomarkers (Van't Veer et al., 2002), wound-response gene biomarkers (Chang et al., 2005), and several of our gene biomarkers (Li et al., 2008; Li et al., 2010; Zhang et al., 2017) are all identified using machine learning algorithms. The 21-gene biomarkers (Van't Veer and Bernards, 2008) and immunotherapy response biomarkers (Ock et al., 2017; Jiang et al., 2018) are based on statistical models.

However, the above methods are only able to identify biomarkers in two groups of samples, not in an individual sample. As cancer is a complex and heterogeneous disease, different patients have differences in pathogenesis and need different treatments. Thus, there is a need for biomarkers for individual patients that reflect their status. Currently, high-throughput biological experiments are usually conducted once or a few times for a single patient, which makes it a challenging problem to analyze single samples and, in particular, to identify biomarkers in individual patients. Some algorithms have been developed to analyze single samples. Rezwan et al. (2015) used the

Crawford-Howell t-test to analyze methylation data of single samples and identified hypomethylation at different sites. However, this method could only detect differences in a single molecular element among different samples and may ignore the relationships of different molecular elements in the same sample. Liu et al. (2017) proposed the sDNB (single-sample dynamic network biomarkers) method to detect early-warning signals or critical states in individual patients using gene expression data. sDNB detects changes in gene expression levels of a pair of genes relative to reference samples and considers the local information of a gene in network. Drier et al. (2013) proposed an algorithm to analyze single tumor samples using pathway-level information instead of gene-level information. Pathways were detected that were significantly associated with survival of glioblastoma and colorectal cancer patients. However, a set of genes in the same pathway have similar functions; this means that models based on redundant features (biomarkers) are usually more complex.

Here, we propose a novel method, IBI (identification of biomarker genes in individual tumor samples), to identify biomarker genes in individual tumor samples using gene expression data. An overview of the IBI method is given in **Figure 1**. First, DEGs in tumor and normal samples are identified. Then, regression models are constructed using the selected DEGs, and residuals of each gene in different samples are analyzed using the kernel density estimation (KDE). Finally, we assess the degree of change of each gene according to the credibility interval (CI) of its residuals to decide which genes are biomarkers of the individual sample.

MATERIALS AND METHODS

Data Collection and Preprocessing

The proposed method was used to analyze three gene expression data sets: TCGA-BRCA (Tomczak et al., 2015), GSE63557 (Lesterhuis et al., 2015), and GSE35640 (Ulloa-Montoya et al., 2013). TCGA-BRCA consists of 1,090 breast cancer samples and 113 normal tissue samples. GSE63557 contains AB1-HA tumor data from mice during immunotherapy with 10 anti-CTLA-4 immunotherapeutic response samples and 10 non-response samples, and GSE35640 consists of advanced melanoma data with 22 MAGE–A3 immunotherapeutic response and 34

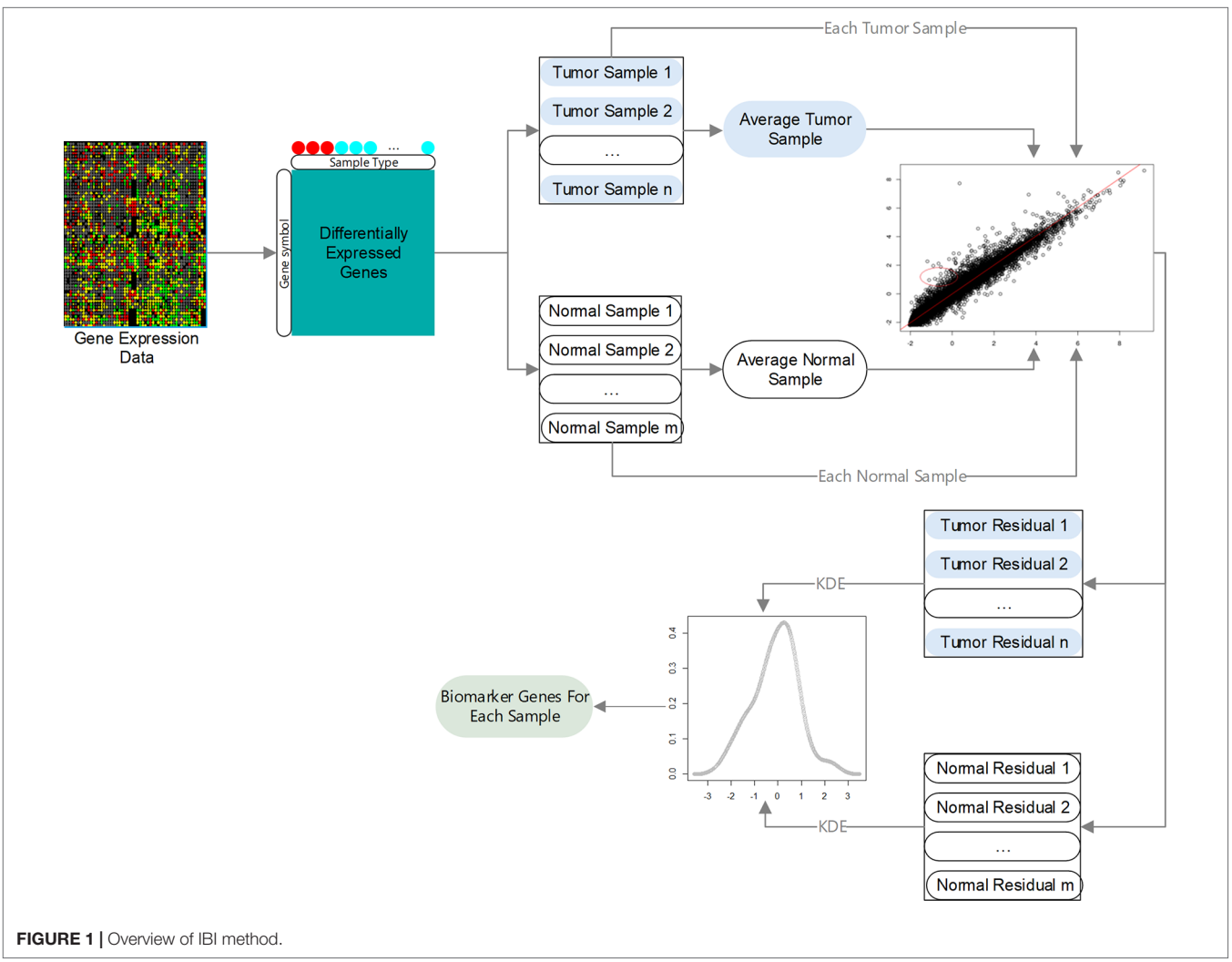


FIGURE 1 | Overview of IBI method.

non-response samples. The first data set contains RNA-seq data, which was preprocessed using DESeq2 (Love et al., 2014), and the latter two data sets were preprocessed using the z-score.

Identification of Differentially Expression Genes

Assuming we have gene expression data with two types of samples and genes, let each sample be labeled with either "+" or "−"; n_1 and n_2 are the number of samples with label "+" and "−", respectively ($n = n_1 + n_2$). y_{ji} is the expression value of the jth gene of the ith sample with label "+", and x_{ji} is the expression value of the jth gene of the ith sample with label "−". q DEGs are obtained using the robust algorithm (Love et al., 2014) or GEO2R (Smyth, 2004).

Average Sample

Let average samples with label "+" and "−" be $u^+ = \left[u_1^+, u_2^+ \ldots u_q^+\right]$ *and* $u^- = \left[u_1^-, u_2^- \ldots u_q^-\right]$, respectively.

$$u_j^+ = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{ji}, \quad q \geq j \geq 1 \tag{1}$$

$$u_j^- = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{ji}, \quad q \geq j \geq 1 \tag{2}$$

Regression Model Based on Average and Single Samples

Let y_{ji} be the expression value of the jth DEG of the ith sample with label "+" and $x_{ji}^{'}$ the expression value of the jth DEG of the ith sample with label "−." For the ith sample with label "+," $S_i^+ = \left[y'_{1i}, y'_{2i} \ldots y'_{qi},\right]$, y'_{ji} can be predicted using the following regression model according to u_j^+:

$$\widehat{y'_{ji}} = \beta_0^+ + \beta_1^+ u_j^+, \quad q \geq j \geq 1 \tag{3}$$

where β_0^+ and β_1^+ are the regression coefficients estimated according to a set of data $\left(y_{1i}, u_1^+\right), \left(y_{2i}, u_2^+\right), \ldots, \left(y_{qi}, u_2^+\right)$, using the least squares method.

Similarity, for the ith sample with label "−", $S_i^- = \left[x_{1i}^{'}, x_{2i}^{'} \ldots x_{qi}^{'}\right]$, $x_{ji}^{'}$ can be predicted using the following regression model according to u_j^-:

$$\widehat{x_{ji}^{'}} = \beta_0^- + \beta_1^- u_j^-, \quad q \geq j \geq 1 \tag{4}$$

where β_0^- and β_1^- are the regression coefficients estimated according to a set of data $\left(x_{1i}, u_1^-\right), \left(x_{2i}, u_2^-\right), \ldots, \left(x_{qi}, u_q^-\right)$ using the least squares method.

Algorithm for Identifying Biomarker Genes of a Single Sample

Among q DEGs, expression values of some genes of a single sample may undergo very significant changes compared with their average values, i.e., the observed values of these genes are far from regression line. These genes are called biomarker genes of the single sample. The degree of the significant difference can be calculated using the residual value between the predicted value and observed value.

For the ith sample with label "+," the residual value of its the jth DEG is:

$$e_{ji}^+ = y'_{ji} - \widehat{y'_{ji}}, \quad q \geq j \geq 1 \tag{5}$$

Similarity, for the ith sample with label "−", the residual value of it's the jth DEG is:

$$e_{ji}^- = x_{ji}^{'} - \widehat{x_{ji}^{'}}, \quad q \geq j \geq 1 \tag{6}$$

To obtain biomarker genes of the ith sample with label "+", the KDE is introduced to estimate the probability density function $\widehat{f_i}\left(e_i\right)$ of residual values: $\left(e_{1i}^+, e_{2i}^+, ..., e_{qi}^+\right)$. Its kernel density estimator with Gaussian kernel K is as follows:

$$\widehat{f_i}\left(e_i\right) = \frac{1}{qh}\sum_{j=1}^{q} K\left(\frac{e_i - e_{ji}^+}{h}\right) \tag{7}$$

$$K\left(x\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{8}$$

where h is a smoothing parameter called the bandwidth (h > 0). Let Φ be the cumulative distribution function of the kernel density estimator; then, the CI at confidence level α is

$$CI_\alpha = \left(0, \Phi\left(\frac{\alpha}{2}\right)\right) \bigcup \left(\Phi\left(1 - \frac{\alpha}{2}\right), 1\right) \tag{9}$$

The jth gene is considered a biomarker gene of the ith sample with label "+" ($n_1 \geq i \geq 1$) if $\Phi\left(e_{ji}^+\right) \in CI_\alpha$. Similarity, we can obtain the biomarker gene of the ith sample with label "−"($n_2 \geq i \geq 1$).

RESULTS

Performance Evaluation

It was somewhat difficult to directly evaluate the performance of the proposed method. Three methods were employed to evaluate the power of the method.

1) Statistical test: The biomarker genes of each sample should be specific, that is, their expression values in the sample should be significantly different from those of other samples. We

designed a method to test such differences, as follows. First, biomarker genes of sample S_i are selected and their expression values extracted from all samples. Then, the expression values of each biomarker gene in different samples are sorted respectively and used to construct a rank matrix. The ith row vector, R_i, of the matrix denotes orders of biomarker genes of S_i. Finally, the Kolmogorov-Smirnov test is performed to determine whether there is a significant difference between R_i and R_j ($j \neq i$).

2) Survival analysis: The biomarker genes of each tumor sample should reflect its characteristics, namely, it should be possible to use biomarker genes to classify tumor samples into high- and low-risk groups and predict the survival risk of tumor patients.
3) Validation *via* biological evidence: The biomarker genes of each tumor sample should reflect the pathogenesis of cancer, that is, they should have been reported to be associated with tumor development in the published literature.

Experimental Results for TCGA-BRCA

The experiments on TCGA-BRCA were performed as follows. First, 6120 DEGs in two groups of samples were identified using DESeq2 (Love et al., 2014) at a 95% confidence level and absolute value of log fold change > 1. Next, average tumor and normal samples based on 6120 DEGs were obtained using Equations. (1) and (2). Then, 1,090 (113) regression models were constructed based on average tumor (normal) samples and 1,090 tumor (113 normal) samples, respectively; an example is shown in **Figure 2**. The residuals of the genes of each sample were calculated using Equations (5) and (6); **Figure 3** shows residual values of biomarker genes from two samples. Finally, biomarker genes for each sample were identified using Equations (7), (8), and (9). The distribution of the number of biomarker genes in the 1,090 (113) tumor (normal) samples is shown in **Figure 4**.

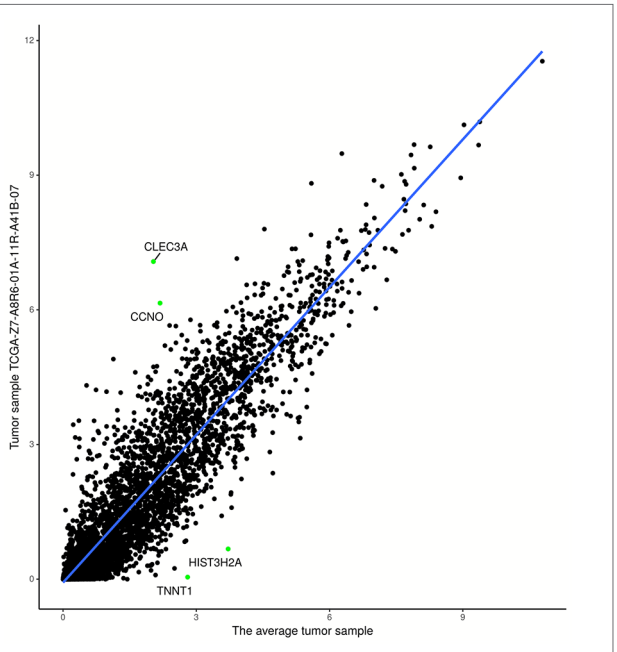


FIGURE 2 | Regression model based on tumor sample TCGA-Z7-A8R6-01A-11R-A41B-07 and average tumor sample. The points in the upper-left (lower-right) partition are two biomarker genes with the highest (lowest) expression levels.

As shown clearly in **Figures 2** and **3**, genes were distributed in two main areas. The genes scattered in the upper-left of the plots are those with higher expression levels, whereas genes in the lower-right portion have lower expression values, in the single tumor/normal sample. In **Figure 2**, there are several spots that are distant from the regression lines. These spots represent biomarker genes of the single sample. **Figure 3** shows more clearly which genes had very significant variation in expression. For example, the residuals of *CLEC3A* and *CCNO* were 4.92 and 3.83, respectively, significantly higher than the values for other genes; while the residuals of *HIST3H2A* and *TNNT1* were −3.33 and −2.95, respectively, significantly lower than those of other genes.

It can also be seen from **Figure 4** that the number of biomarker genes varied among different samples. Some tumor samples had more than 315 biomarker genes, while others had about 290. The mean numbers of biomarker genes in the tumor samples and normal samples were 304.9 and 305, respectively. In addition, the biomarker genes of different samples were also different. In 1090 tumor samples and 113 normal samples, the biomarker genes had different frequencies (a biomarker gene has higher frequency if it is found in more samples). The top 15 biomarker genes with significantly different frequencies in tumor and normal samples are listed in **Supplementary Table 1**. These genes were common biomarkers of most tumor samples, and they had higher frequency in tumor samples than in normal samples. Therefore, these genes were likely to be related to the development of breast cancer. To test our hypothesis, we searched the literature using public databases and found that 14 of the 15 genes were indeed related to the development of breast cancer. The top gene was *S100A7*, which has been found to be expressed in several tissues including breast adenocarcinomas and squamous carcinomas of the head and neck, the cervix, and the lung (Emberley et al., 2004); *S100A7* is also related survival of breast cancer patients (Emberley, 2003). *CLEC3A* had the highest frequency in tumor samples; its overexpression promotes tumor progression and poor prognosis in breast invasive ductal cancer (IDC) and is related to higher lymph node and poorer overall survival (OS) of breast IDC (Ni et al., 2018). *PRAME* has a tumor-promoting role in triple-negative breast cancer, increasing cancer cell motility through the epithelial-to-mesenchymal transition (EMT) gene reprogramming. Therefore, *PRAME* could serve as a prognostic biomarker and/or therapeutic target in triple-negative breast cancer (Al-Khadairi et al., 2019). Kammerer et al. (2016) suggested that patients with estrogen receptor-positive breast cancer might be stratified into high- and low-risk groups based on the *KCNJ3* levels in the tumor. *CST1* was found to be generally upregulated in breast cancer at both the mRNA and the protein level. Furthermore, OS and disease-free survival in the low *CST1* expression subgroup were significantly superior to those in the high *CST1* expression subgroup, indicating that *CST1* could be a prognostic indicator and a potential therapeutic target for breast cancer (Dai et al., 2017). Xuan et al. (2015) reported that higher expression of *MMP1* in

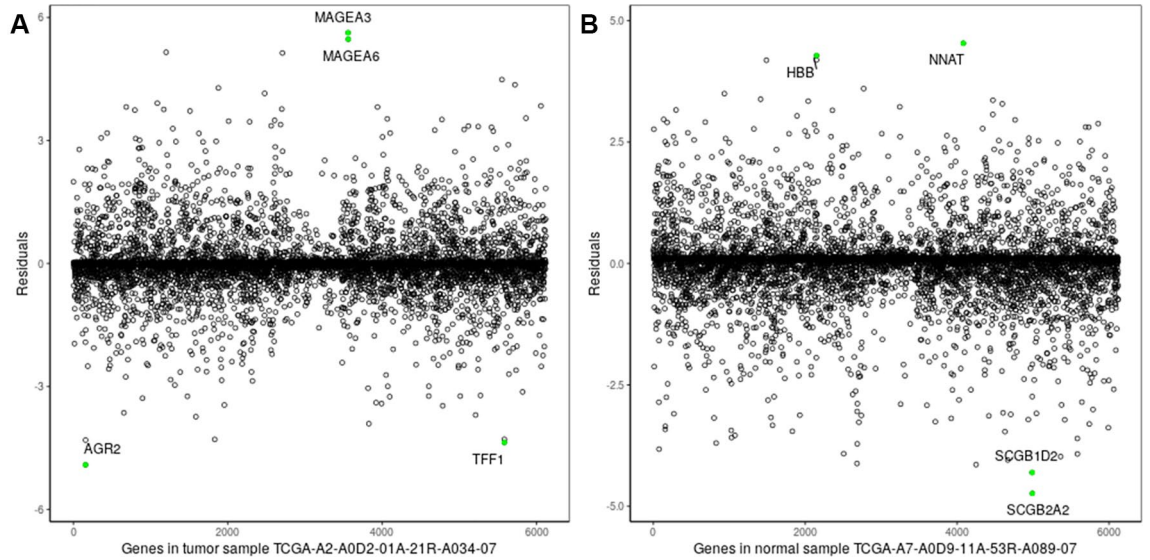


FIGURE 3 | Residuals of genes of a single sample. **(A)** Breast tumor sample TCGA-A2-A0D2-01A-21R-A034-07; **(B)** normal tissue sample: TCGA-A7-A0D9-11A-53R-A089-07. The green points denote the two biomarker genes with the highest/lowest expression levels in the two samples.

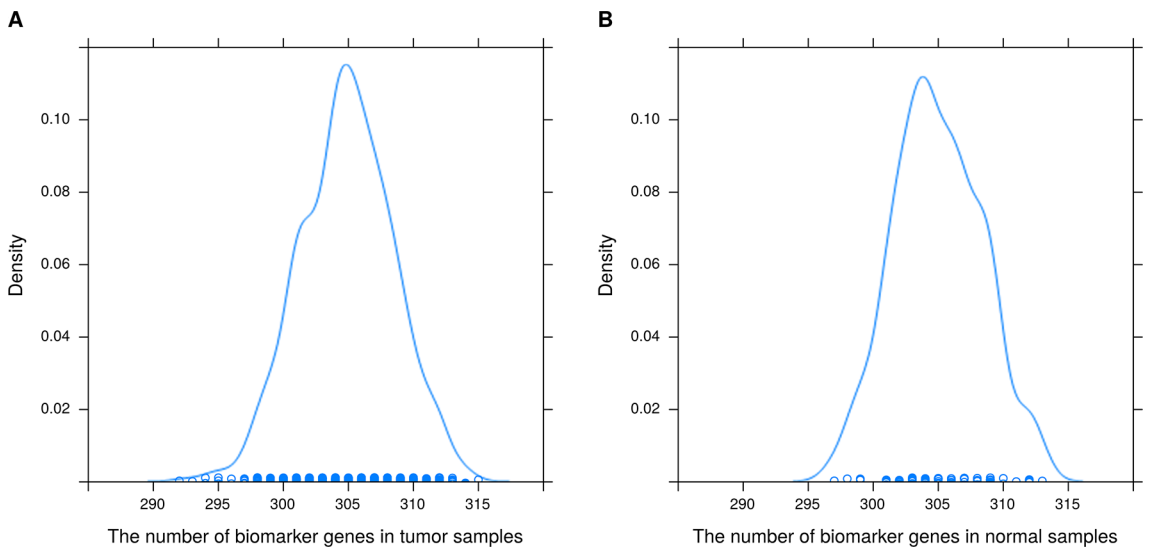


FIGURE 4 | Distribution of the number of biomarker genes in **(A)** 1090 breast tumor samples and **(B)** 113 normal tissue samples.

breast cancer might play a crucial part in promoting breast cancer metastasis. Powell et al. (2018) demonstrated that *CEACAM5* was a clinically relevant driver of breast cancer metastasis. *NKAIN1* is associated with OS in breast cancer (Su et al., 2019). *DSCAM-AS1* promotes tumor growth in breast cancer by reducing miR-204-5p and upregulating *RRM2* (Liang et al., 2019). Overexpression of *CEACAM6* promotes migration and invasion of estrogen-deprived breast cancer cells (Lewis-Wambi et al., 2008). Bhakta et al. (2018) suggested that anti-GFRA1-vcMMAE ADC might provide a targeted therapeutic opportunity for luminal A breast cancer patients. *BMPR1B* is related to proliferation of breast cancer cells (Bokobza et al., 2009). Jia et al. (2016) identified *COL11A1* as a highly specific biomarker of activated cancer-associated fibroblasts (CAFs), which could promote breast cancer and inhibit pancreatic cancer. In summary, 14 of the top 15 biomarker genes have been reported to be associated with breast cancer. Therefore, these results demonstrate that the proposed method can effectively identify biomarkers related to cancer.

Statistical tests were performed to evaluate whether expression levels of biomarker genes of a sample were significantly different compared with those of other samples. As the biomarker gene set of each sample was represented by a p-value vector with dimension n, 1,090*1,089 [n(n−1)], where n is the number of samples) p-values were obtained for the 1090 tumor samples, and

113*112 p-values for the 113 normal samples; 1,186,999 (99.99%) and 12,626 (99.76%) of these p-values were less than 0.05 for the tumor samples and normal samples, respectively. These results indicate that there were significant differences between the expression levels of the identified biomarker genes of a sample and those of other samples, that is, the proposed method can effectively identify the biomarker genes of a single sample.

The frequencies of biomarker genes in tumor and normal samples were different. Here, we mainly analyzed biomarker genes whose frequency was higher in tumor samples than in normal samples, to explore which genes might have important roles in survival prediction and development of breast cancer. We selected 305 biomarker genes with higher frequency in tumor samples, and clustered the tumor samples into two groups using the multiple survival screening (MSS) algorithm (Li et al., 2010). Survival was significantly different between the two groups (p-value = 0.0089) (**Figure 5**). This means these biomarker genes are important features of breast cancer and can be used to distinguish tumor patients into high- and low-risk groups (here, we removed two samples with the negative follow-up-time, so there were 1,088 samples participating in survival analysis).

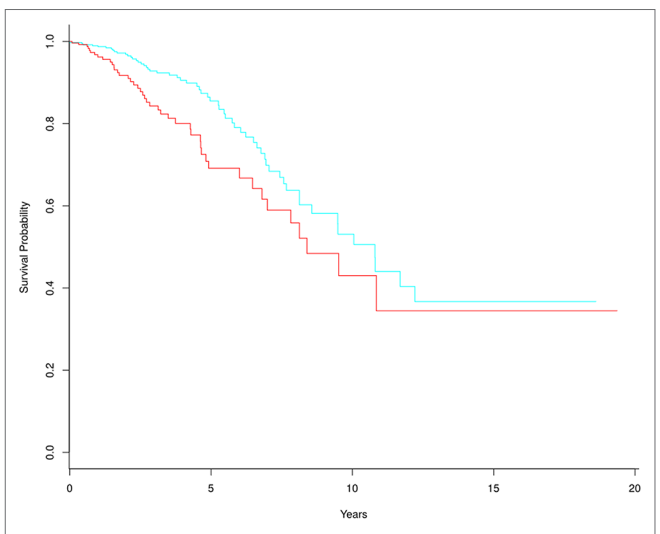


FIGURE 5 | Kaplan-Meier survival curves based on 305 tumor biomarker genes. In the high-risk group (red line), there are 329 tumor samples. In the low-risk group (blue line), there are 759 tumor samples.

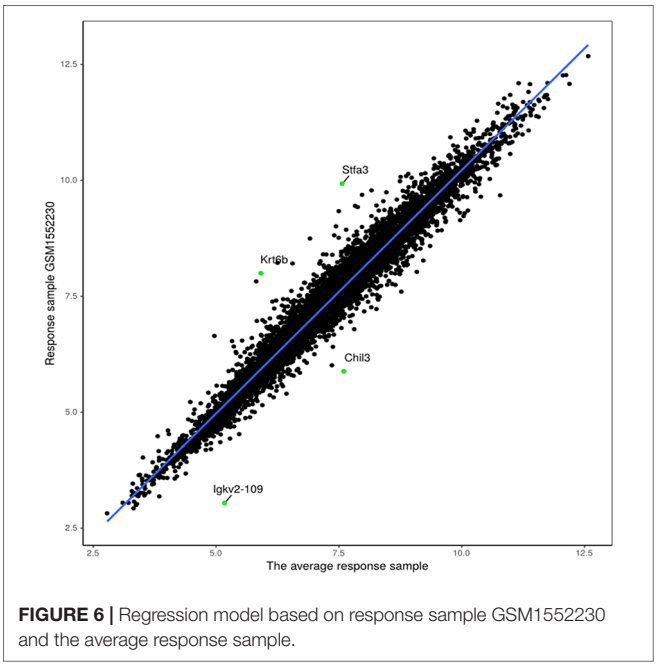


FIGURE 6 | Regression model based on response sample GSM1552230 and the average response sample.

Experimental Results for Immunotherapeutic Response Samples

The proposed method was also used to analyze mouse AB1-HA tumor data: GSE63557. A total of 8,042 DEGs in two groups of samples were identified using GEO2R (Smyth, 2004) at a 95% confidence level. Regression models of 10 anti-CTLA-4 immunotherapeutic response samples and 10 non-response samples were constructed; one of these is shown in **Figure 6**. **Figure 7** shows residual values of biomarker genes from two samples. The number of biomarker genes of 10 response samples and 10 non-response samples is shown in **Figure 8**. In **Figures 6** and **7**, there are several genes that are far from the regression lines. For example, the residuals of *Krt6b* and *Stfa3* were 2.07 and 2.26, respectively, significantly higher than those of other genes; the residuals of *Chil3* and *Igkv2-109* were −1.82 and −2.10, respectively, significantly lower than those of other genes.

The number of biomarker genes of different samples is shown in **Figure 8**, illustrating the variation between samples. The biomarker genes from different samples were also different. For 10 response samples and 10 non-response samples, the top 15 genes with the most significant differences in frequency are shown in **Supplementary Table 2**. Four of these genes, *Gzme*, *CD38*, *CD3D*, and *Chil3*, appeared in the important cancer modules identified by Lesterhuis et al. (2015) However, the top gene, *Jchain*, had not been identified as a member of these important cancer modules; notably, *Jchain* was also found to be the most important of the anti-CTLA-4 immunotherapeutic response biomarker genes in our study, with frequencies in response and non-response samples of 80% and 0%, respectively. This suggests that *Jchain* is related to immunotherapeutic response. GeneCards (https://www.genecards.org/) indeed confirms that *Jchain* has an important role in immune response. Moreover, *Iglj1*, *Cd38*, and *Cd3d* are also immune response related. This demonstrates that the IBI method can detect important genes contributing to the immunotherapeutic response mechanism.

According to the statistical tests, 100% of p-values were less than 0.05 in both response and non-response samples. The rank matrix of each response sample is shown in **Figure 9A**. These results indicate that there are significant differences between the identified response biomarker genes of a sample and those of other samples, that is, the proposed method also can effectively identify biomarker genes of individual samples even when fewer samples are used. We wanted to analyze biomarker genes whose frequency was higher in response samples than in non-response samples, and estimate their ability to predict survival in AB1-HA tumor samples. However, there was no follow-up information for AB1-HA mice. The selected 392 biomarker genes with higher

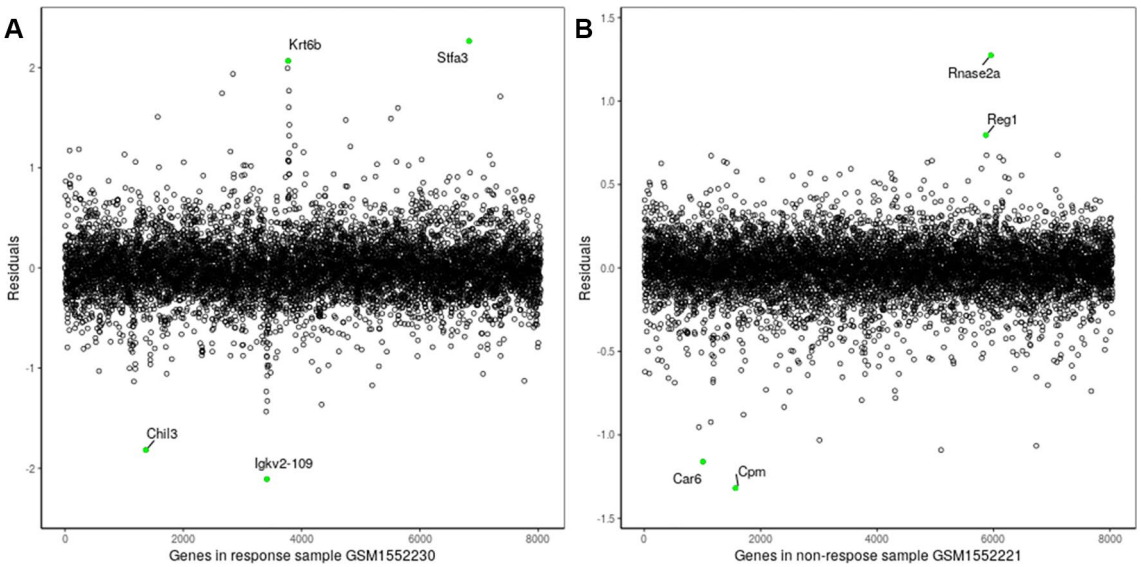


FIGURE 7 | Residuals of biomarker genes **(A)** GSM1552230, **(B)** GSM1552221.

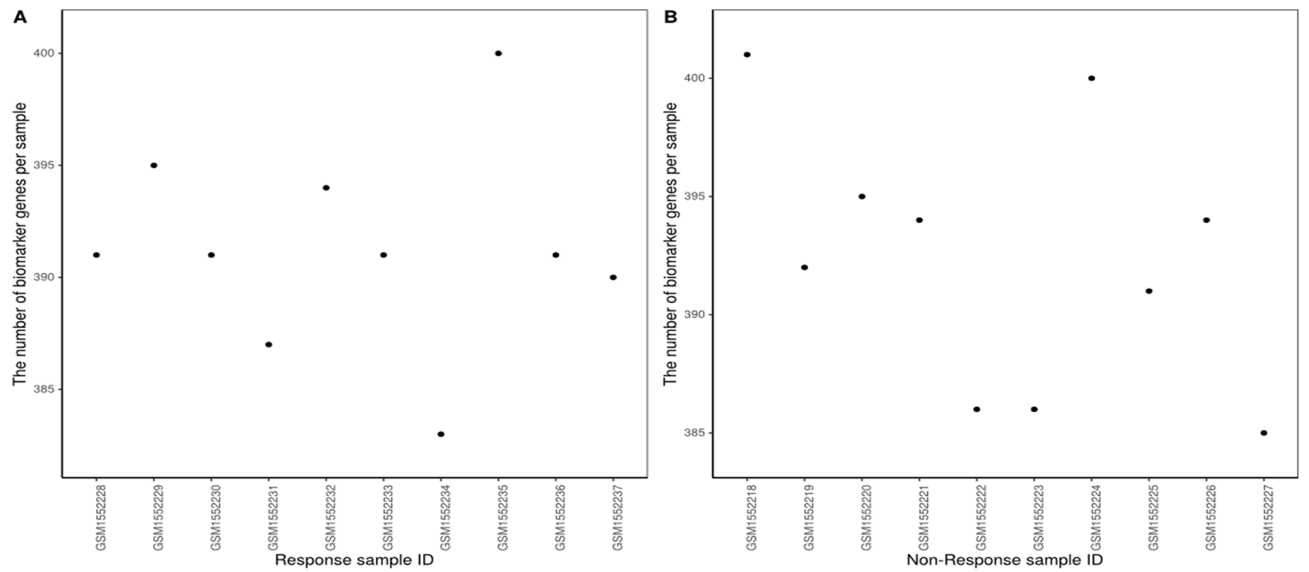


FIGURE 8 | Number of biomarker genes in **(A)** response samples and **(B)** non-response samples.

frequency were tested against a human mesothelioma data set (TCGA-MESO, https://portal.gdc.cancer.gov). Notably, these biomarker genes could still effectively distinguish all patients into high- and low-risk groups (**Figure 9B**) with a p-value of 1.57×10^{-5}. These results further support the validity of the proposed method.

Experimental Results for Advanced Melanoma Data

The proposed method was used to analyze advanced melanoma data: GSE35640. A total of 1420 DEGs were identified in 22 MAGE−A3 immunotherapeutic response and 34 non-response samples using GEO2R (Smyth, 2004) at a 95% confidence level. Regression models of 22 MAGE−A3 immunotherapeutic response and 34 non-response samples were constructed; one of these is shown in **Figure 10**. **Figure 11** shows residual values of biomarker genes from two samples. The number of biomarker genes of 22 response samples and 34 non-response samples is shown in **Figure 12**.

As shown in **Figure 12**, there were small differences in the number of biomarkers from different samples. The mean number of biomarker genes in response samples was 70. The

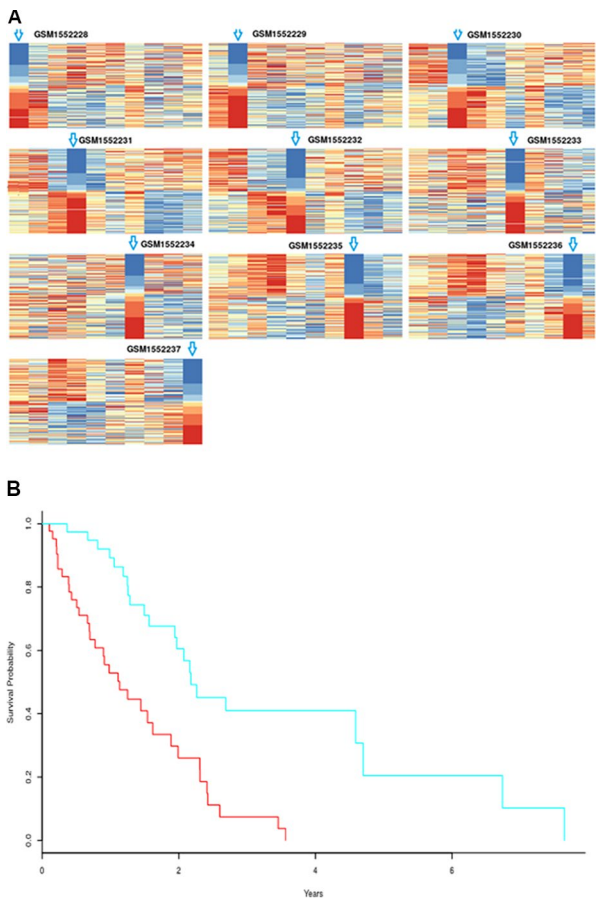


FIGURE 9 | (A) Rank matrix of each response sample. **(B)** Kaplan-Meier survival curves for human mesothelioma tumor samples based on biomarker genes from mouse AB1-HA tumor samples; p-value=1.57×10^{-5}. High-risk group includes 44 samples; low-risk group consists of 40 samples.

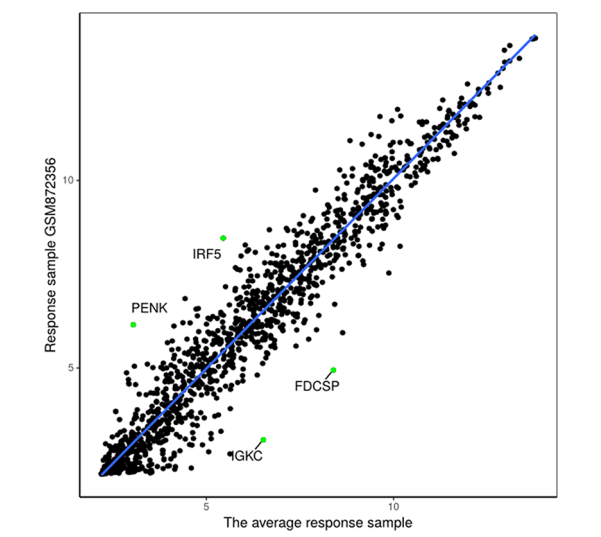


FIGURE 10 | Regression model based on response sample GSM872356 and the average response sample from GSE35640 gene expression data.

top 15 genes with the most significant difference of frequency in 22 response samples and 34 non-response samples are shown in **Supplementary Table 3**. We proposed that these genes were likely to be mainly immune or tumor related. To test our hypothesis, we searched GeneCards for these genes and found that some of them play important roles in the development of immune-related cells. For example, *MS4A1* is associated with the development of B-cells into plasma cells; *CD37* may play a part in T-cell–B-cell interactions; *CD5L* participates in obesity-associated autoimmunity; *MMP8*, *IRF5*, and *RHOF* are related to innate immune pathways; *MMP9* has a role in tumor-associated tissue remodeling; and *TRAM1L1* is related to the well-known cancer-related NF-kB pathway. This demonstrated that the IBI method could detect important genes contributing drug response mechanisms and help to elucidate immunotherapeutic response mechanisms. In the statistical tests, 96.96 and 95.72% of p-values were less than 0.05 in the response and non-response samples, respectively. These results also indicate that biomarker genes of a sample show significant differences compared with those of other samples, that is, the proposed method can also effectively identify MAGE–A3 immunotherapeutic response biomarker genes in individual advanced melanoma samples even with fewer samples.

We wanted to analyze biomarker genes whose frequency was higher in response samples than in non-response samples, and estimate their ability to predict survival in advanced melanoma. However, there was no follow-up information in GSE35640, so we used skin cutaneous melanoma gene expression data (TCGA-SKCM) for the survival analysis. The selected 70 biomarker genes were tested against TCGA-SKCM, showing that these biomarker genes could effectively distinguish skin cutaneous melanoma patients into high- and low-risk groups (**Figure 13**), with a p-value of 0.016. These results indicate that the proposed method performs well. In their original paper, Ulloa-Montoya et al. (2013) identified 84 gene expression signatures associated with response to MAGE-A3 immunotherapy in metastatic melanoma and non-small-cell lung cancer, whereas 61 of the 84 genes were chosen as biomarker genes by our proposed method (e.g., *CD86*, *CCL5*, and *IRF1*). These genes were mainly immune related and were involved in interferon gamma pathways and specific chemokines. Experimental results showed that pretreatment MAGE-A3 immunotherapy in metastatic melanoma influenced the tumor's immune microenvironment and the patient's clinical response. The proposed method could be used to identify these biomarker genes and predict the influence of MAGE-A3 immunotherapy on survival in metastatic melanoma (**Figure 13**).

Experimental Results for the Simulated Data

In order to further test the performance of the proposed method, we added a supplemental experiment on the simulated gene expression data. First, the simulated gene expression data with 10 samples 1000 genes is generated using *simulateGEdata* function in the *RUVcorr* (Freytag et al., 2015) package. Then, 1,000 genes are divided into 10 groups, we increase/decrease gene expression value of the *i*th group of genes in the *i*th sample by an up or down

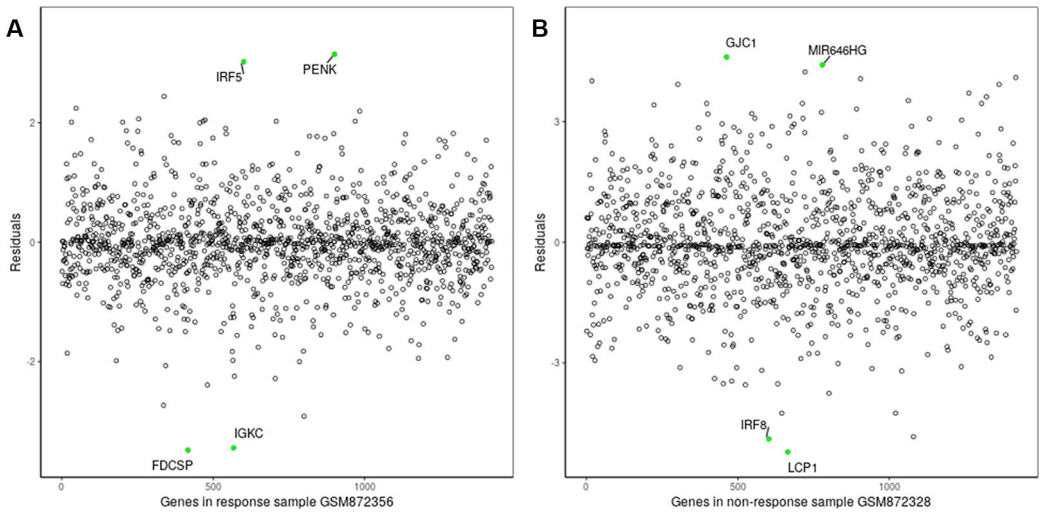


FIGURE 11 | Residuals of biomarker genes. **(A)** GSM872356, **(B)** GSM872328.

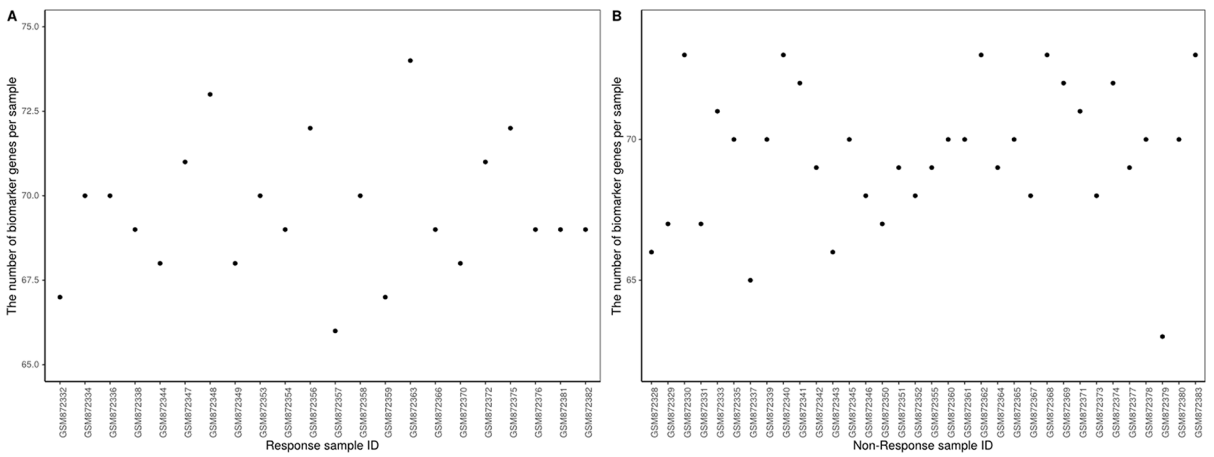


FIGURE 12 | Number of biomarker genes in **(A)** response samples and **(B)** non-response samples.

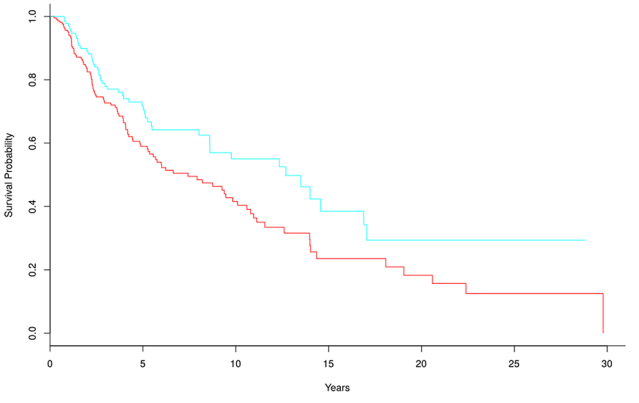


FIGURE 13 | Kaplan-Meier survival curves for TCGA-SKCM based on biomarker genes from GSE35640; p-value = 0.016. There were 281 and 166 samples in the high-risk and low-risk groups, respectively.

perturbation value. The range of perturbation value is from 0 to mean value of the corresponding gene in 10 samples. Thus, the *i*th group of genes can be considered as biomarker genes of the *i*th sample. Finally, experiment is performed on the simulated data to observe whether the proposed method can find these markers. We repeated the above steps ten times and experimental results shown that the proposed method can effectively identify the biomarker genes of 10 samples. The 99% biomarker genes identified by the proposed method are the predefined biomarkers when the perturbation value is twice (see **Supplementary Figure 1**).

DISCUSSION

Precision medicine is an active area of cancer research. The key to cancer precision medicine is to find biomarker genes with high performance, and various approaches to identify

such genes have been developed. However, identification of biomarker genes for individual tumor samples remains a challenging problem; for many reasons, there is a lack of effective approaches to identify biomarkers in individual patients. Here, we developed a novel approach to address this issue. Experimental results based on several different data sets show that the proposed method can effectively identify biomarker genes of individual human tumor samples, not only from several hundred samples but also from a few samples without clinical information, and even from mouse samples.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study: TCGA-BRCA data (found at The Cancer Genome Atlas), GSE63557 (found at Gene Expression Omnibus) and GSE35640 (found at Gene Expression Omnibus).

AUTHOR CONTRIBUTIONS

JL and DW designed and implemented the algorithm. JL and DW analyzed the results and wrote the manuscript, and YW made suggestions. All authors read and approved the final manuscript.

FUNDING

This work was partially supported by National Key Research and Development Program of China (Grant No.2016YFC0901905) and the Natural Science Foundation of Heilongjiang Province (Grant No. F2016016).

ACKNOWLEDGMENTS

The authors acknowledge the contributions of colleagues in the group.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01236/full#supplementary-material

REFERENCES

Al-Khadairi, G., Naik, A., Thomas, R., Al-Sulaiti, B., Rizly, S., and Decock J. (2019). PRAME promotes epithelial-to-mesenchymal transition in triple negative breast cancer. *J. Transl. Med.* 17, 9. doi: 10.1186/s12967-018-1757-3

Bhakta, S., Crocker, L. M., Chen, Y., Hazen, M., Schutten, M. M., Li, D., et al. (2018). An anti-GDNF family receptor alpha 1 (GFRA1) antibody–drug conjugate for the treatment of hormone receptor-positive breast cancer. *Mol. Cancer Ther.* 17, 638–649. doi: 10.1158/1535-7163.MCT-17-0813

Bokobza, S. M., Ye, L., Kynaston, H. E., Mansel, R. E., and Jiang, W. G. (2009). Reduced expression of BMPR-IB correlates with poor prognosis and increased proliferation of breast cancer cells. *Cancer Genomics Proteomics* 6, 101.

Chang, H. Y., Nuyten, D. S., Sneddon, J. B., Hastie, T., Tibshirani, R., Sørlie, T., et al. (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci.* 102, 3738–3743. doi: 10.1073/pnas.0409462102

Dai, D.-N., Li, Y., Chen, B., Du, Y., Li, S.-B., Lu, S.-X., et al. (2017). Elevated expression of CST1 promotes breast cancer progression and predicts a poor prognosis. *J. Mol. Med.* 95, 873–886. doi: 10.1007/s00109-017-1537-1

Dembélé, D., and Kastner, P. (2014). Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinf.* 15, 14. doi: 10.1186/1471-2105-15-14

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Emberley, E. D., Murphy, L. C., and Watson, P. H. (2004). S100A7 and the progression of breast cancer. *Breast Cancer Res.* 6, 153–159. doi: 10.1186/bcr816

Emberley, E. D. (2003). Psoriasin (S100A7) expression is associated with poor outcome in estrogen receptor-negative invasive breast cancer. *Clin. Cancer Res.* 9, 2627–2631.

Freytag, S., Gagnon-Bartsch., J., Speed., T. P., Bahlo, M. (2015). Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics* 16 (1), 309. doi: 10.1186/s12859-015-0745-3

Jia, D., Liu, Z., Deng, N., Tan, T. Z., Huang, R. Y.-J., Taylor-Harding, B., et al. (2016). A COL11A1-correlated pan-cancer gene signature of activated fibroblasts for the prioritization of therapeutic targets. *Cancer Lett.* 382, 203–214. doi: 10.1016/j.canlet.2016.09.001

Jiang, P., Gu, S., Pan, D., Fu, J., Sahu, A., Hu, X., et al. (2018). Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* 24, 10. doi: 10.1038/s41591-018-0136-1

Kammerer, S., Sokolowski, A., Hackl, H., Platzer, D., Jahn, S. W., El-Heliebi, A., et al. (2016). KCNJ3 is a new independent prognostic marker for estrogen receptor positive breast cancer patients. *Oncotarget* 7, 84705. doi: 10.18632/oncotarget.13224

Lesterhuis, W. J., Rinaldi, C., Jones, A., Rozali, E. N., Dick, I. M., Khong, A., et al. (2015). Network analysis of immunotherapy-induced regressing tumours identifies novel synergistic drug combinations. *Sci. Rep.* 5, 12298. doi: 10.1038/srep12298

Lewis-Wambi, J. S., Cunliffe, H. E., Kim, H. R., Willis, A. L., and Jordan, V. C. (2008). Overexpression of CEACAM6 promotes migration and invasion of oestrogen-deprived breast cancer cells. *Eur. J. Cancer* 44, 1770–1779. doi: 10.1016/j.ejca.2008.05.016

Li, J., Tang, X., Zhao, W., and Huang, J. (2007). A new framework for identifying differentially expressed genes. *Pattern Recognit.* 40, 3249–3262. doi: 10.1016/j.patcog.2007.01.032

Li, J., Tang, X., Liu, J., Huang, J., and Wang, Y. (2008). A novel approach to feature extraction from classification models based on information gene pairs. *Pattern Recognit.* 41, 1975–1984. doi: 10.1016/j.patcog.2007.11.019

Li, J., Lenferink, A. E. G., Deng, Y., Collins, C., Cui, Q., Purisima, E. O., et al. (2010). Identification of high-quality cancer prognostic markers and metastasis network modules. *Nat. Commun.* 1, 34. doi: 10.1038/ncomms1033

Liang, W.-H., Li, N., Yuan, Z.-Q., Qian, X.-L., and Wang, Z.-H. (2019). DSCAM-AS1 promotes tumor growth of breast cancer by reducing miR-204-5p and up-regulating RRM2. *Mol. Carcinog.* 58, 461–473. doi: 10.1002/mc.22941

Liu, X., Chang, X., Liu, R., Yu, X., Chen, L., and Aihara, K. (2017). Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PloS Comput. Biol.* 13, e1005633. doi: 10.1371/journal.pcbi.1005633

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi: 10.1186/s13059-014-0550-8

Moore, S. G., Pryce, J. E., Hayes, B. J., Chamberlain, A. J., Kemper, K. E., Berry, D. P., et al. (2016). Differentially expressed genes in endometrium and corpus luteum of Holstein cows selected for high and low fertility are enriched for sequence variants associated with fertility. *Biol. Reprod.* 94, 11–19. doi: 10.1095/biolreprod.115.132951

Ni, J., Yun, P., Fu-Lan, Y., Xun, X., Xing-Wei, H., and Chun, H. (2018). Overexpression of CLEC3A promotes tumor progression and poor prognosis in breast invasive ductal cancer. *Oncotargets Ther.* 11, 3303–3312. doi: 10.2147/OTT.S161311

Ock, C.-Y., Hwang, J.-E., Keam, B., Kim, S.-B., Shim, J.-J., Jang, H.-J., et al. (2017). Genomic landscape associated with potential response to anti-CTLA-4 treatment in cancers. *Nat. Commun.* 8, 1050. doi: 10.1038/s41467-017-01018-0

Powell, E., Shao, J., Picon, H. M., Bristow, C., Ge, Z., Peoples, M., et al. (2018). A functional genomic screen in vivo identifies CEACAM5 as a clinically relevant driver of breast cancer metastasis. *NPJ Breast Cancer* 4, 9. doi: 10.1038/s41523-018-0062-x

Rezwan, F. I., Docherty, L. E., Poole, R. L., Lockett, G. A., Arshad, S. H., Holloway, J. W., et al. (2015). A statistical method for single sample analysis of HumanMethylation450 array data: genome-wide methylation analysis of patients with imprinting disorders. *Clin. Epigenet.* 7, 48. doi: 10.1186/s13148-015-0081-5

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 3. doi: 10.2202/1544-6115.1027

Su, J., Miao, L.-F., Ye, X.-H., Cui, M.-S., and He, X.-F. (2019). Development of prognostic signature and nomogram for patients with breast cancer. *Medicine* 98, 11. doi: 10.1097/MD.0000000000014617

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68. doi: 10.5114/wo.2014.47136

Ulloa-Montoya, F., Louahed, J., Dizier, B., Gruselle, O., and Brichard, V. G. (2013). Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J. Clin. Oncol.* 31, 2388. doi: 10.1200/JCO.2012.44.3762

Van't Veer, L. J., and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature* 452, 564. doi: 10.1038/nature06915

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530. doi: 10.1038/415530a

Wang, D., Li, J.-R., Zhang, Y.-H., Chen, L., Huang, T., and Cai, Y.-D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 9, 155. doi: 10.3390/genes9030155

Xuan, J., Zhang, Y., Zhang, X., and Hu, F. (2015). Matrix metalloproteinase-1 expression in breast cancer and cancer-adjacent tissues by immunohistochemical staining. *Biomed. Rep.* 3, 395–397. doi: 10.3892/br.2015.420

Zhang, Q., Li, J., Wang, D., and Wang, Y. (2017). Finding disagreement pathway signatures and constructing an ensemble model for cancer classification. *Sci. Rep.* 7, 10044. doi: 10.1038/s41598-017-10258-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 December 2019
doi: 10.3389/fbioe.2019.00311

Identifying Acetylation Protein by Fusing Its PseAAC and Functional Domain Annotation

Wang-Ren Qiu[1,2], *Ao Xu*[1], *Zhao-Chun Xu*[1], *Chun-Hua Zhang*[1] *and Xuan Xiao*[1]*

[1] School of Information and Engineering, Jingdezhen Ceramic Institute, Jingdezhen, China, [2] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Acetylation is one of post-translational modification (PTM), which often reacts with acetic acid and brings an acetyl radical to an organic compound. It is helpful to identify acetylation protein correctly for understanding the mechanism of acetylation in biological systems. Although many acetylation sites have been identified by high throughput experimental studies via mass spectrometry, there still are lots of acetylation sites need to be discovered. Computational methods have showed their power for identifying acetylation sites with informatics techniques which usually reduce experiment cost and improve the effectiveness and efficiency. In fact, if there is an approach can distinguish the acetylated proteins from the non-acetylated ones, it is no doubt a very meaningful and effective method for this issue. Here, we proposed a novel computational method for identifying acetylation proteins by extracting features from the conservation information of sequence via gray system model and KNN scores based on the information of functional domain annotation and subcellular localization. The authors have performed the 5-fold cross-validation on three datasets along with much analysis of features and the Relief feature selection algorithm. The obtained accuracies are all satisfactory, as the mean performance, the accuracy is 77.10%, the Matthew's correlation coefficient is 0.5457, and the AUC value is 0.8389. These works might provide useful insights for the related experimental validation, and further studies of other PTM process. For the convenience of related researchers, the web-server named "iACetyP" was established and is accessible at http://www.jci-bioinfo.cn/iAcetyP.

Keywords: acetylation, Random Forest, family and domain databases localization, post-translational modification, identification

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Tianfan Fu,
Georgia Institute of Technology, United States
Pu-Feng Du,
Tianjin University, China

***Correspondence:**
Xuan Xiao
jdzxiaoxuan@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *12 September 2019*
Accepted: *22 October 2019*
Published: *06 December 2019*

Citation:
Qiu W-R, Xu A, Xu Z-C, Zhang C-H and Xiao X (2019) Identifying Acetylation Protein by Fusing Its PseAAC and Functional Domain Annotation. Front. Bioeng. Biotechnol. 7:311. doi: 10.3389/fbioe.2019.00311

INTRODUCTION

To date, more than 450 unique protein modifications have been identified (Han et al., 2018), including phosphorylation, acetylation, ubiquitination, and sumoylation, which are regulatory mechanisms of cellular proteins with a number of biological functions, and also are very important for regulating the function of many prokaryotic and eukaryotic proteins (Yang et al., 2017). Among these post-translational modification (PTM), acetylation is a dynamic and highly conserved PTM (**Figure 1**) that plays a vital role in the regulating processes of diverse cellular. The role of acetylation in histones were first discovered in histones (Allfrey et al., 1964), and the first deacetylase activity was identified back in 1969 (Inoue and Fujimoto, 1969). Owing to its important involvement in some relevant biological processes, acetylation becomes one of the most important reversible

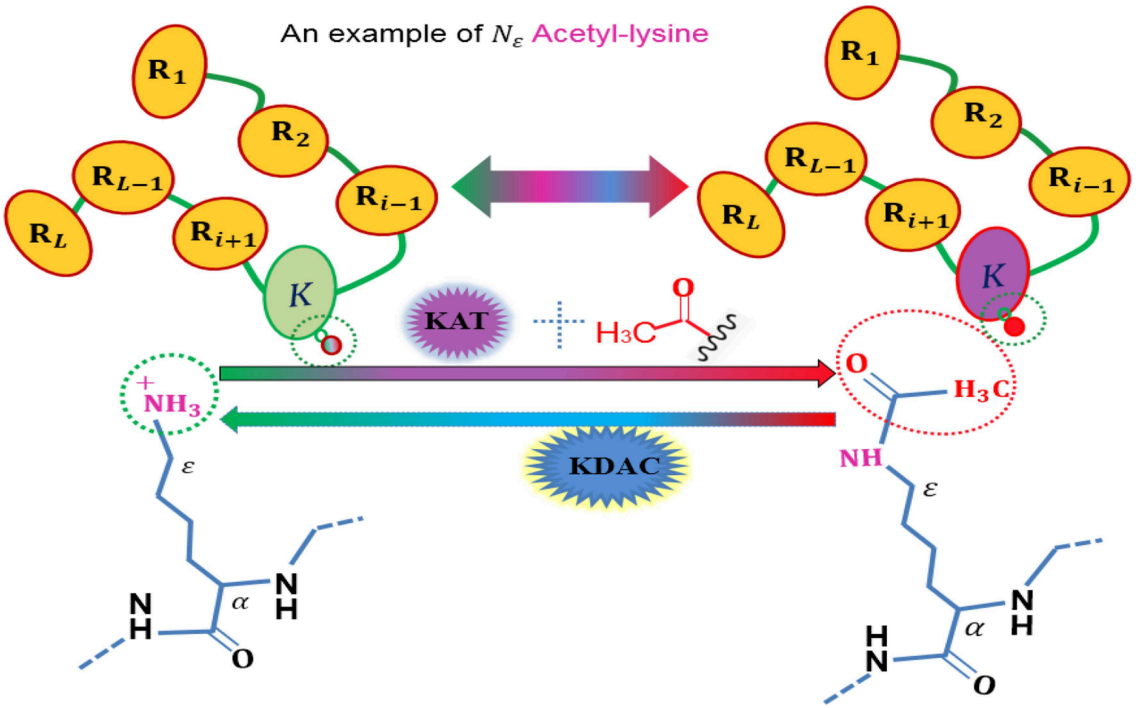


FIGURE 1 | An illustration to show the acetylation protein.

protein posttranslational modifications, hence, more and more acetylated proteins are discovered with the help of high-throughput technologies. Thus, it is a piece of very interesting work to identify the potential acetylation sites for finding the underlying molecular mechanisms, and is helpful for basic bioresearch and drug development.

However, due to the importance and complexity of acetylation, identifying acetylation sites is a great challenge to fully understand the regulatory roles and the molecular mechanism of acetylation regulation. Actually, it is a time-consuming, expensive and labor-intensive process for purifying acetylation sites due to that the acetylation process is dynamic, rapid and reversible (Li et al., 2017; Yang et al., 2017). Fortunately, some studies had showed that experimental methods and computational models can be used to identify underlying PTMs sites (Hershko and Ciechanover, 1998; Haglund and Dikic, 2005; Tung and Ho, 2008; Radivojac et al., 2010), such as ubiquitination model of Radivojac et al. (2010), Zhao et al. (2011), and Cai et al. (2012), phosphorylation model of Ingrell et al. (2007), Yao et al. (2012, 2015), Chen et al. (2015), Li et al. (2015), Trost et al. (2015), and Xu et al. (2015), sumoylation model of Beauclair et al. (2015), Xu et al. (2016), and Han et al. (2018), acetylation model of Zhao et al. (2010), Wang et al. (2012), Hou et al. (2014), and Wuyun et al. (2016), and so on. Although these researchers did make much contribution to this issue, there is still a lot of room for improving the prediction quality. However, most of these efforts are on identifying some determinate PTMs sites for a given protein sequence, and few of computational method was proposed for distinguishing the acetylated proteins from the non-acetylated ones. This study was an attempt for the issue.

For a given protein represented with amino acid sequence, how to identify whether it may be one of some certain PTM proteins or may not? This may be the first step for identifying PTM sites and then is helpful and meaningful for basic research and drug development. In fact, we have made some preliminary exploration and attempt on identifying phosphorylated proteins. In Qiu et al. (2017a,b), we presented a method for identifying human phosphorylated proteins and a multi-label classifying model for different type of phosphorylated proteins with the help of the General PseAAC concept and gray system theory. Although the results are not so perfect, we still argue that the formulations and models can be applied to this issue, and it may be more powerful when some structure, function or localization information of proteins were added into the model. This site may be a fruitful opportunity for bioinformatics. For example, Gene Ontology (GO) (Ashburner et al., 2000) was proposed by Ashburner to reposit the concepts denoted as GO Terms that are associated to other gene products, and it has been widely used in describing the attributes for gene products (Agapito et al., 2016; Peng et al., 2016).

The dataset we used here was fully extracted from Uniprot (The UniProt, 2017). The present study tried to construct a classifying model for potential acetylation proteins by fusing the digital features which are come from its evolution information, Subcellular localization (noted as **SL**) (Nakai and Horton, 1999) information and functional domain annotation (noted as **FDA**) databases including GO (Harris et al., 2004), Pfam (Bateman

et al., 1999), Smart (Letunic et al., 2004), InterPro (Hunter et al., 2009), PRINTS (Attwood et al., 2012), PROSITE (Sigrist et al., 2010), SUPFAM (Pandit et al., 2004). As for subcellular localization (Du et al., 2012), it was retrieved from the original UniProt database, which was reorganized by UniProt build-in hierarchical subcellular localization table. This paper proposed a new computational model for identifying potential acetylation proteins only on the basis of a query amino acid sequence by using its evolution information obtained with gray system model (Gray-PSSM) (Kaur and Raghava, 2004; Jones, 2007) and KNN scores calculated with the fuzzy distance by using its FDA and subcellular localization information. There are 80 amino acid sequence features extracted by incorporating the sequence evolution information were fused into PseAAC feature set and KNN scores, all of these features are combined according to different coefficients on the basis of its importance. To highlight the advantages of the proposed model, the model was trained and tested with three sub-datasets and cross-validations methods. In addition to some discussion of protein abovementioned features, some hypotheses for distinguishing acetylation proteins from non-acetylation ones were also depicted with the aid of training dataset.

MATERIALS AND METHODS

Benchmark Dataset

It is fundamental and important that a stringent benchmark dataset be stablished for testing the proposed statistical predictor. Luckily, the UniProtKB/Swiss-Prot database is accepted by most of bioinformatics researchers, and has been using more and more widely. The data used in the current study to support this work are established on the basis of web http://www.uniprot.org.

In this study, we assume that our work is to identify whether an uncharacterized amino acid sequence is acetylation protein. As we known, the input sequence is comprised by amino acids and can be expressed as

$$\mathbf{P} = P_1 P_2 P_3 \cdots P_i \cdots P_L \tag{1}$$

where P_i represents the i-th residue of amino acids sequence $\mathbf{P}$, L is the length of $\mathbf{P}$.

Here, we separate a benchmark dataset into a training dataset noted as S. Thus, the datasets can be formulated as:

$$\mathbf{S_{all}} = S_{posi} \cup S_{nega} \quad S_{nega} = S_1^- \cup S_2^- \cup S_3^- \tag{2}$$

where S_{posi} is composed of the acetylation proteins, S_{nega} is composed of the non-acetylation proteins, $\mathbf{S_i^-} \cap \mathbf{S_j^-} = \varnothing$ $(\mathbf{i} \neq \mathbf{j}; \mathbf{i,j} = \mathbf{1,2,3})$. $\cup$ and $\cap$ represent the symbol for "set union" and "set intersection," respectively.

The version of protein data used in the current study was released in May 2017. The positive dataset was generated according to the following criteria: (1) The potential acetylated proteins should be noted by anyone keyword of the set, i.e. {N_acetylcysteine, N_acetylserine, N_acetylglutamate, N_acetylglycine, N_acetylproline, N_acetylthreonine, N_acetylvaline, N_acetylmethionine, N_acetyltyrosine, N2_acetylarginine, N6_acetyllysine, O_acetylserine, O_acetylthreonine}. (2) The collected proteins are labeled by "Evidence" for the item of "Any assertion method." (3) Only the proteins which consisting of 30 and more amino acid residues can be included, and the redundant proteins were removed with the threshold of 50% by using CD-HIT software.

The negative dataset was generated similar to the positive one except that those proteins should not be labeled none member of the mentioned above keyword-set. Since there are mass number of candidates here, we randomly collected negative datasets which have the balance samples size with positives.

Under the aforementioned standards, we obtained 2,925 protein samples, of which, the numbers of positive and negative samples are 725 and 2,175, respectively. In terms of Equation (2), we have 725 positive samples in S_{posi} and 2,175 negative samples in S_{nega}. Here, we test the models with cross-validation on the three datasets with 1,450 samples, i.e., $S_{posi} \cup S_1^-$, $S_{posi} \cup S_2^-$ and $S_{posi} \cup S_3^-$, of which, the positive and negative ones are equal, i.e., 725 samples.

Feature Construction

General Pseudo Amino Acid Composition (PseAAC)

Most of traditional machine-learning algorithms, such as Random Forest, SVM, and K nearest Neighbor, are not so powerful, the input should be vectors instead of sequence samples for biological issue. To overcome this problem, the researchers trying their best to improve the discrete or vector model by formulating the amino acids sequence into all kinds of pseudo amino acid composition (PseAAC), encoding method (Zhang et al., 2006; Chen et al., 2011; Shi et al., 2012; Jiao and Du, 2017) or other approaches.

Here, the proposed model followed the idea of PseAAC (Chou, 2011), and formulated an amino acids sequence $\mathbf{P}$ as:

$$P = \begin{bmatrix} p_1 & p_2 & \cdots & p_u & \cdots & p_N \end{bmatrix}^{\mathbf{T}} \tag{3}$$

Here, the symbol $\mathbf{T}$ means the transpose operator for a matrix, N is an integer representing the number of features which depend on the method(s) used for extracting information from protein $\mathbf{P}$ (cf. Equation 1). P is a vector for representing amino acids sequence $\mathbf{P}$ and p_i $(i = 1, 2, \cdots, N)$ is the ith element of the vector. Below, we will describe how to extract functional domain annotation and subcellular localization information as well as pseudo amino acid composition, which are used in this study, from a query sequence to define the components for amino acids sequence $\mathbf{P}$.

Protein Sample Formulation With KNN Score Based on FDA and Subcellular Localization (SL)

In addition to GO database, "Pfam," "Smart," "PROSITE," "SUPFAM," "InterPro," and "PRINTS" are established according to cellular component, molecular function, biological process or some other characteristics. For example, the Pfam database is a large collected protein families generated by using hidden Markov models. SMART is abbreviation of Simple Modular Architecture Research Tool which can be used for research on the protein domains and architectures. PROSITE consists of entries describing the protein families, domains and functional sites as well as amino acid patterns and profiles. InterPro provides a

functional analysis of protein sequences, and PRINTS also is a resource of detailed annotation for protein families in addition to a diagnostic tool for newly determined sequences. Subcellular localization feature is a key functional characteristic of potential gene products such as proteins, especially for plant.

Actually, in the GO database, proteins are clustered in a way in which their subcellular locations can be reflected fully. To incorporate more information, most of the approaches need to formulate a long list of the GO numbers, and a great part of the GO numbers make meaningless as a whole. In literatures (Gao et al., 2010; Yao et al., 2012), the authors show us that local sequence clusters often appear in the neighborhood of PTM sites for the reason that the same PTM family generally have some similarities in local sequences. As a better choice for depicting the character, K nearest neighbor score was proposed. To take advantage of such cluster information of GO and other FDA databases as well as subcellular localization for predicting acetylation proteins, for a given potential acetylation protein, we took the characteristics around the query neighborhood and extracted the KNN scores features from the training dataset containing both positive and negative samples. The algorithm is listed as follows.

Step 1. For a query protein sequence find its k nearest neighbors, which can be positive or negative samples, in the whole set according to local sequence similarity. For a given protein $\boldsymbol{p}$, $FDA_j(p) = \{N_1^{p,j}, N_2^{p,j}, \cdots, N_{n_p}^{p,j}\}$ represents the keywords set of the jth FDA. The j (=1, 2, …, 7, 8) represents "GO," "Pfam," "Smart," "PROSITE," "SUPFAM," "InterPro," "PRINTS," or "subcellular localization," respectively), $\mathrm{FDA}_j(q) = \{N_1^{q,j}, N_2^{q,j}, \cdots, N_{n_q}^{q,j}\}$ is the similar mean for protein $\boldsymbol{q}$. The similarity distance $\mathrm{Dist}_j(\boldsymbol{p}, \boldsymbol{q})$ between $\boldsymbol{p}$ and $\boldsymbol{q}$ can be defined as follows:

$$\mathrm{Dist}_j(\boldsymbol{p}, \boldsymbol{q}) = \mathrm{w}_1 \cdot \left(1 - \frac{|\mathrm{FDA}_p(j) \bigcap \mathrm{FDA}_q(j)|}{|\mathrm{FDA}_p(j) \bigcup \mathrm{FDA}_q(j)|}\right) + \mathrm{w}_2 \cdot dist(\boldsymbol{p}, \boldsymbol{q}) \quad (4)$$

Where $\bigcap, \bigcup$ and $|\ |$ are the operators "union," "intersection," and "norm" of the set theory, respectively. Here, $|\ |$ is defined as the number of its elements. $dist(\boldsymbol{p}, \boldsymbol{q})$ is the Euclidean distance on the basis of PseAAC. w_1 and w_2 are the weights of the two distances.

Step 2. A corresponding KNN feature is then extracted by calculating the KNN score, noted it as the percentage of acetylation proteins in its k nearest neighbors.

Step 3. To obtain diverse and enough properties of neighbors with KNN scores, the above two steps were repeated for different k values. For the jth member of FDA, the protein **P** can be formulated as:

$$\boldsymbol{P}_{\mathrm{FDA}_j} = [\varphi_1(j),\ \varphi_2(j), \cdots,\ \varphi_K(j)\ ]^T \quad (5)$$

In this paper, the number of features is 50 and k was defined to be 0.1, 0.4, 0.7, …, 14.5 and 14.8 percent of the size of the involved dataset. In this way, 50 KNN scores were extracted as features for identifying acetylation proteins. To be more precisely, $\varphi_1(j)$ is the ratio of positive neighbors to whole concerned samples, i.e., 0.1 percent of the size of the training data set, $\varphi_2(j)$ is the ratio of positive neighbors to whole concerned samples whose value is the product of 0.004 and the size of the training data set, and so forth, when $K = 50$, $\varphi_{50}(j)$ is the ratio of positive neighbors to 14.8 percent of the size of the training data set.

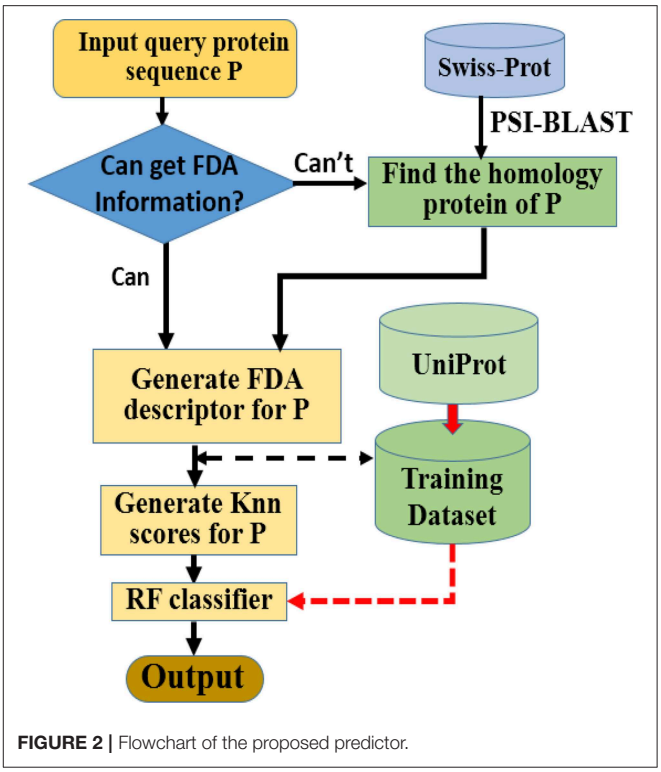


FIGURE 2 | Flowchart of the proposed predictor.

In a word, a query protein sequence can be formulated with seven 50-Dimension vectors, i.e., $\boldsymbol{P}_{\mathrm{FDA}} = [\boldsymbol{P}_{\mathrm{FDA}_1}, \boldsymbol{P}_{\mathrm{FDA}_2}, \ldots, \boldsymbol{P}_{\mathrm{FDA}_7}]$, by using FDA database. Since Chou's pseudo amino acid composition (PseAAC) (Chou, 2001; Mondal and Pai, 2014) have showing so great powerful for identifying structure and function of protein, the proposed method took it into account according to the style of reference (Shen and Chou, 2008) (we select type 1 and let $\lambda = 5$). Thus, a given protein sequence can be expressed as 375-dimension vector, and these digital representations served as the input of the query protein for the prediction model.

Operation Engine and Evaluation

Algorithms

Here we choose Random Forest as the operation engine as the predictor, and named the final predictor as "iAcet-PseFDA." This name is an acronym created from its description, and **Figure 2** would show how iAcet-PseFDA working.

As shown in **Figure 2**, the first step is to input the query amino acid sequence **P**. And then, the PSI-BLAST software was used to find the most similar protein to **P**, which is used to determine the most likely GO or other information of FDA set and generate the KNN scores with it. With the descriptor of **P**, the desired result can be obtained with the framework of Random Forest classifier trained on the benchmark.

Metrics and Test Method

The predictor iAcet-PseFDA was evaluated with cross-validation tests in the terms of following seven widely-accepted measurements: accuracy (or Acc, for short), Mathew's correlation coefficient abbreviated as Mcc, sensitivity (abbreviated as Sn, i.e., the fraction of the relevant documents that are successfully retrieved), specificity (i.e., Sep), Precision (i.e., Pre, a description of random errors), F-measure (or F-m, the harmonic mean of precision and recall), and G-mean. Since the area under the receiver operating characteristic curve (auROC, for short) is another important measurement of the performance of a given model, it was also calculated and plotted in this study. In view of the traits of validation method trait, cross-validation method was applied on three datasets for evaluating the proposed predictor.

RESULTS AND DISCUSSION

Investigating the Performances of KNN Score of FDA Represent

Figure 3 depicted the comparisons of the KNN scores of acetylation and non-acetylation proteins on all of the FDA features, and there really are some differences between the positive and negative samples. **Figure 3A** showed the comparison of PAAC represents between acetylation proteins and non-acetylation proteins, **Figure 3B** showed those of KNNScore-GO, and so forth, **Figure 3I** showed those of Subcellular localization. Overall, acetylation proteins gained obvious larger KNN scores than non-acetylation proteins on GO and Subcellular localization, and a little larger gap between the KNNScores of positive and negative datasets, all of the average KNN scores are nearly merged in 0.5 with the growth of features.

Specifically, for acetylation proteins with the view of GO evaluated on different sizes of nearest neighbors, the average values shown in **Figure 3B** are within 0.6–0.8, however, the average digits are within 0.2–0.4 for non-acetylation proteins. From the view of Subcellular localization as showed in **Figure 3I**, most of the average KNN scores of acetylation proteins are waved within 0.5–0.7 while those of non-acetylation proteins fluctuating around 0.4. From the view of Smart, Supfam, InterPro Pfam, Prosite and PRINTS as showed in **Figures 3C–H**, there are clearly gaps between the acetylation proteins and non-acetylation proteins, and the gaps are narrowing with the growth of KNNScores number.

We tested the eight kinds of features on the three datasets with RF, and the mean performances are depicted in the first 11th lines of **Table 1**, while the compared measurements obtained from the proposed model, in which the features were selected with Relief, are attached in the last line. As showed in the table, the features of Subcellular localization reached the best results with Acc is 73.95%, Mcc is 0.4843, Sn is 81.24%, Recall is 81.24%, F-measure is 75.72%, and G-mean is 73.59%. As regards for Sp and Precison, GO gained the best result which are 68.37 and

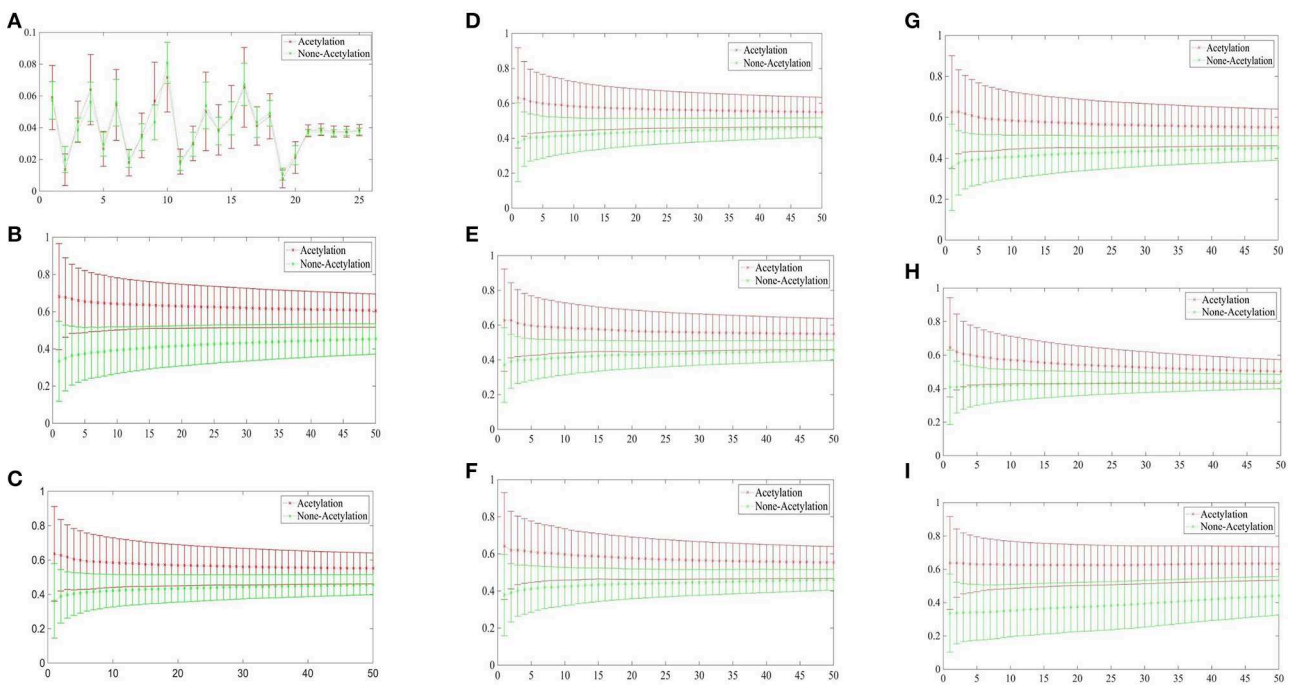


FIGURE 3 | (A) Comparison of KNNScore-PAAC represents between acetylation proteins and non-acetylation proteins. **(B)** Comparison of KNNScore-GO represents between acetylation proteins and non-acetylation proteins. **(C)** Comparison of KNNScore-Pfam represents between acetylation proteins and non-acetylation proteins. **(D)** Comparison of KNNScore-Smart represents between acetylation proteins and non-acetylation proteins. **(E)** Comparison of KNNScore-Prosite represents between acetylation proteins and non-acetylation proteins. **(F)** Comparison of KNNScore-Supfam represents between acetylation proteins and non-acetylation proteins. **(G)** Comparison of KNNScore-InterPro represents between acetylation proteins and non-acetylation proteins. **(H)** Comparison of KNNScore-PRINTS represents between acetylation proteins and non-acetylation proteins. **(I)** Comparison of KNNScore-SL represents between acetylation proteins and non-acetylation proteins.

TABLE 1 | Mean performance comparison with different KNN score feature tested with RF.

	Acc%	Mcc%	Sn%	Sp%	Pre%	F_m%	Gmean%
PAAC	69.49	0.3909	73.01	65.98	68.22	70.53	69.40
GO	73.61	0.4754	78.85	68.37	71.39	74.90	73.39
Pfam	68.21	0.3647	70.99	65.43	67.27	69.07	68.14
Smart	67.01	0.3410	70.11	63.91	66.04	68.01	66.93
PROSITE	68.37	0.3679	70.85	65.89	67.52	69.14	68.31
SUPFAM	68.67	0.3738	71.17	66.16	67.79	69.44	68.62
InterPro	68.25	0.3658	71.40	65.10	67.17	69.22	68.18
PRINTS	66.11	0.3234	69.52	62.71	65.10	67.21	65.99
Subcellular localization	73.95	0.4843	81.24	66.67	70.91	75.72	73.59
All-MeanJK	74.64	0.4980	81.38	67.91	**71.78**	76.24	74.30
This paper	**77.55**	**0.5883**	**96.41**	**71.26**	52.79	**68.23**	**82.89**

**The bold value means the largest element of the column.*

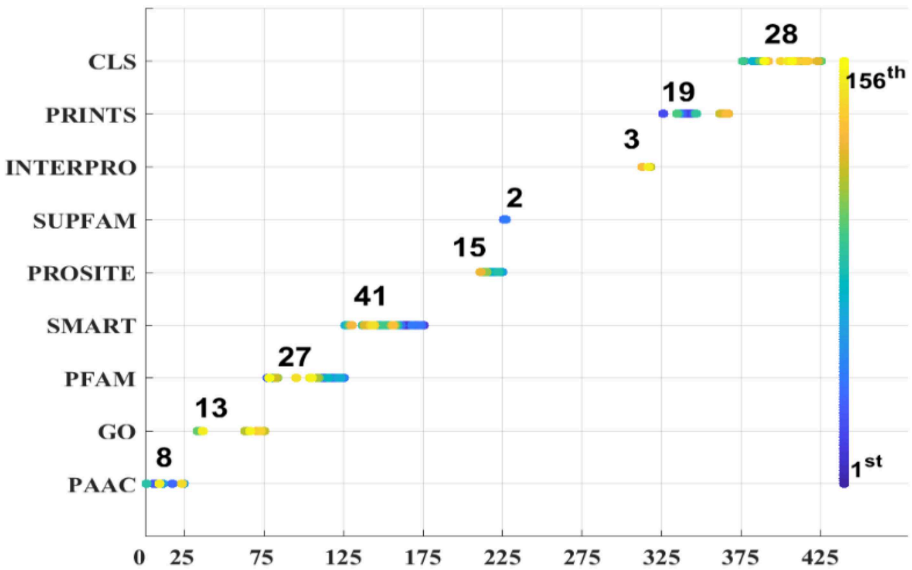


FIGURE 4 | Features selected by Relief.

71.39%, respectively. Thus, the features of GO gained the second place. The other six performances are not satisfactory and worse than those of GO and Subcellular localization, all Accs of them are <0.7 except for GO and Subcellular localization. The results obtained with the enhanced model are discussed below.

Performance of Proposed Model

Based on the above discussions, we argue that the local amino acids surrounding acetylation sites, which have been verified, would share in similar pattern(s) with positive set on average as expected. These findings confirm that there are some acetylation-related clusters in acetylated proteins and hence may be used to distinguish them from the non-acetylation protein. Accordingly, the KNN scores were used to encode query sequence for predicting acetylation proteins in this study.

As we known, the Relief algorithm as a feature weighting algorithm was first proposed by Kira and Rendell (1992). In the algorithm, the features were allocated different weights in light of the relevance of characteristics and categories. The feature will be removed when its weight less than a threshold by this method. Since the combined features generated a high-dimensional vector, and the Relief method can rank the values of features, this work thus used Relief to reduce feature redundancy. With the help of Relief, we tested the predictor on different features sets and listed the mean performances in the last line of **Table 1**. The Acc is 77.55% which is better than 74.64%, the result obtained by using all of the eight features, and better than that of subcellular localization. The Relief model gained the better results according to the other seven measurements. **Figure 4** depicted the selected features by Relief algorithm which containing 156 potential features (of which, there are 8 PSSM-gray features, 13 GO KNNScores, 27 for PFAM, 41 for SMART, 15 for PROSITE, 2 for SUPFAM, 3 for INTEPRO, 19 for PRINTS, and 28 for Subcellular localization KNNScores). From the figure, we can see that the importance of PAAC, SMART and PRINTS are obvious since a lot of features are noted as blue which means their rank

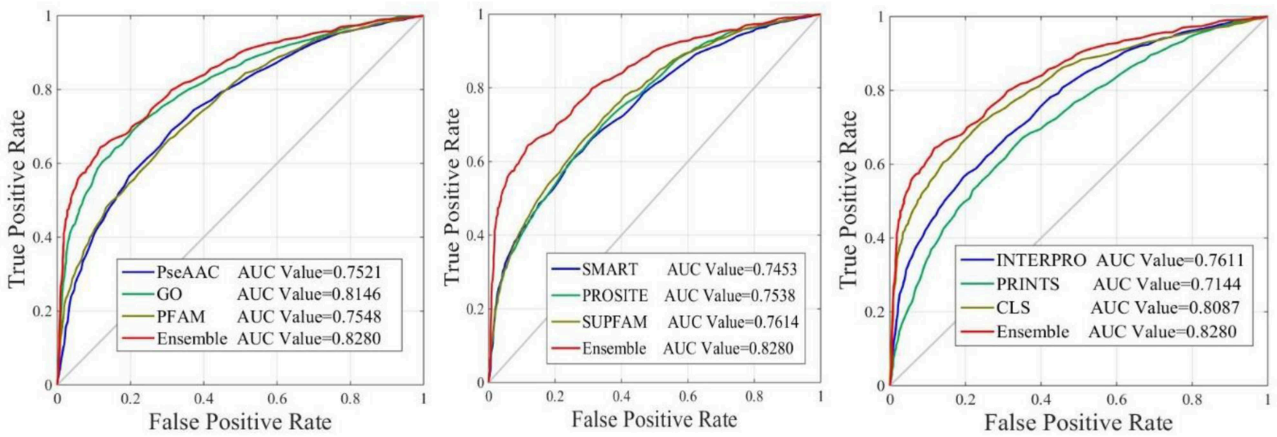


FIGURE 5 | The ROC curves of predictor with different features.

in the selected feature set. The predictor obtains the best result at 156, which means there are 156 features were selected here, with Acc is 77.55%, Mcc is 0.5883, Sn is 96.41%, Sp is 71.26%, Precision is 52.79% which isn't the best performance unfortunately, Recall is 96.41%, F-measure is 68.23% and G-mean is 82.89%. These obtained results are better than anyone of **Table 1**.

The performance of iAcet-PseFDA was also depicted with ROC curves shown in **Figure 5** in which the graphic lines are represent for GO, Subcellular localization and other Domain notations' KNNScores along with PseAAC's. As shown in first subfigure of **Figure 5**, the proposed model's AUC value is 0.8280 while those of PseAAC, GO, PFAM are 0.7521, 0.8146, 0.7548, respectively. Thus the proposed model obtained best result of the four methods. With similar analysis depicted in the last two subfigures of **Figure 5**, the AUC values of SMART, PROSITE, SUPFAM, INTERPRO, PRINTS, and Subcellular localization KNNScores are 0.7453, 0.7538, 0.7614, 0.7611, 0.7144, and 0.8087, respectively. In conclusion, all of the values are <0.8280, and there still are gaps between them and that of the proposed model. It shows that the feature set enhanced with Relief would obtain more satisfactory results than those of the independent FDA features.

CONCLUSION

In order to detect acetylation proteins, this study developed a method on the basis of Random Forest algorithm and Relief. Our approach considered information of sequence conservation extracted by PSI-BLAST besides with PseACC. The involved features are extracted from the sequence conservation information and "GO," "Pfam," "Smart," "PROSITE," "SUPFAM," "InterPro," "PRINTS" and Subcellular localization information of the given query amino acid sequence. This work may cope with the expensive and time-consuming process of identifying acetylation proteins because that the features only incorporated the sequence conservation via gray system model and Knn scores based on FDA databases. All of these processes only need computational model instead of any physical chemistry experiment.

Also, our result manifested that it appears that using FDAs is essential for the prediction of acetylation functional class, which had been reported in previous research (Qiu et al., 2016a,b, 2017b), and the information related to subcellular is also important for identifying the PTM proteins. As the growing demand of verification of acetylation sites, we argue that more effort should be input in developing organism-specific predictors for this issue. The reason for presenting the model here then is for the improving the predictor used in similar research, and it may be helpful for those researchers who would like to deal with bioinformatics problems with computational models. In addition, the involved features may provide important clues of the acetylation mechanism and guide the related experimental validations.

Additionally, a web-server has been established at http://www.jci-bioinfo.cn/iAcetyP which is user-friendly and convenient for the researchers who are working in distinguishing acetylated proteins from non-acetylated proteins.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.uniprot.org/.

AUTHOR CONTRIBUTIONS

AX and Z-CX carried out the extraction of annotation features, model construction, model training, and evaluation, also drafted the related initial manuscript version. C-HZ carried out the preparation of the data, GO enrichment, subcellular localization analysis, comparison with deep learning tool, and drafted the subsequent manuscript. XX led this project and guided this work and shared his idea to implement and discussion. All authors involved in discussion and conclusion section.

FUNDING

This work was supported by the grants from the National Natural Science Foundation of China (31860312, 31760315, 31560316), the Natural Science Foundation of Jiangxi Province, China (No. 20171ACB20023), and the Scientific Research plan of the Department of Education of Jiangxi Province (GJJ180703, GJJ180733).

REFERENCES

Agapito, G., Milano, M., Guzzi, P. H., and Cannataro, M. (2016). Extracting cross-ontology weighted association rules from gene ontology annotations. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 13, 197–208. doi: 10.1109/TCBB.2015.2462348

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and methylation of histones and their possible role in the regulation of rna synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 51, 786–794. doi: 10.1073/pnas.51.5.786

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., et al. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford).* 2012:bas019. doi: 10.1093/database/bas019

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D., and Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* 27, 260–262. doi: 10.1093/nar/27.1.260

Beauclair, G., Bridier-Nahmias, A., Zagury, J. F., Saïb, A., and Zamborlini, A. (2015). JASSA: a comprehensive tool for prediction of SUMOylation sites and SIMs. *Bioinformatics* 31, 3483–3491. doi: 10.1093/bioinformatics/btv403

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 42, 1387–1395. doi: 10.1007/s00726-011-0835-0

Chen, X., Shi, S. P., Suo, S. B., Xu, H. D., and Qiu, J. D. (2015). Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity. *Bioinformatics* 31, 194–200. doi: 10.1093/bioinformatics/btu598

Chen, Z., Chen, Y. Z., Wang, X. F., Wang, C., Yan, R. X., and Zhang, Z. (2011). Prediction of ubiquitination sites by using the composition of k-spaced amino acid Pairs. *PLoS ONE* 6:e22930. doi: 10.1371/journal.pone.0022930

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Du, P., Tian, Y., and Yan, Y. (2012). Subcellular localization prediction for human internal and organelle membrane proteins with projected gene ontology scores. *J. Theor. Biol.* 313, 61–67. doi: 10.1016/j.jtbi.2012.08.016

Gao, J., Thelen, J. J., Dunker, A. K., and Xu, D. (2010). Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol. Cell. Proteomics* 9, 2586–2600. doi: 10.1074/mcp.M110.001388

Haglund, K., and Dikic, I. (2005). Ubiquitylation and cell signaling. *EMBO J.* 24, 3353–3359. doi: 10.1038/sj.emboj.7600808

Han, Z. J., Feng, Y. H., Gu, B. H., Li, Y. M., and Chen, H. (2018). The post-translational modification, SUMOylation, and cancer (Review). *Int. J. Oncol.* 52, 1081–1094. doi: 10.3892/ijo.2018.4280

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32, D258–D261. doi: 10.1093/nar/gkh036

Hershko, A., and Ciechanover, A. (1998). The ubiquitin system. *Annu. Rev. Biochem.* 67, 425–479. doi: 10.1146/annurev.biochem.67.1.425

Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., et al. (2014). LAceP: lysine acetylation site prediction using logistic regression classifiers. *PLoS ONE* 9:e89575. doi: 10.1371/journal.pone.0089575

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

Ingrell, C. R., Miller, M. L., Jensen, O. N., and Blom, N. (2007). NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics.* 23, 895–897. doi: 10.1093/bioinformatics/btm020

Inoue, A., and Fujimoto, D. (1969). Enzymatic deacetylation of histone. *Biochem. Biophys. Res. Commun.* 36, 146–150. doi: 10.1016/0006-291X(69)90661-5

Jiao, Y. S., and Du, P. F. (2017). Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.* 416, 81–87. doi: 10.1016/j.jtbi.2016.12.026

Jones, D. T. (2007). Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics* 23, 538–544. doi: 10.1093/bioinformatics/btl677

Kaur, H., and Raghava, G. P. (2004). A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* 20, 2751–2758. doi: 10.1093/bioinformatics/bth322

Kira, K., and Rendell, L. A. (1992). "The feature selection problem: traditional methods and a new algorithm," in *Tenth National Conference on Artificial Intelligence* (San Jose, CA: San Jose Convention Center).

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., et al. (2004). SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32, D142–D144. doi: 10.1093/nar/gkh088

Li, C., Choi, H. P., Wang, X., Wu, F., Chen, X., Lü, X., et al. (2017). Post-translational modification of human histone by wide tolerance of acetylation. *Cells* 6:34. doi: 10.3390/cells6040034

Li, Z., Wu, P., Zhao, Y., Liu, Z., and Zhao, W. (2015). Prediction of serine/threonine phosphorylation sites in bacteria proteins. *Adv. Exp. Med. Biol.* 827, 275–285. doi: 10.1007/978-94-017-9245-5_16

Mondal, S., and Pai, P. P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.* 356, 30–35. doi: 10.1016/j.jtbi.2014.04.006

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36. doi: 10.1016/S0968-0004(98)01336-X

Pandit, S. B., Bhadra, R., Gowri, V. S., Balaji, S., Anand, B., and Srinivasan, N. (2004). SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics* 5:28. doi: 10.1186/1471-2105-5-28

Peng, J., Wang, T., Wang, J., Wang, Y., and Chen, J. (2016). Extending gene ontology with gene association networks. *Bioinformatics* 32, 1185–1194. doi: 10.1093/bioinformatics/btv712

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, D., and Chou, K. C. (2017a). iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. *Mol. Inform.* 36:1600010. doi: 10.1002/minf.201600010

Qiu, W. R., Sun, B. Q., Xiao, X., Xu, Z. C., and Chou, K. C. (2016b). iPTM-mLys: identifying multiple lysine PTM sites and their different types. *Bioinformatics* 32, 3116–3123. doi: 10.1093/bioinformatics/btw380

Qiu, W. R., Xiao, X., Xu, Z. C., and Chou, K. C. (2016a). iPhos-PseEn: identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier. *Oncotarget* 7, 51270–51283. doi: 10.18632/oncotarget.9987

Qiu, W. R., Zheng, Q. S., Sun, B. Q., and Xiao, X. (2017b). Multi-iPPseEvo: a multi-label classifier for identifying human phosphorylated proteins by incorporating evolutionary information into Chou's general PseAAC via grey system theory. *Mol. Inform.* 36:1600085. doi: 10.1002/minf.201600085

Radivojac, P., Vacic, V., Haynes, C., Cocklin, R. R., Mohan, A., Heyen, J. W., et al. (2010). Identification, analysis, and prediction of protein ubiquitination sites. *Proteins* 78, 365–380. doi: 10.1002/prot.22555

Shen, H. B., and Chou, K. C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Anal. Biochem.* 373, 386–388. doi: 10.1016/j.ab.2007.10.012

Shi, S. P., Qiu, J. D., Sun, X. Y., Suo, S. B., Huang, S. Y., and Liang, R. P. (2012). A method to distinguish between lysine acetylation and lysine methylation from protein sequences. *J. Theor. Biol.* 310, 223–230. doi: 10.1016/j.jtbi.2012.06.030

Sigrist, C. J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., et al. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–D166. doi: 10.1093/nar/gkp885

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

Trost, B., Napper, S., and Kusalik, A. (2015). Case study: using sequence homology to identify putative phosphorylation sites in an evolutionarily distant species (honeybee). *Brief. Bioinform.* 16, 820–829. doi: 10.1093/bib/bbu040

Tung, C. W., and Ho, S. Y. (2008). Computational identification of ubiquitylation sites from protein sequences. *BMC Bioinform.* 9:310. doi: 10.1186/1471-2105-9-310

Wang, L., Du, Y., Lu, M., and Li, T. (2012). ASEB: a web server for KAT-specific acetylation site prediction. *Nucleic Acids Res.* 40, W376–W379. doi: 10.1093/nar/gks437

Wuyun, Q., Zheng, W., Zhang, Y., Ruan, J., and Hu, G. (2016). Improved species-specific lysine acetylation site prediction based on a large variety of features set. *PLoS ONE* 11:e0155370. doi: 10.1371/journal.pone.0155370

Xu, X., Li, A., and Wang, M. (2015). Prediction of human disease-associated phosphorylation sites with combined feature selection approach and support vector machine. *IET Syst. Biol.* 9, 155–163. doi: 10.1049/iet-syb.2014.0051

Xu, Y., Ding, Y. X., Deng, N. Y., and Liu, L. M. (2016). Prediction of sumoylation sites in proteins using linear discriminant analysis. *Gene* 576(1 Pt 1), 99–104. doi: 10.1016/j.gene.2015.09.072

Yang, Y., Tong, M., Bai, X., Liu, X., Cai, X., Luo, X., et al. (2017). Comprehensive proteomic analysis of lysine acetylation in the foodborne pathogen *Trichinella spiralis*. *Front. Microbiol.* 8:2674. doi: 10.3389/fmicb.2017.02674

Yao, Q., Gao, J., Bollinger, C., Thelen, J. J., and Xu, D. (2012). Predicting and analyzing protein phosphorylation sites in plants using musite. *Front. Plant Sci.* 3:186. doi: 10.3389/fpls.2012.00186

Yao, Q., Schulze, W. X., and Xu, D. (2015). Phosphorylation site prediction in plants. *Methods Mol. Biol.* 1306, 217–228. doi: 10.1007/978-1-4939-2648-0_17

Zhang, Z. H., Wang, Z. H., Zhang, Z. R., and Wang, Y. X. (2006). A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* 580, 6169–6174. doi: 10.1016/j.febslet.2006.10.017

Zhao, C., Liu, H., Li, J., Deng, Y., and Shi, T. (2010). Nucleosome structure incorporated histone acetylation site prediction in Arabidopsis thaliana. *BMC Genom.* 11(Suppl 2):S7. doi: 10.1186/1471-2164-11-S2-S7

Zhao, X., Li, X., Ma, Z., and Yin, M. (2011). Prediction of lysine ubiquitylation with ensemble classifier and feature selection. *Int. J. Mol. Sci.* 12, 8347–8361. doi: 10.3390/ijms12128347

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a shared affiliation, though no other collaboration, with one of the authors W-RQ.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 19 December 2019
doi: 10.3389/fbioe.2019.00407

Copy Number Variation Pattern for Discriminating MACROD2 States of Colorectal Cancer Subtypes

ShiQi Zhang[1,2†], XiaoYong Pan[3†], Tao Zeng[4†], Wei Guo[5], Zijun Gan[6], Yu-Hang Zhang[6], Lei Chen[7,8], YunHua Zhang[9], Tao Huang[6] and Yu-Dong Cai[1*]*

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark, [3] Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai, China, [4] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, [5] Institute of Health Sciences, Chinese Academy of Sciences, Shanghai Jiao Tong University School of Medicine and Shanghai Institutes for Biological Sciences, Shanghai, China, [6] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [7] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [8] Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai, China, [9] Anhui Province Key Laboratory of Farmland Ecological Conservation and Pollution Prevention, School of Resources and Environment, Anhui Agricultural University, Hefei, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Qinghua Jiang,
Harbin Institute of Technology, China
Ran Su,
Tianjin University, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 21 October 2019
Accepted: 27 November 2019
Published: 19 December 2019

Citation:
Zhang S, Pan X, Zeng T, Guo W, Gan Z, Zhang Y-H, Chen L, Zhang Y, Huang T and Cai Y-D (2019) Copy Number Variation Pattern for Discriminating MACROD2 States of Colorectal Cancer Subtypes. Front. Bioeng. Biotechnol. 7:407. doi: 10.3389/fbioe.2019.00407

Copy number variation (CNV) is a common structural variation pattern of DNA, and it features a higher mutation rate than single-nucleotide polymorphisms (SNPs) and affects a larger fragment of genomes. CNV is related with the genesis of complex diseases and can thus be used as a strategy to identify novel cancer-predisposing markers or mechanisms. In particular, the frequent deletions of mono-ADP-ribosylhydrolase 2 (*MACROD2*) locus in human colorectal cancer (CRC) alters DNA repair and the sensitivity to DNA damage and results in chromosomal instability. The relationship between CNV and cancer has not been explained. In this study, on the basis of the genome variation profiling by the SNP array from 651 CRC primary tumors, we computationally analyzed the CNV data to select crucial SNP sites with the most relevance to three different states of *MACROD2* (heterozygous deletion, homozygous deletion, and normal state), suggesting that these CNVs may play functional roles in CRC tumorigenesis. Our study can shed new insights into the genesis of cancer based on CNV, providing reference for clinical diagnosis, and treatment prognosis of CRC.

Keywords: copy number variation, MACROD2, colorectal cancer, subtype, classification

INTRODUCTION

Copy number variation (CNV) is a common structural variation pattern of DNA; it is defined as a >1 kb genomic segment with a different copy number compared with the reference genome, leading to gains, or losses of multiple DNA sites that are either microscopic or submicroscopic (Redon et al., 2006). CNV features a higher mutation rate than single-nucleotide polymorphisms (SNPs) and affects a larger fragment of genomes (Zhang et al., 2009). For a large number of CNVs generated in the human genome, one of the known mechanisms is DNA recombination, which includes non-allelic homologous recombination and non-homologous end-joining. Recently, a new mechanism based on DNA error replication has been discovered. Named the "Fork stalling and switching" model, this mechanism can explain complex-structure CNVs that do not conform to non-allelic homologous recombination or non-homologous end-joining.

With the development of high-resolution SNP arrays, identifying large-scale CNVs in thousands of samples has been possible (Beroukhim et al., 2010). Studies have demonstrated that CNV is related to the genesis of Mendelian diseases, sporadic diseases, and susceptibility to complex diseases (Yang et al., 2008; De Cid et al., 2009; Willer et al., 2009; Sato et al., 2014; Zhang et al., 2014). CNVs also play a potential role in cancer risk, and the genome-wide copy number analysis can be used as a strategy to identify novel cancer-predisposing markers or mechanisms (Kuiper et al., 2010). Ding et al. (2010) reported that the genome of primary tumors is diverse and frequently includes gene rearrangements and copy number variations. Shlien et al. (2008) used high-density oligonucleotide arrays to compare the genomes of healthy population and a Li–Fraumeni cancer predisposition disorder (LFS) cohort and observed that CNV in the cell adhesion gene mixed-lineage leukemia translocated 4 (*MLLT4*) is associated with LFS, in which patients always harbor a germline heterozygous mutation of the tumor suppressor gene *TP53* and experience a high probability of developing early-stage breast, sarcoma, brain, and other tumors. Scrima et al. (2012) revealed that 24, 31, and 26% of patients with lung adenocarcinoma achieved a copy number gain in adenylate kinase (*AK*) 1, *AK2*, and phosphoinositide-3-kinase, catalytic, alpha polypeptide (*PI3KCA*), respectively, via fluorescence *in situ* hybridization.

Evidence has recognized CNV as one of the most important genomic alterations affecting cancer pathogenesis (Hermsen et al., 2002), whereas chromosomal instability and allelic imbalance at certain chromosomal loci play crucial roles in most sporadic cases of colorectal cancer (CRC) (Zanke et al., 2007). CRC is the fourth most common cancer and the second leading cause of cancer death worldwide, with over 1.1 million new cancer cases and 880,000 deaths estimated in 2018 (Bray et al., 2018). For better assessment of the progression of CRC, the Dukes staging system was proposed as a common classification system for CRC (Dukes, 1932). Four stages of CRC are defined by such system depended on the degree of colorectal pathology. Dukes A represents the invasion of tumor cells into but not through the bowel wall. Patients in Dukes A stage usually have better outcomes with over 90% 5-year survival. When tumor grows through the muscle layer of the bowel but not infiltrate into lymph nodes, it will be identified as Dukes B stage. Dukes C refers to the spread of cancer to at least one lymph node close to the bowel. And lastly, widespread metastases of tumor cells in CRC, also called advanced CRC, indicate the stage of Dukes D. The clear stage of CRC contributes to the decision making in clinical treatment, and also provides a detailed description for the pathology research.

Frequent deletions of the mono-ADP-ribosylhydrolase 2 *(MACROD2*) locus in human CRC alter DNA repair and sensitivity to DNA damage and result in chromosomal instability (Sakthianandeswaren et al., 2018). In addition, *MACROD2* deletion in CRC is significantly associated with the extent of malignancy, indicating that *MACROD2* acts as a haploin-sufficient tumor suppressor, with the loss of function promoting chromosome instability and thereby driving cancer evolution.

In this study, based on the genomic variation profiling by SNP array from 651 CRC primary tumors (Sakthianandeswaren et al., 2018), the log R ratio (LRR) and B allele frequency data (BAF) of each SNP site were exported using two types of hybridization probes specific to two types of known alleles (Wang et al., 2007), and the SNP genotype also can be determined by the ratios of the hybridization intensities of two types of probes. The genotype of SNPs located in the region of *MACROD2* was used to represent the genotype state of *MACROD2*, which means that the individuals with the loss of both alleles in at least one SNP site in *MACROD2* will be classified into the state of homozygous deletion, and the deletion of only one allele indicates the heterozygous deletion status. A wild-type stage or normal stage refers to no deletion happened in *MACROD2*. Following that, each patient was classified into one of the three states: heterozygous deletion, homozygous deletion, and normal state in our study. We computationally analyzed the CNV data to select the crucial SNP sites showing the most relevance to the four Dukes stages of CRC (A, B, C, and D) and three different states of *MACROD2* (heterozygous deletion, homozygous deletion, and normal state), suggesting that these CNVs may play functional roles in CRC tumorigenesis. We constructed a classifier with high accuracy to group individuals into the corresponding state categories. This classification model also provides a meaningful list of genomic loci that perform important functions in the development and progression of cancers. To date, the relationship between CNV and cancer has not been exactly explained. Our study can shed new light on the genesis of cancer based on CNV, providing reference for the clinical diagnosis and treatment prognosis of CRC.

MATERIALS AND METHODS

In this study, we first used the minimum redundancy and maximum relevance (mRMR) method (Peng et al., 2005) to analyze all features. Irrelevant features were discarded and the rest features were ranked in a feature list, which was further fed into the incremental feature selection (IFS) (Liu and Setiono, 1998) to obtain the optimum features and extract the classification rules for readable explanation. We adopted the same computational pipeline to separately analyze four kinds of carefully organized datasets, including the CRC stage with LRR or BAF and the *MACROD2* status with LRR or BAF.

Datasets

The LRR and BAF data on 651 CRC primary tumors obtained using the Illumina Human610-Quad v1.0 BeadChip were downloaded from Gene Expression Omnibus under the accession number GSE115145 (Sakthianandeswaren et al., 2018). The LRR and BAF were calculated with GenomeStudio (Illumina). The 651 CRC samples can be divided into four stages: 60 stage A samples, 208 stage B samples, 297 stage C samples, and 86 stage D samples. Based on *MACROD2* status, 441 wild-type samples, 137 heterozygous deletion samples, and 73 homozygous deletion samples were obtained. Each sample was represented by 620,901 SNP features.

Feature Selection

As mentioned above, each sample was represented by lots of SNP features. Clearly, not all of them were highly related to classification of these samples. Thus, we employed some powerful feature selection methods to analyze all features. The analysis procedures included three stages. The first stage was to exclude irrelevant features; the second one was to sort rest features; the last stage was to construct optimal classifier with optimum features and classification rules with the help of IFS method, support vector machine (SVM) (Corinna Cortes, 1995), and repeated incremental pruning to produce error reduction (RIPPER) (Cohen, 1995).

The purpose of the first stage was to exclude irrelevant features. To this end, all features were evaluated by the mRMR method. The mRMR method was a mutual information (MI)-based feature selection method (Peng et al., 2005; Li et al., 2019). The importance of each feature was evaluated by its MI to class labels. It is clear that the higher the MI values were, the more important the features were. After a threshold for MI value was set, irrelevant features can be excluded.

After irrelevant features were excluded, rest features were assessed by mRMR method in another way in the second stage. In detail, rest features were ranked in a feature list in terms of their relevance to class labels and redundancies to other features. The feature subset consisting of some top features in the list can be deemed to be the optimal feature combination with highest relevance to class labels and lowest redundancies among these features, which can provide a powerful discrimination. In this study, we used the mRMR program downloaded from http://home.penglab.com/proj/mRMR/index.htm. Default parameters were adopted.

In the third stage, we ran a two-stage IFS with a classification algorithm to select the optimum features for building the optimal classifier or construct classification rules. In the first stage, a series of feature subsets with a step 10 was generated, where feature subset 1 consists of the top 10 features, feature subset 2 consists of the top 20 features, and so on. Then, for each feature subset, a classifier was trained on the samples consisting of the features from the feature subset, and this classifier was evaluated using 10-fold cross-validation (Kohavi, 1995). An interval [min, max] with a good performance was determined. In the second stage, a series of feature subsets within the interval [min, max] was generated to further select the final optimum features or construct classification rules. Based on these optimum features, an optimal classifier can be built.

SVM

SVM attempts to identify a hyper plane with a maximum margin between two groups of samples, and it has been widely used in biological data studies (Pan and Shen, 2009; Mirza et al., 2015; Cai et al., 2018; Chen et al., 2018, 2019; Zhou et al., 2019). In this work, we used a multi-class SVM with a one vs. rest strategy. The multi-class SVM consists of multiple binary SVMs, and each SVM classifies the samples of one class from the rest of the classes. When predicting the class for a new sample, the SVM predicts the sample's label corresponding to the class with the highest probability. This study adopted the SVM implemented by a tool "SMO" in Weka.

Rule Learning

To understand how a classification model makes a prediction, we used rule learning to extract the readable classification rules. A rule consists of an IF-THEN relationship between features and output labels, such as IF SNP1 <= 0.7 AND SNP2 >= 1.02; THEN stage = "A." In this study, we applied RIPPER (Cohen, 1995), which is implemented by a tool "JRip" in Weka. RIPPER consists of two stages, including the rule building stage and rule optimization stage.

SMOTE

As mentioned in the *Datasets* section, 651 CRC samples were classified into three or four classes. The sizes of classes varied a lot. Thus, investigated datasets were imbalanced. For this type of dataset, the performance of an ordinary classifier is dependent on the biggest class. To tackle this problem, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002; Wang et al., 2018; Zhang et al., 2019) was employed in this study, which is a oversampling method. This method can produce some new samples and pour into minority class, thereby making all classes having equal sizes. In this study, for the BAF/LRR dataset of CRC stage, new samples were generated by SMOTE for classes of stages A, B, and D, while new samples were yielded by SMOTE for classes of heterozygous deletion and homozygous deletion for BAF/LRR dataset of *MACROD2* status.

In this study, we adopted the SMOTE program implemented by python, which was downloaded at https://github.com/scikit-learn-contrib/imbalanced-learn.

RESULTS

In this study, we separately analyzed the four kinds of carefully organized datasets with a three-stage feature selection method. Whole procedures are illustrated in **Figure 1**.

For the first stage, we set the threshold of MI values to be 0.01; i.e., features receiving the MI values larger than 0.01 were kept. The number of remaining features for BAF/LRR dataset of CRC stage was 47515/44931, while it was 20839/20973 for BAF/LRR dataset of *MACROD2* status. Then, in the second stage, remaining features in each dataset were ranked by the mRMR method. Obtained feature lists are provided in **Tables S1–S4**. The third stage employed the IFS method and classification algorithms to extract optimum features and construct classification rules. The key results are provided in **Tables 1–4**.

Results on BAF Dataset of CRC Stage

We first ran the computational pipeline on the first BAF dataset of CRC stage. Key results are provided in **Table 1** and **Figure 2**. For the first stage of IFS with a step 10, results are provided in **Table S5** and a curve with Matthews correlation coefficient (MCC) (Matthews, 1975; Gorodkin, 2004; Zhao et al., 2018, 2019; Cui and Chen, 2019) as Y-axis and number of features as X-axis was plot, as shown in **Figure 3A**. The SVM yielded the highest

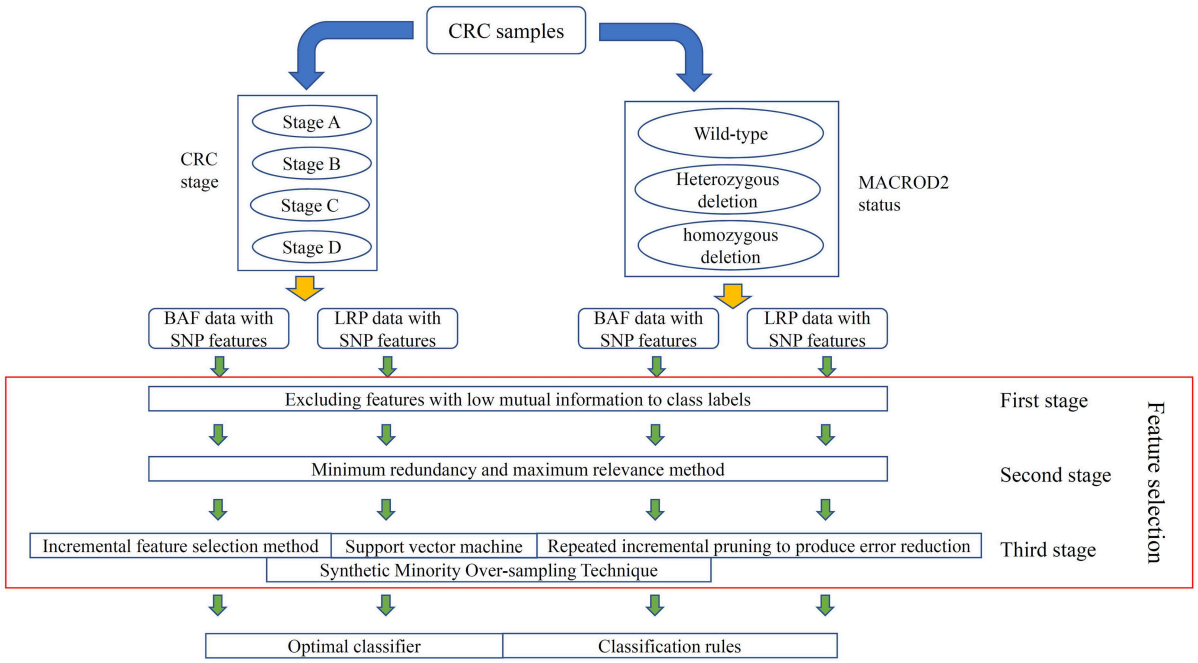


FIGURE 1 | Entire procedures to analyze the log R ratio (LRR) and B allele frequency (BAF) data on colorectal cancer (CRC) primary tumor samples. CRC samples are classified into four stages; at the same time, they can also be classified into three classes according to their MACROD2 status. For each classification, two datasets with LRR and BAF, respectively, were constructed. Four datasets were obtained in total, in which single-nucleotide polymorphism (SNP) features were used to represent each CRC sample. A feature selection procedure, including three stages, was adopted to analyze all SNP features. Finally, an optimal classifier and several classification rules were accessed for each dataset.

TABLE 1 | Performance of classification models on BAF dataset of CRC stage with IFS method.

Classifier	1st-stage IFS*		2nd-stage IFS*	Number of rules
	Highest point	Turning point		
SVM	0.9653 (35,440)	0.9007 (8,790)	0.9008 (8,797)	—
RIPPER	0.2932 (8,500)	0.2692 (2,170)	0.2745 (2,075)	30

**These performances are measured by MCC; numbers of used features are listed in brackets.*
BAF, B allele frequency; CRC, colorectal cancer; IFS, incremental feature selection; SVM, support vector machine; RIPPER, repeated incremental pruning to produce error reduction; MCC, Matthews correlation coefficient.

TABLE 2 | Performance of classification models on LRR dataset of CRC stage with IFS method.

Classifier	1st-stage IFS*		2nd-stage IFS*	Number of rules
	Highest point	Turning point		
SVM	0.7542 (20,400)	0.7143 (3,960)	0.7231 (3,967)	—
RIPPER	0.3420 (18,530)	0.3417 (3,040)	0.3490 (2,841)	32

**These performances are measured by MCC; numbers of used features are listed in brackets.*
LRR, log R ratio.

TABLE 3 | Performance of classification models on BAF dataset of MACROD2 status with IFS method.

Classifier	1st-stage IFS*		2nd-stage IFS*	Number of rules
	Highest point	Turning point		
SVM	0.9683 (5,610)	0.9406 (2,080)	0.9436 (2,064)	—
RIPPER	0.3923 (18,460)	0.3677 (5,530)	0.3677 (5,530)	23

**These performances are measured by MCC; numbers of used features are listed in brackets.*

TABLE 4 | Performance of classification models on LRR dataset of MACROD2 status with IFS method.

Classifier	1st-stage IFS*		2nd-stage IFS*	Number of rules
	Highest point	Turning point		
SVM	0.9069 (5,540)	0.8759 (1,030)	0.8785 (1,022)	—
RIPPER	0.6953 (410)	—	0.7385 (306)	17

**These performances are measured by MCC; numbers of used features are listed in brackets.*

MCC value of 0.9653 (**Table 1**) when the top 35,440 features were used. Considering this extremely large number, we used another turning point (top 8,790 features), which still yielded a high MCC value of 0.9007. Thus, in the second IFS stage, we ran the same pipeline with the interval [1, 8800] with a step 1. Results are collected in **Table S6**, and a curve was also plotted, as shown in **Figure 3B**. The best MCC value was 0.9008 when the

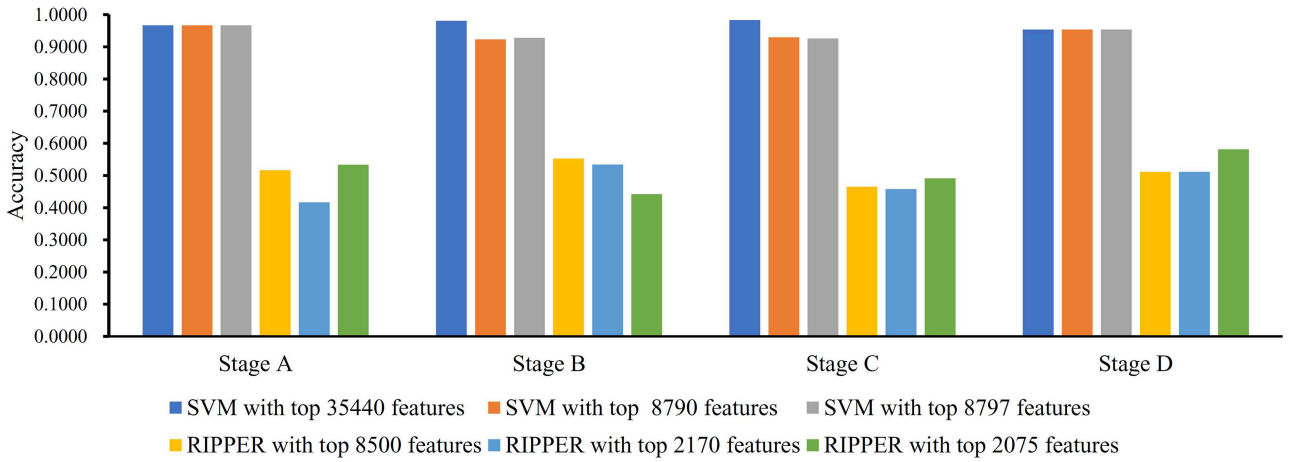


FIGURE 2 | Bar chart to show accuracies on four CRC stages yielded by key support vector machine (SVM) and repeated incremental pruning to produce error reduction (RIPPER) classifiers on BAF data of CRC stage.

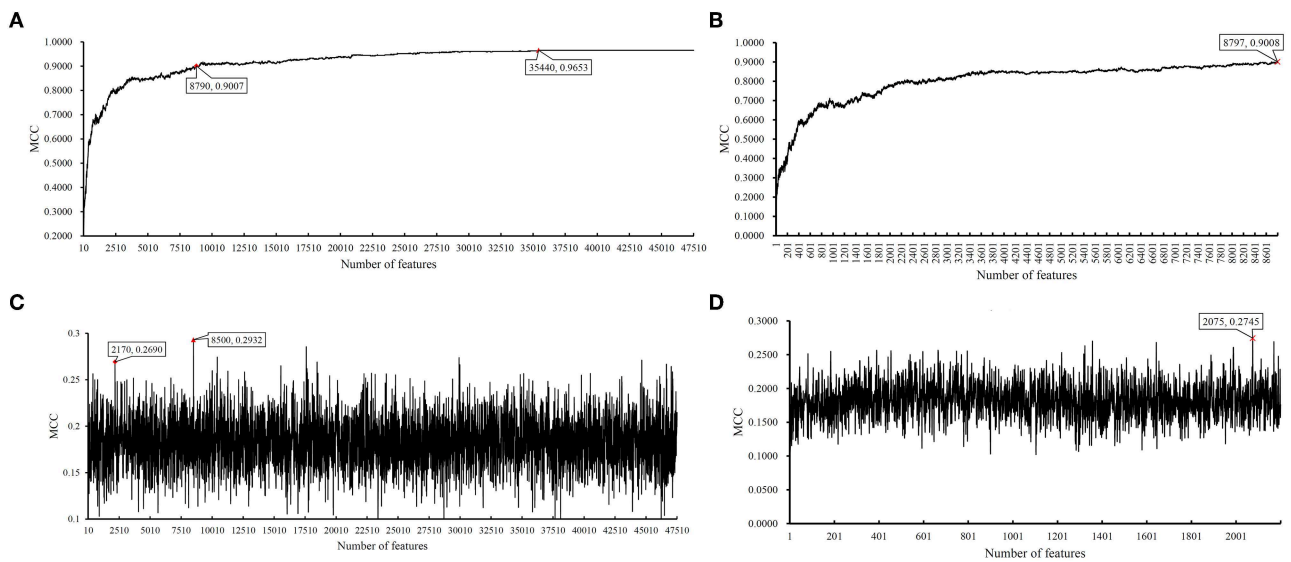


FIGURE 3 | Incremental feature selection (IFS) results on BAF data of CRC stage yielded by SVM and RIPPER. **(A)** First-stage IFS results on BAF data of CRC stage yielded by SVM. **(B)** Second-stage IFS results on BAF data of CRC stage yielded by SVM. **(C)** First-stage IFS results on BAF data of CRC stage yielded by RIPPER. **(D)** Second-stage IFS results on BAF data of CRC stage yielded by RIPPER.

top 8,797 features were used. Accordingly, we built an optimal SVM classifier with the top 8,797 features.

In addition to SVM, we applied the interpretable rule learning method RIPPER to evaluate the selected features' performance in a rule manner. After running RIPPER on the samples consisting of features from individual feature subsets with a step 10, we obtained the performance of RIPPER on different feature subsets, as shown in **Table S5** and **Figure 3C**. We obtained the best MCC value of 0.2932 when the top 8,500 features were used. A turning point was observed (top 2,170 features), yielding an MCC value of 0.2692. To further select the optimum features, we ran the IFS with RIPPER within the interval [1, 2,200]. Results are available in **Table S6** and displayed in **Figure 3D**. We obtained the best MCC value of 0.2745 when the top 2,075 features were used.

Although RIPPER showed a poorer performance than SVM in this case, one advantage of RIPPER is that it can generate classification rules, which help us understand how the model makes a prediction on a subgroup of samples. Considering these data, the RIPPER produced 30 classification rules, which are given in **Table S7**.

Results on LRR Dataset of CRC Stage

We ran the above same pipeline on the second dataset. Key results are provided in **Table 2** and **Figure 4**. When running the

IFS with an SVM on the samples consisting of features from individual feature subsets, we obtained the best MCC value of 0.7542 when the top 20,400 features were used. We adopted a smaller turning value (top 3,960 features), which yielded an MCC value of 0.7143. Then, we ran the second stage of IFS on the interval [1, 4000] and obtained the best MCC value of 0.7231 when the top 3,967 features were used. These results are given in **Tables S8**, **S9** and illustrated in **Figures 5A,B**. Accordingly, an optimal SVM classifier was built based on the top 3,967 features.

Similarly, IFS with RIPPER was also used on this dataset. All results are provided in **Tables S8**, **S9** and displayed in **Figures 5C,D**. We obtained the best MCC value of 0.3420 when using the top 18,530 features. Of note, when 3,040 features were used, the performance showed a notable change as a performance turning point. Thus, in the second stage of IFS, we ran the RIPPER on the interval [1, 3100] and obtained the best MCC value of 0.3490 when using the top 2,841 features. The 32 learned classification rules are given in **Table S10**.

Results on BAF Dataset of MACROD2 Status

Instead of analyzing the association between the CRC stages and CNV states, we used the same pipeline to analyze the *MACROD2* status associated with particular CNV types. For the BAF dataset of *MACROD2* status, key results are provided in **Table 3** and **Figure 6**. Results of the first stage of IFS with SVM are available in

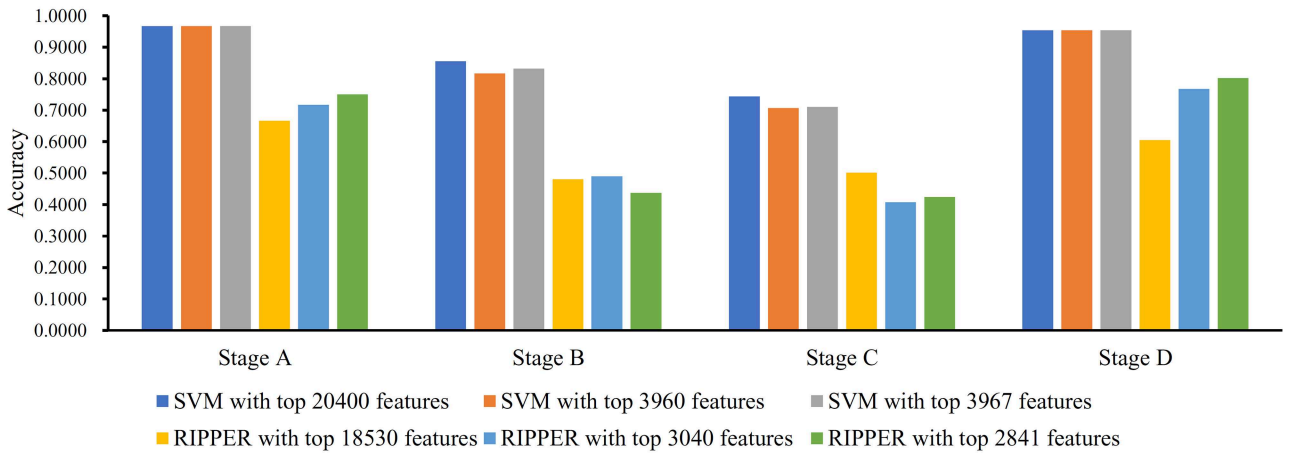


FIGURE 4 | Bar chart to show accuracies on four CRC stages yielded by key SVM and RIPPER classifiers on LRR data of CRC stage.

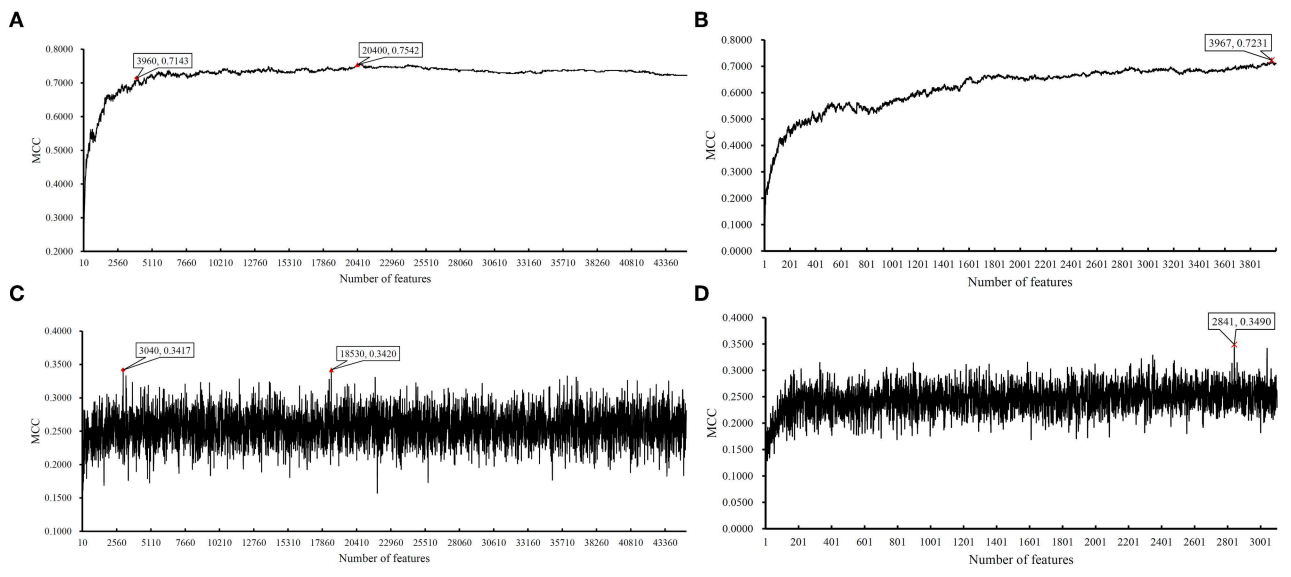


FIGURE 5 | IFS results on LRR data of CRC stage yielded by SVM and RIPPER. **(A)** First-stage IFS results on LRR data of CRC stage yielded by SVM. **(B)** Second-stage IFS results on LRR data of CRC stage yielded by SVM. **(C)** First-stage IFS results on LRR data of CRC stage yielded by RIPPER. **(D)** Second-stage IFS results on LRR data of CRC stage yielded by RIPPER.

Table S11, and a curve was plotted in **Figure 7A**. We obtained the best MCC value of 0.9683 when the top 5,610 features were used. We detected the turning point 2,080, which yielded an MCC value of 0.9406. In the second stage of IFS, we ran the SVM on the interval [1, 2080]. Results are collected in **Table S12**, and a curve was plotted in **Figure 7B**. The best MCC value was 0.9436 when the top 2,064 features were used, which can be used to build an optimal SVM classifier.

We also ran the IFS with RIPPER on this dataset. The first-stage results are provided in **Table S11**. A curve was plotted in **Figure 7C**. RIPPER yielded the best MCC value of 0.3923 when the top 18,460 features were used. We also selected the turning point 5530 for the second stage of IFS, which yielded an MCC value of 0.3677. For the second stage of IFS within the interval [1, 5530], results are available in **Table S12** and a curve was shown in **Figure 7D**. We still obtained the best MCC value of 0.3677 when the top 5,530 features were used. The 23 classification rules generated by RIPPER are listed in **Table S13**.

Results on LRR Dataset of MACROD2 Status

We did the similar procedures for the LRR dataset of *MACROD2* status. Key results are provided in **Table 4** and **Figure 8**. For the first stage of IFS with SVM, results are provided in **Table S14** and

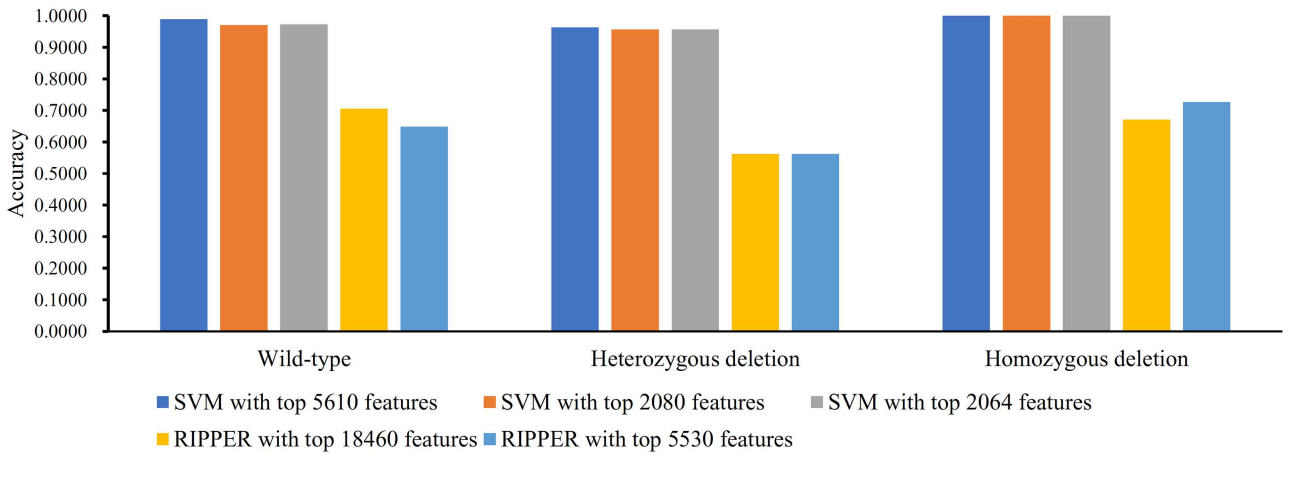


FIGURE 6 | Bar chart to show accuracies on three MACROD2 status yielded by key SVM and RIPPER classifiers on BAF data of MACROD2 status.

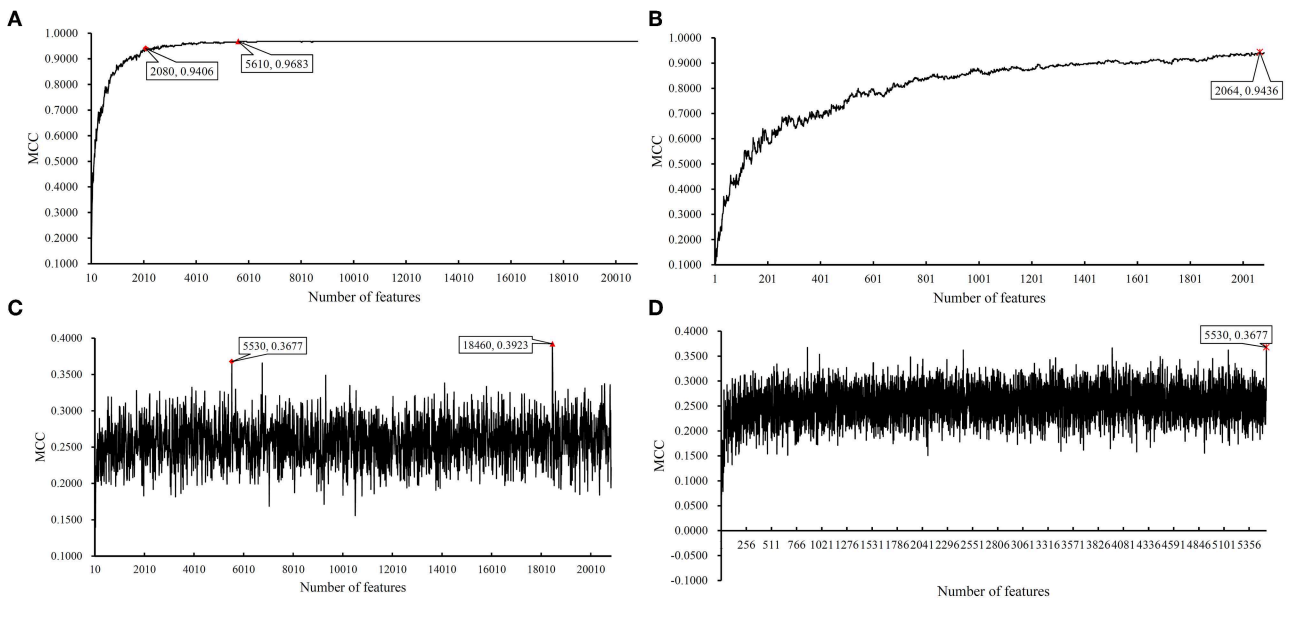


FIGURE 7 | IFS results on BAF data of MACROD2 status yielded by SVM and RIPPER. **(A)** First-stage IFS results on BAF data of MACROD2 status yielded by SVM. **(B)** Second-stage IFS results on BAF data of MACROD2 status yielded by SVM. **(C)** First-stage IFS results on BAF data of MACROD2 status yielded by RIPPER. **(D)** Second-stage IFS results on BAF data of MACROD2 status yielded by RIPPER.

a curve was plotted in **Figure 9A**. We obtained the best MCC value of 0.9069 when using the top 5,540 features. Similarly, a smaller turning point 1,030 was used for the second stage of IFS, because it still yielded a satisfactory MCC value of 0.8759. In the second stage of IFS, we set the interval [1, 1,100]. Results are collected in **Table S15**, and a curve was plotted in **Figure 9B**. We obtained the best MCC value of 0.8785 when the top 1,022 features were adopted. The optimal SVM classifier was built using the top 1,022 features.

We ran the IFS with RIPPER again. Results are provided in **Table S14**. A curve was plotted in **Figure 9C**, from which we can see that the best MCC value was 0.6953 when the top 410 features were used. Then, we ran the second stage of IFS within the interval [1, 410]. Results are available in **Table S15**. A curve was plotted in **Figure 9D**. It can be seen that the best MCC value was 0.7385 when using the top 306 features. **Table 5** lists the 17 classification rules generated by RIPPER.

DISCUSSION

On each of four datasets, a group classification rules were generated by RIPPER. According to the performance of RIPPER listed in **Table 4**, rules on the LRR data of *MACROD2* status were

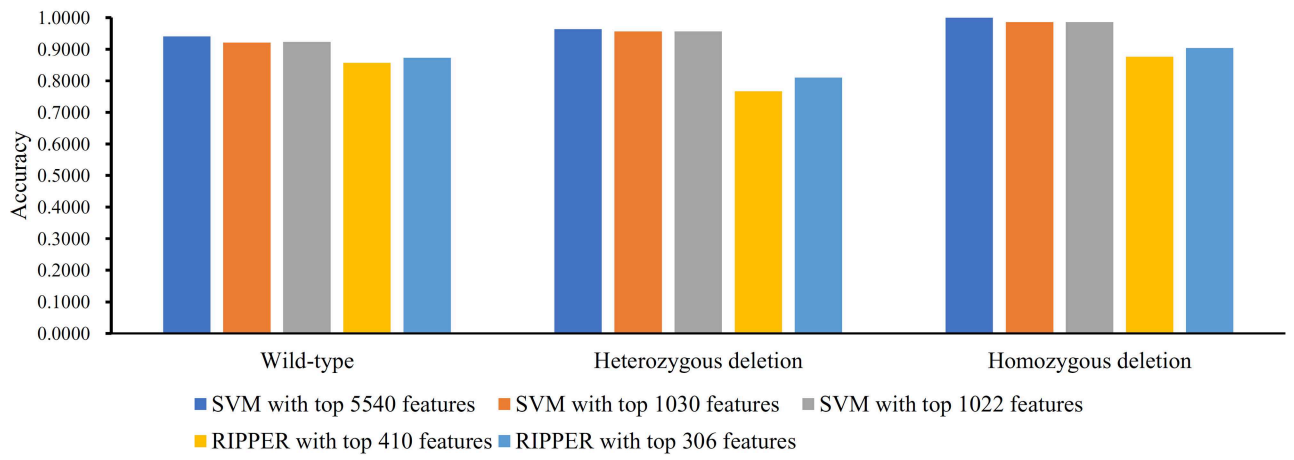


FIGURE 8 | Bar chart to show accuracies on three MACROD2 status yielded by key SVM and RIPPER classifiers on LRR data of MACROD2 status.

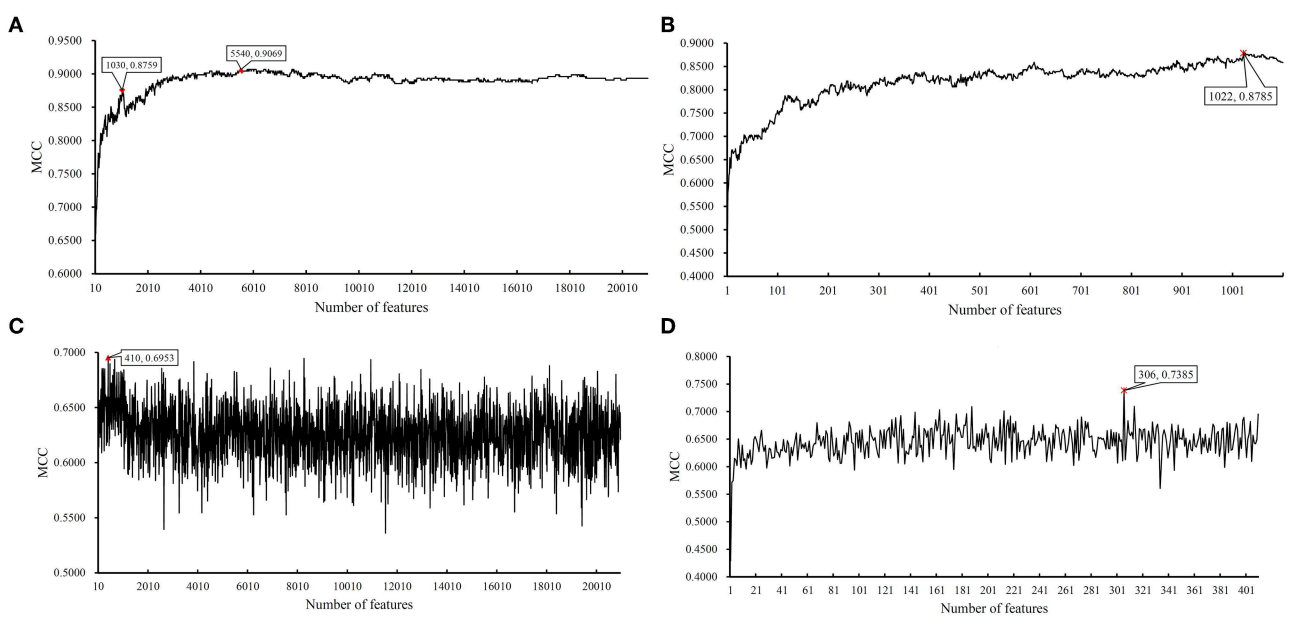


FIGURE 9 | IFS results on LRR data of MACROD2 status yielded by SVM and RIPPER. **(A)** First-stage IFS results on LRR data of MACROD2 status yielded by SVM. **(B)** Second-stage IFS results on LRR data of MACROD2 status yielded by SVM. **(C)** First-stage IFS results on LRR data of MACROD2 status yielded by RIPPER. **(D)** Second-stage IFS results on LRR data of MACROD2 status yielded by RIPPER.

TABLE 5 | Classification rules on dataset of MACROD2 status with LRR.

Index	Condition	Result	Support[a]%	Accuracy[b]%
1	rs353149 <= −0.3811 rs6034087 <= −0.2724	Homozygous deletion	5.84	92.11
2	rs445945 <= −0.3040 rs6712905 >= 0.1367 rs377954 <= −0.3691	Homozygous deletion	3.23	90.48
3	rs6135314 <= −0.5109 rs6685801 <= −0.0057 rs2900712 <= −0.0619 rs9444675 <= 0.1020	Homozygous deletion	4.61	93.33
4	rs6135362 <= −0.2468 rs2100272 >= 0.1398 rs700029 <= 0.0035	Homozygous deletion	1.84	91.67
5	rs6110500 <= −0.2528 rs10500528 <= −0.0094	Homozygous deletion	7.37	83.33
6	rs6079537 <= −0.2319 rs2900712 <= −0.0981 rs6043173 >= −0.0832	Homozygous deletion	0.92	100.00
7	rs9355387 <= −0.2856 rs11905979 <= −0.3878	Homozygous deletion	1.84	91.67
8	rs199305 <= −0.4455 rs377201 >= −0.2189	Homozygous deletion	0.77	100.00
9	rs6135314 >= −0.0746 rs1998086 >= 0.0340 rs381053 >= −0.0576	Wild-type	35.48	98.70
10	rs1475531 >= −0.0454 rs365516 >= 0.0220	Wild-type	31.80	96.14
11	rs2423866 >= −0.1223 rs385770 >= −0.0670 rs7241111 >= −0.1500	Wild-type	24.42	98.11
12	rs449849 >= −0.0559 rs716316 >= −0.0107 rs5904713 >= −0.1428	Wild-type	27.80	97.24
13	rs1327323 <= −0.2719 rs6135269 <= −0.1044	Wild-type	6.76	75.00
14	rs353149 >= −0.0059 rs13011654 >= 0.0742 rs445945 <= 0.067	Wild-type	5.07	96.97
15	rs6034046 >= −0.015 rs6135314 >= −0.0323 rs6034011 <= 0.0668	Wild-type	19.05	95.16
16	rs6043173 >= 0.131 rs449849 >= −0.0689	Wild-type	23.81	94.84
17	Others	Heterozygous deletion	20.28	85.61

[a] *The support of a rule is the percentage of samples satisfying the rule.*
[b] *The accuracy of a rule is the proportion of the corrected classified samples among samples satisfying the rule.*

with the highest performance (MCC = 0.7385). Thus, we mainly discussed these rules, which are listed in **Table 5**. Each rule can cover some CRC samples and give high accuracies.

Given that the status of *MACROD2* is significantly relevant to the intestinal tumorigenesis and plays a crucial role in cancer development (Sakthianandeswaren et al., 2018), our classifiers are expected to be prognostic indicators for evaluating the malignancy of intestinal tumor. On LRR data, 17 decision rules were generated by RIPPER, which can distinguish the three status of *MACROD2* with LRR with a classification accuracy of 0.7385. Depending on the CNV profiles of selected loci, predicting whether a heterozygous, or homozygous depletion of *MACROD2* exists in CRC patients is possible. To validate the reliability of these rules, we examined existing experimental evidence through a literature review.

We focused on the 17 decision rules and a few top-ranked features on data of *MACROD2* status with LRR. Such rules and features described specific CNV characteristics contributing to the identification of *MACROD2* status and CRC classification, indicating their crucial roles in cancer development. Especially, several top-ranked features showed strong biological and biomedical relevance with *MACROD2*, indicating that they also play relevant functions in cancer progression.

Among the 17 rules, 8 rules could identify the homozygous deletion of *MACROD2*, and the other 8 decision rules can identify the normal non-depletion status of *MACROD2*. The last one indicates the heterozygous deletion, which means that if the CNV profiles in patients failed to meet any criteria of the other 16 rules, they were predicted to carry the heterozygous deletion of *MACROD2*.

Rules for Homozygous Deletion

In the eight rules identifying the homozygous deletion of *MACROD2* (see first eight rules in **Table 5**), 21 criteria involving 20 SNP sites were located in different regions of six genes. Notably, 12 of these SNP sites were located in the genomic regions of *MACROD2*, and the LRR of specific regions near these SNP sites featured a low value, which is naturally and logically consistent given that the CNV loss in *MACROD2* leads to homozygous deletion. Thus, our analysis actually highlights the potential core roles of specific SNP sites, suggesting its capability to identify the overall state of *MACROD2* based on the CNV conditions of a few loci. In detail, the 12 SNPs (rs353149, rs6034087, rs445945, rs377954, rs6135314, rs6135362, rs6110500, rs6079537, rs6043173, rs11905979, rs199305, and rs377201) were distributed in different locations of the intron regions of *MACROD2* and displayed strong relevance to the overall status of *MACROD2*. By the detection of CNV in these selected loci markers, we can identify the deletion state of *MACROD2* in patients. We will find the corresponding therapy methods for the treatment targets in the future. Further research about these incompletely elucidated SNP sites may reveal the mechanisms of tumor development at the genomic level. The biological and biomedical significance of several SNPs is summarized below.

The SNP site rs6685801 located in chr1:3547887 required a low value of LRR to identify the homozygous deletion of *MACROD2* in our decision rules. This position is in the intron region of multiple EGF-like-domains 6 (*MEGF6*) gene, which was reported to play a critical role in cell adhesion and involved in many disorders of neural system development (Sunnerhagen et al., 1993). Recent publications have confirmed that *MEFG6* can promote the epithelia-to-mesenchymal transition in CRC metastasis (Hu et al., 2018). This gene is also significantly upregulated in tumor tissue and results in the poor survival of a colon adenocarcinoma cohort. *MEGF6* can also accelerate the cell growth and inhibit apoptosis in CRC as demonstrated by the

experiment *in vitro*. All these results suggest that *MEGF6* may serve as an oncogene, and its overexpression may contribute to the tumorigenesis in CRC patients. We inferred that the copy number loss in this specific intron region caused the upregulated expression of *MEGF6* as it may perform inhibitory effects on transcription. Thus, the low LRR of the SNP site rs6685801 can indicate the severe extent of CRC, consistent with the homozygous deletion state of *MACROD2*.

Another important SNP site rs9444675, which displayed strong relevance to the status of *MACROD2* in our classifier, is located in the intron region of gamma-aminobutyric acid receptor subunit rho-1 (*GABRR1*). *GABRR1*, also called GABA(A) receptor, is a member of the rho subunit family and acts as the receptor of major inhibitory neurotransmitters in the mammalian brain (Cutting et al., 1992). A recent study has shown that *GABRR1* is significantly upregulated by the transcriptome of chemokine (C-X-C motif) ligand 1-(*CXCL1*) treated colon cancer cells (Hsu et al., 2018). Further analysis via bioinformatics methods reported that high expression of *GABRR1* showed a significant correlation with reduced overall survival rates, suggesting the crucial role of *GABRR1* in the progression of colon cancer. In addition, another research reported the upregulation of *GABRR1* in cancer cohorts compared with the controls with regard to gene expression profiles of medullary thyroid carcinoma (Oczko-Wojciechowska et al., 2006). These pieces of evidences support the decision rule that copy number loss of specific region located in *GABRR1* will lead to the upregulation of *GABRR1* and contribute to the carcinogenesis of CRC, resulting in the similar consequence as the homozygous deletion state of *MACROD2*.

One important criterion identified in the decision rules suggests the high value of LRR near the specific SNP site rs2100272. This site is located in the intron regions of *VWA3B*, which showed a tendency toward malignancy development. *VWA3B* encodes an intracellular protein thought to function in transcription, DNA repair, and membrane transport (Kawarai et al., 2016; Huttlin et al., 2017), playing a role similar to *MACROD2,* which was reported to influence DNA repair and sensitivity to DNA damage and result in chromosome instability (Sakthianandeswaren et al., 2018). In the patients of bladder urothelial carcinoma, evident copy number alterations were observed in the 2q12 regions in which the *VWA3B* was mapped (E. Pontes et al., 2013), in line with the suggestion that VWA3B plays a crucial role in bladder carcinogenesis. In addition, *VWA3B* is significantly differentially expressed in tongue squamous cell carcinoma samples at the transcriptome level (Song et al., 2019). These results confirm our decision rules, which indicate that the copy number gain of the specific regions near rs2100272 will alter the expression of VWA3B and contribute to the development of certain cancers including CRC.

Another criterion was found in the experimental findings, and it required a low LRR near the SNP site rs700029 to identify the homozygous deletion state of *MACROD2*. This SNP site is located in chr1:81805339 and was mapped in the intron region of adhesion G protein-coupled receptor L2 (*ADGRL2*), which encodes a member of the latrophilin subfamily of G-protein coupled receptors. *ADGRL2* functions as a p53 target gene and regulator of neuronal exocytosis (Hamann et al., 2015). Recent research has shown the low expression level of ADGRL2 in genomic sequencing analyses of both gastric cancer and colon cancer cell lines due to the hypermethylation of CpG islands within the gene (Jeon et al., 2016). ADGRL2 is also associated with lung squamous cell carcinoma and may serve as the diagnostic marker for *small cell lung cancer* (Huang et al., 2018). The rules that require the copy number loss of specific intron region in ADGRL2 may result in the alteration of expression profile and lead to the development of CRC.

We also identified a critical SNP site rs9355387 located in the intron region of gene Parkin RBR E3 ubiquitin protein ligase (*PRKN*), which according to the rules indicates the homozygous deletion state of *MACROD2*. The gene *PRKN,* best known as *PARK2,* is a key component of a multiprotein E3 ubiquitin ligase complex, which mediates the targeting of substrate proteins for proteasomal degradation. Mutations occurring in this gene cause Parkinson's disease (Oczkowska et al., 2013). The loss of *PRKN* at both the DNA copy number and mRNA expression levels contributes to cancer progression via redox-mediated inactivation of phosphatase and tensin homolog (*PTEN*) (Gupta et al., 2017). The depletion of *PRKN* also enhanced pancreatic tumorigenesis in KRAS-driven engineered mouse models based on its role in mediating the degradation of mitochondrial iron importers (Kang et al., 2019), implying that *PRKN* can be a potential target for pancreatic cancer therapy. These results highlight the crucial role of *PRKN* in cancer progression and confirm our predicted rules, indicating that the loss of copy number near rs9355387 would be an indicator of severe status of cancer.

Rules for Wild-Type

The eight rules for identifying the non-deletion or wild-type status of *MACROD2* included 21 criteria with 19 SNP sites, 15 of which are located in the intron regions of *MACROD2*. The LRR of these specific regions requires a high value opposite that of the homozygous deletion state. Among the 15 SNP sites located in *MACROD2* and with built non-deletion status, 4 SNPs (rs6135314, rs353149, rs445945, and rs6043173) have been applied in the identification of the homozygous deletion state of *MACROD2* with relatively low values as mentioned before. The other 11 SNP sites (rs1998086, rs381053, rs1475531, rs365516, rs2423866, rs385770, rs449849, rs716316, rs6135269, rs6034046, and rs6034011) showed different distributions in varying locations in the intron regions of *MACROD2*, displaying a significant correlation with the overall state of *MACROD2* and implying that these selected loci may play unexplained functional roles in regulating DNA replication. The candidate SNP sites identified by our prediction model can be applied as biomarkers for the pathologic evaluation of CRC, given that the state of *MACROD2* has been confirmed to be a significant signal in intestinal cancers.

The copy number loss of the regions near the SNP site rs1327323 can indicate the non-deletion state of *MACROD2* in one decision rule. This site is located in chr13:52296316 and mapped in the intron regions of transmembrane phosphoinositide 3-phosphatase and tensin homolog 2

pseudogene 2 (*TPTE2P2*), which is considered a putative promoter in human genome (Kimura et al., 2006). By the whole-exome sequencing analysis of 42 tumor–normal paired samples, highly frequent sites of increased copy number were found in the specific position of chromosome arm 13q (Corraliza Márquez, 2014), the gains in which have been associated with a poor prognosis and metastasis in CRC (Leary et al., 2008). *TPTE2P2* is present in the segments with copy number loss, suggesting that it probably facilitates defect in tumorigenesis. Another publication also reported *TPTE2P2* as one of the key genes identified in gastric cancers (Zeng et al., 2018), implying its crucial role in certain cancers. We inferred that the copy number gain in the specific intron region of *TPTE2P2* results in the progression of CRC, and the loss of copy number in our decision rules identifies the normal status of *MACROD2* and the absence of CRC.

Some SNP sites (rs5904713 and rs13011654) are located in the intron regions of the non-coding RNA gene or the intergenic regions in our decision rules. They have not been reported in current research literature but show strong relevance to the progression of CRC at the CNV level, implying their potential roles in the regulation of oncogenes.

Numerous top-ranked features display the significant relevance to the classification of three status of *MACROD2*, most of which are located in the intron regions of *MACROD2*. Coincident with the relevant information and our inferred decision rules, the CNVs in *MACROD2* resulted in the direct altered states (e.g., cancer). In addition, our approach provides an effective method to evaluate the malignancy extent by detecting a few biomarkers (e.g., SNP sites) rather than conducting an overall detailed analysis of the large gene *MACROD2*, which is more than two million base pairs in size. In summary, our study has proposed for the first time that specific SNP sites can be applied as biomarkers in cancer diagnosis, and further research on these sites will shed light on the molecular mechanism on how these specific DNA regions contribute to the progression of CRC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE115145.

AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. SZ, XP, and LC performed the experiments. TZ, WG, ZG, Y-HZ, and YZ analyzed the results. SZ, XP, and TZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This study was supported by the National Key R&D Program of China (2018YFD1100104, 2018YFC0910403), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Natural Science Foundation of China (31701151, 318724180), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program (16YF1413800), Youth Innovation Promotion Association of Chinese Academy of Sciences (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), and the Science and Technology Commission of Shanghai Municipality (18dz2271000).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00407/full#supplementary-material

Table S1 | Ranked features with MaxRel scores for CRC stage with BAF.

Table S2 | Ranked features with MaxRel scores for CRC stage with LRR.

Table S3 | Ranked features with MaxRel scores for MACROD2 status with BAF.

Table S4 | Ranked features with MaxRel scores for MACROD2 status with LRR.

Table S5 | Performance of 1st-stage IFS with SVM and RIPPER for CRC stage with BAF.

Table S6 | Performance of 2nd-stage IFS with SVM and RIPPER for CRC stage with BAF.

Table S7 | Classification rules learned by RIPPER for CRC stage with BAF.

Table S8 | Performance of 1st-stage IFS with SVM and RIPPER for CRC stage with LRR.

Table S9 | Performance of 2nd-stage IFS with SVM and RIPPER for CRC stage with LRR.

Table S10 | Classification rules learned by RIPPER for CRC stage with LRR.

Table S11 | Performance of 1st-stage IFS with SVM and RIPPER for MACROD2 status with BAF.

Table S12 | Performance of 2nd-stage IFS with SVM and RIPPER for MACROD2 status with BAF.

Table S13 | Classification rules learned by RIPPER for MACROD2 status with BAF.

Table S14 | Performance of 1st-stage IFS with SVM and RIPPER for MACROD2 status with LRR.

Table S15 | Performance of 2nd-stage IFS with SVM and RIPPER for MACROD2 status with LRR.

REFERENCES

Beroukhim, R., Mermel, C. H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., et al. (2010). The landscape of somatic copy-number alteration across human cancers. *Nature* 463, 899–905. doi: 10.1038/nature08822

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Cai, Y.-D., Zhang, S., Zhang, Y.-H., Pan, X., Feng, K., Chen, L., et al. (2018). Identification of the gene expression rules that define the subtypes in glioma. *J. Clin. Med.* 7:350. doi: 10.3390/jcm7100350

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Pan, X., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2019). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* 120, 7068–7081. doi: 10.1002/jcb.27977

Cohen, W. W. (1995). "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning* (Tahoe City, CA: Elsevier), 115–123. doi: 10.1016/B978-1-55860-377-6.50023-2

Corinna Cortes, V. V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Corraliza Márquez, A. M. (2014). *Copy Number Variations of Colorectal Cancer by Whole Exome Sequencing Data* (Master's thesis). University of VIC, Barcelona, Spain.

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteomics* 16, 381–389. doi: 10.2174/1570164616666190126103036

Cutting, G. R., Curristin, S., Zoghbi, H., O'hara, B., Seldin, M. F., and Uhl, G. R. (1992). Identification of a putative gamma-aminobutyric acid (GABA) receptor subunit rho2 cDNA and colocalization of the genes encoding rho2 (GABRR2) and rho1 (GABRR1) to human chromosome 6q14-q21 and mouse chromosome 4. *Genomics* 12, 801–806. doi: 10.1016/0888-7543(92)90312-G

De Cid, R., Riveira-Munoz, E., Zeeuwen, P. L. J. M., Robarge, J., Liao, W., Dannhauser, E. N., et al. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* 41, 211–215. doi: 10.1038/ng.313

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* 464, 999–1005. doi: 10.1038/nature08989

Dukes, C. E. (1932). The classification of cancer of the rectum. *J. Pathol. Bacteriol.* 35, 323–332. doi: 10.1002/path.1700350303

E. Pontes, M. G. N., Da Silveira, S. M., De Souza Trindade Filho, J. C., Rogatto, S. R., and De Camargo, J.L.V. (2013). Chromosomal imbalances in successive moments of human bladder urothelial carcinoma, *Urologic Oncology: Seminars and Original Investigations* 31, 827–835. doi: 10.1016/j.urolonc.2011.05.015

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Gupta, A., Anjomani-Virmouni, S., Koundouros, N., and Poulogiannis, G. (2017). PARK2 loss promotes cancer progression via redox-mediated inactivation of PTEN. *Mol. Cell. Oncol.* 4:e1329692. doi: 10.1080/23723556.2017.1329692

Hamann, J., Aust, G., Araç, D., Engel, F. B., Formstone, C., Fredriksson, R., et al. (2015). International union of basic and clinical pharmacology. xciv. adhesion g protein–coupled receptors. *Pharmacol. Rev.* 67, 338–367. doi: 10.1124/pr.114.009647

Hermsen, M., Postma, C., Baak, J., Weiss, M., Rapallo, A., Sciutto, A., et al. (2002). Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology* 123, 1109–1119. doi: 10.1053/gast.2002.36051

Hsu, Y.-L., Chen, Y.-J., Chang, W.-A., Jian, S.-F., Fan, H.-L., Wang, J.-Y., et al. (2018). Interaction between tumor-associated dendritic cells and colon cancer cells contributes to tumor progression via CXCL1. *Int. J. Mol. Sci.* 19:2427. doi: 10.3390/ijms19082427

Hu, H., Wang, M., Wang, H., Liu, Z., Guan, X., Yang, R., et al. (2018). MEGF6 promotes the epithelial-to-mesenchymal transition via the TGFβ/SMAD signaling pathway in colorectal cancer metastasis. *Cell. Physiol. Biochem.* 46, 1895–1906. doi: 10.1159/000489374

Huang, B., Zhong, N., Cao, H., and Yu, G. (2018). A curated target gene pool assisting disease prediction and patient-specific biomarker selection for lung squamous cell carcinoma. *Oncol. Lett.* 16, 5140–5146. doi: 10.3892/ol.2018.9241

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* 545, 505–509. doi: 10.1038/nature22366

Jeon, M.-S., Song, S.-H., Yun, J., Kang, J.-Y., Kim, H.-P., Han, S.-W., et al. (2016). Aberrant epigenetic modifications of LPHN2 function as a potential cisplatin-specific biomarker for human gastrointestinal cancer. *Cancer Res Treat.* 48, 676–686. doi: 10.4143/crt.2015.153

Kang, R., Xie, Y., Zeh, H. J., Klionsky, D. J., and Tang, D. (2019). Mitochondrial quality control mediated by PINK1 and PRKN: links to iron metabolism and tumor immunity. *Autophagy* 15, 172–173. doi: 10.1080/15548627.2018.1526611

Kawarai, T., Tajima, A., Kuroda, Y., Saji, N., Orlacchio, A., Terasawa, H., et al. (2016). A homozygous mutation of VWA3B causes cerebellar ataxia with intellectual disability. *J. Neurol. Neurosurg. Psychiatr.* 87, 656–662. doi: 10.1136/jnnp-2014-309828

Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., et al. (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* 16, 55–65. doi: 10.1101/gr.4039406

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in: *International Joint Conference on Artificial Intelligence: Lawrence Erlbaum Associates Ltd* (Montreal), 1137–1145.

Kuiper, R. P., Ligtenberg, M. J., Hoogerbrugge, N., and Van Kessel, A. G. (2010). Germline copy number variation and cancer risk. *Curr. Opin. Genet. Dev.* 20, 282–289. doi: 10.1016/j.gde.2010.03.005

Leary, R. J., Lin, J. C., Cummins, J., Boca, S., Wood, L. D., Parsons, D. W., et al. (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16224–16229. doi: 10.1073/pnas.0808041105

Li, J., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* 120, 405–416. doi: 10.1002/jcb.27395

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intellig.* 9, 217–230. doi: 10.1023/A:1008363719778

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Mirza, A. H., Berthelsen, C. H., Seemann, S. E., Pan, X., Frederiksen, K. S., Vilien, M., et al. (2015). Transcriptomic landscape of lncRNAs in inflammatory bowel disease. *Genome Med.* 7:39. doi: 10.1186/s13073-015-0162-2

Oczko-Wojciechowska, M., Włoch, J., Wiench, M., Fujarewicz, K., Simek, K., Gala, G., et al. (2006). Gene expression profile of medullary thyroid carcinoma-preliminary results. *Endokrynol Pol.* 57, 420–426.

Oczkowska, A., Kozubski, W., Lianeri, M., and Dorszewska, J. (2013). Mutations in PRKN and SNCA genes important for the progress of Parkinson's disease. *Curr. Genom.* 14, 502–517. doi: 10.2174/1389202914666131210205839

Pan, X. Y., and Shen, H. B. (2009). Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection. *Protein Pept. Lett.* 16, 1447–1454. doi: 10.2174/092986609789839250

Peng, H. C., Long, F. H., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/TPAMI.2005.159

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

Sakthianandeswaren, A., Parsons, M. J., Mouradov, D., Mackinnon, R. N., Catimel, B., Liu, S., et al. (2018). MACROD2 haploinsufficiency impairs catalytic activity of PARP1 and promotes chromosome instability and growth of intestinal tumors. *Cancer Discov.* 8, 988–1005. doi: 10.1158/2159-8290.CD-17-0909

Sato, S., Yamamoto, K., Matsushita, T., Isobe, N., Kawano, Y., Iinuma, K., et al. (2014). A genome-wide copy number variation study identified T-cell receptor as a susceptibility gene for multiple sclerosis and neuromyelitis optica. *Multiple Scler. J.* 20, 251–252. doi: 10.1002/ana.24511

Scrima, M., De Marco, C., Fabiani, F., Franco, R., Pirozzi, G., Rocco, G., et al. (2012). Signaling networks associated with AKT activation in non-small cell lung cancer (NSCLC): new insights on the role of phosphatydil-inositol-3 kinase. *PLoS ONE* 7:e30427. doi: 10.1371/journal.pone.0030427

Shlien, A., Tabori, U., Marshall, C. R., Pienkowska, M., Feuk, L., Novokmet, A., et al. (2008). Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11264–11269. doi: 10.1073/pnas.0802970105

Song, Y., Pan, Y., and Liu, J. (2019). Functional analysis of lncRNAs based on competitive endogenous RNA in tongue squamous cell carcinoma. *PeerJ* 7:e6991. doi: 10.7717/peerj.6991

Sunnerhagen, M. S., Persson, E., Dahlqvist, I., Drakenberg, T., Stenflo, J., Mayhew, M., et al. (1993). The effect of aspartate hydroxylation on calcium binding to epidermal growth factor-like modules in coagulation factors IX and X. *J. Biol.Chem.* 268, 23339–23344.

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674. doi: 10.1101/gr.6861907

Wang, T., Chen, L., and Zhao, X. (2018). Prediction of drug combinations with a network embedding method. *Comb. Chem. High Throughput Screen.* 21, 789–797. doi: 10.2174/1386207322666181226170140

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* 41, 25–34. doi: 10.1038/ng.287

Yang, T. L., Chen, X. D., Guo, Y., Lei, S. F., Wang, J. T., Zhou, Q., et al. (2008). Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. *Am. J. Hum. Genet.* 83, 663–674. doi: 10.1016/j.ajhg.2008.10.006

Zanke, B. W., Greenwood, C. M., Rangrej, J., Kustra, R., Tenesa, A., Farrington, S. M., et al. (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* 39:989–994. doi: 10.1038/ng2089

Zeng, W., Rao, N., Li, Q., Wang, G., Liu, D., Li, Z., et al. (2018). Genome-wide analyses on single disease samples for potential biomarkers and biological features of molecular subtypes: a case study in gastric cancer. *Int. J. Biol. Sci.* 14, 833–842. doi: 10.7150/ijbs.24816

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi: 10.1146/annurev.genom.9.081307.164217

Zhang, F., Guo, X., Zhang, Y. P., Wen, Y., Wang, W. Z., Wang, S., et al. (2014). Genome-wide copy number variation study and gene expression analysis identify ABI3BP as a susceptibility gene for Kashin-Beck disease. *Hum. Genet.* 133, 793–799. doi: 10.1007/s00439-014-1418-4

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/ACCESS.2019.2944177

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* doi: 10.2174/1574893614666190220114644. [Epub ahead of print].

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2019). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical (ATC) classes of drugs. *Bioinformatics* btz757. doi: 10.1093/bioinformatics/btz757. [Epub ahead of print].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 20 December 2019
doi: 10.3389/fgene.2019.01231

Predicting Endoplasmic Reticulum Resident Proteins Using Auto-Cross Covariance Transformation With a U-Shaped Residue Weight-Transfer Function

Yang-Yang Miao[1,2], *Wei Zhao*[1], *Guang-Ping Li*[1], *Yang Gao*[3]* and *Pu-Feng Du*[1]*

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] School of Chemical Engineering, Tianjin University, Tianjin, China, [3] School of Medicine, Nankai University, Tianjin, China

OPEN ACCESS

Edited by:
Yanjie Wei,
Shenzhen Institutes of Advanced Technology (CAS), China

Reviewed by:
Xiaoqi Zheng,
Shanghai Normal University, China
Pengmian Feng,
North China University of Science and Technology, China

***Correspondence:**
Yang Gao
gaoy@nankai.edu.cn
Pu-Feng Du
pdu@tju.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *09 October 2019*
Accepted: *06 November 2019*
Published: *20 December 2019*

Citation:
Miao Y-Y, Zhao W, Li G-P, Gao Y and Du P-F (2019) Predicting Endoplasmic Reticulum Resident Proteins Using Auto-Cross Covariance Transformation With a U-Shaped Residue Weight-Transfer Function. Front. Genet. 10:1231. doi: 10.3389/fgene.2019.01231

Background: The endoplasmic reticulum (ER) is an important organelle in eukaryotic cells. It is involved in many important biological processes, such as cell metabolism, protein synthesis, and post-translational modification. The proteins that reside within the ER are called ER-resident proteins. These proteins are closely related to the biological functions of the ER. The difference between the ER-resident proteins and other non-resident proteins should be carefully studied.

Methods: We developed a support vector machine (SVM)-based method. We developed a U-shaped weight-transfer function and used it, along with the positional-specific physiochemical properties (PSPCP), to integrate together sequence order information, signaling peptides information, and evolutionary information.

Result: Our method achieved over 86% accuracy in a jackknife test. We also achieved roughly 86% sensitivity and 67% specificity in an independent dataset test. Our method is capable of identifying ER-resident proteins.

Keywords: pseudo-amino acid composition, support vector machine, endoplasmic reticulum resident protein, leave-one-out cross-validation, weight transfer

INTRODUCTION

The endoplasmic reticulum (ER) is an important subcellular organelle in eukaryotic cells. Two major functions are usually recognized for ER. One is that it selectively transports secreted proteins and membrane proteins. The other is that it retains some proteins to maintain its own structure and function (Lavoie and Paiement, 2008). The ER proteins are sorted precisely with quality controls (Ellgaard and Helenius, 2003; Araki and Nagata, 2011). An understanding of these processes contributes to the elucidation of endoplasmic reticulum function and the pathogenesis of many diseases (Paschen and Frandsen, 2001; Verkhratsky, 2002).

ER-resident proteins are an important topic in ER-related studies. Some of the ER-resident proteins possess sorting signals, such as KDEL or KXXX, while some others do not (Stornaiuolo et al., 2003). Over the last two decades, several efforts have been made to determine the ER sorting

signals experimentally. For example, Teasdale and Jackson (1996) found that UGT2 localizes to the endoplasmic reticulum when they studied the UDP-galactosyl transporter (UGT). They also reported that the C-terminal sequence "LLTKVKGS" of the UGT2 is useful in the sorting process. Kabuss et al. (2005) proved that mutating this part of the sequence will result in re-localization of UGT2 to the Golgi apparatus. Although wet experiments for detecting protein localization signals can provide clear evidence and distinguish between maintenance and return signals, performing these experiments is always costly and time-consuming. Therefore, computational predictions are recognized as an alternative approach that provides useful and informative guidance to the experimental methods.

Computational predictions of protein subcellular localizations have been heavily studied in bioinformatics. In the early 1990s, computational systems were developed to recognize the sorting signals from the primary sequences of proteins (Nakai and Kanehisa, 1991; Nakai and Horton, 1999; Wang et al., 2014). When statistical sequence features were introduced to represent protein sequences, machine learning-based algorithms were employed to predict protein sorting destinations. Many studies have tried to apply various algorithms to predict protein subcellular localizations at different levels in different contexts. Several online services have proved useful in this regard. These services include ProLoc-GO (Huang et al., 2007; Huang et al., 2008), KnowPredsite (Lin et al., 2009), SlocX (Ryngajllo et al., 2011), iLoc-Animal (Lin et al., 2013), iLoc-Euk (Chou et al., 2011), Cello v-2.5 (Yu et al., 2006), HybridGO-Loc (Wan et al., 2014), mGOASVM (Wan et al., 2012), Hum-mPloc (Shen and Chou, 2007; Shen and Chou, 2009; Zhou et al., 2017), Euk-mPloc (Chou and Shen, 2007; Chou and Shen, 2010), HPSLPred (Wan et al., 2017), and many others (Chou and Shen, 2008; Briesemeister et al., 2010; Du et al., 2011; Du and Xu, 2013; Almagro Armenteros et al., 2017; Wei et al., 2018; Chen et al., 2019).

The general-purpose protein subcellular location predictors take ER as only one of many subcellular locations. The dataset used for training and testing these methods does not distinguish between ER-resident proteins and non-ER-resident proteins. Since both of these types of proteins may be annotated with subcellular localization ER, constructing a high-quality dataset that is capable of separating them is important. Kumar et al. (2017) proposed the ERPred method, using a carefully curated dataset to distinguish the ER-resident proteins from the non-ER-resident proteins. By using split amino acid compositions (SAAC), they achieved a very promising result. Their results confirmed that the peptide sequences at the terminals of proteins are very informative in guiding the protein sorting process in the ER. Moreover, their results revealed that even if no known sorting signals were found on the sequence, the terminal peptides were still very useful in identifying ER-resident proteins (Kumar et al., 2017).

Pseudo-amino acid composition, which was proposed by Chou (2001), has been widely applied in representing protein sequences for predicting various attributes of proteins. By coupling this with many different machine-learning algorithms, a series of consecutive successes have been achieved. These successful efforts provide consolidated evidence that the pseudo-amino acid compositions are capable of representing protein sequences of various lengths using a fixed-length numerical vector without losing much of the sequential information (Chou, 2011; Chou, 2013; Chou, 2015).

In this study, we introduced a U-shaped weight-adjustment function to improve the pseudo-amino acid compositions. The U-shaped weight-adjustment function transfers weights from the middle-positioned residues to those at the terminals. Besides the weight-adjustment function, we have made two more augmentations to the original pseudo-amino acid compositions. One is to introduce the auto-cross covariance pseudo-factor form, which has been applied in finding protein folding patterns (Dong et al., 2009). The other is to incorporate positional-specific physicochemical properties, which have been applied in predicting protein submitochondrial locations and sub-Golgi locations (Du and Yu, 2013; Jiao and Du, 2017; Zhao et al., 2019).

Our method actually emphasizes the terminal signaling peptide information in pseudo-amino acid compositions. We expect that our approach can be applied not only in predicting ER-resident proteins but also in other topics associated with analyzing protein sorting and localization processes.

MATERIALS AND METHODS

Benchmarking Datasets

In this study, we took the ERPred dataset as our benchmarking dataset. Kumar et al. (2017) released this dataset along with their ERPred study. The ERPred dataset contains two parts: the training set and the independent testing set. **Table 1** gives a breakdown of the entire ERPred dataset. The training set contains 124 ER-resident proteins and 1200 non-ER-resident proteins. The independent testing set contains 65 ER-resident proteins and 2900 non-ER-resident proteins. It is obvious that this dataset is highly imbalanced. The number of non-ER-resident proteins is about 10 times that of the ER-resident proteins in the training set and over 40 times that in the independent testing set. The identifiers of the proteins in the benchmarking dataset are listed in the supplementary materials (**Tables S1–S3**).

Sequence Representations

The ERPred study applied SAAC sequence representations. The result of ERPred implied that the terminal peptides contain more

TABLE 1 | Breakdown of the dataset.

Data set	ERRP[a]	non-ERRP[b]
Training set	124	1200
Independent testing set	65	2900

[a]*ERRP, Endoplasmic reticulum resident proteins.*
[b]*non-ERRP, Non-endoplasmic reticulum resident proteins.*

information for sorting proteins to ER (Kumar et al., 2017). Therefore, we introduced a U-shaped weight-adjustment function to transfer weights from those residues in the middle part of the sequence to those at the terminals of the sequence. Besides this improvement, we incorporated the sequential evolution information using the positional-specific physicochemical properties (PSPCP) (Du and Yu, 2013; Jiao and Du, 2017), as well as the auto-cross covariance form pseudo-factors (Dong et al., 2009).

In order to explain our method properly, we developed a new set of matrix-based notations to describe the Type-II classic pseudo-amino acid compositions, also known as the amphiphilic pseudo-amino acid compositions (Chou, 2005). These new formulations, in mathematics, equal the original ones but with a much simpler appearance. We first give the definitions of the all-ones vector and the shifting matrix.

An n-D all-ones vector is defined as follows:

$$\mathbf{J}_n = [\,\delta_1\ \delta_2\ \cdots\ \delta_n\,]^T, \tag{1}$$

where $\delta_i = 1$ ($i = 1, 2, \ldots, n$).

An n-sized shifting matrix is defined as:

$$\mathbf{M}_n = \left\{m_{i,j}\right\}_{n\times n}, \tag{2}$$

where

$$m_{i,j} = \begin{cases} 1 & \text{, when i} - j = 1; \\ 0 & \text{, otherwise} \end{cases} (i = 1, 2, \ldots, n, j = 1, 2, .., n) \tag{3}$$

A given protein sequence p with length l can be represented as a string:

$$p = r_1 r_2 \cdots r_l, \tag{4}$$

where r_j ($j = 1, 2, \ldots, l$) is the j-th residue on the protein sequence. Every residue represents one of twenty different kinds of amino acids. We use a 20-D binary vector $\mathbf{A}_j$ to represent r_j ($j = 1, 2, \ldots, l$):

$$\mathbf{A}_j = \left[\,a_{1,j}\ a_{2,j}\ \cdots\ a_{20,j}\,\right]^T, \tag{5}$$

where

$$a_{i,j} = \begin{cases} 1 & \text{, when } r_j \text{ is the } i\text{–th type amino acid;} \\ 0 & \text{, otherwise} \end{cases} (i = 1, 2, \ldots, 20, j = 1, 2, \ldots, l) \tag{6}$$

The whole sequence can be represented using a matrix, as follows:

$$\mathbf{A}(p) = [\mathbf{A}_1 \quad \mathbf{A}_2 \cdots \mathbf{A}_l]^T, \tag{7}$$

where $\mathbf{A}(p)$ is a matrix-based sequence representation, and $\mathbf{A}_j$ ($j = 1, 2, \ldots, l$) as in Eq. (5).

When the PSSM can be created using the PSI-BLAST program for protein p, we can obtain a normalized PSSM scoring matrix for p, as elaborated in (Du and Yu, 2013). The normalized PSSM scoring matrix is denoted as follows:

$$\mathbf{B}(p) = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,l} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,l} \\ \vdots & \vdots & \ddots & \vdots \\ b_{20,1} & b_{20,2} & \cdots & b_{20,l} \end{bmatrix}^T, \tag{8}$$

where the following normalization condition is satisfied:

$$\sum_{i=1}^{20} b_{i,j} = 1 (j = 1, 2, \ldots, l) \tag{9}$$

We define the following matrix to combine matrix $\mathbf{B}(p)$ and matrix $\mathbf{A}(p)$:

$$\mathbf{S}(p) = \begin{cases} \mathbf{EB}(p) & \text{, when PSSM can be created for protein } p; \\ \mathbf{EA}(p) & \text{, otherwise.} \end{cases} \tag{10}$$

where matrix $\mathbf{E}$ is a weight-adjustment matrix. It can be defined as a diagonal matrix, as follows:

$$\mathbf{E} = diag(\,\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_l\,), \tag{11}$$

where ε_j ($j = 1, 2, \ldots, l$) is a weight-adjustment factor for the j-th residue on the sequence. It is computed by a U-shaped function, as follows:

$$\varepsilon_j = l\ \frac{\exp[k(2j-l)/l] + \exp[k(l-2j)/l]}{\sum_{j=1}^{l}(\exp[k(2j-l)/l] + \exp[k(l-2j)/l])}\ (j = 1, 2, \ldots, l), \tag{12}$$

where k is a weight distribution parameter, exp(.) is the exponential function, l is the length of the sequence, and j is the j-th residue.

Given a type of physicochemical property H, the values for 20 different types of amino acids can be represented using a 20-D vector.

$$\mathbf{H} = [\,h_1\ h_2\ \cdots\ h_{20}\,]^T, \tag{13}$$

where h_i ($i = 1, 2, \ldots, 20$) is the physicochemical property value of the i-th type amino acid. We use the following method to standardize the physicochemical property vector:

$$\hat{\mathbf{H}} = (\mathbf{H} - m(\mathbf{H})\mathbf{J}_{20})/sd(\mathbf{H}), \tag{14}$$

where $\mathbf{J}_{20}$ is a 20-D all-ones vector,

$$m(\mathbf{H}) = \mathbf{H}^T\mathbf{J}_{20}/20, \tag{15}$$

and

$$sd(\mathbf{H}) = \sqrt{\mathbf{H}^T\mathbf{H}/20 - m^2(\mathbf{H})}\,. \tag{16}$$

In this study, we took two different kinds of physicochemical properties into consideration: the hydrophobicity and hydrophilicity of amino acids. We denote them as $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively. We define the sequence auto-cross covariance matrix of physicochemical properties as:

$$\mathbf{R}_{u,v}(p) = \mathbf{S}(p)\hat{\mathbf{H}}_u\hat{\mathbf{H}}_v^T\mathbf{S}^T(p), \tag{17}$$

where u, $v \in \{1,2\}$.

The k-th order covariance factor can be defined as:

$$\tau_{k,u,v}(p) = tr\left(\mathbf{R}_{u,v}(p)\mathbf{M}_l^k\right)/(l-k), \tag{18}$$

where $tr(.)$ computes the trace of a matrix, $\mathbf{M}_l$ the l-sized shifting matrix, and u, v as in Eq. (17). For every given value of k, a 4-D covariance vector can be generated as:

$$\boldsymbol{\theta}_k(p) = [\,\tau_{k,1,1}(p)\ \tau_{k,1,2}(p)\ \tau_{k,2,1}(p)\ \tau_{k,2,2}(p)\,]^T. \tag{19}$$

By setting the maximum value of k, which is denoted as λ, we can use a 4λ-D vector to contain all covariance factors as:

$$\mathbf{V}_\lambda(p) = \left[\,\boldsymbol{\theta}_1^T(p)\ \boldsymbol{\theta}_2^T(p)\ \ldots\ \boldsymbol{\theta}_\lambda^T(p)\,\right]^T. \tag{20}$$

Considering the weight-adjustment factors, the 20-D conventional amino acid composition vector can be constructed as follows:

$$\mathbf{C}(p) = \mathbf{S}(p)\mathbf{J}_l/l\,. \tag{21}$$

We can combine the $\mathbf{V}_\lambda(p)$ and the $\mathbf{C}(p)$ to create a $(20 + 4\lambda)$-D vector to represent the protein sequence p, as follows:

$$\mathbf{F}(p) = \left[\frac{\mathbf{C}^T(p)}{\mathbf{C}^T(p)\mathbf{J}_{20}+\omega\mathbf{V}_\lambda^T(p)\mathbf{J}_{4\lambda}}\ \frac{\omega\mathbf{V}_\lambda^T(p)}{\mathbf{C}^T(p)\mathbf{J}_{20}+\omega\mathbf{V}_\lambda^T(p)\mathbf{J}_{4\lambda}}\right]^T, \tag{22}$$

where ω is a balancing parameter between 0 and 1. We use $\mathbf{F}(p)$ to represent protein p in this study.

Prediction Algorithm

We employed a support vector machine (SVM) as the prediction algorithm. The SVM searches for an optimal separating hyper-plane in the high-dimensional feature space, which is widely used in bioinformatics problems (Liao et al., 2018; Meng et al., 2019a; Meng et al., 2019b). The hyper-plane can maximize the margin in the feature space. We applied the radial basis function (RBF) as the kernel function in SVM, because the RBF kernel function is the most flexible and the most widely used of such functions. It can be defined as follows:

$$K(\mathbf{F}(p),\mathbf{F}(q)) = \exp\left(-\gamma|\mathbf{F}(p)-\mathbf{F}(q)|^2\right), \tag{23}$$

where p and q are two proteins, and $|.|$ is the operator that computes the Euclidean length of a vector.

Due to the dataset imbalance, we developed a voting scheme to use all samples in the dataset. We partitioned the negative samples into m subsets. The first m - 1 subsets have an equal number of negative samples as that of all the positive samples. The remaining subset contains all the remaining negative samples. For each of these m subsets, all the positive samples were replicated to compose a training subset. We trained the SVM classifier on each of these training subsets. The final prediction result is the majority result of these m classifiers. **Figure 1** is a flowchart of the entire algorithm.

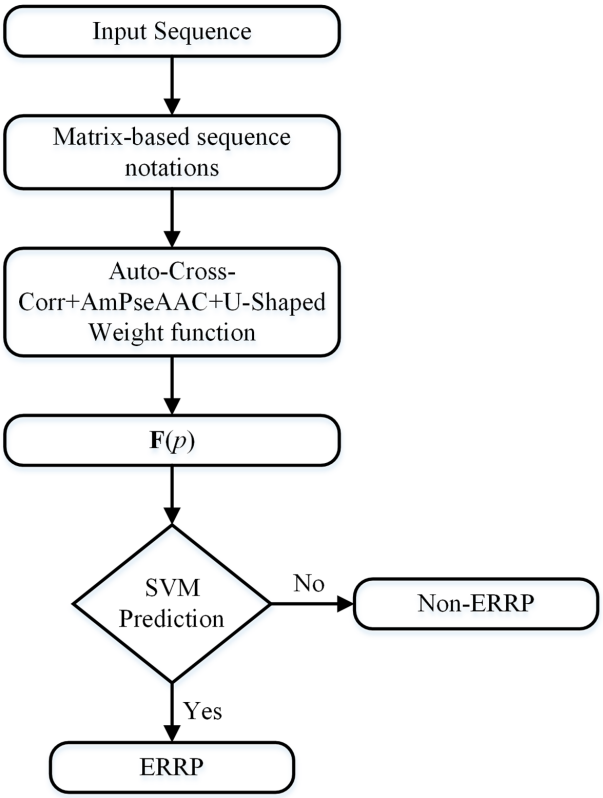


FIGURE 1 | Flowchart of the algorithm. The input sequence will be first converted to matrix-based notations. These notations will be converted into fixed-length numerical vectors, which can represent the sequence order information, the evolutionary information, and the importance of the terminal signaling peptides.

Evaluation Method

Three validation methods are commonly applied in evaluating a bioinformatics predictor. They are known as the self-consistency test, jackknife test, and independent dataset test (Jiao and Du, 2016). Of them, the jackknife test is usually considered as the most objective and rigorous (Chou and Zhang, 1995). However, some recent studies have shown that the independent dataset test can provide even better estimation to the true performance if a sufficiently large testing dataset can be given (Jiao and Du, 2016). Due to the limited size of the training dataset and the fact that our training dataset is highly imbalanced, we applied the jackknife test to estimate the prediction performance of our method. We also evaluated our method using the independent testing dataset, which allowed us to compare our method to the state-of-the-art methods in a fair manner.

Four statistics were applied to measure the prediction performances of our method quantitatively. They are the

sensitivity, specificity, overall accuracy, and the Matthew's Correlation Coefficient (MCC). They are defined as follows:

$$Sen = \frac{TP}{TP + FN}, \quad (24)$$

$$Spe = \frac{TN}{TN + FP}, \quad (25)$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (26)$$

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP + FP)(TP + FN)(FP + FN)(TN + FN)}}, \quad (27)$$

where *Sen* is the sensitivity, *Spe* the specificity, *Acc* the overall accuracy, *MCC* the Matthew's Correlation Coefficient, and *TP*, *TN*, *FP*, and *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively.

Parameter Calibrations

Several parameters can be adjusted in our method. The values of these parameters affect the prediction performances. We applied a grid-search strategy to optimize the jackknife test performance by scanning different combinations of the values of k, λ, and ω. The parameter k was scanned in the set {0, 0.01, 0.1, 1, 1.5}, the parameters λ from 2 to 20 with a step of 1, and the parameter ω from 0.05 to 0.95 with a step of 0.05. For each parameter combination, we use another grid-search to find the best values of c, γ, and w, where c is the cost parameter of SVM, γ is the parameter in the RBF kernel, and w is the class weight ratio between two classes. In this study, we applied the SVM functions in the *scikit-learn* python package. The grid search of SVM parameters was conducted automatically with a python script.

RESULTS AND DISCUSSION

Performance Analysis and Comparison

We obtained the optimized combination of parameters when k = 0.1, λ = 16, ω = 0.55, c = 1000, γ = 0.01, and w = 1.2. The PSSM matrix was created using the PSI-BLAST program with three iterations and 0.001 as the threshold of e-values.

In the jackknife test, our method can correctly identify 111 out of all 124 ER-resident proteins. The prediction performance values are recorded in **Table 2**, with comparison to the ERPred method.

According to these performance values, our method performed better than the ERPred method. Our method achieved a sensitivity of 83.06% and a specificity of 86.38%, which are both higher than the values for ERPred on the same dataset.

TABLE 2 | Prediction performance estimations using a jackknife test.

Methods	Sensitivity	Specificity	Accuracy	MCC
This work	83.1%	86.4%	86.1%	50.6%
ERPred	79.8%	81.6%	81.4%	42.0%

Independent Dataset Test

The training dataset of our work is identical to that used for ERPred. This dataset is highly imbalanced. To further eliminate the concern of over-estimated performances, we performed testing with an independent dataset. We took the same independent testing dataset as used in the ERPred method. The independent testing dataset was processed by the predictor that was trained with the training dataset. The prediction performances of our method are recorded in **Table 3**. Although the specificity is lower than that from the jackknife test, the sensitivity value remains almost unchanged. Therefore, we think the prediction performance is not over-estimated.

We also entered the same testing dataset into several other predictors for comparison. The compared predictors include ERPred (Kumar et al., 2017), Cello v2.5 (Yu et al., 2006), iLoc-Euk (Chou et al., 2011) and Euk-mPLoc 2.0 (Chou and Shen, 2007; Chou and Shen, 2010), which all provide the option to identify ER proteins. According to the prediction performance values, our method has the best sensitivity. However, the specificity of our method is lower. The results indicate that Cello and iLoc-Euk tend to assign non-ER locations to an ER-resident protein. They increase the specificity by severely sacrificing the sensitivity. As the nature of the ER-resident proteins is that the number of non-ER resident proteins is much larger than the resident ones, we think it is acceptable to sacrifice some specificity for the balance to the sensitivity. The ERPred method, Euk-mPLoc 2.0, and our method have a better balance between sensitivity and specificity. Particularly, it seems that the Euk-mPLoc 2.0 method has the best performance, as it achieves over 66% sensitivity while maintaining over 99% specificity. However, it should be noted that Euk-mPLoc 2.0 is not specifically designed to identify ER-resident proteins. Some of the proteins in the testing dataset may have already been used as training samples when Euk-mPLoc 2.0 was developed. This may result in an over-estimated performance value in the comparison. Another factor that should be noticed for Euk-mPLoc 2.0 is that it relies on GO annotations, which makes it not an *ab initio* predictor. Although using GO annotations is common in developing this kind of predictor (Du and Xu, 2013), comparing an *ab initio* predictor with a homology search-based method is not a fair comparison. Therefore, we believe that our method has, at least, comparable prediction performance to other existing methods. Especially in identifying ER-resident proteins, our method should be considered with a higher priority than general-purpose subcellular location predictors.

TABLE 3 | Prediction performance comparison using the independent dataset.

Methods	Sensitivity	Specificity
This work	85.7%	67.2%
ERPred	72.3%	83.7%
Cello 2.5	16.9%	99.9%
iLoc-Euk	15.4%	99.8%
Euk-mPloc 2.0	66.2%	99.0%

Effects of the Residue Weight-Transfer Function

The ER-resident proteins can be roughly divided into two different types. One type is proteins with a specific C-terminal tetra-peptide signal, which usually has a form like KDEL or HDEL. The other type is proteins without this kind of signaling peptide on either its C-terminal or N-terminal. The latter types of proteins usually have an N-glycan modification or similar modifications like cereal prolamin storage proteins (Stornaiuolo et al., 2003). In our training dataset, we searched for the tetra-peptide signals by using ProSite. We found only 41 signaling peptides in all of the 124 ER-resident proteins. In our testing dataset, we performed the same search. We found only 11 singling peptides in all of the 65 non-ER-resident proteins. Therefore, it is not practical to identify ER-resident proteins using only the signaling peptide information. This observation is consistent with the motivation of the ERPred study.

ERPred is a very powerful and useful computational method. It introduces SAAC sequence representations, which successfully emphasize the terminal signaling sequence information. However, the sequence order information is lost in the amino acid composition representations. Although the pseudo-amino acid composition representation can preserve the sequence order information, it cannot emphasize the terminal signaling peptides in the protein sequence. Therefore, we introduced a U-shaped weight-transfer function into the pseudo-amino acid composition in this study. The purpose of this weight-transfer function is to emphasize the terminal signaling information and also to incorporate the sequence order information. However, it is difficult to decide how many weights should be transferred to the terminals from the middle part of a sequence. We formulate this factor as a parameter k in Eq. (12). **Figure 2** illustrates the shape of the function with different k values. **Figure 2** enables an intuitive understanding of this U-shaped weighting function. The larger the value of k, the more weights are transferred to the terminals of a sequence. Please also note that **Figure 2** is only an intuitive illustration of the U-shaped function when the length of a protein is 100. The crossing point under this condition cannot be extended to other cases.

To find an optimized k value, we trained and tested predictors with different k values. **Figure 3** plots the performance values with different k. The sensitivity increases slightly with an increase in k. The specificity peaks when $k = 0.1$. Therefore, at least for predicting ER-resident proteins, $k = 0.1$ creates a good weight-transfer function.

The choice of using a U-shaped function rather than another shape is not easy. Since we do not know how much weight should be transferred, this must be an adjustable parameter in the function. Besides, we need to make the function satisfy the following conditions at the same time: (1) all weights are positive; (2) the sum of all weights equals the sequence length; (3) the portion of the weight-increased part and weight-decreased part remains almost unchanged when we adjust the amount of weight that is transferred. This will make the function only transfer weights among residues, not create or remove total weight. The U-shaped function not only satisfies all these conditions but also provides us with a simple way to implement it.

Sequence Representation Augments

Besides the U-shaped weight-transfer function, we augmented the classic amphiphilic pseudo-amino acid compositions in two

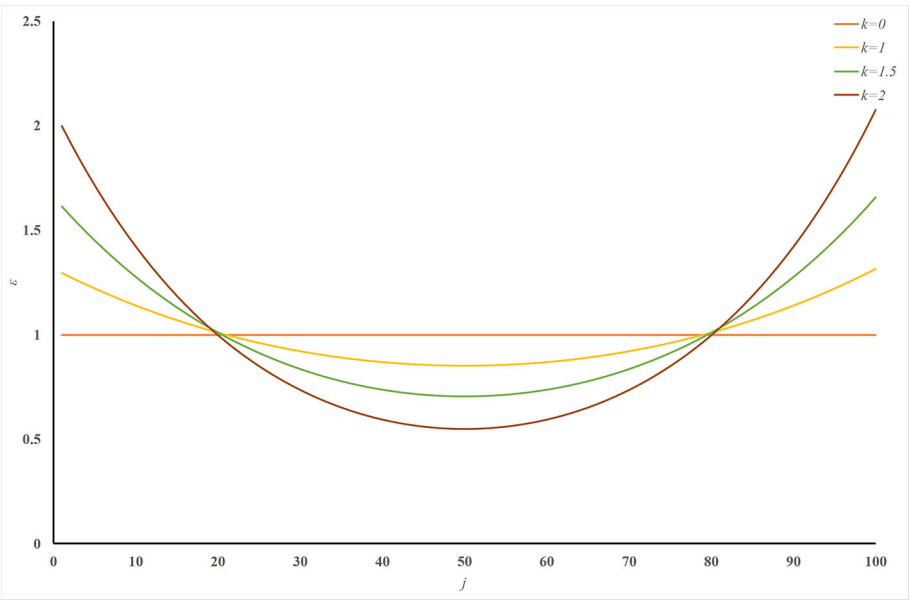


FIGURE 2 | Illustration of the U-shaped weight-transfer function with various k values. The U-shaped function transfers weights from the middle part of a sequence to its terminals. The total weight of a sequence does not change after applying the U-shaped weight-transfer function. When the parameter k is 0, every residue on the sequence has equal weights, which will produce identical results as where there is no weight-transfer function. When the value of k increases, more and more weights are transferred from the residues in the middle part of a sequence to the residues on its terminals.

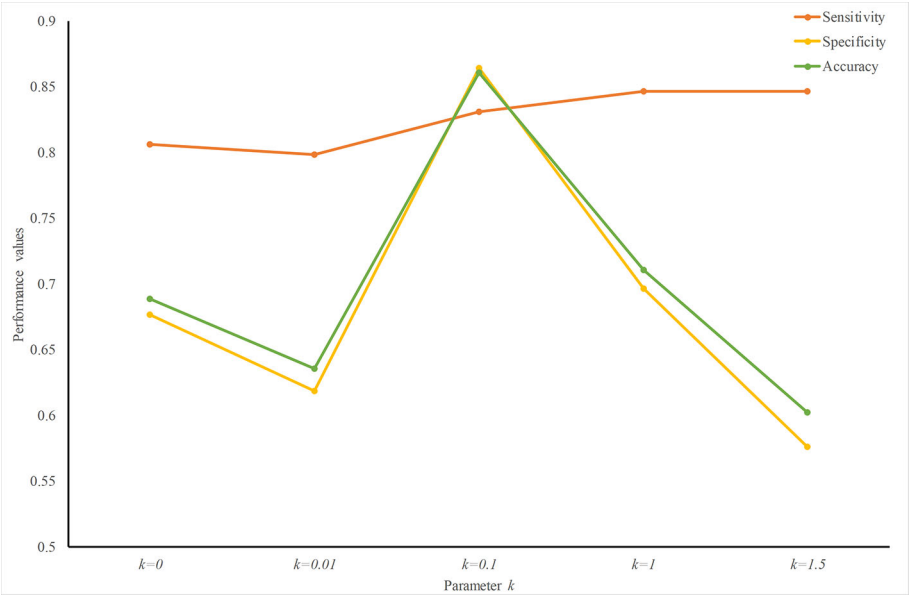


FIGURE 3 | Performance analysis with different weight-transfer functions. Prediction performance varies with the value of parameter *k* in the weight-transfer function. When $k = 0.1$, the performance value peaks. This means that the residues on the terminals are slightly more important than those in the middle part in predicting ER-resident proteins.

ways. One is to use auto-cross correlation to replace the auto-correlation in the classic amphiphilic pseudo-amino acid compositions, while the other is to use matrix-based notations to represent the sequence itself.

The advantage of using auto-cross correlation over auto-correlation has been proved in predicting protein folding patterns (Dong et al., 2009). The matrix-based sequence notations see each residue on the sequence as a 20-D composition vector. The original sequence can then be represented using the one-hot encoding scheme, which can be unified with the normalized PSSM. Since PSI-BLAST cannot generate PSSM for every protein sequence, the matrix-based notation actually provides a mathematically compatible way to compensate for the missing PSSM using the one-hot encodings. As elaborated in Du and Yu (2013), when the PSSM is available for a protein sequence, this matrix-based notation also adjusts the weights of residues according to the evolutionary information.

Therefore, our sequence representation actually encoded the sequence order information and the evolutionary information with emphasis on the terminal signaling peptides in a $(20 + 4\lambda)$-D numerical vector. Compared to other studies, our sequence representation has a much lower number of dimensions. On a dataset with limited samples, the risk of over-estimated performance increases with the number of dimensions of the representation. Our method should be a better choice when the number of samples is limited.

CONCLUSIONS

Many existing methods can predict protein subcellular locations. However, only the ERPred method can specifically identify ER-resident proteins. The ER may be the most important type of subcellular organelle, linking all the major subcellular structures, including the nucleus, cytoplasm, and cell membrane. In this study, we present a new method for predicting ER-resident proteins. Although establishing a web server for a predictive method is good practice, it is not easy for us to do so due to the limitations of our resources and the complexity of this new method. We will establish a web server for this method in the future. The most important part of this work is to introduce a U-shaped weight-transfer function into the pseudo-amino acid compositions. Since the signaling peptide information is useful in analyzing many different subcellular processes and this is the first time that the signaling peptide information has been emphasized in pseudo-amino acid composition representations, we believe that our method has great potential for application in predicting various attributes of proteins.

DATA AVAILABILITY STATEMENT

In this study, we took the ERPred dataset as our benchmarking dataset. Kumar et al released this dataset, along with their ERPred study (Kumar et al., 2017).

AUTHOR CONTRIBUTIONS

Y-YM curated the dataset, designed the algorithm, implemented the algorithm, and partially calibrated the parameters. WZ performed the experiments, partially calibrated the parameters, and collected the results. G-PL partially collected the results and analyzed the results. YG and P-FD investigated the question, designed the whole study,

conceptualized the algorithm, analyzed the results, and wrote the manuscript.

FUNDING

This work was supported by National Key R&D Program of China (2018YFC0910405); National Natural Science Foundation of China (NSFC 61872268); and Open Project Funding of CAS Key Lab of Network Data Science and Technology, Institute of Computing Tech-nology, Chinese Academy of Sciences (CASNDST201705).

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 3387–3395. doi: 10.1093/bioinformatics/btx431

Araki, K., and Nagata, K. (2011). Protein folding and quality control in the ER. *Cold Spring Harb. Perspect. Biol.* 3, a007526. doi: 10.1101/cshperspect.a007526

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010). Going from where to why–interpretable prediction of protein subcellular localization. *Bioinformatics* 26, 1232–1238. doi: 10.1093/bioinformatics/btq115

Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug Metab.* 20, 224–228. doi: 10.2174/1389200219666181031105916

Chou, K.-C., and Shen, H.-B. (2007). Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. *J. Proteome Res.* 6, 1728–1734. doi: 10.1021/pr060635i

Chou, K.-C., and Shen, H.-B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* 3, 153–162. doi: 10.1038/nprot.2007.494

Chou, K.-C., and Shen, H.-B. (2010). A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. *PloS One* 5, e9931. doi: 10.1371/journal.pone.0009931

Chou, K. C., and Zhang, C. T. (1995). Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* 30, 275–349. doi: 10.3109/10409239509083488

Chou, K.-C., Wu, Z.-C., and Xiao, X. (2011). iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PloS One* 6, e18258. doi: 10.1371/journal.pone.0018258

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot1035

Chou, K.-C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21, 10–19. doi: 10.1093/bioinformatics/bth466

Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273, 236–247. doi: 10.1016/j.jtbi.2010.12.024

Chou, K.-C. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Mol. Biosyst.* 9, 1092–1100. doi: 10.1039/c3mb25555g

Chou, K.-C. (2015). Impacts of bioinformatics to medicinal chemistry. *Med. Chem.* 11, 218–234. doi: 10.2174/1573406411666141229162834

Dong, Q., Zhou, S., and Guan, J. (2009). A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* 25, 2655–2662. doi: 10.1093/bioinformatics/btp500

Du, P., and Xu, C. (2013). Predicting multisite protein subcellular locations: progress and challenges. *Expert Rev. Proteomics* 10, 227–237. doi: 10.1586/EPR.13.16

Du, P., and Yu, Y. (2013). SubMito-PSPCP: predicting protein submitochondrial locations by hybridizing positional specific physicochemical properties with pseudoamino acid compositions. *BioMed. Res. Int.* 263829. doi: 10.1155/2013/263829

Du, P., Li, T., and Wang, X. (2011). Recent progress in predicting protein sub-subcellular locations. *Expert Rev. Proteomics* 8, 391–404. doi: 10.1586/EPR.11.20

Ellgaard, L., and Helenius, A. (2003). Quality control in the endoplasmic reticulum. *Nat. Rev. Mol. Cell Biol.* 4, 181–191. doi: 10.1038/nrm1052

Huang, W.-L., Tung, C.-W., Huang, H.-L., Hwang, S.-F., and Ho, S.-Y. (2007). ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *BioSystems* 90, 573–581. doi: 10.1016/j.biosystems.2007.01.001

Huang, W.-L., Tung, C.-W., Ho, S.-W., Hwang, S.-F., and Ho, S.-Y. (2008). ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinf.* 9, 80. doi: 10.1186/1471-2105-9-80

Jiao, Y., and Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* 4, 320–330. doi: 10.1007/s40484-016-0081-2

Jiao, Y.-S., and Du, P.-F. (2017). Predicting protein submitochondrial locations by incorporating the positional-specific physicochemical properties into Chou's general pseudo-amino acid compositions. *J. Theor. Biol.* 416, 81–87. doi: 10.1016/j.jtbi.2016.12.026

Kabuss, R., Ashikov, A., Oelmann, S., Gerardy-Schahn, R., and Bakker, H. (2005). Endoplasmic reticulum retention of the large splice variant of the UDP-galactose transporter is caused by a dilysine motif. *Glycobiology* 15, 905–911. doi: 10.1093/glycob/cwi085

Kumar, R., Kumari, B., and Kumar, M. (2017). Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. *Peer J.* 5, e3561. doi: 10.7717/peerj.3561

Lavoie, C., and Paiement, J. (2008). Topology of molecular machines of the endoplasmic reticulum: a compilation of proteomics and cytological data. *Histochem. Cell Biol.* 129, 117–128. doi: 10.1007/s00418-007-0370-y

Liao, Z., Li, D., Wang, X., and Zou*, L. L. Q. (2018). Cancer diagnosis through IsomiR expression with machine learning method. *Curr. Bioinf.* 13 (1), 57–63. doi: 10.2174/1574893611666160609081155

Lin, H.-N., Chen, C.-T., Sung, T.-Y., Ho, S.-Y., and Hsu, W.-L. (2009). Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. *BMC Bioinf.* 10 Suppl 15, S8. doi: 10.1186/1471-2105-10-S15-S8

Lin, W.-Z., Fang, J.-A., Xiao, X., and Chou, K.-C. (2013). iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. *Mol. Biosyst.* 9, 634–644. doi: 10.1039/C3MB25466F

Meng, C., Jin, S., Wang, L., Guo, F., and Zou, Q. (2019a). AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front. Bioeng. Biotechnol.* 7, 224. doi: 10.3389/fbioe.2019.00224

Meng, C., Wei, L., and Zou, Q. (2019b). SecProMTB: support vector machine-based classifier for secretory proteins using imbalanced data sets applied to mycobacterium tuberculosis. *Proteomics* 19, 1900007. doi: 10.1002/pmic.201900007

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36. doi: 10.1016/S0968-0004(98)01336-X

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* 11, 95–110. doi: 10.1002/prot.340110203

Paschen, W., and Frandsen, A. (2001). Endoplasmic reticulum dysfunction–a common denominator for cell injury in acute and degenerative diseases of the brain? *J. Neurochem.* 79, 719–725. doi: 10.1046/j.1471-4159.2001.00623.x

Ryngajllo, M., Childs, L., Lohse, M., Giorgi, F. M., Lude, A., Selbig, J., et al. (2011). SLocX: predicting subcellular localization of arabidopsis proteins leveraging gene expression data. *Front. Plant Sci.* 2, 43. doi: 10.3389/fpls.2011.00043

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01231/full#supplementary-material

TABLE S1 | Training Dataset.

TABLE S2 | Independent testing Dataset-Positive.

TABLE S3 | Independent testing Dataset-Negative.

Shen, H.-B., and Chou, K.-C. (2007). Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochem. Biophys. Res. Commun.* 355, 1006–1011. doi: 10.1016/j.bbrc.2007.02.071

Shen, H.-B., and Chou, K.-C. (2009). A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.* 394, 269–274. doi: 10.1016/j.ab.2009.07.046

Stornaiuolo, M., Lotti, L. V., Borgese, N., Torrisi, M.-R., Mottola, G., Martire, G., et al. (2003). KDEL and KKXX retrieval signals appended to the same reporter protein determine different trafficking between endoplasmic reticulum, intermediate compartment, and Golgi complex. *Mol. Biol. Cell* 14, 889–902. doi: 10.1091/mbc.e02-08-0468

Teasdale, R. D., and Jackson, M. R. (1996). Signal-mediated sorting of membrane proteins between the endoplasmic reticulum and the golgi apparatus. *Annu. Rev. Cell Dev. Biol.* 12, 27–54. doi: 10.1146/annurev.cellbio.12.1.27

Verkhratsky, A. (2002). The endoplasmic reticulum and neuronal calcium signalling. *Cell Calcium* 32, 393–404. doi: 10.1016/S0143416002001896

Wan, S., Mak, M.-W., and Kung, S.-Y. (2012). mGOASVM: multi-label protein subcellular localization based on gene ontology and support vector machines. *BMC Bioinf.* 13, 290. doi: 10.1186/1471-2105-13-290

Wan, S., Mak, M.-W., and Kung, S.-Y. (2014). HybridGO-Loc: mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins. *PloS One* 9, e89545. doi: 10.1371/journal.pone.0089545

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17, 1700262. doi: 10.1002/pmic.201700262

Wang, Z., Zou, Q., Jiang, Y., Ju, Y., and Zeng, X. (2014). Review of protein subcellular localization prediction. *Curr. Bioinf.* 9, 331–342. doi: 10.2174/1574893609666140212000304

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217. doi: 10.1016/j.jpdc.2017.08.009

Yu, C.-S., Chen, Y.-C., Lu, C.-H., and Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins* 64, 643–651. doi: 10.1002/prot.21018

Zhao, W., Li, G.-P., Wang, J., Zhou, Y.-K., Gao, Y., and Du, P.-F. (2019). Predicting protein sub-Golgi locations by combining functional domain enrichment scores with pseudo-amino acid compositions. *J. Theor. Biol.* 473, 38–43. doi: 10.1016/j.jtbi.2019.04.025

Zhou, H., Yang, Y., and Shen, H.-B. (2017). Hum-mPLoc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics* 33, 843–853. doi: 10.1093/bioinformatics/btw723

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fgene.2019.01319

Application of MCMC-Based Bayesian Modeling for Genetic Evolutionary and Dynamic Change Analysis of Zika Virus

Tong Shao[1], *Jiahui Pan*[1], *Shiwei Zhang*[1], *Zhuoyuan Xin*[1] *and Guoqing Wang*[1,2*]

[1] *Department of Pathogenobiology, The Key Laboratory of Zoonosis, Chinese Ministry of Education, College of Basic Medical Science, Jilin University, Changchun, China,* [2] *The Key Laboratory for Bionics Engineering, Ministry of Education, China, Jilin University, Changchun, China*

Zika virus was first discovered in 1947. For a long time afterward, no large-scale outbreaks occurred. However, more recently, in 2007 and 2016, there were two episodes of ZIKV outbreak that have produced serious public health problems. By analyzing the evolution of the viral genome, we can understand the potential for its outbreak. In this study, we constructed a maximum clade credibility (MCC) tree for the ZIKV non-structural protein 5 (NS5) gene using the Bayesian method. A total of 108 whole-NS5 sequences were retrieved from the GeneBank. We carried out an analysis of potential glycosylation and phosphorylation sites of the ZIKV virus NS5 gene and dynamic analysis of the evolutionary characteristics of the gene. Phylogenetic analysis revealed the presence of two sequence lineages: African and Asian. The sequence of the strains obtained from GeneBank has high homology of 85% to 100%. There are 35 potential phosphorylation sites and glycosylation sites in the ZIKV-NS5 sequences. This article analyzes the possible causes of ZIKV virus outbreaks from the perspective of genetic evolution and analyzes the dynamic trends of virus outbreaks to provide a theoretical basis for predicting the outbreak of the virus.

Keywords: zika virus, NS5, evolution, dynamic changes, Bayesian method

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Yuan Zhou,
Peking University, China
Wang Guohua,
Harbin Institute of Technology, China

***Correspondence:**
Guoqing Wang
qing@jlu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *20 October 2019*
Accepted: *03 December 2019*
Published: *10 January 2020*

Citation:
Shao T, Pan J, Zhang S, Xin Z and Wang G (2020) Application of MCMC-Based Bayesian Modeling for Genetic Evolutionary and Dynamic Change Analysis of Zika Virus. Front. Genet. 10:1319. doi: 10.3389/fgene.2019.01319

INTRODUCTION

In 1947, Zika virus (ZIKV) was first isolated from a monkey in Zika forest, Uganda. (Dick et al., 1952). ZIKV is a member of the virus family *Flaviviridae* and genus *Flavivirus* and is a mosquito-transmitted virus. The virus particles are spherical, with diameters of about 40 ~ 70 nm. Zika virus is a type of single-stranded, positive-sense RNA virus. The whole-genome length is about 10.8 kb, and its single ORF encodes three structural proteins and seven non-structural proteins (NS1, NS2A, NS2A, NS4A, NS4A, NS4B, and NS5) (Kuno and Chang, 2007). The nonstructural protein 5 (NS5) is necessary for genomic replication of zika virus. The N-terminal of NS5 contains methyltransferase (MT), followed by the RNA-dependent RNA polymerase (RdRp).

Abbreviations: ZIKV, Zika virus; NS5, non-structural protein 5; MCC, maximum clade credibility.

The methyltransferase domain at the N-terminal stabilizes the viral RNA genome through 5' capping, while the RdRp domain at the end of C-terminal is very important for the RNA replication of the virus (Decroly et al., 2011; Lu and Gong, 2013; Zhao et al., 2015).

The main means of transmission of ZIKV is through *Aedes* mosquito bites, perinatal transmission, sexual contact, and blood transfusion (Besnard et al., 2014; Musso et al., 2014; Franchini and Velati, 2016). Since the first discovery of Zika virus in 1947, it has gradually spread to become a large-scale problem in the world. The first strain isolated from Asia was named "P6–740" and was isolated from Aededon in Malaysia in 1966 (Haddow et al., 2012). Molecular biology and bioinformatics analysis showed that there are two subtypes of ZIKV, the African and Asian lineages. However, from 1966 to 2007, confirmed cases were scarce, and there was no associated sequence data regarding the Asian linkage. That was until 2007, when 49 cases of ZIKV infection were confirmed in Yip Island and became the first large-scale human infection event in history (Duffy et al., 2009). Now, more than 30 countries have reported ZIKV infections, and these infections have led to multiple imported cases. The ZIKV epidemic has become an important public issue of concern to the whole world (Gong et al., 2016). Base variation, including base recombination, conversion and deletion, will affect the codon usage pattern of the virus, and changes in the codon usage pattern will affect the encoded protein. It is reported that there are potential mutation sites associated with microcephaly (Wang et al., 2017). Studies have shown that envelope protein and NS1 protein of zika virus are predicted to have glycosylation modification sites (Lanciotti et al., 2008; Haddow et al., 2012; Faye et al., 2014). Recently, it has been suggested that correlation between the polymorphism of glycosylation sites and vectors has caused the evolution of Zika virus (Faye et al., 2014). The prediction of viral mutation sites and glycosylation sites is of great significance for understanding the evolution of the virus and the spread of the disease.

Bayesian Inference (BI) is based on using the evolutionary model of sequence evolution to reconstruct the statistical method of the system tree. The resulting tree not only reflects the best estimate of the phylogenetic relationship but also provides the exact support for the branch (Battaglia et al., 2016). Because of the important function of the NS5 gene and the previous construction of an evolutionary tree using the NS5 gene (Gong et al., 2016; Shen et al., 2016), this article uses the Bayesian method to analyze the evolution of the Zika virus NS5 protein, with simultaneous analysis of possible mutation sites. The research result provides a significant theoretical guide to the prevention and treatment of the disease.

MATERIALS AND METHODS

Sequence Collection

The total of 108 NS5 gene sequences that had been added to GenBank before October 2017 were downloaded for Bayesian analysis. These gene sequences are the complete NS5 sequences. The detailed sequence information is listed in Additional **Table S1**.

Sequence Analysis and Comparisons of NS5

The nucleotide sequences of Zika virus NS5 were analyzed using DNASTAR Lasergen 7.0 software to compare their homology.

Analysis of Potential Protein Modification Sites

We used the NetOGlyc 4.0 Server (http://www.cbs.dtu.dk/services/NetOGlyc/) (Steentoft et al., 2013) to estimate the O-glycosylation status of these NS5 sequences (Boon et al., 2016). GlycoMine was used to predict the C-linked and N-linked glycosylation (http://glycomine.erc.monash.edu/Lab/GlycoMine/).

Analysis of the Obtained Viral Genome Data by the Bayesian Method

The Bayesian analysis method was used to study the present evolution rate and evolution model of the epidemic ZIKV strain. The complete NS5 sequence alignment of the ZIKV was disposed carefully with the Clustal W program in MEGA. The RDP3 recombination package was used to detect the recombination of all the sequences. The saturation monitoring was also tested by screening sequences with DAMBE software. If ISS < ISS.c, it means that the sequence substitution is not saturated and meets the requirements for building a phylogenetic tree using Bayesian methods. Finally, the best evolution model was selected with jModelTest software. BEAST v1.8.0 was employed under the GTR +I+G model of nucleotide substitutions and with the Relaxed clock: Uncorrelated Log-normal setting to perform 80 million MCMC runs to construct a maximum clade credibility tree (effective sampling size >200). The analysis was sampled at every 8000 states. Posterior probabilities were calculated with a burn-in of 8 million states. The analysis of the sampling data was output by Tracer v1.6, and the Tree Annotator program was employed to output the results of the MCC tree model. FigTree program was then used to plot the MCC molecular evolutionary tree.

RESULTS

Homologous Comparison of Zika Virus Sequences

Zika virus is a member of the family *Flaviviridae* and genus Flavivirus and is a mosquito-transmitted virus. In the phylogenetic tree, it is close to Dengue virus, Japanese encephalitis virus, and West Nile virus;, the closest virus is Spondweni (**Figure 1**). The Zika virus strains used in this study were 108 strains collected from 20 districts. The results showed that the nucleotide homology of the 108 strains of Zika virus was between 85.4% and 100% and that some of them were 100% homologous. Twenty-four strains from the United States (**Figure 2A**) and seventeen strains from Brazil are compared

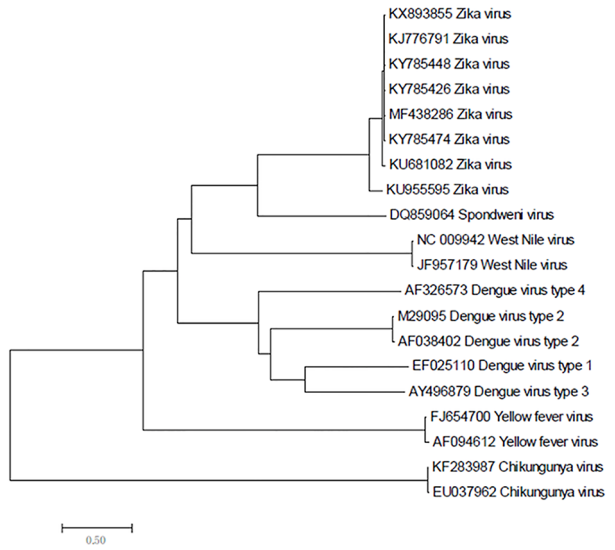


FIGURE 1 | Phylogenetic analysis of Zika virus, Dengue virus, Spondweni virus, West Nile virus, Yellow fever virus and Chikungunya virus based on NS5 gene.

respectively (**Figures 2B**). We can see the homogeneity of the twenty-four strains of the NS5 gene from the United States is 96.3%~100%, and there are many sequences of the strains of the NS5 amino acid that have a homology of 100%. The same result was found in the Brazil strains. The nucleotide homology of the Brazil strains is 99.4%~100%. This result shows that the mutation rate in the 108 strains is low and that the NS5 gene is relatively conservative.

Prediction of Glycosylation Sites

We performed three types of glycosylation site prediction for NS5 sequences of 100 strains. We used GlycoMine to predict C-linked and N-linked glycosylation sites and NetOGlyc to predict O-glycosylation. It can be seen from the results that there are 10 sites that are potentially modified by O-glycosylation (**Figure 3**), and the number of sites of N-linked and C-linked glycosylation that may occur in different strains is not much different (**Figure 3**). For example, comparing one strain of KY014296 from Brazil with other strains from Brazil that lack a C-linked site 654 (**Figure 3**), we can see in the sequence alignment that the amino acid of the strain at this position is arginine, whereas the amino acid of the other strains at this position is tryptophan.

Recombinant Analysis of Virus Strains

In order to identify whether recombination occurs between different strains in the same region, we used SimPlot to analyze the sequences of different strains in the same region. As shown in **Figure 4**, there is no recombination in the strains of Brazil. The sequences of all of the NS5 genes were then grouped by region, with the strains from the same region grouped together. SimPlot was then used to verify the occurrence of recombination events further. From **Figure 4**, we can see that

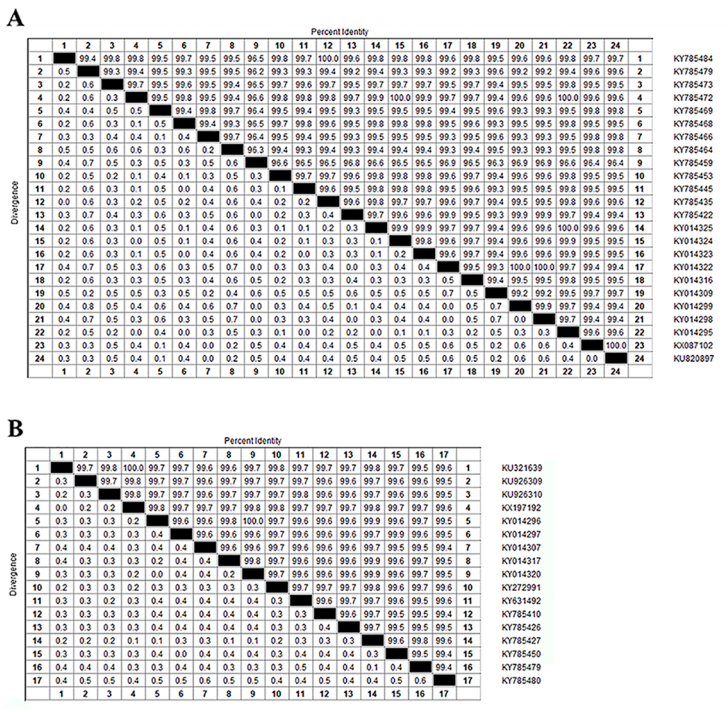


FIGURE 2 | Homologous comparison. **(A)** Homology alignment analysis of 24 Zika strains from the United States using the DNASTAR software package. **(B)** Homology alignment analysis of 17 Zika strains from Brazil using the DNASTAR software package.

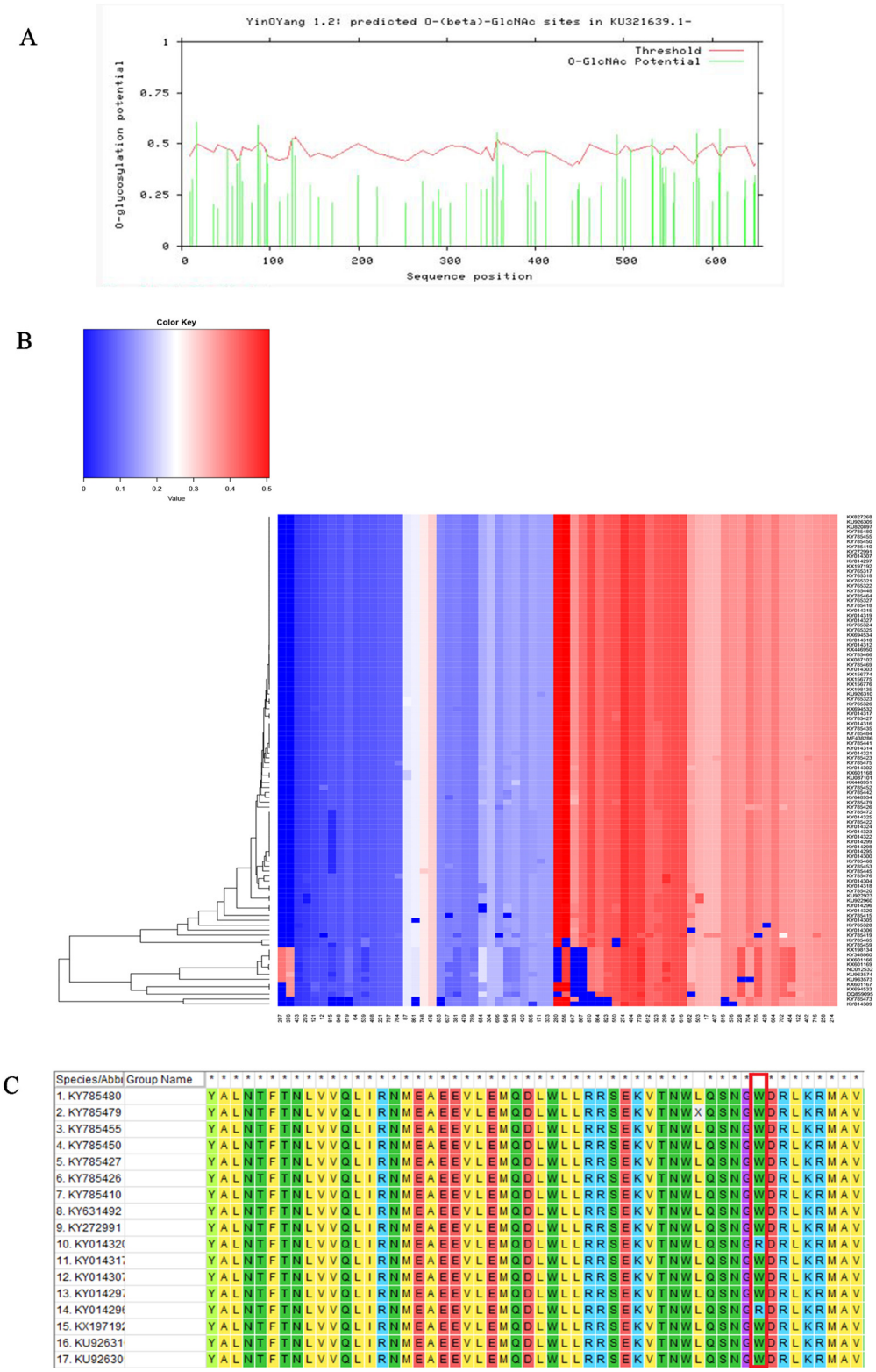


FIGURE 3 | Functional site prediction. **(A)** Prediction of O-glycosylation site of Zika virus from Brazil using the NetOGlyc website. **(B)** Heat map of predicted N-linked and C-linked glycosylation sites. The ordinate of the heat map represents the sequence of different strains; the abscissa represents all amino acid sites of the NS5 gene. Glycosylation predictions are performed on 100 sequences. The darker the red color, the higher the score. **(C)** Sequence alignment result.

there is no recombination event in any of the sequences. The above results indicate that no recombination occurred in the selected strain sequences.

Evolutionary Tree Construction Based on the Bayesian-Markov Chain Method

A total of 108 complete NS5 gene sequences were used in the phylogenetic analysis. Samples were collected from 20 regions. Although these strains were from 20 different regions, they were eventually divided into two groups, the African lineage and the Asian lineage. After 2015, the isolated strains were very close to each other and the new outbreak strains selected in this study are all Asian-type. Indicating that during this time, the NS5 gene sequence of ZIKV was conservative. There is no extra genotype from after the outbreak of the ZIKV epidemics in 2015 and 2016. The recent outbreak was predominantly in Asia, and the contemporary epidemics are dominantly evolved from Asian strains. Neither the American strains nor the Brazilian strains have a very specific genotype (**Figure 5C**). Moreover, among these strains, there is no clear dividing line between the strains of each country. As can be seen from **Figure 5**, strains from the United States, Brazil, and the Dominican Republic are cross-distributed in the phylogenetic tree, with the closest ancestor being a tree root. Strains from Honduras, Nicaragua, and Mexico are closely spaced, and these are the strains most distant from the Africa linkage. From a temporal perspective, the kinship strains we collected at different times from the same area were the most recent. This shows that the strains in each region are from local ancestors, and there is no cross-infection with other regions.

The Bayesian-Markov chain method was used to determine the codon mutation rate of the Zika NS5 gene, and the BEAST results were analyzed by Trace. The results showed that the

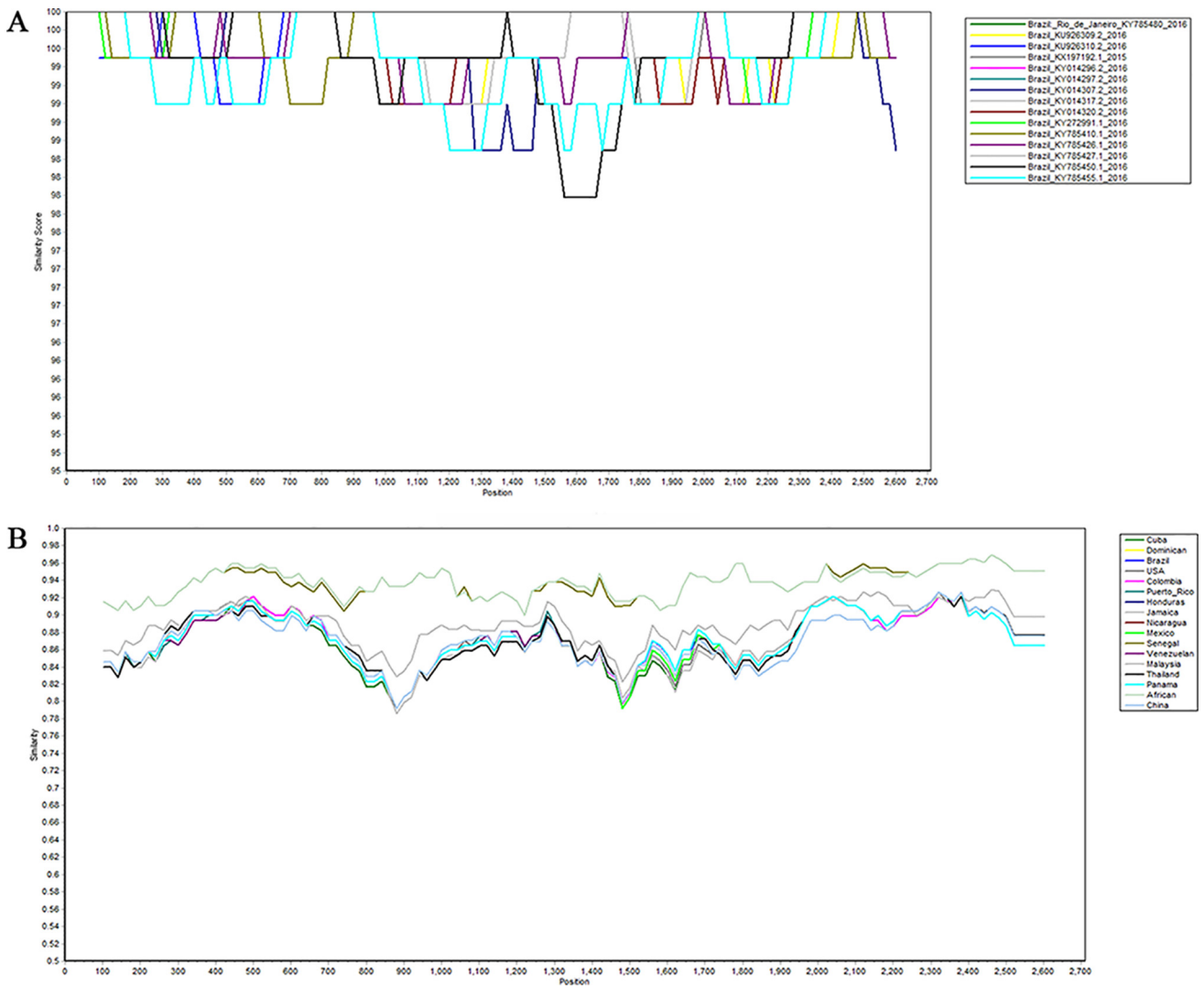


FIGURE 4 | Recombination analysis of Zika virus. Plots of the similarity (generated by SimPlot) of a set of reference sequences. **(A)** Sequences of Zika virus from Brazil. **(B)** Comparison after grouping the Zika virus into different regions.

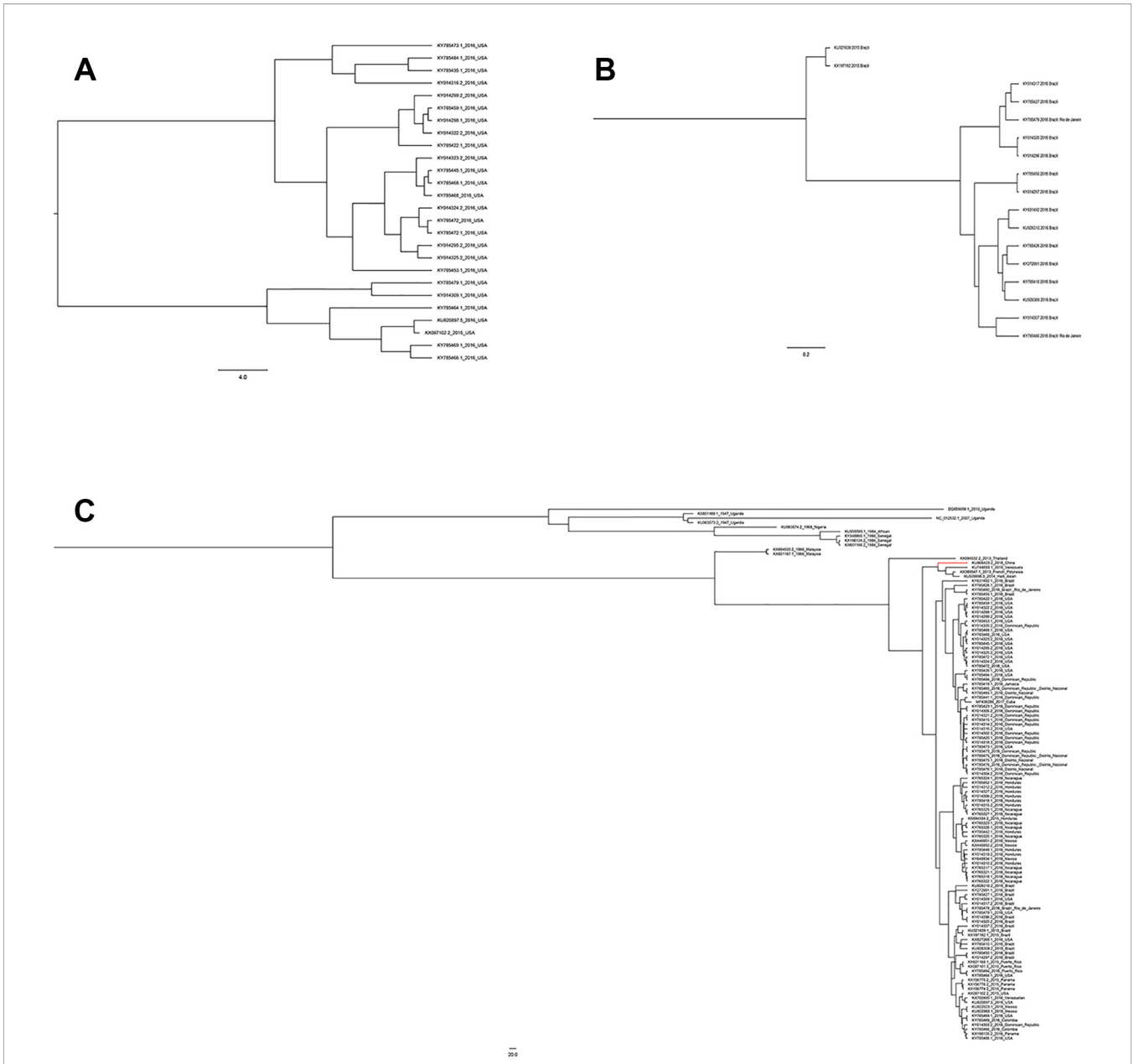


FIGURE 5 | Phylogenetic analysis of Zika virus. **(A)** Evolutionary development of 24 strains of Zika virus NS5 genes from the United States. **(B)** Evolutionary development of 17 strains of Zika virus NS5 genes from Brazil. **(C)** Evolutionary development of 108 strains of Zika virus NS5 genes.

codon mutation rates of the amino acids encoded of ZIKV NS5 were different, and, respectively, the mutation rates of the three codons were 0.3695, 0.1596, and 2.4709 (**Figures 6A**). Thus, the mutation rate of the third codon was the highest. Since the mutation rate of the third codon is the highest and the codon has degeneracy, some mutations do not change the amino acids of the encoded protein, which makes the homology between the Zika virus strains very high. These values indicate that during this period, there was a base mutation of the NS5 gene, and this may be associated with the recent outbreak of the Zika virus. The geographical distribution of Zika viruses is steadily growing. As can be seen from the skyline plot (**Figure 6**), the effective size of the Zika virus has decreased since its discovery, but it also increased somewhat in 2015, coinciding with the Zika virus outbreak.

DISCUSSION

The Zika virus, which was discovered in 1947, returned 70 years after its discovery, unexpectedly appearing in the Pacific Islands and Latin America. Pathogenic changes, including microcephaly

A

Summary Statistic	CP1.mu	CP2.mu	CP3.mu
Mean	0.3695	0.1596	2.4709
95% HPD lower	0.2976	0.1133	2.3819
95% HPD upper	0.4482	0.2111	2.5546
Effective sample size（ESS）	7300.39	8101	7532.37

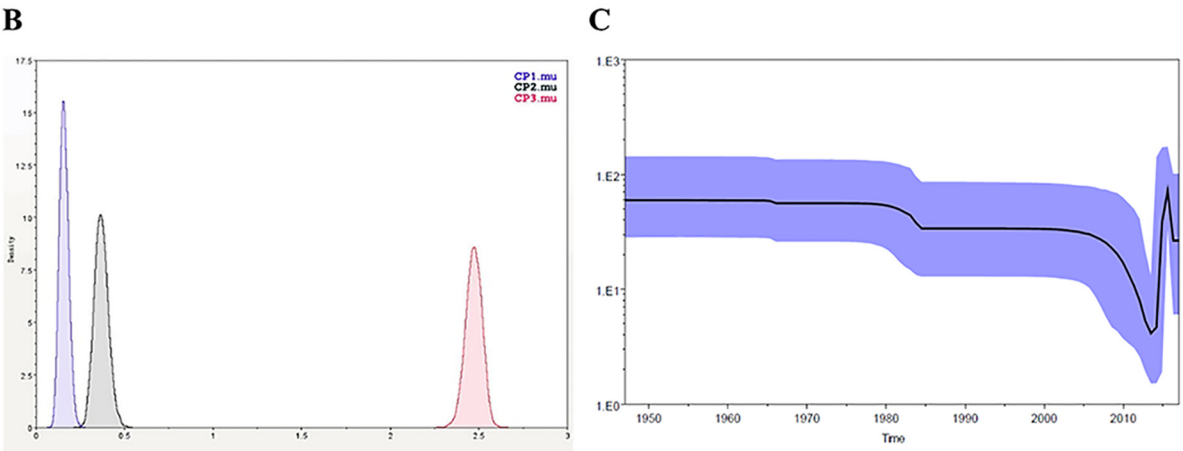


FIGURE 6 | Zika virus NS5 codon mutation rate and skyline plot **(A**, **B).** The codon mutation rate of the Zika virus NS5 gene was estimated by the Bayes-Markov chain method. The codon mutation rate is the result of a BEAST run using Trace analysis. **(C)** Dynamic study of Zika virus NS5 gene genetic diversity by Bayesian skyline plot. The thick solid line is the median estimate, and the dotted line shows the 95% confidence interval. The abscissa is time, and the ordinate is the effective population size. The curve shows that the Zika virus NS5 gene has been in a stable state and the population gradually began to grow in 2015.

and Guillain-Barre syndrome, have caused widespread concern. One possible reason for this is the objective environmental conditions of an increased global population and an increased mosquito vector population (Pettersson et al., 2016; Shi et al., 2018). Another possibility is that amino acid substitution occurs that affects the rate of transmission and the pathogenicity of the virus. The effect of amino acid substitution on pathogenicity has been reported previously. For example, it was found that there was a substitution from S to N at position 139 of the prM protein before the French Polynesian outbreak of 2013 and that the subsequent strains in the Americas were all 139N. In vitro experiments showed that amino acid substitution enhanced infectivity and induced more severe microcephaly. Interaction between the virus and the host can lead to different infection outcomes (Yuan et al., 2017). Yang Liu et al. showed that spontaneous mutations on NS1 proteins increase their own antigenemia (Liu et al., 2017). Hongjie Xia et al. believe that mutations in Zika NS1 protein increase the body's ability to evade the immune response and increase the possibility of infection and epidemic (Xia et al., 2018). The replacement of one amino acid site has the potential to improve pathogenicity and transmission efficiency, which may explain why Zika virus has re-emerged after so many years. This is of great significance to study this mutation.

Compared with other gene fragments, NS5 and envelope gene fragments still had higher variability, although the non-structural proteins NS3 and NS5 were relatively conserved compared with other gene fragments according to homologous modeling analysis (Koh, 2014; Mazeaud et al., 2018), which in turn affects the genetic stability of the protein, making it easier for the virus to invade the human body (Yuan et al., 2015). We can observe an obvious cluster of NS5 genes consisting of only Chinese strains (**Figure 4** red), and the genetic distance between Chinese strains and French Polynesian strain is small. In 2013, a study showed that this strain from China and the Latin American strains have a common ancestor. (Faria et al., 2016). This suggests that this Chinese lineage may have evolved from an ancestor that erupted in the Pacific islands in 2013. Asian strains form an independent cluster, and the recent outbreaks of the Zika virus are of Asian lineage, indicating that Asian strains are more diversified than African Zika virus strains. There is a certain degree of mutation in the NS5 genes of Zika virus strains collected from Brazil and the United States. However, these mutations did not alter the glycosylation and phosphorylation sites of the NS5-encoded protein, suggesting that though there are mutations in the NS5 gene, these mutations did not impair the stability of the virus, and the protein structure, which plays an important role in the protein structure, remained

stable. From **Figure 5**, we can see that the mutation rate of the third codon is the highest. Because of the degeneracy of the codon, mutations that occur on the third codon may not cause amino acid changes, which could explain why the various glycosylation and phosphorylation sites did not change after mutation. Glycosylation sites in the Zika virus genome display polymorphisms and may have adaptive value in evolutionary processes (Singh et al., 2016). It has been reported that Zika virus has a loss of glycosylation sites (Hanna et al., 2005; Lee et al., 2010). Mutations at amino acid sites play an important role in the pathogenicity of Zika virus, so analysis of the virus evolution is critical to better understand the pathogenesis of viral infection and the variability of its clinical phenotype.

The data we selected included the NS5 protein sequence of the Zika virus that broke out in 2016 and previously. The relatively stable NS5 gene nucleotide sequence will provide a great opportunity to develop a vaccine for this disease. We predicted the dynamic phylogenetic trends, which indicate the outbreak trends of ZIKV and provide theoretical foundations for clinical prevention. The potential glycosylation and phosphorylation sites of the NS5 gene were predicted and discussed in conjunction with existing functional assays.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/nuccore/?term = DQ859059, KU321639, KU509998, KU744693, KU820897, KU866423, KU922923, KU922960, KU926309, KU926310, KU955595, KU963573, KU963574, KX087101, KX087102, KX156774, KX156775, KX156776, KX197192, KX198134, KX198135, KX369547, KX446950, KX446951, KX601166, KX601167, KX601168, KX601169, KX694532, KX694533, KX694534, KX702400, KX827268, KY014295, KY014296, KY014297, KY014298, KY014299, KY014300, KY014302, KY014303, KY014304, KY014305, KY014306, KY014307, KY014309, KY014310, KY014312, KY014314, KY014315, KY014316, KY014317, KY014318, KY014319, KY014320, KY014321, KY014322, KY014323, KY014324, KY014325, KY014327, KY272991, KY348860, KY631492, KY648934, KY765317, KY765318, KY765320, KY765321, KY765322, KY765323, KY765324, KY765325, KY765326, KY765327, KY785410, KY785415, KY785418, KY785419, KY785420, KY785422, KY785423, KY785426, KY785427, KY785435, KY785441, KY785442, KY785445, KY785448, KY785450, KY785452, KY785453, KY785455, KY785459, KY785464, KY785465, KY785466, KY785468, KY785469, KY785472, KY785473, KY785475, KY785476, KY785479, KY785480, KY785484, MF438286, NC_012532.

AUTHOR CONTRIBUTIONS

GW designed this study. TS analyzed the data and wrote the manuscript. JP, SZ, and ZX contributed to data collection. All of the authors approved the final manuscript.

FUNDING

This work was supported by grants from Epidemiology, Early Warning and Response Techniques of Major Infectious Diseases in the Belt and Road Initiative (#2018ZX10101002), the National Natural Science Foundation of China (#81871699), and the Foundation of Jilin Province Science and Technology Department (#172408GH010234983).

ACKNOWLEDGMENTS

We thank Medjaden Bioscience Limited (Hong Kong, China) for editing and proofreading this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01319/full#supplementary-material

TABLE S1 | List of the 108 complete NS5 sequences of ZIKV strains available from GenBank that were selected to generate the ZIKV MCC (XLSX 12kb).

REFERENCES

Battaglia, V., Gabrieli, P., Brandini, S., Capodiferro, M. R., Javier, P. A., Chen, X. G., et al. (2016). The worldwide spread of the tiger mosquito as revealed by mitogenome haplogroup diversity. *Front. Genet.* 7, 208. doi: 10.3389/fgene.2016.00208

Besnard, M., Lastere, S., Teissier, A., Cao-Lormeau, V., and Musso, D. (2014). Evidence of perinatal transmission of Zika virus, French Polynesia, December 2013 and February 2014. *Euro. Surveill.* 19 (13), 20751.

Boon, L., Ugarte-Berzal, E., Vandooren, J., and Opdenakker, G. (2016). Glycosylation of matrix metalloproteases and tissue inhibitors: present state, challenges and opportunities. *Biochem. J.* 473 (11), 1471–1482. doi: 10.1042/BJ20151154

Decroly, E., Ferron, F., Lescar, J., and Canard, B. (2011). Conventional and unconventional mechanisms for capping viral mRNA. *Nat. Rev. Microbiol.* 10 (1), 51–65. doi: 10.1038/nrmicro2675

Dick, G. W., Kitchen, S. F., and Haddow, A. J. (1952). Zika virus. I. Isolations and serological specificity. *Trans. R. Soc. Trop. Med. Hyg.* 46 (5), 509–520. doi: 10.1016/0035-9203(52)90042-4

Duffy, M. R., Chen, T. H., Hancock, W. T., Powers, A. M., Kool, J. L., Lanciotti, R. S., et al. (2009). Zika virus outbreak on Yap Island, Federated States of Micronesia. *N. Engl. J. Med.* 360 (24), 2536–2543. doi: 10.1056/NEJMoa0805715

Faria, N. R., Azevedo, R., Kraemer, M. U. G., Souza, R., Cunha, M. S., Hill, S. C., et al. (2016). Zika virus in the Americas: early epidemiological and genetic findings. *Science* 352 (6283), 345–349. doi: 10.1126/science.aaf5036

Faye, O., Freire, C. C., Iamarino, A., Faye, O., de Oliveira, J. V., Diallo, M., et al. (2014). Molecular evolution of Zika virus during its emergence in the 20(th) century. *PloS Negl. Trop. Dis.* 8 (1), e2636. doi: 10.1371/journal.pntd.0002636

Franchini, M., and Velati, C. (2016). Blood safety and zoonotic emerging pathogens: now it's the turn of Zika virus! *Blood Transfus.* 14 (2), 93–94. doi: 10.2450/2015.0187-15

Gong, Z., Gao, Y., and Han, G. Z. (2016). Zika Virus: two or three lineages? *Trends Microbiol.* 24 (7), 521–522. doi: 10.1016/j.tim.2016.05.002

Haddow, A. D., Schuh, A. J., Yasuda, C. Y., Kasper, M. R., Heang, V., Huy, R., et al. (2012). Genetic characterization of Zika virus strains: geographic expansion of the Asian lineage. *PloS Negl. Trop. Dis.* 6 (2), e1477. doi: 10.1371/journal.pntd.0001477

Hanna, S. L., Pierson, T. C., Sanchez, M. D., Ahmed, A. A., Murtadha, M. M., and Doms, R. W. (2005). N-linked glycosylation of west nile virus envelope proteins influences particle assembly and infectivity. *J. Virol.* 79 (21), 13262–13274. doi: 10.1128/JVI.79.21.13262-13274.2005

Koh, P. O. (2014). Estradiol ameliorates the reduction in parvalbumin expression induced by ischemic brain injury. *Neurosci. Lett.* 574, 36–40. doi: 10.1016/j.neulet.2014.05.006

Kuno, G., and Chang, G. J. (2007). Full-length sequencing and genomic characterization of Bagaza, Kedougou, and Zika viruses. *Arch. Virol.* 152 (4), 687–696. doi: 10.1007/s00705-006-0903-z

Lanciotti, R. S., Kosoy, O. L., Laven, J. J., Velez, J. O., Lambert, A. J., Johnson, A. J., et al. (2008). Genetic and serologic properties of Zika virus associated with an epidemic, Yap State, Micronesia, 2007. *Emerg. Infect. Dis.* 14 (8), 1232–1239. doi: 10.3201/eid1408.080287

Lee, E., Leang, S. K., Davidson, A., and Lobigs, M. (2010). Both E protein glycans adversely affect dengue virus infectivity but are beneficial for virion release. *J. Virol.* 84 (10), 5171–5180. doi: 10.1128/JVI.01900-09

Liu, Y., Liu, J., Du, S., Shan, C., Nie, K., Zhang, R., et al. (2017). Evolutionary enhancement of Zika virus infectivity in Aedes aegypti mosquitoes. *Nature* 545 (7655), 482–486. doi: 10.1038/nature22365

Lu, G., and Gong, P. (2013). Crystal Structure of the full-length Japanese encephalitis virus NS5 reveals a conserved methyltransferase-polymerase interface. *PloS Pathog.* 9 (8), e1003549. doi: 10.1371/journal.ppat.1003549

Mazeaud, C., Freppel, W., and Chatel-Chaix, L. (2018). the multiples fates of the Flavivirus RNA genome during pathogenesis. *Front. Genet.* 9, 595. doi: 10.3389/fgene.2018.00595

Musso, D., Nhan, T., Robin, E., Roche, C., Bierlaire, D., Zisou, K., et al. (2014). Potential for Zika virus transmission through blood transfusion demonstrated during an outbreak in French Polynesia, November 2013 to February 2014. *Euro. Surveill.* 19 (14).

Pettersson, J. H., Eldholm, V., Seligman, S. J., Lundkvist, A., Falconar, A. K., Gaunt, M. W., et al. (2016). How did Zika virus emerge in the Pacific Islands and Latin America? *MBio* 7 (5), e01239–16. doi: 10.1128/mBio.01239-16

Shen, S., Shi, J., Wang, J., Tang, S., Wang, H., Hu, Z., et al. (2016). Phylogenetic analysis revealed the central roles of two African countries in the evolution and worldwide spread of Zika virus. *Virol. Sin.* 31 (2), 118–130. doi: 10.1007/s12250-016-3774-9

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., and Yu, B. (2018). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 111 (6), 1839–1852. doi: 10.1016/j.ygeno.2018.12.007

Singh, R. K., Dhama, K., Malik, Y. S., Ramakrishnan, M. A., Karthik, K., Tiwari, R., et al. (2016). Zika virus - emergence, evolution, pathology, diagnosis, and control: current global scenario and future perspectives - a comprehensive review. *Vet. Q.* 36 (3), 150–175. doi: 10.1080/01652176.2016.1188333

Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T., et al. (2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* 32 (10), 1478–1488. doi: 10.1038/emboj.2013.79

Wang, A., Thurmond, S., Islas, L., Hui, K., and Hai, R. (2017). Zika virus genome biology and molecular pathogenesis. *Emerg. Microbes Infect.* 6 (3), e13. doi: 10.1038/emi.2016.141

Xia, H., Luo, H., Shan, C., Muruato, A. E., Nunes, B. T. D., Medeiros, D. B. A., et al. (2018). An evolutionary NS1 mutation enhances Zika virus evasion of host interferon induction. *Nat. Commun.* 9 (1), 414. doi: 10.1038/s41467-017-02816-2

Yuan, L., Wu, J., Liu, J., Li, G., and Liang, D. (2015). Intermittent hypoxia-induced parvalbumin-immunoreactive interneurons loss and neurobehavioral impairment is mediated by NADPH-Oxidase-2. *Neurochem. Res.* 40 (6), 1232–1242. doi: 10.1007/s11064-015-1586-1

Yuan, L., Huang, X. Y., Liu, Z. Y., Zhang, F., Zhu, X. L., Yu, J. Y., et al. (2017). A single mutation in the prM protein of Zika virus contributes to fetal microcephaly. *Science* 358 (6365), 933–936. doi: 10.1126/science.aam7120

Zhao, Y., Soh, T. S., Lim, S. P., Chung, K. Y., Swaminathan, K., Vasudevan, S. G., et al. (2015). Molecular basis for specific viral RNA recognition and 2'-O-ribose methylation by the dengue virus nonstructural protein 5 (NS5). *Proc. Natl. Acad. Sci. U. S. A.* 112 (48), 14834–14839. doi: 10.1073/pnas.1514978112

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 10 January 2020
doi: 10.3389/fgene.2019.01316

Three-Layer Heterogeneous Network Combined With Unbalanced Random Walk for miRNA-Disease Association Prediction

Limin Yu[1,2], *Xianjun Shen*[1,2*], *Duo Zhong*[1,2] *and Jincai Yang*[1]

[1] School of Computer, Central China Normal University, Wuhan, China, [2] Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning, Central China Normal University, Wuhan, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Zengyou He,
Dalian University of Technology 047, China
Lei Deng,
Central South University, China
Yan Wang,
Hubei University of Chinese Medicine, China

***Correspondence:**
Xianjun Shen
xjshen@mail.ccnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 28 September 2019
Accepted: 02 December 2019
Published: 10 January 2020

Citation:
Yu L, Shen X, Zhong D and Yang J (2020) Three-Layer Heterogeneous Network Combined With Unbalanced Random Walk for miRNA-Disease Association Prediction. Front. Genet. 10:1316. doi: 10.3389/fgene.2019.01316

miRNA plays an important role in many biological processes, and increasing evidence shows that miRNAs are closely related to human diseases. Most existing miRNA-disease association prediction methods were only based on data related to miRNAs and diseases and failed to effectively use other existing biological data. However, experimentally verified miRNA-disease associations are limited, there are complex correlations between biological data. Therefore, we propose a novel Three-layer heterogeneous network Combined with unbalanced Random Walk for MiRNA-Disease Association prediction algorithm (TCRWMDA), which can effectively integrate multi-source association data. TCRWMDA based not only on the known miRNA—disease associations, also add the new priori information (lncRNA–miRNA and lncRNA–disease associations) to build a three-layer heterogeneous network, lncRNA was added as the transition path of the intermediate point to mine more effective information between networks. The AUC value obtained by the TCRWMDA algorithm on 5-fold cross validation is 0.9209, compared with other models based on the same similarity calculation method, TCRWMDA obtained better results. TCRWMDA was applied to the analysis of four types of cancer, the results proved that TCRWMDA is an effective tool to predict the potential miRNA-disease association. The source code and dataset of TCRWMDA are available at: https://github.com/ylm0505/TCRWMDA.

Keywords: miRNA-disease association prediction, three-layer heterogeneous network, unbalanced random walk, *LncRNA*, Laplace normalization

INTRODUCTION

MiRNAs are widely found in eukaryotes and regulate the expression of other genes. miRNA is very important for the control of animal development and physiology (Victor, 2004). miRNA is involved in regulating cell differentiation (Lee et al., 1993)and plays an important role in many biological processes, including cell cycle progression and apoptosis (Brennecke et al., 2003). Mutations and biogenic dysfunction of miRNA and disorders of miRNA and its targets may lead to a variety of diseases. Calin et al. published the first study that microRNAs linked to cancer in 2002, there was a

significant association between decreased levels of both miRNAs and chronic lymphoblastic leukemia, suggesting a potential relationship between miRNA and cancer (Calin et al., 2002). miRNA is an important factor in tumorigenesis, and the artificial regulation of some miRNAs may lead to the occurrence or apoptosis of tumors, which depends on the regulation of miRNA (Yang et al., 2009). With the development of miRNA research, the association between miRNA and disease has been extended to many types of cancer, including leukemia and lung cancer (Johnson et al., 2005; Bandyopadhyay et al., 2010), breast cancer, and colon cancer (Michael et al., 2003), and so on, exploring the relationship between miRNA and disease has become the subject of many kinds of cancer research. More and more evidence proving that miRNA is closely related to diseases, understanding relationships between miRNA and disease is conducive to understanding the pathogenesis of diseases at the molecular level, but more importantly is conducive to prognosis, diagnosis, evaluation, treatment, and prevention of diseases and the promotion of human medical progress. Traditional experiments are costly, time consuming, and only suitable for small-scale data, with the development of biology, mass biological data about miRNA have been generated. There is an urgent need to develop a powerful computational method to predict the potential disease-related miRNAs, possible candidate miRNAs with higher prediction score were obtained by computational methods can reduce the time and cost of biological experiment.

In the early research methods of miRNA-disease association prediction, under the assumption that functionally related miRNAs are often related to diseases with similar phenotypes (Lu et al., 2008), A computational model based on hypergeometric distribution to predict the miRNA-disease association was proposed (Jiang et al., 2010), and constructed a heterogeneous phenome-microRNAome network for human phenome-microRNAome by combining the miRNA functional similarity network and the disease phenotype similarity network with the known miRNA-disease association, However, this method relies on the neighbor point information of the predicted miRNA, and the false positive and false negative rates are relatively high, so the prediction accuracy of this method is not high. With the development of miRNA-disease research, the restart random walk algorithm was used to predict the miRNA-disease association (RWRMDA) based on the similarity model, which is the first to use the global network to predict miRNA-disease association (Chen et al., 2012b). A restart random walk was performed on the MiRNA functional similarity network to predict potential MiRNA disease interactions, but RWRMDA did not work on any known related MiRNA disease. A semi-supervised classification method RLSMDA to predict the potential miRNA-disease association based on regularized least squares is proposed (Chen and Yan, 2015), RLSMDA is a semi-supervised model that does not require negative samples and a global approach that prioritizing the association of all diseases at the same time. Combined Within-Score with Between-Score for miRNA-disease association prediction (WBSMDA) was proposed (Chen et al., 2016), WBSMDA based on the basis of known miRNA-disease association data and assuming that miRNAs with similar functions are more likely to be associated with diseases with similar phenotypes may lead to bias (preference) on miRNAs with more known diseases, In addition, the accuracy of the model is still not very high. Then, a KNN model based on rank to predict potential related miRNAs for diseases (RKNNMDA) was proposed (Chen et al., 2017), which based on miRNA functional similarity, disease semantic similarity, Gaussian interaction profile kernel similarity and known miRNA-disease association. In RKNNMDA, k-nearest neighbor algorithm was used to search k-nearest neighbor of miRNA and disease, and these k-nearest neighbors were reordered and reweighted according to the support vector machine model to obtain the final predicted results. Random walk has also been further developed in the prediction of miRNA-disease association. The random walk technique has also been developed in association prediction, unbalanced bi-random walk on the heterogeneous networks (BRWH) based on RWR was proposed (Luo and Xiao, 2017) to predict the miRNA-disease Association. From the matrix, making use of matrix completion algorithm (MCMDA) to update the adjacency matrix based on the known miRNA-disease association data to predict its potential association proposed in (Li et al., 2017). In 2018, there is a KATZMDA model for miRNA-disease association prediction (Qu et al., 2018), which based on KATZ model to calculate miRNA similarity and disease similarity to predict the association between miRNA and disease, and KATZMDA yields better results than the previous algorithms mentioned. Based on the idea of MCMDA, a new induction matrix completion model (IMCMDA) for MiRNA-Disease Association prediction was proposed (Chen et al., 2018). Different from MCMDA, IMCMDA uses disease similarity and miRNA similarity as the characteristics of disease and miRNA to complete the missing miRNA-disease association. Recently, a kernel-based soft-neighborhood similarity model combined with similar network fusion for miRNA-disease association prediction was proposed (Ma et al., 2018a). The improvement of the similarity model improves the accuracy of predicting miRNA-disease. Ha et al. predict miRNA and disease associations based on matrix decomposition, which has been widely used in recommendation systems (Ha et al., 2019). Based on the heterogeneous network of miRNA and disease, structural perturbation method is also applied to the prediction of miRNA-disease correlation, and the final perturbed matrix represents the correlation score between the two (Zeng et al., 2018). However, these methods mentioned above only considered the miRNA-disease association data sets and functional similarity, without extracting more information from other data sets related to them to improve the accuracy and reliability of the model.

With the development of biomedicine, the number of biological databases increases, and the association between biological data is gradually excavated, which enables us to combine different information from different databases to reliably predict the miRNA-disease association. In view of the

limitations of the above methods, in this paper, we put forward a novel prediction model of three-layer network combining unbalanced random walk for miRNA-disease association prediction (TCRWMDA). Based on the known associated data of miRNA-disease, lncRNA–miRNA and lncRNA-disease, TCRWMDA build a three-layer heterogeneous network and performs unbalanced random walk between networks and on heterogeneous networks to obtain the final prediction results.

To evaluate the effectiveness of the TCRWMDA, we compared it with other classical and advanced algorithms based on the same similarity measure on 5-fold cross-validation. In addition, compared with the latest model based on the kernel-based soft neighborhood network fusion similarity model. In order to verify the applicability of TCRWMDA algorithm, four diseases were studied by TCRWMDA algorithm. Experimental results and case studies show that this method can be effectively used to predict the potential association between miRNA and disease.

MATERIALS AND METHODS

The Dataset

The associated data sets used in this article are from (Chen, 2015). The dataset mainly consists of three association data sets. First, miRNA-disease association data set is from HMDDV2.0 (Li et al., 2013), finally, 5,430 miRNA-disease associations were obtained, including 383 diseases and 495 miRNAs. A represents the known association between miRNA and disease, $A(i,j)$=1. denotes miRNA $m(i)$ is related to disease $d(j)$, otherwise, $A(i,j)$=0.

$$A(i,j) = \begin{cases} 1, \textit{if miRNA } m(i) \textit{ is associated with lncRNA } l(j) \\ 0, \textit{otherwise} \end{cases}$$

Second, the lncRNA–miRNA association dataset was derived from the star-base v2.0 database (Yang et al., 2011). Repeated associations of different evidences were deleted, as well as the lncRNA–miRNA associations that did not exist in 5,430 known miRNA-disease associations and their corresponding lncRNA–miRNA associations in the lncRNA-disease association. Finally, 704 lncRNA–miRNA associations were obtained. B represents the known relationship between lncRNA–miRNA, $B(i,j)$=1 represents miRNA $m(i)$ is related to lncRNA $l(j)$,otherwise, $B(i,j)$=0.

$$B(i,j) = \begin{cases} 1, \textit{if miRNA } m(i) \textit{ is associated with IncRNA } I(j) \\ 0, \textit{otherwises} \end{cases}$$

Third, the lncRNA-disease association data set in the lncRNA Disease database (Geng Chen et al., 2012a) was downloaded, and the repeated association of different evidences and the association of lncRNA-disease related to the disease or lncRNA were removed. After removing the data of diseases not shown in the above data set, 182 lncRNA-disease associations of 34 lncRNAs were finally obtained. C represents association matrix between lncRNA and disease, $C(i,j)$=1 denotes lncRNA $l(i)$ is related to disease $d(j)$, otherwise, $C(i,j)$=0.

$$C(i,j) = \begin{cases} 1, \textit{if lncRNA } l(i) \textit{ associated with disease } d(j) \\ 0, \textit{otherwise} \end{cases}$$

TCRWMDA

Based on the idea of unbalanced bi-random walk, we proposed three-layer heterogeneous network combined with unbalanced random walk for miRNA-disease association prediction algorithm. TCRWMDA algorithm includes three random walks, including the random walk on miRNA–miRNA network, disease similarity network, and the mapping relationship of miRNA–lncRNA-disease. **Figure 1** shows the flow chart of TCRWMDA algorithm to predict miRNA-disease association. In the dotted black box above **Figure 1**, blue dots represent miRNA, yellow dots represent disease, and red dots represent lncRNA. A three-layer heterogeneous network consist of the similar networks formed by same color nodes with straight lines and the heterogeneous networks formed by nodes of different colors with dotted lines. The similarity measure can be obtained by calculating the similarity of association data, the similarity measure was use to obtain the transition probability matrix by Laplace normalization, finally, TCRWMDA algorithm using the transition probability matrix to unbalanced random walk on heterogeneous network to get the potential association scores between the disease and its associated miRNAs and sorting. The feasibility and effectiveness of the algorithm is verified by whether the predicted results already exist in the existing database.

Construction of Similarity Networks

The similarity networks in this paper consist of lncRNA similarity network, Disease similarity network, miRNA similarity network.

lncRNA Similarity Network

Genes can be mutated, inserted and deleted, it is difficult to achieve a complete match of two sequences, so we use sequence information as its feature. We extract the sequence features by considering sequence composition (Zhang et al., 2018). For lncRNA sequences, we calculated the proportion of four nucleotide types (A, C, G, T) and 16 dinucleotide types (AA, AG, AC…) in each lncRNA sequence, every lncRNA $l(i)$ can get a 20−dimensional eigenvector, where (i) is its component, named as lncRNA sequence composition. The sequence data of 34 selected lncRNA were downloaded from LNCipedia5 (Volders et al., 2019). Use cosine similarity method to calculate the lncRNA similarity sl, the formula of lncRNA similarity is as follows:

$$sl(i,j) = \frac{\sum_{i=1}^{20} \mathcal{L}(i) \times \mathcal{L}(j)}{\sqrt{\sum_{i=1}^{20} (\mathcal{L}(i))^2} \times \sqrt{\sum_{j=1}^{20} (\mathcal{L}(j))^2}}$$

Disease Similarity Network

In this paper, we used the same method as in literature (Wang et al., 2010) to calculate the disease similarity.

Disease semantic similarity model 1: Directed acyclic graph (DAG) was constructed to describe the disease based on MeSH

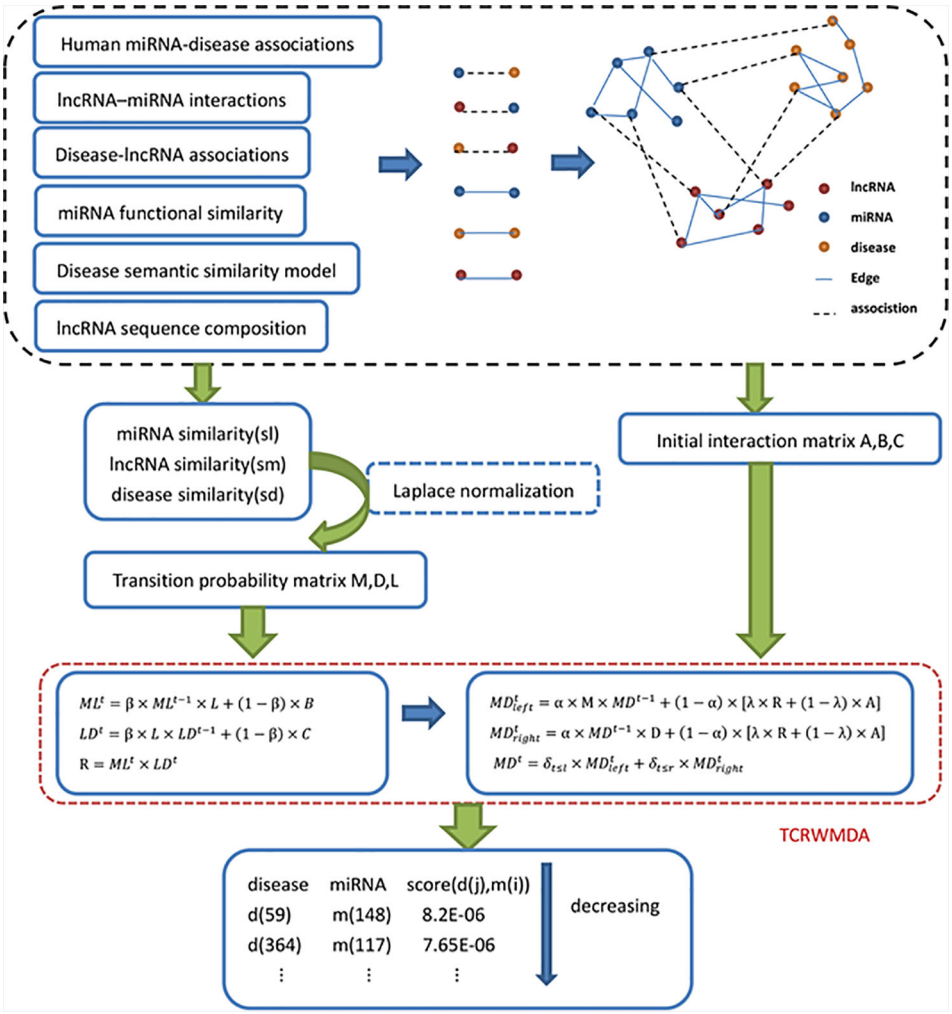


FIGURE 1 | Flow chart of TCRWMDA algorithm. The steps of TCRWMDA for the association prediction between miRNA and disease are divided into four stages: the construction of similarity network, the calculation of transition probability matrix and the random walk on the three-layer heterogeneous network. Finally, the final prediction score is obtained to analyze the association probability of a certain disease and a certain miRNA. In the black dotted box is the construction of similarity network, which are based on association data and related data from the available database. The red dotted line shows that an unbalanced random walk on a three-layer heterogeneous network.

descriptor downloaded from national library of medicine (Lipscomb, 2000) (http://www.nlm.nih). According to DAG, the contribution of disease d to the semantic value of disease d DAG (d) is expressed as:

$$\begin{cases} D1_D(d) = 1 & if\ d = D \\ D1_D(d) = \{\Delta * D1_D(d') | d' \in children\ \ of\ \ d\} & if\ d\ \ \neq D \end{cases}$$

Δ denotes attenuation coefficient of semantic contribution. The self-semantic value of disease D was defined as follows:

$$\mathrm{DV1(D)} = \sum_{d \in T(D)} D1_D(d)$$

Where T(D) represents all ancestor nodes of D and D itself. Based on the assumption that the two diseases share a large part of DAG and their semantic similarity is large, the semantic similarity between disease $d(i)$. and disease $d(j)$ can be defined as:

$$\mathrm{SS1}(d(i), d(j)) = \frac{\sum_{d \in T(d(i)) \cap T(d(j))} \left(D1_{d(i)}(d) + D1_{d(j)}(d)\right)}{\mathrm{DV1}(d(i)) + \mathrm{DV1}(d(j))}$$

Disease semantic similarity model 2: It is unreasonable to give the same contribution value for diseases in the same layer of DAG (D). Therefore, according to the model proposed by Xuan et al., we define the contribution of disease d to the semantic value of disease d in DAG (d) as follows:

$$D2_D(d) = -\log\left[\frac{the\ \ number\ \ of\ \ DAGs\ \ including\ \ d}{the\ \ number\ \ of\ \ diseases}\right]$$

We define the semantic similarity of diseases $d(i)$,$d(j)$ as the ratio of share ancestor node contributions to all ancestor node contributions. The semantic similarity model 2 is calculated as follows:

$$SS2(d(i),d(j)) = \frac{\sum_{d\in T(d(i))\cap T(d(j))}\left(D2_{d(i)}(d) + D2_{d(j)}(d)\right)}{DV2(d(i)) + DV2(d(j))}$$

Among them,

$$DV2(D) = \sum_{d\in T(D)} D2_D(d)$$

miRNA Similarity Network

Wang et al. (2010) proposed the method of MISIM and miRNA functional similarity based on the hypothesis that miRNAs with similar functions are more likely to be associated with diseases with similar characteristics. The miRNA function similarity data downloaded from http://www.cuilab.cn/files/images/cuilab/misim.zip. We use $FS(m(i),m(j))$ to represent association score between miRNA $m(i)$ and miRNA $m(j)$

Gaussian Kernel Similarity

Based on the basic assumption that similar diseases are often associated with miRNAs with similar functions (Wang et al., 2010), we calculated the Gaussian kernel similarity for miRNA and disease to obtain the miRNA similarity and disease similarity. First, we use vector $IP(d(i))$ to represent there is or is not an association between each miRNA and disease $d(i)$ and regard $IP(d(i))$ as interaction profile of the disease $d(i)$, then, the gaussian interaction profile kernel similarity between disease $d(i)$ and $d(j)$ was calculated:

$$kd(i,j) = exp\left(-\gamma_d \,||\, IP(d(i)) - IP(d(j)) \,||^2\right)$$
$$\gamma_d = \gamma_d' / \left(\frac{1}{nd}\sum_{i=1}^{nd} ||IP(d(i))||^2\right)$$

γ_d controls kernel bandwidth. Similarly, the Gaussian kernel similarity between disease $m(i)$ and disease $m(j)$ can be obtained as follows:

$$km(i,j) = exp\left(-\gamma_m \,||\, IP(m(i)) - IP(m(j)) \,||^2\right)$$
$$\gamma_{m=}\gamma_m' / \left(\frac{1}{nm}\sum_{i=1}^{nm} ||IP(m(i))||^2\right)$$

Integrated Similarity for Diseases and miRNAs

We could not obtain the DAGs of all diseases, that is, for a specific disease without DAG, we could not calculate the semantic similarity score of this disease with other diseases. Therefore, for the disease pairs with semantic similarity score, we used the semantic similarity score to express the disease similarity, and for other disease pairs, we used the gaussian kernel interaction profile similarity to represent the disease similarity. The disease similarity matrix of disease $d(i)$ and disease $d(j)$ was constructed as follows:

$$sd(i,j) = \begin{cases} \frac{SS1(d(i),d(j))+SS2(d(i),d(j))}{2} & d(i) \text{ and } d(j) \text{ has semantic similarity} \\ kd(d(i),d(j)) & \text{otherwise} \end{cases}$$

Similarly, the similarity matrix of miRNA can be obtained:

$$sm(i,j) = \begin{cases} FS(m(i),m(j)) & m(i) \text{ and } m(j) \text{ has functional similarity} \\ km(m(i),m(j)) & \text{otherwise} \end{cases}$$

The similarity between the two miRNAs is the weight of edge in the miRNA similarity network, in the same way, the similarity between the two diseases is the weight of edge in the disease similarity network.

Calculation of Transition Probability Matrix

To perform a random walk on three-layer heterogeneous networks, the state transition between networks must be considered and transition probability matrix needs to be created. To calculate the transition probability in the miRNA similarity network, we make use of the Laplace normalization (Zhao et al., 2015) to calculate transition probability matrix in the miRNA similarity network, and the exit degree of nodes and the entry degree of nodes were taken into account.

Laplace normalization: Assuming that $Z=[(i,j)],i,j=1,2,\ldots,N$ is a symmetric matrix, Y is a diagonal matrix, defined as: Y (i,i) is the sum of the i row of Z, When i is not equal to j,Y $(i,j)=0$. Matrix normalization: $\mathbb{Z}=Y^{-1/2}AY^{-1/2}$ also a symmetric matrix, The elements in can be defined as:

$$\mathbb{Z}(i,j) = \frac{Z(i,j)}{\sqrt{Y(i,i)Y(j,j)}}$$

Then the transition probability matrix M in the miRNA similarity network can be expressed as:

$$M(i,j) = \begin{cases} \frac{sm(i,j)}{\sqrt{\sum_i sm(i,j)\sum_j sm(i,j)}} & if \sum_i sm(i,j) and \sum_j sm(i,j) \neq 0 \\ 0 & otherwise \end{cases}$$

Similarly, we can obtain the transition probability matrix D and L in the disease similarity network and lncRNA similarity network as follows:

$$D(i,j) = \begin{cases} \frac{sd(i,j)}{\sqrt{\sum_i sd(i,j)\sum_j sd(i,j)}} & if \sum_i sd(i,j) and \sum_j sd(i,j) \neq 0 \\ 0 & otherwise \end{cases}$$

$$L(i,j) = \begin{cases} \frac{sl(i,j)}{\sqrt{\sum_i sl(i,j)\sum_j sl(i,j)}} & if \sum_i sl(i,j) and \sum_j sl(i,j) \neq 0 \\ 0 & otherwise \end{cases}$$

TCRWMDA Algorithm Process

Specifically, TCRWMDA algorithm can be divided into two parts: one is random walk on heterogeneous networks, and the other is random walk between networks. **Table 1** introduces the process of TCRWMDA algorithm in predicting miRNA-disease association, and **Table 2** introduces the process of unbalanced random walk between networks.

Random Walk on Three-Layer Heterogeneous Networks

Where MD represents the predicted correlation matrix between miRNA and disease, MD^t represents t-step random walk were

TABLE 1 | The description of the TCRWMDA algorithm.

Algorithm 1 TCRWMDA (Random Walk on three-layer heterogeneous network)

Input: Transition probability matrix M, D, L; Initial association matrix A, B, C; Parameter α, λ, β, *l, r, s*.
Output: Predicted miRNA-disease association matrix MD
1: MD^0 =A/sum(A)
2: for t=1 to max (*l, r*)
3: MD'=MD
4: if $t \leq l$ then
5: $MD^t_{left} = \alpha \times M \times MD^{t-1} + (1-\alpha) \times [\lambda \times BNetWalk(B, C, L, \beta, S) + (1-\lambda) \times A]$
6: end if $t \leq r$ then
7: $MD^t_{right} = \alpha \times MD^{t-1} \times D + (1-\alpha) \times [\lambda \times BNetWalk(B, C, L, \beta, S) + (1-\lambda) \times A]$
8: $MD^t = \delta_{t \leq l} \times MD^t_{left} + \delta_{t \leq r} \times MD^t_{right}$
9: end for
10: return MD

E is identity matrix, if ≤x, $\delta_{t \leq x}$ is 1, and 0 otherwise.

TABLE 2 | The description of the BNetWalk algorithm.

Algorithm 2 BNetWalk (Random Walk between networks)

Input: Transition probability matrix L; Initial association matrix B and C; parameter β,s

Output: Predicted miRNA-disease association matrix R
1: ML^0 =B/sum(B), LD^0 =C/sum(C)
2: for t=1 to s
3: R′=*R*
4: $ML^t = \beta \times ML^{t-1} \times L + (1-\beta) \times B$
5: $LD^t = \beta \times L \times LD^{t-1} + (1-\beta) \times C$
6: $R = ML^t \times LD^t$
7: end for
7: return R

performed MD, *A*、 *B*、 *C* denotes matrix of prior knowledge. TCRWMDA algorithm has six parameters: α, β, *l, r, s. l, r.* s represents the number of steps random walk on miRNA-miRNA network, disease-disease network and networks respectively. α controls network walk or return to the proportion of prior knowledge; The function of λ is to provide a new priori knowledge, there is a linearly combination of the new state form by a random walk between networks and the known initial state by λ. That is, if the current particle is in the miRNA network, then the particle has probability of α to perform the *l*-step random walk in the miRNA network, to perform the *l*-step random walk (1-α)×λ perform the s-step random walk into disease network, and has probability of (1-)×(1-λ) to return the start node. If the current particle is in the disease network, then the particle has probability of α to perform the *r*-step random walk in the disease network, has probability of (1-α)×λ perform the *s*-step random walk into miRNA network.

Random Walk Between Networks

ML represents the predicted association score between miRNA and lncRNA, while LD represents the probability matrix of disease generation on lncRNA. β notes the probability of controlling the random walk on the lncRNA network or returning to prior knowledge during random walk among networks. R represents the miRNA-disease association matrix formed through Random Walk between networks.

ML^t and LD^t represents t-step random walks were performed ML and LD, respectively. In equation (18), the association matrix between miRNA and lncRNA is multiplied by the right transition probability matrix L on the lncRNA network, which represents a random walk on lncRNA network to update ML. Similarly, the left multiplication probability transition matrix L represents a random walk on lncRNA network to update LD, finally, we can obtain association between miRNA and disease.

RESULTS AND ANALYSIS

Parameter Analysis

Receiver operating characteristic curve (ROC curve) takes true positive rate (sensitivity) as the vertical coordinate and false positive rate (1-specificity) as the horizontal coordinate. The area under the ROC curve is the AUC value, which can be used as the evaluation index to intuitively evaluate the classifier. The higher the AUC value, the better the performance of the algorithm. In the process of parameter selection, AUC value is selected as the index to evaluate the influence of parameters. For an algorithm, if the parameters are set with different values, it corresponds to different models. For which model to choose, the best way is to use the model with the minimum generalization error. However, it is generally impossible to directly obtain the generalization error of the model, we select the model parameter when the AUC value is the largest.

TCRWMDA has six parameters, set step size of α, β and λ is 0.1, with values ranging from 0 to 1. For *l, r* and *s*, set the step size to 1 and the value range to 1–5. The known association between 495 miRNAs and 383 diseases verified by 5-fold cross validation. First, fix some parameters, change the value of a parameter, and then the influence of parameters on the model performance was determined according to the change of AUC value. In the process of parameter selection, the value of s was changed in the experiment, and the AUC value did not change much. The increase in the number of steps in the network could not provide us with more information, and the information that could be mined was limited. Moreover, the larger *s* was, the higher the algorithm complexity, and the performance of the model barely changed as s increased, so we set $s = 1$ in this paper, which also indicates that the data volume in the lncRNA data set is too small to provide more network structure information.

Change the values of *l* and *s* and fix other parameters. The change result of AUC is shown in **Figure 2**. For parameters *l* and *r*, the results are significantly better when $l \geq r$ than when $l < r$. Fixed *l*, with the increase of *r*, the AUC value decreased significantly, which indicated that excessive walking on the disease network would lead to a certain false positive, and the overall performance decreased. According to the results of parameter analysis, we set $l = 1$ and $r = 1$.

Next, fix $l = 1$, $r = 1$, $s = 1$, Change the values of α, β, and λ, the experimental results are shown in **Figure 3**. α denotes restart probability, when $\alpha = 0$, only random walk between networks

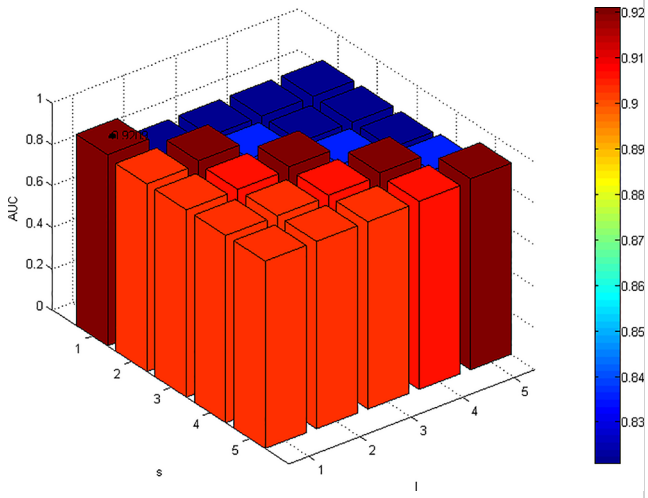


FIGURE 2 | Effects of parameters l and r on the result of 5-fold cross validation. (α=0.1, β=0.1, s=1, λ=0.9). When the value of α, β, s, and λ are fixed, the AUC value is maximized when *l* and *r* are both equal to 1.

played a role, ignoring the random walk between the miRNA network itself and the heterogeneous network on the disease network. Therefore, the results of the model were not ideal, but the remaining values of AUC were 0.9205~0.9209, with no significant fluctuation. When β = 0.1, the AUC value is the maximum and the model performance is the best. When the parameter β is larger, the probability of prior knowledge is reduced. The known association information is gradually ignored, and the results presented are reduced, which indicates that the known association information plays an important role in the algorithm itself and cannot be ignored. Parameter λ has little influence on the model, when λ = 0.9, AUC is the largest. From what has been discussed above, we select $l = 1$, $r = 1$, $s = 1$, $\alpha = 0.1$, $\beta = 0.1$, $\lambda = 0.9$.

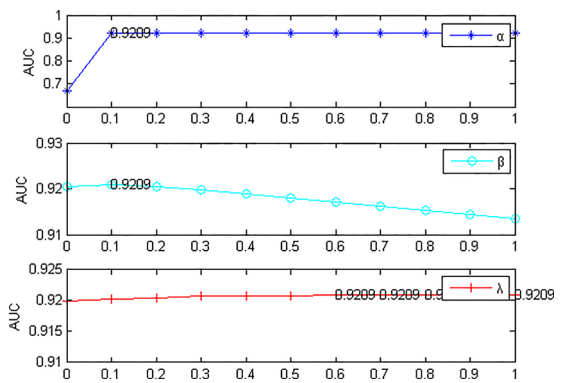


FIGURE 3 | Effects of parameters α, β and λ on the result of 5-fold cross validation. (l=1, r=1, s=1). When α=0, AUC value is the lowest. In this case, only the random walk between the networks is at work. AUC is relatively stable with the variation of parameters β and λ.

Algorithm Performance Comparison

In this paper, we take the AUC (Area under Curve) value as the evaluation index, all known miRNA-disease associations were divided into five groups of the same size, four of which were used as training set for model learning, then, the similarity calculation method mentioned above was used to calculate miRNA and disease similarity, we compare TCRWMDA with IMCMDA (Chen et al., 2018), RWRMDA(Xing Chen et al., 2012), KATZMDA (Qu et al., 2018), BRWH (Luo and Xiao, 2017) for 5-fold cross validation. The results of TCRWMDA and other methods for 5-fold cross validation are shown in **Figure 4**. True positive rate (sensitivity) is the percentage of a test sample ranked above a given threshold. False positive rate (1-specificity) is the percentage of samples below the threshold. In this paper, for the specified threshold, the true positive rate is the percentage that accurately predicts the miRNA associated with a known disease, and the false positive rate is the percentage that predicts the miRNA unrelated to the disease. When AUC = 1, the performance of the model is the best. When AUC = 0.5, it indicates that the classification method is completely ineffective and has no classification value.

It can be seen from **Figure 4**, the area under the ROC curve of TCRWMDA algorithm is the largest, that is, the prediction performance of this algorithm is better than other methods. The AUC values obtained by IMCMDA (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018) (Chen et al., 2018), RWRMDA, KATZMDA, BRWH, and TCRWMDA on 5-fold cross validation are respectively 0.8351、0.8676、0.9088、0.9106、0.9209. The AUC value of the TCRWMDA algorithm was 1.3% higher than that of the BRWH, which indicates add new related dataset and perform a random walk on constructed multi-layer network and then is effective. TCRWMDA is 10.3% better than IMCMDA, 6.1% better than RWRMDA, and 1.1% better than KATZMDA.

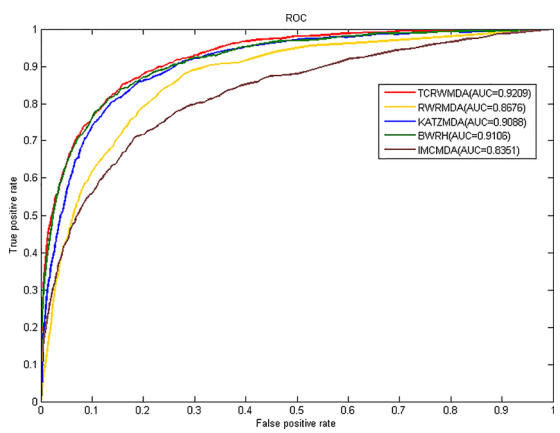


FIGURE 4 | The AUC value of TCRWMDA and other methods for 5-fold cross validation.

Based on Kernel-Based Soft-Neighborhood Network Fusion Similarity Model

Ma et al. considered the distance factor and the reconstruction relationship between samples to establish the nuclear soft neighborhood similarity model (Ma et al., 2018b), and combined the nuclear soft neighborhood similarity matrix of miRNA (disease) with the functional similarity (disease semantic similarity) of miRNA using similarity network fusion (SNF) (Wang et al., 2014), proposed kernel-based soft-neighborhood network fusion similarity model, and obtained good results. The following analysis based on kernel-based soft-neighborhood network fusion similarity model. After parameter analysis, the final selection is $l = 1$, $r = 1$, $s = 1$, $\alpha = 0.2$, $\beta = 0.1$, $\lambda = 0.9$.

Figure 5 shows the results of TCRWMDA and LKSNF soft neighborhood network of nuclear fusion based similarity model on 5-fold cross validation. In **Figure 5**, the red solid line represents the result of TCRWMDA algorithm for 5-fold cross validation, the green dotted line represents the result of TCRWMDA algorithm based on kernel-based soft-neighborhood network fusion similarity model, and the black dotted line represents the result of the LKSNF algorithm on 5-fold cross-validation. Based on kernel-based soft-neighborhood network fusion similarity model, the AUC value of the TCRWMDA algorithm is improved by 0.99%. However, the association data of lncRNA-miRNA and lncRNA-disease are sparse, the number of lncRNAs that can be considered is also small, resulting in a certain deviation in the prediction results, the AUC value obtained by TCRWMDA algorithm is almost the same as that obtained by LKSNF algorithm.

Case Study

Globally, breast cancer is the most common cancer in women, accounting for 25% of all cancers in women. In 2012, there were 1.68 million cases of breast cancer and 520,000 deaths due to breast cancer. Mir-200c inhibits the growth and differentiation of cancer cells, and strongly inhibits the ability of normal breast stem cells to form mammary ducts and human breast cancer-driven tumorigenesis *in vivo* (Shimono et al., 2009). In addition, miRNA may be abnormally downregulated or upregulated in colon cancer tissues. In 2003, the first study on miRNAs was published in colon cancer (Michael et al., 2003), identifying mir-143 and mir-145 as new misaligned miRNAs in colon cancer.

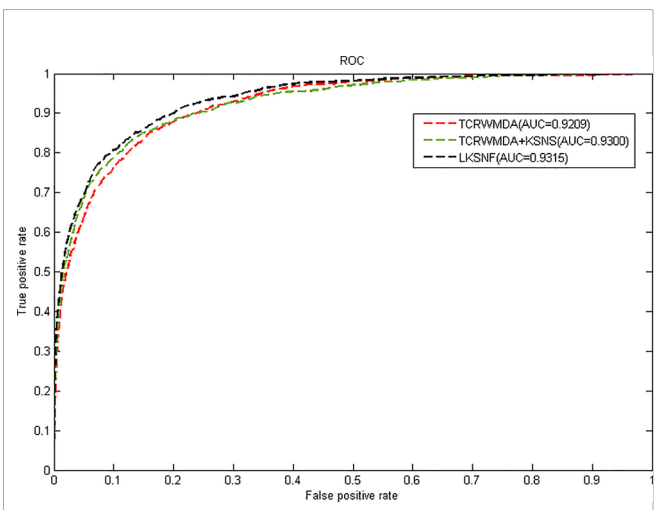


FIGURE 5 | The AUC value of TCRWMDA and LKSNF for 5-fold cross validation based on kernel-based soft-neighborhood network fusion similarity model.

In order to further prove the predictive performance of TCRWMDA in predicting miRNA-disease association, we used TCRWMDA algorithm to carry out analysis of breast cancer and colon cancer, as shown in **Tables 3** and **4**.

The predicted results were verified by dbDEMC database (Yang et al., 2017) and HMDD (Li et al., 2013), for breast tumor diseases, 44 of the first 50 predicted miRNAs were verified in dbDEMC and 45 of the top 50 predicted colon tumor diseases were verified by dbDEMC. In order to enhance the persuasion, we also listed two other cases (lung neoplasms and lymphoma), whose prediction results were verified as shown in the **Supplementary Tables 1** and **2**.

CONCLUSION

With the development of bioinformatics, more and more experiments and evidence show that miRNA is closely related to the generation and development of human diseases, and the discovery of miRNA that may be related to diseases has attracted much attention. The experiment is time-consuming and costly, the new and effective miRNA-disease association prediction

TABLE 3 | The top 50 potential miRNAs predicted by TCRWMDA for breast neoplasms and their associations confirmed by database (column 1: top 1–25; Column 3: top 26–50).

miRNA	Evidence	miRNA	Evidence
hsa-mir-106a	dbDEMC	hsa-mir-454	dbDEMC
hsa-mir-130a	dbDEMC	hsa-mir-421	dbDEMC
hsa-mir-15b	dbDEMC	hsa-mir-181d	dbDEMC
hsa-mir-150	dbDEMC	hsa-mir-216a	dbDEMC
hsa-mir-192	dbDEMC	hsa-mir-330	dbDEMC
hsa-mir-142	unconfirmed	hsa-mir-451	dbDEMC
hsa-mir-130b	dbDEMC	hsa-mir-544a	dbDEMC
hsa-mir-372	dbDEMC	hsa-mir-181c	dbDEMC
hsa-mir-196b	dbDEMC	hsa-mir-198	dbDEMC
hsa-mir-98	dbDEMC	hsa-mir-376a	dbDEMC
hsa-mir-92b	dbDEMC	hsa-mir-211	dbDEMC
hsa-mir-30e	unconfirmed	hsa-mir-363	dbDEMC
hsa-mir-32	dbDEMC	hsa-mir-455	unconfirmed
hsa-mir-186	dbDEMC	hsa-mir-490	unconfirmed
hsa-mir-99b	dbDEMC	hsa-mir-494	dbDEMC
hsa-mir-424	dbDEMC	hsa-mir-381	dbDEMC
hsa-mir-212	dbDEMC	hsa-mir-154	dbDEMC
hsa-mir-449a	dbDEMC	hsa-mir-216b	dbDEMC
hsa-mir-449b	dbDEMC	hsa-mir-370	dbDEMC
hsa-mir-99a	dbDEMC	hsa-mir-520e	dbDEMC
hsa-mir-491	unconfirmed	hsa-mir-484	dbDEMC
hsa-mir-28	dbDEMC	hsa-mir-217	dbDEMC
hsa-mir-151	HMDD	hsa-mir-302e	dbDEMC
hsa-mir-144	dbDEMC	hsa-mir-590	unconfirmed
hsa-mir-95	dbDEMC	hsa-mir-377	dbDEMC

TABLE 4 | The top 50 potential miRNAs predicted by TCRWMDA for colon cancer (colon neoplasms) and confirmed by database (column 1: top 1–25; Column 3: top 26–50).

miRNA	Evidence	miRNA	Evidence
hsa-mir-21	dbDEMC	hsa-mir-200a	unconfirmed
hsa-mir-20a	dbDEMC	hsa-mir-31	dbDEMC
hsa-mir-16	dbDEMC	hsa-mir-137	dbDEMC
hsa-mir-155	dbDEMC	hsa-mir-205	dbDEMC
hsa-mir-29a	dbDEMC	hsa-mir-148a	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-10b	dbDEMC
hsa-mir-143	dbDEMC	hsa-mir-125a	dbDEMC
hsa-mir-19a	dbDEMC	hsa-mir-486	dbDEMC
hsa-mir-146a	dbDEMC	hsa-let-7b	dbDEMC
hsa-mir-18a	dbDEMC	hsa-let-7f	dbDEMC
hsa-let-7a	dbDEMC	hsa-mir-375	dbDEMC
hsa-mir-200c	unconfirmed	hsa-mir-22	dbDEMC
hsa-mir-34a	dbDEMC	hsa-mir-24	dbDEMC
hsa-mir-92a	dbDEMC	hsa-mir-27a	dbDEMC
hsa-mir-9	dbDEMC	hsa-mir-214	dbDEMC
hsa-mir-222	dbDEMC	hsa-mir-183	dbDEMC
hsa-mir-125b	dbDEMC	hsa-mir-18b	dbDEMC
hsa-mir-196a	dbDEMC	hsa-mir-140	dbDEMC
hsa-let-7c	dbDEMC	hsa-mir-7	dbDEMC
hsa-mir-107	dbDEMC	hsa-mir-142	unconfirmed
hsa-let-7e	dbDEMC	hsa-let-7i	dbDEMC
hsa-mir-141	dbDEMC	hsa-mir-25	dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-199a	unconfirmed
hsa-mir-93	dbDEMC	hsa-mir-133b	dbDEMC
hsa-mir-223	unconfirmed	hsa-mir-29c	dbDEMC

algorithm can effectively provide research directions and reduce the cost and time of biological experiments.

In this paper, we propose a novel TCRWMDA algorithm, which is different from the traditional prediction methods based on heterogeneous network and incorporates new prior knowledge (lncRNA information related to miRNA and disease) to effectively make the best use of the information that we have. TCRWMDA is a framework for integrating multiple sources of information, which may yield better results when the data set is large. TCRWMDA is applied to miRNA-disease association prediction, which implements unbalanced random walk on three-layer heterogeneous networks and integrate the related similarity information to predict disease-related miRNAs. TCRWMDA is efficient because it makes use of multi-source information from reliable data sources. Considering the association between lncRNA and disease and the association between miRNA and disease, TCRWMDA mines the association information on between data and topological information in the network to improve the prediction accuracy. Experimental results and case studies prove that the TCRWMDA algorithm is an effective tool for predicting the potential miRNA-disease association. If more data sets are added, the increase and optimization of parameters is a problem worth thinking about. In the future, we hope to conduct more stable data integration and seek methods for optimizing parameter selection.

DATA AVAILABILITY STATEMENT

All datasets for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

LY and XS designed and implemented the computing framework. LY and XS analyzed the results and wrote the manuscript. LY, XS, DZ and JY revised the manuscript. LY prepared the computational codes and carried out. All the authors wrote, reviewed and approved the final manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (61532008, 61872157, 61932008), the Self-determined Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (CCNU19QD003) and the National Language Commission Key Research Project (ZDI135-61).

ACKNOWLEDGMENTS

This article was included in the CBC2019.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01316/full#supplementary-material

REFERENCES

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. Q. (2010). Development of the human cancer microRNA network. *Silence* 1, 6–6. doi: 10.1186/1758-907X-1-6

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., and Cohen, S. M. (2003). bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell* 113, 25–36. doi: 10.1016/S0092-8674(03)00231-9

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., et al. (2002). Frequent Deletions and Down-Regulation of Micro-RNA Genes miR15 and miR16 at 13q14 in Chronic Lymphocytic Leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15524–15529. doi: 10.1073/pnas.242606799

Chen, X., and Yan, G. (2015). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.-UK* 4, 5501. doi: 10.1038/srep05501

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., et al. (2012a). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986. doi: 10.1093/nar/gks1099

Chen, X., Liu, M., and Yan, G. (2012b). RWRMDA: predicting novel human microRNA–disease associations. *Mol. Biosyst.* 8 (2792), 21106. doi: 10.1039/c2mb25180a

Chen, X., Yan, C. C., Zhang, X., You, Z., Deng, L., Liu, Y., et al. (2016). WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. *Sci. Rep.-UK* 6, 21106. doi: 10.1038/srep21106

Chen, X., Wu, Q. F., and Yan, G. Y. (2017). RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction. *RNA Biol.* 14, 952–962. doi: 10.1080/15476286.2017.1312226

Chen, X., Wang, L., Qu, J., Guan, N., and Li, J. (2018). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

Chen, X. (2015). Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.-UK* 5, 13186. doi: 10.1038/srep13186

Ha, J., Park, C., and Park, S. (2019). PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach. *BMC Syst. Biol.* 13, 33. doi: 10.1186/s12918-019-0700-4

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4 Suppl 1, S2–S2. doi: 10.1186/1752-0509-4-S1-S2

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., et al. (2005). RAS Is Regulated by the let-7 MicroRNA Family. *Cell* 120, 635–647. doi: 10.1016/j.cell.2005.01.014

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843. doi: 10.1016/0092-8674(93)90529-y

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2013). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Li, J. Q., Rong, Z. H., Chen, X., Yan, G. Y., and You, Z. H. (2017). MCMDA: Matrix completion for MiRNA-disease association prediction. *Oncotarget* 8, 21187. doi: 10.18632/oncotarget.15061

Lipscomb, C. E. (2000). "Medical Subject Headings (MeSH).", in.).

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An Analysis of Human MicroRNA and Disease Associations. *PloS One* 3, e3420. doi: 10.1371/journal.pone.0003420

Luo, J., and Xiao, Q. (2017). A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. *J. Biomed. Inform.* 66, 194–203. doi: 10.1016/j.jbi.2017.01.008

Ma, Y., Ge, L., Ma, Y., Jiang, X., He, T., and Hu, X. (2018a). Kernel soft-neighborhood network fusion for miRNA-disease interaction prediction, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 197–200. doi: 10.1109/BIBM.2018.8621122

Ma, Y., Yu, L., He, T., Hu, X., and Jiang, X. (2018b). Prediction of long non-coding RNA-protein interaction through kernel soft-neighborhood similarity, in: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 193–196. doi: 10.1109/BIBM.2018.8621460

Michael, M. Z., Young, G. P., and James, R. J. (2003). Reduced accumulation of specific microRNAs in colorectal neoplasia. *Mol. Cancer Res.* 1, 882–891. doi: 10.1007/s00268-008-9865-5

Qu, Y., Zhang, H., Liang, C., and Dong, X. (2018). KATZMDA: prediction of miRNA-disease associations based on KATZ model. *IEEE Access* 6, 3943–3950. doi: 10.1109/ACCESS.2017.2754409

Shimono, Y., Zabala, M., Cho, R. W., Lobo, N., Dalerba, P., Qian, D., et al. (2009). Downregulation of miRNA-200c links breast cancer stem cells with normal stem cells. *Cell* 138, 592–603. doi: 10.1016/j.cell.2009.07.011

Victor, A. (2004). The functions of animal microRNAs. *Nature* 431, 350–355. doi: 10.1038/nature02871

Volders, P., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., et al. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* 47, D135–D139. doi: 10.1093/nar/gky1031

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Yang, L., Belaguli, N., and Berger, D. H. (2009). MicroRNA and Colorectal Cancer. *World J. Surg.* 33, 638–646. doi: 10.1007/s00268-008-9865-5

Yang, J., Li, J., Shao, P., Zhou, H., Chen, Y., and Qu, L. (2011). starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.* 39, D202–D209. doi: 10.1093/nar/gkq1056

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* 45, D812–D818. doi: 10.1093/nar/gkw1079

Zeng, X., Liu, L., Lü, L., Zou, Q., and Valencia, A. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics*. 34. doi: 10.1093/bioinformatics/bty112.

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhao, Z. Q., Han, G. S., Yu, Z. G., and Li, J. (2015). Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem.* 57, 21–28. doi: 10.1186/s12918-018-0660-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 15 January 2020
doi: 10.3389/fbioe.2019.00479

Schizophrenia Identification Using Multi-View Graph Measures of Functional Brain Networks

Yizhen Xiang[1], *Jianxin Wang*[1,2], *Guanxin Tan*[1], *Fang-Xiang Wu*[3] *and Jin Liu*[1]*

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, China, [3] Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada

Schizophrenia (SZ) is a functional mental disorder that seriously affects the social life of patients. Therefore, accurate diagnosis of SZ has raised extensive attention of researchers. At present, study of brain network based on resting-state functional magnetic resonance imaging (rs-fMRI) has provided promising results for SZ identification by studying functional network alteration. However, previous studies based on brain network analysis are not very effective for SZ identification. Therefore, we propose an improved SZ identification method using multi-view graph measures of functional brain networks. Firstly, we construct an individual functional connectivity network based on Brainnetome atlas for each subject. Then, multi-view graph measures are calculated by the brain network analysis method as feature representations. Next, in order to consider the relationships between measures within the same brain region in feature selection, multi-view measures are grouped according to the corresponding regions and Sparse Group Lasso is applied to identify discriminative features based on this feature grouping structure. Finally, a support vector machine (SVM) classifier is employed to perform SZ identification task. To evaluate our proposed method, computational experiments are conducted on 145 subjects (71 schizophrenic patients and 74 healthy controls) using a leave-one-out cross-validation (LOOCV) scheme. The results show that our proposed method can obtain an accuracy of 93.10% for SZ identification. By comparison, our method is more effective for SZ identification than some existing methods.

Keywords: Schizophrenia identification, fMRI, functional brain networks, multi-view graph measures, SVM

OPEN ACCESS

Edited by:
Fa Zhang,
Institute of Computing Technology (CAS), China

Reviewed by:
Leyi Wei,
Tianjin University, China
Renmin Han,
Shandong University, China

***Correspondence:**
Jin Liu
liujin06@csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *05 September 2019*
Accepted: *23 December 2019*
Published: *15 January 2020*

Citation:
Xiang Y, Wang J, Tan G, Wu F-X and Liu J (2020) Schizophrenia Identification Using Multi-View Graph Measures of Functional Brain Networks. Front. Bioeng. Biotechnol. 7:479. doi: 10.3389/fbioe.2019.00479

1. INTRODUCTION

Schizophrenia (SZ) is a functional mental disorder which caused by genetic factors and environmental effects. Patients with SZ (SZs) share some common symptoms which include depression, hallucinations, cognitive dysfunction and disorganized thinking (Marín, 2012). Impairments of this disorder cover multiple cognitive areas, including memory (He et al., 2012), attention and executive function (Heinrichs and Zakzanis, 1998). One percent of the population is affected by the serious psychiatric disease worldwide (Ripke et al., 2013). The clinical diagnosis of SZ relies mainly on mental state examination rather than any biomarker (Arbabshirani et al., 2013; Liu et al., 2017d) since the cause and mechanism of the disease are not clearly revealed. However, this diagnosis method is usually subjective and not completely effective. Therefore, it is urgent to find an objective method to realize the automatic diagnosis of SZ and improve the accuracy of recognition.

Nowadays, Magnetic resonance imaging technology has been widely used in various studies related to brain disease diagnosis (Nieuwenhuis et al., 2012; Liu et al., 2016, 2017b,c, 2018a; Yang and Wang, 2018). Since SZ is reported to be a functional disease, functional magnetic resonance imaging (fMRI) is increasingly used to study brain dysfunction in patients with mental illness (Castro et al., 2011; Huang et al., 2018; Liu et al., 2018b; Moghimi et al., 2018; Chen et al., 2019). In addition, fMRI provides a database for functional analysis of these brain diseases owing to it's massive spatial and temporal information.

In recent years, the number of neurobiological literatures using fMRI to study SZ disease has increased significantly. fMRI is usually applied to discover anomalous patterns present in activation maps [i.e., Regional Homogeneity (REHO), Amplitude of Low Frequency Fluctuations (ALFF), fractional Amplitude of Low Frequency Fluctuations (FALFF)] (Guo et al., 2014; Chyzhyk et al., 2015; Huang et al., 2018) of SZ. These activation maps are widely used as potential clinical biomarkers for the diagnosis of SZ. For example, Huang et al. (2018) used tree-guided group sparse learning method to perform feature selection on fALFF data in multi-frequency bands, and then used multi-kernel learning (MKL) method to achieve an accuracy of 91.10% on 34 subjects. Chyzhyk et al. (2015) combined these activation maps by using extreme learning machines and successfully distinguished SZs from healthy controls (HCs). However, these methods focus on the voxel-wise information in these maps rather than the connectivity between regions of interest (ROIs).

Functional connectivity has been reported to analyze the differences in the functional organization of brain networks between patients and HCs (Lynall et al., 2010; Pettersson-Yeo et al., 2011). Functional connectivity networks are usually derived from fMRI data (Van Den Heuvel and Pol, 2010; Craddock et al., 2013). Nodes of a functional brain network could be the voxels of fMRI data, ROIs defined by brain atlas or the discrete regions with similar size by randomly parcellating the brain (Fornito et al., 2013). Links of a functional brain network could be determined by the correlations estimated from time courses between pairs of nodes (Liu et al., 2017a). For example, Yu et al. (2015) created functional brain network using group ICA and Pearson correlation coefficient, and they found the new evidence about altered dynamic brain graphs in SZ. Abraham et al. (2017) investigated the most predictive biomarkers for Autism spectrum disorders (ASD) by building participant-specific connectomes from functionally-defined brain areas. For these methods, the connections between all pairs of nodes in a brain network are employed as features, but the topological measures of connectivity networks are not considered.

To quantitatively analyze functional brain networks, graph theoretical analysis is employed for investigating the topological organization of functional connectivity (Anderson and Cohen, 2013; Brier et al., 2014). The most commonly used graph measures include betweenness centrality, degree, local efficiency, participation coefficient, average clustering coefficient, average path length, global efficiency, and small-worldness (Liu et al., 2017a). These topological measures have been applied in the brain disease classifications (Cheng et al., 2015; Khazaee et al., 2015, 2017; Moghimi et al., 2018). For example, Moghimi et al. (2018) calculated a set of 25 graph measures including global and local measures for each subject and obtained a classification accuracy of 80% with a double-cross validation scheme. Cheng et al. (2015) achieved an accuracy of 79% by using betweenness centrality measure in SZ identification, and they found that changes in functional hubs were associated with SZ. Overall, these methods using graph measures for SZ identification have not achieved a good classification performance.

In this paper, we propose an improved method based on multi-view graph measures to identify SZs from HCs. Functional brain networks are constructed based on fMRI scans. Nodes of functional brain network are brain regions parcellated with the Brainnetome atlas (Fan et al., 2016), and edges of functional brain networks are determined by Pearson's correlation coefficients. Five local graph measures are calculated from functional brain networks by graph theoretical approach as features. The five local graph measures include betweenness centrality, nodal clustering coefficient, local efficiency, degree and participation coefficient. In order to consider the relationship of features within the same region, firstly we need to group graph measures according to brain regions defined by Brainnetome atlas. Then Sparse Group Lasso feature selection method is employed to select the most important regions as well as discriminative features within the selected regions. Finally, support vector machine (SVM) is trained for SZ identification. Our experiments are conducted on 145 samples with fMRI data, including 74 HCs and 71 SZs. Our proposed method achieves a mean classification accuracy of 93.10% using a leave-one-out cross-validation (LOOCV) scheme. The overall framework of our proposed method is shown in **Figure 1**, which consists of four main components include image preprocessing, feature representation, feature selection, and classification with SVM classifier. The code for this classification framework is available for download at https://github.com/xyzxzj/SZClassification.

2. MATERIALS AND METHODS

2.1. Subject Description and Image Preprocessing

The data involved in this study is collected by the Center for Biomedical Research Excellence (COBRE). COBRE[1] dataset consists of 148 subjects with functional and anatomical MRI data. 74 HCs and 71 SZs of the dataset are employed for our subsequent experiments owing to the class labels of the other three subjects are not given. During the scan, all participants are asked to remain relaxed and keep their eyes open. A brief summary of demographic information of subjects is listed in **Table 1**.

All of the fMRI images are preprocessed by using Data Processing & Analysis for Brain Imaging (DPABI) (Yan et al., 2016). The preprocessing procedure is as follows: the first 10 volumes of functional runs are removed owing to the fMRI signal instability. Then, the rest volumes are performed slice time correction, head-motion correction, and co-registration of T1-weighted MRI images and fMRI images. After that, the fMRI images are normalized to Montreal Neurological Institute (MNI)

[1]http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html

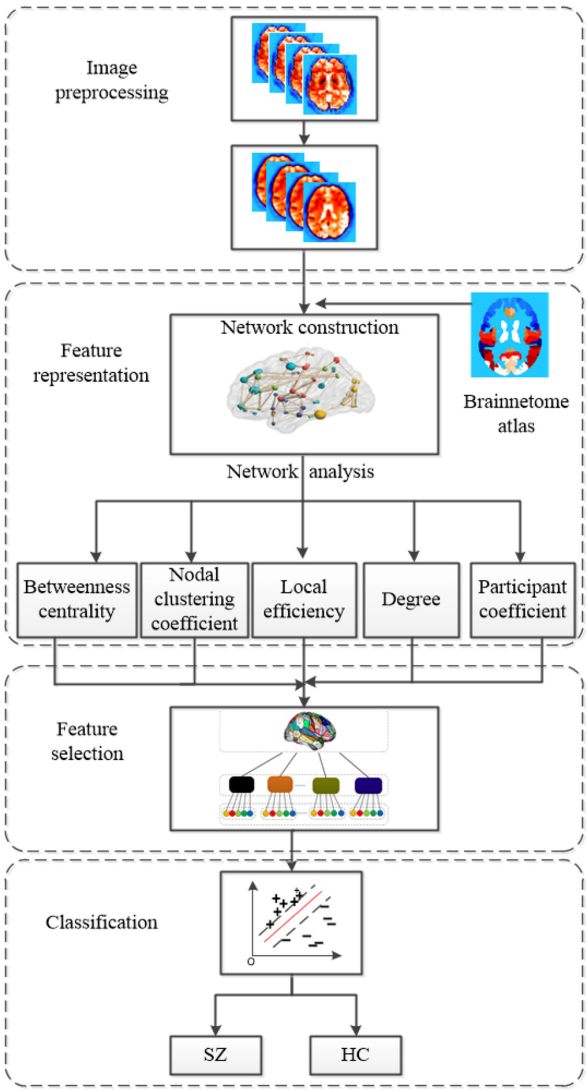


FIGURE 1 | The overall framework of our proposed method using multi-view graph measures of functional brain network for SZ/HC classification.

space and resampled to $3 \times 3 \times 3\,mm^3$ voxels. Smooth (4-mm FWHM) and band-pass filter (0.01–0.1Hz) are applied to the images which are transformed to MNI space.

In order to construct time series matrices for all subjects, first all brain images are parcellated into 246 regions by registering images to the Brainnetome atlas after fMRI data preprocessing. Then we extract the averaged time series for each of 246 brain regions for each subject. The time series of each brain region is derived from averaging fMRI signals of all voxels within the region. Finally, a time series matrix consists of 246 regional time series.

2.2. Feature Representation

2.2.1. Brain Network Construction

A network is composed of a collection of nodes and links. It can be described as a graph $G = (V, E)$, where V denotes the set of nodes and E is the set of links. There are four types of network topology, including weighted undirected, weighted directed, binary undirected and binary directed. In this study, the functional connectivity network is represented by an weighted undirected graph. The nodes in functional connectivity network usually are defined by brain regions, and links can represent temporal correlation in activity between pairs of nodes. Given a time series matrix, we can construct a functional connectivity network by calculating Pearson correlation coefficients (Pedersen et al., 2018) between signals of all pairs of regions. The generated functional brain network has $246 \times (246 - 1)/2 = 30,315$ weighted edges under the condition of 246 brain regions and the strength of each edge is the Pearson correlation coefficient between a pair of connected nodes.

TABLE 1 | Demographic information of 145 subjects from COBRE dataset.

Type	Number	Age	Gender (M/F)
SZ	71	38.1 ± 13.9	57/14
HC	74	35.8 ± 11.5	51/23

2.2.2. Brain Network Analysis

A great deal of functional connections in the network may lead to feature redundancy. A threshold t is employed in the dense network to keep a certain proportion of edges with the highest correlation. Graph-theoretic measures can quantify topological organization of network. Thus, we can extract some measures which can characterize the global or local functional connectivity from the threshold network. We compute 5 local graph measures using brain network analysis as feature representations, including degree, betweenness centrality, nodal clustering coefficient, local efficiency, and participation coefficient.

Degree is the most fundamental and important measure to characterize the centrality of nodes. In general, nodes with a higher degree are more important in networks. Betweenness centrality can also reflect the centrality of nodes. The betweenness centrality of a brain region can measure its ability on information transmission. Nodal clustering coefficient represents the possibility that any two neighbors of a given node are also neighbors of each other. It measures the ability of the node on functional segregation. Local efficiency measures the efficiency of a subnetwork formed by a given node and all its direct neighbors to transfer information. Local efficiency is related to the shortest path length of the node, the shorter the shortest path length, the greater the local efficiency of the node, the faster the information transmission in the subnetwork. Participation coefficient of a node measures its diversity of intermodular interconnections. The nodes with low participation coefficient but high degree in the module are regarded as provincial hubs, it indicates that the hubs are likely to have a great impact on the modular segregation. These five local measures play an important role in information exchange of functional networks. They can be calculated as follow:

$$K(i) = \sum_{j \in N} a_{ij} \tag{1}$$

$$B(i) = \frac{1}{(N-1)(N-1)} \sum_{m \neq j \neq i} \frac{n_{mj}(i)}{n_{mj}} \quad (2)$$

$$C(i) = \frac{2sw_i}{(K_i(K_i-1))} \quad (3)$$

$$E_{loc}(i) = \frac{1}{N_{G_i}(N_{G_i}-1)} \sum_{j \neq h \neq G_i} \frac{1}{l_{jh}} \quad (4)$$

$$PC(i) = 1 - \sum_{m \in M} \left(\frac{k_i(m)}{k_i}\right)^2 \quad (5)$$

where $K(i)$, $B(i)$, $C(i)$, $E_{loc}(i)$, and $PC(i)$ are the degree, betweenness centrality, clustering coefficient, local efficiency, and participation coefficient of node i, respectively. N is the number of nodes in a network, $a_{ij} = 1$ if node i and node j are connected, $a_{ij} = 0$ otherwise; $n_{mj}(i)$ is the number of shortest paths between m and j that pass through node i, and n_{mj} is the number of shortest paths between m and j; sw_i is the sum of the weights of all the connected edges between the neighbors of node i; G_i is the subnetwork that contains node i and its all direct neighbors, N_{G_i} is the number of nodes in the subnetwork G_i, l_{jh} is the length of shortest path between node j and node h in the subgraph; M denotes the set of modules, k_i is determined as the number of links between i and the nodes within module m.

In this study, we adopt the Brain Connectivity Toolbox (http://www.brain-connectivity-toolbox.net) (Rubinov and Sporns, 2010) to calculate these five local graph measures. For each local graph measure (gm), we compute 246 values corresponding to the 246 brain regions. Therefore, the dimension of the final feature vector for each subject is 1,230.

2.3. Feature Selection

The raw feature matrices have high dimension, multiple redundancy and multi-noise characteristics. Thus, applying a suitable feature selection algorithm to identify features related to SZ/HC identification and remove unnecessary information appears especially important. Least absolute shrinkage and selection operator (Lasso) (Chan et al., 2015) is widely used in various areas due to the very low data requirements. In addition, lasso can filter variables and reduce the complexity of the model. It aims to select the most important features from dense data matrix by using l_1 norm constraint. The optimization problem can be formulated as follow:

$$\min_{\alpha} ||y - X\alpha||^2 + \lambda_1 ||\alpha||_1 \quad (6)$$

where X denotes an $n \times p$ feature matrix, and n is the number of subjects, p represents the dimension of a feature vector. y is defined as a class label, α is a coefficient vector, and λ_1 is a regularization parameter.

Graph measures within the same region usually have a certain correlation. However, Lasso has not consider the relationship between graph measures derived in the same brain region. Hence we use the priori information of brain regions to group measures and then perform feature selection based on this feature grouping structure. Group Lasso (GLasso) (Yuan and Lin, 2006), a group variable selection method, is the extension of Lasso. It can select

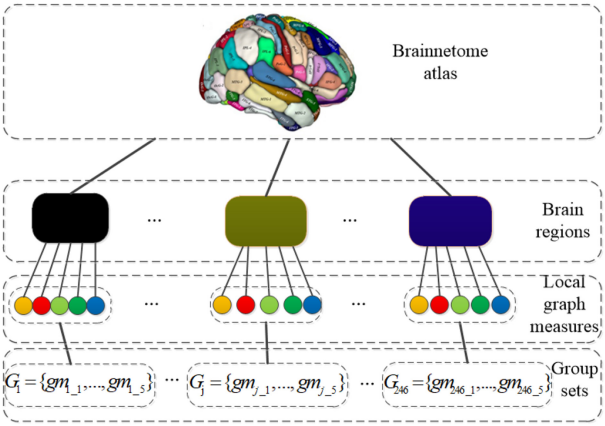


FIGURE 2 | The grouping structure: the nodes in the third layer represent local graph measures and the blocks in the second layer represent brain regions; $G_j = \{gm_{j\_1}, ..., gm_{j\_5}\}$ is a group set which consists of 5 local graph measures calculated for j_th region.

the most important groups by grouping all the variables and penalizing the l_2 norm of each group. The effect is that we can eliminate the entire set of coefficients into zero at the same time and then this set of features are excluded. The objective function of GLasso is as follow:

$$\min_{\alpha} ||y - X\alpha||^2 + \lambda_2 \sum_{j=1}^{M} w_j ||\alpha_{G_j}||_2 \quad (7)$$

where α_{G_j} denotes the set of coefficients of all features in the group G_j, w_j is a weight for group G_j.

Actually, there are also many redundant features in the important groups selected by GLasso. It is necessary to perform another feature selection to choose the most important features from these selected groups. Sparse Group Lasso (SGLasso) (Liu et al., 2009) is introduced to select the most significant groups as well as the discriminative features within the selected groups by adding l_1 and l_2 penalties. The objective function of the SGLasso can be written as:

$$\min_{\alpha} ||y - X\alpha||^2 + \lambda_1 ||\alpha||_1 + \lambda_2 \sum_{j=1}^{M} w_j ||\alpha_{G_j}||_2 \quad (8)$$

Before performing SGLasso, 1230-dimensional feature vector for each subject is grouped as $G = \{G_1, ..., G_j, ..., G_M\}$ according the brain regions defined by Brainnetome atlas. M is the number of groups. $G_j = \{gm_{j\_1}, gm_{j\_2}, gm_{j\_3}, gm_{j\_4}, gm_{j\_5}\}$ is a group consists of 5 local graph measures calculated for j_th region. The grouping structure is shown in **Figure 2**. In addition, z-score transformation is used to normalize the feature matrix before feature selection. It is worth noting that, after feature selection, those features are kept with corresponding regression coefficients greater than the mean value of absolute values of all elements in coefficient vectors.

2.4. Classification

SVM (Chang and Lin, 2011) is widely applied in various fields such as natural language processing, target detection, pattern classification due to its good performance as a supervised machine learning approach. The choice of SVM kernel functions is critical to their performance. In this study, we choose the linear kernel SVM (LSVM) to identify SZs from HCs. Linear kernel is mainly used in linear separability cases, and the dimension of the feature space and input space is the same. It performs good classification in most linear separable problems owing to the less parameters and fast calculation. The formulation of SVM model and linear kernel function are as follows:

$$\begin{aligned} &\max_{\lambda} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j K(x_i,x_j)+\sum_{i=1}^{N}\lambda_i \\ &s.t.\sum_{i=1}^{N}\lambda_i y_i = 0 \\ &0 \le \lambda_i \le C, i = 1,2,...,N \end{aligned} \tag{9}$$

$$K(x_i,x_j) =< x_i,x_j > \tag{10}$$

where λ is the Lagrange multiplier, N is the number of samples, x_i represents the feature vector of the i-th sample, and y_i is the label corresponding to x_i, $K(.,.)$ denotes the kernel function, C is determined as the soft margin parameter.

After feature selection, the optimal feature set $X = \{x_1,...,x_i,...,x_n\}$ is used as the input to SVM classifier, i = 1,..., n. Giving a test subject x, the trained SVM will predict its label based on a decision function P(x) as follows:

$$P(x) = sign(\sum_{i=1}^{N}\lambda_i y_i K(x_i,x)) \tag{11}$$

3. EXPERIMENTS AND RESULTS

3.1. Experiment Settings

In our study, the classification performance of our proposed method is estimated by adopting LOOCV scheme. LOOCV scheme is not affected by the random sample partitioning because n samples are only divided into n subsets in a unique way, each subset contains one sample. Each subset will be tested as a test data in turn while remaining subjects as the training data. In addition, we usually adopt the LIBSVM library (Chang and Lin, 2011) to solve SVM classification. We further calculate classification accuracy (ACC), sensitivity (SEN), specificity (SPE) to measure the performance of the method. These three metrics can be written as follows:

$$ACC = \frac{TP+TN}{TP+FP+FN+TN} \tag{12}$$

$$SPE = \frac{TN}{TN+FP} \tag{13}$$

$$SEN = \frac{TP}{TP+FN} \tag{14}$$

where true positive (TP), true negative (TN), false negative (FN), and false positive (FP) are defined as the number of correctly classified SZs, HCs and misidentified SZs, HCs, respectively.

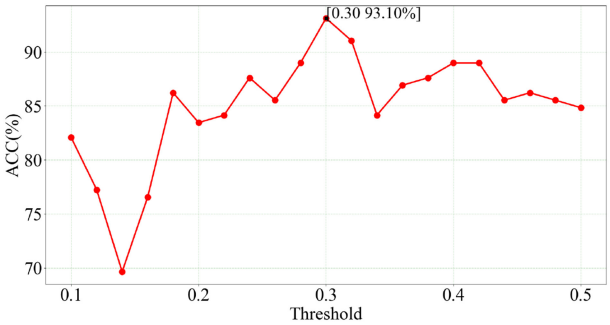


FIGURE 3 | Classification accuracies for SZ identification based on different network thresholds.

In addition, the area under receiver operating characteristic (ROC) curve (AUC) is also used to evaluate overall classification performance of our method.

At the stage of feature representation, we set $t = [0.1, 0.12, ..., 0.48, 0.5]$ to represent a collection of threshold values from 0.1 to 0.5 by the step of 0.02, and then calculate the 5 local graph measures at these 21 thresholds. The two regularization parameters for SGLasso are set as $\lambda_1 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$ and $\lambda_2 = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]$, which are optimized with the grid search algorithm.

3.2. Identification Performance for SZ

We use LSVM to perform SZ/HC classification on the optimal feature set obtained from feature selection of SGLasso at each of 21 thresholds. The classification results corresponding to 21 thresholds are showed in **Figure 3**.

According to **Figure 3**, we can see that the best accuracy (93.10%) is achieved at $t = 0.30$. Furthermore, the classification accuracies at these 21 thresholds are all higher than 70%. In addition, the number of selected features is 55 and SEN, SPE, AUC values are 92.96%, 93.24%, 0.950, respectively. The experimental results indicate that the feature combination of five local measures extracted at $t = 0.30$ has a relatively strong correlation with SZ identification.

4. DISCUSSION

4.1. Comparison With Different Feature Selection Methods

In order to demonstrate the SGLasso method is more effective than the common feature selection methods based on these five local measures for SZ classification, we compare four feature selection methods. The first one is t-test which is the one of the most basic feature selection method and the most critical part of this method is selecting features based on the p-value (i.e., 0.05). The rest methods are Lasso, GLasso and Elastic Net (Enet). These three methods are based on linear sparse models. GLasso and Enet are the extension of Lasso. GLasso is used to solve l_1/l_q-norm regularized problem. Enet is used for the situations

TABLE 2 | Classification with different feature selection methods.

Methods	Number of selected features	ACC (%)	SEN (%)	SPE (%)
t-test	153	78.62	80.28	77.03
Lasso	123	83.45	88.73	78.38
GLasso	225	86.21	85.92	86.49
ENet	64	85.52	84.51	86.19
SGLasso	55	93.10	92.96	93.24

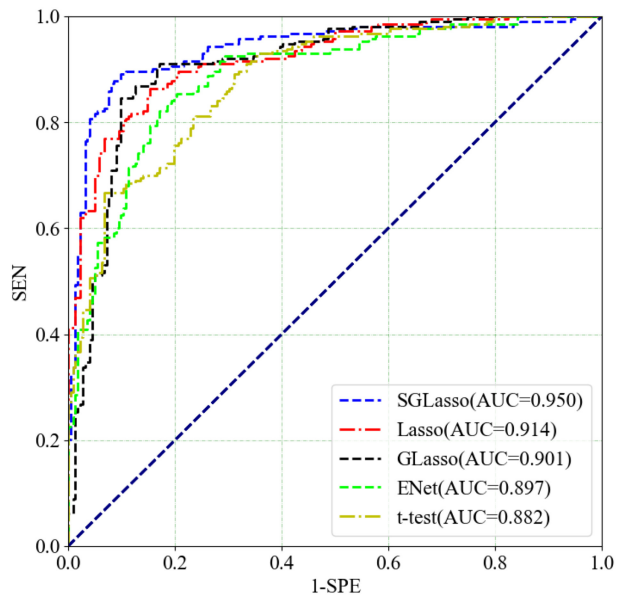


FIGURE 4 | ROC curves for SZ/HC classification for different feature selection methods.

where features are related to each other and always produce valid solution.

These four feature selection methods perform the same experimental procedure as SGLasso for the sake of fairness. It's worth noting that the five local graph measures are extracted at the threshold of 0.30. **Table 2** shows the experimental results of the above mentioned four methods and SGLasso feature selection method. As we can see that SGLasso method selects the least features (55) but achieves the best ACC (93.10%), SEN (92.96%), SPE (93.24%). The ROC curves for SZ/HC classification for different feature selection methods as shown in **Figure 4**. We notice that SGLasso achieves the highest AUC (0.950) than other four feature selection methods. Experimental result shows that considering within- and between- group sparsity is likely helpful for selecting significant features that are effective for SZ identification.

4.2. Comparison With Different Classifiers

In order to prove that LSVM is optimal to conduct classification in this context, a series of comparative experiments using several SVMs with different kernels including Radial Basis Function kernel (RBF), Ploynomial kernel (Poly), Sigmoid kernel (Sigm) under the same condition as the LSVM have been performed. These SVMs are denoted as RBF-SVM, Poly-SVM, Sigm-SVM, respectively. The experimental results of SVMs with different kernels are shown in **Table 3**. It is worth mentioning that bold text indicates that the best result is obtained on a certain evaluation metric.

TABLE 3 | Comparison with other SVMs using different kernels.

Methods	ACC (%)	SEN (%)	SPE (%)	AUC
RBF-SVM	80.00	76.06	83.78	0.8601
Poly-SVM	82.07	77.46	86.49	0.8506
Sigm-SVM	87.59	83.10	91.89	0.9393
LSVM	**93.10**	**92.96**	**93.24**	**0.950**

Bold text indicates that the best result is obtained on a certain evaluation metric.

TABLE 4 | Comparison with other commonly used classifiers.

Methods	ACC (%)	SEN (%)	SPE (%)	AUC
KNN	82.07	74.65	89.19	0.7912
RForest	77.93	74.65	81.08	0.8378
NBayes	84.83	83.10	86.49	0.9069
LDA	90.34	87.32	93.24	0.9418
LSVM	**93.10**	**92.96**	**93.24**	**0.950**

Bold text indicates that the best result is obtained on a certain evaluation metric.

In addition, we also compare four commonly used classifiers, such as k-nearest neighbors (KNN), Random Forest (RForest), NaiveBayes (NBayes), and Linear Discriminant Analysis (LDA). These classifiers are all implemented on the platform of Matlab2016a. We evaluate the performance of the above four classifiers under the same conditions as LSVM. The experimental results of these five classifiers are shown in **Table 4**. As can be seen from **Tables 3**, **4**, LSVM can achieve the best classification performance than other classifiers.

4.3. Regularization Parameter Selection

The regularization parameters of SGLasso have a great influence on the results of feature selection. Using different regularization parameters, the selected features are also different. It affects not only the feature dimension, but also the final classification performance. Therefore, selecting the appropriate regularization parameters can improve the efficiency of SGLasso method and obtain more effective features associated with the labels.

The two regularization parameters of SGLasso are λ_1 and λ_2. λ_1 is used to control the model sparseness, and λ_2 can control the sparse constraint of each feature group. We use the grid search algorithm to find the optimal combination of regularization parameters. **Figure 5** shows the classification results using different combination of λ_1, λ_2. According to **Figure 5**, when the parameter combination is (λ_1=9, λ_2=0.1), the features obtained from SGLasso feature selection method are the most effective for SZ/HC classification.

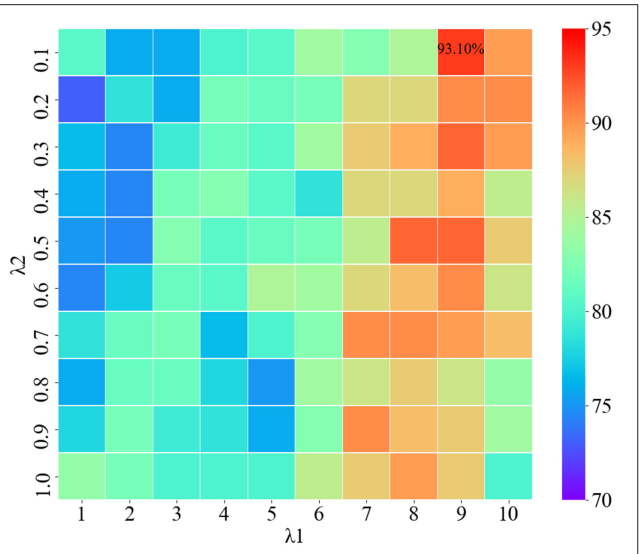


FIGURE 5 | Classification results using different combination of λ_1,λ_2.

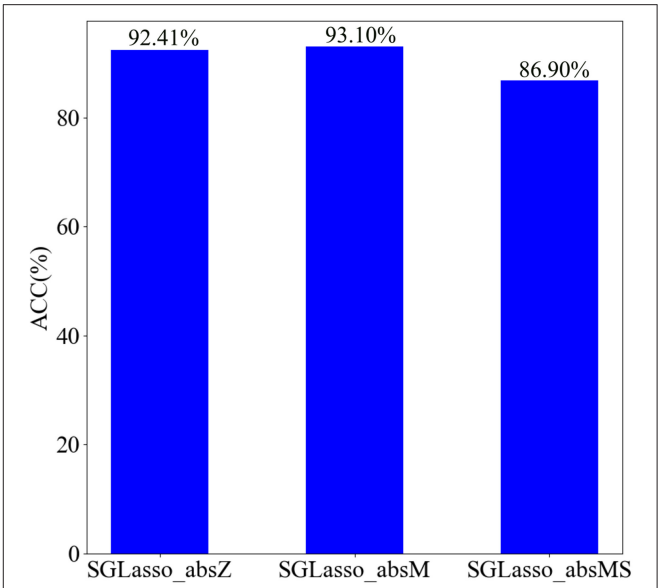


FIGURE 6 | Classification results using three different regression coefficient selection strategies.

4.4. Regression Coefficient Selection

In general, the non-zero elements in the coefficient vector α generated from the SGLasso feature selection algorithm indicate that the corresponding features are selected. In order to retain the least but most informative features according to α, we test the impact of the three coefficient selection strategies on classification performance. We named these three strategies as SGLasso_absZ, SGLasso_absM, and SGLasso_absMS. The description of these three strategies is as follows:

- SGLasso_absZ is a common strategy to retain non-zero coefficients of α.
- SGLasso_absM strategy is to retain those coefficients which are greater than the mean value of absolute values of all elements in α.
- SGLasso_absMS strategy is more strict for selecting coefficients, since it retains the coefficients which are larger than the mean value of absolute values of all non-zero coefficients in α.

We apply the above mentioned three strategies to feature selection, and then select the corresponding features according to the retained coefficients in α. SVM performs SZ identification using these selected features. The classification results using three different regression coefficient selection strategies are shown in **Figure 6**. According to **Figure 6**, the classification accuracy is the best when using SGLasso_absM strategy. Experimental result indicates that using SGLasso_absM strategy in feature selection can select the most effective features for SZ/HC classification. Therefore, we finally choose the SGLasso_absM strategy to select the regression coefficients.

4.5. Classification Comparison Using Different Feature Combinations

In order to explore the impact of different feature combinations on SZ/HC identification, we combine these five local measures extracted at the threshold of 0.30 in $C_5^2 + C_5^3 + C_5^4 + C_5^5 = 26$ ways. Furthermore, we don't consider individual graph measure because we only investigate multiple measures in this study. We evaluate these 26 feature sets under the same experimental settings. The classification results are shown in **Figure 7**.

As can be seen from **Figure 7**, the combination of 5 local graph measures achieves the best classification performance compared to other feature sets. At the same time, we also find that the classification accuracies obtained by using feature sets including two measures are lower than the classification accuracies obtained by using feature sets including three measures, four measures and five measures. It indicates that using fewer measures may not be enough to characterize brain network alternation, and we find that the combination of five local measures can provide more useful information for SZ identification.

4.6. Comparison With Existing Classification Methods

To verify the effectiveness of our proposed classification method, we compare some recently proposed methods for SZ classification using fMRI in the literature. Huang et al. (2018) proposed a tree-guided group sparse learning method to select the most important information from FALFF data in four frequency bands and get a classification accuracy of 91.1% by using multi-kernel SVM. Cheng et al. (2015) calculated only betweenness centrality measure to characterize the network. They used the rank of betweenness centrality of all nodes as feature representations and used SVM to classify SZs from HCs.

The two above mentioned methods are performed on the COBRE dataset. The classification results and ROC curves for SZ/HC classification of the two methods and our proposed method are shown in **Table 5** and in **Figure 8**, respectively.

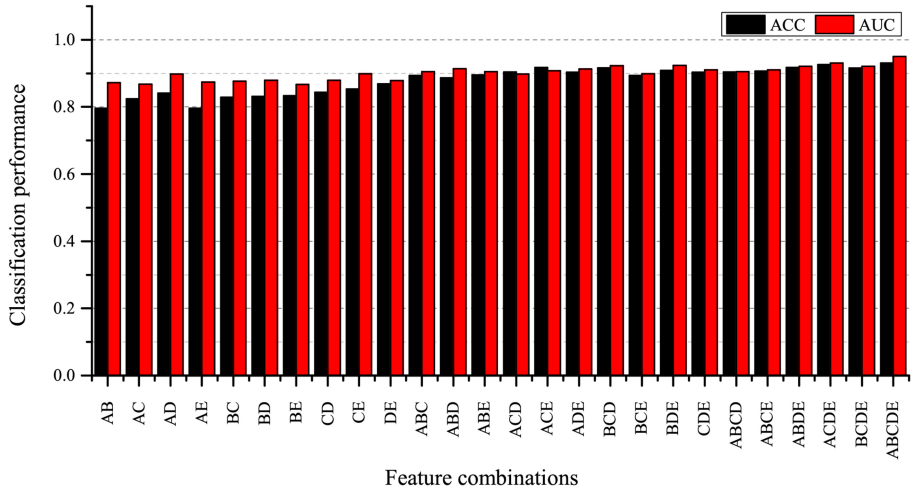


FIGURE 7 | Classification result for different feature combinations. A: betweenness centrality, B: nodal clustering coefficient, C: local efficiency, D: degree, E: participation coefficient.

TABLE 5 | Comparison with some existing methods for SZ/HC classification.

Methods	ACC (%)	SEN (%)	SPE (%)	AUC
Huang et al. (2018)	77.24	77.46	76.58	0.815
Cheng et al. (2015)	74.48	73.53	69.12	0.792
Proposed	**93.10**	**92.96**	**93.24**	**0.950**

Bold text indicates that the best result is obtained on a certain evaluation metric.

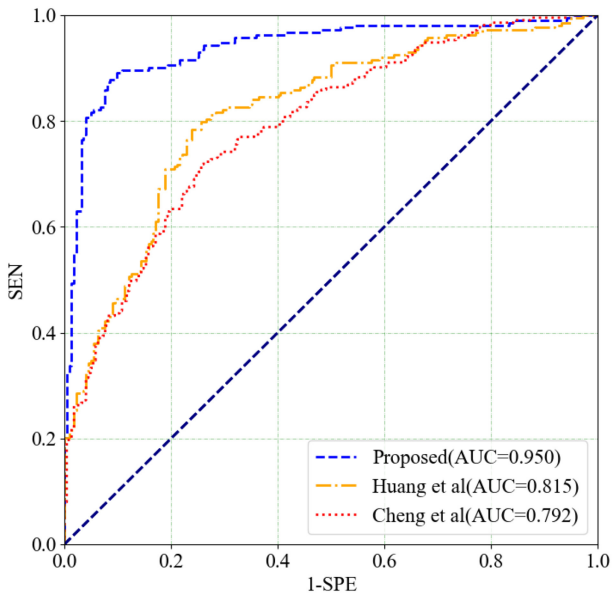


FIGURE 8 | ROC curves for SZ/HC classification for different classification methods.

According to **Table 5** and **Figure 8**, Our proposed method gets the best ACC (93.10%), SEN (92.96%), SPE (93.24%), and AUC (0.950) values. The experimental result illustrates that our proposed method has made a significant improvement in classification performance on the COBRE dataset.

4.7. Analysis of Discriminative Graph Measures and Corresponding Regions

The graph measures selected in the feature selection stage are considered to be related to their corresponding brain regions. Our method can select the most discriminative brain regions as the biomarkers to guide the disease-induced interpretation. There is a total of 145 experiments in the LOOCV scheme due to 145 subjects. And the number of feature occurrence in 145 experiments is introduced to indicate the contribution of the feature to classification. We assume that if the occurrence number of a local graph measure extracted from a certain brain region is greater than 140 in a total of 145 experiments, the brain region is considered to have the most discriminative power to distinguish between SZs and HCs. Based on this hypothesis, 21 salient brain regions have been found. These significant brain regions are shown in **Table 6**. Five brain regions include left superior frontal gyrus (SFG_L_7_2), right inferior temporal gyrus (ITG_R_7_7), right inferior parietal lobule (IPL_R_6_4), right postcentral gyrus (PoG_R_4_1), and right thalamus (Tha_R_8_7) are related to more than one local graph measure.

These findings on discriminative brain regions are in agreement with the following studies: superior frontal gyrus,cingulate gyrus, postcentral gyrus (Szeszko et al., 1999; Gur et al., 2000; Arbabshirani et al., 2013; Chyzhyk et al., 2015), parahippocampal gyrus (Shenton et al., 1992; Chyzhyk et al., 2015), middle temporal gyrus, fusiform gyrus and thalamus (Chyzhyk et al., 2015; Li et al., 2019), inferior parietal lobule, inferior temporal gyrus (Peng et al., 1994; Goldstein et al., 1999; Li et al., 2019). However, we cannot report agreement with these

TABLE 6 | The most discriminative graph measures and corresponding Brainnetome regions.

Graph measures	Hemisphere	Brainnetome regions	Occurrence number
Nodal clustering coefficient	SFG_L_7_2	Superior Frontal Gyrus	144
Degree	SFG_L_7_2	Superior Frontal Gyrus	145
Nodal clustering coefficient	SFG_R_7_2	Superior Frontal Gyrus	140
Participation coefficient	SFG_R_7_7	Superior Frontal Gyrus	144
Betweenness centrality	IFG_L_6_3	Inferior Frontal Gyrus	143
Betweenness centrality	OrG_L_6_2	Orbital Gyrus	143
Betweenness centrality	OrG_R_6_6	Orbital Gyrus	145
Betweenness centrality	PrG_L_6_3	Precentral Gyrus	142
Degree	MTG_L_4_4	Middle Temporal Gyrus	145
Betweenness centrality	MTG_L_4_1	Middle Temporal Gyrus	141
Participation coefficient	ITG_R_7_7	Inferior Temporal Gyrus	145
Betweenness centrality	ITG_R_7_7	Inferior Temporal Gyrus	145
Betweenness centrality	FuG_R_3_3	Fusiform Gyrus	145
Betweenness centrality	PhG_L_6_3	Parahippocampal Gyrus	144
Degree	PhG_R_6_5	Parahippocampal Gyrus	145
Local efficiency	IPL_R_6_4	Inferior Parietal Lobule	145
Participation coefficient	IPL_R_6_4	Inferior Parietal Lobule	145
Degree	IPL_R_6_2	Inferior Parietal Lobule	145
Degree	PCun_L_4_3	Precuneus	145
Nodal clustering coefficient	PoG_R_4_1	Postcentral Gyrus	145
Betweenness centrality	PoG_R_4_1	Postcentral Gyrus	145
Local efficiency	PoG_R_4_1	Postcentral Gyrus	143
Degree	PoG_R_4_1	Postcentral Gyrus	145
Participation coefficient	CG_L_7_4	Cingulate Gyrus	145
Betweenness centrality	CG_R_7_3	Cingulate Gyrus	145
Participation coefficient	LOcC_L_4_3	lateral Occipital Cortex	145
Degree	BG_R_6_1	Basal Ganglia	145
Betweenness centrality	BG_R_6_4	Basal Ganglia	145
Participation coefficient	Tha_L_8_8	Thalamus	145
Degree	Tha_L_8_5	Thalamus	145
Degree	Tha_R_8_8	Thalamus	145
Nodal clustering coefficient	Tha_R_8_7	Thalamus	140
Local efficiency	Tha_R_8_7	Thalamus	141

regions:inferior frontal gyrus, orbital gyrus, precentral gyrus, precuneus, lateral occipital cortex and basal ganglia.

5. CONCLUSION

In this paper, we propose a method to classify SZs from HCs using multi-view graph measures of functional brain networks. We get five local network measures using graph theoretical approach from multiple views. These measures paly an important role in the information exchange of brain networks. Our proposed method achieves a good classification performance on the COBRE dataset. Experimental results demonstrate that this approach is efficient for the clinical diagnosis of SZ. Furthermore, multiple measures have the potential to be used as clinical biomarkers to differentiate SZs from HCs.

DATA AVAILABILITY STATEMENT

The imaging data and phenotypic information was collected and shared by the Mind Research Network and the University of New Mexico funded by a National Institute of Health Center of Biomedical Research Excellence (COBRE) grant 1P20RR021938-01A2. The dataset for this study can be found in this website: http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

AUTHOR CONTRIBUTIONS

JW and JL conceived the project. YX, GT, F-XW, and JL designed the experiments. YX and GT performed the experiments. YX and JL wrote the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61802442, 61877059, the Natural Science Foundation of Hunan Province under Grant No. 2019JJ50775, the 111 Project (No. B18059), the Hunan Provincial Science and Technology Program (2018WK4001). The funding bodies did not influence the design of the study and collection, analysis, and interpretation of data or writing the manuscript.

ACKNOWLEDGMENTS

A partial version of this work was accepted at the Fourth CCF Bioinformatics Conference (CBC 2019) in Guangzhou, China (Aug 23-25, 2019). We would like to thank the reviewers for their detailed suggestions which greatly improved the quality and readability of this work.

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *Neuroimage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

Anderson, A., and Cohen, M. S. (2013). Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. *Front. Hum. Neurosci.* 7:520. doi: 10.3389/fnhum.2013.00520

Arbabshirani, M. R., Kiehl, K., Pearlson, G., and Calhoun, V. D. (2013). Classification of schizophrenia patients based on resting-state functional network connectivity. *Front. Neurosci.* 7:133. doi: 10.3389/fnins.2013.00133

Brier, M. R., Thomas, J. B., Fagan, A. M., Hassenstab, J., Holtzman, D. M., Benzinger, T. L., et al. (2014). Functional connectivity and graph theory in preclinical Alzheimer's disease. *Neurobiol. Aging* 35, 757–768. doi: 10.1016/j.neurobiolaging.2013.10.081

Castro, E., Martínez-Ramón, M., Pearlson, G., Sui, J., and Calhoun, V. D. (2011). Characterization of groups using composite kernels and multi-source

fmri analysis data: application to schizophrenia. *Neuroimage* 58, 526–536. doi: 10.1016/j.neuroimage.2011.06.044

Chan, M., Krebs, M., Cox, D., Guest, P., Yolken, R. H., Rahmoune, H., et al. (2015). Development of a blood-based molecular biomarker test for identification of schizophrenia before disease onset. *Transl. Psychiatry* 5:e601. doi: 10.1038/tp.2015.91

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technology* 2:27. doi: 10.1145/1961189.1961199

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.-P. P., et al. (2019). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2019.2936476. [Epub ahead of print].

Cheng, H., Newman, S., Goñi, J., Kent, J. S., Howell, J., Bolbecker, A., et al. (2015). Nodal centrality of functional network in the differentiation of schizophrenia. *Schizophrenia Res.* 168, 345–352. doi: 10.1016/j.schres.2015.08.011

Chyzhyk, D., Savio, A., and Graña, M. (2015). Computer aided diagnosis of schizophrenia on resting state fMRI data by ensembles of ELM. *Neural Netw.* 68, 23–33. doi: 10.1016/j.neunet.2015.04.002

Craddock, R. C., Jbabdi, S., Yan, C.-G., Vogelstein, J. T., Castellanos, F. X., Di Martino, A., et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* 10:524. doi: 10.1038/nmeth.2482

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., et al. (2016). The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb. Cortex* 26, 3508–3526. doi: 10.1093/cercor/bhw157

Fornito, A., Zalesky, A., and Breakspear, M. (2013). Graph analysis of the human connectome: promise, progress, and pitfalls. *Neuroimage* 80, 426–444. doi: 10.1016/j.neuroimage.2013.04.087

Goldstein, J. M., Goodman, J. M., Seidman, L. J., Kennedy, D. N., Makris, N., Lee, H., et al. (1999). Cortical abnormalities in schizophrenia identified by structural magnetic resonance imaging. *Arch. Gen. Psychiatry* 56, 537–547. doi: 10.1001/archpsyc.56.6.537

Guo, W., Su, Q., Yao, D., Jiang, J., Zhang, J., Zhang, Z., et al. (2014). Decreased regional activity of default-mode network in unaffected siblings of schizophrenia patients at rest. *Eur. Neuropsychopharmacol.* 24, 545–552. doi: 10.1016/j.euroneuro.2014.01.004

Gur, R. E., Cowell, P. E., Latshaw, A., Turetsky, B. I., Grossman, R. I., Arnold, S. E., et al. (2000). Reduced dorsal and orbital prefrontal gray matter volumes in schizophrenia. *Arch. Gen. Psychiatry* 57, 761–768. doi: 10.1001/archpsyc.57.8.761

He, H., Sui, J., Yu, Q., Turner, J. A., Ho, B.-C., Sponheim, S. R., et al. (2012). Altered small-world brain networks in schizophrenia patients during working memory performance. *PLoS ONE* 7:e38195. doi: 10.1371/journal.pone.0038195

Heinrichs, R. W., and Zakzanis, K. K. (1998). Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology* 12:426. doi: 10.1037/0894-4105.12.3.426

Huang, J., Zhu, Q., Hao, X., Shi, X., Gao, S., Xu, X., and Zhang, D. (2018). Identifying resting-state multifrequency biomarkers via tree-guided group sparse learning for schizophrenia classification. *IEEE J. Biomed. Health Informatics* 23, 342–350. doi: 10.1109/JBHI.2018.2796588

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2015). Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin. Neurophysiol.* 126, 2132–2141. doi: 10.1016/j.clinph.2015.02.060

Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., for the Alzheimer's Disease Neuroimaging Initiative (2017). Classification of patients with mci and ad from healthy controls using directed graph measures of resting-state fMRI. *Behav. Brain Res.* 322, 339–350. doi: 10.1016/j.bbr.2016.06.043

Li, J., Sun, Y., Huang, Y., Bezerianos, A., and Yu, R. (2019). Machine learning technique reveals intrinsic characteristics of schizophrenia: an alternative method. *Brain Imaging Behav.* 13, 1386–1396. doi: 10.1007/s11682-018-9947-4

Liu, J., Ji, S., and Ye, J. (2009). *Slep: Sparse Learning With Efficient Projections*. Arizona State University, 7.

Liu, J., Li, M., Lan, W., Wu, F.-X., Pan, Y., and Wang, J. (2016). Classification of Alzheimer's disease using whole brain hierarchical network. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 624–632. doi: 10.1109/TCBB.2016.2635144

Liu, J., Li, M., Pan, Y., Lan, W., Zheng, R., Wu, F.-X., and Wang, J. (2017a). Complex brain network analysis and its applications to brain disorders: a survey. *Complexity* 2017:8362741. doi: 10.1155/2017/8362741

Liu, J., Li, M., Pan, Y., Wu, F.-X., Chen, X., and Wang, J. (2017b). Classification of schizophrenia based on individual hierarchical brain networks constructed from structural MRI images. *IEEE Trans. Nanobiosci.* 16, 600–608. doi: 10.1109/TNB.2017.2751074

Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., et al. (2018a). Applications of deep learning to MRI images: a survey. *Big Data Mining Analyt.* 1, 1–18. doi: 10.26599/BDMA.2018.9020001

Liu, J., Wang, J., Hu, B., Wu, F.-X., and Pan, Y. (2017c). Alzheimer's disease classification based on individual hierarchical networks constructed with 3-D texture features. *IEEE Trans. Nanobiosci.* 16, 428–437. doi: 10.1109/TNB.2017.2707139

Liu, J., Wang, J., Tang, Z., Hu, B., Wu, F.-X., and Pan, Y. (2017d). Improving Alzheimer's disease classification by combining multiple measures. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 1649–1659. doi: 10.1109/TCBB.2017.2731849

Liu, J., Wang, X., Zhang, X., Pan, Y., Wang, X., and Wang, J. (2018b). MMM: classification of schizophrenia using multi-modality multi-atlas feature representation and multi-kernel learning. *Multimedia Tools Appl.* 77, 29651–29667. doi: 10.1007/s11042-017-5470-7

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Marín, O. (2012). Interneuron dysfunction in psychiatric disorders. *Nat. Rev. Neurosci.* 13:107. doi: 10.1038/nrn3155

Moghimi, P., Lim, K. O., and Netoff, T. I. (2018). Data driven classification of fMRI network measures: application to schizophrenia. *Front. Neuroinform.* 12:71. doi: 10.3389/fninf.2018.00071

Nieuwenhuis, M., van Haren, N. E., Pol, H. E. H., Cahn, W., Kahn, R. S., and Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *Neuroimage* 61, 606–612. doi: 10.1016/j.neuroimage.2012.03.079

Pedersen, M., Omidvarnia, A., Zalesky, A., and Jackson, G. D. (2018). On the relationship between instantaneous phase synchrony and correlation-based sliding windows for time-resolved fMRI connectivity analysis. *Neuroimage* 181, 85–94. doi: 10.1016/j.neuroimage.2018.06.020

Peng, L. W., Lee, S., Federman, E. B., Chase, G. A., Barta, P. E., and Pearlson, G. D. (1994). Decreased regional cortical gray matter volume in schizophrenia. *Am. J. Psychiatry* 151:843. doi: 10.1176/ajp.151.6.842

Pettersson-Yeo, W., Allen, P., Benetti, S., McGuire, P., and Mechelli, A. (2011). Dysconnectivity in schizophrenia: where are we now? *Neurosci. Biobehav. Rev.* 35, 1110–1124. doi: 10.1016/j.neubiorev.2010.11.004

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., et al. (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* 45, 1150–1159. doi: 10.1038/ng.2742

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Shenton, M. E., Kikinis, R., Jolesz, F. A., Pollak, S. D., LeMay, M., Wible, C. G., et al. (1992). Abnormalities of the left temporal lobe and thought disorder in schizophrenia: a quantitative magnetic resonance imaging study. *N. Engl. J. Med.* 327, 604–612. doi: 10.1056/NEJM199208273270905

Szeszko, P. R., Bilder, R. M., Lencz, T., Pollack, S., Alvir, J. M. J., Ashtari, M., et al. (1999). Investigation of frontal lobe subregions in first-episode schizophrenia. *Psychiatry Res. Neuroimaging* 90, 1–15. doi: 10.1016/S0925-4927(99)00002-5

Van Den Heuvel, M. P., and Pol, H. E. H. (2010). Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* 20, 519–534. doi: 10.1016/j.euroneuro.2010.03.008

Yan, C.-G., Wang, X.-D., Zuo, X.-N., and Zang, Y.-F. (2016). Dpabi: data processing & analysis for (resting-state) brain

imaging. *Neuroinformatics* 14, 339–351. doi: 10.1007/s12021-016-9299-4

Yang, Y., and Wang, H. (2018). Multi-view clustering: a survey. *Big Data Mining and Analyt.* 1, 83–107. doi: 10.26599/BDMA.2018.9020003

Yu, Q., Erhardt, E. B., Sui, J., Du, Y., He, H., Hjelm, D., et al. (2015). Assessing dynamic brain graphs of time-varying connectivity in fMRI data: application to healthy controls and patients with schizophrenia. *Neuroimage* 107, 345–355. doi: 10.1016/j.neuroimage.2014.12.020

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 17 January 2020
doi: 10.3389/fgene.2019.01349

DSNetwork: An Integrative Approach to Visualize Predictions of Variants' Deleteriousness

Audrey Lemaçon, Marie-Pier Scott-Boyer, Régis Ongaro-Carcy, Penny Soucy, Jacques Simard and Arnaud Droit *

Genomics Center, Centre Hospitalier Universitaire de Quebec—Université Laval Research Center, Quebec, QC, Canada

One of the most challenging tasks of the post-genome-wide association studies (GWAS) research era is the identification of functional variants among those associated with a trait for an observed GWAS signal. Several methods have been developed to evaluate the potential functional implications of genetic variants. Each of these tools has its own scoring system, which forces users to become acquainted with each approach to interpret their results. From an awareness of the amount of work needed to analyze and integrate results for a single locus, we proposed a flexible and versatile approach designed to help the prioritization of variants by aggregating the predictions of their potential functional implications. This approach has been made available through a graphical user interface called DSNetwork, which acts as a single point of entry to almost 60 reference predictors for both coding and non-coding variants and displays predictions in an easy-to-interpret visualization. We confirmed the usefulness of our methodology by successfully identifying functional variants in four breast cancer and nine schizophrenia susceptibility loci.

Keywords: fine-mapping analysis, variant prioritization, decision support, deleteriousness prediction, network visualization

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Richa Gupta,
Independent Researcher, United States
Tianfan Fu,
Georgia Institute of Technology, United States

***Correspondence:**
Arnaud Droit
arnaud.droit@crchudequebec.ulaval.ca

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *25 September 2019*
Accepted: *10 December 2019*
Published: *17 January 2020*

Citation:
Lemaçon A, Scott-Boyer M-P, Ongaro-Carcy R, Soucy P, Simard J and Droit A (2020) DSNetwork: An Integrative Approach to Visualize Predictions of Variants' Deleteriousness. Front. Genet. 10:1349. doi: 10.3389/fgene.2019.01349

INTRODUCTION

Since 2006, thousands of susceptibility loci have been identified through Genome-Wide Association Studies (GWAS) for numerous traits and complex diseases, including breast cancer (MacArthur et al., 2017). GWAS build on the concept of linkage disequilibrium (LD) to identify statistical associations between genetic variants and diseases (Visscher et al., 2017). While this approach is powerful for locus discovery, it cannot distinguish between truly causal variants and non-functional highly correlated neighboring variants. Thus, for the vast majority of these loci, the causal variant(s) and their functional mechanisms have not yet been elucidated.

Statistical fine-mapping analyses combined with the functional annotation of genetic variants can help pinpoint the genetic variant (or variants) responsible for complex traits, or at least narrow down the number of variants underlying the observed association for further functional studies. In this regard, tremendous efforts have been put forth to assist the functional assessment of variants at

risk loci and numerous scoring methods and tools have been developed to predict the deleteriousness of variants based on a number of characteristics such as sequence conservation, characteristics of amino acid substitution, and location of the variant within protein domains or three-dimensional protein structure.

In recent years, efforts have been made towards the aggregation of many different functional annotations resulting from these scoring methods in a single integrative value called metascore (Ionita-Laza et al., 2016; Feng, 2017), an approach that seems to yield better performances than any predictor individually (Dong et al., 2015). Although these methods demonstrate themselves to be useful, they have some limitations, notably not being directly comparable to one another due to integration of different sets of annotations or different weighting of these annotations, and sometimes having contradictory results.

In order to allow a quick survey of a wide range of predictors for a given list of variants and assist in the interpretation of the resulting prediction scores, we propose a flexible and integrative method capable of gathering information from multiple sources in an easy-to-interpret representation rather than a static new metascore. For this purpose, we created a single point of entry fetching predictors for coding and non-coding variants and presenting them as a network, where the nodes illustrate the scores of each predictor for a given variant and the edges the LD between variants. The network is built with the aim of rendering the predictor results easier to peruse during analyses involving multiple variants, and therefore, assist in the variant prioritization process in the context of fine-mapping analyses.

This approach has been made available through a graphical user interface (GUI) stand-alone application called DSNetwork. The tool is freely available *via* bitbucket repository and is also accessible through our portal for demonstration purpose at: http://romix.genome.ulaval.ca/dsnetwork/.

MATERIALS AND METHODS

Annotations Retrieval

Variant annotations and scoring data are fetched on-the-fly from MyVariant.info high-performance web services (Xin et al., 2016) using their third-party R package. SNPnexus (Dayem Ullah et al., 2018) scorings are fetched upon request through a Python script kindly provided by the SNPnexus team. Due to their novelty and relevance for our purpose, three complementary whole genome resources are included: LINSIGHT (Huang et al., 2017), BayesDel (Feng, 2017), and predictions and sequence constraint data (di Iulio et al., 2018), which can be used as a proxy to score functionality and the consequences of mutations. BayesDel, LINSIGHT, and Context-Dependent Tolerance scores were extracted from a local copy. A description of the integrated predictors is available in the **Supplementary Material**.

LD data are computed from 1000 Genomes Phase 3 (1000 Genomes Project Consortium et al., 2015).

Visual Integration

Prediction result for variants of interest are displayed as a network, whose components, namely, the edges and nodes, are used to convey different types of information in an easy-to-comprehend way.

The following paragraphs describe DSNetwork's approach through the hypothetical analysis of a loci containing five variants rs4233486, rs35054111, rs11808410, rs11804913, and rs7554973 using the deleteriousness scores of five distinct fictive predictors A, B, C, D, and E. **Table 1** summarizes the scores generated by these five predictors, reflecting their predictions regarding the functional impacts of the candidate variants.

DSNetwork integrates the characteristics of the different predictors and creates a reference frame containing the lower and upper boundaries as well as the direction [ascending (ASC) or descending (DESC)] of their prediction scores (**Figure 1A**). The direction is used to rank variants from the most deleterious to the least deleterious on the basis of their respective scores. The boundaries are used to establish the absolute deleteriousness level of each variant. Once the different reference frames are integrated, they can be used to prioritize the variants according to three types of representations: the intra-predictor relative ranks, the intra-predictor absolute scores, and the global ranks.

Intra-Predictor Ranks

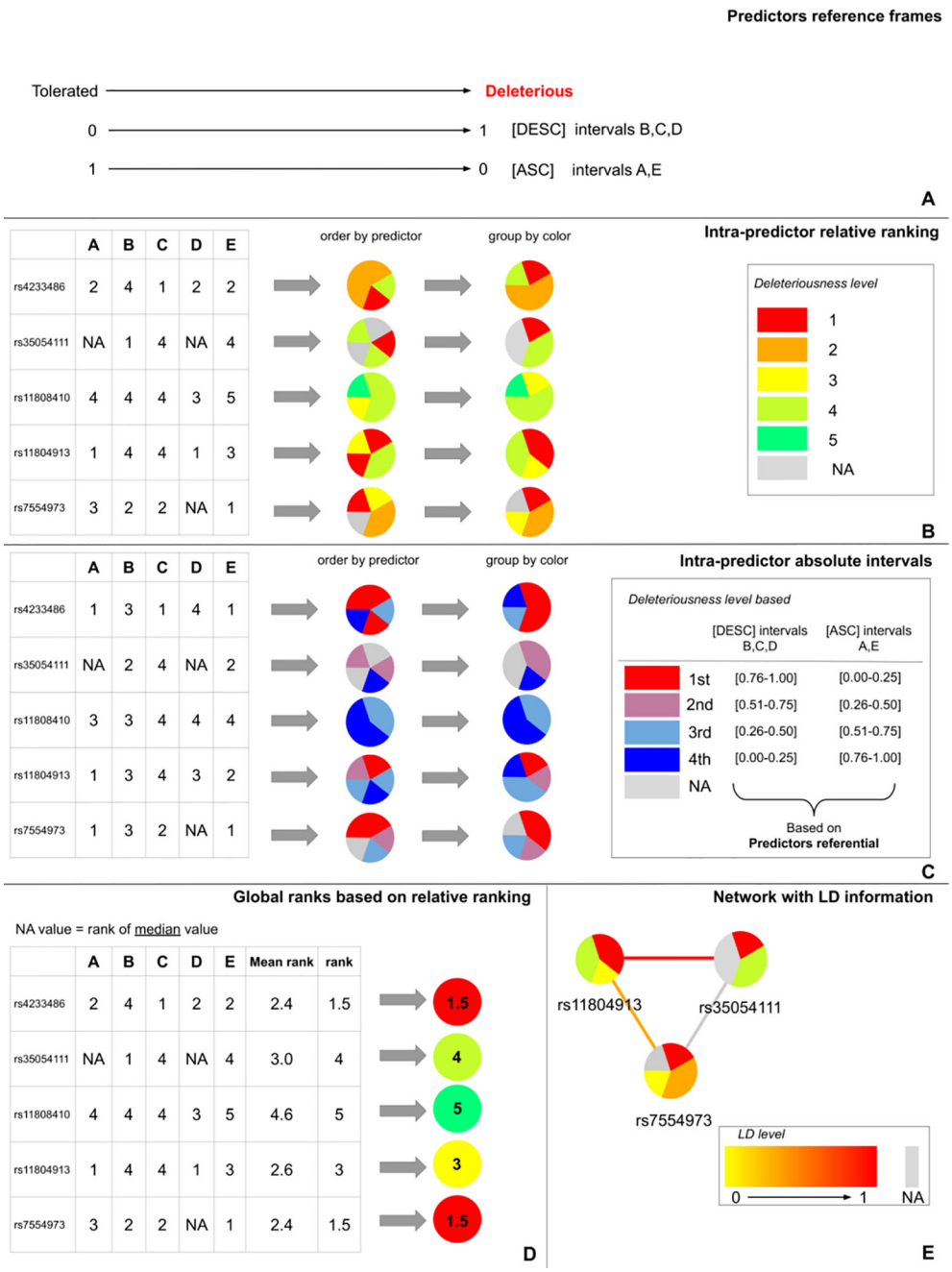
Intra-predictor ranks allow the prioritization of a list of variants relative to one another. According to the reference frames illustrated in **Figure 1A**, the five predictors produce scores ranging from 0 to 1. We can classify the five variants of interest from the most deleterious (rank 1) to the least deleterious (rank 5) with each predictor. In order to summarize this information in an easy-to-interpret representation, each variant is depicted as a pie chart where each slice represents the rank of the variant for one of the predictors. Thus, in the current analysis, five pie charts are generated and each pie chart is divided by five slices of the same size. We used a color gradient ranging from red to green, where red corresponds to the most deleterious variant (rank 1) among the candidates for a given predictor. The gray color represents missing data. **Figure 1B** depicts the pie charts generated for the five candidate variants. The slices can be ordered by color to allow easy identification of variants that appear the most deleterious across predictors.

Intra-Predictor Absolute Scores

Intra-predictor absolute scores allow prediction of variant deleteriousness in reference to the thresholds established for a particular predictor. Given these boundaries, we can determine

TABLE 1 | Deleterious scores generated by five different approaches.

	A	B	C	D	E
rs4233486	0.13	0.4	0.78	0.23	0.12
rs35054111	NA	0.7	0.21	NA	0.43
rs11808410	0.51	0.4	0.21	0.2	0.77
rs11804913	0.01	0.4	0.21	0.3	0.37
rs7554973	0.2	0.5	0.55	NA	0.01



	A	B	C	D	E
rs4233486	2	4	1	2	2
rs35054111	NA	1	4	NA	4
rs11808410	4	4	4	3	5
rs11804913	1	4	4	1	3
rs7554973	3	2	2	NA	1

	A	B	C	D	E
rs4233486	1	3	1	4	1
rs35054111	NA	2	4	NA	2
rs11808410	3	3	4	4	4
rs11804913	1	3	4	3	2
rs7554973	1	3	2	NA	1

	A	B	C	D	E	Mean rank	rank
rs4233486	2	4	1	2	2	2.4	1.5
rs35054111	NA	1	4	NA	4	3.0	4
rs11808410	4	4	4	3	5	4.6	5
rs11804913	1	4	4	1	3	2.6	3
rs7554973	3	2	2	NA	1	2.4	1.5

FIGURE 1 | DSNetwork visual approach. **(A)** Representation of predictors reference frames illustrating each approach boundaries and direction. **(B)** Representation of intra-predictors ranking based on the predictors reference frame. **(C)** Representation of intra-predictors absolute score intervals based on the predictors reference frame. **(D)** Representation of the global mean rank. **(E)** The edges between the nodes can be used to map Linkage Disequilibrium (LD) levels between two variants.

where each variant is located on the deleteriousness spectrum for each predictor. We chose to divide the score range of each approach into 20 equal intervals. This number of intervals was chosen as a compromise between granularity and readability. The first interval contains the most deleterious scores and the 20th, the least deleterious. Thus, the annotation scores obtained for each variant are translated into their corresponding intervals. This allows the user to know if a variant is predicted as deleterious by a particular approach without having to know the implementation details of this approach. For clarity purposes, in this example the range of scores has been divided into four intervals (instead of 20) (**Figure 1C**).

As for intra-predictor ranks, each variant is depicted as a pie chart where each slice represents the score interval of the variant

for a particular predictor. We used a color gradient ranging from red to blue. The red color represents the most deleterious interval for a given predictor. The gray color represents missing data. **Figure 1C** depicts the pie charts generated for the five candidate variants. The slices can be ordered by color to easily identify variants with the most predictions of deleteriousness.

Global Ranking

In order to further facilitate the prioritization, we propose to summarize the information regarding the relative ranks in an overall rank for each variant. To do so, we calculate the average rank of each variant based on its intra-predictor ranks. Then, we order the variants according to their average rank. Variants with the lowest average ranks are considered as the best candidates for being deleterious. Because in some cases there may be missing values for some of the predictors when analyzing a specific set of variants, we propose three strategies for calculating a consistent average rank, which will be comparable between variants and which will take into account these missing values: 1) replace missing values with the median value (default one); 2) replace missing values with the average value; or 3) systematically attribute missing values the "worst" rank. Once the necessary substitutions are made, the average ranks can be calculated and the global ranks generated. As for the intra-predictor scores and ranks, the global ranks are made available for each variant under the form of a pie chart where the rank is represented by a color gradient ranging from red to green. The color red represents the most deleterious variant among the candidates for all approaches (**Figure 1D**).

Variants Network

DSNetwork offers the possibility to simply visualize scores and LD between variants in order to identify potential haplotypes through an interactive interface. Users can interact with the network using the mouse by scrolling in and out to zoom, or double-click on a variant node to display variant annotation details among other features. They can also update the predictors used to prioritize the variants. As displayed in **Figure 1E**, edges between nodes can be used to map LD levels between two variants. LD (squared correlation r^2) is based on a user-chosen 1000 genomes population and is represented by an absolute color gradient ranging from yellow to red. Red indicates a high disequilibrium. The gray color represents the missing information. By default, no LD data are shown. To map LD on the network edges, users have to choose a population from 1000 Genomes and can restrict the LD range to display for a particular variant.

Implementation

DSNetwork was created using the Shiny framework (Chang et al., 2017). This tool provides users with deleteriousness predictions for a selected set of coding and non-coding human Single Nucleotide Variants (SNVs) and short inserts and deletions (InDels) (hg19 build) and generates a set of prioritized results for further analysis. These prediction scores are recovered from several trusted sources and presented in a cross-platform, user-friendly web interface. The interface is organized in three sections, namely, Input, Selection, and Visualization, as illustrated and described in **Figure 2**. For complete usage guide, see the **Supplementary Material**. DSNetwork is encapsulated using Docker platform to guarantee the cross-platform compatibility. The source code and installation procedure are available at https://bitbucket.org/vmtrap/dsnetwork_deploy/src/master/. The tool can be installed on all operating systems supporting Docker Engine (see supported platforms at https://docs.docker.com/install/) and is also accessible through our portal for demonstration purpose at: http://romix.genome.ulaval.ca/dsnetwork/.

Case Studies

We chose to demonstrate the utility of DSNetwork in the context of the functional analysis of four breast cancer susceptibility loci identified through the latest published breast cancer association study (full description in Michailidou et al., 2017) and nine loci reported in the latest published study on schizophrenia susceptibility (full description in Huo et al., 2019). Michailidou et al. (2017) report the discovery of 65 new breast cancer risk loci and deepens the functional characterization for four regions, namely, 1p36, 1p34, 7q22, and 11p15. For each of these regions, the authors defined sets of credible risk variants (CRVs) and investigated their impact through functional assays in order to identify the functional variants. Huo et al. (2019) investigated over 180 loci reported to be associated with schizophrenia in several GWA studies and prioritized regulatory single-nucleotide polymorphisms (SNPs) at these risk loci. They deepen the functional validation of 10 variants from nine different loci.

RESULTS AND DISCUSSION

Prioritization of Four Breast Cancer Susceptibility Loci

The original study by Michailidou et al. (2017) reported 65 novel breast cancer susceptibility loci. For each of these regions, they defined a set of CRV containing variants with P-values within two orders of magnitude of the most significant SNPs in this region. They then selected four loci for further evaluation, namely, 1p36, 1p34, 7q22, and 11p15. Initially, these four regions contained, respectively, 54, 13, 19, and 85 significantly associated variants. The p-value cutoff enabled them to reduce the number of variants to, respectively, 1, 4, 6, and 19 CRVs. The list of variants for these loci was extracted from the original paper's Supplementary Tables 8 and 13 in the context of the current analysis. Following data extraction, the analysis procedure was: 1) upload the variants of interest on the web tool, 2) fetch the annotations, 3) visualize the variants through the overview plot, 4) visualize the available deleteriousness scores through the relative ranking in the decision network, 5) use absolute interval visualizations to identify the best candidates, and finally 6) conclude.

Locus 1p36

This region contains a single CRV, rs2992756 ($P = 1.6\times10^{-15}$). For demonstration purposes, we selected the 30 most associated

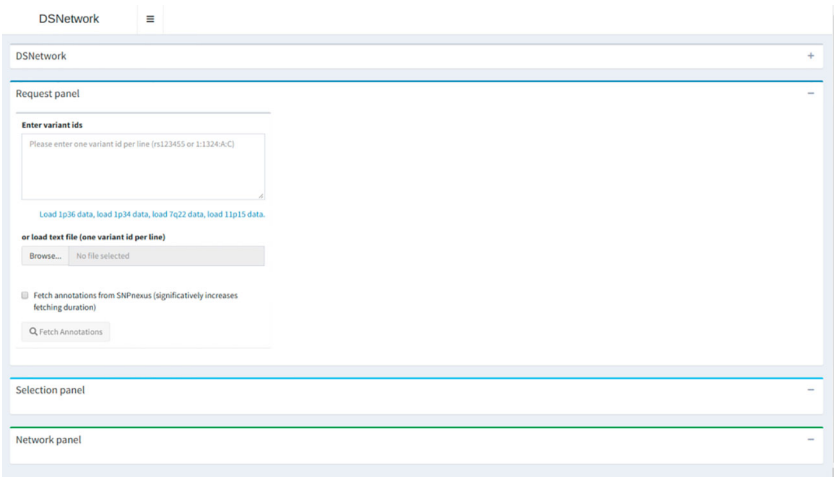


FIGURE 2 | Architecture overview. The first section is dedicated to user input and parameters for data retrieval. The middle panel presents a relevant subset of annotations for each submitted variant and enables the selection of variants to be integrated in the final visualization. The bottom part on the interface is dedicated to the integrated visualization of the deleteriousness predictions displayed as a network.

variants in this region to put to the test. Among these 30 variants, 2 variants (rs200439143, rs71018084) weren't annotated by DSNetwork because of their absence from MyVariant.info service, and 24 were identified as regulatory variants and 4 as non-synonymous variants. For the purposes of our analysis, we focused on the regulatory variants.

Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify rs2992756 as the best candidate. Indeed, the node for this variant contained the largest proportion of red, indicating a high ranking for most of the scoring approaches (**Figure 3A**). To confirm this observation, we used the relative rank visualization (**Figure 3B**). The mean rankings of variants, clearly materialized by both the color code and the values, enabled the confirmation of rs2992756 as the best candidate among the 30 most breast cancer-associated variants at the 1p36 locus. Using reporter assays, Michailidou et al. (2017) demonstrated that the presence of the risk T-allele of this variant within *KLHDC7A* promoter significantly lowers its activity.

Locus 1p34

This region contains four CRVs among 13 significantly associated variants. All the variants were found by DSNetwork and identified as regulatory variants.

Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify two variants, rs42334486 and rs7554973, as the best candidates. Indeed, the nodes for these variants contained the largest proportion of red and orange indicating a good ranking of these variants for most of the scoring approaches (**Figure 4A**). The sorting by color (**Figure 4B**) facilitated the prioritization of these two variants, which initially seemed to present the same proportion of high ranks. The visualization of the mean ranking confirms rs4233486 as the most credible candidate among the CRVs (**Figure 4C**). This observation is in accordance with results from Michailidou et al. (2017), which demonstrated, using reporter assays, that the presence of the risk T-allele of this variant within a putative regulatory element (PRE) reduced *CITED4* promoter activity.

Locus 7q22

This region contains six CRVs among 19 significantly associated variants. All the variants were found by DSNetwork and identified as regulatory variants.

Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify two variants, rs6961094 and rs71559437, as the best candidates. Indeed, the nodes for these variants contained the largest proportion of red, indicating a good ranking for most of the scoring approaches (**Figure 5A**). The visualization of the mean ranking confirms rs6961094 and rs71559437 as the most credible candidates among the CRVs (**Figure 5B**). These observations are supported by the functional experiments performed by Michailidou et al. (2017), which demonstrated, using allele-specific Chromatin Conformation Capture (3C) assays, that the presence of the risk haplotype (rs6961094 combined with rs71559437) is associated with chromatin looping between *CUX1*, *RASA4*, and *PRKRIP1* promoters suggesting that the protective alleles abrogate this phenomenon.

Locus 11p15

This region contains 19 CRVs among 85 candidate variants. Among the 19 CRVs, five variants, located in the proximal promoter of *PIDD1* (a gene implicated in DNA-damage-induced apoptosis and tumorigenesis; Lin et al., 2000), namely, rs7484123, rs7484068, rs11246313, rs11246314, and rs11246316, were further analyzed by Michailidou et al. (2017). They demonstrated, using reporter assays, that these variants, incorporated in a construct, significantly increased *PIDD1* promoter activity.

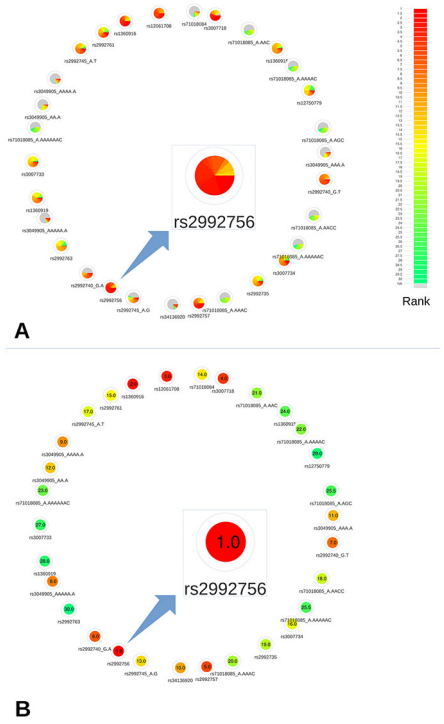


FIGURE 3 | Networks representing the 30 most significant variants associated with breast cancer at the 1p36 locus. **(A)** All available predictions represented under the form of relative rank grouped by color. **(B)** Global ranking representing the mean relative ranks with missing values substituted by the median value. Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify rs2992756 as the best candidate.

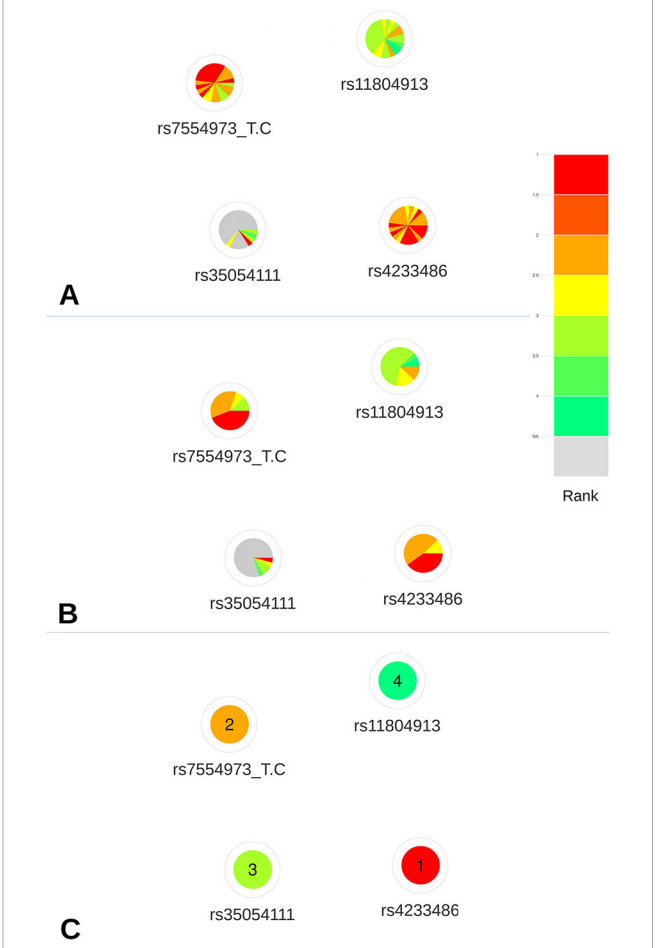


FIGURE 4 | Networks representing the four CRVs associated variants with breast cancer at the 1p34 locus. **(A)** All available predictions represented under the form of relative rank ordered by predictors. **(B)** All available predictions represented under the form of relative rank grouped by color. **(C)** Global ranking representing the mean relative ranks with missing values substituted by the median value. Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify two variants, rs42334486 and rs7554973, as the best candidates.

A quick overview of the relative and absolute metascores visualization allowed to easily prioritize the 19 CRVs (**Figures 6A** and **B**). First, the prioritized list based on the metascores confirms the selection of these five variants as functional credible SNPs. Indeed they are ranked at the first, second, third, fifth, and eighth place out of 19. Moreover, we notice that variants rs7484123 and rs11246314 demonstrate a higher level of coloration, confirming them as the best candidates among the variants located in the proximal promoter of *PIDD1*. The variant rs7484123 particularly stands out as a very promising candidate for subsequent experiments.

Prioritization of Nine Schizophrenia Susceptibility Loci

As a second example, we have applied DSNetwork to data from an extensive study by Huo et al. (2019) investigating over 180 loci reported to be associated with schizophrenia in several GWAS. This study has prioritized regulatory SNPs at these risk loci using five annotation methods (CADD, Eigen, LINSIGHT, GWAVA, and RegulomeDB) and expression quantitative loci (eQTL) annotation. Potentially causal SNPs have further been identified using functional genomics data such as CHIP-Seq experiments performed on brain tissues. Doing so and using reporter gene assays, they have validated the regulatory effect of nine transcription factor binding-disrupting SNPs from nine different loci.

The list of credible causal variants (CCV) for these nine loci was downloaded from the Psychiatric Genomics Consortium portal (https://www.med.unc.edu/pgc/results-and-downloads/scz/). These regions contained, respectively, 37 CCV on chromosome 1, 73 CCV on chromosome 3, 51 CCV on chromosome 6, 55 CCV on chromosome 7, 32 CCV on chromosome 12, 14 and 5 CCV on chromosome 15, 75 CCV on chromosome 16, and 128 CCV on chromosome 17.

The list of CCV for each locus was uploaded on the DSNetwork tool to identify the best functional candidates.

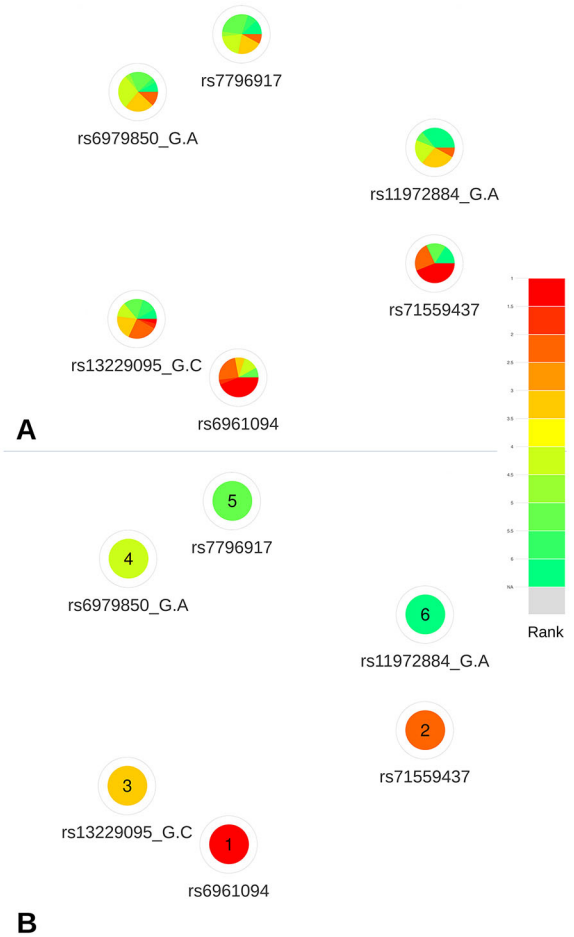


FIGURE 5 | Networks representing the six CRVs associated variants with breast cancer at the 7q22 locus. **(A)** All available predictions represented under the form of relative rank grouped by color. **(B)** Global ranking representing the mean relative ranks with missing values substituted by the median value. Based on the deleteriousness scores available for this subset of variants, a quick overview of variant nodes has allowed to easily identify two variants, rs6961094 and rs71559437, as the best candidates.

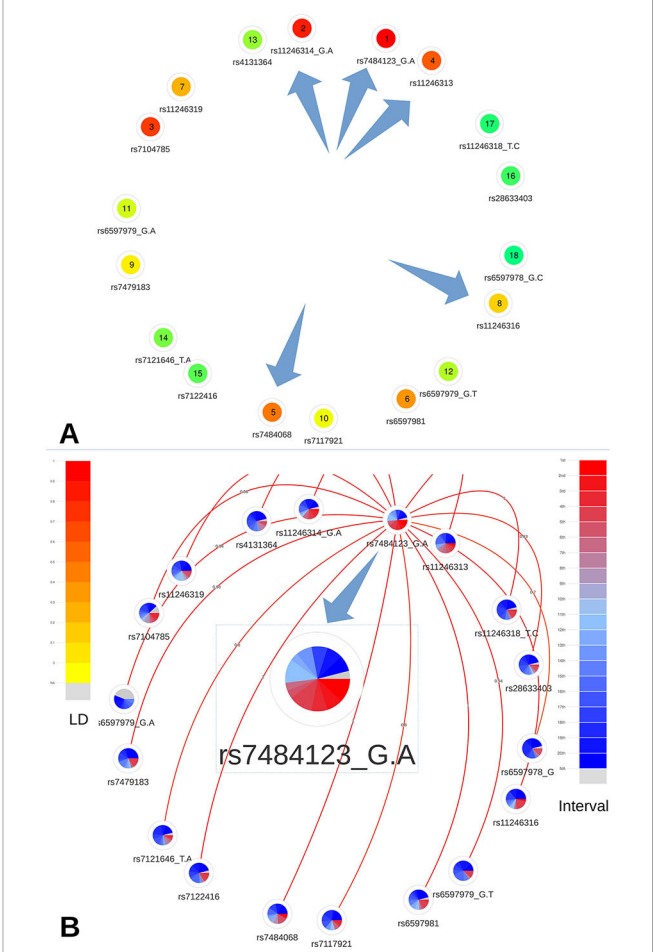


FIGURE 6 | Networks representing the 19 CRVs associated variants with breast cancer at the 11p15 locus. **(A)** Global ranking representing the mean relative ranks with missing values substituted by the median value. The purple arrows highlight the five credible causal variants identified by Michailidou et al. **(B)** The absolute intervals show rs7484123 and rs11246314 as the best candidates with regard to deleteriousness predictions. The best candidate variant rs7484123 sports a high level of linkage disequilibrium (depicted by the red links emanating from rs7484123's node) with the other candidate variants in the European population.

Table 2 presents, for each of the nine loci, the SNP that was prioritized in the original paper and in the DSNetwork analysis. In cases where results diverged, we also present the ranking provided by DSNetwork for the SNP prioritized in the original paper. From these analyses, we can conclude that DSNetwork found the same top SNP in the majority of cases (five SNPs ranked first and two SNPS ranked in the top 3). Two SNPs ranked in the top 10 but one of them rs696520 was not functionally validated in the original paper. Finally, rs17821573 on the chromosome 16 locus ranked 22nd with DSNetwork. It is important to note that fine-mapping analyses aim at reducing the list of candidate variants and not identifying the causal variant (Cannon and Mohlke, 2018). Furthermore, there is a difference between causal and functional variants: a variant showing a regulatory effect in functional assays does not confirm its implication in a phenotypic variation. Therefore, it would be interesting to test if the top SNP identified by DSNetwork (rs17854029) could also be functional.

These examples demonstrate the ability of DSNetwork to effectively reduce the amount of CCV despite a large number of candidate variants.

Furthermore, compared to other existing methods for prioritization, DSNetwork has the advantage of being scalable and flexible. Indeed, as a majority voting based approach where each predictor is a crowdsourcing annotator proposing its prioritized list, DSNetwork enables the addition of an infinite number of annotators. However, in practice, one drawback of usual crowdsourcing systems is that the annotators are

TABLE 2 | Summarized results from DSNetwork analysis for the nine schizophrenia susceptibility loci.

Locus	# of CCV	Huo et al. top SNP	Validated	DSNetwork top SNP	Huo et al. top SNP in DSNetwork
chr1	37	rs301791	Yes	rs301791	1
chr3	73	rs696520	No	rs9845457	7
chr6	51	rs7752421	Yes	rs7752421	1
chr7	55	rs37718	Yes	rs37718	1
chr12	32	rs7304782	Yes	rs7304782	1
chr15 1	14	rs28676999	No	rs62021888	3
chr15 2	5	rs4702	No	rs4702	1
chr16	75	rs17821573	Yes	rs17854029	22
chr17	128	rs11655813	Yes	rs216172	3
chr17	128	rs9908888	Yes	rs2281727	7

anonymous. Therefore, their expertise levels are often unknown and uneven, which makes it difficult for the end-user to trust the final vote. In DSNetwork, the annotations are derived from several databases and their reliability level can be estimated through their performance reported in the literature. By default, all the available predictors are used to produce an optimal decision. However, we enable users to adjust the list of predictors used according to their preferences and expertise. As explained in Ribeiro et al. (2016), "explaining the rationale behind individual predictions would make us better positioned to trust or mistrust the prediction, or the classifier as a whole." For this reason, in order to assist the users in their decision, we provide a short description of each predictor and the list of the annotations they use. Another way to take into account annotator reputation is to add a weight to each vote, the weights representing the competence levels (Tao et al., 2019). This explicit way to incorporate weight in the voting process could be included in further development.

CONCLUSION

We analyzed four breast cancer risk loci through DSNetwork and were able to pinpoint the same most plausible causal variants than those proposed in the original paper. In a similar way, we were able to efficiently circumscribe the number of credible candidate variants throughout the prioritization of nine schizophrenia susceptibility loci. DSNetwork provides a user-friendly interface integrating predictors for both coding and non-coding variants in an easy-to-interpret visualization to assist the prioritization process. The use of DSNetwork greatly facilitates the selection process of potentially deleterious variants by aggregating the results of nearly 60 prediction approaches and easily highlighting the best candidate variants for further functional analysis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: PMID:29059683, PMID:30737407, and https://www.med.unc.edu/pgc/results-and-downloads/scz/.

AUTHOR CONTRIBUTIONS

AL designed and implemented DSNetwork software, conducted literature searches, researched data, and selected relevant articles. AL also created figures and tables, and wrote, formatted, and finalized the article for submission. AL, PS, M-PS-B, and RO-C were in charge to test the software and report all bugs. AL, PS, M-PS-B, and RO-C helped to optimize DSNetwork. AL, PS, JS, and AD supervised and reviewed the design of the study. All authors contributed to writing and reviewing the manuscript.

FUNDING

The PERSPECTIVE and PERSPECTIVE I&I projects were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH1293344, grant GP1-155865), the Ministere de l'Économie, Science et Innovation du Québec through Genome Quebec and the Quebec Breast Cancer Foundation.

ACKNOWLEDGMENTS

We thank the BCAC for providing the impetus to create DSNetwork. We thank the SNPnexus team (Barts Cancer Institute, Queen Mary University of London) for their assistance in integrating SNPnexus data. We also thank Pr. Bing-Jian Feng (University of Utah School of Medicine) for his help and advice regarding BayesDel integration. We are also grateful to all the personnel of Arnaud Droit's lab and particularly Gwenaelle Lemoine, Julien Prunier, and Benjamin Vittrant for their advice and assistance in the preparation of this article. We wish to extend our thanks to all the current and future DSNetwork users. This manuscript has been released as a Pre-Print at (Lemaçon et al., 2019)".

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01349/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Cannon, M. E., and Mohlke, K. L. (2018). Deciphering the emerging complexities of molecular mechanisms at GWAS loci. *Am. J. Hum. Genet.* 103, 637–653. doi: 10.1016/j.ajhg.2018.10.001

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., and McPherson, J. (2017). shiny: Web application framework for R [computer software]. URL http://CRAN.R-project.org/package=shiny (R package version 1. 0. 0).

Dayem Ullah, A. Z., Oscanoa, J., Wang, J., Nagano, A., Lemoine, N. R., and Chelala, C. (2018). SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Res.* 46, W109–W113. doi: 10.1093/nar/gky399

di Iulio, J., Bartha, I., Wong, E. H. M., Yu, H.-C., Lavrenko, V., Yang, D., et al. (2018). The human noncoding genome defined by genetic diversity. *Nat. Genet.* 50, 333–337. doi: 10.1038/s41588-018-0062-7

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi: 10.1093/hmg/ddu733

Feng, B.-J. (2017). PERCH: a unified framework for disease gene prioritization. *Hum. Mutat.* 38, 243–251. doi: 10.1002/humu.23158

Huang, Y.-F., Gulko, B., and Siepel, A. (2017). Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* 49, 618–624. doi: 10.1038/ng.3810

Huo, Y., Li, S., Liu, J., Li, X., and Luo, X.-J. (2019). Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nat. Commun.* 10, 670. doi: 10.1038/s41467-019-08666-4

Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* 48, 214–220. doi: 10.1038/ng.3477

Lemaçon, A., Scott-Boyer, M., Soucy, P., Ongaro-Carcy, R., Simard, J., and Droit, A. (2019). DSNetwork: an integrative approach to visualize predictions of variants' deleteriousnessBioRvix [Preprint]. Available at: https://doi.org/10.1101/526335 (Accessed January 21, 2019).

Lin, Y., Ma, W., and Benchimol, S. (2000). Pidd, a new death-domain-containing protein, is induced by p53 and promotes apoptosis. *Nat. Genet.* 26, 122–127. doi: 10.1038/79102

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.* 45, D896–D901. doi: 10.1093/nar/gkw1133

Michailidou, K., Lindstrom, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94. doi: 10.1038/nature24284

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier, in: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (New York, NY, United States: Association for computing machinery). pp. 1135–1144.

Tao, D., Cheng, J., Yu, Z., Yue, K., and Wang, L. (2019). Domain-weighted majority voting for crowdsourcing. *IEEE Trans. Neural Netw. Learn Syst.* 30, 163–174. doi: 10.1109/TNNLS.2018.2836969

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., et al. (2016). High-performance web services for querying gene and variant annotation. *Genome Biol.* 17, 91. doi: 10.1186/s13059-016-0953-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 January 2020
doi: 10.3389/fgene.2019.01353

Non-Negative Symmetric Low-Rank Representation Graph Regularized Method for Cancer Clustering Based on Score Function

Conghai Lu[1], Juan Wang[1], Jinxing Liu[1], Chunhou Zheng[2], Xiangzhen Kong[1] and Xiaofeng Zhang[3]*

[1] School of Information Science and Engineering, Qufu Normal University, Rizhao, China, [2] College of Electrical Engineering and Automation, Anhui University, Hefei, China, [3] School of Information and Electrical Engineering, Ludong University, Yantai, China

As an important approach to cancer classification, cancer sample clustering is of particular importance for cancer research. For high dimensional gene expression data, examining approaches to selecting characteristic genes with high identification for cancer sample clustering is an important research area in the bioinformatics field. In this paper, we propose a novel integrated framework for cancer clustering known as the non-negative symmetric low-rank representation with graph regularization based on score function (NSLRG-S). First, a lowest rank matrix is obtained after NSLRG decomposition. The lowest rank matrix preserves the local data manifold information and the global data structure information of the gene expression data. Second, we construct the Score function based on the lowest rank matrix to weight all of the features of the gene expression data and calculate the score of each feature. Third, we rank the features according to their scores and select the feature genes for cancer sample clustering. Finally, based on selected feature genes, we use the K-means method to cluster the cancer samples. The experiments are conducted on The Cancer Genome Atlas (TCGA) data. Comparative experiments demonstrate that the NSLRG-S framework can significantly improve the clustering performance.

Keywords: cancer gene expression data, low-rank representation, feature selection, score function, clustering

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Shanfeng Zhu,
Fudan University, China
Yushan Qiu,
Shenzhen University, China

****Correspondence:***
Juan Wang
wangjuansdu@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *08 September 2019*
Accepted: *10 December 2019*
Published: *22 January 2020*

Citation:
Lu C, Wang J, Liu J, Zheng C, Kong X and Zhang X (2020) Non-Negative Symmetric Low-Rank Representation Graph Regularized Method for Cancer Clustering Based on Score Function. Front. Genet. 10:1353. doi: 10.3389/fgene.2019.01353

INTRODUCTION

High-throughput DNA microarray technology has long been used to collect biomedical cancer gene expression data (Russo et al., 2003). In general, gene expression data contain a notably large number of genes (high dimension), a small number of samples (low sample size), irrelevant genes and noisy genes caused by complex processing (Mohamad et al., 2010). Therefore, it is necessary to select feature genes or informative genes that contribute to identifying different cancers and the cancerous state (Mohamad et al., 2013; Ge and Hu, 2014; Tang et al., 2014). The selected genes have potential for use in developing cancer treatment strategies (Rappoport and Shamir, 2018). However, the high-dimensional and low-sample-size characteristics of the cancer gene expression dataset present a

challenge for researchers in terms of data mining. To mitigate this problem, researchers have proposed many methods (Cui et al., 2013; Ge and Hu, 2014; Wang et al., 2016; Wang et al., 2018; Xu et al., 2019). Among the existing methods, feature selection is a reasonable method that has achieved great success.

Feature selection is an important data processing method that can select the most important feature subset from a set of features and reduce the dimension of the feature space. The existing feature selection methods can be divided into two groups: "wrapper" methods and "filter" methods (Kohavi and John, 1997). Wrapper methods use the learning algorithm to evaluate the candidate features. However, because wrapper methods are highly complex with a large amount of calculation, they are not suitable for large-scale datasets (Langley, 1994). Filter methods select a feature subset via the evaluation function. Construction of an evaluation function is based on the correlations between the features and properties of the raw data, such as the distance measures, information measures, dependence measures or others (Dash and Liu, 1997; Talavera, 2005; He et al., 2006). Among the existing evaluation functions, as a criterion, the data variance might be the simplest evaluation for feature selection. The main idea of the data-variance-based approach is to capture the directions of the maximum variance in the data, which reflects the major power of the data. The Principal Component Analysis (PCA) method and its variants belong to the filter methods and are used to find features that are useful for recovering data. However, there is no reason to confirm that selected features can effectively discriminate between data points in different classes. He et al. proposed the Laplacian Score (LS) method to select features with high identification, and the LS method is a "filter" method that is independent of other methods (He et al., 2006). The LS method constructs a nearest neighbour graph to preserve the local geometric structure. The selected features can reflect the local structure of the data space.

As we know, the global structure plays an important role in clustering when the data contain multiple subspaces (Liu et al., 2010). The LS method focuses excess attention on the relationships between local data points but ignores the influence of global data structures. This drawback might lead to reduced discrimination effects of the selected feature when the given data contain multiple subspaces. For the feature selection method, it is a challenge to satisfactorily characterize and represent global data structures from a dataset with multiple subspaces. Fortunately, the Low-Rank Representation (LRR) method solves this issue nicely. The LRR method can find a low-rank matrix to capture and represent the global structure of the raw dataset (Liu et al., 2010). The key to the LRR method is that the high-dimensional data can be represented by potential low-dimensional subspaces (You et al., 2016). In bioinformatics, LRR has achieved great success in gene expression data mining. For example, Cui et al. used the LRR method to identify subspace gene clusters and obtained good results (Cui et al., 2013). To preserve the intrinsic geometric structures of gene expression data, Wang et al. introduced graph regularization into LRR and proposed the Laplacian regularized LRR (LLRR) method (Wang et al., 2016). Recently, LLRR was applied to cancer sample clustering (Wang et al., 2019a). Furthermore, Wang et al. introduced the mixed-norm to increase the robustness of the LLRR method and proposed the mixed-norm Laplacian regularized LRR (MLLRR) method for tumour sample clustering based on penalized matrix decomposition (Wang et al., 2018). However, cancer sample clustering is processed on the obtained low-rank matrix, which is the global structural representation of the raw data. These LRR-based approaches mainly consider the global structure of data, but sometimes they ignore the single feature gene.

Motivated by the above insights, we propose a novel framework that integrates the advantages of the LRR and LS methods. Based on the multi-cancer gene expression dataset, the proposed framework is used to select the feature gene for cancer sample clustering.

First, we incorporate the constraints of the non-negative symmetric low-rank matrix and graph regularization in the LRR method and propose a non-negative symmetric low-rank representation graph regularized method, or NSLRG method for short. The NSLRG method considers the property and structure of the gene expression data. The NSLRG method obtains the lowest rank matrix, which preserves the local data manifold information and the global data structure information of the raw data.

Second, according to the lowest rank matrix, we construct a Score function to evaluate each gene for selection of the feature genes. The importance level of a gene depends on its significance for the global and local structures of the raw data. We integrate the NSLRG method with the Score function to achieve the aim of evaluating and selecting feature genes, and we refer to it as the NSLRG-S framework.

Finally, we apply the K-means method to cluster cancer samples based on the selected feature genes. Based on the different multi-cancer gene expression data, the experimental results suggest that the performance of the NSLRG-S framework is better than that of other methods.

In summary, the contributions of this paper include the following main aspects:

(1.) We propose a novel data mining method known as the NSLRG method. The NSLRG method operates under graph regularization and non-negative symmetric low-rank matrix constraints. The NSLRG method can learn the lowest rank matrix to satisfactorily represent the gene expression data and can capture the global structures and local geometric structures of the raw data. Non-negativity is more consistent with biological modelling. The symmetric constraint improves the interpretability of the lowest rank matrix. The constraints of non-negativity and symmetry facilitate the lowest rank matrix to learn the structure of the gene expression data.

(2.) Based on the lowest rank matrix, we propose a Score function to select the feature genes for cancer sample clustering. The selected feature genes have important significance to the raw data. In the clustering of cancer samples, the selected genes have strong discriminability to realize the classification of different samples.

(3.) We present a novel feature selection framework, known as NSLRG-S, that is designed to evaluate and select the feature genes for cancer sample clustering. Based on this framework, the selected result of the gene expression dataset has lower dimensionality. In multi-cancer sample clustering, this method has a high recognition rate to find subsets using the selected result as experimental data. We conduct extensive experiments to demonstrate that the feature gene subset selected by NSLRG-S has good performance in cancer sample clustering.

The remainder of this paper is organized as follows. In section *Related Work*, we briefly review the original LRR and several related variants as well as the LS method. In section *Method*, we first present the NSLRG method and its optimal solution, and based on the Score function, the NSLRG-S framework is clearly given for modelling of multi-cancer gene expression data. Section *Experiments* analyses and discusses the NSLRG method based on multiple evaluation indicators and convergence analysis. The performance of the NSLRG-S framework is validated by experiments based on synthetic data and multi-cancer gene expression data. Section *Conclusions Work* presents the conclusion of our work.

RELATED WORK

In this section, we briefly introduce the original Low-Rank Representation (LRR) (Liu et al., 2010), the related variants based on the original LRR method, and the Laplacian Score method (He et al., 2006).

Low-Rank Representation

Original LRR Method

The Low-Rank Representation (LRR) method is an efficient method for exploring observed data and subspace clustering. The main idea is that each data sample can be represented as a linear combination of the dictionary data. In general, the matrix $\boldsymbol{X} = [x_1,x_2,\ldots,x_n] \in \mathbb{R}^{m\times n}$ represents the observed data, of which each column is a data sample. Therefore, the matrix $\mathbf{X}$ contains n data samples drawn from independent subspaces. The matrix $\boldsymbol{D} = [d_1,d_2,\ldots,d_k] \in \mathbb{R}^{m\times k}$ represents the dictionary data and is overcomplete. The general model of the LRR method is formulated as follows.

$$\min_{\mathbf{Z}} rank(\mathbf{Z}) \quad s.t. \quad \mathbf{X} = \mathbf{DZ}, \tag{1}$$

where the matrix $\boldsymbol{Z} \in \mathbb{R}^{k\times n}$ is the coefficient matrix. The aim of this model is to learn a lowest rank matrix $\mathbf{Z}^{\star}$ to represent the observed data $\mathbf{X}$. In the actual application, the matrix $\mathbf{X}$ always replaces $\mathbf{D}$ as the dictionary data (Liu et al., 2010; Liu et al., 2013). Therefore, $\mathbf{Z}$ becomes a square matrix and $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$. The element $z_{ij} \in \mathbf{Z}^{*}_{n\times n}$ can denote the confidence of sample i and j in the same subspace (Wang et al., 2019b). Hence, the matrix $\mathbf{Z}^{\star}$ can be used in subspace clustering that clusters data samples into several sets, with each set corresponding to a subspace.

The problem of $\min_{\mathbf{Z}} rank(\mathbf{Z})$ is a rank function, which is difficult to optimize with an NP-hard nature. To mitigate this problem, the best alternative is convex relaxation on problem (1), and it is written as follows.

$$\min_{\mathbf{Z}} \| \mathbf{Z} \|_{\star} \quad s.t. \quad \mathbf{X} = \mathbf{XZ}, \tag{2}$$

where $\|\cdot\|_{\star}$ is the nuclear norm, and $\|\boldsymbol{Z}\|_{\star}$ is defined as $\| \mathbf{Z} \|_{\star} = \sum_i^n \delta_i$, where δ_i is the singular value of matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times n}$. It has been confirmed in the literature (Cai et al., 2010) that matrix $\mathbf{Z}$ of the LRR can capture the global structure of the raw data using the nuclear norm item. Furthermore, to address the real data under the noise and outliers, a more reasonable formula is applied after adjustment, and it is expressed as follows.

$$\min_{\mathbf{Z},\mathbf{E}} \| \mathbf{Z} \|_{\star} + \lambda \| \mathbf{E} \|_{P} \quad s.t. \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \tag{3}$$

where $\|\boldsymbol{E}\|_P$ is the error term, and it selects a different P to model special noise or outliers based on error prior information, such as l_1-norm ($\|\boldsymbol{E}\|_1$) and $l_{2,1}$- norm ($\|\boldsymbol{E}\|_{2,1}$) (Chen and Yang, 2014), and $\lambda > 0$ is the parameter that trades off the effect of the error item.

Many researchers have attempted and proposed variants based on the original LRR method. The main idea is to introduce constraint items to optimize or improve existing methods. For example, the original LRR method is improved by considering the geometric structures within the data, including the graph regularization method (Lu et al., 2013) and k-nearest neighbour graph method (Yin et al., 2016). The different norm items are used to improve the robustness of the original LRR method (Wang et al., 2018) and others.

LRR With Graph Regularization

Under certain conditions, the geometric structure within the data is crucial for the result that we desire. To address this issue, researchers introduced graph regularization into the LRR method to create the graph-regularized low-rank representation (GLRR) method (Lu et al., 2013). The equation of GLRR is written as follows.

$$\min_{\mathbf{Z},\mathbf{E}} \| \mathbf{Z} \|_{*} + \lambda_1 tr\left(\mathbf{ZLZ}^{\mathrm{T}}\right) + \lambda_2 \| \mathbf{E} \|_{2,1} \quad s.t. \quad \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \tag{4}$$

where the error item uses the $l_{2,1}$-norm and $\| \mathbf{E} \|_{2,1} = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} ([\mathbf{E}]_{ij})^2}$, $tr(\cdot)$ is the trace of the matrix, $\mathbf{L}$ is the graph Laplacian, and λ_1 and λ_2 are two parameters used to balance the graph-regularized item and the error item. Based on manifold learning, the graph-regularized item achieves the aim that representative data points z_i and z_j can hold the property of the data points x_i and x_j of $\mathbf{X}$, which are closed in the intrinsic manifold. Therefore, the inherent geometric structure in the raw data is preserved in the low-rank matrix $\mathbf{Z}$.

Non-Negative LRR With Sparsity

The non-negativity constraint ensures that every data point is in the convex hull of its neighbours. The sparse constraint ensures that each sample is associated with only a few samples. The non-

negative and sparse low-rank matrix supplies a well discriminated weight for the subspace and information group.

Inspired by the above insights, Zhuang et al. proposed the non-negative low rank and sparse graph (NNLRS) method (Zhuang et al., 2012). The formula is given as follows.

$$\min_{\mathbf{Z},\mathbf{E}} \| \mathbf{Z} \|_* + \lambda_1 \| \mathbf{Z} \|_1 + \lambda_2 \| \mathbf{E} \|_{2,1} \ \ s.t. \ \ \mathbf{X} = \mathbf{XZ} + \mathbf{E}, \ \ \mathbf{Z} > 0, \quad (5)$$

where $\|\boldsymbol{Z}\|_1$ is the l_1-norm to guarantee the sparsity of coefficient matrix. In real-world applications, the sparsity and non-negativity matrix $\mathbf{Z}$ obtained by the NNLRS method can offer a basis for semi-supervised learning by constructing the discriminative and informative graph (You et al., 2016).

Laplacian Score Method

According to the Laplacian eigenmaps (Belkin and Niyogi, 2001) and the locality preserving projection (He and Niyogi, 2005), the aim of the Laplacian Score (LS) method is to evaluate features based on their locality preserving power (He et al., 2006). The LS is defined as follows.

$$LS(r) = \frac{\sum_{ij} \left(x_{ri} - x_{rj}\right)^2 S_{ij}}{Var\left(x_{r,:}\right)}, \ (1 \le r \le m, 1 \le i \le j \le n), \quad (6)$$

where the heat kernel function $S_{ij} = e^{-\frac{\| x_i - x_j \|^2}{t}}$ is used to obtain weight matrix S, and t is a suitable constant, which is set empirically. The matrix S is used to model the local structure of the raw data space. Additionally, $Var(x_{r,:})$ is the estimated variance of the r-th feature in all data points, and the larger the $Var(x_{r,:})$, the more information held by the r-th feature. The $\sum_{ij}(x_{ri} - x_{rj})^2$ is the sum of differences in the expression of r-th feature between all samples. For larger values of S_{ij} and the smaller values of $\sum_{ij} (x_{ri} - x_{rj})^2$, the value of $LS(r)$ tends to be smaller, meaning that the importance level of the feature is higher. Therefore, the important features are selected according to $LS(r)$.

METHOD

In this section, we propose a novel feature selection framework to select the feature genes for cancer clustering. This framework is set up based on the NSLRG method and the Score function. We refer to this approach as the NSLRG-S Subsection *NSLRG Method* presents the NSLRG method and its optimization algorithm. In subsection *NSLRG With Score Function*, we introduce the NSLRG method with the Score function. The last subsection *Framework of NSLRG-S* is devoted to clustering of cancer samples based on NSLRG-S modelling of gene expression data.

NSLRG Method

Graph Regularization

Because graph regularization can preserve the intrinsic local geometric structure in the original data, it has received much attention from researchers. The theory of graph regularization is based on the principle that the representation of the intrinsic local geometric structure that is distributed in the original data is inherited by a graph under the new basis mapping. In the graph, the vertices correspond to the data points, and the edge weights represent the relationships between the data points (Du et al., 2017). Thus far, graph theory has been widely applied and developed (Chen et al., 2018).

For this paper, in the step of graph construction, we assume that if data points x_i and x_j are "close", an edge exists between x_i and x_j. In this work, we use the K-nearest neighbour method to find the connection of x_i and x_j. In other words, if x_i or x_j is among the K-nearest neighbours of each other, the data points x_i and x_j are located on the same edge. This construction strategy is simpler for determination of connected edges, which tends to lead to a connected graph. In the next step, the edge weights are defined to represent the affinity between the data points. In current study, we define a symmetric weighting matrix $\mathbf{W}$ by the heat kernel weighting function (Cai et al., 2005). The weighting formula is defined as follows.

$$\boldsymbol{W_{ij}} = \begin{cases} \boldsymbol{e^{-\frac{\| x_i - x_j \|^2}{t}}}, \ \boldsymbol{if \ x_i \ and \ x_j \ are \ connected} \\ \mathbf{0}, \qquad \boldsymbol{otherwise} \end{cases}, \quad (7)$$

where the parameter t is defined as the mean value of the Euclidean distance for all data points, which can be automatically adjusted based on the different dataset. Therefore, the degree matrix $\mathbf{D}$ is defined as $\mathbf{D}_{ii} = \sum_j W_{ij}$, which is a diagonal matrix. Finally, based on the connected graph, we obtain the graph Laplacian matrix $\mathbf{L}$, which is defined as follows.

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \quad (8)$$

Accordingly, a reasonable minimize objective function exists to satisfy our assumption, and it is defined as follows.

$$\begin{aligned} \min_{\mathbf{z}} \sum_{ij} \| \boldsymbol{z_i} - \boldsymbol{z_j} \|^2 \boldsymbol{W_{ij}} &= \min_{\mathbf{z}} \boldsymbol{tr}\left(\mathbf{Z}(\mathbf{D} - \mathbf{W})\mathbf{Z}^{\mathbf{T}}\right) \\ &= \min_{\mathbf{Z}} \left(\mathbf{ZLZ}^{\mathbf{T}}\right), \end{aligned} \quad (9)$$

where $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ are mappings of x_i and x_j under the new basis, which are also close to each other if x_i and x_j are close. The objective function is known as the graph regularization item.

Objective Function

We introduce graph regularization and sparse items into the original LRR. Furthermore, we impose the non-negative and symmetric constraints on the low-rank matrix $\mathbf{Z}$. This method is known as the non-negative symmetric low-rank representation graph regularized (NSLRG) method, and its objective function is written as follows.

$$\min_{\mathbf{Z},\mathbf{E}} \| \mathbf{Z} \|_* + \lambda_1 tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \lambda_2 \| \mathbf{E} \|_1 + \lambda_3 \| \mathbf{Z} \|_0 \quad (10)$$
$$s.t. \ \ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \ \ \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \ \ \mathbf{Z} > 0.$$

In the NSLRG method, we represent a given set of data points as a linear combination of other points using a low-rank matrix **Z**. The low-rank matrix should be sparse to improve the recognition ability. Therefore, the matrix **Z** with a sparse constraint could make the result of the representation more discriminative. However, the $\|Z\|_0$ item of problem (10) is NP-hard. Thus, as suggested by matrix completion methods (Candès et al., 2011), we use $\|Z\|_1$, a proper relaxed convex item, to replace $\|Z\|_0$, and the objective function of NSLRG can be rewritten as follows.

$$\min_{\mathbf{Z},\mathbf{E}} \| \mathbf{Z} \|_* + \lambda_1 tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \lambda_2 \| \mathbf{E} \|_1 + \lambda_3 \| \mathbf{Z} \|_1 \quad (11)$$
$$s.t. \ \ \mathbf{X} = \mathbf{X}\mathbf{Z} + \mathbf{E}, \ \ \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \ \ \mathbf{Z} > \mathbf{0}.$$

The matrix $\mathbf{Z}^\star$ is learned by the NSLRG method, and matrix $\mathbf{Z}^\star$ is a non-negative symmetric lowest rank matrix. The element $\mathbf{z}_{ij}$ of $\mathbf{Z}^\star$ can be treated as the degree of similarity between the data points x_i and x_j. In addition, the obtained matrix $\mathbf{Z}^\star$ has good interpretability, for which the element of matrix $\mathbf{Z}^\star$ can be directly converted to similar-degree weights. The symmetry constraint can strictly guarantee the consistency of similarity of data pairs. The similarity of data points i and j corresponding to the similar-degree weights elements $\mathbf{z}_{ij}$ and $\mathbf{z}_{ji}$ is equal, as shown as **Figure 1**. The non-negative constraint is more adaptive for the property of the gene expression data. In other words, the NSLRG method avoids the situation in which the lowest rank matrix might be negative and asymmetric, and it also avoids symmetrization of itself, as suggested in (Liu et al., 2010), i.e., $\mathbf{Z}^\wedge = (|\mathbf{Z}^\star| + |\mathbf{Z}^\star|^{\mathrm{T}})/2$. Therefore, we refer to the matrix $\mathbf{Z}^\star$ as the similar-degree matrix.

Optimization

As we know, many algorithms are based on convex relaxation to solve the high-dimension optimization problem, such as Singular Value Thresholding (SVT) (Cai et al., 2010), Accelerated Proximal Gradient (APG) (Toh and Yun, 2010), Alternating Direction Method (ADM) (Lin et al., 2009) and Linearized Alternating Direction Method with Adaptive Penalty (LADMAP) (Lin et al., 2011). As an extended ADM, the LADMAP algorithm adds the quadratic penalty term linearization and the penalty self-adaption change, which leads to use of fewer auxiliary variables and avoids matrix inversions to solve the problem. Specifically, LADMAP reduces the complexity of the LRR from $O(n^3)$ to $O(rn^3)$, where r is the rank of low-rank matrix **Z**. This algorithm makes it possible for LRR to be applied on large-scale dataset, such as video surveillance, digital images, and gene expression data. Therefore, the LADMAP algorithm has been recognized as the most efficient algorithm for solving the problem of convex relaxation of low-rank and sparse matrices. Similarly, we also adopt LADMAP to solve (11).

First, to easily and effectively obtain matrix **Z**, we use an auxiliary variable **Q** to separate the variables, i.e., nuclear norm ($\|Z\|_*$) and l_1-norm ($\|Z\|_1$). The objective function can be rewritten as equation (12) using the Augmented Lagrange Multiplier method (Lin et al., 2010).

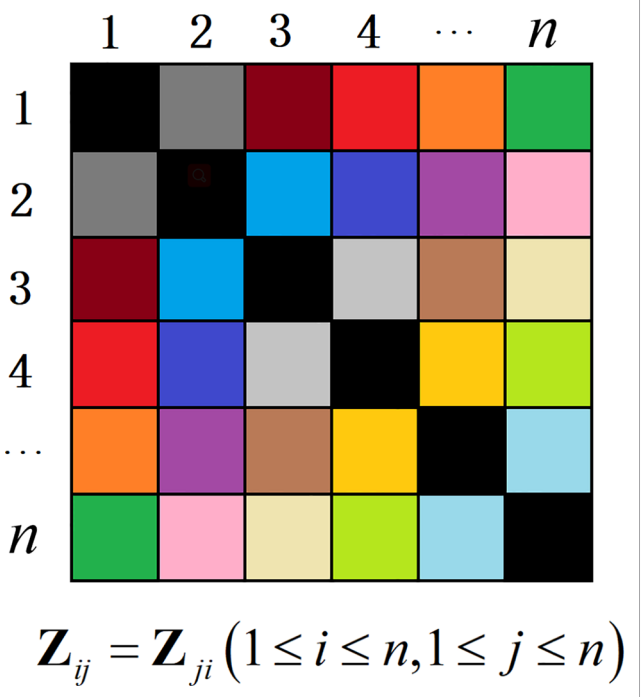


FIGURE 1 | The matrix **Z** with the symmetry constraint.

$$\ell(\mathbf{Z},\mathbf{E},\mathbf{Q},\mathbf{Y}_1,\mathbf{Y}_2,\mu) = \min_{\mathbf{Z},\mathbf{E},\mathbf{Q}} \| \mathbf{Z} \|_* + \lambda_1 tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right)$$
$$+\lambda_2 \| \mathbf{E} \|_1 + \lambda_3 \| \mathbf{Q} \|_1 + \mathbf{Y}_1, \ \mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E} + \mathbf{Y}_2, \ \mathbf{Z} - \mathbf{Q} \quad (12)$$
$$+\frac{\mu}{2} \| \mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E} \|_F^2 + \frac{\mu}{2} \| \mathbf{Z} - \mathbf{Q} \|_F^2 \ \ s.t. \ \ \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} \ge 0,$$

where λ_1, λ_2, and λ_3 are positive weighting parameters; $\mu > 0$ is the penalty parameter; $\mathbf{Y}_1$,$\mathbf{Y}_2$ are Lagrangian multipliers; **A**, **B**=$tr(\mathbf{A}^{\mathrm{T}}\mathbf{B})$ is the Euclidean inner product between the matrices **A** and **B**; and $\|\cdot\|_F$ is the Frobenius-norm. Mathematically, equation (12) is equivalent to equation (13) after applying a small transformation. Equation (13) facilitates processing of the next step.

$$\ell(\mathbf{Z},\mathbf{E},\mathbf{Q},\mathbf{Y}_1,\mathbf{Y}_2,\mu) = \min_{\mathbf{Z},\mathbf{E},\mathbf{Q}} \| \mathbf{Z} \|_*$$
$$+\lambda_1 tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + \lambda_2 \| \mathbf{E} \|_1 + \lambda_3 \| \mathbf{Q} \|_1 \quad (13)$$
$$+f(\mathbf{Z},\mathbf{E},\mathbf{Q},\mathbf{Y}_1,\mathbf{Y}_2,\mu) \ \ s.t. \ \ \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} \ge 0.$$

Hence, $f(\mathbf{Z},\mathbf{E},\mathbf{Q},\mathbf{Y}_1,\mathbf{Y}_2,\mu) = \mu(\| \mathbf{X} - \mathbf{X}\mathbf{Z} - \mathbf{E} + \mathbf{Y}_1/\mu \|_F^2 + \| \mathbf{Z} - \mathbf{Q} + \mathbf{Y}_2/\mu \|_F^2)/2$.

We divide equation (13) into three subproblems and solve it in three steps. The three subproblems are written as follows.

$$\ell_1 = \min_{\mathbf{Z}} \| \mathbf{Z} \|_* + \lambda_1 tr\left(\mathbf{Z}\mathbf{L}\mathbf{Z}^{\mathrm{T}}\right) + f(\mathbf{Z},\mathbf{E},\mathbf{Q},\mathbf{Y}_1,\mathbf{Y}_2,\mu) \quad (14)$$
$$s.t. \ \ \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} \ge 0$$

$$\ell_2 = \min_{\mathbf{E}} \lambda_2 \| \mathbf{E} \|_1 + \mu \| \mathbf{X} - \mathbf{XZ} - \mathbf{E} + \mathbf{Y}_1/\mu \|_F^2 /2 \quad (15)$$

$$\ell_3 = \min_{\mathbf{Q}} \lambda_3 \| \mathbf{Q} \|_1 + \mu \| \mathbf{Z} - \mathbf{Q} + \mathbf{Y}_2/\mu \|_F^2 /2 \quad (16)$$

Finally, we solve the above subproblems to find the optimal solution. The specific steps are given as follows.

Step 1. Update **Z**: The matrix **Z** can be obtained by solving subproblem ℓ_1 while keeping **E** and **Q** fixed. First, we define the following formula (17) based on ℓ_1.

$$\begin{aligned} &\ell_1^k\left(\mathbf{Z}_k, \mathbf{E}_k, \mathbf{Q}_k, \mathbf{Y}_1^k, \mathbf{Y}_2^k, \mu_k\right) \\ &= \lambda_1 tr\left(\mathbf{ZLZ}^{\mathrm{T}}\right) + f\left(\mathbf{Z}_k, \mathbf{E}_k, \mathbf{Q}_k, \mathbf{Y}_1^k, \mathbf{Y}_2^k, \mu_k\right). \end{aligned} \quad (17)$$

By setting the first derivative of ℓ_1^k with respect to $\mathbf{Z}_k$, we can obtain the following formula (18).

$$\begin{aligned} \frac{\partial \ell_1^k}{\partial \mathbf{Z}_k} &= \lambda_1\left(\mathbf{Z}_k\mathbf{L} + \mathbf{Z}_k\mathbf{L}^{\mathrm{T}}\right) + \mu_k\mathbf{X}^{\mathrm{T}}\left(\mathbf{XZ}_k - \mathbf{X} + \mathbf{E}_k - \mathbf{Y}_1^k/\mu_k\right) \\ &+ \mu_k\left(\mathbf{Z}_k - \mathbf{Q}_k + \mathbf{Y}_2^k/\mu_k\right). \end{aligned} \quad (18)$$

According to LADMAP, subproblem ℓ_1 can be replaced by solving the following problem (19).

$$\begin{aligned} &\min_{\mathbf{Z}} \| \mathbf{Z} \|_* + \frac{\partial \ell_1^k}{\partial \mathbf{Z}_k}, \mathbf{Z} - \mathbf{Z}_k + \frac{\eta_1}{2} \| \mathbf{Z} - \mathbf{Z}_k \|_F^2 \\ &s.t. \quad \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} \geq 0, \end{aligned} \quad (19)$$

where $\eta_1 = 2\lambda_1 \| \mathbf{L} \|_2 + \mu_k(1 + \| \mathbf{X} \|_2^2)$.

Equation (19) can be transformed into the following formula (20).

$$\begin{aligned} &\min_{\mathbf{Z}} \frac{1}{\eta_1} \| \mathbf{Z} \|_* + \frac{1}{2} \| \mathbf{Z} - \left(\mathbf{Z}_k - \frac{\partial \ell_1^k}{\partial \mathbf{Z}_k}/\eta_1\right) \|_F^2 \\ &s.t. \quad \mathbf{Z} = \mathbf{Z}^{\mathrm{T}}, \mathbf{Z} \geq 0. \end{aligned} \quad (20)$$

To solve the symmetric and non-negative constraints of low-rank matrix **Z**, we adopt ***Lemma 1*** of (Chen et al., 2017) and the non-negative operator, i.e., equation (24), respectively. ***Lemma 1*** is defined as follows, and the detailed proofs have been given in the literature (Chen et al., 2017).

Lemma 1: If there is an expression similar to equation (21), its closed solution is equation (22).

$$\arg\min_{\mathbf{G}} \frac{1}{\beta} \| \mathbf{G} \|_* + \frac{1}{2} \| \mathbf{G} - \mathbf{H} \|_F^2 \quad s.t. \quad \mathbf{G} = \mathbf{G}^{\mathrm{T}}, \quad (21)$$

$$\mathbf{G}^* = \mathbf{U}_r\left(\Sigma_r - \frac{1}{\beta}\mathbf{I}_r\right)\mathbf{V}_r^{\mathrm{T}}. \quad (22)$$

In this work, $\mathbf{U}_r$, Σ_r and $\mathbf{V}_r$ are the members of the skinny singular value decomposition (SVD) of the matrix $\tilde{\mathbf{G}} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$; $\Sigma_r = diag(\delta_1,\delta_2,\ldots,\delta_r)$; δ_r is the singular value for which the positive singular values are greater than $\frac{1}{\beta}$, i.e., $\{r : \delta_r > \frac{1}{\beta}\}$; $\tilde{\mathbf{G}}$ is defined as $\tilde{\mathbf{G}} = (\mathbf{H} + \mathbf{H}^{\mathrm{T}})/2$; and $\mathbf{I}_r$ is an identity matrix with size $r \times r$.

Based on ***Lemma 1***, we make $\tilde{\mathbf{Z}}_k = \frac{1}{2}[(\mathbf{Z}_k - \frac{\partial \ell_1^k}{\partial \mathbf{Z}_k}/\eta_1) + (\mathbf{Z}_k - \frac{\partial \ell_1^k}{\partial \mathbf{Z}_k}/\eta_1)^{\mathrm{T}}]$. We solve the $\mathbf{Z}_{k+1}$ using the singular value thresholding operator $\boldsymbol{\theta}_\epsilon(\mathbf{A}) = \mathbf{U}_r\mathbf{S}_\epsilon(\Sigma_r - \frac{1}{\eta_1}\mathbf{I}_r)\mathbf{V}_r^{\mathrm{T}}$, where $\boldsymbol{S}_\epsilon = \mathrm{sgn}(x)\max(|x|-\epsilon,0)$. The iterative formula is written as follows.

$$\mathbf{Z}_{k+1} = \boldsymbol{\theta}_{\frac{1}{\eta_1}}(\mathbf{Z}_k), \quad (23)$$

where $\eta_1 = 2\lambda_1 \| \mathbf{L} \|_2 + \mu_k(1 + \| \mathbf{X} \|_2^2)$. After obtaining matrix $\mathbf{Z}_{k+1}$ by equation (23), the non-negative constraint is imposed on matrix $\mathbf{Z}_{k+1}$ through a non-negative operator. The non-negative operator is defined as follows.

$$F\left(\mathbf{Z}_{k+1}^{*(i,j)}\right) = \begin{cases} \mathbf{Z}_{k+1}^{(i,j)}, & \mathbf{Z}_{k+1}^{(i,j)} > 0 \\ 0, & otherwise \end{cases}. \quad (24)$$

Finally, the non-negative symmetric low-rank matrix $\mathbf{Z}_{k+1}^*$ is obtained.

Step 2. Update **E**: The matrix **E** can be obtained by solving subproblem ℓ_2 while keeping **Z** and **Q** fixed. Analogously, following equation (18), the first derivative of ℓ_2 is set with respect to $\mathbf{E}_k$, i.e., $\frac{\partial \ell_2}{\partial \mathbf{E}_k}$, and set $\frac{\partial \ell_2}{\partial \mathbf{E}_k} = 0$. Thus, we obtain equation (25).

$$\begin{aligned} \frac{\partial \ell_2}{\partial \mathbf{E}_k} &= \mu_k\left(\mathbf{E}_k - \mathbf{X} + \mathbf{XZ}_{k+1} - \mathbf{Y}_1^k/\mu_k\right) = 0 \\ &\rightarrow \mathbf{E}_k = \mathbf{X} - \mathbf{XZ}_{k+1} + \mathbf{Y}_1^k/\mu_k. \end{aligned} \quad (25)$$

According to the NSHLRR method (Yin et al., 2016), the iterative formula of **E** is given as follows.

$$\mathbf{E}_{k+1} = \Psi_{\frac{\lambda_2}{\mu_k}}\left(\mathbf{X} - \mathbf{XZ}_{k+1} + \mathbf{Y}_1^k/\mu_k\right). \quad (26)$$

Step 3. Update **Q**: The matrix **Q** can be obtained by solving subproblem ℓ_3 while keeping **Z** and **E** fixed. Similar to **Step 2**, we set the first derivative of ℓ_3 with respect to $\mathbf{Q}_k$, i.e., $\frac{\partial \ell_3}{\partial \mathbf{Q}_k}$, and set $\frac{\partial \ell_3}{\partial \mathbf{Q}_k} = 0$. Thus, we obtain the following equation.

$$\begin{aligned} \frac{\partial \ell_3}{\partial \mathbf{Q}_k} &= \mu_k\left[\mathbf{Q}_k - \left(\mathbf{Z}_{k+1} + \mathbf{Y}_2^k/\mu_k\right)\right] = 0 \\ &\rightarrow \mathbf{Q}_k = \mathbf{Z}_{k+1} + \mathbf{Y}_2^k/\mu_k \end{aligned} \quad (27)$$

According to the NSHLRR method (Yin et al., 2016), the iterative formula of **Q** is written as follows.

$$\mathbf{Q}_{k+1} = \max\left\{\Psi_{\frac{\lambda_3}{\mu_k}}\left(\mathbf{Z}_{k+1} + \mathbf{Y}_2^k/\mu_k\right), 0\right\} \quad (28)$$

Algorithm 1 clearly summarizes the above solution steps. The initialization parameter values are set based on experimental experience and the existing relevant research recommendations (Yin et al., 2016).

NSLRG With Score Function

It is known that both local structure and global structure can influence the importance of features in raw data. However, the LS method primarily focuses on the locality preserving power of data to evaluate the features. Inspired by the lowest rank matrix

$\mathbf{Z}^*$ of the NSLRG method, which can capture the global and local structure of the raw data, we believe that the important feature of high-dimension data can be extracted based on the matrix $\mathbf{Z}^*$. Therefore, we propose a Score function that is established on the lowest rank matrix $\mathbf{Z}^*$ for selection of the important feature. The formula is defined as follows.

$$Score(r) = \frac{\sum_{ij}\left(x_{ri} - x_{rj}\right)^2 \mathbf{Z}_{ij-NSLRG}}{Var\left(x_{r,:}\right)},\ (1 \le r \le m, 1 \le i \le j \le n), \tag{29}$$

where the $\mathbf{Z}_{ij\text{-}NSLRG}$ is the element of $\mathbf{Z}^*$ obtained by the NSLRG method, and $\mathbf{Z}_{ij\text{-}NSLRG}$ denotes the similarity degree of the i-th and j-th samples and is used to measure the r-th feature between two samples. The property of the global and local structure captured by the lowest rank matrix can be used as a constraint for feature selection. The selected feature results are quite useful for capturing the subspace structures of raw data. In different classes, this constraint can guarantee the selected feature with high discrimination.

Based on the result of the Score function, all features are arranged in ascending order to form a score curve. The number of selected features is τ ($\tau < m$), which occurs before the first inflection point of the score curve. Thus, we cluster the cancer samples based on the selected feature genes.

We refer to the NSLRG method with the Score function as the NSLRG-S framework for short. In a nutshell, the NSLRG-S framework can be divided into four steps. In the first step, the lowest rank matrix is obtained by the NSLRG method. In the second step, the Score function is used to evaluate and rank features based on the lowest-rank matrix of the first steps. In the third step, the feature genes are selected according to the results of the Score function. In the fourth step, cancer sample clustering is processed based on the selected feature genes. This novel framework delivers better reliability in selection of the most important feature for cancer sample clustering according to the global and local structure of the raw data.

ALGORITHM 1 | The NSLRG method.

Input: data $\mathbf{X}$; parameters λ_1, λ_2 and λ_3; the number of k-nearest-neighbors.
Initialization: $\mathbf{Z}_0 = \mathbf{E}_0 = \mathbf{Q}_0 = \mathbf{Y}_1^0 = \mathbf{Y}_2^0 = 0$, ρ_0=2.5, μ_0=10^{-3}, $\mu_{\max}$=10^6, ϵ_1=10^{-6}, ϵ_2=10^{-2}, $\mathbf{L}$.
While not converged do
1. Update $\mathbf{Z}$ by **Step1**.
2. Update $\mathbf{E}$ by **Step2**.
3. Update $\mathbf{Q}$ by **Step3**.
4. Update Lagrangian multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$:
$\mathbf{Y}_1^{k+1} = \mathbf{Y}_1^k + \mu_k(\mathbf{X} - \mathbf{X}\mathbf{Z}_{k+1} - \mathbf{E}_{k+1})$
$\mathbf{Y}_2^{k+1} = \mathbf{Y}_2^k + \mu_k(\mathbf{Z}_{k+1} - \mathbf{Q}_{k+1})$
5. Update μ_{k+1}:
μ_{k+1}=min($\mu_{\max}$,$\rho_k\mu_k$),

where ρ_k = {

$$\rho_0, \quad if\ max\ \{\eta_1 \| \mathbf{Z}_{k+1} - \mathbf{Z}_k \|, \mu_k \| \mathbf{E}_{k+1} - \mathbf{E}_k \|, \quad \mu_k \| \mathbf{Q}_{k+1} - \mathbf{Q}_k \|\} \le \epsilon_2$$
$$1, \quad otherwise$$

Checking convergence:
if $\|\boldsymbol{X}-\boldsymbol{X}\boldsymbol{Z}_{k+1}-\boldsymbol{E}_{k+1}\|/\|\boldsymbol{X}\| < \epsilon_1$ or
max{ $\eta_1\|\boldsymbol{Z}_{k+1}-\boldsymbol{Z}_k\|, \mu_k\|\boldsymbol{E}_{k+1}-\boldsymbol{E}_k\|, \mu_k\|\boldsymbol{Q}_{k+1}-\boldsymbol{Q}_k\|$ } $< \epsilon_2$
End while
Output: The lowest rank matrix $\mathbf{Z}^*$.

Framework of NSLRG-S

Based on the proposed NSLRG-S framework, our goal is to model the gene expression data and cluster the cancer samples according to the selected feature genes.

The modelling process is shown in **Figure 2**. At the start, the matrix $\mathbf{X}_{m\times n}$ represents the gene expression data with size $m \times n$, and one row represents the expression level of a same gene in different samples. The totals of genes and samples are m and n, respectively. Usually, m is notably large and n is rather small. The matrix $\mathbf{Z}^*_{n\times n}$ is the lowest-rank matrix obtained by the NSLRG method as the basis for the Score function. Second, according to the score result, all of the genes are ranked in ascending order. The total number of τ ($\tau < m$) feature genes are selected based on the first inflection point of the score curve. Finally, we cluster the cancer samples based on the selected feature genes to demonstrate the selected genes with efficient discrimination. The result is compared with those of different methods, including the K-means, Graph Regularized Nonnegative Matrix Factorization (GNMF) (Cai et al., 2011), Robust Principal Component analysis (RPCA) (Candès et al., 2011), Sparse Principal Component Analysis (SPCA) (Journée et al., 2010), Graph-Laplacian PCA (GLPCA) (Jiang et al., 2013), LS (He et al., 2006), and LLRR (Wang et al., 2016) methods. The details of the experimental result are described in subsection *Experiments on Gene Expression Data*. **Algorithm 2** is the framework of the NSLRG-S for clustering of gene expression data.

EXPERIMENTS

To evaluate the performance of the NSLRG-S framework, we compare the NSLRG-S framework with multiple typical methods, including the K-means, GNMF (Cai et al., 2011), RPCA (Candès et al., 2011), SPCA (Journée et al., 2010), GLPCA (Jiang et al., 2013), LS (He et al., 2006), and LLRR (Wang et al., 2016) methods. In subsection *Evaluation and Quantitative Benchmarks*, we select three quantitative benchmarks to evaluate the experimental results. In subsection *Experiments on Synthetic Data* and subsection *Experiments on Gene Expression Data*, comparative experiments are conducted on synthetic data and cancer gene expression data, respectively.

ALGORITHM 2 | Framework of NSLRG-S for clustering gene expression data.

Input: Gene expression data $\mathbf{X}$ clustering number k
Step:
1) Learn a lowest rank matrix $\mathbf{Z}^*$ by the **Algorithm 1**;
2) Obtain the ranked feature genes by the Score-function;
3) Obtain the selected feature genes.
4) Obtain the clustering cancer samples results using the K-means method.
Output: Clustering results

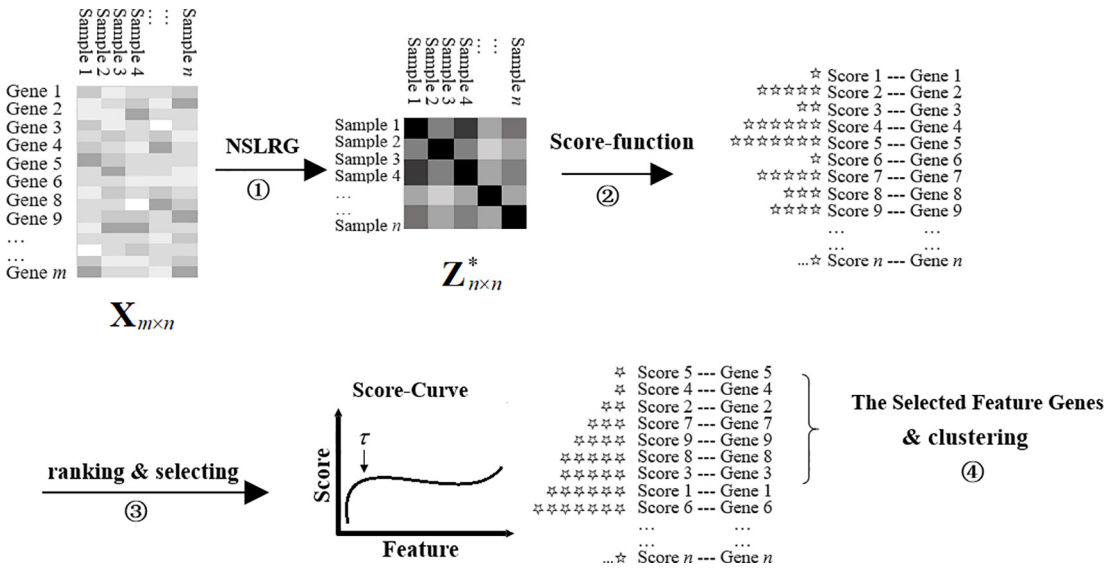


FIGURE 2 | Framework of NSLRG-S for clustering gene expression data.

Evaluation and Quantitative Benchmarks

To evaluate the performance of the clustering results based on comparison methods, we select three quantitative benchmarks: the clustering accuracy rate (Acc) (Cui et al., 2013), F1 measurement (F1) (Rijsbergen, 1979), and Rand Index (RI) (Rand, 1971).

Clustering Accuracy Rate

The Acc is defined as follows.

$$Acc = \frac{\sum_{i=1}^{N} \Xi(\xi_i, map(r_i))}{N} \times 100\,\% \tag{30}$$

where N is the total number of samples, and $\Xi(\xi_i, map(r_i))$ is used to identify whether ξ_i and r_i are matched. The ξ_i and r_i are the actual label and clustering label of the i-th sample, respectively, and if they are matched, the value of $\Xi(\xi_i, map(r_i))$ is equal to one; otherwise, its value is equal to zero. The $map(r_i)$ is the mapping function based on the Kuhn-Munkres method (Lovász and Plummer, 1986).

F1 Measurement

The F1 measurement is a special form of the *F-Measure* under a certain parameter. The *F-Measure* is also referred to as the *F-Score* and is the weighted harmonic mean of the *Precision* rate and *Recall* rate of the result of clustering. The *F-Measure*, *Precision* rate, and *Recall* rate are defined as follows.

$$F = \frac{\left(\phi^2 + 1\right) \times P \times R}{\phi^2 \times (P + R)}, \tag{31}$$

$$P = \frac{tp}{tp + fp}, \tag{32}$$

$$\mathrm{R} = \frac{tp}{tp + fn}, \tag{33}$$

where F is the *F-Measure*, P is the *Precision* rate and R is the *Recall* rate. The tp (true positives) is the item that records the number of positive samples that are clustered into their own positive class, fp (false positives) is the item that records the number of negative samples that are clustered into the positive class, and fn (false negatives) is the item that records the number of positive samples that are clustered into negative class. **Figure 3** clearly shows tp, fp and fn. The *F-Measure* can balance the contribution of fn by weighting *Recall* through the parameter $\phi > 0$. When the parameter $\phi = 1$, *F-Measure* becomes the most common form, i.e., F1 measurement, and equation (31) is rewritten as follows.

$$F1 = \frac{2 \times P \times R}{P + R}. \tag{34}$$

F1 measurement reaches its best value at 1 and its worst score at 0. The relative contributions of the *Precision* rate and *Recall* rate to the F1 measurement are equal.

Rand Index

The given data have two partitions: one is the actual classification, and the other is the clustered result (returned by our **Algorithm 2**). The Rand Index (RI) is used to compute how similar the result of clustering is to the actual classification. The RI is defined as follows.

$$RI = \frac{a + b}{C^2_{n_{samples}}}, \tag{35}$$

where a indicates the number of pairs of data points belonging to the same class in both the actual classification and the clustered result, b indicates the number of pairs of data points belonging to

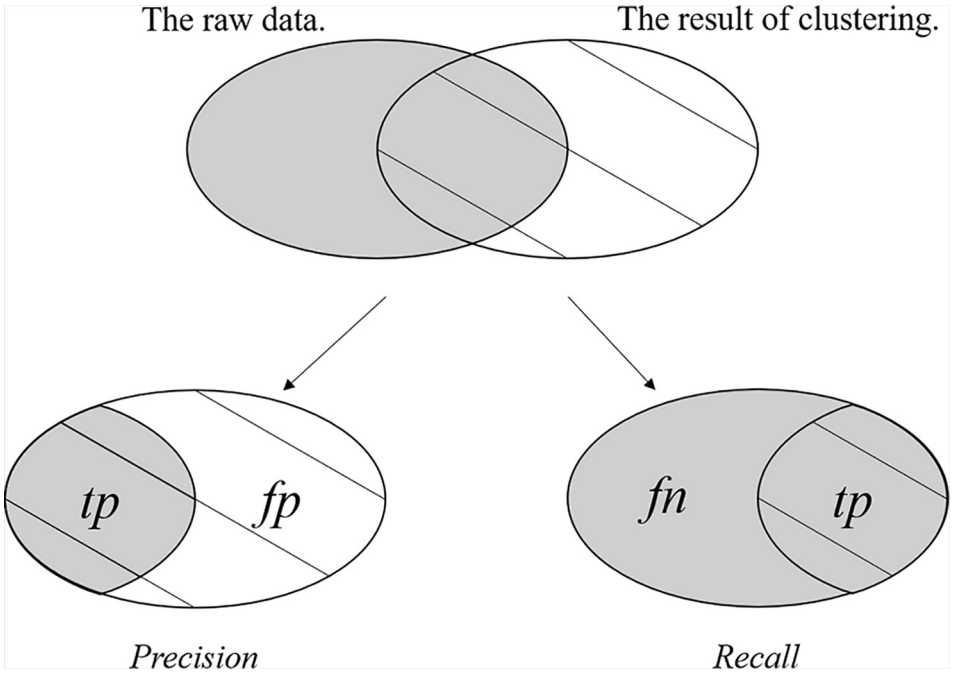


FIGURE 3 | The *tp*, *fp*, and *fn* of the clustering result.

the different class in both the actual classification and the clustered result, and $C^2_{n_{samples}}$ represents the total number of data pairs obtained from the given data. The range of RI is [0,1], and the larger the value, the more the clustering results are in accordance with reality.

Experiments on Synthetic Data

In this subsection, comparison experiments are conducted on synthetic data. In subsection *Synthetic Data*, we construct the synthetic data. In subsection *Convergence Analysis*, we perform convergence analysis to compare the NSLRG-S framework and other methods. In subsection *Clustering Results*, we analyze the performance of comparison methods on clustering data samples.

Synthetic Data

The synthetic data are constructed by the following steps (1) and (2). These synthetic data contain ten independent subspaces.

(1.) Construction of 10 original databases by $\boldsymbol{O}_{i+1} = \boldsymbol{TO}_i$, $1 \leq i \leq 9$. The value of the database ranges from 0 to 1, $\mathbf{T}$ is the transform random rotation matrix, and $\mathbf{O}_1$ is a random orthogonal matrix of 1000×100. The rank of each original database is 100.

(2.) We extract 10 data vectors from each original database by $\boldsymbol{X}_i = \boldsymbol{O}_i\boldsymbol{Q}_i, 1 \leq i \leq 10$, where the matrix $\mathbf{Q}_i^{100\times 10}$ is an independent identical distribution matrix $N(0,1)$, and its size is 100×10. All extracted data vectors are combined in synthetic data $\mathbf{X}^{1000\times 100}_{Synthetic\ \ data} = [\mathbf{X}_1,\ \mathbf{X}_2, \ldots, \mathbf{X}_{10}]$.

Convergence Analysis

We define an Error-Values function $\mathbf{F}_{E\text{-}V}(k)$ based on the loss function value to calculate the convergence rate. In the same iterations, the smaller the value of the Error-Values, the faster the convergence rate. The formula is given as follows.

$$\mathbf{F}_{E-V}(k) = \| \mathbf{X} - (\mathbf{X}\mathbf{Z}_k + \mathbf{E}_k) \|_F, \tag{36}$$

where the minimum value of $\mathbf{F}_{E\text{-}V}(k)$ is equal to zero. To clearly characterize the convergence rate, **Figures 4A, B** show the convergence trends of the NSLRG-S and the compared methods GNMF, RPCA, SPCA, and LLRR in 100 iterations. In **Figure 4B**, we find that the convergence rate of the NSLRG method is faster than those of the other methods.

Clustering Results

Table 1 lists the results of the GNMF, RPCA, SPCA, GLPCA, LS, LLRR, and NSLRG-S methods on the three quantitative benchmarks as Acc, F1, and RI. The results show that the performance of NSLRG-S is better than those of other methods.

Experiments on Gene Expression Data

In this subsection, we conduct experiments on gene expression datasets. The experimental datasets are downloaded from the famous gene expression database The Cancer Genome Atlas (TCGA). We cluster the cancer samples based on the feature genes obtained by the NSLRG-S framework. The experimental results demonstrate that we can improve the performance in cancer samples clustering by applying the selected feature genes.

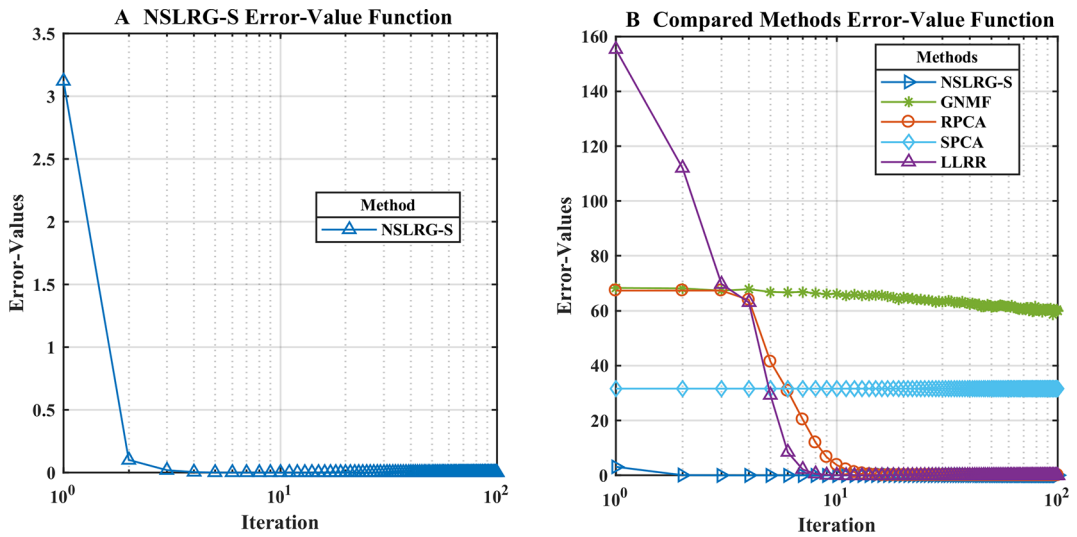


FIGURE 4 | (A and **B)**: The convergence analysis of different methods in 100 iterations.

TABLE 1 | The clustering results of compared methods and NSLRG-S method on synthetic data.

Method	Acc (%)	F1 (%)	RI (%)
GNMF	72.44	68.42	93.01
RPCA	80.68	78.82	95.57
SPCA	70.42	67.6	91.07
GLPCA	67.28	64.45	89.84
LS	80.62	78.37	96.12
LLRR	81.04	78.67	96.12
NSLRG-S	**82.00**	**79.21**	**96.27**

Acc, clustering accuracy rate; F1, F1 measurement; and RI, Rand Index; GNMF, Graph Regularized Nonnegative Matrix Factorization; SPCA, Sparse Principal Component Analysis; GLPCA, Graph-Laplacian PCA; LS, Laplacian Score; and LLRR, Laplacian regularized Low-Rank Representation; NSLRG-S, non-negative symmetric low-rank representation with graph regularization based on score function.
The bolded texts mean the results are better than the others.

Gene Expression Datasets

The TCGA database is a source of experimental data and is an important project for accelerating and comprehensively understanding cancer genetics using innovative genome analysis technologies (Tomczak et al., 2015). This database is one of the invaluable sources for gene expression datasets. Therefore, we select the TCGA database as the data source to research the clustering performance of the NSLRG-S framework.

We downloaded five cancer gene expression datasets, namely, esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD) and pancreatic adenocarcinoma (PAAD). Each type of gene expression dataset contains cancer tissue samples and normal tissue samples. There are 20,502 genes in each tissue sample. The distribution of each gene expression dataset is listed in **Table 2**.

In addition, to find the feature gene with a high recognition rate between different cancers for cancer sample clustering, we construct seven mixed datasets. The mixed datasets are HN-PA, ES-PA, CO-ES and HN-CH; HN-PA-CH, ES-PA-CH, and CO-PA-CH. The construction rule combines tumour tissue samples that come from different gene expression data, and the combined datasets contain two or three types of cancers. For example, in the HN-PA data, HN represents all of the cancer tissue samples of the HNSC data, and PA represents the total of the cancer tissue samples of the PAAD data. The cancer tissue samples of HN and PA are combined to construct the new mixed data, i.e., HN-PA, which contain two types of cancers and have 574 cancer tissue samples. For the other mixed datasets, the distributions are listed in **Table 3**.

The five original datasets and seven mixed datasets are used in experiments. We classify all datasets into three categories according to the number of cancers they contain. The datasets that contain one type of cancer belong to Category I. Thus, Category I contains PAAD, HNSC, ESCA, COAD, and CHOL. Datasets that contain two types of cancers belong to Category II, and they are HN-PA, ES-PA, CO-ES, and HN-CH. The datasets that contain three types of cancers belong to Category III, and the names of these datasets are HN-PA-CH, ES-PA-CH, and CO-PA-CH. **Table 4** clearly lists the category results.

Parameter Selection

In the experiments, we need to select the optimal parameters of the different datasets. For the three parameters (λ_1, λ_2, λ_3) of the

TABLE 2 | The distribution of five gene expression datasets.

Dataset	Cancer tissue samples	Normal tissue samples	Total samples	Total genes
PAAD	176	4	180	20502
HNSC	398	20	418	20502
ESCA	183	9	192	20502
COAD	262	19	281	20502
CHOL	36	9	45	20502

TABLE 3 | The distribution of mixed datasets.

Dataset	Cancer tissue and the number	Total number
HN-PA	398 from HNSC; 176 from PAAD;	574
ES-PA	183 from ESCA; 176 from PAAD;	359
CO-ES	262 from COAD; 183 from ESCA;	445
HN-CH	398 from HNSC; 36 from CHOL;	434
HN-PA-CH	398 from HNSC; 176 from PAAD; 36 from CHOL;	610
ES-PA-CH	183 from ESCA; 176 from PAAD; 36 from CHOL;	395
CO-PA-CH	262 from COAD; 176 from PAAD; 36 from CHOL;	474

ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; and PAAD, pancreatic adenocarcinoma.

NSLRG method, we assume that the optimal value of each parameter exists within an estimation range of $10^t(t = \{-5,-4,-3,-2,-1,0,1,2,3,4,5\})$. We study the influence of each parameter on feature selection and select the optimal parameters according to the different datasets. First, our main task is to determine the sensitivity of each parameter to the different datasets. We change one parameter within the candidate interval while holding the other two parameters fixed to explore the influence degree of this parameter on the dataset. We find that the parameter λ_3 is insensitive for all datasets. Therefore, the NSLRG method is robust for the parameter λ_3, and we select the $\lambda_3 = 10^{-3}$ according to experimental experience. The details of selection of the other two parameters are listed in **Table 5**.

Results and Discussion

In this subsection, based on the datasets of subsection *Gene Expression Datasets*, we apply the NSLRG-S to cluster the cancer samples. We adopt seven clustering methods, including K-means, GNMF, RPCA, SPCA, GLPCA, LS, and LLRR, for comparison with NSLRG-S.

Typically, gene expression data mining can be recognized as addressing a small sample size and high-dimensional problem. The applied methods must face and suffer from what is known as the curse of dimensionality. This situation occurs because the more dimensions contained in the data (20,502 in our case), the more unstable the result. Therefore, in our experiments, we improve the reasonableness of the result by running the experiment 50 times. The mean of the results is taken as the measurement of the clustering results.

Table 6 clearly lists the experimental results of all methods. Based on **Table 6**, we obtain the mean metrics of each category dataset, and they are listed in **Table 7**. Furthermore, to clearly show the experimental results on different categories of dataset

TABLE 4 | The category result of experimental datasets.

Category	I	II	III
Dataset	PAAD	HN-PA	HN-PA-CH
	HNSC	ES-PA	ES-PA-CH
	ESCA	CO-ES	CO-PA-CH
	COAD	HN-CH	/
	CHOL	/	/

ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; and PAAD, pancreatic adenocarcinoma.

TABLE 5 | The parameter selection.

Dataset	λ_1	λ_2	λ_3
PAAD	10^{-5}	10^{-2}	10^{-3}
HNSC	10^{-3}	10^{-4}	10^{-3}
ESCA	10^{4}	10^{-1}	10^{-3}
COAD	10^{4}	10^{0}	10^{-3}
CHOL	10^{-1}	10^{-1}	10^{-3}
HN-PA	10^{-4}	10^{1}	10^{-3}
ES-PA	10^{-2}	10^{-1}	10^{-3}
CO-ES	10^{2}	10^{5}	10^{-3}
HN-CH	10^{-1}	10^{5}	10^{-3}
HN-PA-CH	10^{-5}	10^{-2}	10^{-3}
ES-PA-CH	10^{-4}	10^{0}	10^{-3}
CO-PA-CH	10^{1}	10^{-2}	10^{-3}

ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; and PAAD, pancreatic adenocarcinoma.

and different methods, **Figure 5** presents a broken-line graph for the three category datasets corresponding to different methods. **Figure 6** presents a histogram for the different methods corresponding to the three category datasets.

By comparing the clustering results of NSLRG-S and other methods, we find that the results of the NSLRG-S method are the best of all methods in most datasets. According to **Table 6**, for the Category I dataset, the clustering performance of NSLRG-S for the HNSC and ESCA datasets is higher than that of other methods. In the COAD and CHOL dataset, NSLRG-S achieves the same best results as the other methods. For the Category II dataset, the clustering performance of NSLRG-S is the best of all methods. For the Category III dataset, except for the metrics of Acc and F1 on HN-PA-CH and Acc on CO-PA-CH, which are obtained by GNMF, and F1 on HN-PA-CH obtained by LLRR, the clustering performance of NSLRG-S is better than that of other methods.

In addition to the numerical comparison, we also find that the NSLRG-S method has different advantages after comparing it with different comparison methods. In the next section, we conduct a more detailed comparison and analysis between NSLRG-S and the other comparison methods.

In the seven comparison methods (K-means, GNMF, RPCA, SPCA, GLPCA, LS, and LLRR), K-means is the traditional clustering method; GNMF belongs to matrix factorization techniques, which extend the nonnegative matrix factorization with preservation of the intrinsic geometric structure (Cai et al., 2011); RPCA, SPCA, and GLPCA are variant methods of principal component analysis, which is a well-established descending dimension method for mining high dimensional data (Journée et al., 2010); LS is the feature selection method; and the LLRR is the subspace clustering method. In addition, the NSLRG-S framework combines the NSLRG method and Score function. Therefore, this framework belongs to a mixed method that combines the advantage of both sides.

First, we compare the NSLRG-S framework with K-means. Based on **Table 6**, we find that a higher clustering result is obtained by NSLRG-S. This comparison result shows that the proposed NSLRG-S framework is better than the traditional clustering method in cancer sample clustering. This result occurs because the NSLRG-S considers the local and global

TABLE 6 | The result of comparison experiment.

Category	Dataset	Measure	K-means	GNMF	RPCA	SPCA	GLPCA	LS	LLRR	NSLRG-S
I	PAAD	Acc	69.50%	74.67%	63.49%	56.47%	76.53%	**97.78%**	81.46%	97.22%
		F1	43.28%	46.69%	41.42%	40.31%	45.53%	**66.10%**	48.45%	49.30%
		RI	63.77%	61.96%	55.23%	50.58%	64.45%	**95.63%**	69.73%	94.57%
	HNSC	Acc	69.50%	81.72%	64.52%	62.20%	90.71%	93.54%	81.44%	**94.37%**
		F1	46.78%	44.97%	47.34%	46.59%	68.51%	48.33%	48.43%	**48.55%**
		RI	59.44%	70.05%	54.19%	52.86%	83.68%	87.89%	69.69%	**89.36%**
	ESCA	Acc	62.01%	54.69%	53.65%	53.97%	84.90%	94.79%	67.47%	**94.91%**
		F1	43.97%	40.00%	40.22%	41.15%	46.74%	48.66%	46.97%	**64.18%**
		RI	58.34%	50.18%	50.01%	50.06%	76.19%	90.07%	56.41%	**90.40%**
	COAD	Acc	74.71%	**99.29%**	86.39%	81.28%	84.42%	87.09%	88.20%	**99.29%**
		F1	60.02%	**97.31%**	71.08%	65.41%	68.68%	47.54%	73.40%	**97.31%**
		RI	65.22%	**98.58%**	76.45%	69.48%	73.60%	78.08%	79.15%	**98.58%**
	CHOL	Acc	85.72%	97.78%	**100.00%**	**100.00%**	**100.00%**	63.82%	**100.00%**	**100.00%**
		F1	66.16%	96.66%	**100.00%**	**100.00%**	**100.00%**	44.81%	**100.00%**	**100.00%**
		RI	75.03%	95.56%	**100.00%**	**100.00%**	**100.00%**	53.36%	**100.00%**	**100.00%**
II	HN-PA	Acc	97.66%	99.83%	99.48%	99.30%	98.95%	68.95%	99.65%	**100.00%**
		F1	95.99%	99.80%	99.39%	99.19%	98.78%	41.77%	99.59%	**100.00%**
		RI	96.38%	99.65%	98.96%	98.61%	97.93%	57.11%	99.30%	**100.00%**
	HN-CH	Acc	85.42%	98.39%	82.56%	89.59%	92.06%	90.12%	94.14%	**99.54%**
		F1	73.89%	94.18%	71.16%	77.83%	81.62%	47.40%	86.08%	**98.45%**
		RI	76.94%	96.82%	72.33%	81.36%	85.37%	82.15%	89.46%	**99.08%**
	ES-PA	Acc	96.41%	97.21%	98.25%	99.16%	99.16%	50.86%	99.16%	**99.72%**
		F1	73.89%	97.21%	97.95%	99.16%	99.16%	34.37%	99.16%	**99.72%**
		RI	95.44%	94.57%	97.37%	98.34%	98.34%	49.89%	98.34%	**99.44%**
	CO-ES	Acc	96.58%	80.67%	97.53%	96.85%	96.18%	59.10%	97.30%	**98.65%**
		F1	96.07%	77.59%	97.45%	96.75%	96.06%	37.65%	97.21%	**98.60%**
		RI	93.95%	68.75%	95.17%	93.89%	92.63%	51.55%	94.74%	**97.33%**
III	HN-PA-CH	Acc	81.01%	**92.79%**	77.20%	78.83%	80.13%	65.25%	87.71%	**88.62%**
		F1	62.79%	63.16%	61.82%	63.15%	65.25%	26.69%	**70.03%**	63.36%
		RI	84.14%	**94.79%**	81.99%	81.85%	81.76%	51.20%	87.74%	89.98%
	ES-PA-CH	Acc	81.14%	68.86%	73.91%	72.78%	72.52%	46.51%	86.03%	**89.37%**
		F1	65.98%	52.42%	63.41%	66.55%	66.13%	22.30%	69.23%	**72.11%**
		RI	86.29%	77.41%	82.73%	80.33%	80.29%	42.64%	85.98%	**90.58%**
	CO-PA-CH	Acc	80.24%	**89.45%**	74.04%	74.63%	75.40%	55.59%	85.57%	83.74%
		F1	68.56%	63.60%	61.77%	63.27%	64.27%	26.89%	70.44%	**73.56%**
		RI	84.22%	84.00%	82.27%	84.02%	83.65%	45.84%	84.53%	**85.52%**

ESCA, esophageal carcinoma; HNSC, head and neck squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; and PAAD, pancreatic adenocarcinoma. The bolded texts mean the results are better than the others.

TABLE 7 | The mean metrics of result for all methods on Category dataset I, II, III.

Metrics	Category	K-means	GNMF	RPCA	SPCA	GLPCA	LS	LLRR	NSLRG-S
ACC	I	72.29%	81.63%	73.61%	70.78%	87.31%	87.40%	83.71%	**97.16%**
	II	94.02%	94.03%	94.45%	96.23%	96.59%	67.26%	97.56%	**99.48%**
	III	80.80%	83.70%	75.05%	75.42%	76.02%	55.78%	86.44%	**87.24%**
F1	I	52.04%	65.13%	60.01%	58.69%	65.89%	51.09%	63.45%	**71.87%**
	II	84.96%	92.20%	91.49%	93.23%	93.91%	40.30%	95.51%	**99.19%**
	III	65.78%	59.73%	62.34%	64.32%	65.21%	25.29%	**69.90%**	69.67%
RI	I	64.36%	75.27%	67.18%	64.60%	79.58%	81.01%	75.00%	**94.58%**
	II	90.68%	89.95%	90.96%	93.05%	93.57%	60.17%	95.46%	**98.96%**
	III	84.88%	85.40%	82.33%	82.07%	81.90%	46.56%	86.08%	**88.70%**

Acc, clustering accuracy rate; F1, F1 measurement; and RI, Rand Index; GNMF, Graph Regularized Nonnegative Matrix Factorization; SPCA, Sparse Principal Component Analysis; GLPCA, Graph-Laplacian PCA; LS, Laplacian Score; and LLRR, Laplacian regularized Low-Rank Representation; NSLRG-S, non-negative symmetric low-rank representation with graph regularization based on score function.
The bolded texts mean the results are better than the others.

structure of the raw data. This framework can select feature genes with a high recognition rate for cancer sample clustering. In addition, the K-means method performs cancer sample clustering based on the raw data, which ignores the contents considered in NSLRG-S. **Figure 5** clearly shows that the NSLRG-S is superior to the K-means method.

Second, we compare the NSLRG-S with the GNMF method. In GNMF, a nearest neighbour graph is constructed by encoding the geometrical information of the data space. The method seeks matrix factorization, which incorporates the graph structure (Cai et al., 2011). Based on **Table 5**, the GNMF method obtains good results, and a subset of them are even better than those of

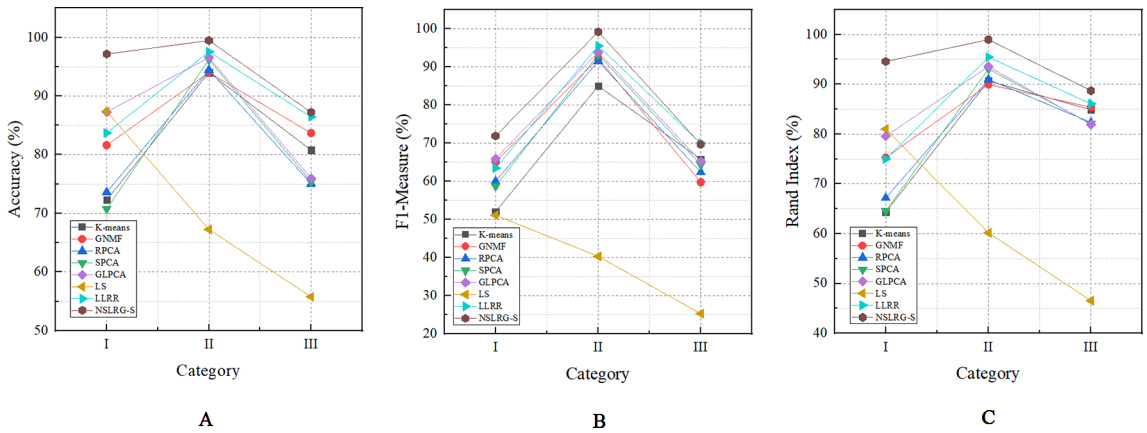


FIGURE 5 | The mean metrics of experimental result for Category I, II, and III. **(A)** Accuracy-Category **(B)** F1-Category **(C)** Rand Index-Category.

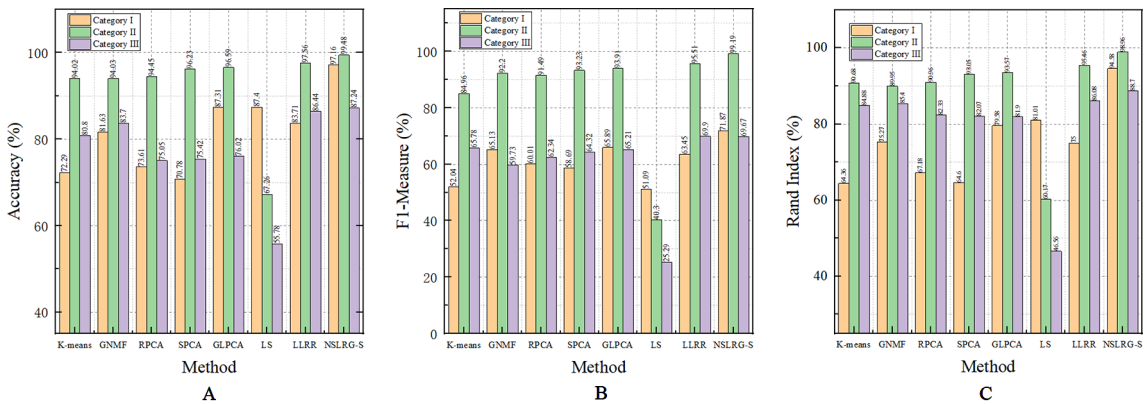


FIGURE 6 | The mean metrics of experimental result for all methods. **(A)** Accuracy-Method **(B)** F1-Method **(C)** Rand Index-Method.

NSLRG-S method. For most of the datasets, the results of NSLRG-S are still better than those of GNMF. The reason for this result is that the NSLRG-S method can obtain the characteristics of the subspace structure of the raw data, and the corresponding subspace of different types of cancer can be satisfactorily distinguished.

Third, we compare the NSLRG-S with the RPCA, SPCA, and GLPCA methods. RPCA, SPCA, and GLPCA belong to principal component analysis methods and are suitable for processing high-dimensional gene expression data by learning a low-dimensional representation. The results of NSLRG-S are better than those of three methods, except for the CHOL dataset. We can conclude that the NSLRG-S method is better than the variant methods of principal component analysis in clustering of multiple cancer samples.

Fourth, we compare the NSLRG-S with the LS method. Based on **Figure 5**, we find that the performance of LS decreases gradually on the Category I, Category II and Category III datasets, and this trend is different with other methods. The reason for this result is that the feature genes selected by the LS method have locality-preserving power attributes but do not have good multi-subspace separation attributes. In the framework of the NSLRG-S, feature genes are obtained under the Score function based on the low-rank matrix obtained by the NSLRG method. This low-rank matrix can preserve the global and local structure of the raw data, and after further processing the low-rank matrix through the Score function, the selected genes have a strong discrimination in multi-subspace clustering. Therefore, the performance of NSLRG-S is better than that of LS.

Finally, we compare the NSLRG-S with the LLRR method. Based on **Figure 5**, the broken line of the NSLRG-S is always above that of the LLRR method except for F1 on the Category III dataset. The comparison results show that the Score function plays an important role in further mining of the low-rank matrix of the NSLRG method.

Furthermore, we note an interesting trend in the results of three categories of datasets for each method, as shown in **Figure 6**. Other than the LS method, which shows a

downward trend, the other methods show an upward trend first followed by a downward trend. In other words, except for the LS method, after comparing all of the results of the other methods, we note that the experimental results of the Category II datasets are the best, followed by the Category III datasets or the Category I datasets, and this trend occurs in all metrics. According to **Tables 2**–**4**, the distributions of sample size in the Category II datasets are more balanced than those in Category I and Category III. Therefore, the result of the Category II dataset is more reasonable and stable than the results of Category I and Category III. However, with an increasing number of subspaces, the structure of the data is more complex, and the global and local structures of raw data are more difficult to capture. Therefore, compared with the experimental results of the Category II datasets, the experimental results of the Category III datasets decrease. Fortunately, according to **Table 7**, the NSLRG-S is still better than other methods. This observation demonstrates that the NSLRG-S framework has better advantages in cancer sample clustering than other methods when working with unbalanced and multi-subspace datasets. Based on the above discussion and analysis, we conclude that the NSLRG-S framework has a good effect for cancer sample clustering based on a gene expression dataset.

CONCLUSIONS WORK

In this paper, we cluster the cancer samples of multi-cancer gene expression datasets based on select feature genes obtained by the NSLRG-S framework. In addition, NSLRG-S simultaneously considers the local and global structure of the raw gene expression dataset. The selected feature genes have a high recognition rate in subspace clustering. The comparison experimental results suggest that the NSLRG-S framework can significantly improve the cancer samples clustering performance.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the [The Cancer Genome Atlas (TCGA)] https://cancergenome.nih.gov/. We have uploaded scripts and examples on GitHub to adhere standards for reproducibility. The URL is https://github.com/guoguoguolu/NSLRG-S-method-scripts-and-example-files.

AUTHOR CONTRIBUTIONS

JW and CL conceived the original research plans and methodology. JL and XK performed synthetic data analysis. JW, CL and XK performed experiments on gene expression data. JW and CL supervised and wrote the original draft. JW, CZ, and XZ reviewed and revised the writing.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61872220, 61702299, and 61873117.

ACKNOWLEDGMENTS

We thank to the contributions of CBC 2019 that aided the efforts of the authors.

REFERENCES

Belkin, M., and Niyogi, P. (2001). "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, British Columbia, Canada: MIT Press).

Cai, D., He, X., and Han, J. (2005). Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* 17 (12), 1624–1637. doi: 10.1109/TKDE.2005.198

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A Singular Value Thresholding Algorithm for Matrix Completion. *SIAM J. Optim.* 20 (4), 1956–1982. doi: 10.1137/080738970

Cai, D., He, X., Han, J., and Huang, T. S. (2011). Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (8), 1548–1560. doi: 10.1109/TPAMI.2010.231

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust Principal Component Analysis? *ACM* 58 (3), 1–37. doi: 10.1145/1970392.1970395

Chen, J., and Yang, J. (2014). Robust Subspace Segmentation Via Low-Rank Representation. *IEEE Trans. Cybern.* 44 (8), 1432–1445. doi: 10.1109/TCYB.2013.2286106

Chen, J., Mao, H., Sang, Y., and Yi, Z. (2017). Subspace clustering using a symmetric low-rank representation. *Knowl.-Based Syst.* 127, 46–57. doi: 10.1016/j.knosys.2017.02.031

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2018). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35 (4), 602–610. doi: 10.1093/bioinformatics/bty662%JBioinformatics

Cui, Y., Zheng, C. H., and Yang, J. (2013). Identifying subspace gene clusters from microarray data using low-rank representation. *PloS One* 8 (3), e59377. doi: 10.1371/journal.pone.0059377

Dash, M., and Liu, H. (1997). Feature selection for classification. *Intell. Data Anal.* 1 (1), 131–156. doi: 10.1016/S1088-467X(97)00008-5

Du, S., Ma, Y., and Ma, Y. (2017). Graph regularized compact low rank representation for subspace clustering. *Knowl.-Based Syst.* 118, 56–69. doi: 10.1016/j.knosys.2016.11.013

Ge, H., and Hu, T. (2014). "Genetic Algorithm for Feature Selection with Mutual Information," in *2014 Seventh International Symposium on Computational Intelligence and Design*. (Piscataway, NJ: IEEE), 116–119.

He, X., and Niyogi, P. (2005). In *Advances in neural information processing systems 16 (NIPS)*. (Cambridge, MA: MIT Press),153–160.

He, X., Cai, D., and Partha, N. (2006). "Laplacian Score for Feature Selection," in *the Neural Information Processing Systems Conference (NIPS)* (Cambridge, MA: MIT Press), 507–514.

Jiang, B., Ding, C., Luo, B., and Tang, J. (2013). "Graph-Laplacian PCA: Closed-Form Solution and Robustness," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Piscataway, NJ: IEEE).

Journée, M., Nesterov, Y., Richtárik, P., and Sepulchre, R. (2010). Generalized Power Method for Sparse Principal Component Analysis. *J. Mach. Learn. Res.* 11 (2), 517–553.

Kohavi, R., and John, G. H. (1997). Wrappers for feature subset selection. *Artif. Intell.* 97 (1), 273–324. doi: 10.1016/S0004-3702(97)00043-X

Langley, P. (1994). "Selection of relevant features in machine learning," in *Proc of the AAAI Fall Symposium on Relevance.* (Menlo Park, CA: AAAI), 1–5.

Lin, Z., Chen, M., and Ma, Y. (2009). *The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices* (University of illinois at urbana-champaign technical report). (UILU-ENG-09-2215).

Lin, Z., Chen, M., and Ma, Y. (2010). The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *Eprint Arxiv*. 2010 (v1). doi: 10.1016/j.jsb.2012.10.010

Lin, Z., Liu, R., and Su, Z. (2011). Linearized Alternating Direction Method with Adaptive Penalty for Low-Rank Representation. In *Advances in Neural Information Processing Systems (NIPS 2011).* (New York: Curran Associates), 612–620.

Liu, G., , Lin, Z., , and Yu, Y. (2010). "Robust Subspace Segmentation by Low-Rank Representation," in *Proceedings of the 27th International Conference on Machine Learning* (Madison, Wisconsin, USA: Omnipress), 663–670.

Liu, G. C., Lin, Z. C., Yan, S. C., Sun, J., Yu,Y. , and Ma, Y. (2013). Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1), 171–184. doi: 10.1109/TPAMI.2012.88

Lovász, L., and Plummer, M. D. (1986). Matching Theory. *J. Appl. Math. Mech.* 68 (3), 146–146. doi: 10.1002/zamm.19880680310

Lu, X., Wang, Y., and Yuan, Y. (2013). Graph-Regularized Low-Rank Representation for Destriping of Hyperspectral Images. *IEEE Trans. Geosci. Remote Sens.* 51 (7), 4009–4018. doi: 10.1109/TGRS.2012.2226730

Mohamad, M. S., Omatu, S., Deris, S., and Yoshioka, M. (2010). "A Three-Stage Method to Select Informative Genes from Gene Expression Data in Classifying Cancer Classes," in *2010 International Conference on Intelligent Systems, Modelling and Simulation.* (Piscataway, NJ: IEEE), 158–163.

Mohamad, M. S., Omatu, S., Deris, S., and Yoshioka, M. (2013). "A Constraint and Rule in an Enhancement of Binary Particle Swarm Optimization to Select Informative Genes for Cancer Classification," in *Revised Selected Papers of PAKDD 2013 International Workshops on Trends and Applications in Knowledge Discovery and Data Mining - Volume 7867*. (Berlin, Heidelberg: Springer), 168–178.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* 66 (336), 846–850. doi: 10.1080/01621459.1971.10482356

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 46 (20), 10546–10562. doi: 10.1093/nar/gky889

Rijsbergen, C. J. V. (1979). *Information Retrieval* (London: Butterworths).

Russo, G., Zegar, C., and Giordano, A. (2003). Advantages and limitations of microarray technology in human cancer. *Oncogene* 22, 6497–6507. doi: 10.1038/sj.onc.1206865

Talavera, L. (2005). *An Evaluation of Filter and Wrapper Methods for Feature Selection in Categorical Clustering* (Berlin Heidelberg: Springer), 440–451.

Tang, P., Tang, X., Tao, Z., and Li, J. (2014). "Research on feature selection algorithm based on mutual information and genetic algorithm," in *2014 11th International Computer Conference on Wavelet Actiev Media Technology and Information Processing (ICCWAMTIP).* (Piscataway, NJ: IEEE), 403–406.

Toh, K., and Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pac. J. Optim.* 6 (3), 615–640.

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Poznan Poland)* 19 (1A), A68–A77. doi: 10.5114/wo.2014.47136

Wang, Y. X., Liu, J. X., Gao, Y. L., Zheng, C. H., and Shang, J. L. (2016). Differentially expressed genes selection via Laplacian regularized low-rank representation method. *Comput. Biol. Chem.* 65, 185–192. doi: 10.1016/j.compbiolchem.2016.09.014

Wang, J., Liu, J. X., Zheng, C. H., Wang, Y. X., Kong, X. Z., and Weng, C. G. (2018). A mixed-norm laplacian regularized low-rank representation method for tumor samples clustering. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16 (1), 172–182. doi: 10.1109/TCBB.2017.2769647

Wang, J., Liu, J.-X., Kong, X.-Z., Yuan, S.-S., and Dai, L.-Y. (2019a). Laplacian regularized low-rank representation for cancer samples clustering. *Comput. Biol. Chem.* 78, 504–509. doi: 10.1016/j.compbiolchem.2018.11.003

Wang, J., Zheng, R., Liang, Z., Li, M., Wu, F.-X., and Pan, Y. (2019b). SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics* 35 (19), 3642–3650. doi: 10.1093/bioinformatics/btz139

Xu, A., Chen, J., Peng, H., Han, G., and Cai, H. (2019). Simultaneous interrogation of cancer omics to identify subtypes with significant clinical differences. *Front. Genet.* 10, 236. doi: 10.3389/fgene.2019.00236

Yin, M., Gao, J., and Lin, Z. (2016). Laplacian Regularized Low-Rank Representation and Its Applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (3), 504–517. doi: 10.1109/TPAMI.2015.2462360

You, C.-Z., Wu, X.-J., Palade, V., and Altahhan, A. (2016). "Manifold locality constrained low-rank representation and its applications," in *2016 International Joint Conference on Neural Networks (IJCNN).* (Piscataway, NJ: IEEE), 3264–3271.

Zhuang, L., Gao, H., Lin, Z., Ma, Y., Zhang, X., and Yu, N. (2012). "Non-negative low rank and sparse graph for semi-supervised learning," in *2012 IEEE Conference on Computer Vision and Pattern Recognition.* (Piscataway, NJ: IEEE), 2328–2335.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 January 2020
doi: 10.3389/fgene.2019.01328

Analysis of the Differentially Expressed Genes Induced by Cisplatin Resistance in Oral Squamous Cell Carcinomas and Their Interaction

Hua-Tao Wu[1†], *Wen-Tian Chen*[2†], *Guan-Wu Li*[3], *Jia-Xin Shen*[4], *Qian-Qian Ye*[2,5], *Man-Li Zhang*[5], *Wen-Jia Chen*[2,5] *and Jing Liu*[2,5*]

[1] Department of General Surgery, The First Affiliated Hospital of Shantou University Medical College, Shantou, China, [2] Department of Physiology, Shantou University Medical College, Shantou, China, [3] Open Laboratory for Tumor Molecular Biology, Department of Biochemistry, The Key Lab of Molecular Biology for High Cancer Incidence Coastal Chaoshan Area, Shantou University Medical College, Shantou, China, [4] Department of Hematology, The First Affiliated Hospital of Shantou University Medical College, Shantou, China, [5] Chang Jiang Scholar's Laboratory/Guangdong Provincial Key Laboratory for Diagnosis and Treatment of Breast Cancer, Shantou University Medical College, Shantou, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Kevin Tak-Pan Ng,
The University of Hong Kong, Hong Kong
Ningxia Zhu,
Baylor College of Medicine, United States

***Correspondence:**
Jing Liu
jliu12@stu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *14 October 2019*
Accepted: *05 December 2019*
Published: *23 January 2020*

Citation:
Wu H-T, Chen W-T, Li G-W, Shen J-X, Ye Q-Q, Zhang M-L, Chen W-J and Liu J (2020) Analysis of the Differentially Expressed Genes Induced by Cisplatin Resistance in Oral Squamous Cell Carcinomas and Their Interaction. Front. Genet. 10:1328. doi: 10.3389/fgene.2019.01328

Background: Oral squamous cell carcinoma (OSCC) is a solid tumor, which originates from squamous epithelium, with about 400,000 new-cases/year worldwidely. Presently, chemoradiotherapy is the most important adjuvant treatment for OSCC, mostly in advanced tumors. However, clinical resistance to chemotherapy still leads to poor prognosis of OSCC patients. Via high-throughput analysis of gene expression database of OSCC, we investigated the molecular mechanisms underlying cisplatin resistance in OSCC, analyzing the differentially expressed genes (DEGs) and their regulatory relationship, to clarify the molecular basis of OSCC chemotherapy resistance and provide a theoretical foundation for the treatment of patients with OSCC and individualized therapeutic targets accurately.

Methods: Datasets related to "OSCC" and "cisplatin resistance" (GSE111585 and GSE115119) were downloaded from the GEO database and analyzed by GEO2R. Venn diagram was used to obtain drug-resistance-related DEGs. Functional enrichment analysis and Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis were performed on DEGs using The Database for Annotation, Visualization and Integrated Discovery (DAVID) software. Protein–protein interaction (PPI) network was constructed by STRING (search tool for recurring instances of neighbouring genes) database. Potential target genes of miRNA were predicted *via* miRDB, and cBioportal was used to analyze the function and survival of the potential functional genes.

Results: Forty-eight upregulated DEGs and 49 downregulated DEGs were obtained from the datasets, with cutoff as $p < 0.01$ and $|\log FC| > 1$. The DEGs in OSCC mainly enriched in cell proliferation regulation, and chemokine activity. In PPI network with hub score > 300, the hub genes were identified as *NOTCH1*, *JUN*, *CTNNB1*, *CEBPA*, and *ETS1*.

Among miRNA–mRNA targeting regulatory network, hsa-mir-200c-3p, hsa-mir-200b-3p, hsa-mir-429, and hsa-mir-139-5p were found to simultaneously regulate multiple hub genes. Survival analysis showed that patients with high *CTNNB*1 or low *CEBPA* expression had poor outcome.

Conclusions: In the OSCC cisplatin-resistant cell lines, *NOTCH1*, *JUN*, *CTNNB1*, *CEBPA*, and *ETS1* were found as the hub genes involved in regulating the cisplatin resistance of OSCC. Members of the miR-200 family may reverse drug resistance of OSCC cells by regulating the hub genes, which can act as potential targets for the treatment of OSCC patients with cisplatin resistance.

Keywords: differentially expressed genes, resistance, oral squamous cell carcinomas, cisplatin, miRNA

INTRODUCTION

Head and neck squamous cell carcinoma (HNSCC), the sixth most common malignant tumor in the world (Kim et al., 2011a), is an important public health issue worldwide. Among the total HNSCC cases, 30% are oral squamous cell carcinoma (OSCC) cases (World Health, 2003; Petersen, 2003a; Petersen, 2003b). In 2012, about 145,000 patients with OSCC died worldwide, with a mortality rate of 1.8% (Petersen, 2005; Kim et al., 2011a; Ong et al., 2016). Interestingly, OSCC is one of the three most common malignancies in Central and South Asia. In India, the age-standardized incidence of OSCC is 12.6 per 100,000 people (Petersen, 2005). According to statistics, the incidence of OSCC has increased sharply in several countries and regions, including Denmark, France, Germany, Scotland, and Central and Eastern Europe (Petersen, 2005).

OSCC can occur in different areas of the mouth and tongue, including lips, alveolar ridge, oral floor, oral tongue, hard palate, posterior molars triangle, and buccal mucosa, lined by squamous epithelium and scattered in smaller salivary glands and lymphatic drainage pathways. OSCC is common in the elder people with a history of tobacco and alcohol usage, with malignant tumors or somatic cell mutation by inducing DNA damage (Leemans et al., 2011). Although surgery is the main treatment strategy for OSCC, chemoradiotherapy is also an effective method, especially for advanced tumors. However, drug resistance due to unraveled molecular mechanisms significantly reduces the survival of OSCC patients.

Since the first miRNA— lin-4 was identified in 1993, miRNAs have attracted the attention of researchers in the field of gene expression regulation and gene therapy (Liang et al., 2014). By inhibition of RNA translation or degradation of target mRNA, miRNAs act as negative gene regulators at the post-transcriptional level (Sakai et al., 2013). Importantly, miRNAs can simultaneously modulate many target genes, such as tumor suppressors or oncogenes, widely influencing the phenotype of malignant tumors. Since miRNAs have been found to have important role in various aspects of malignant tumors, including oncogenesis, proliferation, metastasis, multidrug resistance, self-renewal, and differentiation of malignant stem cells (Wu et al., 2014), they may represent a new set of therapeutic target biomarkers for finding multidrug resistance in malignant tumors (Hong et al., 2013).

In this study, the potential molecular mechanisms of cisplatin resistance of OSCC were studied by using high-throughput gene expression database. The differentially expressed genes (DEGs) in OSCC and their regulatory relationships were analyzed, in order to elucidate the molecular basis of OSCC chemotherapy resistance, and to provide theoretical basis and individualized precise therapeutic targets for the treatment of OSCC patients.

MATERIALS AND METHODS

Microarray Datasets

"OSCC" and "cisplatin resistance" were used as the keywords for searching the GEO database, and GSE111585 and GSE115119 were downloaded as the gene expression data sets for cisplatin resistance in OSCC; the platforms used were GPL14715 and GPL16955.

GSE111585 included six samples of SCC9 cells and was divided into normal group and drug resistance group (Lin et al., 2018). GSE115119 contained four samples of CAL-27 cells, with normal group and drug resistant group. Both SCC9 and CAL-27 are human OSCC cell lines.

Data Analysis and Differential Expressed Gene Acquisition

Limma package of R software (GEO2R) was used for analysis of the original datasets. $|\log FC| > 1$ and p value < 0.01 were defined as the cutoff values for further analysis of DEGs. Volcano maps were constructed by SangerBox software.

Furthermore, the list of oncogenes (http://ongene.bioinfo-minzhao.org/) (Liu et al., 2017) and tumor-suppressor genes (https://bioinfo.uth.edu/TSGene/index.html) (Zhao et al., 2016) provided potential functional roles of genes in cancer process. To obtain DEGs in cisplatin-resistant OSCC cells, Venn package (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to draw the intersection of the up-regulated or down-regulated genes in the datasets with oncogenes or tumor-suppressor genes, respectively.

Functional Enrichment Analysis of DEGs

Gene Ontology (GO) provides a computational model of biological systems, from the molecular to the organism level, across different species in the following three categories: biological process (BP), molecular function (MF), and cellular component (CC) (Thomas, 2017). Kyoto encyclopedia of genes and genomes (KEGG) is a database for high-level functions and utilities of the biological systems, based on molecular-level information of genome sequencing and other high-throughput experimental technologies (Kanehisa et al., 2017). DAVID Bioinformatics Resources 6.8 (https://david.ncifcrf.gov/) comprises a comprehensive set of functional annotation tools for functional enrichment analysis of gene groups (Huang da et al., 2009). To identify the biological significance of DEGs in cisplatin-resistant OSCC cells, DAVID 6.8 was used to analyze GO function and KEGG pathway enrichment, with the enrichment standard as $p < 0.05$.

Protein–Protein Interaction Network of DEGs

Protein–protein interaction (PPI) network analysis is helpful to investigate the molecular mechanisms of diseases and discover new drug targets from a systematic perspective. STRING 11.0 (https://string-db.org/), covering more than 5,000 organisms with known and predicted protein–protein interactions, provides direct (physical) and indirect (functional) association (Szklarczyk et al., 2019). The PPI analysis of DEGs was performed by STRING 11.0, and the results were analyzed by Cytoscape 3.7.1. Furthermore, the cytoHubba plug in was used to calculate the interaction coefficient score between the DEGs. The top genes with hub score > 300 were identified as the hub genes with high connectivity in the PPI network.

Predicting Hub Gene-Related miRNAs

MicroRNAs (miRNAs), small non-coding RNA molecules with highly conserved regions, regulate the expression of target genes by binding to the 3'-untranslated regions (3'-UTR) of specific mRNAs, involved in many physiological and disease processes. Each miRNA is thought to regulate multiple genes with enormous potential regulatory circuitry afforded by miRNA (Lim et al., 2003). To identify the potential miRNA–mRNA interaction in the network of the hub genes, miRDB (http://mirdb.org/), an online resource for miRNA target prediction and functional annotation (Wong and Wang, 2015), was used to predict the hub gene-related miRNAs, and the miRNA–mRNA regulatory network was constructed by Cytoscape 3.7.1.

Expression and Survival Analysis of Hub Genes

The Oncomine database (https://www.oncomine.org/resource/login.html), an online cancer microarray database-mining platform (Rhodes et al., 2004), was used to investigate the difference in transcriptional levels of the hub genes in HNSCC and normal tissues.

As mutations of oncogenes and/or tumor-suppressor genes are frequent in tumor tissues, the Human Protein Atlas (http://www.proteinatlas.org/) was analyzed for the prognostic values of the hub genes (Uhlen et al., 2017), and cBioportal database (http://www.cbioportal.org/), an open-access online resource for multi-dimension analysis of data from The Cancer Genome Atlas (TCGA) (Gao et al., 2013), was used to analyze the effects of mutations in hub genes on the survival of patients with OSCC (MD Anderson, Cancer Discov, 2013).

RESULTS

Difference of Gene Expression Between Parental and Cisplatin-Resistant OSCC Cells

The gene expression microarray datasets, GSE111585 and GSE115119 were downloaded from GEO datasets with paired parental and cisplatin-resistant OSCC cells. As shown in **Figure 1**, the expression of most genes in cisplatin-resistant OSCC cells was similar to that of the parental OSCC cells. Cluster analysis by R software ($|\log FC| > 1$ and p value < 0.01 as the cutoff) revealed 1,386 up-regulated genes and 643 down-regulated genes in cisplatin-resistant OSCC cells compared with parental OSCC cells in GSE111585 (**Figure 1A**), and 757 up-regulated genes and 625 down-regulated genes in cisplatin-resistant OSCC cells compared with parental OSCC cells in GSE115119 (**Figure 1B**).

The intersection between DEGs and the list of oncogenes drawn by Venn software showed 48 up-regulated DEGs (**Figure 1C**), and 49 down-regulated DEGs were obtained *via* intersection between down-regulated genes and the list of tumor-suppressor genes (**Figure 1D**).

Close Association of the DEGs With the Regulation of Transcription and microRNAs in Cancers

Using the DAVID analysis software, functional enrichment analyses (BP, MF, and CC) of the DEGs were done. BP enrichment showed that the up-regulated DEGs were mainly enriched in cell proliferation regulation, inflammatory reaction, lipopolysaccharide, cells in response to growth factors to stimulate, neuronal migration, transmembrane receptor protein tyrosine kinase signaling pathway, and transcription of RNA polymerase II promoter (**Figure 2A**), whereas down-regulated DEGs were significantly enriched mainly in the following GO terms: response to X-ray, RNA polymerase II promoter negative transcription regulation, and DNA damage response (**Figure 2B**).

MF enrichment showed that the up-regulated DEGs were significantly enriched in chemokine activity, transcription factor activity, sequence specific DNA binding, non-membrane crossing protein tyrosine kinase activity, and sequence specific DNA binding (**Figure 2C**), and the down-regulated DEGs were enriched in p53 binding, sequence specific DNA binding, transcriptional activator activity, and RNA polymerase II hub promoter proximal region sequence specific binding (**Figure 2D**).

CC analysis predicted close association between the up-regulated DEGs and the following GO terms: mRNA cutting, polyadenylation specific factor complex, extracellular space,

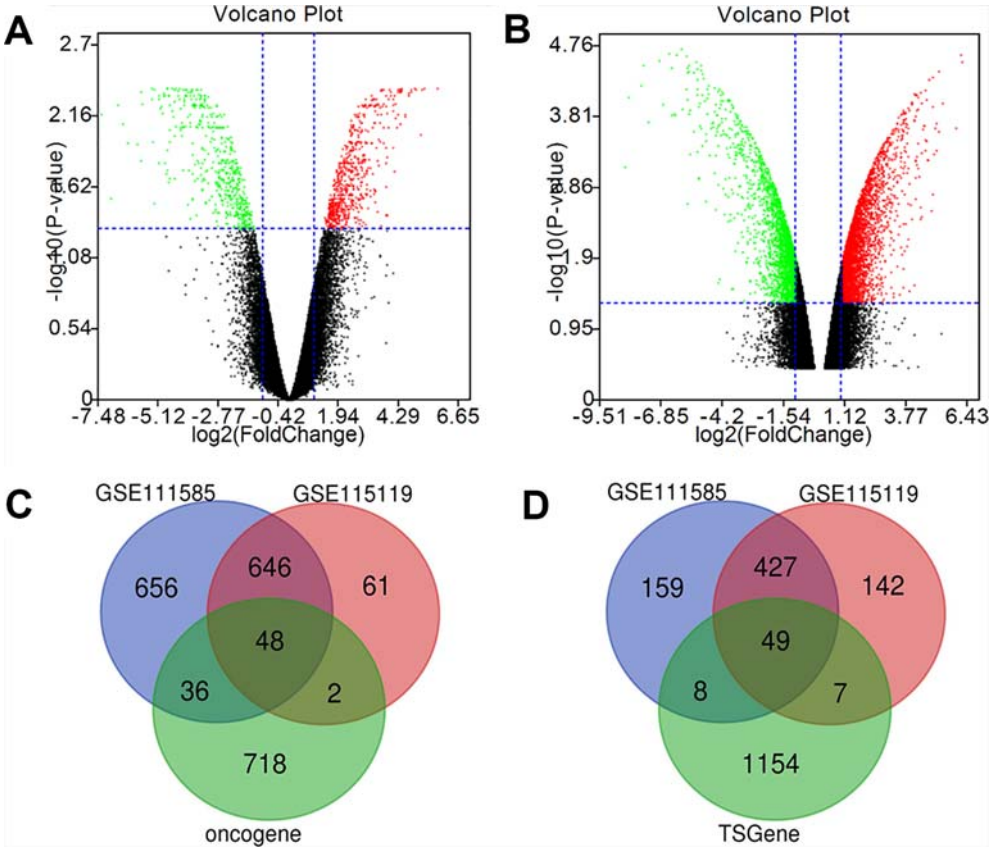


FIGURE 1 | Identification of cisplatin-resistant DEGs in OSCC. **(A)** Volcano map of GSE111585. **(B)** Volcano map of GSE115119. **(C)** 48 up-regulated DEGs was selected based on the intersection between up-regulated gene in GSE111585/GSE115119 and oncogenes. **(D)** Forty-nine down-regulated DEGs was selected based on the intersection between down-regulated gene in GSE111585/GSE115119 and tumor-suppressor genes.

promyelocytic leukemia proteome, and transcription factor complex (**Figure 2E**), and significant relation was found between the down-regulated DEGs and the following GO terms: junction complex, desmosomes, ciliated tips, cytoplasm, nuclear cytoplasm, and plasma membrane (**Figure 2F**).

KEGG pathway analysis provided the potential function cluster of DEGs, showing that the up-regulated DEGs were clustered in malaria, human T-cell leukemia virus type I, the way of malignant tumor, legionella infection disease, TNF signaling pathways, and T-cell receptors signaling pathways (**Figure 3A**), whereas the down-regulated DEGs were significantly concentrated in axon guidance and microRNAs in cancers (**Figure 3B**).

Identification of Hub Genes Through PPI Network of DEGs

To further analyze the correlation between DEGs in cisplatin-resistant OSCC cells, STRING was used to construct PPI network showing close relationship between the DEGs (**Supplemental Figure 1**), and their hub score was calculated. The genes with high hub score were predicted to have a strong association with other genes (shown in dark color in the figures). As shown in **Figure 4**, based on the cutoff hub score > 300, the following five genes were selected as the hub genes: *NOTCH1*, *JUN*, *CTNNB1*, *CEBPA*, and *ETS1*.

Construction of miRNA–mRNA Network Based on Predicting miRNA-Target Genes

As the DEGs in cisplatin-resistant OSCC cells were closely related to tumor-related miRNA, miRDB database was used to predict potential miRNAs that might participate in the transcriptional regulation of the hub genes in this process. The prediction scores were also collected from the miRDB database, and the miRNA–mRNA with high score meant close potential function of miRNA in regulation of the target mRNA. After setting cutoff > 80, Cytoscape software was used to construct the miRNA–mRNA network (**Figure 5**). Interestingly, hsa-miR-200c-3p, hsa-miR-200b-3p, hsa-miR-429, and hsa-miR-139-5p could simultaneously regulate multiple hub genes, which may be the key miRNAs involved in this process. Interestingly, hsa-miR-200c-3p, hsa-miR-200b-3p, and hsa-miR-429 belong to miR-200 family members, with similar functions; suppression of ZEB1/2, followed by inhibition of epithelial–mesenchymal transition (EMT).

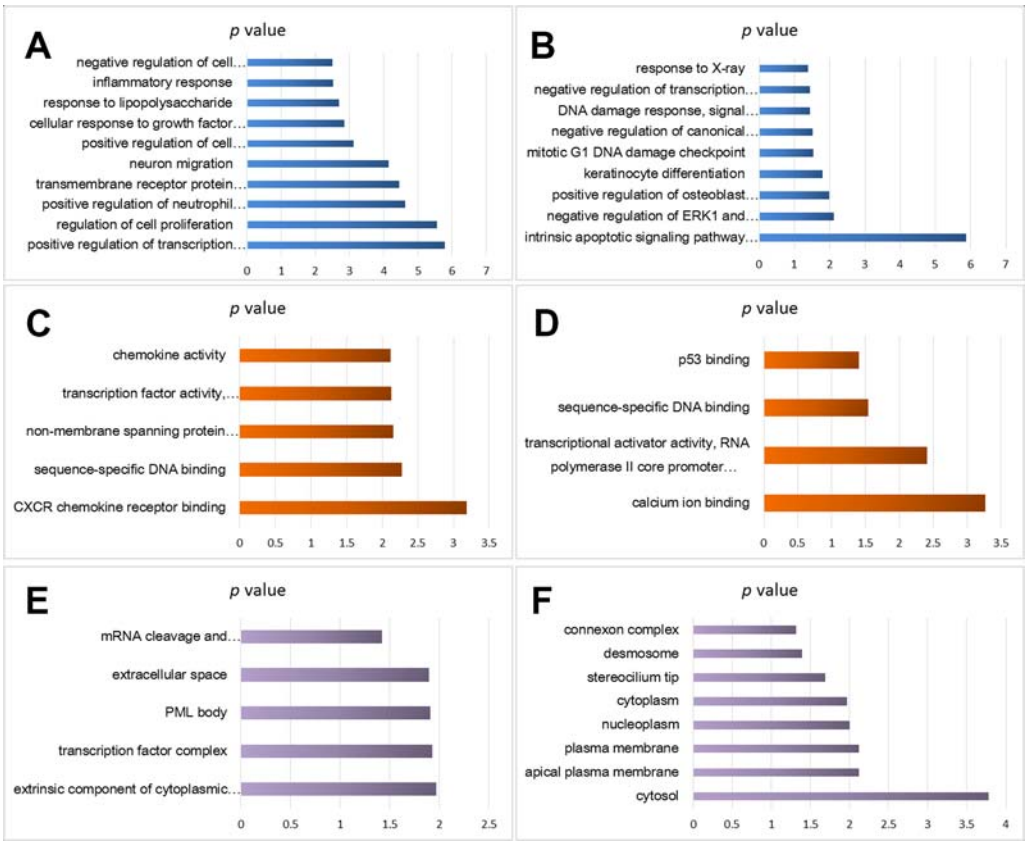


FIGURE 2 | Functional enrichment analysis of cisplatin-resistant DEGs in OSCC. **(A)** BP analysis of up-regulated DEGs. **(B)** BP analysis of down-regulated DEGs. **(C)** MF analysis of up-regulated DEGs. **(D)** MF analysis of down-regulated DEGs. **(E)** CC analysis of up-regulated DEGs. **(F)** CC analysis of down-regulated DEGs.

The Expression Pattern of Hub Genes in OSCC

To investigate the potential function of the hub genes in OSCC, Oncomine database was used to analyze the difference in the expression levels of the hub genes. However, due to limited research on OSCC, only one study revealed that the expression of *CTNNB1* and *ETS1* in tumor tissues was higher than that in normal tissues, with 2.285 and 2.111 fold change, respectively, while no difference was found in the expression of *NOTCH1* and *JUN* genes in the two tissues (**Figure 6**).

Survival Value of Hub Genes in OSCC

For survival analysis, cBioportal based on TCGA database was used, which revealed that low expression of *CTNNB1* in patients with OSCC showed better overall survival ($p = 0.01$) (**Figure 7C**), and low expression of *CEBPA* predicted poor overall survival in

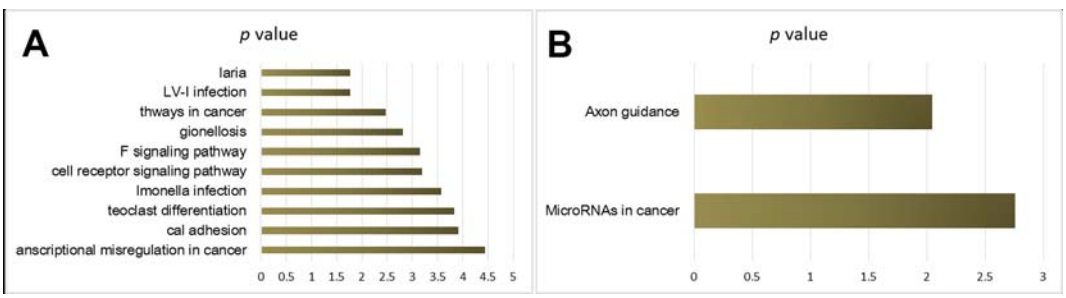


FIGURE 3 | KEGG pathway analysis of cisplatin-resistant DEGs in OSCC. **(A)** KEGG of up-regulated DEGs. **(B)** KEGG of down-regulated DEGs.

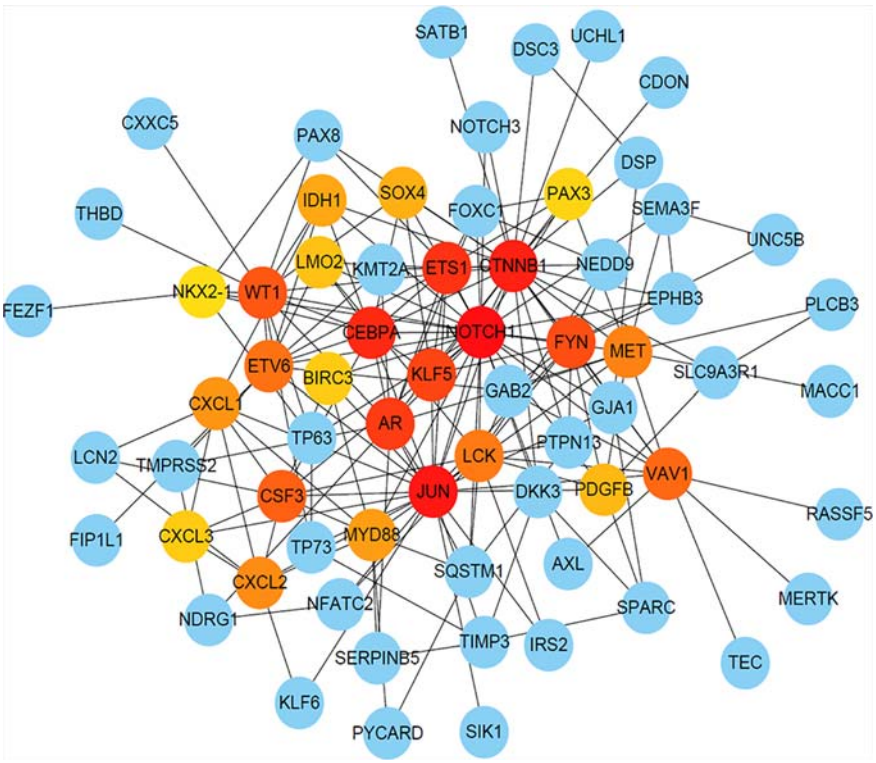


FIGURE 4 | The PPI network of DEGs with Hub score. The dark color indicates high hub score, and the light color predicts low hub score.

OSCC patients ($p = 0.04$) (**Figure 7D**). Although the expression of other hub genes did not show a significant relationship with the survival status of OSCC patients ($p > 0.05$), the OSCC patients with high expression of *NOTCH1* (**Figure 7A**) and *ETS1* (**Figure 7E**) or low expression of *JUN* (**Figure 7B**) tended to have long lifespans.

The dataset obtained from MD Anderson, Cancer Discov 2013, showed that the median overall survival of all OSCC patients was 78.8 months. Except for *NOTCH1*, no mutation was found in the other hub genes in the OSCC patients. And the mutations in *NOTCH1* showed no significant association with the overall survival of patients with OSCC (**Figure 8**), suggesting that the regulation, without mutation of the hub genes was the main mechanism of cisplatin resistance in OSCC.

DISCUSSION

Worldwide, OSCC is an important public health issue with limited therapy strategies and researches; systemic drug resistance has aggravated this situation. In this study, high-throughput screening was used to explore the potential genes involved in cisplatin resistance of OSCC, and *NOTCH1*, *JUN*, *CTNNB1*, *CEBPA*, and *ETS1* were identified as the hub genes in the occurrence of cisplatin resistance. These genes were found to be regulated by the members of the miR-200 family. Regulation of the corresponding hub genes by miRNAs may reverse cisplatin resistance of OSCC, and the sensitivity of tumor cells to cisplatin maybe restored; thus, providing a novel potential target for anticancer therapy.

Studies have shown that changes in NOTCH signaling pathway are associated with many human cancers (Villanueva et al., 2012). *NOTCH1* is reported to be both a tumor suppressor gene and a tumor oncogene. The tumorigenic or anti-tumor activity of NOTCH family members in different types of tumors displays its role in promoting or inhibiting the undifferentiated state of stem cells in the corresponding tissues (Wang et al., 2012). Carcinogenic action of NOTCH has been found in many cancers, including non-small cell lung cancer (Lenhart et al., 2015), acute T lymphoblastic leukemia (Weng et al., 2004), and malignant gliomas (Purow et al., 2005). In contrast, *NOTCH1* signaling is inhibited in neuroendocrine tumor cells, including small cell lung cancer (Platta et al., 2008). This suggests that induction of *NOTCH1* expression is an effective strategy for treating these tumors. NOTCH signaling pathway is also involved in chemotherapy resistance. For example, *NOTCH1* plays an important role in cisplatin resistance mechanism of head and neck squamous cell tumor, colorectal tumor, ovarian cancer (Wang et al., 2010), and other malignant tumors. In this study also, expression of *NOTCH1* gene was found to be significantly inhibited in cisplatin-resistant OSCC cell lines as compared to that in normal or tumor tissues, but no effect was observed on the overall survival of patients. These results suggest that NOTCH1 signaling molecules may be involved in different

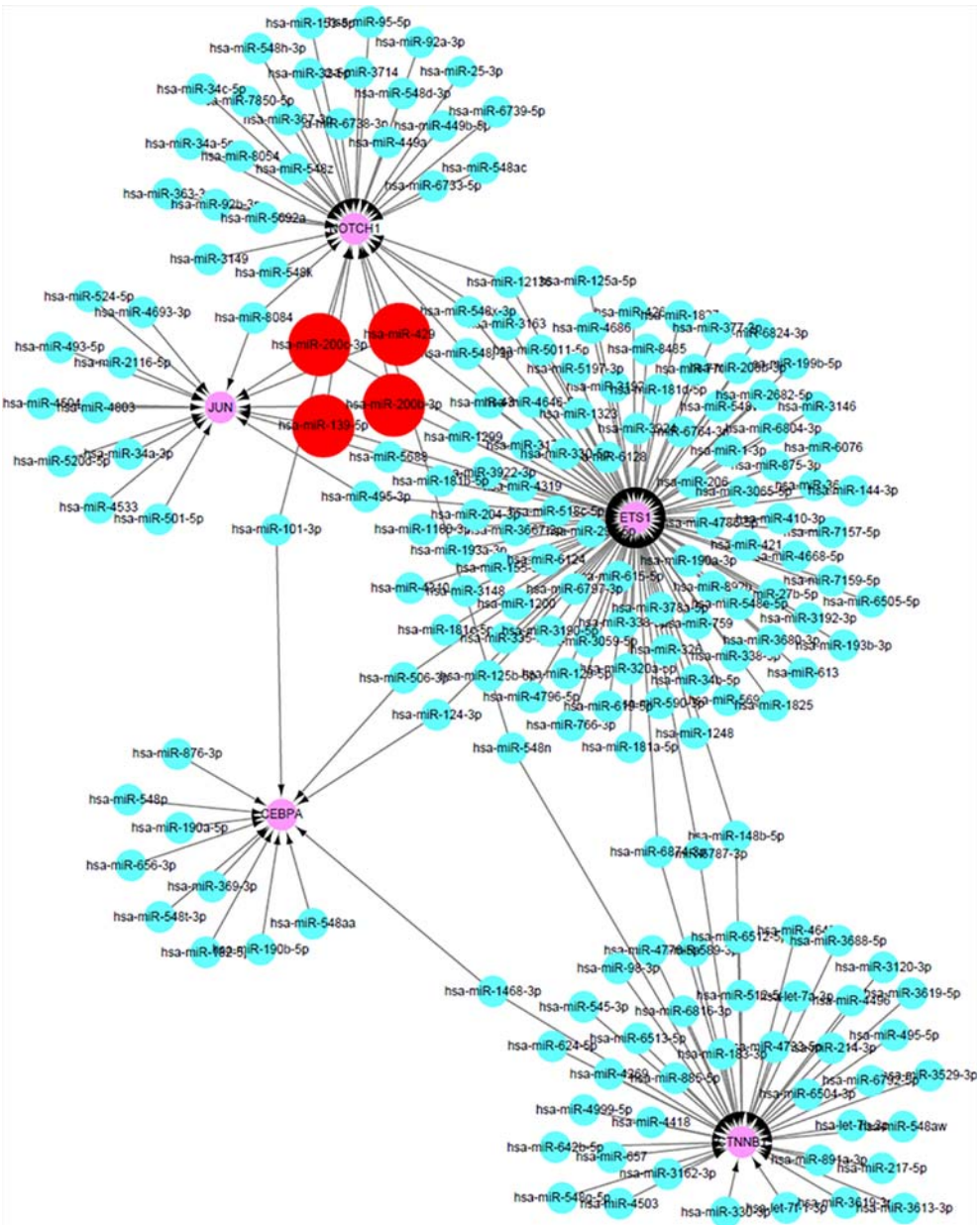


FIGURE 5 | The construction of miRNA–mRNA network of hub genes in OSCC. The red circles predicted the potential miRNAs that can regulate multiple hub genes in OCSS.

biological processes of malignant tumor development through different molecular pathways, and could play an important role in resistance of OSCC against cisplatin and other chemotherapy drugs.

JUN is a protein-coding gene and has no introns; it is located in 1p32-p31: a chromosomal region involved in human malignant translocations and deletions (Fazal et al., 2017) JUN-related diseases, include sarcomas and whooping cough (Syc-Mazurek et al., 2017). *JUN* is involved in the following pathways: apoptosis regulation, signal transduction, tacrolimus/cyclosporine pathway, and pharmacodynamics. *JUN* is also associated with sequence-specific DNA binding (GO annotation). In this study, although *JUN* molecular expression was significantly changed, its correlation with malignant tumor tissues and its influence on patient survival were not found. Therefore, its function and molecular mechanism will be explored in future studies.

The protein encoded by *CTNNB1* is a part of the protein complex that forms the adhesive junctions. Adhesion is necessary to create and maintain the epithelial layer (Li et al., 2017a). The coding proteins, which also include the actin cytoskeleton, are responsible for signaling contact inhibition, and once the upper cortex completes signaling, the cell stops dividing. Finally, the

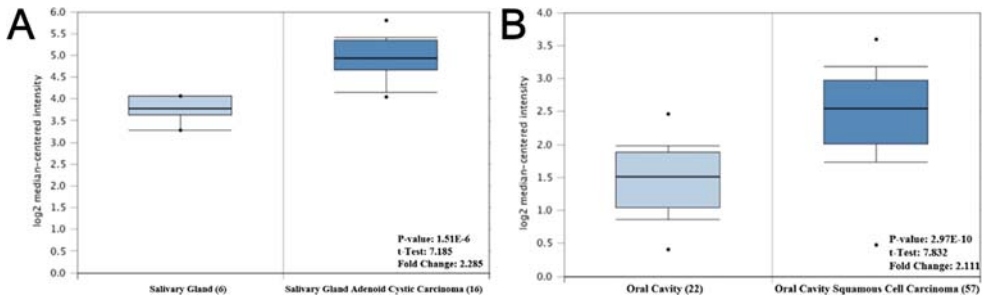


FIGURE 6 | The mRNA expression pattern of hub genes in OSCC. **(A)** The expression of CTNNB1 was increased in OSCC tissues, compared with normal tissues. **(B)** The expression of ETS1 was increased in OSCC tissues, compared with normal tissues.

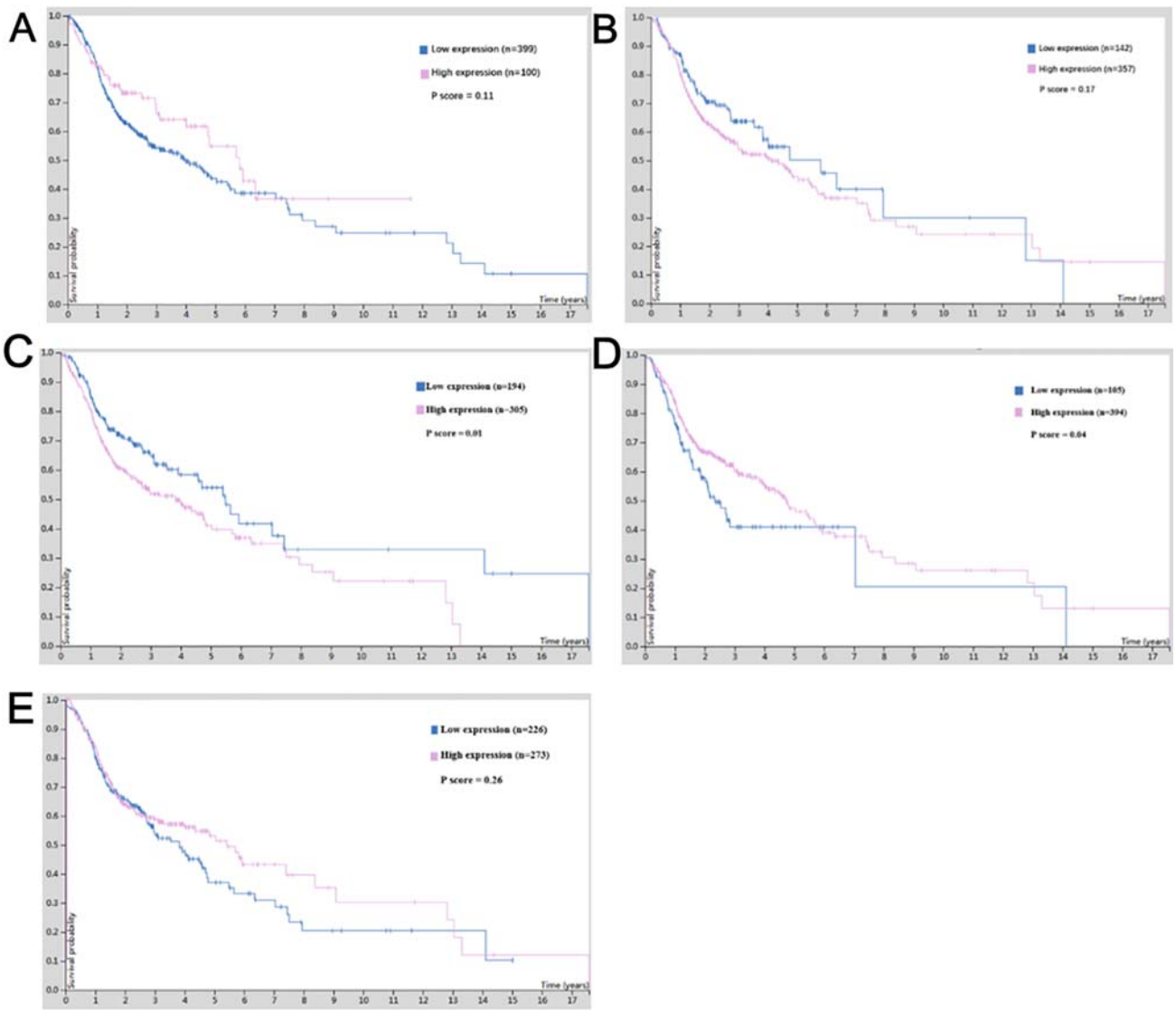


FIGURE 7 | The survival value of the expression of hub genes in cisplatin-resistant OSCC. **(A)** NOTCH1. **(B)** JUN. **(C)** CTNNB1. **(D)** CEBPA. **(E)** ETS1.

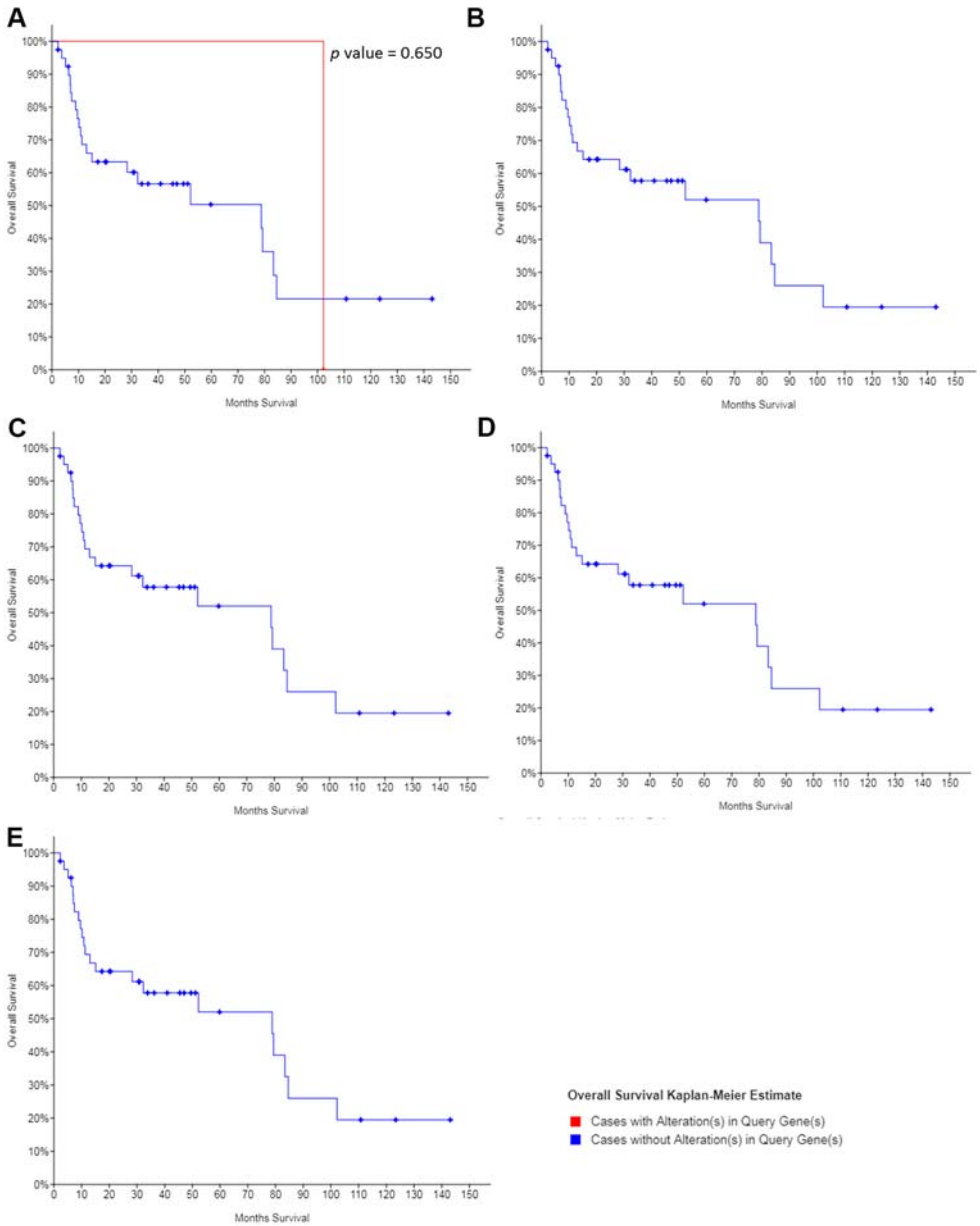


FIGURE 8 | The survival value of the mutations of hub genes in cisplatin-resistant OSCC. **(A)** NOTCH1. **(B)** JUN. **(C)** CTNNB1. **(D)** CEBPA. **(E)** ETS1.

protein binds to the product of the *APC* gene, which is mutated in colorectal adenomatous polyposis. The mutation is a cause of colorectal cancer, hairy tumors, medulloblastoma, and ovarian cancer (Li et al., 2017b). Selective splicing of *CTNNB1* RNA leads to multiple transcript variants. Diseases associated with *CTNNB1*, include hairy tumors and intellectual disability, both being 19 autosomal dominant (Lee et al., 2018). The pathways associated with *CTNNB1* are beta-adrenergic signaling and blood-brain barrier pathways. Because it inhibits the expression of downstream signals, GO annotations associated with it, include DNA-binding transcription factor activity and binding. In this study, it was found that the expression of *CTNNB1* in tumor tissues was significantly higher than that in normal tissues, and the survival period of patients with high expression of *CTNNB1* was significantly shortened. These results suggest that *CTNNB1* also plays an important role in the occurrence and development of OSCC, but the mechanism of its influence on cisplatin chemotherapy resistance needs to be further studied and explored.

CEBPA is an intron-free transcription factor that contains a basic leucine zipper domain and recognizes the CCAAT motif in the target gene promoter (Mannelli et al., 2017). The coding proteins act in homodimers and heterodimers with CCAAT or enhancer binding proteins, beta and gamma. The activity of CEBPA protein regulates the expression of genes,

which are involved in cell cycle regulation and weight balance. Mutation in the *CEBPA* gene has been linked to acute myeloid leukemia (Avellino et al., 2016). *CEBPA* mutations are particularly associated with cytogenetically-normal AML (Taskesen et al., 2011). *CEBPA* is necessary for granulocyte formation in mice. Mutations in *CEBPA* are associated with longer survival of OSCC patients. CEBPA-related diseases, include leukemia, acute myeloid leukemia, and myeloid leukemia. The pathways associated with *CEBPA* are adenoid cystic carcinoma and the adipogenesis pathway. CEBPA has important DNA binding transcription factor activity and can bind to sequence specific DNA. However, there are no relevant studies on *CEBPA* and cisplatin resistance of OSCC at present. In the current study, we found that high expression of *CEBPA* is closely related to the prognosis of OSCC patients.

ETS1 is a member of the encoding transcription factor ETS family, which has a conserved DNA binding domain of ETS that recognizes the hub consistent DNA sequence GGAA/T in the target gene (Poon and Kim, 2017). These proteins act as transcriptional activators or inhibitors of many genes and are involved in stem cell development, cell aging and death, and tumorigenesis. Splicing transcriptional variants encoding different subtypes have also been previously described. Jacobsen syndrome and estrogen receptor negative breast cancer are the diseases associated with *ETS1* (Carpinelli et al., 2015). The pathways involving *ETS1* include focal adhesion and focal adhesion kinase mediated signal transduction events. The gene also has important DNA-binding transcription factor activity and transcription factor binding. We found that *ETS1* is an important cisplatin resistant gene based on high-throughput data analysis, PPI network, and expression verification. Studies have shown that overexpression of *ETS1* induces IKK alpha mRNA and protein expression as well as IKK alpha activity (Gu et al., 2004). In a previous study, ETS1 protein expression and IKK alpha were significantly upregulated in 231 cisplatin-resistant cell lines. Inhibition of *ETS1* expression has been reported to enhance cisplatin sensitivity of resistant cell lines. *ETS1* knockout increases the stability of cisplatin in mouse xenograft models (Zhang et al., 2018). These results are similar to the results obtained in the current study. *ETS1* was highly expressed in cisplatin-resistant OSCC cell lines as compared to that in the normal tissues; ETS1 was highly expressed in tumor tissues, suggesting that it is an important molecule in this process.

Based on previous studies on hub genes and members of the miR-200 family, miR-200b/a/429 transcription is known to be regulated by different transcriptional factors in tissue-specific manner (Kim et al., 2011b). ZEB1/2 is the classical target gene of miR-200s, and many other potential factors have also been reported as the genes regulated by miR-200s (Nagalla et al., 2011). In the current study, new potential target genes were reported as the hub genes in cisplatin-resistant OSCC cells. In 2018, Liu *et al.* reported a smart miRNA-reporter gene for *in vitro* and *in vivo* imaging of biogenesis of miRNA and their related functions (Liu et al., 2018). Further study involving the reporter system could be helpful in investigation of the relationship between miR-200s and the hub genes in OSCC. And as the researches related to OSCC are limited, the relationship between the expression of hub genes and clinicopathological parameters in OSCC patients will be collected and analyzed in the further, to confirm their roles in the occurrence of cisplatin resistance in OSCC.

CONCLUSION

We found that *NOTCH1*, *JUN*, *CTNNB1*, *CEBPA*, and *ETS1* were the key genes regulating cisplatin resistance in OSCC drug-resistant cell lines, and the miR-200 family may be capable of reversing OSCC cell resistance by regulating *NOTCH1*, *JUN*, and *ETS1*, which could also act as potential targets for treating cisplatin resistant OSCC patients.

DATA AVAILABILITY STATEMENT

GSE111585 and GSE115119 were downloaded from Gene Expression Omnibus.

AUTHOR CONTRIBUTIONS

JL conceptualized and designed the study. W-TC, H-TW, and JL organized the database, searched literature, and structured and drafted the manuscript. G-WL, J-XS, Q-QY, M-LZ, and W-JC analyzed and interpreted the data, and drafted and revised the manuscript. JL revised the original manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 81501539 and 81320108015), the Natural Science Foundation of Guangdong Province (No. 2016A030312008), and Li Ka Shing Foundation Grant for Joint Research Program between Shantou University and Technion-Israel Institute of Technology (No. 43209501).

ACKNOWLEDGMENTS

We are thankful to the Fourth CCF Bioinformatics Conference (CBC 2019) for their valuable comments and recommendation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01328/full#supplementary-material

SUPPLEMENTAL FIGURE 1 | The PPI network of DEGs in cisplatin-resistant DEGs. **(A)** The PPI network of up-regulated DEGs. **(B)** The PPI network of down-regulated DEGs.

REFERENCES

Avellino, R., Havermans, M., Erpelinck, C., Sanders, M. A., Hoogenboezem, R., van de Werken, H. J., et al. (2016). An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood* 127 (24), 2991–3003. doi: 10.1182/blood-2016-01-695759

Carpinelli, M. R., Kruse, E. A., Arhatari, B. D., Debrincat, M. A., Ogier, J. M., Bories, J. C., et al. (2015). Mice haploinsufficient for ETS1 and FLI1 display middle ear abnormalities and model aspects of Jacobsen syndrome. *Am. J. Pathol.* 185 (7), 1867–1876. doi: 10.1016/j.ajpath.2015.03.026

Fazal, S. V., Gomez-Sanchez, J. A., Wagstaff, L. J., Musner, N., Otto, G., Janz, M., et al. (2017). Graded elevation of c-Jun in Schwann cells *in vivo*: gene dosage determines effects on development, remyelination, tumorigenesis, and hypomyelination. *J. Neurosci.* 37 (50), 12297–12313. doi: 10.1523/JNEUROSCI.0986-17.2017

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* 6 (269), pl1. doi: 10.1126/scisignal.2004088

Gu, L., Zhu, N., Findley, H. W., Woods, W. G., and Zhou, M. (2004). Identification and characterization of the IKKalpha promoter: positive and negative regulation by ETS-1 and p53, respectively. *J. Biol. Chem.* 279 (50), 52141–52149. doi: 10.1074/jbc.M407915200

Hong, L., Yang, Z., Ma, J., and Fan, D. (2013). Function of miRNA in controlling drug resistance of human cancers. *Curr. Drug Targets* 14 (10), 1118–1127. doi: 10.2174/13894501113149990183

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4 (1), 44–57. doi: 10.1038/nprot.2008.211

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092

Kim, S. Y., Nam, S. Y., Choi, S. H., Cho, K. J., and Roh, J. L. (2011a). Prognostic value of lymph node density in node-positive patients with oral squamous cell carcinoma. *Ann. Surg. Oncol.* 18 (8), 2310–2317. doi: 10.1245/s10434-011-1614-6

Kim, T., Veronese, A., Pichiorri, F., Lee, T. J., Jeon, Y. J., Volinia, S., et al. (2011b). p53 regulates epithelial–mesenchymal transition through microRNAs targeting ZEB1 and ZEB2. *J. Exp. Med.* 208 (5), 875–883. doi: 10.1084/jem.20110235

Lee, Y. H., Huang, W. C., and Hsieh, M. S. (2018). CTNNB1 mutations in basal cell adenoma of the salivary gland. *J. Formos. Med. Assoc.* 117 (10), 894–901. doi: 10.1016/j.jfma.2017.11.011

Leemans, C. R., Braakhuis, B. J., and Brakenhoff, R. H. (2011). The molecular biology of head and neck cancer. *Nat. Rev. Cancer* 11 (1), 9–22. doi: 10.1038/nrc2982

Lenhart, R., Kirov, S., Desilva, H., Cao, J., Lei, M., Johnston, K., et al. (2015). Sensitivity of small cell lung cancer to BET inhibition is mediated by regulation of ASCL1 gene expression. *Mol. Cancer Ther.* 14 (10), 2167–2174. doi: 10.1158/1535-7163.MCT-15-0037

Li, N., Xu, Y. F., Li, G. Q., Yu, T. T., Yao, R. E., Wang, X. M., et al. (2017a). Exome sequencing identifies a *de novo* mutation of CTNNB1 gene in a patient mainly presented with retinal detachment, lens and vitreous opacities, microcephaly, and developmental delay. *Medicine* 96 (20), e6914. doi: 10.1097/Md.0000000000006914

Li, Y. K., Zhang, F. Q., and Yang, D. H. (2017b). Comprehensive assessment and meta-analysis of the association between CTNNB1 polymorphisms and cancer risk. *Biosci. Rep.* 37 (6), BSR20171121. doi: 10.1042/Bsr20171121

Liang, H., Gong, F., Zhang, S., Zhang, C. Y., Zen, K., and Chen, X. (2014). The origin, function, and diagnostic potential of extracellular microRNAs in human body fluids. *Wiley Interdiscip. Rev. RNA* 5 (2), 285–300. doi: 10.1002/wrna.1208

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. (2003). Vertebrate microRNA genes. *Science* 299 (5612), 1540. doi: 10.1126/science.1080372

Lin, Z., Sun, L., Xie, S., Zhang, S., Fan, S., Li, Q., et al. (2018). Chemotherapy-induced long non-coding RNA 1 promotes metastasis and chemo-resistance of TSCC *via* the Wnt/beta-catenin signaling pathway. *Mol. Ther.* 26 (6), 1494–1508. doi: 10.1016/j.ymthe.2018.04.002

Liu, Y., Sun, J., and Zhao, M. (2017). ONGene, a literature-based database for human oncogenes. *J. Genet. Genomics* 44 (2), 119–121. doi: 10.1016/j.jgg.2016.12.004

Liu, J., Shen, J. X., He,, and Zhang, G. J. (2018). Bioluminescence imaging for monitoring miR-200c expression in breast cancer cells and its effects on epithelial–mesenchymal transition progress in living animals. *Mol. Imaging Biol.* 20 (5), 761–770. doi: 10.1007/s11307-018-1180-4

Mannelli, F., Ponziani, V., Bencini, S., Bonetti, M. I., Benelli, M., Cutini, I., et al. (2017). CEBPA-double-mutated acute myeloid leukemia displays a unique phenotypic profile: a reliable screening method and insight into biological features. *Haematologica* 102 (3), 529–540. doi: 10.3324/haematol.2016.151910

Nagalla, S., Shaw, C., Kong, X., Kondkar, A. A., Edelstein, L. C., Ma, L., et al. (2011). Platelet microRNA–mRNA coexpression profiles correlate with platelet reactivity. *Blood* 117 (19), 5189–5197. doi: 10.1182/blood-2010-09-299719

Ong, W., Zhao, R., Lui, B., Tan, W., Ebrahimi, A., Clark, J. R., et al. (2016). Prognostic significance of lymph node density in squamous cell carcinoma of the tongue. *Head Neck* 38 (1), E859–E866. doi: 10.1002/hed.24113

Petersen, P. E. (2003a). Global framework convention on tobacco control: the implications for oral health. *Community Dent. Health* 20 (3), 137–138. doi: 10.1126/science.106.2757.419

Petersen, P. E. (2003b). Tobacco and oral health—the role of the World Health Organization. *Oral Health Prev. Dent.* 1 (4), 309–315.

Petersen, P. E. (2005). Strengthening the prevention of oral cancer: the WHO perspective. *Community Dent. Epidemiol.* 33 (6), 397–399. doi: 10.1111/j.1600-0528.2005.00251.x

Platta, C. S., Greenblatt, D. Y., Kunnimalaiyaan, M., and Chen, H. (2008). Valproic acid induces Notch1 signaling in small cell lung cancer cells. *J. Surg. Res.* 148 (1), 31–37. doi: 10.1016/j.jss.2008.03.008

Poon, G. M. K., and Kim, H. M. (2017). Signatures of DNA target selectivity by ETS transcription factors. *Transcription* 8 (3), 193–203. doi: 10.1080/21541264.2017.1302901

Purow, B. W., Haque, R. M., Noel, M. W., Su, Q., Burdick, M. J., Lee, J., et al. (2005). Expression of Notch-1 and its ligands, Delta-like-1 and Jagged-1, is critical for glioma cell survival and proliferation. *Cancer Res.* 65 (6), 2353–2363. doi: 10.1158/0008-5472.Can-04-1890

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6 (1), 1–6. doi: 10.1016/s1476-5586(04)80047-2

Sakai, N. S., Samia-Aly, E., Barbera, M., and Fitzgerald, R. C. (2013). A review of the current understanding and clinical utility of miRNAs in esophageal cancer. *Semin. Cancer Biol.* 23 (6 Pt B), 512–521. doi: 10.1016/j.semcancer.2013.08.005

Syc-Mazurek, S. B., Fernandes, K. A., and Libby, R. T. (2017). JUN is important for ocular hypertension-induced retinal ganglion cell degeneration. *Cell Death Dis.* 8, e2945. doi: 10.1038/Cddis.2017.338

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi: 10.1093/nar/gky1131

Taskesen, E., Bullinger, L., Corbacioglu, A., Sanders, M. A., Erpelinck, C. A. J., Wouters, B. J., et al. (2011). Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* 117 (8), 2469–2475. doi: 10.1182/blood-2010-09-307280

Thomas, P. D. (2017). The gene ontology and the meaning of biological function. *Methods Mol. Biol.* 1446, 15–24. doi: 10.1007/978-1-4939-3743-1_2

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al (2017). A pathology atlas of the human cancer transcriptome. *Science* 357 (6352). pii: eaan2507. doi: 10.1126/science.aan2507

Villanueva, A., Alsinet, C., Yanger, K., Hoshida, Y., Zong, Y. W., Toffanin, S., et al. (2012). Notch signaling is activated in human hepatocellular carcinoma and induces tumor formation in mice. *Gastroenterology* 143 (6), 1660–166+. doi: 10.1053/j.gastro.2012.09.002

Wang, Z., Li, Y., Ahmad, A., Azmi, A. S., Banerjee, S., Kong, D., et al. (2010). Targeting Notch signaling pathway to overcome drug resistance for cancer therapy. *Biochim. Biophys. Acta* 1806 (2), 258–267. doi: 10.1016/j.bbcan.2010.06.001

Wang, J., Sullenger, B. A., and Rich, J. N. (2012). Notch signaling in cancer stem cells. *Adv. Exp. Med. Biol.* 727, 174–185. doi: 10.1007/978-1-4614-0899-4_13

Weng, A. P., Ferrando, A. A., Lee, W., Morris, J. P. T., Silverman, L. B., Sanchez-Irizarry, C., et al. (2004). Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science* 306 (5694), 269–271. doi: 10.1126/science.1102160

Wong, N., and Wang, X. (2015). miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res.* 43, D146–D152. doi: 10.1093/nar/gku1104

World Health, O. (2003). World Health Assembly adopts historic tobacco control pact. *Indian J. Med. Sci.* 57 (8), 377–378. doi: 10.1016/j.medmal.2005.06.009

Wu, Q., Yang, Z., Nie, Y., Shi, Y., and Fan, D. (2014). Multi-drug resistance in cancer chemotherapeutics: mechanisms and lab approaches. *Cancer Lett.* 347 (2), 159–166. doi: 10.1016/j.canlet.2014.03.013

Zhang, Y. Z., Wu, J. J., Ye, M. N., Wang, B., Sheng, J. Y., Shi, B. L., et al. (2018). ETS1 is associated with cisplatin resistance through IKK alpha/NF-kappa B pathway in cell line MDA-MB-231. *Cancer Cell Int.* 18, 86. doi: 10.1186/s12935-018-0581-4

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: a literature-based database of tumor suppressor genes for pan-cancer analysis. *Nucleic Acids Res.* 44 (D1), D1023–D1031. doi: 10.1093/nar/gkv1268

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fgene.2019.01376

Identification of AIDS-Associated Kaposi Sarcoma: A Functional Genomics Approach

Peng Zhang[1,2†], *Jiafeng Wang*[3†], *Xiao Zhang*[4†], *Xiaolan Wang*[5], *Liying Jiang*[6] *and Xuefeng Gu*[6*]

[1] School of Clinical Medicine, Shanghai University of Medicine & Health Sciences, Shanghai, China, [2] Department of Public Health, Shanghai General Practice Medical Education and Research Center, Shanghai, China, [3] Stem Cell Research and Cellular Therapy Center, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China, [4] Department of Implant Dentistry, Ninth People's Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China, [5] College of Nursing and Health Management, Shanghai University of Medicine & Health Sciences, Shanghai, China, [6] Shanghai Key Laboratory of Molecular Imaging, Collaborative Research Center, Shanghai University of Medicine & Health Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Chenqi Wang,
University of South Florida, United States
Biju Issac,
Leidos Biomedical Research, Inc., United States
Lei Chen,
Shanghai Maritime University, China

***Correspondence:**
Xuefeng Gu
guxf@sumhs.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 03 September 2019
Accepted: 17 December 2019
Published: 24 January 2020

Citation:
Zhang P, Wang J, Zhang X, Wang X, Jiang L and Gu X (2020) Identification of AIDS-Associated Kaposi Sarcoma: A Functional Genomics Approach. Front. Genet. 10:1376. doi: 10.3389/fgene.2019.01376

Background: Kaposi sarcoma-associated herpes virus (KSHV) is one of the most common causal agents of Kaposi Sarcoma (KS) in individuals with HIV-infections. The virus has gained attention over the past few decades due to its remarkable pathogenic mechanisms. A group of genes, ORF71, ORF72, and ORF73, are expressed as polycistronic mRNAs and the functions of ORF71 and ORF72 in KSHV are already reported in the literature. However, the function of ORF73 has remained a mystery. The aim of this study is to conduct comprehensive exploratory experiments to clarify the role of ORF73 in KSHV pathology and discover markers of AIDS-associated KSHV-induced KS by bioinformatic approaches.

Methods and Results: We searched for homologues of ORF-73 and attempted to predict protein-protein interactions (PPI) based on GeneCards and UniProtKB, utilizing Position-Specific Iterated BLAST (PSI-BLAST). We applied Gene Ontology (GO) and KEGG pathway analyses to identify highly conserved regions between ORF-73 and p53to help us identify potential markers with predominant hits and interactions in the KEGG pathway associated with host apoptosis and cell arrest. The protein p53 is selected because it is an important tumor suppressor antigen. To identify the potential roles of the candidate markers at the molecular level, we used PSIPRED keeping the conserved domains as the major parameters to predict secondary structures. We based the FUGE interpretation consolidations of the sequence-structure comparisons on distance homology, where the score for the amino acids matching the insertion/deletion (indels) detected were based on structures compared to the FUGE database of structural profiles. We also calculated the compatibility scores of sequence alignments accordingly. Based on the PSI-BLAST homologues, we checked the disordered structures predicted using PSI-Pred and DISO-Pred for developing a hidden Markov model (HMM). We further applied these HMMs models based on the alignment of constructed 3D models between the known structure and the HMM of our sequence. Moreover, stable homology and

structurally conserved domains confirmed that ORF-73 maybe an important prognostic marker for AIDS-associated KS.

Conclusion: Collectively, similar variants of ORF-73 markers involved in the immune response may interact with targeted host proteins as predicted by our computational analysis. This work also suggests the existence of potential conformational changes that need to be further explored to help elucidate the role of immune signaling during KS towards the development of therapeutic applications.

Keywords: herpesvirus, immune evasion, sequence homology, protein–protein interactions, AIDS, ORF-73

INTRODUCTION

Pre-existing human immunodeficiency virus (HIV) infections affect the immune system increasing the risk for development of Kaposi sarcoma (KS). Since the discovery of Kaposi sarcoma-associated herpesvirus (KSHV), also termed human herpesvirus 8 (HHV8), the tumor development and oncogenesis were associated with co-expression of different genes (Barré-Sinoussi et al., 1983; Gelmann et al., 1983). KS is a common type of cancer associated with blood vessels and lymph nodes. Soon after the discovery of HIV-1, scientists discovered γ-herpesvirus in KS lesions (Chang et al., 1994). Now that the full KSHV genome has been sequenced, it fulfils Koch's modern postulates linking the KS cancer initiation to the oncogenic virus (Russo et al., 1996; zur Hausen, 2001). KSHV is a key viral pathogen in cancer biology affecting humans and its discovery promoted clinical and epidemiological research into viral oncology (Chang et al., 1994). However, many questions remain unanswered due to the significant mortality and rapid morbidity of those affected by HIV-1 and KSHV (Parkin, 2006; Sinfield et al., 2007; Dittmer and Damania, 2019; Gaur et al., 2019).

In fact, KS was named after Dr. Moritz Kaposi, a prominent Hungarian dermatologist, who described KS as an 'idiopathic pigmented sarcoma of the skin' in 1872 (Kaposi, 1872). The evolved gamma-herpesviruses have been classified into many subfamilies (Roizman et al., 1981) and produce many viral gene products capable of subverting the normal cellular machinery through processes involving apoptosis, cell cycle progression, antiviral responses, and immune surveillance resulting in alterations in master cell signaling pathways to establish a persistent host infection. The double-stranded KSHV genome (124–174 kb) is enclosed in an icosahedral capsid composed of 162 capsomeres with many of its ORFs being conserved in alpha- and beta-herpesviruses, but absent from other herpesviruses.

The KSHV is closely related to the subfamily Rhadinoviridae (gamma-2-herpesviruses), which is also close to the Herpes virus saimiri (HVS); therefore, similarities between ORFs of KSHV and HVS may influence the pathogenesis of KS (Schäfer et al., 2003). The HVS genome exists as a stable non-integrated circular episome in altered human and simian T cells. A group of genes, ORF71, ORF72, and ORF73, are located at the right end of the L-DNA and are expressed as polycistronic mRNAs (Fickenscher et al., 1996). Initial studies discerned that both KSHV and HVS ORF71 encode the anti-apoptotic FLICE inhibitory protein (vFLIP) (Thome et al., 1997), although HVS ORF71 is not mandatory for viral replication, transformation, or pathogenicity (Glykofrydes et al., 2000). ORF72 produces a v-Cyclin D homolog which is important for transformation of human T lymphocytes (Ensser et al., 2001). However, the function of ORF73 has remained a mystery. Therefore, developing and conducting comprehensive exploratory experiments to clarify the role of ORF73 in KSHV pathology is important.

Typically, the phenotypic features of KS initially appear on the face, legs, or feet as painless red spots but, in severe cases, the lesions also appear in the lungs and digestive tract (Bhutani et al., 2015; Yarchoan et al., 2015). KSHV is considered an oncogenic human virus (Martin et al., 1998). People with weak immune systems are more susceptible to HHV-8 infection (triggering KS development). Even with the availability of the anti-retroviral treatment [HAART], the prevalence of AIDS-associated KS has not declined significantly (Nguyen et al., 2008). Although KSHV infection is important for the onset of KS, additional factors must be present to allow the establishment of the lesions. The chance of infection is one in 100,000 among the general population, but only around one in 20 among HIV-infected individuals (La Ferla et al., 2013). The chance of acquiring the infection was one in three among HIV-infected individuals before the introduction of HAART (Beral et al., 1990; Gallo, 1998). Epidemiological observations from incidence rates in endemic areas suggest that HIV-negative individuals with KSHV infections never develop KS due to the role of immunological host factors including immune-response genes and genetic polymorphisms of the inflammatory modulators (Cottoni et al., 2004; Gazouli et al., 2004; Dorak et al., 2005).

KSHV infection of endothelial and/or hematopoietic progenitors (Della Bella et al., 2008) alter their morphology (Moses et al., 1999), growth rate, gene expression (Flore et al., 1998; Ciufo et al., 2001), and glucose metabolism (Delgado et al., 2010), leading to development of KS. Antibody titers specific for KSHV correlate with its viral load. Among individuals with low viral load, antibody titer concentrations may be too low for current serological assays to identify them. Identification of circulating biomarkers in KSHV-associated disease may help in predicting clinical outcomes (Aka et al., 2015). Immune modulatory and evasion proteins of KSHV modulate cellular responses associated with complement activation, autophagy, IFN family signaling, chemokines, natural killer cells, and

apoptosis (Liang et al., 2008). They are located in a region of the viral capsid that is rich in a protein known as tegument. Six tegument proteins have been identified: ORF21, ORF33, ORF45, ORF63, ORF64, ORF73 and ORF75. Among these, the roles of ORF63 and ORF64 in immune evasion have been elucidated (Zhu et al., 2005; Gregory et al., 2011). We focused on the identification of the role of ORF73 in KSHV. The ORF73 gene encodes the HHV-LANA1 viral proteins that have been linked with AIDS-associated KS, indicating an association between HIV and ORF73. For our computational study, we hypothesized that ORF-73 is a viral proliferation factor based on studies on KS and on its interactions with the host gene p53 (Woodberry et al., 2005). The importance of ORF-73 for cellular host apoptosis through the p53 signaling pathway and p53 is in order of ORF-73 which illustrates the molecular mechanism of this key biomarker associated with KS (Duus et al., 2004).

The variability in KS lesions observed in histopathological assays include spindle cell hemangiomas, cutaneous angiosarcomas, vascular leiomyomas, and fibrous histiocytomas (Hunt et al., 2004). Endothelial biomarkers, such as CD31 and CD34, bcl-2, c-kit, Ki-67, and p53, have been used to distinguish nonvascular spindle sarcomas from angiosarcomas (Weeden, 2002; Fukunaga, 2005). Hence, investigating the HHV-latent associated nuclear antigen-1 (LANA-1) viral protein encoded by ORF-73 is important to identify markers for AIDS-associated KS. Also, studying its interactions may help in the development of preventive strategies and therapeutic options against KS. In this study, we used advanced bioinformatics tools and approaches to identify KS markers **Supplementary Figure 1**.

MATERIALS AND METHODS

Selection of Markers

We used publicly available databases including the National Centre for Biotechnology Information (NCBI), GeneCards (Hou et al., 2017) and UniProtKB (Tang et al., 2013) to identify potential markers of KS and selected the most specific ones using "Kaposi's sarcoma" as a keyword. Human protein markers were further ran through a BLAST search for homology sequences. We extracted ORF-73 sequences from the NCBI database search using the accession number AAC57158.1. These are the exact URLs of the searched databases we used to identify markers associated with KS : GeneCards https://genecards.weizmann.ac.il/v3/index.php?path=/Search/keyword/kaposi%20sarcoma%20markers/0/20; UniPortKB https://www.uniprot.org/uniprot/?query=kaposi+sarcoma&sort=score; and NCBI https://www.ncbi.nlm.nih.gov/protein/?term=ORF-73%20kaposi%20sarcoma).

Bioinformatics: Sequence Computational Analysis

We used publicly available internet-based protein search tools and bioinformatics programs with default settings, unless otherwise stated in the text, for the analysis. We tested selected protein sequences to identify conserved domains from NCBI and BLAST algorithms, and we used the PSIPRED program to predict the secondary structure of proteins based on the conserved domain sequences. We further executed a position specific iterative BLAST (PSI-BLAST) search to build a PSSMs (position specific score matrix), which could predict the secondary structure of the input sequences (Majerciak et al., 2015) to predict secondary structures of the selected conserved domains based on multiple sequence alignment related proteins spanning a variety of organisms to reveal sequence regions containing the same, or similar, patterns of amino acids. We submitted the primary sequence of ORF-73 to FUGUE to show the sequence-structural homology by identifying distant sequence-structure homologues and alignments comparing amino acid insertions/deletions (Shi et al., 2001). We used BLASTp and PSI-BLAST (non-redundant protein databases) for pattern specific profiling (Bujnicki and Rychlewski, 2001).

Gene Ontology and Pathway Enrichment Analysis

We chose the ORF-73 target effector to perform a Gene Ontology (GO) search, is a hierarchical graph-based annotation system where the terms closer to the root describe more general information while those away from the root provide more specific information about a given GO category and all the GO terms associated with a protein sequence were obtained from the GO database. The KEGG network pathway enrichment analysis by collecting data of related genomes and their pathways associated with diseases (Yan et al., 2013) and we set a P value <0.05 as the cut-off criterion.

Protein–Protein Interaction (PPI) Network Analysis

We used the online Search Tool for the Retrieval of Interacting Genes (STRING) (Franceschini et al., 2013) and GeneMania (https://genemania.org/) to analyze interactions associated with KS among the proteins encoded by the DEGs. The two parts of GeneMania algorithm consists of an algorithm based on linear regression to calculate functional association from multiple networks from different data sources; and a label predicting gene function of composite network. We employed keywords such as—ORF73 to determine interacting partners. This was pursued using downstream regulator p53 as an apoptosis marker during pathogenesis in the host. Moreover, the marker protein was used for transient interaction study.

PPI Biochemical Analysis

We immobilized His-tag, GST-tag, or biotin-tag bait proteins to an affinity resin and incubated them with solution expressed proteins as prey proteins. We then captured the bound bait and pulled down the cell lysate flow through. Subsequently, we used mass spectrometry (MS) or Western blots to confirm interactions. Using this technique, we determined interacting protein partners of relevant proteins (Einarson, 2001; Arifuzzaman et al., 2006).

RESULTS

Homology Search and KS Marker Identification

Annotations used to search for the KS-associated markers in the UniProtKB database quoted about 137 entries, which we then screened to find those with computationally annotated data. Search engine GeneCards reported about 369 KS markers with a relevance score. **Table 1** lists the markers with the top ten scores.

We found61 ORF-73 marker homologous hits related to the family of human gamma herpes virus 8 with varied E-values. Out of these, we used only the most identical sequence (based on sequence identity was measured by matched by dividing the length of region aligned match), AAC57158.1, for our computational analyses. A search for proteins similar to the selected marker ORF-73 resulted in8 protein accessions (ORF21, ORF33, ORF45, ORF63, ORF64, and ORF75), and 2 CDS regions (accession numbers AAC57158.1 and AAC55944.1).

Domain Prediction and Structural Profile

We looked for conserved domains in the marker protein ORF-73 based on hypothetical domain sequences using literature recapitulation NCBI's Conserve Domain Database (CDD). To identify potential marker roles at the molecular level, we focused on its predicted secondary structure. Therefore, we searched for hypothetical protein having conserved domain and used accession number AAC5744 of gi.1633572 in an NCBI domain search and found only one significant hypothetical conserved domain (PHA03169) with the same accessison number (**Figure 1**). We then used PSIPRED to predict the secondary structure, noted the conserved domains (**Figure 2**) and highlighted the regions with different markers to predict the secondary structures. FUGE interpretation consolidations of the sequence-structure comparison were based on distance homology, where the score for the amino acids matching the insertion/deletion (indels) detected were based on structures compared to the FUGE database of structural profiles and we calculated the compatibility scores of sequence alignment accordingly (**Table 2**).

TABLE 1 | GeneCards and UniPortKB databases used to choose the top-most scored identities of markers associated with KS.

GeneCard database

Sl. No	Symbol	Description	GC id	Score
1	KRT15	Keratin 15	GC17M039675	1.58
2	OSM	Oncostatin M	GC22M030658	1.58
3	TAT	Tyrosine aminotransferase	GC16M071599	1.27
4	MKI67	Marker of proliferation Ki-67	GC10M129894	1.14
5	CD34	CD34 molecule	GC01M208057	1.11
6	PTX3	Pentraxin 3, long	GC03P157154	1.09
7	PECAM1	Platelet/endothelial cell adhesion molecule 1	GC17M062399	1.01
8	FLI1	Fli-1 proto-oncogene, ETS transcription factor	GC11P128596	1.01
9	IFNA2	Interferon, alpha 2	GC09M021374	1.01
10	ACTC1	Actin, alpha, cardiac muscle 1	GC15M035080	0.99

Uniport KB database

Sl. No.	Entry name	Protein name	Entry	Gen name
1	MIR1_HHV8P	E3 ubiquitin-protein ligase MIR1	P90495	K3
2	MIR2_HHV8P	E3 ubiquitin-protein ligase MIR2	P90489	K5
3	GB_HHV8P	Envelope glycoprotein B	F5HB81	gBORF8
4	ARBH_HHV8P	Apoptosis regulator Bcl-2 homolog	F5HGJ3	vBCL2 ORF16
5	SCAF_HHV8P	Capsid scaffolding protein	Q2HRB6	ORF17
6	OX2V_HHV8P	OX-2 membrane glycoprotein homolog	P0C788	K14
7	GN_HHV8P	Envelope glycoprotein N	F5HFQ0	gN ORF53
8	GM_HHV8P	Envelope glycoprotein M	F5HDD0	gM ORF39
9	ORF45_HHV8P	Protein ORF45	F5HDE4	ORF45
10	VMI2_HHV8P	Viral macrophage inflammatory protein	Q98157	ORF K4
11	VIRF1_HHV8P	VIRF-1	F5HF68	vIRF-1
12	ICP27_HHV8P	mRNA export factor ICP27 homolog	Q2HR75	ORF57
13	GH_HHV8P	Envelope glycoprotein H	F5HAK9	gH ORF22
14	AN_HHV8P	Shutoff alkaline exonuclease	Q2HR95	ORF37
15	LANA1_HHV8P	Protein LANA1	Q9QR71	LANA1 ORF73

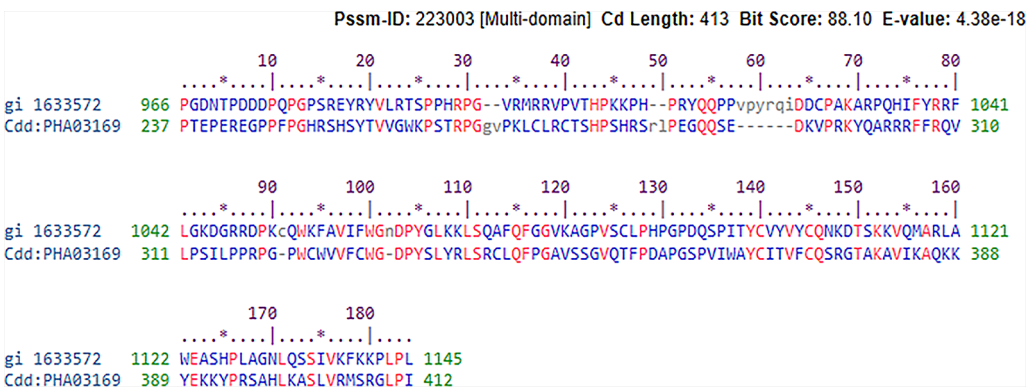


FIGURE 1 | Conserved hypothetical protein domain of PHA03169 in reference to the ORF-73 of Human gamma herpesvirus 8,E-value 38e–18.

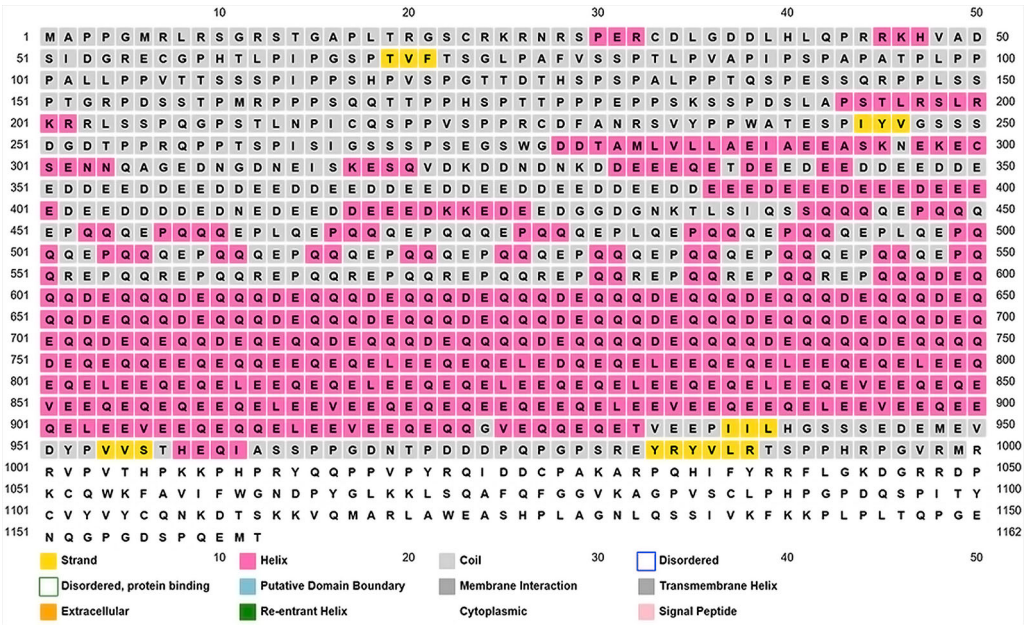


FIGURE 2 | Overview of the ORF-73 secondary structure prediction. The predicted structural positions incorporate two feed-forward neural networks obtained from PSI-BLAST.

TABLE 2 | Structure of Kaposi sarcoma marker ORF-73 predicted based on an environmental-specific substitution table and its structure-dependent gap penalties.

Sl. No.	Profile Hit	PLEN	RAWS	RVN	ZSCORE
1	hs4blga	121	−755	247	24.21
2	hs2ap3a	191	215	8	17.29
3	hs2qiha	136	−822	10	16.57
4	hs2p03a	323	249	21	14.78
5	hs1i4da	188	157	33	14.61
6	hs4cgka	351	325	115	13.67
7	hs2eqbb	93	−880	5	13.53
8	hs1fxka	103	168	19	13.45
9	hs1owaa	156	166	6	13.28
10	hs4hpqc	396	−555	5	12.92

PLEN, Profile length; RAWS, Raw alignment score; RVN, (Raw score)-(Raw score for NULL model); ZSCORE, Z-score normalized by sequence divergence (evolutionary relationship associated with a score >5.0 to the sequences are compared to each other); ZORI, Original Z-score (before normalization).

Using PSI-BLAST, we confined the search of HHV-latency-associated nuclear antigen homology to ORF-73 homologs. The DNA binding of viral protein associated with HHV-8 LANA sheltered 134 residues covering 12% of the sequence with 100% confidence based on the single highest scoring template of c4k2jB (**Figures 3** and **4**). 598 residues covering 51% could be modelled at >90% confidence using multiple-templates. We submitted the top-ranking model of the protein (c4k2jB, 100.0% confidence) to the 3DLigandSite (Wass et al., 2010) server to predict potential binding sites. Based on PSI-BLAST homologues, the predicted disordered structures were checked using PSI-Pred (Jones, 1999) and DISO-Pred (Jones and Cozzetto, 2015) for generating a hidden Markov model (HMM). The models were based on the alignment of the constructed 3D models between the known structure and the HMM of our sequence predicting the3-states—α-helix, β-strand or coil ("SS" indicates the predicted confidence; middle orange, yellow, and green indicate the confidence of prediction).

Gene Expression and Pathway Prediction

The exclusive over-expression of HHV-8 LANA-1 in KS confirms significant sensitivity and specificity. The domain is conserved in the HHV-8 and ORF-73, suggesting its expression during viral latency and allowing it to interact with p53, thereby inducing the apoptosis pathway. The evidence from another study indicates abnormal expression of p53 in the nodular region and metastatic lesion of angiosarcomas (rather than in the primary lesion) (Yee-Lin et al., 2018). To account for this, the lead p53 in KS was taken with reference to the database for a herpes virus-associated infection model so as to understand the immune evasion with a detailed pathway demonstrating the dominant role of a p53 oncogene in KSHV- (**Figure 5**). The tumor suppressor antigen p53 depends on cellular conditions inducing arrest of the cell growth and controlling cell division. This process inhibits cyclin-dependent kinases mediated by the expression of BAX and FAS antigens or by the repression of the Bcl-2expression (Kanashiro et al., 2003). Addressing the markers involved in the cell-cycle arrest is important to understand the molecular evolution of KS and for work towards its eradication. We examined PPIs to explore the complex biochemical interactions and molecular functions of proteins of interest with cellular components, as reported in **Table 3**. **Table 3** also presents the functional enrichment of p53 including its biological process, molecular functions, and cellular components. The effector p53 is directly involved in the arrest of the

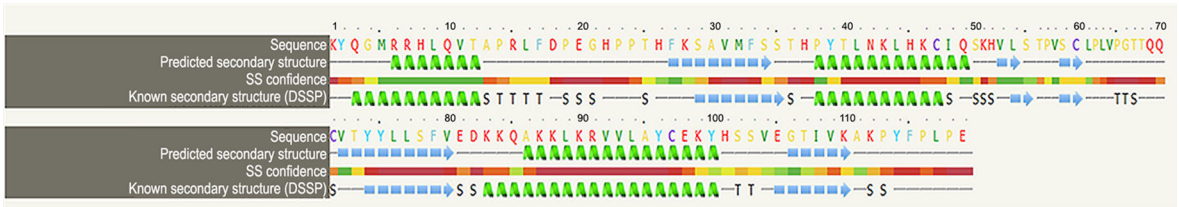


FIGURE 3 | Highest scored template c4k2jB chain B structure.

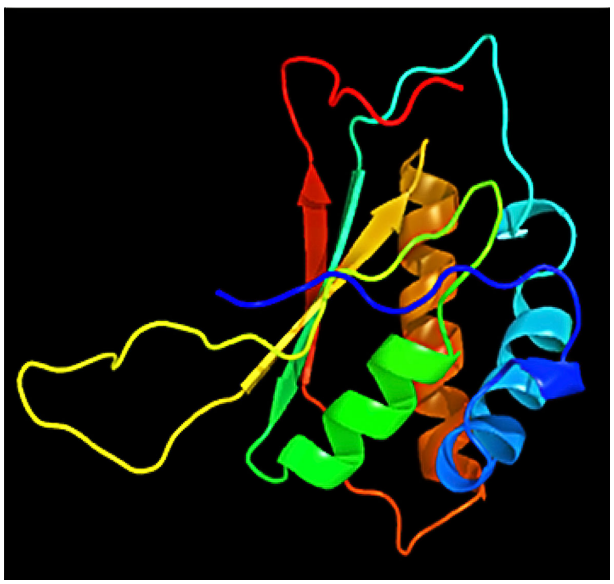


FIGURE 4 | Decameric ring structure of KSHV HHV-LANA DNA binding domain with dimensions (X:40.909, Y:43.389, and Z:44.674).

G1/S cell-cycle progression from normal to cancerous cells (Chen, 2016). Analysis of PPI with STRING showed an enriched p-value of 1.31e−05 with respect to the network having significantly more interactions than expected with 11 nodes, 47 edges, an average node degree of 8.55 and an average local cluster coefficient of 0.919 (**Figure 6**). The functions of the protein p53, a tumor protein, are associated with various expression levels during oncogenesis. GeneMania predicted various valuable functions of the query protein and interacting partners associated with it (**Figure 7**).

Pulldown Strategy and Protein Interaction Prediction for Biomarker Selection

Pull-down assays serve as a complementary method to further validate the predicted interactions in a quantitative manner towards understanding their dissociation constants and relative bindings of proteins and their direct binding sites. However, this is beyond the scope of this study. We believe the following recommendations should be followed by researchers investigating transient protein interactions: First, determining the protein solubility is essential. If the prey protein is at a too-high concentration, it will not be sufficiently soluble. Second, shortening the time and adjusting buffer conditions of incubation help prevent prey protein degradation. Third, checking the prey protein with beads if bait protein is not bound should be done as a control. Fourth, conducting all assays at a constant temperature of 4 °C should be considered if a variation in Kd is found between repeated experiments.

The tumor suppressor antigen p53 depends on specific cellular conditions to induce arrest of cell growth and to control cell division (Pucci et al., 2000; Chen, 2016).

Our network analysis (entry N00170, class nt06164) showed involvement of LANA and other effector markers in KS conditions and helped elucidate their mechanisms of action (**Figure 8**, **Table 4**). Therefore, we suggest that ORF-73 is an important protein that may be a useful biomarker for AIDS-related KS. Studies have suggested a linkage between ORF-73 and host apoptosis through p53 signaling pathways (Tornesello et al., 2018), that could represent a molecular mechanism for the predicted markers associated with KS. Our study discovered KS-associated markers which trigger cancer. ORF-73 encodes LANA-1 virtual proteins of KSHV, linking them with AIDS-associated KS, by their interaction with several cellular processes which include cell apoptosis (through p53) and inhibition of downstream transcriptomic performance. The association between HIV and ORF73 can be inferred by these findings.

DISCUSSION

Many viral genes are homologous to host cellular genes in KSHV (Swanton et al., 1997). The PubMed, Google Scholar, and Scopus searches confirmed the key diagnostic markers for KS based on the available literature. Our computational study on them revealed their importance and evolutionary role in human cancer biology. LANA-1 imparts important immunogenic effects to KSHV, and it specifically interacts with many cellular pathways, including that of cell apoptosis (through its interaction with p53, and repression of downstream transcripts; see **Table 4**). This induces oncogenesis by targeting the protein-E2F transcriptional regulatory pathway (Radkov et al., 2000). The protein homologues identified through our search were structurally different from each other. Therefore, we analyzed selected proteins and compared them using homology searches for the selected domains to prove

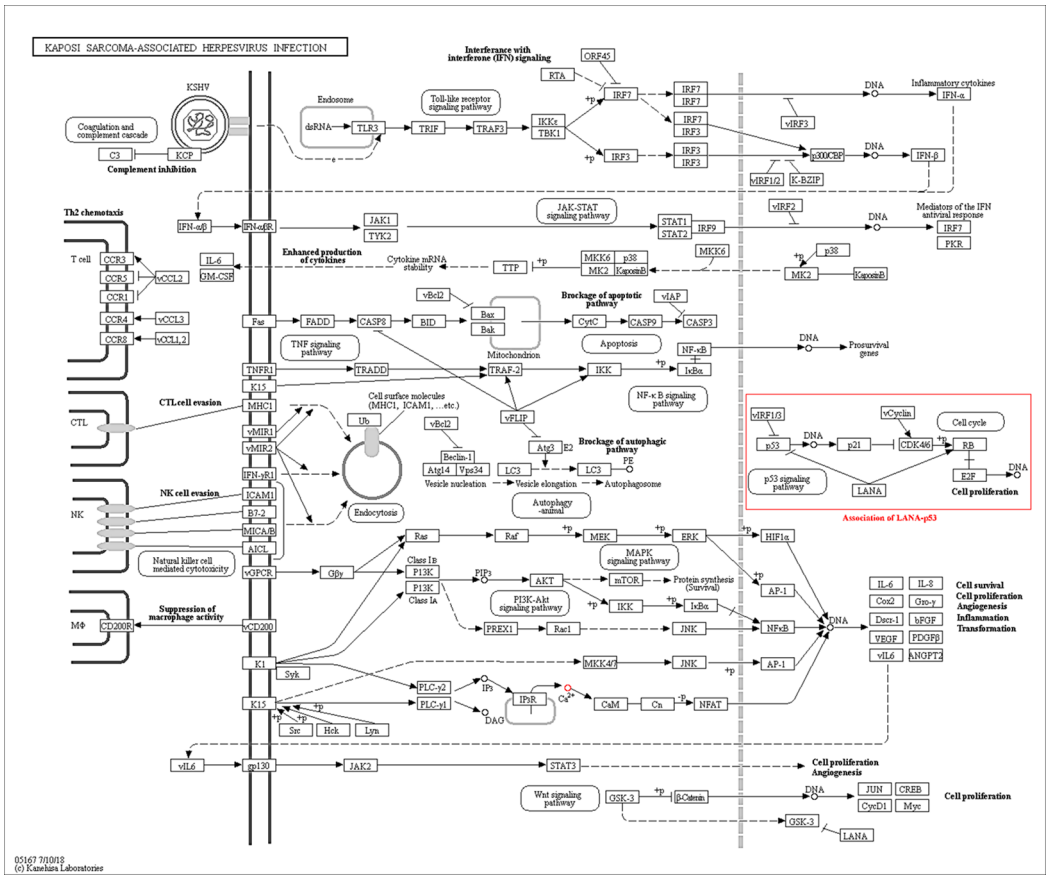


FIGURE 5 | The Kaposi sarcoma-associated herpesvirus infection pathway from KEGG. Reference pathway highlighted using red box shows that LANA is associated with p53 signaling pathway which confirms the predictable role of the ORF-73 protein in the KS associate marker protein.

interactions with other host proteins that trigger and induce cancer in individuals with immunosuppression (Kersse et al., 2011). Hyper mutation and conserved structural sequence similarities help to maintain key aspects of secondary and tertiary structures, which were consistent with the computational analyses in our study (Huang et al., 2002). **Figure 5** shows the KSHV infection pathway from KEGG. We highlighted the reference pathway using a red box that shows that LANA is associated with the p53 signaling pathway. A BLAST homology search confirmed an ORF-73 marker interaction during herpesvirus pathogenesis. The results of STRING and KEGG searches suggested ORF-73 interacts with the host p53.

ORF-73 is not the only protein marker implicated in KS pathology, but much about it remains unknown. It is used as a marker for KSHV; especially, its protein folding and motifs are important for the marker assessment observed in the pattern of structural domains in the selected sequence analyzed with PSI-PRED. The pathogenic interactions in the network-based analysis between LANA and the host p53 suggest that LANA was confirmed by STRING and FUGUE tools. The predicted sequence motifs give detailed interactions that are conserved in the subfamilies of the herpesviruses as discussed in detail on the KEGG pathway with notable mechanisms described in the literature (Schulz, 2000; Direkze and Laman, 2004; Sharma-Walia et al., 2004; Mesri et al., 2010). However, the markers associated with KS need to be incorporated into comprehensive clinical cohort studies, designed using differential protein purification techniques and evidence-based knowledge on protein interactions with bait proteins to develop practical medical applications in the future.

Many PPIs have been elucidated using pull-down assays to map the genomes of many organisms, such as yeast (Valente et al., 2009), *Escherichia coli* (Arifuzzaman et al., 2006) *Caenorhabditis elegans* (Remmelzwaal and Boxem, 2019).

Like all other herpesviruses, KSHV displays latency and a lytic life cycle replication that are characteristic of some viral gene expressions. The genes LANA, v-FLIP, v-cyclin, and Kaposins A, B, and C for latency facilitate the establishment of life in its host and survival against host immune mechanisms. During latency, proteins expressed as K1, K15, vIL6, vGPCR, vIRFs, and vCCLs participate in inflammatory and angiogenic processes evident in KS lesions. Many other lytic and latent viral proteins are involved in the transformation of KSHV host cells into malignant cells. Also, Bcl-2 is one of the major KS progression factors, and TP53 and c-myc have a role in the progression of disease. KS pathology is

TABLE 3 | Functional enrichment of p53.

Sl. No	GO-term	Description	Count in gene set	False discovery rate
Biological process (GO)				
1	GO:0016579	Protein deubiquitination	10 of 275	3.83e–15
2	GO:0007249	I-kappaB kinase/NF-kappaB signaling	8 of 70	3.83e–15
3	GO:0035666	TRIF-dependent toll-like receptor signaling pathway	6 of 24	8.43e–13
4	GO:0051092	Positive regulation of NF-kappaB transcription factor activity	5 of 2142	6.64e–11
5	GO:0070423	Nucleotide-binding oligomerization domain	5 of 27	4.65e–10
Molecular function (GO)				
1	GO:0031625	Ubiquitin protein ligase binding	5 of 311	4.44e–05
2	GO:0042975	Peroxisome proliferator activated receptor binding	2 of 10	0.00062
3	GO:0019899	Enzyme binding	7 of 2197	0.0012
4	GO:0042802	Identical protein binding	6 of 1754	0.0032
5	GO:0032813	Tumor necrosis factor receptor superfamily binding	2 of 46	0.0052
Cellular components (GO)				
1	GO:0043657	Host cell	4 of 29	2.76e–07
2	GO:0030666	Endocytic vesicle membrane	5 of 152	2.90e–07
3	GO:0098805	Whole membrane	8 of 1554	3.85e–06
4	GO:0012506	Vesicle membrane	6 of 743	1.69e–05
5	GO:0005741	Mitochondrial outer membrane	4 of 181	3.05e–05
KEGG pathway				
1	hsa04668	TNF signaling pathway	4 of 108	1.27e–05
2	hsa04064	NF-kappa B signaling pathway	4 of 93	1.27e–05
3	hsa05160	Hepatitis C	4 of 131	1.60e–05
4	hsa04210	Apoptosis	4 of 135	1.60e–05
5	hsa05167	Kaposi's sarcoma-associated herpesvirus infection	4 of 183	3.53e–05
Reactome pathways				
1	HSA-5357956	TNFR1-induced NFkappaB signaling pathway	9 of 30	3.98e–21
2	HSA-5357905	Regulation of TNFR1 signaling	9 of 32	3.98e–21
3	HSA-5689880	Ub-specific processing proteases	10 of 202	1.94e–17
4	HSA-6804757	Regulation of TP53 Degradation	7 of 35	2.30e–15
5	HSA-5675482	Regulation of necroptotic cell death	6 of 17	2.63e–14
UniPort keywords				
1	KW-0832	Ubl conjugation	9 of 2380	1.28e–05
2	KW-0013	ADP-ribosylation	4 of 100	1.28e–05
3	KW-1017	Isopeptide bond	7 of 1713	0.00017
4	KW-0945	Host–virus interaction	4 of 432	0.00094
5	KW-0963	Cytoplasm	9 of 4972	0.0015
PFAM Protein Domains				
1	PF14560	Ubiquitin-like domain	4 of 14	3.12e–09
2	PF11976	Ubiquitin-2 like Rad60 SUMO-like	4 of 21	6.44e–09
3	PF00240	Ubiquitin family	4 of 46	7.76e–08
4	PF02201	SWIB/MDM2 domain	2 of 5	2.86e–05
5	PF00641	Zn-finger in Ran binding protein and others	2 of 16	0.00017
INTERPRO Protein Domains and Features				
1	IPR019956	Ubiquitin	4 of 12	1.83e–09
2	IPR019954	Ubiquitin conserved site	4 of 10	1.83e–09
3	IPR000626	Ubiquitin domain	4 of 57	3.14e–07
4	IPR016495	p53 negative regulator Mdm2/ Mdm4	2 of 2	1.46e–05
5	IPR029071	Ubiquitin-like domain superfamily	4 of 184	1.75e–05
SMART Protein Domains				
1	SM00213	Ubiquitin homologues	4 of 45	6.77e–08
2	SM00005	DEATH domain, found in proteins involved in cell death	2 of 27	0.00035
3	SM00184	Ring finger	3 of 308	0.0012

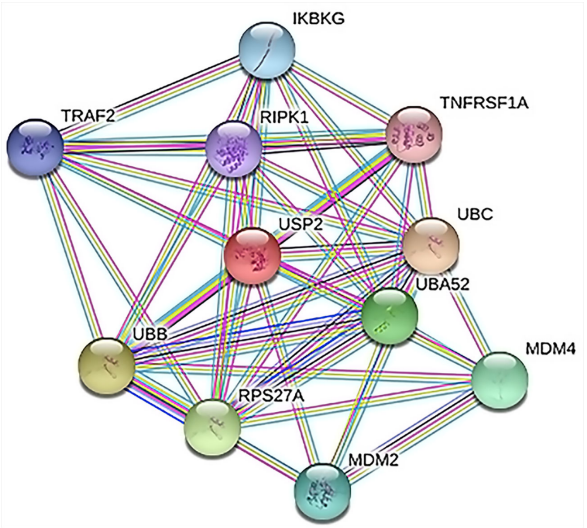


FIGURE 6 | Protein–protein interactions (PPI) between cell arrest marker p53 of cancer cell and Ubiquitin Specific Peptidase 2 (USP2). TNF receptor-associated factor 2 (TRAF2), tumor necrosis factor receptor superfamily member 1A (TNFRSF1A), polyubiquitin-C (UBC), protein Mdm4, E3 ubiquitin-protein ligase Mdm2, ubiquitin-40S ribosomal protein S27a, polyubiquitin-B (UBB), NF-kappa-B essential modulator (IKBKG), receptor-interacting serine/threonine–protein kinase 1 (RIPK1), and Ubiquitin-60S ribosomal protein L40 (UBA52) play important roles in the regulation of cell survival and apoptosis.

interconnected with immune modulation effects such as cell cycle arrest in the host cell, which is required for pathogenic conditions and is mitigated by modulating key factors such as LANA.

Likewise, measuring the expression level and identifying the function of the encoded protein products is important to understand the pathogenesis of KS. We used a methodology similar to that in co-immunoprecipitation (Co-IP) experiments because of our ligand's affinity to capture the strongest interacting proteins (Lapetina and Gil-Henn, 2017). MS identifies subunits and helps explore the structural information associated with the protein of interest (Byrum et al., 2012). Dynamic PPI machines assemble or disassemble the ever-changing inter-, intra-, and extracellular influx cues as a preliminary step towards understanding the structure of proteins and to determine their functions to identify the relevant pathways of interacting proteins (Einarson, 2001; Vikis and Guan, 2004; Einarson et al., 2007). The role and important reason to select ORF-73 in the study is that

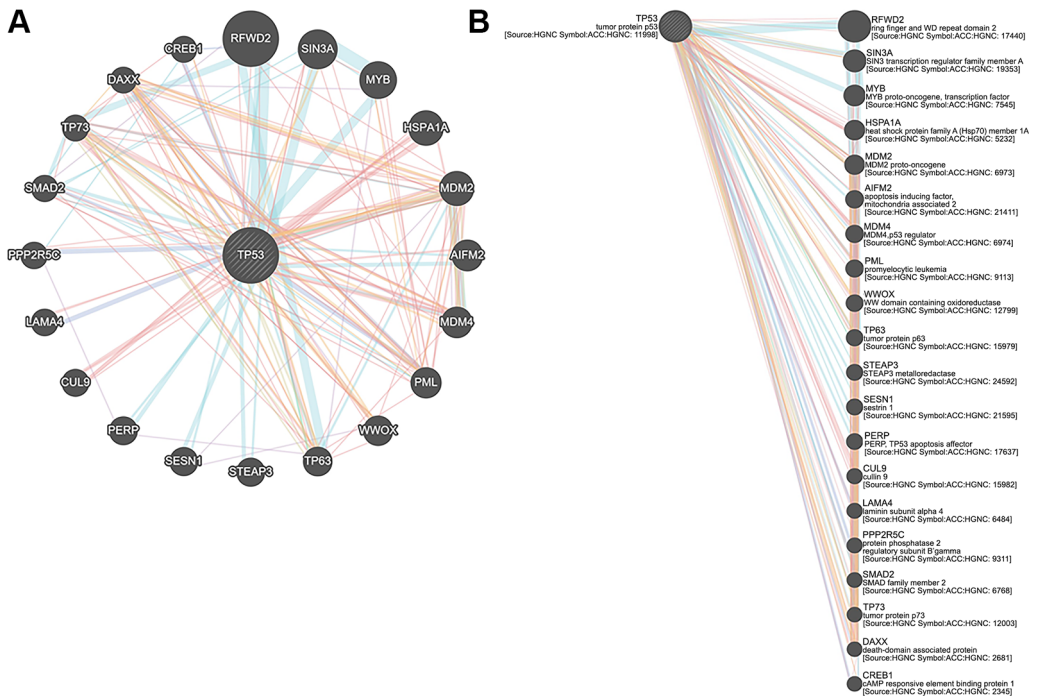


FIGURE 7 | (A) Tumor protein 53 (TP53) network analysis and **(B)** members of the complex pathway and genes with co-expression, co-localization, genetic interactions and specific functions.

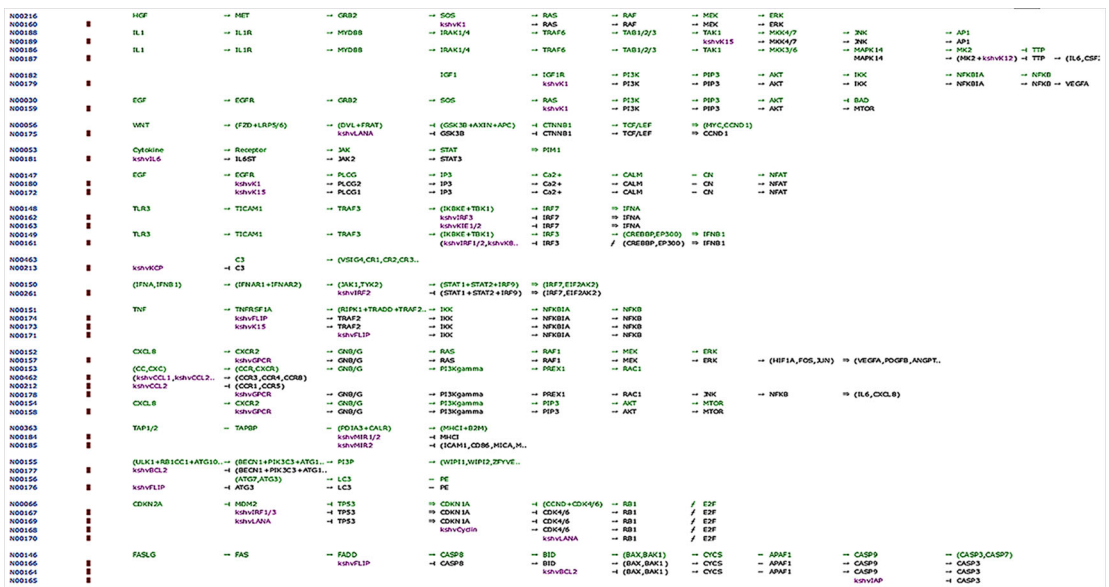


FIGURE 8 | Network map of KEGG for the selected KS protein marker LANA. Protein downstream effect in the cell cycle of disease progression with pooled effectors in cell cycle arrest at G1/S and KS activating mechanisms.

encoding LANA protein distinct domain induces a putative nuclear localization signal (NLS), which product shown interacting with many co-cellular p53, pRb, and ATF4/CREB2. LANA also modulates transcriptional activity of HIV-1 long terminal repeat and to understand the how ORF-73 appears to prevent activity of KS-associated genes was new to know to make

TABLE 4 | Identities of associated markers, downstream signaling candidates, and linked pathways during Kaposi sarcoma pathogenesis.

Sl. No	Entry	Description
1	N00216	HGF-MET-RAS-ERK signaling pathway
2	N00160	KSHV K1 to RAS-ERK signaling pathway
3	N00188	IL1-IL1R-JNK signaling pathway
4	N00189	KSHV K15 to JNK signaling pathway
5	N00186	IL1-IL1R-p38 signaling pathway
6	N00187	KSHV Kaposin B to p38 signaling pathway
7	N00182	IGF-IGFR-PI3K-NFKB signaling pathway
8	N00179	KSHV K1 to PI3K-NFKB signaling pathway
9	N00030	EGF-EGFR-RAS-PI3K signaling pathway
10	N00159	KSHV K1 to PI3K signaling pathway
11	N00056	Wnt signaling pathway
12	N00175	KSHV LANA to Wnt signaling pathway
13	N00053	Cytokine-Jak-STAT signaling pathway
14	N00181	KSHV vIL-6 to Jak-STAT signaling pathway
15	N00147	EGF-EGFR-PLCG-calcineurin signaling pathway
16	N00180	KSHV K1 to PLCG-calcineurin signaling pathway
17	N00172	KSHV K15 to PLCG-calcineurin signaling pathway
18	N00148	TLR3-IRF7 signaling pathway
19	N00162	KSHV vIRF3 to TLR3-IRF7 signaling pathway
20	N00163	KSHV KIE1/2 to TLR3-IRF7 signaling pathway
21	N00149	TLR3-IRF3 signaling pathway
22	N00161	KSHV vIRF1/2 to TLR3-IRF3 signaling pathway
23	N00463	Alternative pathway of complement activation
24	N00213	KSHV Kaposin to alternative pathway of complement activation
25	N00150	Type I IFN signaling pathway
26	N00261	KSHV vIRF2 to IFN signaling pathway
27	N00151	TNF-NFKB signaling pathway
28	N00174	KSHV vFLIP to TNF-NFKB signaling pathway
29	N00173	KSHV K15 to TNF-NFKB signaling pathway
30	N00171	KSHV vFLIP to NFKB signaling pathway
31	N00152	CXCR-GNB/G-ERK signaling pathway
32	N00157	KSHV vGPCR to GNB/G-ERK signaling pathway
33	N00153	CCR/CXCR-GNB/G-PI3K-RAC signaling pathway
34	N00462	KSHV vCCL1/2/3 to CCR signaling pathway
35	N00212	KSHV vCCL2 to CCR signaling pathway
36	N00178	KSHV vGPCR to GNB/G-PI3K-JNK signaling pathway
37	N00154	CXCR-GNB/G-PI3K-AKT signaling pathway
38	N00158	KSHV vGPCR to GNB/G-PI3K-AKT signaling pathway
39	N00363	Antigen processing and presentation by MHC class I molecules
40	N00184	KSHV MIR1/2 to antigen processing and presentation by MHC class I molecules
41	N00185	KSHV MIR2 to cell surface molecule-endocytosis
42	N00155	Autophagy-vesicle nucleation
43	N00177	KSHV vBCL2 to autophagy-vesicle nucleation
44	N00156	Autophagy-vesicle elongation
45	N00176	KSHV vFLIP to autophagy-vesicle elongation
46	N00066	MDM2-p21-Cell cycle G1/S
47	N00167	KSHV vIRF1/3 to p21-cell cycle G1/S
48	N00169	KSHV LANA to p21-cell cycle G1/S
49	N00168	KSHV vCyclin to cell cycle G1/S
50	N00170	KSHV LANA to cell cycle G1/S
51	N00146	Crosstalk between extrinsic and intrinsic apoptotic pathways
52	N00166	KSHV vFLIP to crosstalk between extrinsic and intrinsic apoptotic pathways
53	N00164	KSHV vBCL2 to crosstalk between extrinsic and intrinsic apoptotic pathways
54	N00165	KSHV vIAP to crosstalk between extrinsic and intrinsic apoptotic pathways

preventive strategy (Schäfer et al., 2003). Our findings may help researchers planning cancer prevention strategies, but we used common computational analyses alone, and future studies with expression and interaction analyses should be used to confirm our results and generate treatment options for KS.

CONCLUSION

Our computational studies found that ORF-73 is involved in host apoptosis through p53 signaling pathways and is a key marker associated for Kaposi Sarcoma. This study also identified potential KS-associated genes which are reported to trigger cancer and suggested mechanisms of interaction that may help researcher developing prevention strategies.

ETHICS STATEMENT

We retrieved all data from publicly available resources and we required no ethical approvals for dissemination of this purely academic information.

AUTHOR CONTRIBUTIONS

PZ, JW, and XZ conceived and designed the study. XW, LJ and XG provided study materials and were responsible for the collection and assembly of data, data analysis, and interpretation. PZ was involved in writing of the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the Seed Fund Program of Shanghai University of Medicine and Health Sciences (Grant No. SFP-18-21-01-002), the General Program of Pudong New Area Health and Family Planning Commission of Shanghai, China (Grant No. PW2016A-7), the National Natural Science Foundation of China (No. 81830052), Construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400), Shanghai Municipal Education Commission (Class II Plateau Disciplinary Construction Program for Medical Technology of SUMHS, 2018-2020), and the Natural Science Foundation of Guangdong Province (2016A030313680).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01376/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Flow chart showing the methodology for choosing selective markers for downstream analyses to develop a PPI network.

REFERENCES

Aka, P. V., Kemp, T. J., Rabkin, C. S., Shiels, M. S., Polizzotto, M. N., Lauria, C., et al. (2015). A multiplex panel of plasma markers of immunity and inflammation in classical kaposi sarcoma. *J. Infect. Dis.* 211, 226–229. doi: 10.1093/infdis/jiu410

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., et al. (2006). Large-scale identification of protein-protein interaction of Escherichia coli K-12. *Genome Res.* 16, 686–691. doi: 10.1101/gr.4527806

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., et al. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 220, 868–871. doi: 10.1126/science.6189183

Beral, V., Peterman, T. A., Berkelman, R. L., and Jaffe, H. W. (1990). Kaposi's sarcoma among persons with AIDS: a sexually transmitted infection? *Lancet* 335, 123–128. doi: 10.1016/0140-6736(90)90001-l

Bhutani, M., Polizzotto, M. N., Uldrick, T. S., and Yarchoan, R. (2015). Kaposi sarcoma-associated herpesvirus-associated malignancies: epidemiology, pathogenesis, and advances in treatment. *Semin. Oncol.* 42, 223–246. doi: 10.1053/j.seminoncol.2014.12.027

Bujnicki, J. M., and Rychlewski, L. (2001). The herpesvirus alkaline exonuclease belongs to the restriction endonuclease PD-(D/E)XK superfamily: insight from molecular modeling and phylogenetic analysis. *Virus Genes* 22, 219–230. doi: 10.1023/A:1008131810233

Byrum, S., Smart, S. K., Larson, S., and Tackett, A. J. (2012). Analysis of stable and transient protein-protein interactions. *Methods Mol. Biol.* 833, 143–152. doi: 10.1007/978-1-61779-477-3_10

Chang, Y., Cesarman, E., Pessin, M. S., Lee, F., Culpepper, J., Knowles, D. M., et al. (1994). Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* 266, 1865–1869. doi: 10.1126/science.7997879

Chen, J. (2016). The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression. *Cold Spring HarbPerspect Med.* 6, a026104. doi: 10.1101/cshperspect.a026104

Ciufo, D. M., Cannon, J. S., Poole, L. J., Wu, F. Y., Murray, P., Ambinder, R. F., et al. (2001). Spindle cell conversion by Kaposi's sarcoma-associated herpesvirus: formation of colonies and plaques with mixed lytic and latent gene expression in infected primary dermal microvascular endothelial cell cultures. *J. Virol.* 75, 5614–5626. doi: 10.1128/JVI.75.12.5614-5626.2001

Cottoni, F., Masala, M. V., Santarelli, R., Carcassi, C., Uccini, S., Montesu, M. A., et al. (2004). Susceptibility to human herpesvirus-8 infection in a healthy population from Sardinia is not directly correlated with the expression of HLA-DR alleles. *Br. J. Dermatol.* 151, 247–249. doi: 10.1111/j.1365-2133.2004.06060.x

Delgado, T., Carroll, P. A., Punjabi, A. S., Margineantu, D., Hockenbery, D. M., and Lagunoff, M. (2010). Induction of the Warburg effect by Kaposi's sarcoma herpesvirus is required for the maintenance of latently infected endothelial cells. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10696–10701. doi: 10.1073/pnas.1004882107

Della Bella, S., Taddeo, A., Calabrò, M. L., Brambilla, L., Bellinvia, M., Bergamo, E., et al. (2008). Peripheral blood endothelial progenitors as potential reservoirs of Kaposi's sarcoma-associated herpesvirus. *PloS One* 3, e1520. doi: 10.1371/journal.pone.0001520

Direkze, S., and Laman, H. (2004). Regulation of growth signalling and cell cycle by Kaposi's sarcoma-associated herpesvirus genes. *Int. J. Exp. Pathol.* 85, 305–319. doi: 10.1111/j.0959-9673.2004.00407.x

Dittmer, D. P., and Damania, B. (2019). "Kaposi's Sarcoma-Associated Herpesvirus (KSHV)-Associated Disease in the AIDS Patient: An Update," in *HIV/AIDS-Associated Viral Oncogenesis* (Cham: Springer), 63–80.

Dorak, M. T., Yee, L. J., Tang, J., Shao, W., Lobashevsky, E. S., Jacobson, L. P., et al. (2005). HLA-B, -DRB1/3/4/5, and -DQB1 gene polymorphisms in human immunodeficiency virus-related Kaposi's sarcoma. *J. Med. Virol.* 76, 302–310. doi: 10.1002/jmv.20361

Duus, K. M., Lentchitsky, V., Wagenaar, T., Grose, C., and Webster-Cyriaque, J. (2004). Wild-type Kaposi's sarcoma-associated herpesvirus isolated from the oropharynx of immune-competent individuals has tropism for cultured oral epithelial cells. *J. Virol.* 78, 4074–4084. doi: 10.1128/jvi.78.8.4074-4084.2004

Einarson, M. B., Pugacheva, E. N., and Orlinick, J. R. (2007). Identification of protein-protein interactions with glutathione-S-Transferase (GST) fusion proteins. *CSH Protoc.* 2007. doi: 10.1101/pdb.top11

Einarson, M. (2001). "Detection of Protein-Protein Interactions Using the GST Q21 Fusion Protein Pulldown Technique," in *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Protocols, 2006(1).

Ensser, A., Glykofrydes, D., Niphuis, H., Kuhn, E. M., Rosenwirth, B., Heeney, J. L., et al. (2001). Independence of herpesvirus-induced T cell lymphoma from viral cyclin D homologue. *J. Exp. Med.* 193, 637–642. doi: 10.1084/jem.193.5.637

Fickenscher, H., Biesinger, B., Knappe, A., Wittmann, S., and Fleckenstein, B. (1996). Regulation of the herpesvirus saimiri oncogene stpC, similar to that of T-cell activation genes, in growth-transformed human T lymphocytes. *J. Virol.* 70, 6012–6019.

Flore, O., Rafii, S., Ely, S., O'Leary, J. J., Hyjek, E. M., and Cesarman, E. (1998). Transformation of primary human endothelial cells by Kaposi's sarcoma-associated herpesvirus. *Nature* 394, 588–592. doi: 10.1038/29093

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 41, D808–D815. doi: 10.1093/nar/gks1094

Fukunaga, M. (2005). Expression of D2-40 in lymphatic endothelium of normal tissues and in vascular tumours. *Histopathology* 46, 396–402. doi: 10.1111/j.1365-2559.2005.02098.x

Gallo, R. C. (1998). The enigmas of Kaposi's sarcoma. *Science* 282, 1837–1839. doi: 10.1126/science.282.5395.1837

Gaur, N., Tikla, T., and Kaul, R. (2019). Kaposi sarcoma-associated herpes virus (KSHV) latent protein LANA modulates cellular genes associated with epithelial-to-mesenchymal transition. *Arch. Virol.* 164 (1), 91–104. doi: 10.1007/s00705-018-4060-y

Gazouli, M., Zavos, G., Papaconstantinou, I., Lukas, J. C., Zografidis, A., Boletis, J., et al. (2004). The interleukin-6-174 promoter polymorphism is associated with a risk of development of Kaposi's sarcoma in renal transplant recipients. *Anticancer Res.* 24, 1311–1314.

Gelmann, E. P., Popovic, M., Blayney, D., Masur, H., Sidhu, G., Stahl, R. E., et al. (1983). Proviral DNA of a retrovirus, human T-cell leukemia virus, in two patients with AIDS. *Science* 220, 862–865. doi: 10.1126/science.6601822

Glykofrydes, D., Niphuis, H., Kuhn, E. M., Rosenwirth, B., Heeney, J. L., Bruder, J., et al. (2000). Herpesvirus saimiri vFLIP provides an antiapoptotic function but is not essential for viral replication, transformation, or pathogenicity. *J. Virol.* 74, 11919–11927. doi: 10.1128/jvi.74.24.11919-11927.2000

Gregory, S. M., Davis, B. K., West, J. A., Taxman, D. J., Matsuzawa, S., Reed, J. C., et al. (2011). Discovery of a viral NLR homolog that inhibits the inflammasome. *Science* 331, 330–334. doi: 10.1126/science.1199478

Hou, C., Wang, F., Liu, X., Chang, G., Wang, F., and Geng, X. (2017). Comprehensive analysis of interaction networks of telomerase reverse transcriptase with multiple bioinformatic approaches: deep mining the potential functions of telomere and telomerase. *Rejuvenation Res.* 20, 320–333. doi: 10.1089/rej.2016.1909

Huang, Q., Petros, A. M., Virgin, H. W., Fesik, S. W., and Olejniczak, E. T. (2002). Solution structure of a Bcl-2 homolog from Kaposi sarcoma virus. *Proc. Natl. Acad. Sci. U.S.A.* 99, 3428–3433. doi: 10.1073/pnas.062525799

Hunt, S., Santa Cruz, D., and Barnhill, R. (2004). *Vascular tumors*.

Jones, D. T., and Cozzetto, D. (2015). DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31, 857–863. doi: 10.1093/bioinformatics/btu744

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202. doi: 10.1006/jmbi.1999.3091

Kanashiro, C. A., Schally, A. V., Groot, K., Armatis, P., Bernardino, A. L. F., and Varga, J. L. (2003). Inhibition of mutant p53 expression and growth of DMS-153 small cell lung carcinoma by antagonists of growth hormone-releasing hormone and bombesin. *Proc. Natl. Acad. Sci. U.S.A.* 100, 15836–15841. doi: 10.1073/pnas.2536558100

Kaposi (1872). Idiopathisches multiples Pigmentsarkom der Haut. *Arch. f. Dermat.* 4, 265–273. doi: 10.1007/BF01830024

Kersse, K., Verspurten, J., Vanden Berghe, T., and Vandenabeele, P. (2011). The death-fold superfamily of homotypic interaction motifs. *Trends Biochem. Sci.* 36, 541–552. doi: 10.1016/j.tibs.2011.06.006

La Ferla, L., Pinzone, M. R., Nunnari, G., Martellotta, F., Lleshi, A., Tirelli, U., et al. (2013). Kaposi' s sarcoma in HIV-positive patients: the state of art in the HAART-era. *Eur. Rev. Med. Pharmacol. Sci.* 17, 2354–2365.

Lapetina, S., and Gil-Henn, H. (2017). A guide to simple, direct, and quantitative *in vitro* binding assays. *J. Biol. Methods* 4, e62. doi: 10.14440/jbm.2017.161

Liang, C., Lee, J.-S., and Jung, J. U. (2008). Immune evasion in Kaposi's sarcoma-associated herpes virus associated oncogenesis. *Semin. Cancer Biol.* 18, 423–436. doi: 10.1016/j.semcancer.2008.09.003

Majerciak, V., Pripuzova, N., Chan, C., Temkin, N., Specht, S. I., and Zheng, Z.-M. (2015). Stability of structured Kaposi's sarcoma-associated herpesvirus ORF57 protein is regulated by protein phosphorylation and homodimerization. *J. Virol.* 89, 3256–3274. doi: 10.1128/JVI.03721-14

Martin, J. N., Ganem, D. E., Osmond, D. H., Page-Shafer, K. A., Macrae, D., and Kedes, D. H. (1998). Sexual transmission and the natural history of human herpesvirus 8 infection. *N. Engl. J. Med.* 338, 948–954. doi: 10.1056/NEJM199804023381403

Mesri, E. A., Cesarman, E., and Boshoff, C. (2010). Kaposi's sarcoma and its associated herpesvirus. *Nat. Rev. Cancer* 10, 707–719. doi: 10.1038/nrc2888

Moses, A. V., Fish, K. N., Ruhl, R., Smith, P. P., Strussenberg, J. G., Zhu, L., et al. (1999). Long-term infection and transformation of dermal microvascular endothelial cells by human herpesvirus 8. *J. Virol.* 73, 6892–6902.

Nguyen, H. Q., Magaret, A. S., Kitahata, M. M., Van Rompaey, S. E., Wald, A., and Casper, C. (2008). Persistent Kaposi sarcoma in the era of HAART: characterizing the predictors of clinical response. *AIDS* 22, 937–945. doi: 10.1097/QAD.0b013e3282ff6275

Parkin, D. M. (2006). The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* 118, 3030–3044. doi: 10.1002/ijc.21731

Pucci, B., Kasten, M., and Giordano, A. (2000). Cell cycle and apoptosis. *Neoplasia* 2, 291–299. doi: 10.1038/sj.neo.7900101

Radkov, S. A., Kellam, P., and Boshoff, C. (2000). The latent nuclear antigen of Kaposi sarcoma-associated herpesvirus targets the retinoblastoma-E2F pathway and with the oncogene Hras transforms primary rat cells. *Nat. Med.* 6, 1121–1127. doi: 10.1038/80459

Remmelzwaal, S., and Boxem, M. (2019). Protein interactome mapping in Caenorhabditiselegans. *Curr. Opin. Syst. Biol.* 13, 1–9. doi: 10.1016/j.coisb.2018.08.006

Roizman, B., Carmichael, L. E., Deinhardt, F., de-The, G., Nahmias, A. J., Plowright, W., et al. (1981). Herpesviridae. definition, provisional nomenclature, and taxonomy. The Herpesvirus study group, the international committee on taxonomy of viruses. *Intervirology* 16, 201–217. doi: 10.1159/000149269

Russo, J. J., Bohenzky, R. A., Chien, M.-C., Chen, J., Yan, M., Maddalena, D., et al. (1996). Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *PNAS* 93, 14862–14867. doi: 10.1073/pnas.93.25.14862

Schäfer, A., Lengenfelder, D., Grillhösl, C., Wieser, C., Fleckenstein, B., and Ensser, A. (2003). The latency-associated nuclear antigen homolog of herpesvirus saimiri inhibits lytic virus replication. *J. Virol.* 77, 5911–5925. doi: 10.1128/jvi.77.10.5911-5925.2003

Schulz, T. F. (2000). Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8): epidemiology and pathogenesis. *J. Antimicrob. Chemother.* 45 Suppl T3, 15–27. doi: 10.1093/jac/45.suppl_4.15

Sharma-Walia, N., Naranatt, P. P., Krishnan, H. H., Zeng, L., and Chandran, B. (2004). Kaposi's sarcoma-associated herpesvirus/human herpesvirus 8 envelope glycoprotein gB induces the integrin-dependent focal adhesion kinase-Src-phosphatidylinositol 3-kinase-rho GTPase signal pathways and cytoskeletal rearrangements. *J. Virol.* 78, 4207–4223. doi: 10.1128/jvi.78.8.4207-4223.2004

Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310, 243–257. doi: 10.1006/jmbi.2001.4762

Sinfield, R. L., Molyneux, E. M., Banda, K., Borgstein, E., Broadhead, R., Hesseling, P., et al. (2007). Spectrum and presentation of pediatric malignancies in the HIV era: experience from Blantyre, Malawi, 1998-2003. *Pediatr. Blood Cancer* 48, 515–520. doi: 10.1002/pbc.20917

Swanton, C., Mann, D. J., Fleckenstein, B., Neipel, F., Peters, G., and Jones, N. (1997). Herpes viral cyclin/Cdk6 complexes evade inhibition by CDK inhibitor proteins. *Nature* 390, 184–187. doi: 10.1038/36606

Tang, S., Li, T., Cong, P., Xiong, W., Wang, Z., and Sun, J. (2013). PlantLoc: an accurate web server for predicting plant protein subcellular localization by substantiality motif. *Nucleic Acids Res.* 41, W441–W447. doi: 10.1093/nar/gkt428

Thome, M., Schneider, P., Hofmann, K., Fickenscher, H., Meinl, E., Neipel, F., et al. (1997). Viral FLICE-inhibitory proteins (FLIPs) prevent apoptosis induced by death receptors. *Nature* 386, 517–521. doi: 10.1038/386517a0

Tornesello, M. L., Annunziata, C., Tornesello, A. L., Buonaguro, L., and Buonaguro, F. M. (2018). Human Oncoviruses and p53 Tumor Suppressor Pathway Deregulation at the Origin of Human Cancers. *Cancers (Basel)* 10. doi: 10.3390/cancers10070213

Valente, A. X. C. N., Roberts, S. B., Buck, G. A., and Gao, Y. (2009). Functional organization of the yeast proteome by a yeast interactome map. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1490–1495. doi: 10.1073/pnas.0808624106

Vikis, H. G., and Guan, K.-L. (2004). Glutathione-S-transferase-fusion based assays for studying protein-protein interactions. *Methods Mol. Biol.* 261, 175–186. doi: 10.1385/1-59259-762-9:175

Wass, M. N., Kelley, L. A., and Sternberg, M. J. E. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* 38, W469–W473. doi: 10.1093/nar/gkq406

Weeden, D. (2002). *Vascular tumor In Skin Pathology*.

Woodberry, T., Suscovich, T. J., Henry, L. M., Martin, J. N., Dollard, S., O'Connor, P. G., et al. (2005). Impact of Kaposi sarcoma-associated herpesvirus (KSHV) burden and HIV coinfection on the detection of T cell responses to KSHV ORF73 and ORF65 proteins. *J. Infect. Dis.* 192, 622–629. doi: 10.1086/432103

Yan, Q., Li, W., Tang, Q., Yao, S., Lv, Z., Feng, N., et al. (2013). Cellular microRNAs 498 and 320d regulate herpes simplex virus 1 induction of Kaposi's sarcoma-associated herpesvirus lytic replication by targeting RTA. *PloS One* 8, e55832. doi: 10.1371/journal.pone.0055832

Yarchoan, R., Uldrick, T., and Polizzotto, M. (2015). "117 - HIV-associated malignancies," in *DeVita, Hellman, and Rosenberg"s Cancer: Principles and Practice of Oncology*. Lippincott Williams and Wilkins.

Yee-Lin, V., Pooi-Fong, W., and Soo-Beng, A. K. (2018). Nutlin-3, A p53-Mdm2 Antagonist for Nasopharyngeal Carcinoma Treatment. *Mini Rev Med Chem* 18, 173–183. doi: 10.2174/1389557517666170717125821

Zhu, F. X., Chong, J. M., Wu, L., and Yuan, Y. (2005). Virion proteins of Kaposi's sarcoma-associated herpesvirus. *J. Virol.* 79, 800–811. doi: 10.1128/JVI.79.2.800-811.2005

zur Hausen, H. (2001). Oncogenic DNA viruses. *Oncogene* 20, 7820–7823. doi: 10.1038/sj.onc.1204958

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fgene.2019.01357

A Bipartite Network Module-Based Project to Predict Pathogen–Host Association

Jie Li, Shiming Wang, Zhuo Chen and Yadong Wang*

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Pathogen–host interactions play an important role in understanding the mechanism by which a pathogen can infect its host. Some approaches for predicting pathogen–host association have been developed, but prediction accuracy is still low. In this paper, we propose a bipartite network module-based approach to improve prediction accuracy. First, a bipartite network with pathogens and hosts is constructed. Next, pathogens and hosts are divided into different modules respectively. Then, modular information on the pathogens and hosts is added into a bipartite network projection model and the association scores between pathogens and hosts are calculated. Finally, leave-one-out cross-validation is used to estimate the performance of the proposed method. Experimental results show that the proposed method performs better in predicting pathogen–host association than other methods, and some potential pathogen–host associations with higher prediction scores are also confirmed by the results of biological experiments in the publically available literature.

Keywords: BNMP, bipartite network project, pathogen, host, pathogen–host association

OPEN ACCESS

Edited by:
Huiru Zheng,
Ulster University, United Kingdom

Reviewed by:
Alneu De Andrade Lopes,
Universidade de São Paulo São Carlos, Brazil
Stephen Beckett,
Georgia Institute of Technology, United States

****Correspondence:***
Jie Li
jieli@hit.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *27 September 2019*
Accepted: *11 December 2019*
Published: *24 January 2020*

Citation:
Li J, Wang S, Chen Z and Wang Y (2020) A Bipartite Network Module-Based Project to Predict Pathogen–Host Association. Front. Genet. 10:1357. doi: 10.3389/fgene.2019.01357

INTRODUCTION

Pathogen–host interactions (PHIs) play a crucial role in understanding the mechanisms of infections and identifying potential targets for infection therapeutics. Therefore, various biological experimental or computing methods have been developed to test and predict the interactions between pathogens and hosts. However, it is not only time-consuming and laborious to test PHIs through biological experimentation but also costs a lot of money. Computing methods such as biological reasoning and machine learning are considered as another important approach for predicting PHIs. Three main approaches can be used to predict PHIs: biological reasoning homology-based, structure-based, and domain/motif interaction-based (Nourani et al., 2015). The basis of homology-based prediction is that the interaction between conserved homologous organisms would also be conserved. Lee et al. inferred more than 3000 H. sapiens–P. falciparum protein–protein interactions (PPIs) based on orthologous pairs, revealing that Plasmodium falciparum can utilize calcium regulatory proteins in host cells to maintain Ca2+ levels (Lee et al., 2008). Wuchty et al. used the random forest method to evaluate and filter homology-based prediction results, which further improved prediction accuracy (Stefan, 2011). Structure-based prediction assumes that a pair of proteins with similar protein structures that are known to interact may interact in the same manner. Davis et al. proposed an algorithm for predicting possible

interactions based on the physical structure of the protein by scanning the genome of the pathogen and host to find structurally similar proteins (Davis et al., 2010). Aloy and Russell also proposed a method for inferring the molecular details of interactions that might occur by evaluating a pair of potentially interacting proteins on a complex of known 3D structures (Patrick and Russell, 2002). Doolittle et al. used this method to predict the interaction between HIV and human proteins, providing assistance for further trials and therapeutic intervention targets (Doolittle and Gomez, 2010). Domain/motif interaction-based prediction combines the known intraspecific PPI with the protein domain spectrum to predict the PPI between host and pathogen proteins (Dyer et al., 2007). Evans et al. used the method to predict the interaction between HIV-1 and human proteins, confirming that the linear binding motif shared by the virus and the host protein was an important part of the crosstalk between the virus and the host (Evans et al., 2009). Machine learning methods are widely used in the prediction of pathogen–host interaction relationships. Ahmed et al. used a comparison of a neural network model versus SVM for the prediction of host-pathogen PPI based on a combination of features including amino acid quadruplets, pairwise sequence similarity, and human interactome properties; they found that the neural network achieved a significant improvement in overall performance compared to a predictor using the triplets feature and that it achieved good accuracy in predicting B.anthracis–human interaction (Ahmed et al., 2018). Mei et al. proposed the AdaBoost approach to predict proteome-wide interactions between Salmonella and human proteins based on multi-instance transfer learning (Mei and Zhu, 2014). Subsequently, a new negative data sampling method based on single-class SVM was proposed to predict the protein interaction between HTLV retrovirus and Homo sapiens. Use of this method provided valuable cues for the pathogenesis of HTLV retrovirus (Mei and Zhu, 2015).

Predicting unknown relations between pathogens and hosts in advance is of great significance for detecting changes in their relations and preventing the spread of infectious diseases in hosts. The above methods are used to predict protein–protein interactions of pathogens and hosts based on protein-related information. However, in cases where protein information or other molecular information is unavailable and we only know the relations between pathogens and hosts, we need to develop a new method to predict the potential relations between pathogens and hosts based only on the relations of pathogens and hosts. Zhang et al. developed a bipartite network project (BNP) (Zhou et al., 2007) to predict the relations between an X set and Y set (two sets included in the bipartite network). The experimental results on personal recommendation shown that BNP performed much better than the most commonly used global ranking method. Chen et al. proposed a novel computational model of Bipartite Network Projection for MiRNA–Disease Association prediction (BNPMDA) (Chen et al., 2018) based on the known miRNA–disease associations, integrated miRNA similarity, and integrated disease similarity. BNPMDA could effectively predict the potential miRNA–disease associations with a high accuracy level. Sun et al. developed the NTSMDA method to predict miRNA–disease associations by integrating network topological similarity (Sun et al., 2016). NTSMDA demonstrates excellent predictive performance. Tad et al. developed an algorithm to predict missing links based on conditional probability estimation and associated, node-level features (Dallas et al., 2017). They validated this algorithm on simulated data and then applied it to a desert small mammal host-parasite network. The approach achieved high accuracy on simulated and observed data, providing a simple method for accurately predicting missing links in networks without relying on prior knowledge about the network structure. These methods are based on bipartite network models and are widely used in different fields. However, these methods not only ignore the relations of elements in the X set but also the relations of elements in the Y set, though these relations are important to predict the relations of the X set and Y set. Zhang et al. proposed a weight-based model (Zhang et al., 2015) in a dual-layer network, using the cell line similarity network, drug similarity network, and drug-cell line response network. WBSMDA (Chen et al., 2016a) employed the concepts of within-score and between-score to predict the association score in the association network. These methods consider the relations of elements in the X and Y sets from a global perspective, and collecting the information from a local perspective and then integrating them from the global perspective can detect the information in the network more comprehensively. Based on this idea, we proposed a bipartite network module-based project (BNMP) to predict pathogen–host associations by adding modular information into a bipartite network projection. Firstly, a pathogen–host bipartite network is constructed, and the distances of pathogens and hosts are computed respectively on the basis of the topological structure. Pathogens are then divided into several modules, as are hosts. Finally, the module information of pathogens and hosts, respectively, is applied to BNP to calculate the prediction score.

MATERIALS AND METHODS

Data Collection and Pre-Processing

First, the pathogen–host interaction data were downloaded from PHI-base (Urban et al., 2017) (http://www.PHI-base.org/index.jsp), HPIDB (Ammari et al., 2016) (https://hpidb.igbb.msstate.edu/index.html), and IntAct (Sandra et al., 2014) (https://www.ebi.ac.uk/intact/). These three databases are commonly used molecular interaction databases that cover most of the molecular interaction data in open data sources. We downloaded all of the entire datasets of these three databases on September 8, 2019. These three databases provide downloads of previous version data, and researchers can select the related version for replication. Then, based on the taxonomy ID, we selected bacteria–host interaction data and deleted duplicate data from the data sets. The final dataset comprised data on 997 bacteria–host interactions, covering 243 hosts and 388 bacteria. The number of pathogens and hosts were s and t, respectively. We used them to generate the pathogen–host association matrix A. $A[\ p_i\ ][\ h_j\ ]=1$ means that there is a pathogen–host protein–

protein interaction between the *i*th pathogen and the *j*th host, whereas $A[\,p_i\,][\,h_j\,]=0$ means there is no interaction between the *i*th pathogen and the *j*th host.

Bipartite Network Projection

Here, for a bipartite network $G(P,H,E)$ where $P=\{\,p_1,p_2,\ldots,p_s\,\}$ and $H=\{\,h_1,h_2,\ldots,h_t\,\}$ are pathogen and host sets respectively, $E\subseteq P\times H$ is the edge set between pathogens and hosts, and the association scores between a host and all pathogens can be calculated using the bipartite network projection (Zhou et al., 2007) (BNP) method. If we let a host h_{seed} be the seed vertex, the association scores between h_{seed} and all pathogens are as follows.

$$BNP(P,H,h_{seed}) = \{scp(p_1), scp(p_2), \ldots, scp(p_s)\}$$
$$scp(p_i) = \sum_{j=1}^{t} A[p_i]\left[h_j\right] sch\left(h_j\right)/d\left(h_j\right)$$
$$sch\left(h_j\right) = \sum_{i=1}^{s} A[p_i]\left[h_j\right] A[p_i][h_{seed}]/d(p_i)$$

where $d(h_j)$ and $d(p_i)$ are the degrees of the *j*th host and the *i*th pathogen, respectively. $scp(p_i)$ is the association score between h_{seed} and the *i*th pathogen, which requires $sch(h_1)$, $sch(h_2)$, …, $sch(h_t)$ as the input.

Bipartite Network Module-Based Project

For $G(P, H, E)$ with s pathogens and t hosts, BNMP comprises the following steps (**Figure 1**):

1) Let a host h_{seed} be the seed vertex. Calculate the distance between two pathogens. $Dis(p_i,p_j)$ is the distance between pathogen p_i and p_j in the following formula (**Figure 1A**), where $A[\,p_i\,]$ is the binary vector in the *i*th row in association matrix A.

$$Dis\left(p_i,p_j\right) = 1 - \exp\left(-||A[p_i] - A\left[p_j\right]||^2\right)$$

2) Divide pathogen set $P=\{\,p_1,p_2,\ldots,p_s\,\}$ into m modules $\{M_1, M_2,\ldots,M_m\}$ with $s_1,s_2,\ldots$, and s_m pathogens, respectively (**Figure 1B**) where m is the degree of h_{seed}, namely the number of pathogens associated with h_{seed}, as expressed in the following formula. The intersection between two modules is empty. So $s = \sum_{l=1}^{m} s_l$, $M_l = \{p_r^l | p_r^l \in P,\ 1 \le r \le s_l\}$.

$$\mathrm{m} = \sum_{i=1}^{s} A[p_i][h_{seed}]$$

The process of generating m modules is as follows: (1) m pathogens associated with h_{seed} are divided into m modules respectively and marked as the core vertexes of the corresponding m modules; (2) p_i (i=1,2,…,s) is added to the module whose core vertex has the shortest distance from it; (3) In order to keep a balance of resources received by the h_{seed} from different modules, select $s_l-\lceil s/m\rceil$ ($\lceil s/m\rceil$ means the rounded-up value of the result of s/m) pathogens with the furthest distance from the core vertex of M_l if s_l is larger than $\lceil s/m\rceil$ and reassign them to other modules in which the number of pathogens is less than $\lceil s/m\rceil$. (4) Repeat (3) until the number of pathogens in each module does not exceed $\lceil s/m\rceil$.

3) Calculate the association score set $\overline{score_{M_l}}$ between h_{seed} and $M_l(l=1, 2,\ldots,m)$ (**Figure 1C**).

$$\overline{score_{M_l}} = \frac{\sum_{1\le j\le m, j\ne l} w\left(M_l, M_j\right) \times B_{M_l}}{\sum_{1\le j\le m, j\ne l} w\left(M_l, M_j\right)}$$

where

$$w\left(M_l, M_j\right) = \exp\left(-\frac{\sum_{p_u\in M_l}\sum_{p_v\in M_j} Dis(p_u,p_v)}{|M_l| \times |M_j|}\right)$$
$$B_{M_l,M_j} = BNP\left(\overline{M_{lj}}, \overline{H_{lj}}, h_{seed}\right)$$
$$B_{M_l,M_j} = B_{M_l} \cup B_{M_j}$$
$$\overline{M_{lj}} = M_l \cup M_j$$
$$\overline{H_{lj}} = \{h_n | A[p_k][h_n] = 1, p_k \in \overline{M_{lj}}, 1 \le n \le t\}$$

$w(M_l,M_j)$ is the weight coefficient of resources that M_l receive from M_j ($j\ne l$). B_{M_l,M_j} is the association score set obtained by running the BNP algorithm on $\overline{M_{lj}}$, $\overline{H_{lj}}$, and h_{seed}, which includes two sets: B_{M_l} and B_{M_j}. B_{M_l} and B_{M_j} are the association score sets of pathogens in B_{M_l} and B_{M_j}, respectively.

Finally, the association score set $\{\overline{score_{M_1}}, \overline{score_{M_2}}, \ldots, \overline{score_{M_m}}\}$ between h_{seed} and all pathogens is obtained.

4) Select each host as the seed vertex in turn, and repeat the process above. Obtain r association score sets, and combine them to form a pathogen and host association score matrix $S_{pathogen-host}$ (**Figure 1D**). Each element of $S_{pathogen-host}$ is an association score of a pathogen and a host. Similarly, chose a pathogen as the seed vertex in turn, and obtain another association score matrix, $S_{host-pathogen}$ (**Figures 1E–H**).

5) Finally, take the integrated value of the two matrices, $S_{pathogen-host}$ and $S^T_{host-pathogen}$, as the association score matrix between pathogens and hosts, where x is a parameter to balance $S_{pathogen-host}$ and $S^T_{host-pathogen}$ (**Figure 1I**):

$$\mathrm{S} = x \times S_{pathogen-host} + (1-x) \times S^T_{host-pathogen}$$

RESULTS

Performance Evaluation

Leave-one-out cross-validation (Kohavi, 1995) (LOOCV) is used to evaluate the performance of BNMP relative to previous evaluation methods (Geeleher et al., 2014; Zhang et al., 2015; Chen et al., 2016b; Sun et al., 2016 Fei et al., 2018; Le and Pham, 2018). Specifically, each known pathogen–host interaction is chosen as a test data set in turn, the remaining known interactions are chosen as the training set, and the pathogen–host association score in the training set is calculated using BNMP. After the LOOCV test process is completed, we plot the receiver operating characteristic (ROC) curve and precision recall (PR) curve and use the area under the ROC curve (AUROC) and the area under the PR curve (AUPR) to evaluate the performance of BNMP.

Performance Analysis of BNMP

We constructed the pathogen–host association network, namely network 1, which consists of 388 pathogens, 243 hosts, and 997 associations, as shown in **Table 1**. To clarify the influence of the balance parameter x, AUROC and AUPR values were calculated with different values of x, as shown in **Figures 2A** and **B**. It can be found that the prediction performance with x, $\in (0, 1)$ is better than with $x = 0$ or $x = 1$, demonstrating the effectiveness of the

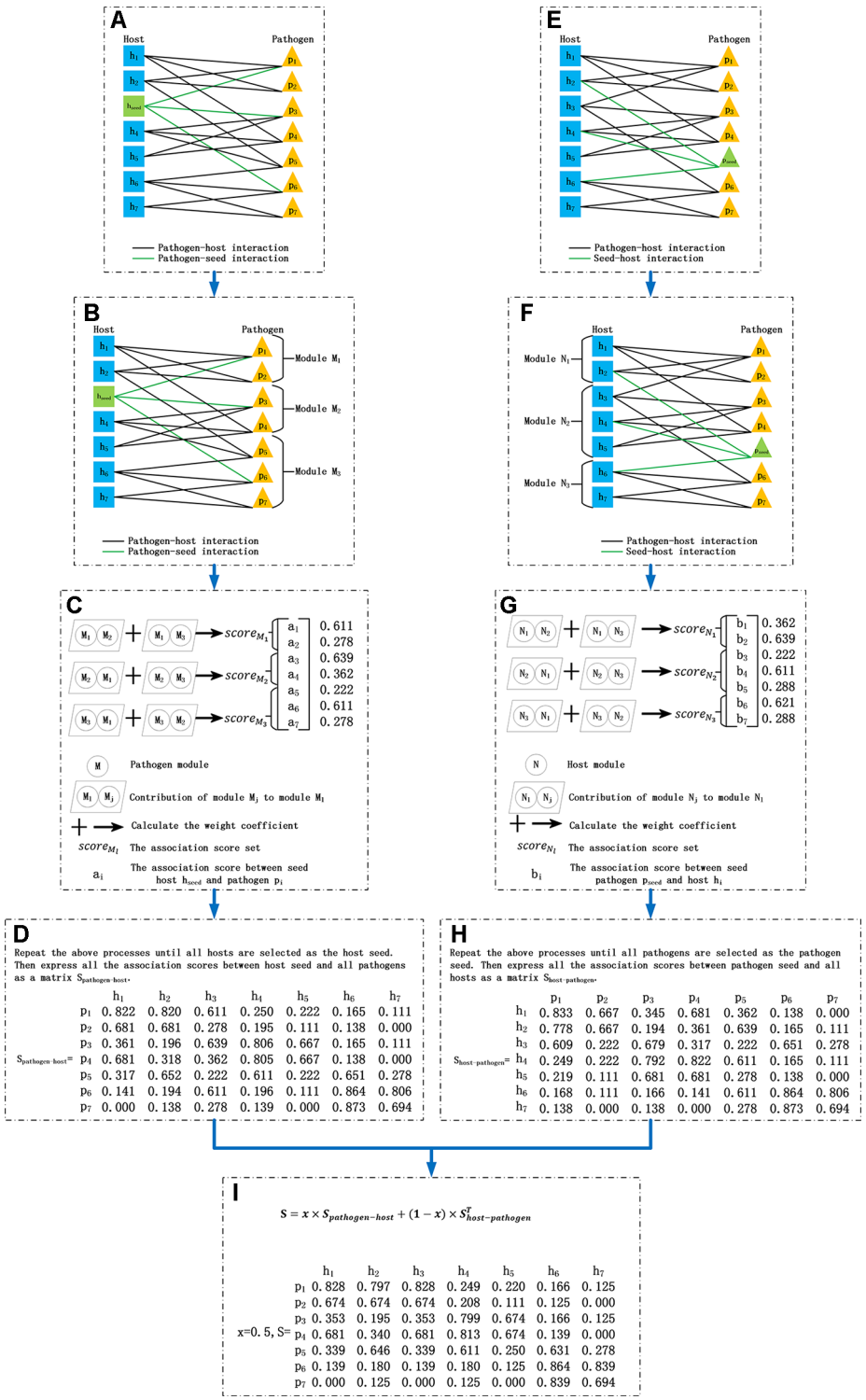


FIGURE 1 | Process of the bipartite network module-based project. **(A)** Construct the pathogen–host bipartite network and choose a host as the seed vertex. **(B)** Divide the pathogen set into several modules. **(C)** Calculate the association score between the seed and pathogens in each module. **(D)** Select each host as the seed vertex in turn and repeat process **(A–C)** then obtain the pathogen–host association score matrix $S_{pathogen-host}$ **(E)** Choose a pathogen as the seed vertex. **(F)** Divide the host set into several modules. **(G)** Calculate the association score between the seed and hosts in each module. **(H)** Select each pathogen as the seed vertex in turn and repeat process **(E–G)** then obtain the host–pathogen association score matrix $S_{host-pathogen}$. **(I)** Integrate matrix $S_{pathogen-host}$ and $S_{host-pathogen}$ as the association score matrix between all pathogens and hosts.

TABLE 1 | The constructed network 1 and network 2.

Network	Number of pathogens	Number of hosts	Number of associations
Network 1	388	243	997
Network 2	167	96	653

integrated association score matrix. When x = 0.575, BNMP acquires the highest AUROC and AUPR values. We plotted the ROC and PR curves when x = 0,0.575, and 1, as shown in **Figures 2C** and **D**. It is noteworthy that the ROC curves take the form of an oblique upward-sloping straight line. We analyzed the results and found that more than half of the hosts are related to only one pathogen. As a result, the association scores between these hosts and pathogens are predicted to be zero in the LOOCV experiment, which has little worth for our prediction and results in the oblique upward-sloping straight line rather than a smooth ROC curve. To evaluate the prediction accuracy of BNMP on hosts (pathogens) that have more than one association with pathogens (hosts), the rows or columns with only one "1" are removed from the pathogen–host association matrix. After processing, 167 pathogens, 96 hosts, and 653 associations remained, namely network 2, and this was used to evaluate the performance of BNMP, as shown in **Table 1**. The analysis regarding x is shown in **Figures 3A** and **B**. When x = 0.675, BNMP achieves the highest AUROC value of 0.8656. When x = 0.825, BNMP achieves the highest AUPR value of 0.4318.

Comparison With Existing Methods

In order to further prove the effectiveness of the proposed method, BNMP is compared with four other methods : Zhang's method (Zhang et al., 2015), NTSMDA (Sun et al., 2016), WBSMDA (Chen et al., 2016a), and BNP (Zhou et al., 2007). BNMP has different prediction performance when x is different (see **Figure 3**). To ensure the fairness of the comparison, we did not select the best prediction performance of BNMP for comparison with the other four methods. Instead, we ranked the AUROC values in **Figure 3A** in descending order and selected the upper quartile (the corresponding x value is 0.8) for comparison with other methods. LOOCV experiments were performed with BNMP, Zhang's method, NTSMDA, WBSMDA, and BNP, and the resulting ROC and PR curves are shown in **Figure 4**. BNMP acquires an AUROC value of 0.8645, exceeding those of NTSMDA (0.8376), BNP (0.8352), Zhang's method (0.7807), and WBSMDA (0.7592). Meanwhile, BNMP obtains an AUPR value of 0.4315, exceeding those of NTSMDA (0.3729), WBSMDA (0.3254), Zhang's method (0.2644), and BNP (0.201). We also calculated the AUROC and AUPR values for each pathogen by these methods, and performed a paired t-test (Demišar and Schuurmans, 2006) between BNMP and the other methods (see **Figure 5**). The result is that all the p-values

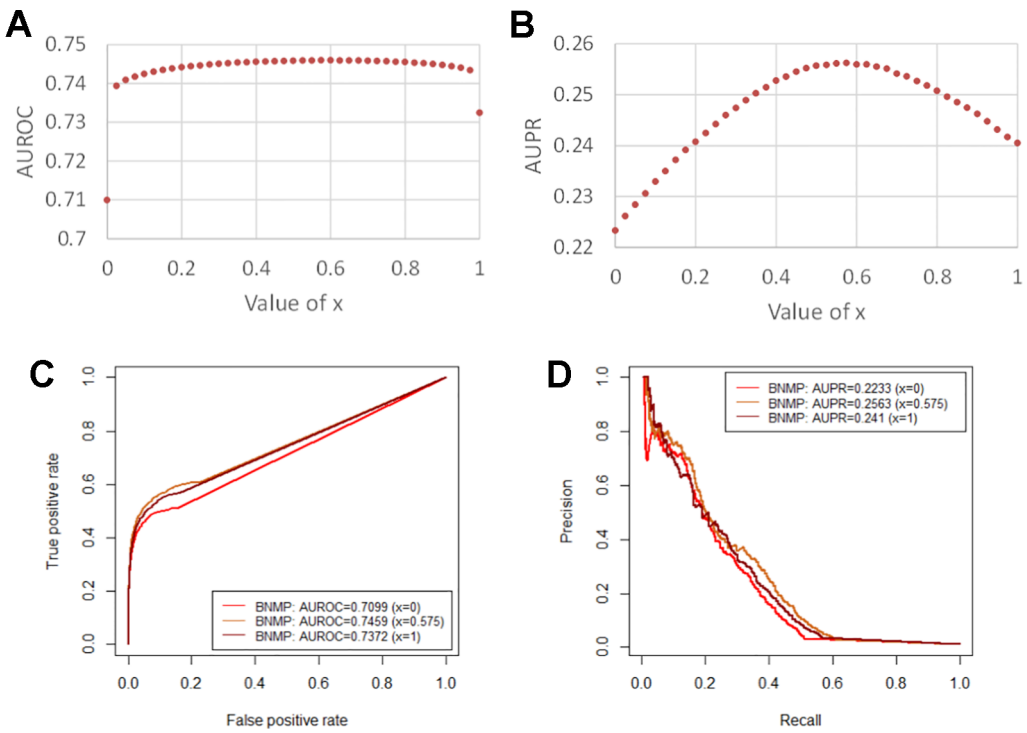


FIGURE 2 | Prediction performance of BNMP with network 1. **(A)** Influence on AUROC values by different balance parameter values. **(B)** Influence on AUPR values by different balance parameter values. **(C)** ROC curves of BNMP with the different balance parameter values. **(D)** PR curves of BNMP with the different balance parameter values.

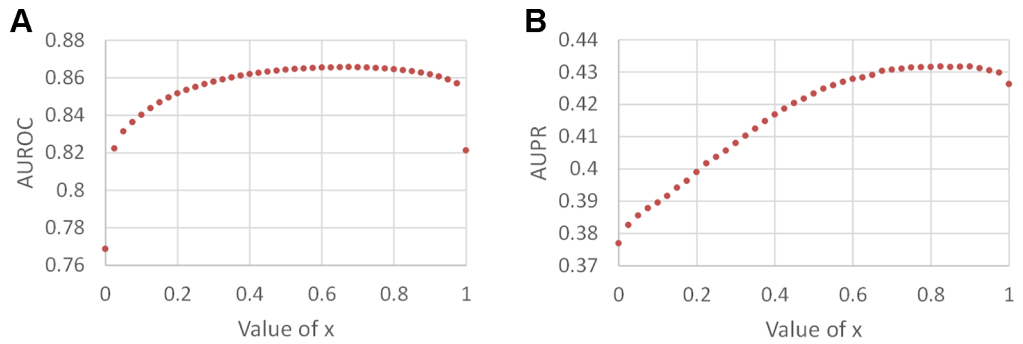


FIGURE 3 | Prediction performance of BNMP with network 2. **(A)** Influence on AUROC values by different balance parameter values. **(B)** Influence on AUPR values by different balance parameter values.

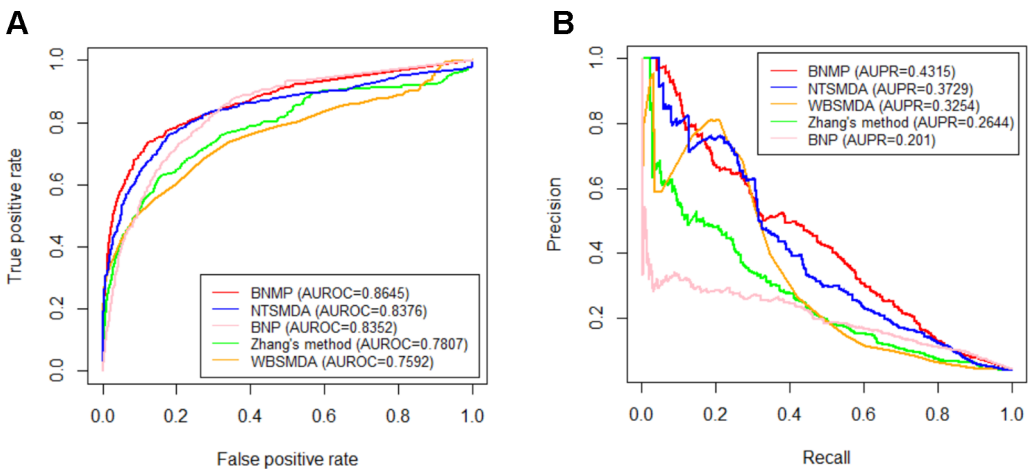


FIGURE 4 | Comparison of five methods. **(A)** ROC curves. **(B)** PR curves.

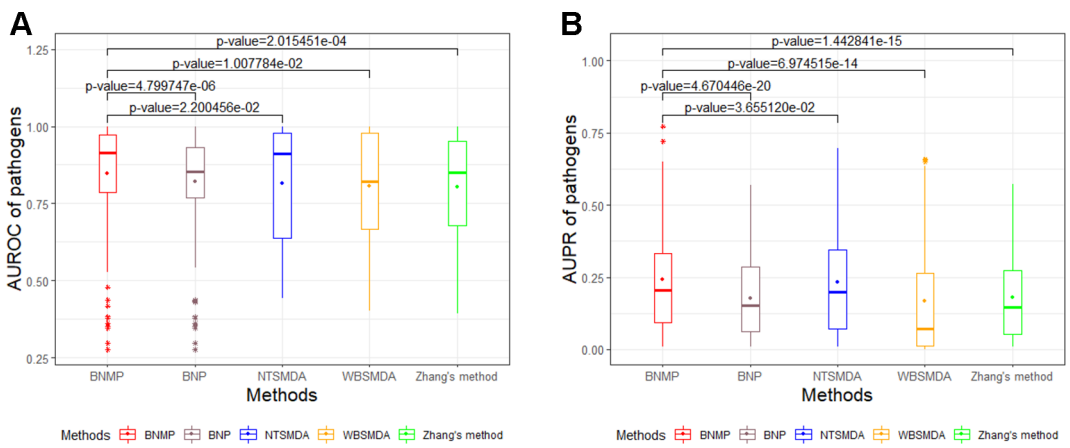


FIGURE 5 | Paired *t*-test for the AUROC and AUPR values of pathogens between BNMP and other methods. **(A)** Box-and-whisker plot of AUROC values with *p*-values. **(B)** Box-and-whisker plot of AUPR values with *p*-values.

are less than 0.05, indicating that the proposed approach is a significant advance over the previous approaches and has better prediction ability.

Validation *via* Biological Evidence

Most data sources use text mining algorithms to obtain the original interaction data. Due to the limitation of the development of pathogen–host interaction text mining algorithms, the existing open data sources can only cover a part of pathogen–host interaction data. To further test the ability of BNMP to predict potential pathogen–host associations, we rank pathogen–host pairs without relations in existing data sets according to association scores and search the public literature to see whether there is evidence that pathogens and hosts with higher association scores have relations. It is found that among the top 20 pathogen–host pairs without relations in the existing data set, biological experiments have verified that 16 pairs have associations (**Table 2**); these 16 pairs are ranked lower by the other four methods. The pair of pathogen Serratia marcescens and host Mus musculus ranks 1st. Iwaya A et al. studied the clinical application and evaluation of rapid and quantitative detection of blood Serratia marcescens by a real-time PCR assay in a mouse infection model (Iwaya et al., 2005). The pair of pathogen Cronobacter turicensis and host Mus musculus ranks 3rd. Tóthová Ľ et al. used Cronobacter turicensis to infect female mice to prove the effects of isolated Cronobacter-specific phages on renal colonization in a model of urinary tract infection in mice (Tóthová et al., 2011). The pair of pathogen Escherichia coli O157:H7 and host Mus musculus ranks 4th. Tanji Y et al. found that repeated oral administration of SP15-21-22 can effectively treat mice infected with Escherichia coli O157:H7 (Tanji et al., 2005). The pair of pathogen Acinetobacter nosocomialis and host Homo sapiens ranks 5th. Visca P et al. discussed the infection mechanism and threats of Acinetobacter nosocomialis and other Acinetobacter species to humans (Visca et al., 2011). The pair of pathogen Stenotrophomonas maltophilia and host Mus musculus ranks 6th. Bacterial adhesion to mouse tracheal mucus as the role of flagella in the adhesion process were investigated using clinical isolates of Stenotrophomonas maltophilia (Zgair and Chhibber, 2011). The pair of pathogen Sclerotinia sclerotiorum and host Nicotiana tabacum ranks 7th. Researchers carried out a preliminary evaluation of the potential of polyamine biosynthesis inhibition a strategy for the control of plant diseases initiated by S. sclerotiorum ascospores, using tobacco (Nicotiana tabacum) leaf discs as an experimental system (Garriz et al., 2010). The 8th-ranking confirmed pair is pathogen Pseudomonas aeruginosa and host Oryctolagus cuniculus. Researchers have determined the pharmacokinetics and adverse effects following SC administration of ceftiofur crystalline free acid (CCFA) in Oryctolagus cuniculus by using Pseudomonas aeruginosa and other bacterium (Gardhouse et al., 2017). The 9th-ranking confirmed pair is pathogen Enterococcus faecalis and host Homo sapiens. A study showed that an 88-kDa secreted protein, endoglycosidase (Endo) E, which is most likely responsible for the activity of the human pathogen Enterococcus faecalis, degrades the N-linked glycans of human RNase B to acquire nutrients (Mattias and Fischetti, 2004). The pair of pathogen Alternaria citri and host Citrus reticulate ranks 10th. Reasearchers found that the phytopathogenic fungus, Alternaria citri (Alternaria alternata pathotype citri), produces a complex of analogous toxins (ACTG-toxin) that selectively damages Dancy tangerine (Citrus reticulata) and other mandarin cultivars (Kohmoto et al., 1979). The pair of pathogen Mycobacterium marinum and host Homo sapiens ranks 12th. Flowers found that a person was infected with Mycobacterium marinum by being bitten by a dolphin and thus associated human mycobacterial infection with an aquatic mammal (Flowers, 1970). The 14th score is the pair of pathogen Mycobacteroides abscessus and host Homo sapiens. Mycobacterium abscessus is one of the common species that causes disseminated infections in patients with cystic fibrosis. It has been reported that NLRP3 inflammasome activation contributed to antimicrobial responses against M. abscessus in human macrophages and that its activation was dependent on dectin-1/Syk signaling (Hye-Mi et al., 2012). The pair of pathogen Alternaria alternata and host Solanum

TABLE 2 | Pathogen–host pairs predicted using BNMP and their rank according to five methods.

Pathogen	Host	BNMP	NTSMDA	BNP	Zhang's method	WBSMDA
Serratia marcescens	Mus musculus (Iwaya et al., 2005)	1	43	15	17	13
Cronobacter turicensis	Mus musculus (Tóthová et al., 2011)	3	10	26	24	109
Escherichia coli O157:H7	Mus musculus (Tanji et al., 2005)	4	38	172	14	10
Acinetobacter nosocomialis	Homo sapiens (Visca et al., 2011)	5	13	251	119	18
Stenotrophomonas maltophilia	Mus musculus (Zgair and Chhibber, 2011)	6	44	124	21	13082
Sclerotinia sclerotiorum	Nicotiana tabacum (Garriz et al., 2010)	7	61	44	540	169
Pseudomonas aeruginosa	Oryctolagus cuniculus (Gardhouse et al., 2017)	8	588	62	960	55
Enterococcus faecalis	Homo sapiens (Mattias and Fischetti, 2004)	9	37	33	109	19
Alternaria citri	Citrus reticulata (Kohmoto et al., 1979)	10	528	57	9021	41
Mycobacterium marinum	Homo sapiens (Flowers, 1970)	12	39	36	115	26
Mycobacteroides abscessus	Homo sapiens (Hye-Mi et al., 2012)	14	20	25	102	20
Alternaria alternata	Solanum lycopersicum (Hai and Gubler, 2012)	15	261	40	447	3045
Enterococcus faecium	Homo sapiens (Lester et al., 2006)	16	40	27	106	121
Fusarium oxysporum	Nicotiana tabacum (Jennings et al., 2001)	17	118	43	537	1313
Pectobacterium carotovorum	Arabidopsis thaliana (Lee et al., 2012)	19	259	74	199	764
Mycoplasma agalactiae	Mus musculus (Smith, 1967)	20	26	201	101	211

lycopersicum ranks 15th. A study evaluated whether 1-MCP treatment could affect postharvest decay caused by A. alternata, B. cinerea, and Fusarium spp. in Solanum lycopersicum (Hai and Gubler, 2012). The 16th-ranking association is the pair of pathogen Enterococcus faecium and host Homo sapiens. A previous study was performed to determine whether resistance genes from an E. faecium isolate of animal origin could be transferred to a human E. faecium isolate in the intestines of human volunteers without any selective antimicrobial pressure (Lester et al., 2006). The 17th pair of pathogen and host is Fusarium oxysporum and Nicotiana tabacum. Jennings et al. found that protein Nep1 from Fusarium oxysporum inducted defense responses in tobacco (Jennings et al., 2001). The 19th potential link is Pectobacterium carotovorum and Arabidopsis thaliana. The study indicated that Arabidopsis thaliana were infected with Pectobacterium carotovorum (Lee et al., 2012). The 20th potential link is pathogen Mycoplasma agalactiae and host Mus musculus. Smith G R. et al. used Mycoplasma agalactiae to infect mice to verify the toxicity of the Mycoplasma agalactiae (Smith, 1967). Based on the above findings, one can argue that BNMP is very efficient in predicting associations between pathogens and hosts.

DISCUSSION

In this study, we focus on the problem of pathogen–host association prediction. To consider the relations of pathogens and hosts comprehensively, we adopt the pattern of local before global, proposing a novel approach, BNMP. The method is based on bipartite network modules and integrates module information of pathogens and hosts, respectively, into a bipartite network projection model to improve prediction performance. Where the host is the seed, the time complexity of acquiring the association score vector between the seed and all pathogens is $O(ms^3t)$, where m is the degree of the seed. Hence, the time complexity of acquiring $S_{pathogen-host}$ is $O(es^3t)$, where e is the number of associations in the host-pathogen association network. Similarly, the time complexity of acquiring $S^T_{host-pathogen}$ is $O(et^3s)$. BNMP has a time complexity of $O(est(s^2+t^2))$, namely $O(es^3t)$ when $s>t$ and $O(et^3s)$ when $t>s$. Experimental results show that BNMP achieved better prediction performance compared with other efficient methods.

Although BNMP is used here in pathogen–host association prediction, it can also be applied to association analysis in other fields, such as miRNA–disease association prediction, drug–target interaction prediction, and drug–cell line response prediction. Hence, our study has a wide range of uses. Module-based information can help improve the score in the bipartite network because more information related to the nodes in a network is included in the predictive model, which avoid missing the information of neighbors. Although BNMP performs well on the existing data set, the number of associations between pathogens and hosts in the data set is insufficient, which affects the performance of the proposed method. As more association relationships are found or added into databases and more information about regulatory modules (Chen et al., 2019a; Chen et al., 2019b) is employed in the future, the prediction performance of BNMP should further improve.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be downloaded from PHI-base (http://www.PHI-base.org/index.jsp), HPIDB (https://hpidb.igbb.msstate.edu/index.html) and intact (https://www.ebi.ac.uk/intact/).

AUTHOR CONTRIBUTIONS

JL and SW designed and implemented the algorithm. ZC and SW analyzed the results and wrote the manuscript, and YW made suggestions. All of the authors read and approved the final manuscript.

FUNDING

This work was partially supported by the Natural Science Foundation of Heilongjiang Province (Grant No. F2016016) and the National Key Research and Development Program of China (Grant No.2016YFC0901905).

ACKNOWLEDGMENTS

The authors acknowledge the contributions of colleagues in the group.

REFERENCES

Ahmed, I., Witbooi, P., and Christoffels, A. (2018). Prediction of human-Bacillus anthracis protein-protein interactions using multi-layer neural network. *Bioinformatics* 34 (24), 4159–4164. doi: 10.1093/bioinformatics/bty504

Ammari, M. G., Gresham, C. R., Mccarthy, F. M., and Nanduri, B. (2016). HPIDB 2.0: a curated database for host–pathogen interactions. *Database* 2016, baw103. doi: 10.1093/database/baw103

Chen, J., Han, G., Xu, A., and Cai, H. (2019a) Identification of Multidimensional Regulatory Modules through Multi-graph Matching with Network Constraints. *IEEE Transact. Biomed. Engr.* doi: 10.1109/TBME.2019.2927157

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019b). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35, 602–610. doi: 10.1093/bioinformatics/bty662

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L., Liu, Y., et al. (2016a). WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. *Sci. Rep.* 6, 21106. doi: 10.1038/srep21106

Chen, X., Ren, B., Chen, M., Wang, Q., Zhang, L., and Yan, G. (2016b). NLLSS: predicting synergistic drug combinations based on semi-supervised learning. *PLoS Computat. Biol.* 12, e1004975. doi: 10.1371/journal.pcbi.1004975

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018). BNPMDA: bipartite network projection for MiRNA-Disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Dallas, T., Park, A. W., and Drake, J. M. (2017). Predicting cryptic links in host-parasite networks. *PLoS Computat. Biol.* 13, e1005557. doi: 10.1371/journal.pcbi.1005557

Davis, F. P., Barkan, D. T., Eswar, N., Mckerrow, J. H., and Sali, A. (2010). Host pathogen protein interactions predicted by comparative modeling. *Protein Sci.* 16, 2585–2596. doi: 10.1110/ps.073228407

Demišar, J., and Schuurmans, D. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Machine Learn. Res.* 7, 1–30.

Doolittle, J. M., and Gomez, S. M. (2010). Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens. *Virol. J.* 7, 82. doi: 10.1186/1743-422x-7-82

Dyer, M., Murali, T., and Sobral, B. (2007). Computational prediction of host-pathogen protein-protein interactions. *Bioinformatics* 23, i159. doi: 10.1093/bioinformatics/btm208

Evans, P., Dampier, W., Ungar, L., and Tozeren, A. (2009). Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. *BMC Med. Genomics* 2, 27. doi: 10.1186/1755-8794-2-27

Fei, Z., Wang, M., Xi, J., Yang, J., and Ao, L. (2018). A novel heterogeneous network-based method for drug response prediction in cancer cell lines. *Sci. Rep.* 8, 3355. doi: 10.1038/s41598-018-21622-4

Flowers, D. (1970). Human infection due to mycobacterium marinum after a dolphin bite. *J. Clin. Pathol.* 23, 475–477. doi: 10.1136/jcp.23.6.475

Gardhouse, S., Guzman, D. S., Cox, S., Kass, P. H., Drazenovich, T. L., Byrne, B. A., et al. (2017). Pharmacokinetics and safety of ceftiofur crystalline free acid in new zealand white rabbits (Oryctolagus cuniculus). *Am. J. Veterin. Res.* 78, 796–803. doi: 10.2460/ajvr.78.7.796

Garriz, A., Dalmasso, M. C., Marina, M., Rivas, E. I., Ruiz, O. A., and Pieckenstain, F. L. (2010). Polyamine metabolism during the germination of sclerotinia sclerotiorum ascospores and its relation with host infection. *New Phytol.* 161, 847–854. doi: 10.1046/j.1469-8137.2003.00983.x

Geeleher, P., Cox, N. J., and Huang, R. S. (2014). Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines. *Genome Biol.* 153, R47. doi: 10.1186/gb-2014-15-3-r47

Hai, S., and Gubler, W. D. (2012). Effect of 1-methylcyclopropene (1-MCP) on reducing postharvest decay in tomatoes (Solanum lycopersicum L.). *Postharvest Biol. Technol.* 64, 133–137. doi: 10.1016/j.postharvbio.2011.06.005

Hye-Mi, L., Jae-Min, Y., Ki-Hye, K., Jichan, J., Gun, K., Bong, P. J., et al. (2012). Mycobacterium abscessus activates the NLRP3 inflammasome *via* Dectin-1-Syk and p62/SQSTM1. *Immunol. Cell Biol.* 90, 601–610. doi: 10.1038/icb.2011.72

Iwaya, A., Nakagawa, S., Iwakura, N., Taneike, I., Kurihara, M., Kuwano, T., et al. (2005). Rapid and quantitative detection of blood Serratia marcescens by a real-time PCR assay: its clinical application and evaluation in a mouse infection model. *FEMS Microbiol. Lett.* 248, 163–170. doi: 10.1016/j.femsle.2005.05.041

Jennings, J. C., Apel-Birkhold, P. C., Mock, N. M., JacynBaker, C., Anderson, J. D., and Bailey, B. A. (2001). Induction of defense responses in tobacco by the protein Nep1 from Fusarium oxysporum. *Plant Sci.* 161, 891–899. doi: 10.1016/s0168-9452(01)00483-6

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *International joint conference on artificial intelligence*. 14 (2), 1137–1145.

Kohmoto, K., Scheffer, R. P., and Whiteside, J. O. (1979). Host-selective toxins from Alternaria citri". *Phytopathology* 69, 667–671. doi: 10.1094/phyto-69-667

Le, D. H., and Pham, V. H. (2018). Drug response prediction by globally capturing drug and cell line information in a heterogeneous network. *J. Mol. Biol.* 430 (18), 2993–3004. doi: 10.1016/j.jmb.2018.06.041

Lee, S. A., Chan, C. H., Tsai, C. H., Lai, J. M., Wang, F. S., Kao, C. Y., et al. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions. *BMC Bioinform.* 9, 2008. doi: 10.1186/1471-2105-9-s12-s11

Lee, H. J., Jin, S. K., Yoo, S. J., Kang, E. Y., Song, H. H., Yang, K. Y., et al. (2012). Different roles of glycine-rich RNA-binding protein7 in plant defense against Pectobacterium carotovorum, Botrytis cinerea, and tobacco mosaic viruses. *Plant Physiol. Biochem. Ppb* 60, 46–52. doi: 10.1016/j.plaphy.2012.07.020

Lester, C. H., Niels, F. M. L., Thomas Lund, S. R., Monnet, D. L., and Hammerum, A. M. (2006). In vivo transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers. *Antimicrob. Agents Chemother.* 50, 596. doi: 10.1128/aac.50.2.596-599.2006

Mattias, C., and Fischetti, V. A. (2004). A novel secreted endoglycosidase from Enterococcus faecalis with activity on human immunoglobulin G and ribonuclease B. *J. Biolog. Chem.* 279, 22558–22570. doi: 10.1074/jbc.m402156200

Mei, S., and Zhu, H. (2014). AdaBoost based multi-instance transfer learning for predicting proteome-wide interactions between Salmonella and human proteins. *PLoS One* 9, e110488. doi: 10.1371/journal.pone.0110488

Mei, S., and Zhu, H. (2015). A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Sci. Rep.* 5, 8034. doi: 10.1038/srep08034

Nourani, E., Khunjush, F., and Durmuş, S. (2015). Computational approaches for prediction of pathogen-host protein-protein interactions. *Front. Microbiol.* 6, 94. doi: 10.3389/fmicb.2015.00094

Patrick, A., and Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. U.S.A.* 99, 5896–5901. doi: 10.1073/pnas.092147999

Sandra, O., Mais, A., Bruno, A., Lionel, B., Leonardo, B., Fiona, B. C., et al. (2014). The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 42, 358–363. doi: 10.1093/nar/gkt1115

Smith, G. (1967). Experimental infection of mice with Mycoplasma agalactiae. *J. Comparative Pathol.* 77, 199–202. doi: 10.1016/0021-9975(67)90011-4

Stefan, W. (2011). Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens. *PLoS ONE* 6, e26960. doi: 10.1371/journal.pone.0026960

Sun, D., Ao, L., Feng, H., and Wang, M. (2016). NTSMDA: Prediction of miRNA-disease associations by integrating network topological similarity. *Mol. Biosyst.* 12, 2224. doi: 10.1039/c6mb00049e

Tanji, Y., Shimada, T., Fukudomi, H., Miyanaga, K., Nakai, Y., and Unno, H. (2005). Therapeutic use of phage cocktail for controlling Escherichia coli O157: H7 in gastrointestinal tract of mice. *J. Biosci. Bioengr.* 100, 280–287. doi: 10.1263/jbb.100.280

Tóthová, Ľ., Celec, P., Bábíčková, J., Gajdošová, J., Al-Alami, H., Kamodyova, N., et al. (2011). Phage therapy of Cronobacter-induced urinary tract infection in mice. *Med. Sci. Monitor: Int. Med. J Exp. Clin. Res.* 17, BR173. doi: 10.12659/msm.881844

Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., et al. (2017). PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.* 45. doi: 10.1093/nar/gkw1089

Visca, P., Seifert, H., and Towner, K. J. (2011). Acinetobacter infection–an emerging threat to human health. *IUBMB Life* 63, 1048–1054. doi: 10.1002/iub.600

Zgair, A. K., and Chhibber, S. (2011). Adhesion of Stenotrophomonas maltophilia to mouse tracheal mucus is mediated through flagella. *J. Med. Microbiol.* 60, 1032–1037. doi: 10.1099/jmm.0.026377-0

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., and Liu, X. S. (2015). Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model. *PLoS Computat. Biol.* 11, e1004498. doi: 10.1371/journal.pcbi.1004498

Zhou, T., Ren, J., Medo, M., and Zhang, Y. C. (2007). Bipartite network projection and personal recommendation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76, 046115. doi: 10.1103/physreve.76.046115

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 28 January 2020
doi: 10.3389/fbioe.2020.00008

Predict New Therapeutic Drugs for Hepatocellular Carcinoma Based on Gene Mutation and Expression

Liang Yu, Fengdan Xu and Lin Gao*

School of Computer Science and Technology, Xidian University, Xi'an, China

Hepatocellular carcinoma (HCC) is the fourth most common primary liver tumor and is an important medical problem worldwide. However, the use of current therapies for HCC is no possible to be cured, and despite numerous attempts and clinical trials, there are not so many approved targeted treatments for HCC. So, it is necessary to identify additional treatment strategies to prevent the growth of HCC tumors. We are looking for a systematic drug repositioning bioinformatics method to identify new drug candidates for the treatment of HCC, which considers not only aberrant genomic information, but also the changes of transcriptional landscapes. First, we screen the collection of HCC feature genes, i.e., kernel genes, which frequently mutated in most samples of HCC based on human mutation data. Then, the gene expression data of HCC in TCGA are combined to classify the kernel genes of HCC. Finally, the therapeutic score (*TS*) of each drug is calculated based on the kolmogorov-smirnov statistical method. Using this strategy, we identify five drugs that associated with HCC, including three drugs that could treat HCC and two drugs that might have side-effect on HCC. In addition, we also make Connectivity Map (CMap) profiles similarity analysis and KEGG enrichment analysis on drug targets. All these findings suggest that our approach is effective for accurate predicting novel therapeutic options for HCC and easily to be extended to other tumors.

Keywords: hepatocellular carcinoma (HCC), drug repositioning, mutated genes, kernel genes, gene expression

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Zengyou He,
Dalian University of Technology (DUT), China
Wen Zhang,
Huazhong Agricultural University, China

****Correspondence:***
Liang Yu
lyu@xidian.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *21 November 2019*
Accepted: *07 January 2020*
Published: *28 January 2020*

Citation:
Yu L, Xu F and Gao L (2020) Predict New Therapeutic Drugs for Hepatocellular Carcinoma Based on Gene Mutation and Expression. Front. Bioeng. Biotechnol. 8:8. doi: 10.3389/fbioe.2020.00008

INTRODUCTION

Identifying a cure for cancer is a difficult, costly and often inefficient process (Adams and Brantner, 2006). Drug repositioning, i.e., the discovery of new indications of existing drugs, beyond their original indications, is an increasingly attractive new-use discovery model. In addition to saving time and money, one advantage of the drug reuse approach is that existing drugs have been reviewed for safety, dose and toxicity (Ashburn and Thor, 2004; Fathima et al., 2018; Su et al., 2019; Yu et al., 2019). As a result, repurposed drugs usually go into clinical trials faster than newly developed drugs (Yu et al., 2017a, 2018). The rapid development of genomics has resulted in the generation of genomic and transcription group data from disease samples, normal tissue samples, animal models and cell lines. Transcriptomic profiles, such as gene expression data, are most widely used for drug repositioning (Yu et al., 2016). A key data source behind several re-use efforts is the Connectivity Map (CMap) (Lamb et al., 2006), which generated large-scale gene expression profiles in human cancer cell lines treated with different drug compounds under different conditions. The CMap method attempts to provide a more comprehensive view of this transcription data and use them to connect expression profiles across conditions (Lamb et al., 2006). In particular, it suggests that if there is a strong negative correlation between disease characteristics and drug expression

characteristics, the drug may have a therapeutic effect on the disease. For example, by systematically comparing the gene expression characteristics of GEO-derived inflammatory bowel disease (IBD) with the gene expression characteristics of a group of 164 drug compounds from CMap, Dudley et al. (2011) predicted several interesting new drug-disease pairs and, in the IBD preclinical model, validated one pair. Yu et al. (2015) proposed a method that discovered the drug-disease association based on protein complexes. In another case, Jahchan et al. (2013) applied a drug repurposing bioinformatics method to identifying antidepressant drugs for the treatment of small cell lung cancer through querying a large compendium of gene expression profiles. Although many machine learning-based methods have been developed by using features (Zhang et al., 2017, 2018a,b, 2019), more and more literature supports the usage of CMap for drug repositioning; despite this, there are still problems. A candidate can often be strengthened using independent disease signatures. But disease signatures are often selected by statistical methods, they are lack of biological information.

Hepatocellular carcinoma (HCC) is the fourth most common primary liver tumor and is an important medical problem worldwide (El-Serag and Mason, 1999; Yu et al., 2017b). HCC is usually caused by infection with hepatitis B virus (HBV) (Chang and Liu, 2016) and hepatitis C (HCV) (Lingala and Ghany, 2015), exposure to aflatoxin B1 from Aspergillus (Kew, 2013), alcohol abuse (Abenavoli et al., 2016), or non-alcoholic fatty hepatitis (Charrez et al., 2016). However, the use of current therapies for HCC is no possible to be cured, and despite numerous attempts and clinical trials, there are not so many approved targeted treatments for HCC. So, it is necessary to identify additional treatment strategies to prevent the growth of HCC tumors.

Many diseases, but especially cancer, are related with abnormal genomes and transcription landscapes (Chakravarthi et al., 2016; Tang et al., 2018). In this study, we seek to use systematic drug repositioning bioinformatics to identify new drug candidates for the treatment of HCC. First, we screen the collection of HCC feature genes that frequently mutated in most samples of HCC based on human mutation data. Then, the gene expression data of HCC in TCGA are combined to classify the gene set of HCC. Finally, the therapeutic score (*TS*) of each drug is calculated based on the kolmogorov-smirnov statistical method. Using this strategy, we identified five drugs that associated with HCC, including three drugs that could cure HCC and two drugs that might have bad effect on HCC. In addition, we also make CMap (Lamb et al., 2006) profiles similarity analysis and KEGG enrichment analysis on drug targets. All these findings suggest that our approach is effective for accurate discovering novel therapeutic options for HCC and easily to be extended to other tumors.

MATERIALS AND METHODS

Datasets

HCC Gene Expression Data

The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015) is a comprehensive and coordinated effort to accelerate our understanding of the molecular basis of cancer by applying genomic analysis techniques, including large-scale genome sequencing. TCGA researchers aim to catalog and discover major changes to the cancer-causing genome to create a comprehensive "atlas" of cancer genomes. So far, the project analyzed groups of more than 30 human tumors through large-scale genome sequencing and integrated multidimensional analysis.

We download the gene expression profiles of HCC from TCGA, and there are 423 samples in the data set. The type of a sample is distinguished by the barcode provided by TCGA. If the fourth part of the barcode of one sample is in the range from 01 to 09, the sample is a cancer sample. If the fourth part of the barcode in the range from 10 to 19, the sample is a normal sample. The specific introduction to the barcode can be found in TCGA help file. First, we obtain gene expression matrix data (20,501 × 423), which contains 373 cancer samples, 50 normal samples, and 20,501 genes. Then, we standardize the expression values of all genes as follows:

$$z_{ij} = \frac{g_{ij} - mean(g_i)}{std(g_i)} \tag{1}$$

where g_{ij} represents the expression value of gene i in sample j, and $mean\left(g_i\right)$and $std\left(g_i\right)$, respectively represent mean and standard deviation of the expression vector for gene i across all samples. Finally, we use Limma (Ritchie et al., 2015) to analyze cancer and normal samples and get the log FC value of each gene. The definition of log FC is as follows:

$$\log FC_i = \log_2 \left(\frac{\frac{1}{|T|} \sum\limits_{k \in T} z_{ik}}{\frac{1}{|N|} \sum\limits_{k \in N} z_{ik}} \right) \tag{2}$$

where $\log FC_i$ is the log FC value of gene i; z_{ik} is the normalized expression of gene i in sample k [see formula (1)]; T is the set of cancer samples ($|T|$ =373); N is the set of normal samples ($|N|$ =50).

For a gene, if its $\left|\log FC\right| \geq 1$ and $p- > value \leq 0.02$, it is a differentially expressed gene. The thresholds of log FC and $p- > value$ refer to Dalman et al. (2012).

Gene Expression Data Related to Drugs

The gene expression data related to drugs is downloaded from the CMap (http://www.broadinstitute.org/cmap/) database. It contains 6,100 instances which cover 1,309 drugs. These instances are measured on five types of human cancer cell lines, including the breast cancer epithelial cell lines MCF7 and ssMCF7, the prostate cancer epithelial cell line PC3, the nonepithelial lines HL60 (leukemia) and SKMEL5 (melanoma).

SNP Mutation Data of HCC

We download the single nucleotide polymorphism (SNP) gene mutation data of HCC from TCGA database. The SNP mutation data contains 373 cancer patient sample files, and each sample file contains the detailed descriptions of all the mutated genes.

Since the mutation frequency of each gene across all samples is different, we select genes with relatively high mutation frequency for further analysis. Here, the mutation frequency is set to be no less than 11 ($11 = 373 \times 3\%$), that is a gene mutated in at least three percent of all samples. These genes are defined as frequently mutated genes. Finally, we find 406 frequently mutated genes.

Methods

Defining the Feature Gene Set of HCC

According to the data analysis we have done in section **Datasets**, we can divide the 20,501 genes into three classes based on their mutation frequency and differential expression value. One category is the kernel genes, which mutate frequently. The second category is the secondary genes, which do not mutate frequently but differentially express. The third category is the marginal genes, which neither mutate frequently nor differentially express.

In our work, we take the 406 kernel genes, i.e., frequently mutated gene, as the feature gene set of HCC.

Calculating the Therapeutic Scores of Drugs

We select kernel genes as the feature genes of HCC and rank them in descending order based on their differential expressions. For a gene, if its $\log FC$ value is >0, it is stored in up-regulated gene set. Otherwise, it is stored in down-regulated gene set. Finally, we get two ordered gene lists for HCC:

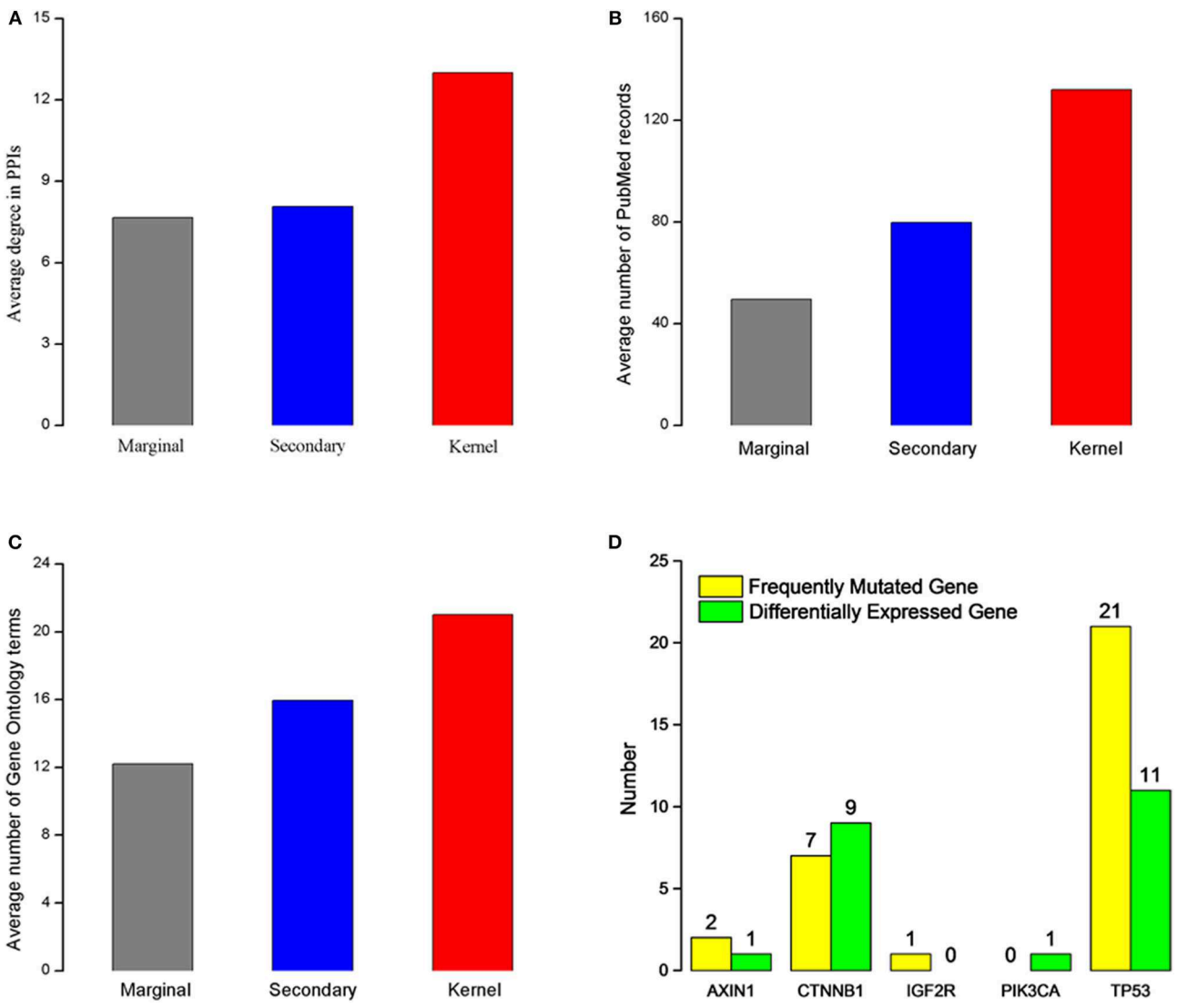


FIGURE 1 | Characteristics of the three gene types. **(A)** Average degree for three different gene types. **(B)** Average PubMed records associated for each gene type. **(C)** Gene Ontology terms annotated for each gene type. **(A–C)** The red rectangle represents kernel genes, the blue rectangle represents secondary genes, the gray rectangle represents marginal genes. **(D)** Type distribution of five kernel genes' direct neighbor genes. Green rectangle represents differentially expressed genes in HCC, and yellow rectangle represents frequently mutated genes in patients with HCC.

the up-regulated gene list (G_{up}) and the down-regulated gene list (G_{down}).

We get 6,100 gene expression instances covered 1,309 drugs from CMap database. In other words, a drug may correspond to multiple instances. We rank the genes in each instance by their differential expression values between drug-treated and drug-untreated cell lines. In this way, we get 6,100 drug-related gene lists. Therefore, based on kernel genes and 6,100 drug-related gene expression instances, we use a non-parameter, ranking-based pattern matching strategy that was originally introduced by Lamb et al. (2006) to evaluate the relationship between drugs and HCC.

TABLE 1 | HCC related genes extracted from OMIM.

Gene names	Gene entrez IDs
IGF2R	3482
CASP8	841
MET	4233
PDGFRL	5157
TP53	7157
PIK3CA	5290
CTNNB1	1499
AXIN1	8312

We take each ranked drug expression profile as reference signature and assess their similarity to HCC. We compute a connectivity score separately for the set of up- or down-regulated genes: ES_{up} or ES_{down}. The computational formulas as follows (Lamb et al., 2006):

$$a = \max_{p=1}^{m}\left[\frac{p}{m} - \frac{V(p)}{n}\right] \tag{3}$$

$$b = \max_{p=1}^{m}\left[\frac{V(p)}{n} - \frac{p-1}{m}\right] \tag{4}$$

$$ES_{up/down} = \begin{cases} a_{up/down}(\text{if } a_{up/down} > b_{up/down}) \\ -b_{up/down}(\text{if } a_{up/down} < b_{up/down}) \end{cases} \tag{5}$$

Where n represents the total number of genes in the reference drug expression profile; m represents the size of G_{up} or the size of G_{down}; p represents the position of the input set ($p = 1...m$); $V(p)$ is the position of the pth input gene in the gene list of drug expression profile.

The therapeutic score (TS) of a drug is calculated as follows:

$$TS = \frac{1}{k}\sum_{j=1}^{k} ES_{up} - ES_{down} \tag{6}$$

If the up-regulated genes are near the top (up-regulated) of the rank-ordered drug gene lists and the down-regulated genes are

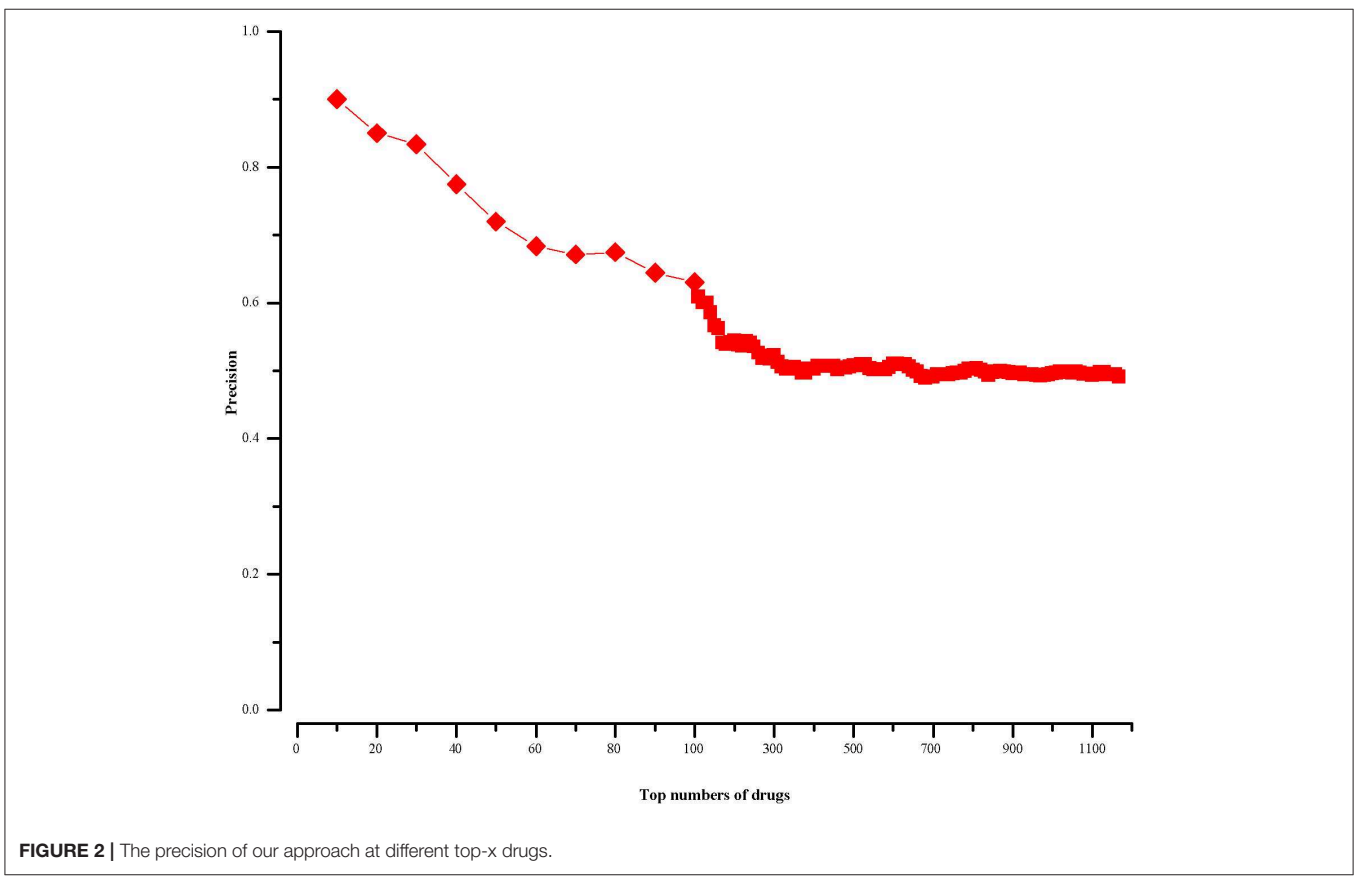


FIGURE 2 | The precision of our approach at different top-x drugs.

near the bottom (down-regulated) of the rank-ordered drug gene lists, we can get high positive therapeutic scores (*TS*), which indicate the drugs and HCC have similar expression profiles and the drugs might aggravate HCC. On the other hand, if the up-regulated pathway genes are near the bottom of the rank-ordered drug gene lists and the down-regulated pathway genes are near the top of the rank-ordered drug gene lists, we can get negative therapeutic scores (*TS*), which imply the given drugs and HCC have adverse expression profiles and the drugs could be treatment candidates for HCC.

RESULTS

Analysis of Disease Characteristics of HCC

We characterize the kernel, secondary, and marginal genes in the context of protein interaction (PPIs) network, PubMed (www.ncbi.nlm.nih.gov/pubmed), and Gene Ontology (Ashburner et al., 2000) term annotation. The Human Protein Reference Database (HPRD) (Prasad et al., 2009) is a protein database for experimentally derived information about human proteomics, including protein and protein interactions (Ding et al., 2016; Wei et al., 2017a), post-translational modifications (PTMs) (Wei et al., 2017b) and other information. We download all human PPIs from this database, containing 15,231 proteins and 38,167 interactions. Interestingly, we find that all three gene types had heterogeneous degree distribution, and that the kernel genes tend to have higher degrees than those of secondary and marginal genes (**Figure 1A**). Similarly, kernel genes are related with more PubMed records and Gene Ontology term annotation than secondary and marginal genes (**Figures 1B,C**).

In order to analyze biological functions of kernel genes, we analysis the nine HCC pathogenic genes obtained from Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005) from two aspects of gene mutation and expression level change. These eight HCC pathogenic genes (**Table 1**) are IGF2R, CASP8, MET, PDGFRL, TP53, PIK3CA, CTNNB1, and AXIN1. We find that five (IGF2R, TP53, PIK3CA, CTNNB1, AXIN1) of these genes are belong to kernel genes, these genes are frequent mutations, but their expression level don't change significantly. For direct neighbors in PPIs of these five genes, we find that there are frequently mutated or differentially expressed genes (see **Figure 1D**) among their direct neighbors. TP53 is a quite important tumor suppressor gene, which can translate and synthesize protein P53. P53 protein is a vital regulator for cell growth, proliferation and injury repair. For the direct neighbors of TP53, there are 27 frequently mutated genes, and 11 differentially expressed genes. CTNNB1 gene can encode β-catenin, a dual function protein that involves in regulation and coordination of cell–cell adhesion and gene transcription (Nollet et al., 1996). Recent study of HCC has shown that CTNNB1 gene mutations and overexpression of its encoded protein are closely related to occurrence, progression and prognosis of tumor (Kitao et al., 2015). CTNNB1 has 7 frequently mutated direct neighbors, and 9 differentially expressed direct neighbors. The above analysis results show that the kernel genes selected by mutation and expression information contain more comprehensive biological knowledge and to some extent, the characteristics of HCC can be depicted.

Choosing Potential HCC Drugs Through CTD Benchmark

To find most likely HCC-related drugs, we need evaluate the precision of our method firstly. We take Comparative Toxicogenomics Database (CTD) (Davis et al., 2015) as benchmark. CTD supplies manual collated information about drug-gene, drug-disease, and gene-disease interactions. Curated chemical-disease relationships are obtained from the published literature by CTD biocurators and inferred relationships are set up via CTD curated chemical-gene associations.

For a drug in CMap, if it cannot find corresponding chemical name in CTD, we will not calculate its therapeutic score (defined in section "Methods"). In this way, we finally get 1168 scored drugs. Because most drug-disease associations in CTD are not marked as positive or negative, we rank the 1168 drugs in descending order by the absolute values of their therapeutic scores. We know the top drugs imply stronger connections with HCC. And then we calculate the precisions of our approach at different top-x drugs, which are shown in **Figure 2**. The precision is calculated as follows:

$$precision = \frac{P_{CTD}}{P} \tag{7}$$

TABLE 2 | Nineteen therapeutic drugs for HCC in the Top-30 drugs.

Rank	Drug name	Evidence	Inferred count
1	Daunorubicin	T	42
2	Chrysin	Inferred	34
3	Topiramate	Inferred	8
4	**Securinine**	**NULL**	**NULL**
5	Piperlongumine	Inferred	8
6	Luteolin	Inferred	28
7	Apigenin	Inferred	36
8	Celastrol	Inferred	19
9	Sirolimus	T	68
10	**Mercaptopurine**	**NULL**	**NULL**
11	Genistein	T	93
12	Irinotecan	Inferred	46
13	Sanguinarine	Inferred	5
14	Tyrphostin Ag-825	Inferred	7
15	Decitabine	M	84
16	Camptothecin	Inferred	28
17	**Reserpine**	**NULL**	**NULL**
18	Mycophenolic Acid	Inferred	7
19	Tyrphostin Ag-1478	Inferred	35

Evidence represents a drug-disease association is curated, inferred or not existed in CTD database. Curated associations include three types: marker/mechanism (Evidence = "M"), therapeutic (Evidence = "T"), marker/mechanism & therapeutic (Evidence = "M&T"). If an association is inferred by CTD, Evidence = "inferred," and if it is not existed in CTD, Evidence = "NULL"; Inferred Count represents the number of inferrederence (s) for the curated and inferred associations. If an association is not existed in CTD, Inferred Count = "NULL".

where P represents the number of top-x drugs, i.e., $P = x$; P_{CTD} represents the number of drugs in the top-x drugs, which can be found related with HCC in CTD database.

We find in the top-10 drugs (x = 10), there are 9 drugs associated with HCC in CTD. That is to say, the precision is 0.9. For the top-20 drugs (x = 20), the precision is 0.85 and there are three potentially HCC-related drugs. When x is 30, its precision is 0.83 and we get five potential drugs with HCC. From the **Figure 2**, we notice that with the increase of x, the precision declines and the number of potential drugs increases. We aim to predict relatively more HCC-related drugs with high precision. Then, we choose top-30 (x = 30) drugs for further analysis.

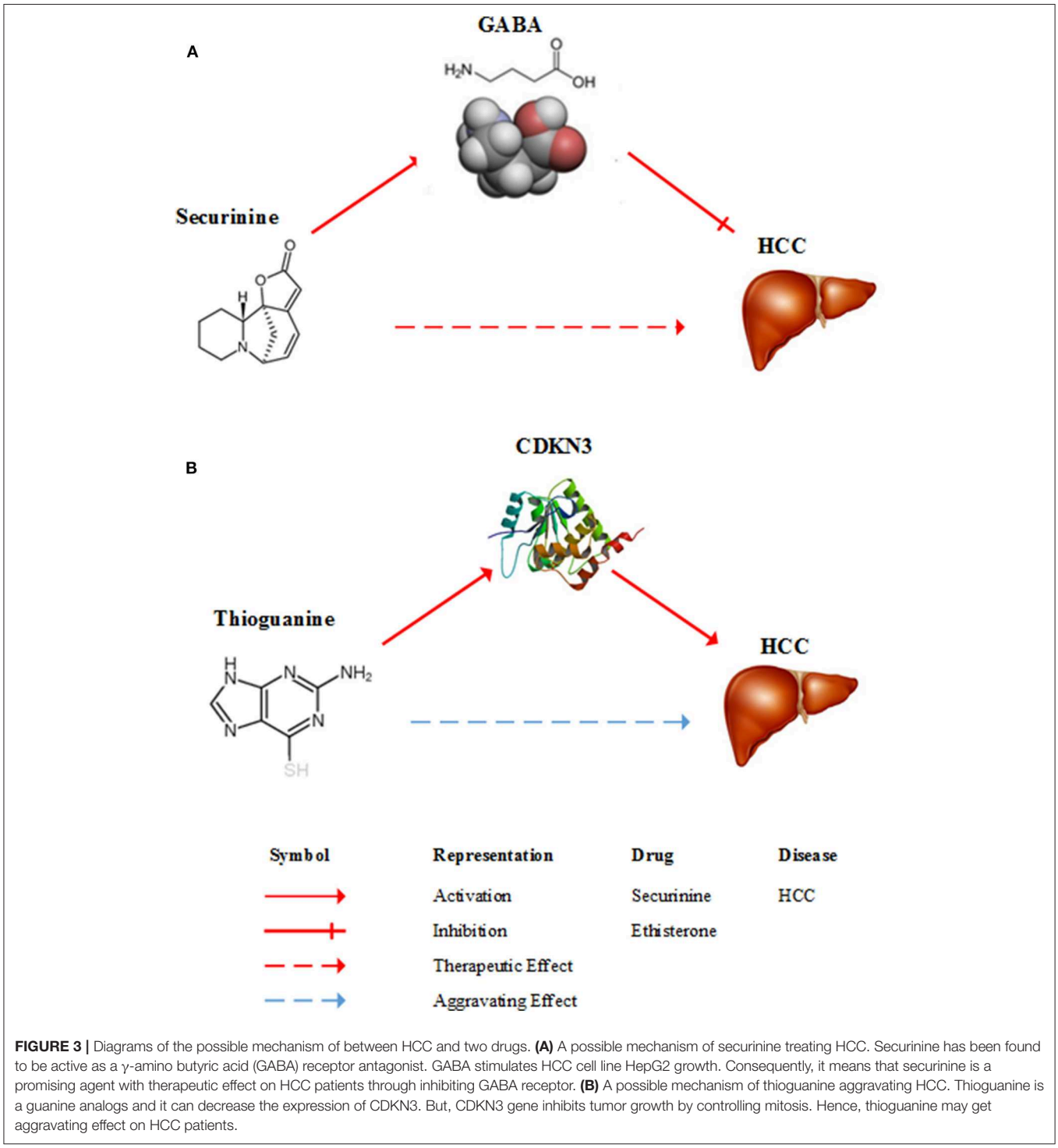


FIGURE 3 | Diagrams of the possible mechanism of between HCC and two drugs. **(A)** A possible mechanism of securinine treating HCC. Securinine has been found to be active as a γ-amino butyric acid (GABA) receptor antagonist. GABA stimulates HCC cell line HepG2 growth. Consequently, it means that securinine is a promising agent with therapeutic effect on HCC patients through inhibiting GABA receptor. **(B)** A possible mechanism of thioguanine aggravating HCC. Thioguanine is a guanine analogs and it can decrease the expression of CDKN3. But, CDKN3 gene inhibits tumor growth by controlling mitosis. Hence, thioguanine may get aggravating effect on HCC patients.

Validating Potentially HCC-Related Drugs Through Pubmed Literature

In the above section, we choose the top-30 drugs (precision = 0.83) for further analysis. There are 19 therapeutic drugs with negative *TS* values in the top-30 drugs, shown in **Table 2.** Sixteen of them can be found having connections with HCC in CTD (Davis et al., 2015). Three of the 16 drugs are marked as therapeutic drug (Rank = 1, Rank = 9, Rank = 11, and Evidence = "T" in **Table 2**) for HCC. Meanwhile, one drug is marked as marker/mechanism drug (Rank = 15, Evidence = "M" in **Table 2**) for HCC and the other 12 inferred drugs are unmarked in CTD. Here, we can indicate these 12 unmarked drugs are possibly therapeutic drugs for HCC. The rest three drugs (Securinine, Mercaptopurine, and Reserpine) are newly predicted ones by our method, which are marked as bold in **Table 2.** Based on PubMed, we analyze the three drugs further.

TABLE 3 | Eleven aggravating drugs for HCC in the Top-30 drugs.

Rank	Drug name	Evidence	Inferred count
1	Cytochalasin B	Inferred	5
2	Exemestane	Inferred	2
3	Spiperone	Inferred	2
4	Cinchonine	Inferred	1
5	Mepacrine	Inferred	8
6	**Tioguanine**	**NULL**	**NULL**
7	**Rifabutin**	**NULL**	**NULL**
8	N-Phenylanthranilic Acid	Inferred	1
9	Valinomycin	Inferred	1
10	Betulin	Inferred	2
11	Puromycin	Inferred	13

TABLE 4 | The relationships of five predicted drugs with known HCC therapeutic drugs in CTD.

Predicted drugs	Known HCC drugs in CTD	Connectivity scores
Securinine	Daunorubicin	0.916
	Troglitazone	0.902
	Paclitaxel	0.844
Mercaptopurine	Estradiol	0.941
	Dexamethasone	0.926
	Sirolimus	0.845
	Troglitazone	0.833
Reserpine	Roxithromycin	0.922
	Resveratrol	0.834
Tioguanine	Genistein	−0.973
	Sirolimus	−0.928
	Indometacin	−0.891
	Paclitaxel	−0.872
Rifabutin	Calcium Folinate	−0.878
	Estradiol	−0.873

The potentially therapeutic drugs of HCC are marked as bold. The other two drugs are potentially aggravating drugs of HCC.

PubMed, a free resource, is developed and maintained by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM). PubMed comprises more than 26 million inferrederences and abstracts on life sciences and biomedical topics.

Securinine (Rank = 4 in **Table 2**), a quinolizine pseudoalkaloid (not from amino acid) from securinega suffurutiosa or securinini nitras, is one of central nervous

TABLE 5 | Pathway enrichment analysis result of five selected drugs.

Drug name	Drug targets	KEGG pathways
Securinine	None	None
Mercaptopurine	HPRT1, PPAT	Purine metabolism; Metabolic pathways; Drug metabolism other enzymes; Alanine aspartate and glutamate metabolism; Biosynthesis of antibiotics
Reserpine	SLC18A2, SLC18A1	Cocaine addiction; Synaptic vesicle cycle; Amphetamine addiction; Serotonergic synapse; Dopaminergic synapse; Parkinson's disease; Alcoholism
Tioguanine	None	None
Rifabutin	rpoA, rpoB, rpoC, HSP90A1, HSP90B1	NOD-like receptor signaling pathway; **Prostate cancer;** Estrogen signaling pathway; Protein processing in endoplasmic reticulum; PI3K-Akt signaling pathway; **Pathways in cancer;** Antigen processing and presentation; Thyroid hormone synthesis; Progesterone-mediated oocyte maturation

The potentially therapeutic drugs of HCC are marked as bold. The other two drugs are potentially aggravating drugs of HCC. "NULL" represents the drug has no targets in DrugBank at present. Thus, its corresponding KEGG pathway is "NULL" too.

TABLE 6 | Twelve enriched tissue-specific KEGG pathways with HCC.

Pathways	Number of HCC-specific genes	*P*-values
Pathways in cancer	**20**	**3.06E-13**
Prostate cancer	**10**	**1.41E-08**
Adherens junction	8	1.44E-06
Endometrial cancer	7	2.15E-06
Colorectal cancer	8	2.62E-06
Apoptosis	8	3.32E-06
Melanoma	7	1.36E-05
Wnt signaling pathway	9	1.41E-05
Cell cycle	7	3.28E-04
Notch signaling pathway	5	4.16E-04
Basal cell carcinoma	5	7.61E-04
Melanogenesis	6	8.61E-04

stimulants and clinically applied to treat amyotrophic lateral sclerosis (ALS) (Buravtseva, 1958), poliomyelitis (Copperman et al., 1973) and multiple sclerosis (Copperman et al., 1974). It is found to be active as a gamma-aminobutyric acid (GABA) receptor antagonist (Perez et al., 2016). GABA is the main inhibitory neurotransmitter of the central nervous system and plays an important role in reducing neuronal excitability throughout the nervous system. Studies show that GABA stimulates HCC cell line HepG2 growth (Lu et al., 2015). Consequently, it means that securinine is a promising agent with therapeutic potential for HCC through inhibiting GABA receptor. **Figure 3A** gives a diagram of the possible mechanism of the treatment of HCC by securinine.

Mercaptopurine(6-MP, Rank = 10 in **Table 2**) is a drug for cancer and autoimmune diseases (Sahasranaman et al., 2008). As a purine analog, mercaptopurine belongs to purine antagonist anti-metabolic drugs (Thackery, 2002). 6-MP nucleotides inhibit the synthesis and metabolism of pure nucleotides by inhibiting an enzyme called phosphoribosyl pyrophosphate (PRPP) amidotransferase PRPP Amidotransferase is a rate-limiting enzyme for pure synthesis (Zollner, 1982). This changes the synthesis and function of RNA and DNA. Mercaptopurine interferes with nucleotide conversion and glycoprotein synthesis. This makes the mercaptopurine can effectively inhibit the synthesis of DNA, thereby inhibiting the growth of tumor cells (Cara et al., 2004). At present, although there is no direct experiment that mercaptopurine can inhibit the growth of HCC cells, it is used to treat acute lymphoblastic leukemia (ALL), chronic myeloid leukemia (CML), Crohn's disease and ulcerative colitis (Joint Formulary Committee, 2011). In summary, mercaptopurine is likely to achieve a certain effect on HCC.

Reserpine (Rank = 17 in **Table 2**) is an antipsychotic and antihypertensive drug (Bridgwater and Sherwood, 1960) used to control hypertension and relieve psychotic symptoms (Arnt et al., 1985). The results of Gwak et al. (2009) showed that reserpine could reduce the expression level of CCND1 gene and its encoded protein. The CCND1 gene encodes the cyclin D1 protein. Cyclin D1 protein is a member of the circulatory protein family, involved in regulating cell cycle progression. This protein plays a key role during the transition from the G1 phase, in which the cell grows, to the S phase, during which DNA is replicated. Overexpression

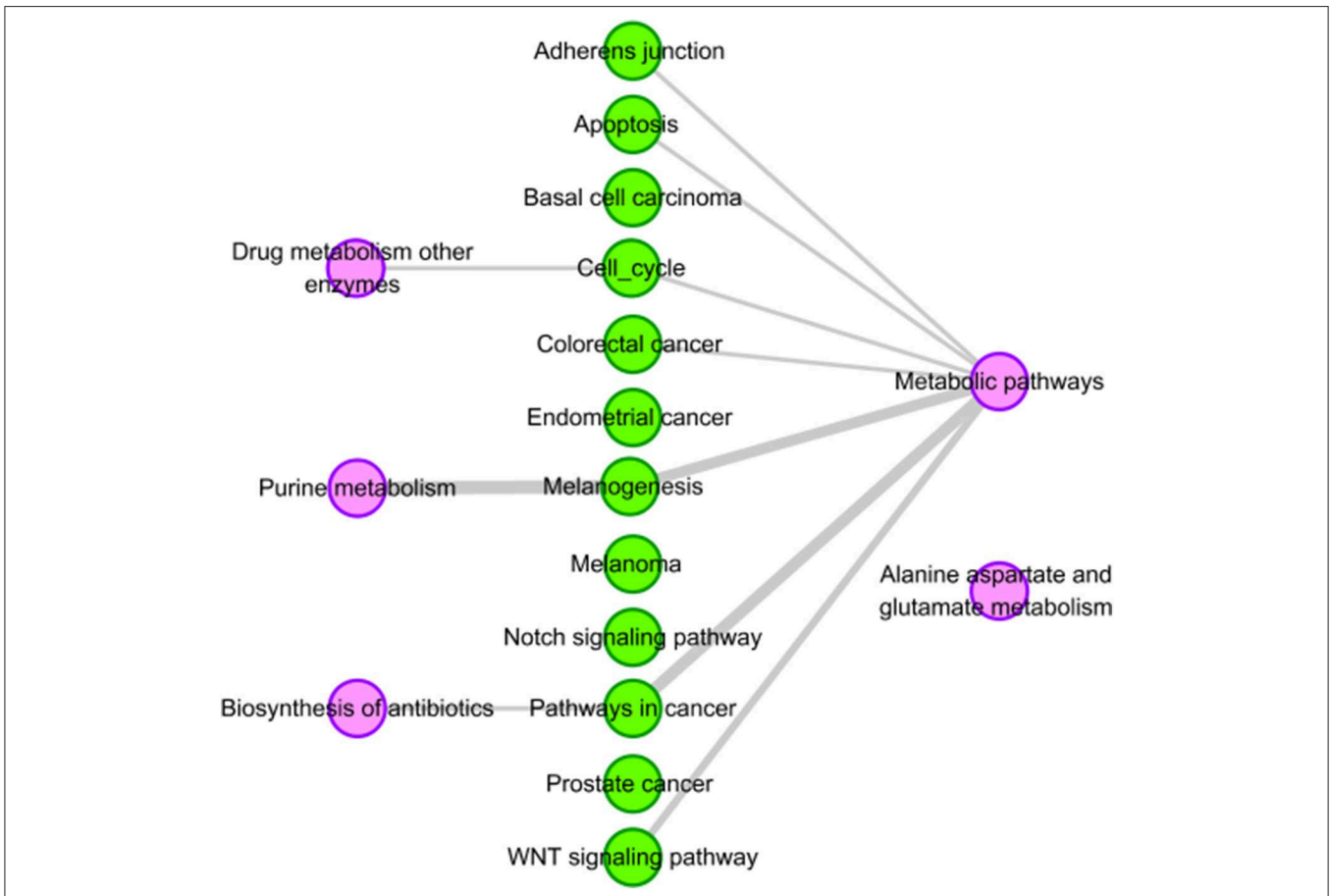


FIGURE 4 | Pathway analysis of mercaptopurine. The purple circles represent pathways related to mercaptopurine, and the green circles represent the tissue-specific KEGG pathways of HCC. The gray edges indicate that there are common genes between two pathways, and the more genes there are, the wider the edges in the network.

of this protein allows cells to be easily crossed G1/S checkpoint that limits the growth of cells, which promotes tumor hyperplasia and is therefore considered to be an oncoprotein (Donnellan and Chetty, 1998). Some studies have found that CCND1 gene is over-expressed in HCC (Xu et al., 2004). Thus, reserpine can potentially be used as an agent against HCC.

The other 11 drugs with negative *TS* values are shown in **Table 3**. They are possible to aggravate HCC. Nine of them have been found having relationships with HCC in CTD database and we can infer these relationships are possibly negative. The remaining 2 drugs (Tioguanine, Rifabutin) are newly potential drugs for aggravating HCC marked as bold in **Table 3**. We will investigate the two drugs (Tioguanine, Rifabutin) based on PubMed.

Tioguanine, also known as thioguanine, (Rank = 6 in **Table 3**) is a guanine analogs, with cell cycle specificity, for the S cycle of the strongest cell sensitivity. In addition, thioguanine can inhibit the synthesis of guanosine nucleoside, by inhibiting the biological activity of guanylate kinase, the drug can inhibit the guanosine monophosphate (GMP) phosphoric acid to guanosine bisphosphate (GDP) transformation process (Golan, 2011). Thibird is a drug used to treat acute myeloid leukemia (AML) (Gill et al., 1982), acute lymphoblastic leukemia (ALL) (Marmont and Damasio, 1973) and chronic myeloid leukemia (CML) (Yang et al., 2006). In 2005, Ganter et al. showed that CDKN3 expression was significantly decreased after a period of administration of thioguanine (Ganter et al., 2005). The CDKN3 gene inhibits tumor growth by controlling mitosis, which is a tumor suppressor gene (Nalepa et al., 2013). Dai et al. found that CDKN3 expression in patients with HCC was significantly lower than that in normal humans. CDKN3 knockout experiments indicated that CDKN3 could inhibit tumor growth (Dai et al., 2016). A possible mechanism of thioguanine aggravating HCC is shown in **Figure 3B**. Therefore, in order to ensure the effectiveness of the treatment, clinical patients should avoid HCC patients taking thioguanine.

Rifabutin (Rank = 7 in **Table 3**) is a piperazine-containing rifamycin derivative, the drug has a broad spectrum of antibacterial activity. It can able to bind to the β-subunit of RNA polymerase and inhibit RNA polymerase activity, thereby reducing the number of RNA synthesis of bacterial (Beard, 2001). Rifabutin has been approved to prevent and treat disseminated infections of mycobacterium mycobacterium complex (MAC) carried by HIV-infected persons (Arevalo et al., 1997), and it is also used to treat multidrug-resistant tuberculosis (Skolik et al., 2005). Kobayashi et al. find that rifabutine will lead to an increase in the expression of cytochrome P450 3A4 (CYP3A4) in liver tissue (Nakajima et al., 2011). CYP3A4 is an important metabolic enzyme, belongs to the cytochrome P450 family. It is also the most important component of adult liver microsomes CYP450, this gene is expressed in the intestinal, liver and kidney (Hashimoto et al., 1993). However, Fanni et al. find a significant increase of expression of CYP3A4 in HCC patients and overexpression of CYP3A4 gene could result in

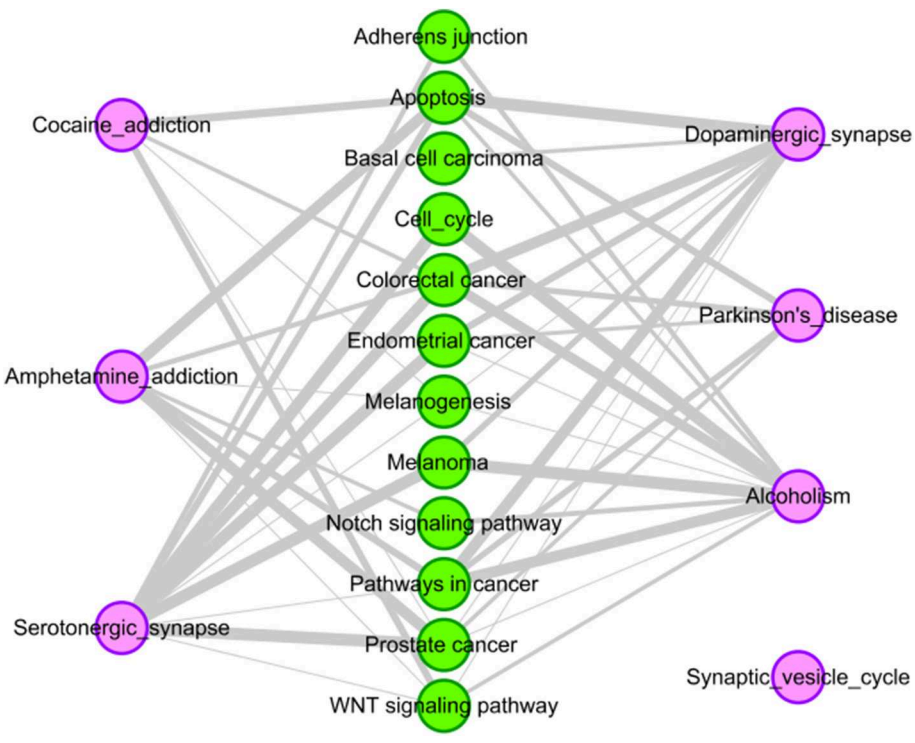


FIGURE 5 | Pathway analysis of reserpine. The purple circles represent pathways of reserpine, and the green circles represent the tissue-specific KEGG pathways of HCC. The gray edges indicate that there are common genes between two pathways, and the more genes there are, the wider the edges in the network.

drug degradation or even a decreased therapeutic effect (Fanni et al., 2016). Therefore, for both suffering from HCC and tuberculosis patients, doctors should avoid using rifabutin to treat tuberculosis.

Analyzing Potentially HCC-Related Drugs Through CMap Database

The CMap database can not only be applied to calculate drug-disease correlations, but also can be used to identify connections between two drugs. In particular, for a same disease, if two drugs have strongly positive relationship, they may have similar effects on this disease. On the contrary, if their relationship is negative, they may have opposite effects. In this section, we further analyze the five predicted drugs (three therapeutic drugs shown in **Table 2**: securinine, mercaptopurine and reserpine; two aggravating drugs shown in **Table 3**: tioguanine and rifabutin) based on CMap and estimate their correlations [evaluated by formula (6)] with known HCC drugs marked as "therapeutic" in CTD database. The results are shown in **Table 4**.

For the three potentially therapeutic drugs (securinine, mercaptopurine and reserpine) marked as bold in **Table 4**, we find that they all have strong positive correlation with known drugs for HCC. Securinine yields highly positive connectivity score [calculated by formula (6)] with drugs daunorubicin, troglitazone and paclitaxel. Mercaptopurine is found having strongly positive relationships with drugs estradiol, dexamethasone, sirolimus, and troglitazone. Reserpine gets high positive connectivity scores with drugs roxithromycin and resveratrol. For the two potentially aggravating drugs (tioguanine and rifabutin) in **Table 4**, they all have negative relationship with known HCC drugs. Tioguanine has high negative connectivity scores with drugs genistein, sirolimus, indomethacin, and paclitaxel. Rifabutin have clear negative connection scores with drugs calcium folinate and estradiol.

Overlap Between Pathways Associated With Predicted Drugs and HCC-Related Tissue-Specific Pathways

In this part, we further analyze the relationship between these five drugs (three therapeutic drugs: securinine, mercaptopurine and reserpine; two aggravating drugs: tioguanine and rifabutin) and HCC from the point of view of drug targets. First, we get the target set of drugs from DrugBank (Law et al., 2014) because DrugBank contains the most complete information on drug and drug targets. Then, we use DAVID (Huang et al., 2009) to obtain all the KEGG (Kanehisa et al., 2010) pathways of the drug target. The p-value is set to be less than or equal to 0.05. The results are shown in **Table 6**.

From **Table 5**, it can be seen that securinine and tioguanine have no corresponding target information in the DrugBank database. So we can't enrich their associated pathways. Mercaptopurine has two drug targets, and we find five KEGG

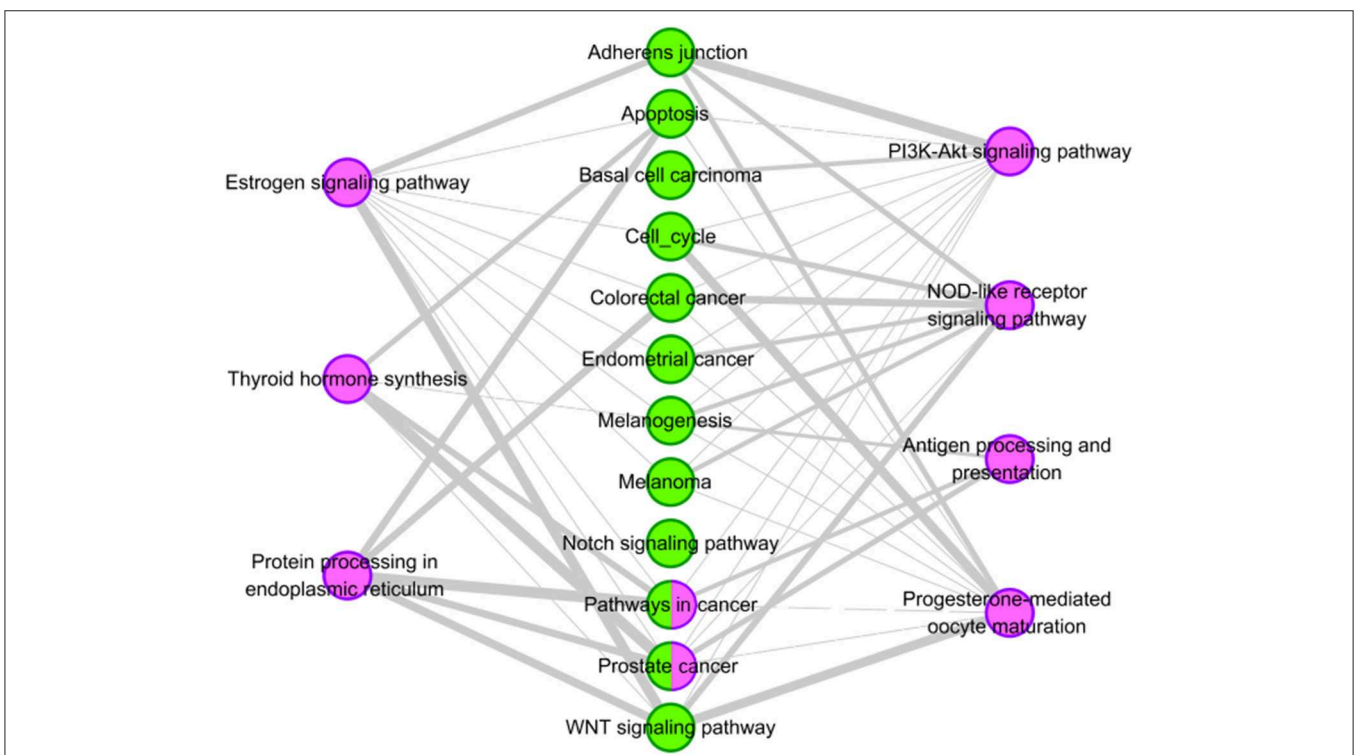


FIGURE 6 | Pathway analysis of rifabutine. The purple circles represent pathways of the rifabutine, and the green circles represent tissue-specific KEGG pathways of HCC. Nodes with two colors represent overlapping pathways for rifabutine and HCC. The gray edges indicate that there are common genes between two pathways, and the more genes there are, the wider the edges in the network.

pathways related to them. Reserpine has two drug targets, which are included in seven KEGG pathways. Rifabutine has five drug targets, and nine KEGG pathways are enriched to them.

In order to obtain the tissue-specific KEGG pathways of HCC, firstly, the eight genes (see **Table 1**) related to HCC are extended through obtaining their direct neighbors in liver-specific protein-protein interaction (PPI) network got from GIANT (Greene et al., 2015). Then, we obtain a subnetwork from the liver PPI network, which contains 57 genes and 838 edges with weight $\geq$ 0.1. Finally, by using DAVID tool, we obtain 12 KEGG pathways related to the 57 genes (see **Table 6**). The parameters of DAVID are fixed as: p-value = 0.001 and count = 5.

We find that there are four pathways related to mercaptopurine have common genes with the 12 tissue-specific KEGG pathway of HCC. The interactions between the four pathways and the 12 tissue-specific KEGG pathway of HCC is shown in **Figure 4**. The gray edges indicate that there are common genes between two pathways, and the more genes there are, the wider the edges in the network. "Metabolic pathways" have common genes with seven tissue-specific KEGG pathways of HCC. Though there is only one edge between "purine metabolism" and HCC related pathway, the edge is very wide, indicating that there are a lot of common genes. These overlap genes between the pathways of mercaptopurine and HCC tissue-specific KEGG pathways show that mercaptopurine has a potential effect on treating HCC.

For drug reserpine, there are six pathways have common genes with the 12 tissue-specific KEGG pathway of HCC. Their relationships are shown **Figure 5**. For example, "serotonergic synapse" has common genes with ten pathways of HCC. "Dopaminergic synapse" has common genes with nine pathways of HCC. Overall, drug reserpine has more overlapping pathways with HCC, and more genes overlap between pathways. The results indicate that drug reserpine is likely to become the treatment of HCC.

For the potential aggravating drug rifabutine, we also analyze its pathway overlap with HCC. We try to explain the possible reasons for its aggravating HCC in terms of pathway overlap. Two pathways of rifabutine ("Pathways in cancer" and "Prostate cancer") are overlapped with pathways of HCC highlighted in **Table 6**. The interactions between the pathways and the 12 tissue-specific pathways of HCC is shown in **Figure 6**. Two overlapping pathway nodes are colored in two colors (purple and green) in **Figure 6**. We find the pathways of rifabutine have a very large number of overlapping genes with the pathways of HCC. This shows a strong correlation between rifabutine and HCC, confirming our prediction on the other hand.

DISCUSSIONS

We propose a method based on the combination of gene mutation data and differential expression data. First, we select the feature genes of hepatocellular carcinoma (HCC) that frequently mutated in most samples of HCC based on human somatic mutation data. Then, the gene expression data of HCC in TCGA are combined to classify the genes related to HCC. Finally, the therapeutic score (*TS*) of each drug is calculated based on the kolmogorov-smirnov statistical method. By this method, five drugs associated with HCC are obtained, including three drugs that could be the potential treatment for HCC and two drugs that might have side effect on HCC. There are advantages in our method. First, we take into account the essential impact of genetic changes on HCC. Secondly, we integrate multiple data to define the type of a gene. Finally, our method can clearly distinguish positive and negative relationships between drugs and HCC.

In the future, as more and more drug-related data continues to be generated, such as cell lines, gene expression and mutation data, we will further improve our computational model and predict more accurate potential drugs for the treatment of HCC.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

LY designed the study. All authors analyzed the data, interpreted the results, wrote the manuscript, read and approved the final manuscript.

FUNDING

This work was supported in part by the National Key Research and Development Program of China (No. 2018YFC0910403), National Natural Science Foundation of China (Nos. 61672406, 61532014, 61672407, and 61772395).

REFERENCES

Abenavoli, L., Masarone, M., Federico, A., Rosato, V., Dallio, M., Loguercio, C., et al. (2016). Alcoholic hepatitis: pathogenesis, diagnosis and treatment. *Rev. Recent Clin. Trials* 11, 159–166. doi: 10.2174/1574887111666160724183409

Adams, C. P., and Brantner, V. V. (2006). Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff.* 25, 420–428. doi: 10.1377/hlthaff.25.2.420

Arevalo, J. F., Russack, V., and Freeman, W. R. (1997). New ophthalmic manifestations of presumed rifabutin-related uveitis. *Ophthalmic Surg. Lasers* 28, 321–324.

Arnt, J., Christensen, A. V., and Hyttel, J. (1985). Pharmacology in vivo of the phenylindan derivative, Lu 19–005, a new potent inhibitor of dopamine, noradrenaline and 5-hydroxytryptamine uptake in rat brain. *Naunyn Schmiedebergs. Arch. Pharmacol.* 329, 101–107. doi: 10.1007/BF00501197

Ashburn, T. T., and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683. doi: 10.1038/nrd1468

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Beard, E. L. Jr. (2001). The American society of health system pharmacists. *JONAS Healthc Law Ethics Regul.* 3, 78–79. doi: 10.1097/00128488-200109000-00003

Bridgwater, W., and Sherwood, E. J. (1960). *The Columbia Encyclopedia*. Parents' Magazine's Education Press, a division of the Parent's Institute.

Buravtseva, G. R. (1958). [Result of application of securinine in acute poliomyelitis]. *Farmakol. Toksikol.* 21, 7–12.

Cara, C. J., Pena, A. S., Sans, M., Rodrigo, L., Guerrero-Esteo, M., Hinojosa, J., et al. (2004). Reviewing the mechanism of action of thiopurine drugs: towards a new paradigm in clinical practice. *Med. Sci. Monit.* 10, RA247–R254.

Chakravarthi, B. V., Nepal, S., and Varambally, S. (2016). Genomic and epigenomic alterations in cancer. *Am. J. Pathol.* 186, 1724–1735. doi: 10.1016/j.ajpath.2016.02.023

Chang, K. M., and Liu, M. (2016). Chronic hepatitis B: immune pathogenesis and emerging immunotherapeutics. *Curr. Opin. Pharmacol.* 30, 93–105. doi: 10.1016/j.coph.2016.07.013

Charrez, B., Qiao, L., and Hebbard, L. (2016). Hepatocellular carcinoma and non-alcoholic steatohepatitis: the state of play. *World J. Gastroenterol.* 22, 2494–2502. doi: 10.3748/wjg.v22.i8.2494

Copperman, R., Copperman, G., and Der Marderosian, A. (1973). From Asia securinine–a central nervous stimulant is used in treatment of amytrophic lateral sclerosis. *Pa. Med.* 76, 36–41.

Copperman, R., Copperman, G., and Marderosian, A. D. (1974). Letter: securinine. *JAMA* 228:288. doi: 10.1001/jama.228.3.288c

Dai, W., Miao, H. L., Fang, S., Fang, T., Chen, N. P., and Li, M. Y. (2016). CDKN3 expression is negatively associated with pathological tumor stage and CDKN3 inhibition promotes cell survival in hepatocellular carcinoma. *Mol. Med. Rep.* 14, 1509–1514. doi: 10.3892/mmr.2016.5410

Dalman, M. R., Deeter, A., Nimishakavi, G., and Duan, Z. H. (2012). Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics* 13 (Suppl. 2):S11. doi: 10.1186/1471-2105-13-S2-S11

Davis, A. P., Grondin, C. J., Lennon-Hopkins, K., Saraceni-Richards, C., Sciaky, D., King, B. L., et al. (2015). The comparative toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.* 43, D914–D920. doi: 10.1093/nar/gku935

Ding, Y., Tang, J., and Guo, F. (2016). Identification of Protein–Protein interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17:1623. doi: 10.3390/ijms17101623

Donnellan, R., and Chetty, R. (1998). Cyclin D1 and human neoplasia. *Mol. Pathol.* 51, 1–7. doi: 10.1136/mp.51.1.1

Dudley, J. T., Sirota, M., Shenoy, M., Pai, R. K., Roedder, S., Chiang, A. P., et al. (2011). Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3:96ra76. doi: 10.1126/scitranslmed.3002648

El-Serag, H. B., and Mason, A. C. (1999). Rising incidence of hepatocellular carcinoma in the United States. *N. Engl. J. Med.* 340, 745–750. doi: 10.1056/NEJM199903113401001

Fanni, D., Manchia, M., Lai, F., Gerosa, C., Ambu, R., and Faa, G. (2016). Immunohistochemical markers of CYP3A4 and CYP3A7: a new tool towards personalized pharmacotherapy of hepatocellular carcinoma. *Eur. J. Histochem.* 60:2614. doi: 10.4081/ejh.2016.2614

Fathima, A. J., Murugaboopathi, G., and Selvam, P. (2018). Pharmacophore Mapping of ligand based virtual screening, molecular docking and molecular dynamic simulation studies for finding potent NS2B/NS3 protease inhibitors as potential anti-dengue drug compounds. *Curr. Bioinform.* 13, 606–616. doi: 10.2174/1574893613666180118105659

Ganter, B., Tugendreich, S., Pearson, C. I., Ayanoglu, E., Baumhueter, S., Bostian, K. A., et al. (2005). Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* 119, 219–244. doi: 10.1016/j.jbiotec.2005.03.022

Gill, R. A., Onstad, G. R., Cardamone, J. M., Maneval, D. C., and Sumner, H. W. (1982). Hepatic veno-occlusive disease caused by 6-thioguanine. *Ann. Intern. Med.* 96, 58–60. doi: 10.7326/0003-4819-96-1-58

Golan, D. E. (2011). *Principles of Pharmacology: The Pathophysiologic Basis of Drug Therapy*. Lippincott, Williams and Wilkins.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Gwak, J., Song, T., Song, J. Y., Yun, Y. S., Choi, I. W., Jeong, Y., et al. (2009). Isoreserpine promotes beta-catenin degradation via Siah-1 up-regulation in HCT116 colon cancer cells. *Biochem. Biophys. Res. Commun.* 387, 444–449. doi: 10.1016/j.bbrc.2009.07.027

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Hashimoto, H., Toide, K., Kitamura, R., Fujita, M., Tagawa, S., Itoh, S., et al. (1993). Gene structure of CYP3A4, an adult-specific form of cytochrome P450 in human livers, and its transcriptional control. *Eur. J. Biochem.* 218, 585–595. doi: 10.1111/j.1432-1033.1993.tb18412.x

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jahchan, N. S., Dudley, J. T., Mazur, P. K., Flores, N., Yang, D., Palmerton, A., et al. (2013). A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.* 3, 1364–1377. doi: 10.1158/2159-8290.CD-13-0183

Joint Formulary Committee (2011). *Britain RPSoG: British National Formulary 61*. Royal Pharmaceutical Society.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi: 10.1093/nar/gkp896

Kew, M. C. (2013). Aflatoxins as a cause of hepatocellular carcinoma. *J Gastrointest Liver*22, 305–310.

Kitao, A., Matsui, O., Yoneda, N., Kozaka, K., Kobayashi, S., Sanada, J., et al. (2015). Hepatocellular carcinoma with beta-catenin mutation: imaging and pathologic characteristics. *Radiology* 275, 708–717. doi: 10.1148/radiol.14141315

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., et al. (2014). DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42, D1091–D1097. doi: 10.1093/nar/gkt1068

Lingala, S., and Ghany, M. G. (2015). Natural history of hepatitis, C. *Gastroenterol. Clin. North Am.* 44, 717–734. doi: 10.1016/j.gtc.2015.07.003

Lu, Y. H., Huang, C., Gao, L., Xu, Y. J., Chia, S. E., Chen, S. S., et al. (2015). Identification of serum biomarkers associated with hepatitis B virus-related hepatocellular carcinoma and liver cirrhosis using mass-spectrometry-based metabolomics. *Metabolomics* 11, 1526–1538. doi: 10.1007/s11306-015-0804-9

Marmont, A. M., and Damasio, E. E. (1973). Letter: Neurotoxicity of intrathecal chemotherapy for leukaemia. *Br. Med. J.* 4:47. doi: 10.1136/bmj.4.5883.47-a

Nakajima, A., Fukami, T., Kobayashi, Y., Watanabe, A., Nakajima, M., and Yokoi, T. (2011). Human arylacetamide deacetylase is responsible for deacetylation of rifamycins: Rifampicin, rifabutin, and rifapentine. *Biochem. Pharmacol.* 82, 1747–1756. doi: 10.1016/j.bcp.2011.08.003

Nalepa, G., Barnholtz-Sloan, J., Enzor, R., Dey, D., He, Y., Gehlhausen, J. R., et al. (2013). The tumor suppressor CDKN3 controls mitosis. *J. Cell Biol.* 201, 997–1012. doi: 10.1083/jcb.201205125

Nollet, F., Berx, G., Molemans, F., and van Roy, F. (1996). Genomic organization of the human beta-catenin gene (CTNNB1). *Genomics* 32, 413–424. doi: 10.1006/geno.1996.0136

Perez, M., Ayad, T., Maillos, P., Poughon, V., Fahy, J., and Ratovelomanana-Vidal, V. (2016). Synthesis and biological evaluation of new securinine analogues as potential anticancer agents. *Eur. J. Med. Chem.* 109, 287–293. doi: 10.1016/j.ejmech.2016.01.007

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database-2009 update. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Sahasranaman, S., Howard, D., and Roy, S. (2008). Clinical pharmacology and pharmacogenetics of thiopurines. *Eur. J. Clin. Pharmacol.* 64, 753–767. doi: 10.1007/s00228-008-0478-6

Skolik, S., Willermain, F., and Caspers, L. E. (2005). Rifabutin-associated panuveitis with retinal vasculitis in pulmonary tuberculosis. *Ocul. Immunol. Inflamm.* 13, 483–485. doi: 10.1080/092739405909 51115

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Thackery, E. (2002). *The Gale Encyclopedia of Cancer*. Gale Cengage.

Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77. doi: 10.5114/wo.2014.47136

Wei, L., Xing, P., Tang, J., Zou, Q., and (2017b). PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans. NanoBioscience*. 16, 240–247. doi: 10.1109/TNB.2017.2661756

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017a). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Xu, J. M., Wen, J. M., Zhang, M., Lu, G. L., Wu, L. Z., and Wang, W. S. (2004). [A study of gene amplification and expression of cyclin D1 in hepatocellular carcinoma]. *Zhonghua Bing Li Xue Za Zhi* 33, 26–30. doi: 10.3760/j.issn:0529-5807.2004.01.007

Yang, M. Y., Chang, J. G., Lin, P. M., Tang, K. P., Chen, Y. H., Lin, H. Y., et al. (2006). Downregulation of circadian clock genes in chronic myeloid leukemia: alternative methylation pattern of hPER3. *Cancer Sci.* 97, 1298–1307. doi: 10.1111/j.1349-7006.2006.00331.x

Yu, L., Huang, J. B., Ma, Z. X., Zhang, J., Zou, Y. P., and Gao, L. (2015). Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8:13. doi: 10.1186/1755-8794-8-S2-S2

Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., et al. (2017b). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE Acm Trans. Comput. Biol. Bioinformatics*14, 966–977. doi: 10.1109/TCBB.2016.2550453

Yu, L., Wang, B., Ma, X., and Gao, L. (2016). The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC Syst. Biol.* 10:111. doi: 10.1186/s12918-016-0364-2

Yu, L., Yao, S. Y., Gao, L., and Zha, Y. H. (2019). Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments. *Front. Genet* 9:745. doi: 10.3389/fgene.2018.00745

Yu, L., Zhao, J., and Gao, L. (2017a). Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artif. Intell. Med.* 77, 53–63. doi: 10.1016/j.artmed.2017.03.009

Yu, L., Zhao, J., and Gao, L. (2018). Predicting potential drugs for breast cancer based on miRNA and tissue specificity. *Int. J. Biol. Sci.* 14, 971–980. doi: 10.7150/ijbs.23350

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019). SFLLN: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inf. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, W., Yue, X., Chen, Y., Lin, W., Li, B., Liu, F., Li, X., and (2017). "Predicting drug-disease associations based on the known association bipartite network," in *IEEE International Conference on Bioinformatics and Biomedicine* (Kansas City, MO), 503–509. doi: 10.1109/BIBM.2017.8217698

Zhang, W., Yue, X., Huang, F., Liu, R., Chen, Y., and Ruan, C. (2018a). Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* 145, 51–59. doi: 10.1016/j.ymeth.2018.06.001

Zhang, W., Yue, X., Lin, W., Wu, W., Liu, R., Huang, F., et al. (2018b). Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics* 19:233. doi: 10.1186/s12859-018-2220-4

Zollner, N. (1982). Purine and pyrimidine metabolism. *Proc. Nutr. Soc.* 41, 329–342. doi: 10.1079/PNS19820048

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 30 January 2020
doi: 10.3389/fgene.2019.01400

Meta-Analysis of SNP-Environment Interaction With Overlapping Data

Qinqin Jin[1,2] *and Gang Shi*[1]*

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China, [2] Applied Science College, Taiyuan University of Science and Technology, Taiyuan, China

Meta-analysis, which combines the results of multiple studies, is an important analytical method in genome-wide association studies. In genome-wide association studies practice, studies employing meta-analysis may have overlapping data, which could yield false positive results. Recent studies have proposed models to handle the issue of overlapping data when testing the genetic main effect of single nucleotide polymorphism. However, there is still no meta-analysis method for testing gene-environment interaction when overlapping data exist. Inspired by the methods of testing the main effect of gene with overlapping data, we proposed an overlapping meta-regulation method to address the issue in testing the gene-environment interaction. We generalized the covariance matrices of the regular meta-regression model by employing Lin's and Han's correlation structures to incorporate the correlations introduced by the overlapping data. Based on our proposed models, we further provided statistical significance tests of the gene-environment interaction as well as joint effects of the gene main effect and the interaction. Through simulations, we examined type I errors and statistical powers of our proposed methods at different levels of data overlap among studies. We demonstrated that our method well controls the type I error and simultaneously achieves statistical power comparable with the method that removes overlapping samples *a priori* before the meta-analysis, i.e., the splitting method. On the other hand, ignoring overlapping data will inflate the type I error. Unlike the splitting method that requires individual-level genotype and phenotype data, our proposed method for testing gene-environment interaction handles the issue of overlapping data effectively and statistically efficiently at the meta-analysis level.

Keywords: meta-regression, meta-analysis, gene-environment interaction, overlapping data, correlation matrix

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Yuriy L. Orlov,
First Moscow State Medical University, Russia
Yang Zhao,
Nanjing Medical University, China

***Correspondence:**
Gang Shi
gshi@xidian.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *24 September 2019*
Accepted: *23 December 2019*
Published: *30 January 2020*

Citation:
Jin Q and Shi G (2020) Meta-Analysis of SNP-Environment Interaction With Overlapping Data. Front. Genet. 10:1400. doi: 10.3389/fgene.2019.01400

INTRODUCTION

Numerous associations between human traits or diseases and single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (GWAS) (Manolio, 2010). Meta-analysis combines the results from multiple studies to increase the effective sample size and statistical power of the association test (Fleiss, 1993; Borenstein et al., 2009). It has played an important role in finding the genetic architectures of complex traits and diseases.

Many meta-analysis methods are used in GWAS (Eleftheria and John, 2009). The fixed effect model is a commonly used method. It assumes that there are the same effect sizes across different studies. This method is effective if the heterogeneity among studies is small (Pfeiffer et al., 2009). Other methods, such as random effect models, are used in GWAS as well. They assume that the effect sizes of the studies follow a probability distribution due to the heterogeneity (Pereira et al., 2009). Recently, we proposed a new random effect method for testing the interaction between SNP and environment factor, which provides a higher power than the fixed effect methods when heterogeneity is large (Jin and Shi, 2019). The P-value based method (Fisher, 1967) was widely used earlier and has been abandoned because it does not include directions of effects under test; thus, it cannot provide an overall estimation of the effect size. The application of this method may lead to false positive results (Evangelou and Ioannidis, 2013). The Z scores method considers the direction of effect and its weight is estimated as the square root of the sample size of each study (Evangelou and Ioannidis, 2013). Bayesian methods (Kraft and Haiman, 2010) depend on the assumption of the prior distribution of the parameters and are usually computationally intensive. The subset method (Morris, 2011; Wen and Stephens, 2014) is similar to the fixed effect methods; however, it assumes that the effect exists only in a subset of the studies. All these classical methods assume that the studies have no overlapping samples, thus helping maintain independence among the summary statistics of the studies.

However, in GWAS practice, overlapping data between studies may occur. This may be caused inadvertently or intentionally by researchers. Spurious association may be achieved if overlapping data exist and are ignored in the meta-analysis (Lin and Sullivan, 2009; Han et al., 2016). Recently, meta-analysis methods, such as the P-value based method (Zaykin and Kozbur, 2010), subset method (Bhattacharjee et al., 2012), Bayesian method (Wen, 2014), fixed effect method (Lin and Sullivan, 2009), and random effect methods (Han and Eskin, 2011; Han et al., 2016) have been proposed for handling the overlapping data issue. All existing methods are for testing the SNP main effect. Lin's method (Lin and Sullivan, 2009) is proposed for combining the results of case-control studies. It has been shown to yield higher and more robust power than the splitting method that removes the overlapped data in studies before calculating the study-level summary statistics. Han's method (Han et al., 2016) involves modeling the covariance matrix of the estimated effects due to the overlapping data in fixed or random effect models and transforming the covariance matrix to be diagonal. The transformed matrix can then be synthesized by regular methods that assume independent data among studies.

Meta-regression (MR) (Xu et al., 2013) is an efficient meta-analysis method for testing SNP-environment interaction assuming independent data among studies. In MR, subjects in each study are divided into groups by the distribution of an environment variable. Then, the SNP main effects, standard errors, and the average environmental variables in each group are estimated using linear or logistic regressions. The SNP main effects and environmental variables across all groups are then collected and synthesized by MR. The overall main effect of the SNP, the effect of SNP-environment interaction, and the corresponding standard errors can be derived. The MR method is also shown to be robust when confounding effects exist (Shi and Nehorai, 2017).

Many complex diseases or traits are owing to the combination of effects of genetic factors, environment factors, and gene-environment interactions and involve in complex regulatory networks (Chen et al., 2019; Chen et al., 2019). Consider CDKN2A/B-rs10811661 as an example, which is associated with dyslipidemia. Researchers used CC/CT genotypes with a low-energy diet and a high frequency of exercise as the control group to study the effect of the interaction between rs10811661 gene polymorphism and energy intake and exercise on the level of blood lipid. The study found that the incidence of hypercholesterolemia was approximately 2 times higher in the TT genotype than in the control group and 1.5 times higher in the CC/CT genotype than in the control group (Mehramiz et al., 2018). The analysis of the genes and environment interactions can provide new insight into complex traits or disease mechanisms. However, a meta-analysis of SNP-environment interaction method with overlapping data does not exist currently. Data have to be split in studies such that every study contributes non-overlapped samples, i.e., the so-called splitting method. The splitting method requires the study-level genotype and phenotype data, which is usually unavailable for the meta-analysis. In addition, different ways of splitting samples may lead to different results.

In this paper, inspired by Lin's method (Lin and Sullivan, 2009) and Han's decoupling method (Han et al., 2016) for testing the SNP main effect, and based on MR, we propose the overlapping MR (OMR) method, which is a fixed effect MR model designed especially for handling overlapping data. The remainder of this paper is organized as follows: In the materials and methods section, we present the correlation matrices for the OMR method and then the method for testing the SNP-environment interaction. We also provide the relationship between MR and OMR. In the *Results* section, we simulate numerical examples and use them to examine the type I error and power of our method and the splitting method. We also show that the type I error is inflated with regular MR without considering overlapping samples. In the discussion and conclusion sections, we discuss the results and conclude the paper.

MATERIALS AND METHODS

Based on Lin's and Han's correlation structures (Lin and Sullivan, 2009; Han et al., 2016), we generalized regular MR model for independent studies to consider studies with correlated summary statistics due to overlapping data. To describe our method clearly, we first briefly introduce the regular MR method.

Regular MR Method

Before the MR analysis, individuals in each study are first stratified into several groups according to their environmental measurements. The main effects of SNP at the group level can be estimated via linear regression as follows:

$$Y = \beta_0 + \beta_G G + \beta_E E + \epsilon,$$

where Y is a quantitative phenotype, G is the code of the SNP, and E is the environmental measurement.

Assume that $\hat{\beta}$ is the estimate of the SNP main effect, and $\hat{\beta}_{ij}$ is the estimate of the SNP main effect for the i-th study and the j-th group where i= 1,2,…,n, j=1,2,…,n_i, The symbol n is the number of studies and n_i denotes the number of groups in the i-th study, and $\hat{e}_{ij}$ denotes the standard error in the j-th group of the i-th study. The mean environmental measurement in the j-th group of the i-th study is E_{ij}. α is the regression coefficient vector of interest. The symbol X is the interest matrix and X_i is the interest matrix for the i-th study. ε is the standard error matrix and the ε_i is the standard error matrix for i-th study. In MR, the SNP effect is regressed on the environmental factor as follows:

$$\hat{\beta} = X\alpha + \epsilon, \tag{1}$$

where

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{pmatrix}, \quad \hat{\beta}_i = \begin{pmatrix} \hat{\beta}_{i1} \\ \hat{\beta}_{i2} \\ \vdots \\ \hat{\beta}_{in_i} \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}, X_i = \begin{pmatrix} 1 & E_{i1} \\ 1 & E_{i2} \\ \vdots & \vdots \\ 1 & E_{in_i} \end{pmatrix},$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \epsilon_i = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in_i} \end{pmatrix}, \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_n \end{pmatrix},$$

$$\Sigma_i = \begin{pmatrix} \hat{e}_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{e}_{in_i} \end{pmatrix}$$

and $\epsilon_{ij} \sim N(0, \hat{e}_{ij}), i = 1, 2, \ldots, n, \ j = 1, 2, \ldots, n_i$.

α and $\mathrm{Cov}(\alpha)$ are estimated by (Xu et al., 2013; Shi and Nehorai, 2017).

$$\begin{gathered} \hat{\alpha} = \left(X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}\hat{\beta} \\ \hat{\alpha}_2 = (0,1)\,\hat{\alpha} \\ \mathrm{Cov}(\hat{\alpha}) = \left(X'\Sigma^{-1}X\right)^{-1} \\ \mathrm{Cov}(\hat{\alpha})_{22} = (0,1)\left(X'\Sigma^{-1}X\right)^{-1}\begin{pmatrix} 0 \\ 1 \end{pmatrix} \end{gathered} \tag{2}$$

Under the null hypothesis H_0:α_2=0, Wald statistic for testing the SNP-environment interaction effect is $\hat{\alpha}_2/\mathrm{Cov}(\hat{\alpha})_{22}$, which follows a 1 degree of freedom (df) χ^2 distribution. Under the null hypothesis of H_0:α=0, the Wald statistic for testing joint effects of the SNP and the interaction is $\hat{\alpha}'\mathrm{Cov}(\hat{\alpha})^{-1}\hat{\alpha}$, which follows a 2 df χ^2 distribution.

The model (1) can be specified as any nonlinear function of the environmental variable as necessary. For example, to test quadratic SNP-environment interaction, the model can be formulated as

$$\hat{\beta} = X^N\alpha^N + \epsilon^N \tag{3}$$

where

$$X^N = \begin{pmatrix} X_1^N \\ X_2^N \\ \vdots \\ X_n^N \end{pmatrix}, X_i^N = \begin{pmatrix} 1 & E_{i1} & E_{i1}^2 \\ 1 & E_{i2} & E_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & E_{in_i} & E_{in_i}^2 \end{pmatrix}, \alpha^N = \begin{pmatrix} \alpha_1^N \\ \alpha_2^N \\ \alpha_3^N \end{pmatrix}.$$

The Wald statistic then follows a 2 df χ^2 distribution when testing the two interaction effects simultaneously. The Wald statistic follows a 3 df χ^2 distribution for testing the SNP main and interactions jointly (Xu et al., 2013).

Overlapping MR Method

Inspired by the methods for testing the SNP main effect with overlapping data (Lin and Sullivan, 2009), based on regular MR, we propose the OMR model for testing the SNP-environment interaction when data among studies are overlapped.

We consider the kernel process for modeling the correlations due to the overlapping data. Following Lin's recommendation, the covariance matrix under the correlated studies can be modeled as follows (Lin and Sullivan, 2009):

$$\Omega = \Sigma^{1/2}C\Sigma^{1/2}, \tag{4}$$

where C is the correlation matrix. The dimensions of this matrix C are related to the number of studies and the group number of each study. The details of the correlation matrix will be presented in the next section.

Alternatively, the variance covariance matrix can be generalized according to Han's suggestion as follows (Han et al., 2016):

$$\Omega = \mathrm{diag}\left(e'\left(\Sigma^{1/2}C\Sigma^{1/2}\right)^{-1}\right)^{-1} \tag{5}$$

where e is a vector of ones whose length is the sum of the number of groups among all studies. After this modification, the correlation matrix becomes a diagonal matrix. This matrix is highly likely to be positive semi-definite and the analysis of the positive semi-definite matrix is similar to the condition of case-control studies (Han et al., 2016).

Lin's variance covariance matrix is equivalent to Han's (Han et al., 2016). The variance covariance matrix based on Han's formula (5) is more flexible. However, it is more computationally intensive. The method of Lin is simple in its mathematical form and calculation. In cases analyzing with existing programs that require studies to be independent, Han's method can be applied.

Correlation Matrices

Lin and Sullivan (2009) developed a correlation matrix C for incorporating correlations among summary statistics of studies due to the overlapping data. The correlation of studies i and j is given as follows:

$$\gamma_{ij} \approx n_{ij}/\sqrt{n_i n_j}, \quad (6)$$

where n_i and n_j are the numbers of studies i and $j$respectively, and n_{ij} is the number of overlapped individuals between the i-th and j-th studies.

When considering the MR method, this correlation can be modeled as follows:

$$\gamma_{i_h j_k} \approx n_{i_h j_k}/\sqrt{n_{i_h} n_{j_k}}, \quad (7)$$

where n_{i_h} and n_{j_k} are the sample sizes of the h-th group of study i and the k-th group of study j, and $n_{i_h j_k}$ is the number of overlapping samples between them. In this correlation structure, the block matrix that corresponds to each study is an identity matrix; that is, the diagonal block matrices of the correlation matrix are all identity matrices.

Hypothesis Testing

With the introduced correlation matrix, linear unbiased estimates $\hat{\alpha}$ and Cov($\hat{\alpha}$) can be found as follows (Becker and Wu, 2007):

$$\begin{aligned} \hat{\alpha} &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}\hat{\beta} \\ \hat{\alpha}_2 &= (0,1)\hat{\alpha} \\ \mathrm{Cov}(\hat{\alpha}) &= \left(X'\Omega^{-1}X\right)^{-1} \\ \mathrm{Cov}(\hat{\alpha})_{22} &= (0,1)\mathrm{Cov}(\hat{\alpha})\begin{pmatrix}0\\1\end{pmatrix} \end{aligned} \quad (8)$$

Under the null hypothesis α_2=0, the Wald statistic for testing the SNP-environment interaction effect is given as follows:

$$S_I = \alpha_2^2/\mathrm{Cov}(\hat{\alpha})_{22} \quad (9)$$

This statistic follows a 1 df χ^2distribution.

Under null distribution α=0 the Wald statistics for testing the SNP and the interaction joint effects are given as follows:

$$S_J = \hat{\alpha}^2/\mathrm{Cov}(\hat{\alpha}) \quad (10)$$

which follows a 2 df χ^2 distribution.

OMR method can also be extended to test nonlinear SNP-environment interaction for overlapping method. This process is similar with model (1), the Wald statistic for the test of SNP-environment interaction and quadratic SNP-environment interaction follows a 2 df χ^2 distribution. The Wald statistic for testing the SNP, SNP-environment interaction, and quadratic SNP-environment interaction interactions jointly follows a 3 df χ^2 distribution.

As can be seen, our models are generalized versions of the regular MR. When the data of studies are independent, correlation matrix C is an identity matrix, and the two covariance matrices become

$$\Omega = \Sigma^{\frac{1}{2}}C\Sigma^{\frac{1}{2}} = \Sigma \quad (11)$$

and

$$\Omega = \mathrm{diag}\left(e'\left(\Sigma^{1/2}C\Sigma^{1/2}\right)^{-1}\right)^{-1} = \Sigma \quad (12)$$

In this case, the covariance matrix is identical to that of the regular MR.

RESULTS

We evaluated the type I error to ensure that the false positive rate is appropriately controlled by our proposed OMR method when overlapping data exist, that is, whether the empirical type I error rate is close to the specified level. We compared our method with the splitting method and regular MR method, which did not consider overlapping data. The power was then compared at different levels of sample overlap. We considered two scenarios where there were 100 and 400 overlapping subjects between every two studies.

Simulation

The quantitative phenotype Y was simulated as being related to G and E, which were the genotypes of the SNP and environment variables, respectively. The simulation model representing this relationship is given as follows:

$$Y = \beta_G G + \beta_{G\times E} G \times E + \beta_E E + \epsilon$$

Here, the SNP was assumed to have an additive genetic effect; the minor allele frequency was 0.3, and G was the code of SNP, which was the number of minor alleles. We generated random numbers by the runif function in R, then the values of G are determined by which intervals the random numbers fall into, and the intervals are determined by genotype frequency. Variable E was normally distributed, E~N(0,1). 10% of the variation in Y was explained by $\beta_E E$. The fixed effects β_G and $\beta_{G\times E}$ varied in our simulated datasets. The random error ε was normally distributed with zero mean and its variance was chosen such that phenotypic variance is unit. The environment variable and error term were generated by the rnorm function in R. In all our numerical experiments, we considered meta-analyses of data from 2, 3, 4, 5, and 6 studies, each of which had 1,000 unrelated individuals. In each study, we simulated three variables: the phenotype Y, environmental E, and genotype SNP. Across studies, there were 100 or 400 overlapping samples between any two studies. Under each simulation setup, data were generated with 1,000 replicates.

We divided 1,000 unrelated individuals in each study into five groups according to the distribution of E, before meta-analyses. In each group, we applied linear regression to estimate the main effects β_G, its corresponding error ε, and the mean environment variable E. Meta-analysis were performed with 2, 3, 4, 5, and 6 studies.

Type I Error

To obtain the type I error of the interaction test, the effect of the SNP-environment interaction was set to be zero and the SNP main effect explained 0.5% variance of the trait variance. The empirical type I error of our method was calculated by transforming the covariance matrix with overlapping data into a diagonal matrix and then using regular MR. Under this simulation, the test of empirical type I error of our method followed a 1 df distribution. The empirical type I error of the splitting method with two studies was estimated by removing 100 or 400 overlapping subjects of study 1, and the data in study

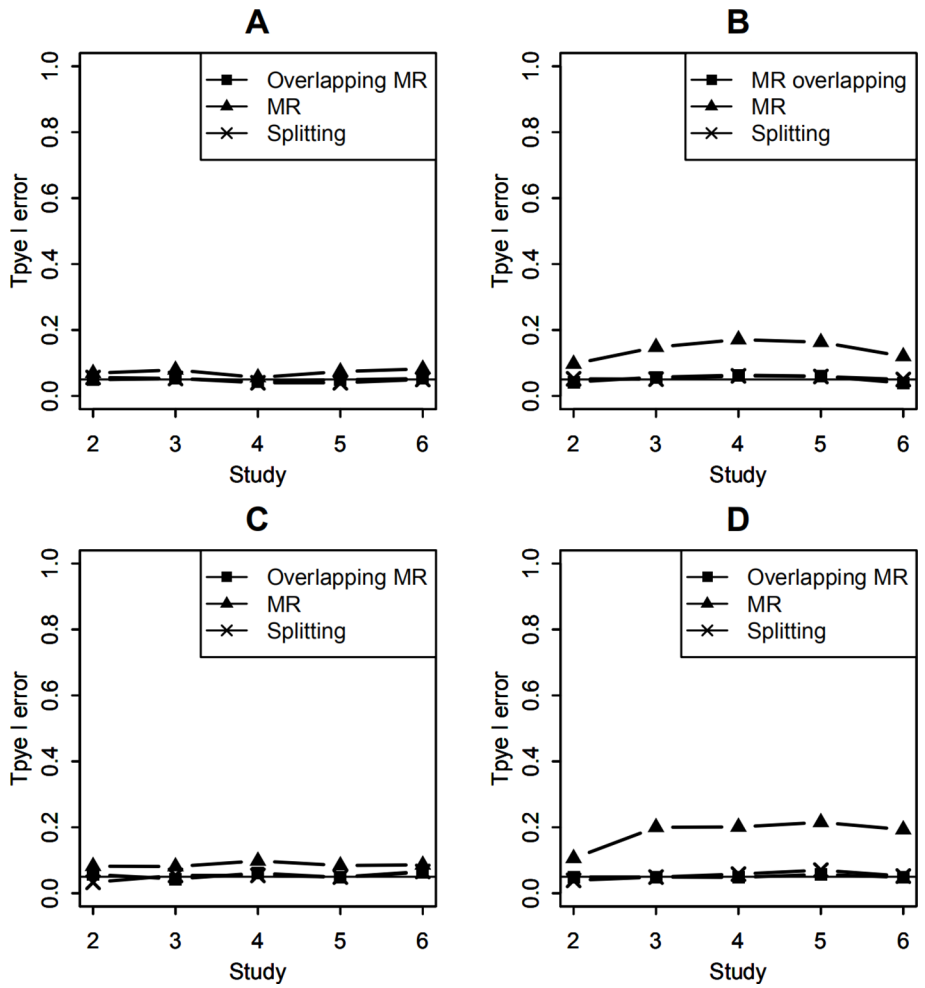


FIGURE 1 | Type I error of testing SNP-environment interaction and jointly testing SNP main effect and the interaction. **(A, B)** are type I errors of the interaction test with 100 and 400 overlapping data, respectively. **(C, D)** are type I errors of the joint test with 100 and 400 overlapping data, respectively. Solid line with crosses is type I errors of the splitting method with 2, 3, 4, 5, and 6 studies. Solid line with filled squares is type I errors of OMR method with 2, 3, 4, 5, and 6 studies. Solid line with filled triangles is type I errors of the regular MR with 2, 3, 4, 5, and 6 studies when overlapping data is ignored.

2 were left unchanged. The empirical type I error of the splitting data method with 3, 4, 5, and 6 studies was estimated by discarding 100 or 400 overlapped subjects from each study. **Figures 1A, B** show the type I error rates of 2, 3, 4, 5, and 6 studies in the test of SNP-environment interaction with 100 and 400 overlapping subjects, respectively. We can see that both our method and the splitting data method yielded type I error results close to the specified 0.05 level. The regular MR method, which did not consider overlapping data, yielded inflated type I error rates. The greater the overlap, the more the inflation was.

To calculate the type I error rates of the joint test of the SNP main effect and the interaction, we set both the SNP and the SNP-environment interaction effects to be zeros. The Wald test statistics followed a 2 df χ^2distribution. **Figures 1C, D** show the type I errors of the joint test under the null hypotheses. We can also see that the results of the two methods were around 0.05 as well; thus, both our OMR method and splitting method treated the overlapping data appropriately. The regular MR method in the joint test yielded a higher type I error than in the interaction test because it included more information on overlapping data.

In real meta-analysis, sample sizes of studies vary and percentages of overlapping may be different for studies. Here, we set the sample sizes of the 6 studies as (1,000, 1,200, 1,400, 1,600, 1,800, 2,000). Let the effect of the SNP-environment interaction to be zero and the SNP main effect explained 0.5% of trait variance. Type I errors of testing the SNP-environment interaction are shown in **Figures 2A, B**, which represent results of testing the interaction with 100 and 400 overlapping individuals in each study, respectively. Setting both the SNP and the SNP-environment interaction effects to be zeros, we conducted joint tests for SNP and SNP-environment interaction. **Figures 2C, D** show type I errors of the joint test with 100 and 400 overlapping individuals, respectively. As the results in **Figure 1**, OMR and the splitting method control type I errors as expected, while inflated type I errors can be observed for the regular MR.

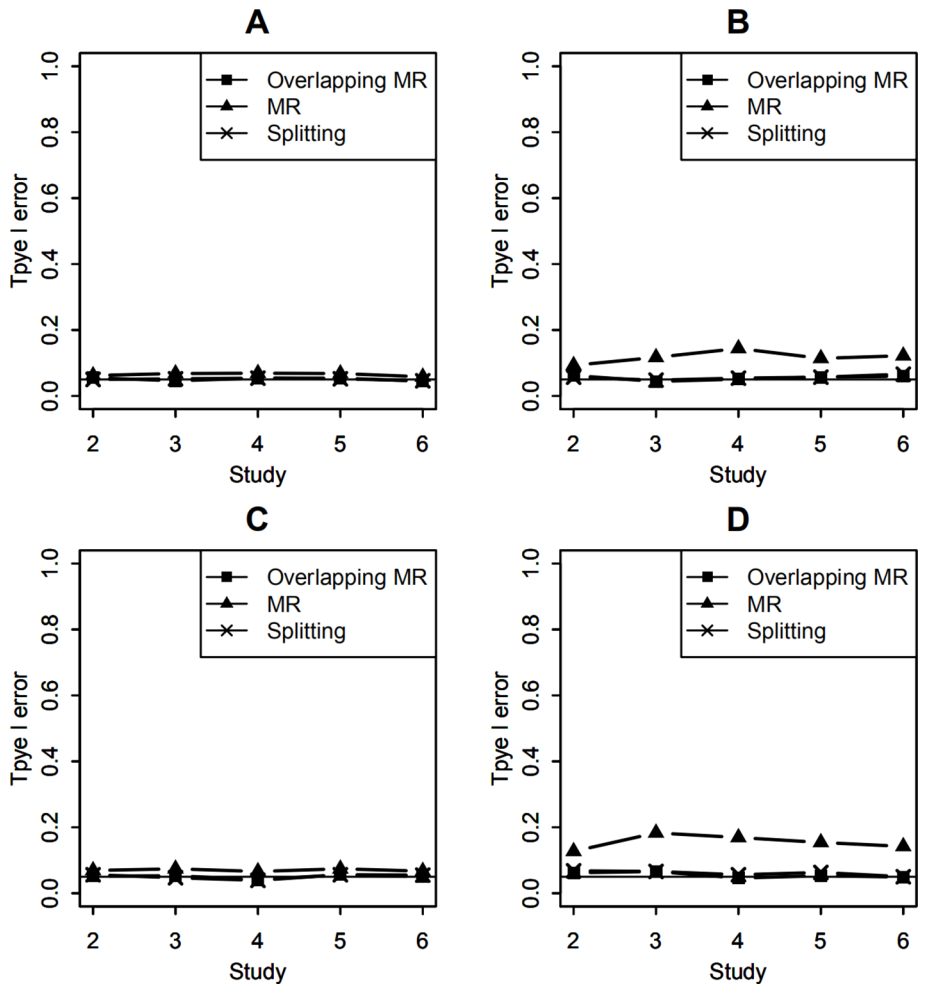


FIGURE 2 | Type I error of testing SNP-environment interaction and jointly testing SNP main effect and the interaction with 6 studies of 1,000, 1,200, 1,400, 1,600, 1,800, 2,000 individuals, respectively. **(A, B)** are type I errors of the interaction test with100 and 400 overlapping data, respectively. **(C, D)** are type I errors of the joint test with 100 and 400 overlapping data, respectively. Solid line with crosses is type I errors of the splitting method with 2, 3, 4, 5, and 6 studies. Solid line with filled squares is type I errors of OMR method with 2, 3, 4, 5, and 6 studies. Solid line with filled triangles is type I errors of the regular MR with 2, 3, 4, 5, and 6 studies when overlapping data is ignored.

Power

To compare the statistical power of testing the SNP-environment interaction, both SNP-environment and SNP effects explained 0.5% variance of the trait variance. In this simulation, statistical significance was determined by the P values of the tests, which were smaller than 0.05. The empirical power was obtained by calculating the proportion of the significant results in 1,000 replicates. The P values were calculated using the Wald test (9), which followed a 1 df χ^2 distribution. **Figures 3A, B** show the power of the SNP-environment interaction with overlapping data of 100 and 400, respectively. We can see that our method yields similar results to those of the splitting method. Note that our method does not require the study-level genotype or phenotype data, which is its major advantage.

In the joint test of the SNP main effect and the SNP-environment interaction effect, both SNP-environment and SNP effects explained 0.5% variance of the trait variance. In this simulation, the P values were again calculated using the Wald test (10) following a 2 df χ^2 distribution. **Figures 3C, D** show the powers of the joint test with 100 and 400 overlapping samples, respectively. We compared our method with the splitting method. These results are similar to those from the SNP-environment interaction test; however, the joint test yielded higher power than the interaction test. This is because the joint test included more effects than the SNP-environment interaction test (Kraft et al., 2007).

For studies with unequal sample sizes (1,000, 1,200, 1,400, 1,600, 1,800, 2,000), power of testing the SNP-environment interaction and power of the joint test for the SNP and the interaction are presented in **Figure 4**. Effects of the SNP and the interaction are the same as those in previous example. We can see that powers in **Figure 4** demonstrate similar patterns as those in **Figure 3**, whereas the former are in general larger than the latter. This is because that total sample size employed in **Figure 4** is larger than that in **Figure 3**.

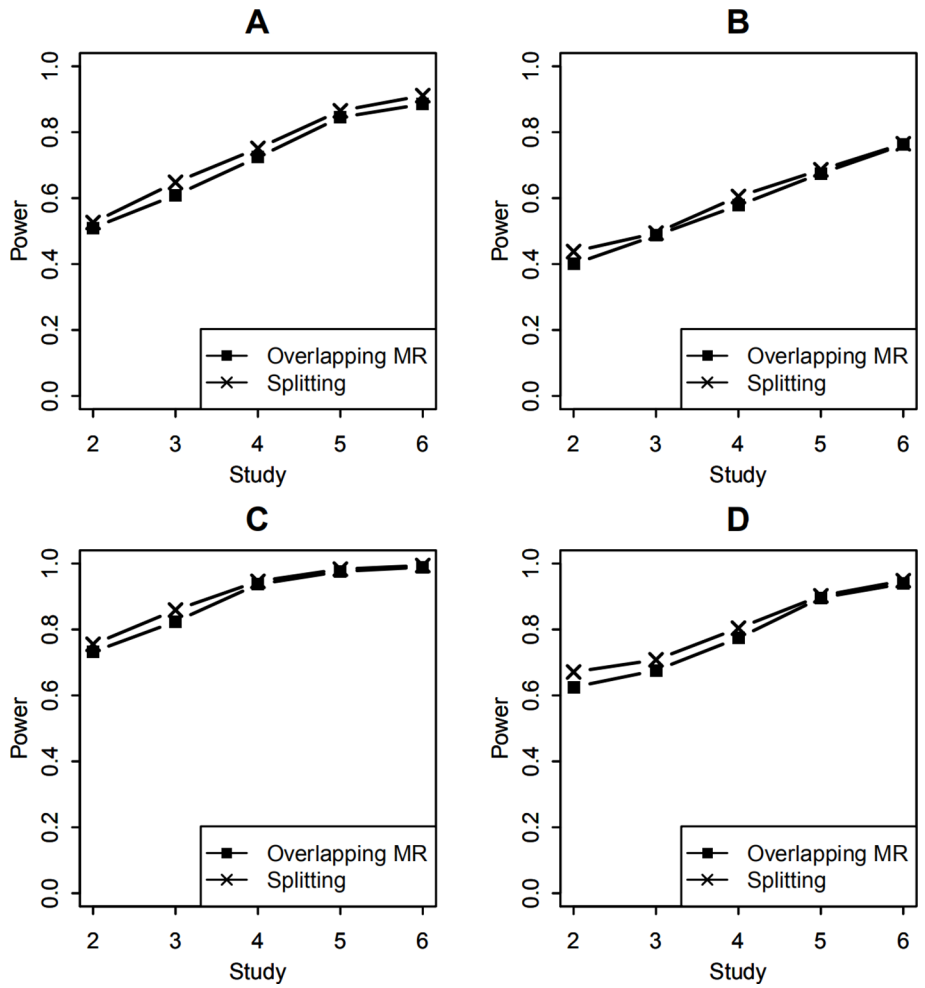


FIGURE 3 | Statistical power of testing SNP-environment interaction and jointly testing SNP main effect and the interaction. **(A, B)** are statistical powers of the interaction test with 100 and 400 overlapping data, respectively. **(C, D)** are statistical powers of the joint test with 100 and 400 overlapping data, respectively. Solid line with crosses is powers of the splitting method with 2, 3, 4, 5, and 6 studies. Solid line with filled squares is powers of the OMR method with 2, 3, 4, 5, and 6 studies.

In GWAS, it is a common phenomenon that effects of the SNP and SNP-environment interaction may have different directions. Here, we consider the scenario that both the SNP and the interaction explained 0.5% variance of the trait variance but the directions of their effects are opposite. As in the previous example, we tested the SNP-environment interaction as well as joint effects of the SNP and the interaction. **Figures 5A, B** show powers of the interaction test with 100 and 400 overlapping samples. **Figures 5C, D** present powers of joint test with 100 and 400 overlapping samples. Compared with the results in **Figure 3**, whose effects of the SNP and interaction have the same direction, we can see that the powers of the two tests are about the same in the two scenarios.

Finally, we added simulation for nonlinear SNP-environment interaction when testing the effect of SNP-environment interaction and the joint effects of SNP and SNP-environment. Both the effect of SNP and the effect of SNP-environment interaction explained 0.5% variance of the trait variance, the effect of nonlinear SNP-environment interaction explained 0.05% variance of the trait variance. We compared the model considering nonlinear SNP-environment as in (Xu et al., 2013). with the model not considering nonlinear SNP-environment. **Figures 6A, B** show the results of this comparison with 100 and 400 overlapping individuals for the test of interaction respectively, in each of the two figures, we can see that the two lines we compared present similar results. From **Figures 6C, D** we can see that the powers under the model considering nonlinear SNP-environment are lower than that not considering with 100 and 400 overlapping individuals for the joint test respectively. That is because the column variables in X are not an orthonormal basis when considering nonlinear interaction. The nonlinear interaction enters the model as part of the SNP main effect (Xu et al., 2013).

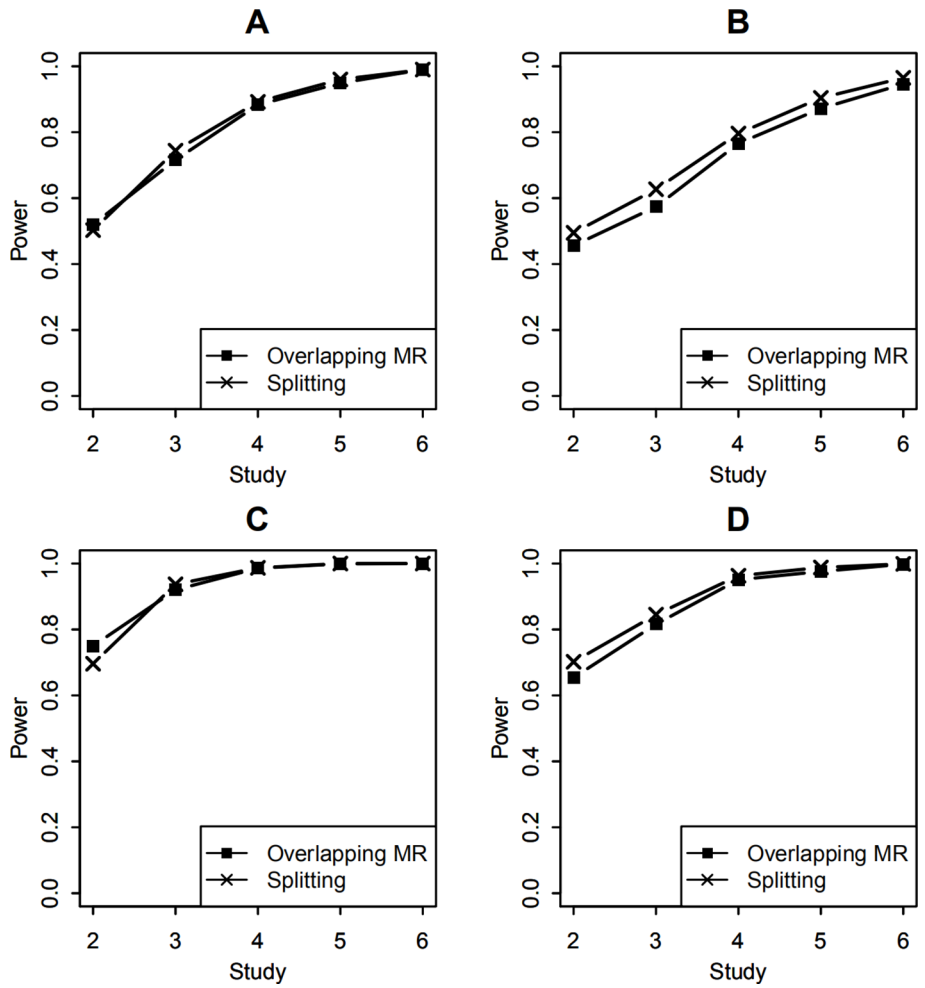


FIGURE 4 | Statistical power of testing SNP-environment interaction and jointly testing SNP main effect and the interaction with 6 studies of 1,000, 1,200, 1,400, 1,600, 1,800, 2,000 individuals, respectively. **(A, B)** are statistical powers of the interaction test with 100 and 400 overlapping data, respectively. **(C, D)** are statistical powers of the joint test with 100 and 400 overlapping data, respectively. Solid line with crosses is powers of the splitting method with 2, 3, 4, 5, and 6 studies. Solid line with filled squares is powers of the OMR method with 2, 3, 4, 5, and 6 studies.

DISCUSSION

SNP may indeed interact with *E* nonlinearly in real biological process. In this case, regressing the main effect of SNP on *E* linearly involved model mis-specification. On the other hand, such linear regression can hopefully capture a portion of the main effect. In this case, we can employ Hermite polynomials to the nonlinear interaction model to avoid this phenomenon (Xu et al., 2013).

The sample sizes of studies vary in real meta-analysis. As explained in the reference (Manning et al., 2011), there are 561 individuals in the FamHS Study, 1,661 in the HealthABC Study, 2,854 in the CHS Study, 8,367 in the ARIC Study, 6,023 in the FHS Study, which gives a total sample size of 19,946. For methodological evaluations, the authors of (Manning et al., 2011) chose to simulate five studies each of 1,000 individuals. In our work, we also adopted a relatively moderate sample size 1,000 to verify the effectiveness of our method. In the revised manuscript, we conducted additional simulations to have studies with different sample sizes to evaluate the sensitivity to the unbalanced sample sizes among studies.

When testing the SNP main effect, the splitting method for case-control studies was reported to yield a lower power than Lin's method, which is because the studies share common controls (Lin and Sullivan, 2009). Splitting these studies such that every subject contributes only once leads to a dramatic decrease in the effective sample size. Our simulation examples based on cohort studies yielded slightly less power than the splitting method because the overlapping structure in our examples differed from that in the case-control studies. The splitting method in the cohort studies drops less data than in case-control studies, so the power loss due to splitting the data is smaller.

Our method is based on the MR in which one divides the studies into several groups according to the environmental variable. Thus, when calculating the correlation matrix, we

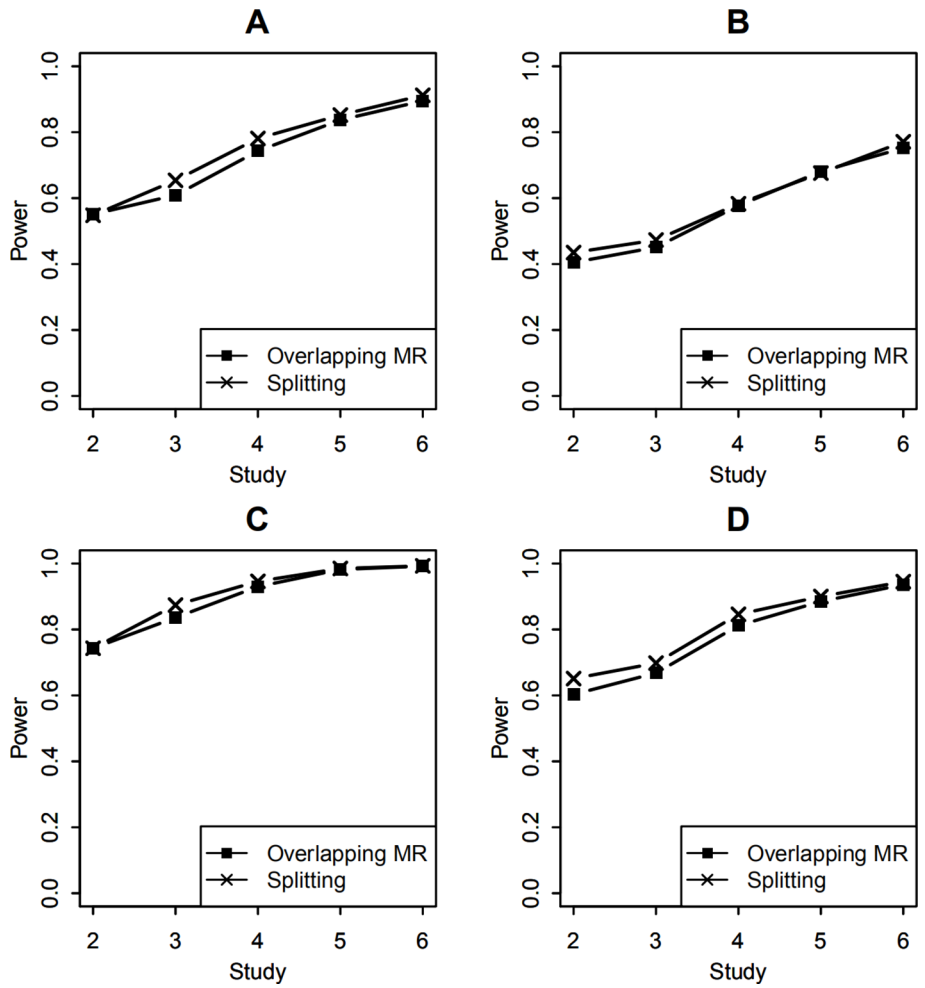


FIGURE 5 | Statistical power of testing SNP-environment interaction and jointly testing SNP main effect and the interaction with opposite directions for the effects of the SNP and the SNP-environment interaction. **(A, B)** are statistical powers of the interaction test with 100 and 400 overlapping data, respectively. **(C, D)** are statistical powers of the joint test with 100 and 400 overlapping data, respectively. Solid line with crosses is powers of the splitting method with 2, 3, 4, 5, and 6 studies. Solid line with filled squares is powers of the OMR method with 2, 3, 4, 5, and 6 studies.

must consider both the number of overlapping data among studies and the number of overlapping data among groups. When the overlaps among groups are unavailable and the data overlap is independent of the environment variable, the overlaps between two groups can be estimated by the overlaps between their studies and the sample proportions of the groups in the two studies. In either case, our method does not require individual-level data as the splitting method does.

To the best of our knowledge, there is still no meta-analysis method for testing SNP-environment interaction with overlapping data among studies. Our OMR method was generalized from regular MR. When evaluating our proposed OMR method, we compared our method with the splitting method and regular MR. **Figure 1** indicates that regular MR yielded inflated type I error rates; the more the amount of overlapping data, the higher the amount of inflation. On the other hand, our OMR method controlled the type I error rates appropriately. Therefore, regular MR is unsuitable for studies that have overlapping data.

CONCLUSION

In this paper, we generalized the regular MR model to OMR by incorporating correlations among studies due to the overlapping data. We proposed a test for the SNP-environment interaction as well as a joint test for the SNP and the interaction under the OMR framework. The two test were compared with the splitting method in terms of their type I error rate and statistical power. Through simulation, we demonstrated that our method yielded comparative power with respect to the splitting method and the type I error rate of the regular MR is inflated when overlapping data are ignored. We also evaluated our OMR method with unequal sample sizes among studies, opposite directions of the SNP effect and the interaction effect, and assessed the robustness of our method when nonlinear interaction effect exists. Our method does not require individual-level genotype and phenotype data, which overcomes the major limitation of the splitting method. In GWAS practice, our OMR method can be

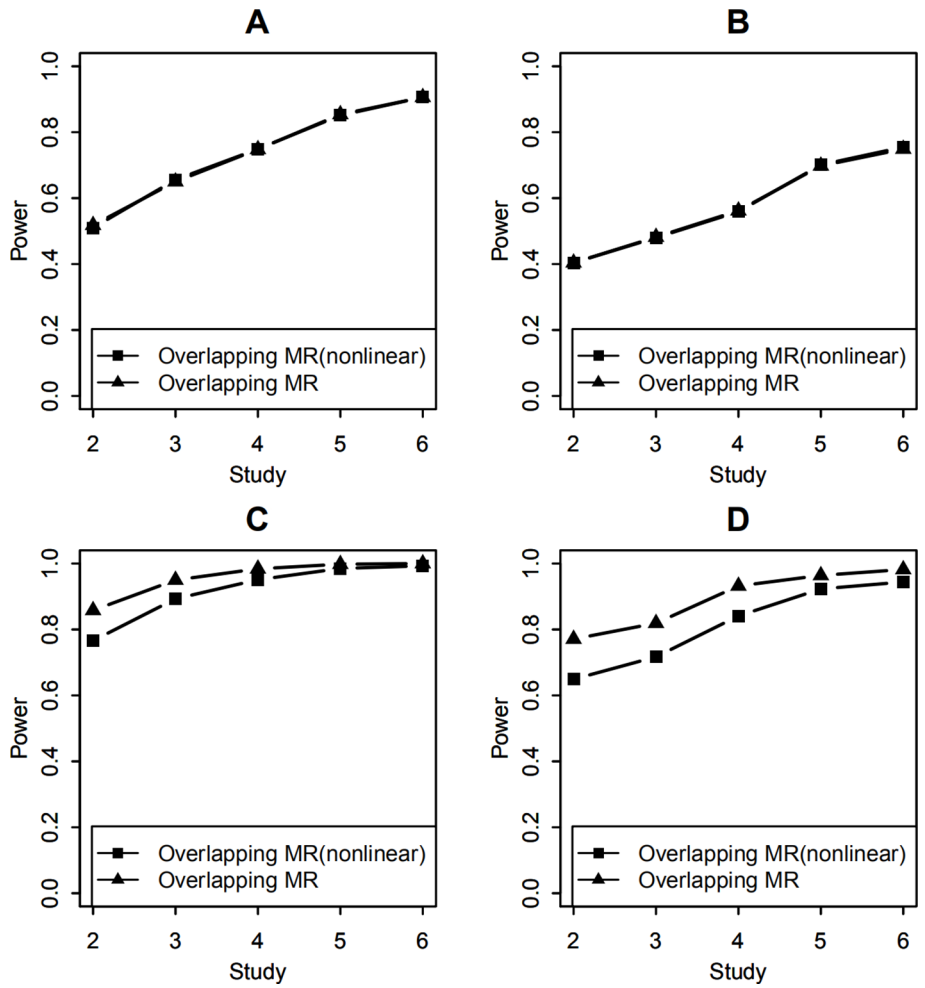


FIGURE 6 | Statistical power of testing SNP-environment interaction and jointly testing SNP main effect and the interaction with nonlinear SNP-environment interaction effect. **(A, B)** are statistical powers of the interaction test with 100 and 400 overlapping data, respectively. **(C, D)** are statistical powers of the joint test with 100 and 400 overlapping data, respectively. Solid line with filled squares shows powers of OMR when nonlinear interaction effect was considered in the model. Solid line with filled triangles is powers of the OMR when nonlinear interaction effect was not considered in the model.

used to control false positive results when the studies with overlapping individuals are included in the meta-analysis, thus improve the probability of finding genuine associations.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

QJ: conceived the concept, designed and conducted the simulation studies, and drafted the manuscript. GS: conceived the concept, supervised the work, and reviewed and revised the manuscript.

FUNDING

This work was supported by the National Thousand Youth Talents Plan.

ACKNOWLEDGMENTS

This work has been accepted by the Fourth China Computer Federation Bioinformatics conference (CBC2019) which was held in Guangzhou, 23–25 August 2019. The authors want to thank for the helpful comments from anonymous reviewers of CBC2019.

REFERENCES

Becker, B. J., and Wu, M. J. (2007). The synthesis of regression slopes in meta-analysis. *StatSci* 22 (3), 414–429. doi: 10.1214/07-STS243

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* 90 (5), 821–835. doi: 10.1016/j.ajhg.2012.03.015

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis* (Chichester: John Wiley & Sons.Ltd).

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene-drug regulatory modules identification. *Bioinformatics* 35 (4), 602–610. doi: 10.1093/bioinformatics/bty662

Chen, J., Han, G., Xu, A., and Cai, H. (2019). Identification of multidimensional regulatory modules through multi-graph matching with network constraints. *IEEE Trans. Biomed. Eng.* doi: 10.1109/TBME.2019.2927157

Eleftheria, Z., and John, P. A. I. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10 (2), 191–201. doi: 10.2217/14622416.10.2.191

Evangelou, E., and Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. 14 (6), 379–89. *Nat. Rev. Genet.* doi: 10.1038/nrg3472

Fisher, R. A. (1967). *Statistical methods for researchworkers* (Edinburgh: Oliver&Boyd).

Fleiss, J. (1993). The statistical basis of meta-analysis. *Stat. Methods Med. Res.* 2 (2), 121–145. doi: 10.1177/096228029300200202

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* 88 (5), 586–598. doi: 10.1016/j.ajhg

Han, B., Duong, D., Sul, J. H., de Bakker, P. I., Eskin, E., and Raychaudhuri, S. (2016). A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.* 25 (9), 1857–1866. doi: 10.1093/hmg/ddw049

Jin, Q., and Shi, G. (2019). Meta-analysis of SNP-environment interaction with heterogeneity. *Hum. Hered.* doi: 10.1159/000504170

Kraft, P., and Haiman, C. A. (2010). GWAS identifies a common breast cancer risk allele among BRCA1 carriers. *Nat. Genet.* 42 (10), 819–820. doi: 10.1038/ng1010-819

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., and Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* 63 (2), 111–119. doi: 10.1159/000099183

Lin, D. Y., and Sullivan, P. F. (2009). Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* 85 (6), 862–872. doi: 10.1016/j.ajhg.2009.11.001

Manning, A. K., LaValley, M., Liu, C. T., Rice, K., An, P., Liu, Y., et al. (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP×environment regression coefficients. *Genet. Epidemiol.* 35 (1), 11–18. doi: 10.1002/gepi.20546

Manolio, T. A. (2010). Genome-wide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363 (2), 166–176. doi: 10.1056/NEJMra0905980

Mehramiz, M., Ghasemi, F., Esmaily, H., Tayefi, M., Hassanian, S. M., Sadeghzade, M., et al. (2018). Interaction between a variant of CDKN2A/B-gene with lifestyle factors in determining dyslipidemia and estimated cardiovascular risk: a step toward personalized nutrition. *ClinNutr.* 37 (1), 254–261. doi: 10.1016/j.clnu.2016.12.018

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* 35 (8), 809–822. doi: 10.1002/gepi.20630

Pereira, T. V., Patsopoulos, N. A., Salanti, G., and Ioannidis, J. P. (2009). Discovery properties of genome-wide association signals from cumulatively combined data sets. *AM. J. Epidemiol.* 170 (10), 1197–1206. doi: 10.1093/aje/kwp262

Pfeiffer, R. M., Mitchell, H. G., and Pee, D. (2009). On combining data from genome-wide association studies to discover disease-associated SNPs. *Statist. Sci.* 24 (4), 547–560. doi: 10.1214/09-STS286

Shi, G., and Nehorai, A. (2017). Robustness of meta-analyses in finding gene × environment interactions. *PLoS One* 12 (3), e0171446. doi: 10.1371/journal.pone.0171446

Wen, X., and Stephens, M. (2014). Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *Ann. Appl. Stat.* 8 (1), 176–203. doi: 10.1214/13-AOAS695

Wen, X. (2014). Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* 70 (1), 73–83. doi: 10.1111/biom.12112

Xu, X., Shi, G., and Nehorai, A. (2013). Meta-regression of gene-environment interaction in genome-wide association studies. *IEEE Trans. Nanobiosci.* 12 (4), 354–362. doi: 10.1109/TNB.2013.2294331

Zaykin, D. V., and Kozbur, D. O. (2010). P-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol.* 34 (7), 725–738. doi: 10.1002/gepi.20536

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 31 January 2020
doi: 10.3389/fgene.2019.01396

MAC: Merging Assemblies by Using Adjacency Algebraic Model and Classification

Li Tang[1], *Min Li*[1]*, *Fang-Xiang Wu*[1,2], *Yi Pan*[1,3] *and Jianxin Wang*[1]

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada, [3] Department of Computer Science, Georgia State University, Atlanta, GA, United States

With the generation of a large amount of sequencing data, different assemblers have emerged to perform de novo genome assembly. As a single strategy is hard to fit various biases of datasets, none of these tools outperforms the others on all species. The process of assembly reconciliation is to merge multiple assemblies and generate a high-quality consensus assembly. Several assembly reconciliation tools have been proposed. However, the existing reconciliation tools cannot produce a merged assembly which has better contiguity and contains less errors simultaneously, and the results of these tools usually depend on the ranking of input assemblies. In this study, we propose a novel assembly reconciliation tool MAC, which merges assemblies by using the adjacency algebraic model and classification. In order to solve the problem of uneven sequencing depth and sequencing errors, MAC identifies consensus blocks between contig sets to construct an adjacency graph. To solve the problem of repetitive region, MAC employs classification to optimize the adjacency algebraic model. What's more, MAC designs an overall scoring function to solve the problem of unknown ranking of input assembly sets. The experimental results from four species of GAGE-B demonstrate that MAC outperforms other assembly reconciliation tools.

Keywords: adjacency algebraic model, contig classification, contig reconciliation, de novo assembly, next-generation sequencing

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Xingpeng Jiang,
Central China Normal University, China
Cuncong Zhong,
University of Kansas, United States

***Correspondence:**
Min Li
limin@mail.csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *23 October 2019*
Accepted: *20 December 2019*
Published: *31 January 2020*

Citation:
Tang L, Li M, Wu F-X, Pan Y and Wang J (2020) MAC: Merging Assemblies by Using Adjacency Algebraic Model and Classification. Front. Genet. 10:1396. doi: 10.3389/fgene.2019.01396

INTRODUCTION

Next-generation sequencing technologies (NGS) offer a large volume of short sequences with relatively short insert size compared to the traditional Sanger sequencing technology and the third generation sequencing technologies, e.g., Pacific Biosciences (Eid et al., 2009) and Oxford Nanopore (Clarke et al., 2009). Although considerable third generation sequencing data has been produced, due to the higher cost per base and higher sequencing errors, NGS sequencing data still plays an important role in tackling an increasing list of biological problems. The de novo genome assembly is a fundamental process for computational biology (Schatz et al., 2010), which drives the generation of many assemblers to complete the construction of genome sequences, such as Velvet (Zerbino and Birney, 2008), ABySS (Simpson et al., 2009), ALLPATHS- LG (Gnerre and Jaffe, 2011), SOAPdenovo (Li et al., 2010), EPGA2(Luo et al., 2015), Miniasm (Li, 2015), BOSS (Luo et al., 2017), SCOP (Li et al., 2018a), ARC (Liao et al., 2018), iLSLS

(Li et al., 2018b), MEC (Wu et al., 2017), EPGA-SC (Liao et al., 2019a), PE-Trimmer (Liao et al., 2019b), and so on.

However, there is no single assembler that could perform optimally in every quality metric, which has been demonstrated repeatedly (Earl et al., 2011; Salzberg et al., 2012; Bradnam et al., 2013). The situation is caused by various factors: (1) Assembly algorithms are mainly based on overlap-layout-consensus graphs or de Bruijn graphs, these two types of algorithms use different strategies to deal with errors, inconsistencies, and ambiguities; (2) NGS genome assemblies suffer from long repeats and duplications, which is the primary reason why some assemblers outperform others in specific regions and specific species (Alkan et al., 2010); (3) the uneven sequencing coverage of NGS data increases the complexity of assembly, which makes the parameters having great influence on the assembly results, such as k-mer size; (4) the sequencing errors and chimeric reads cause direct assembly mistakes. Besides, different sequencing platforms usually introduce different bias (Harismendy et al., 2009), so the assemblies generated by various platforms may present different features, and there is usually complementarity between them (Diguistini et al., 2009). Thus, it is appealing merging different assemblies to generate a high-quality assembly by using complementary, which is first proposed by Zimin et al., called assembly reconciliation. The main goal of assembly reconciliation is to increase the contiguity of assembly results while reducing (or at least not increasing) the errors in assembly.

Many assembly reconciliation algorithms have been proposed, for some earlier ones, such as Reconciliator (Zimin et al., 2005) and GAM (Casagrande et al., 2009). Reconciliator detects apparent errors in the assembly, and then the error regions are modified by using the alternative draft assembly, through which the gaps between sequences are reduced. GAM defined supercontig to facilitate the integration, which takes two assemblies as input, and regards the former one as reference. For some reference-based algorithms, such as eRGA (Francesco et al., 2011), RAGOUT (Kolmogorov et al., 2014), and MAIA (Nijkamp et al., 2010), if there is no corresponding reference or relative reference genome, they cannot work properly, so we don't discuss these methods here. The algorithm CISA is used to integrate the assemblies of bacterial genome in the four major phases (Lin and Liao, 2013). Firstly, CISA extracts the largest contig as a representative contig, and aligns the remaining contigs to the representative coting, then conducts extension with the contig whose alignment rate is more than 80%. This step is repeated iteratively until there is no representative contig found. Secondly, CISA identifies two types of misassemble contigs: for the misjoined error, CISA removes the contig; for the insertion error, CISA splits the contig. Thirdly, CISA merges contigs which have at least 30% overlap, and also estimates the size of repeats. Finally, if the overlap between two contigs is greater than the maximum size of repeats, CISA merges the contigs. CISA could be used to merge more than two assemblies.

The objective of GAA is to generate an accordance assembly from two or more large genome assemblies (Yao et al., 2012). GAA takes a target assembly and a query assembly as input, then uses BLAT aligner (Kent, 2002) to align the query assembly to target assembly. The high scoring matches are used to construct the accordance graph, GAA finds the maximal sub-paths from the graph, and the gaps can be divided into two types, between contigs and inside contigs. For the gaps between contigs, GAA compares the observed value and expected value of gap size, then decides whether to merge two contigs. For the gaps inside contigs, a compression-expansion(CE) statistic module (Zimin et al., 2005) is used to evaluate the gap regions. The 454 and Illumina de novo assemblies are used to examine the performance of GAA.

GAM-NGS (Vicedomini et al., 2013) is the updated version of GAM, GAM-NGS can be used on all NGS-based assemblies, especially for eukaryote genomes. Two assemblies and a SAM alignment file are taken as input, GAM-NGS first searches the mapping file to identify highly similar fragments between two assemblies, which is called "blocks", then a graph is used to record and weight the information of blocks, and the conflicts are resolved in the graph. A semi-global alignment between contigs is computed by GAM-NGS, and two contigs are merged if the identity between them is larger than 95%. The CE statistic module (Zimin et al., 2005) is used to choose which assembly can be merged.

The main purpose of MIX (Soueidan et al., 2013) is to reduce both the fragmentation of contig sets and reduce the time consumption of genome finishing. MIX builds an extension graph where vertices represent the terminals of contigs, and the edges represent the alignment situation between contigs. MIX attempts to solve the maximal independent longest path set, which is NP-hard. The performance of algorithm is evaluated on the GAGE-B (Tanja et al., 2013) bacterial dataset.

Metassembly (Wences and Schatz, 2015) merges all the input assemblies into a final one, which is better than or as good as the original assemblies. Metassembly regards one of the inputs as a "primary" assembly, then the others are "secondary" assemblies, the secondary assemblies are used to add useful information to the primary assembly. A pairwise algorithm is used to merge multiple assemblies, the primary assembly is aligned to the secondary assembly, and the best aligned position can be evaluated by LIS (longest increasing subsequence) function. The CE statistic (Zimin et al., 2005) is used to assess the conflicts and select the locally best sequence.

In general, most of the methods described above are based on the CE statistic (Zimin et al., 2005), which is used to detect compression or expansion misassemblies between two input assemblies. However, the CE statistic is obtained by aligning paired-end or mate-pair reads to the assembly, which is impacted by the alignment quality and the false positive within error detection leads to the misassembly directly. In addition, most of the current reconciliation tools are designed for merging short sequences (<100bp), like CISA and GAM-NGS, which performed poorly in merging longer sequences (>200bp). Therefore, there is an urgent require for the robust reconciliation tool to increase the length and quality of assembly, as well as adapt to longer sequencing data.

In this study, we propose a novel assembly reconciliation tool, named MAC, which uses alignment information and GC-content of paired-end reads to classify all the contigs into two types. Then, consensus blocks between contig sets are identified, and the unreliable fragments caused by uneven sequencing depth or sequencing errors could be filtered out. In addition, MAC utilizes the adjacency algebraic model to facilitate the merging process, in which the adjacent graph is used to fulfill accurate fusions between consensus blocks. The classification result of contigs is used to optimize the model, and the repetitive regions could be eliminated by splitting contigs and reconstructing the adjacent graph. What's more, an overall scoring function is proposed to solve the problem of unknown ranking of input assemblies, the scoring function evaluates the overall quality of assembly sets by alignment quality and coverage information. The experimental results from the datasets of GAGE-B demonstrate that MAC performs better than other reconciliation tools.

METHOD

MAC employs the adjacency algebraic model (Sankoff et al., 2000) and the classification to merge assemblies. The identification of consensus blocks is to filter out the unreliable fragments caused by uneven sequencing depth and sequencing errors; the addition of classification is to optimize the adjacency algebraic model and eliminate the influence of repetitive regions. The outline for the whole algorithm is as follows: (1) Preprocessing: MAC aligns paired-end/mate-pair reads to contig sets, and filters out the low-quality alignment; (2) Ranking input assemblies: MAC designs an overall scoring function to rank the input assemblies; (3) Classifying contigs: MAC utilizes the alignment results and GC-content of paired reads to classify contigs; (4) Adopting the adjacent algebraic model: MAC constructs an adjacent graph to fulfill some accurate fusions of consensus blocks, then uses classification results of contigs to optimize the remaining processing steps. The flowchart of MAC algorithm is shown in **Figure 1**.

Preprocessing

MAC takes multiple contig sets and paired-end/mate-pair reads as input, the aligner Bowtie2 needs to be installed in advance. The input reads are aligned to each contig set, respectively. For reads aligning to multiple positions, MAC only maintains the highest score alignment for each read, and removes the redundant alignments. According to the paper of Luo et al., the length of insert size follows a normal distribution N (μ_{is}, σ_{is}), so the distance between two paired reads, which align to the same contig, should be in the range of [μ_{is}– 3*σ_{is}, μ_{is}+ 3*σ_{is}]. For the reads which violate this rule, MAC removes the corresponding alignment. To reduce the impact of sequencing errors, MAC extracts the sequencing quality of every base in reads, and calculates the average and standard deviation of sequencing quality for the remaining alignment, denoted by M_q and σ_{ra}, respectively. Let Q_i represent the mean value of sequencing quality for the i-th alignment. If $Q_i < M_q - 3*\sigma_{ra}$, MAC removes the alignment information.

Ranking Input Assemblies

As most existing assembly reconciliation tools depend on the ranking of input, and the results usually change when the order of input assemblies change. To achieve better results, users have to evaluate the contiguity and correctness of every input assembly by taking the reference into Quast (Gurevich et al., 2013) or other evaluation tools. In the study, MAC utilizes the mapping quality and read coverage to rank the input assemblies. The compact idiosyncratic gapped alignment report (CIGAR) can be obtained from files in the SAM format, in which "M" represents match/mismatch, "I" represents insertion, "D" represents deletion, and the number before a character represents its corresponding quantity. Assume that the length of contig C is L, j $(1 \le j \le L)$ is the position at C, q_j represents the CIGAR of position j, which is calculated as follows.

$$\mathbf{q_j} = \begin{cases} \mathbf{1}, \ \mathbf{if} \ \ \mathbf{j = M} \\ \mathbf{-1}, \ \mathbf{if} \ \ \mathbf{j = M^{*} or \ I \ or \ D}, \end{cases} \tag{1}$$

where M denotes match and M* denotes mismatch. In fact, we cannot distinguish match and mismatch from a single character "M", so MAC calculates the average mapping score of the SAM file. If the mapping score of the corresponding read is larger than or equal to the average mapping score, the character "M" is thought to be match, otherwise, "M" is thought to be mismatch.

To take the coverage into consideration, MAC extracts the alignment of contig C, to calculate the average rc, and standard deviation σ_{rc} of read coverage. Assume that rc_i is the read coverage of the spanning region of read i, RC is used to indicate whether the coverage of the region deviates too much, which is computed as follows.

$$\mathbf{RC} = \begin{cases} \mathbf{1}, \ \mathbf{if} \ \ (\mathbf{rc_i} > \mathbf{rc} + \mathbf{2} * \boldsymbol{\sigma}_{\mathbf{rc}}) \ \ \mathbf{or} \ \ (\mathbf{rc_i} < \mathbf{rc} - \mathbf{2} * \boldsymbol{\sigma}_{\mathbf{rc}}) \\ \mathbf{-1}, \ \mathbf{otherwise} \end{cases} \tag{2}$$

In order to comprehensively consider the mapping quality and read coverage, MAC employs an overall scoring function to rank the input assemblies, which is calculated as follows.

$$\mathbf{score} = \frac{\sum_{c=1}^{N} \sum_{j=1}^{L} \mathbf{q_j} - \sum_{c=1}^{N} \sum_{i=1}^{K} \mathbf{RC}}{\mathbf{N}} \tag{3}$$

Classifying Contigs

In this step, MAC evaluates the quality of contigs by using the alignment result and GC-content, and then classifies all the contigs into two types. Due to the problems of sequencing errors, uneven depth, existence of repetitive regions and the bias of algorithm strategy, contigs often contain misassemblies, which influence the subsequent assemblies directly. Therefore, MAC estimates the correctness of

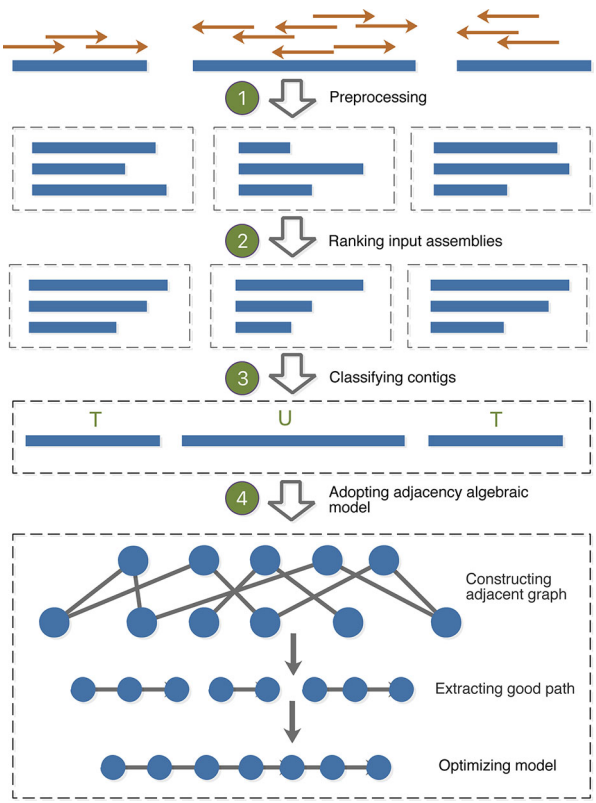


FIGURE 1 | Flowchart of MAC algorithm.

contigs, and marks the type for every contig, and records the potential error positions.

For a contig C, whose length is L, the coordinate of position j is in the range of [1, L]. The fragment coverage $fc(j)$ could be defined as the number of reads with the high alignment scores which span the position j. Because low coverage regions more likely contain error joints, MAC employs a cutoff fc^* to identify the potential error positions, fc^* can be calculated by the average of fragment coverage for all the positions of contig C as follows (Wu et al., 2017).

$$\mathbf{fc^*} = \boldsymbol{\alpha}\star \frac{\sum_{j=1}^{L} \mathbf{fc(j)}}{\mathbf{L}} \tag{4}$$

The parameter α can be set by users. If the fragment coverage of position j is less than the cutoff, that is $fc(j) \le fc^*$, the position j is regarded as a potential error position. If there are multiple continuous potential error positions, the region covering these positions can be group into a region set T ($T = \{[m, n] \mid n \ge m, \forall j \in [m, n], fc(j) \le fc^*\}$). For every region in set T, MAC chooses the position whose fragment coverage is the lowest as the breakpoint, B_p ($m \le B_p \le n$).

Owing to the uneven sequencing depth, some low-depth regions may be mistakenly categorized as containing error positions. Therefore, MAC estimates the coverage condition of the neighbor flanking regions for breakpoint B_p to reduce false positives. M_s is the number of paired reads whose left mate maps to the left flanking region of B_p, and right mate maps to other contigs. M_p is the number of paired reads whose right mate maps to the right flanking region of B_p, and left mate maps to other contigs. Then two rates: P_{cs} and P_{cp}, are calculated as follows (Wu et al., 2017).

$$\mathbf{P_{cs}} = \frac{\mathbf{fc(B_p)}}{\mathbf{fc(B_p)} + \mathbf{M_s}} \tag{5}$$

$$\mathbf{P_{cp}} = \frac{\mathbf{fc(B_p)}}{\mathbf{fc(B_p)} + \mathbf{M_p}} \tag{6}$$

P_{cs} and P_{cp} are used to estimate whether the region [m, n] is low-depth or not. If $P_{cs} > \beta$ or $P_{cp} > \beta$, the region is thought to be a low-depth region, and should be removed from the potential set.

Owing to the GC-content bias, some regions may cover less reads or even no reads, and these regions are mistakenly categorized as containing error positions. Therefore, MAC evaluates whether the GC-content of the neighbor flanking regions for B_p is too high or too low. P_{GC} is the GC-content rate of the potential error region which contains B_p, and P_g is the GC-content of the whole genome, P_g is calculated as follows (Wu et al., 2017).

$$\mathbf{P_g} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{L_i} \mathbf{I_j}}{\sum_{i=1}^{N} \mathbf{L_i}}, \tag{7}$$

where N represents the number of contigs, L_i is the length of the i-th contig, I_j is an indicator variable, when the base at position i is G or C, I_j equals to 1, otherwise, I_j equals to 0. If $P_{GC} \ge P_g + 1$, the region is thought to be GC-rich, otherwise, the region is thought to be GC-poor. Both GC-rich and GC-poor regions are removed from the potential set.

After removing the low-depth regions and GC-bias regions, the remaining single potential positions and potential regions are certainly false. which satisfy the following conditions at the same time:

① $fc(j) \le fc^*$;
② $P_{cs} \le \beta$ and $P_{cp} \le \beta$;
③ $P_{GC} < P_g + 1$ and $P_{GC} > P_g + 1$.

The regions estimated as low-depth or GC-bias are thought to be uncertain regions, and the positions in these regions satisfy the following conditions simultaneously:

① $fc(j) \le fc^*$;
② $P_{cs} > \beta$ or $P_{cp} > \beta$;
③ $P_{GC} \ge P_g + 1$ or $P_{GC} \le P_g + 1$.

After excluding the above two types of positions, the rest positions are certainly true. For the certainly false positions/regions, MAC breaks the corresponding contigs at the false position or the B_p position of the false region, and eliminate certainly false positions. Based on the above evaluation, all the

input contigs can be divided into two types: Uncertain (U) and True (T). If the contig contains one or more uncertain regions, the contig is classified as U contig, while the contig only containing true positions is classified as T contig.

Adopting Adjacency Algebraic Model

The order of merging is determined by the ranking of overall scores, which are calculated in the previous step. MAC merges two assemblies at a time, the next assembly and the resultant assembly are merged iteratively. In the merging process, MAC utilizes an adjacency algebraic model (Sankoff et al., 2000) to find the conjunctions between contigs. The adjacency algebraic model was introduced by Feijã£O and Meidanis to find a permutation to minimize the algebraic rearrangement distance (Feijã£O and Meidanis, 2013), and the adjacency algebraic model was proved to be efficient on the problem of contig ordering (Chen et al., 2018). In this study, MAC uses the adjacent graph to represent the adjacency algebraic model and utilizes the classification of contigs to optimize the model, the pseudo-code of adopting the adjacency algebraic model is as shown in **Algorithm 1 of Supplementary Material**.

Constructing Adjacent Graph

Given two input contig sets *O* and *R*, MAC utilizes the NUCmer package from MUMmer (Kurtz et al., 2004) to identify the high similarity fragments between *O* and *R*, which are called "consensus blocks", and numbers these consensus blocks. Two consensus blocks are thought to be adjacent, if they are next to each other, or if they overlap each other end-to-end with the overlapping length of l ($l \leq l_{cmin}$* 0.01), where l_{cmin} is the smaller one between the lengths of two consensus blocks, l is called the adjacent region. In general, there are two or more consensus blocks in one contig, and the consensus blocks may connect with each other, or maybe contain intervals between them. As described above, every contig can be divided into two types: "U" and "T". For the "U" type of contigs, if potential error positions locate at the adjacent regions of consensus blocks, the position information is retained. Otherwise, if potential error positions locate at the center region of consensus blocks, these positions are thought to be reliable, and could be removed form the potential error set. For the "T" type of contigs, MAC retains its state. MAC distributes the orientation for every consensus block, and uses tail("t") to denote the starting position, head("h") to denote the ending position. As shown in the example of **Figure 2**, 9 consensus blocks are found between two contig sets *O* and *R*, the adjacent relationships are enclosed in brackets, so *O* and *R* can be represented by *O* = {[1, 5], [9], [8, 2], [-3, 7], [6, 4]}, *R* = {[1, 6, 5], [4, 3], [2, 7, 8]}, respectively. As the orientation of consensus block "3" in *O* is reversed (from 3h to 3t), we use "-3" to represent this consensus block in *O*. In **Figure 2**, we suppose that there were uncertain positions between [1,5] in the first contig of O and [7,8] in the third contig of R, so these two contigs were regarded as "U" type, which are marked by red cycles on the contigs, and the corresponding consensus blocks are also marked with underlines in **Figure 2**, the detail classification strategy has been described above.

Then the adjacent graph $G = <V, E>$ is constructed, V is the vertices set of the adjacent graph, the single terminals or conjunctions of consensus blocks are regarded as vertices, in the example of **Figure 2**, 1t, 9t, 9h, 6t, and so on are the single terminals of *O* set, 1h5t, 6h4t, and so on are the conjunctions of *O* set. *E* is the edges set of the adjacent graph, an edge is added between *O* and *R* if two vertices have a terminal in common, such as 1h5t of *O* and 1h6t of *R* both have 1h, so there is an edge between 1h5t and 1h6t.

Extracting Good Paths

The major objective of the adjacent algebraic model is to minimize the algebraic distance between two contig sets, which can be denoted by $d(O,R) = N - C - \frac{P}{2}$ (Feijã£O and Meidanis, 2013), where N represents the number of contigs, C represents the number of cycles, P represents the number of paths in the adjacent graph G.

Through the demonstration of Lu et al., getting the minimum algebraic distance is equivalent to obtaining the maximum number of cycles, and the term of "good path" is defined for the cycle (closing path), which can connect multiple consensus blocks to generate a longer assembly. Here we define the conjunctions between two consensus blocks as adjacency, which are enclosed in square brackets in **Figure 2**. The paths in adjacent graph can be summarized according to the length, whether two ends of the path in the same set or in the same adjacency or not. We list all the possible combinations of the features mentioned above, as shown in **Table 1** there are 9 types of combinations in total. In the adjacent graph, two ends of the path appear in the same contig only if they appear in the same contig set, so if two ends of the path cannot be found in the same contig set, they cannot be found in the same contig or adjacency, thus for the types of No.3 and 7 in **Table 1**, two ends of the paths are in different contig sets, they cannot in the same adjacency, here we use "-" to represent the type is absent. If the length of path is odd, two ends should be found in different contig sets, so types of No.1 and 2 are absent. If the length of path is even, two terminals should be found in the same set, so type of No.8 is absent. However, there is an exceptional case, when two terminals form a circle, they can be found in different sets and different adjacencies.

From **Table 1**, four types are absent, and the type of No.6 is regarded as a good path, whose length is even, both of the ends are in the same set but in different adjacencies. There are two kinds of poor paths: No.4 and No.5. As the example in **Figure 2** shows, the paths of {4h, 4h3t, 3t7t, 2h7t, 2h}, {9t, 3h9t, 3h}, {7h, 7h8t, 8t} are good, which can form the cycles of [4h, 2h], [9t, 3h], and [7h, 8t]. Through the fusion of [4h, 2h], adjacencies [8, 2] and [6, 4] can be joined into [8, 2, -4, -6]. Through the fusion of [9t, 3h], adjacencies [9] and [-3, 7] can be joined into [-9, -3, 7]. Through the fusion of [7h, 8t], adjacencies [8, 2] and [-3, 7] can be joined into [-3, 7, 8, 2], and these two newly generated results can be further merged into [-9, -3, 7, 8, 2, -4, -6], equals to [6, 4, -2, -8, -7, 3, 9].

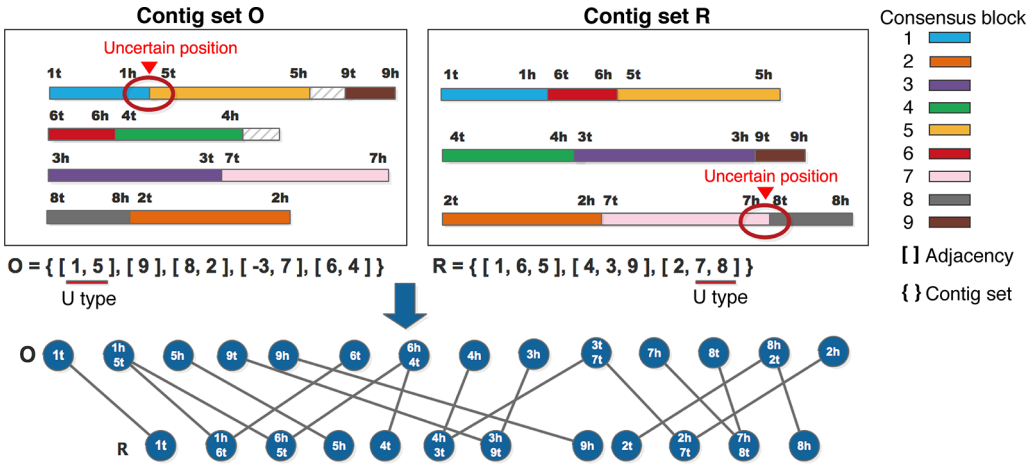


FIGURE 2 | An example for constructing adjacent graph.

Optimizing the Adjacency Algebraic Model

In the study of Lu et al., two odd paths (No.4 in **Table 1**) are chosen to join into a cycle repeatedly, until the odd path graph becomes an alternating cycle with the length of two. The remaining No.4 and No.5 paths can be arbitrarily joined together into two longer paths. However, in the actual implementation process, they found the fusion of these two types of paths resulting in error joints. In this study, MAC utilizes the classification of contigs to optimize the processing of poor paths in the adjacency algebraic model. Due to the circle paths in the graph represent the same adjacencies between two sets, so MAC maintains these paths without any process.

As described above, all the input contigs are divided into two types: Uncertain ("U") and True ("T"), the classification result is stored in the form of a label for every contig together with the potential error positions. After extracting good paths from the adjacent graph, the poor-1 type of paths can be further divided into two sub-types: single path and non-single path. The length of single path is 1, and two terminals are the same, for example, {1t, 1t}, {5h, 5h} and {3h, 3h} in **Figure 2** are single paths, {6t, 1h6t, 1h5t, 6h5t, 6h4t, 4t} is a non-single path. MAC uses the following steps to process poor paths:

- For non-single paths, MAC extracts the adjacencies which are included in the path, then checks the classification of contigs where the adjacencies are located. If contig is "U", and there is potential error position locating at adjacent region *l*, MAC splits the contig at the potential error position, and then reconstructs the sub-graph to extract good paths again.
- For single paths, MAC does not take any treatment, because during the process of graph reconstruction, some single paths would be eliminated automatically.
- For poor-2 type of paths, if both terminals of a poor-2 path appears in any good path, then the poor-2 path is thought to be spurious, and MAC removes this path along with the contig contained in the path. Otherwise, the poor-2 path can be retained temporarily.

MAC repeats these processing steps iteratively until there is no good path added, and single paths are merged into good paths to generate new adjacencies. For the new adjacencies, if there are overlapping blocks, the shorter adjacency is merged into a longer one. Here we use the same example as **Figure 2** to explain the optimization process, and the detail procedure is as shown in **Figure 2**.

After extracting good paths for the first time, the remaining are two types of poor paths. For non-single path {6t, 1h6t, 1h5t, 6h5t, 6h4t, 4t}, the adjacency included in the path are: [1, 5], [6, 4] in set *O* and contig [1, 6, 5] in set *R*, according to the processing steps, the error position of adjacency [1, 5] locates at the adjacent region, so the first contig of *O* should be split at the position. In fact, the head of block 1 and the head of block 6 contain the same repetitive sequences, which cause a misjoin between 1 and 5, as shown in the dashed line box of **Figure 3**. As such MAC could solve the problem of repetitive regions. Then the sub-graph is reconstructed, two more good paths are extracted, and there are four single paths remaining. After merging these single paths into good paths, the final contigs can be represented by the adjacencies: [1, 6, 4, -2, -8, -7, 3, 9] and [1, 6, 5]. MAC identifies there is an overlapping region between two adjacencies, and thus merges [1, 6, 5] into [1, 6, 4, -2, -8, -7, 3, 9] to get the final adjacency [1, 6, 5, 6, 4, -2, -8, -7, 3, 9].

TABLE 1 | Nine types of paths in the adjacent graph.

No.	Length of path	In the same set	In the same adjacency	Type
1	Odd	Y	Y	–
2	Odd	Y	N	–
3	Odd	N	Y	–
4	Odd	N	N	Poor-1
5	Even	Y	Y	Poor-2
6	Even	Y	N	Good
7	Even	N	Y	–
8	Even	N	N	–
9	Even	N	N	Circle

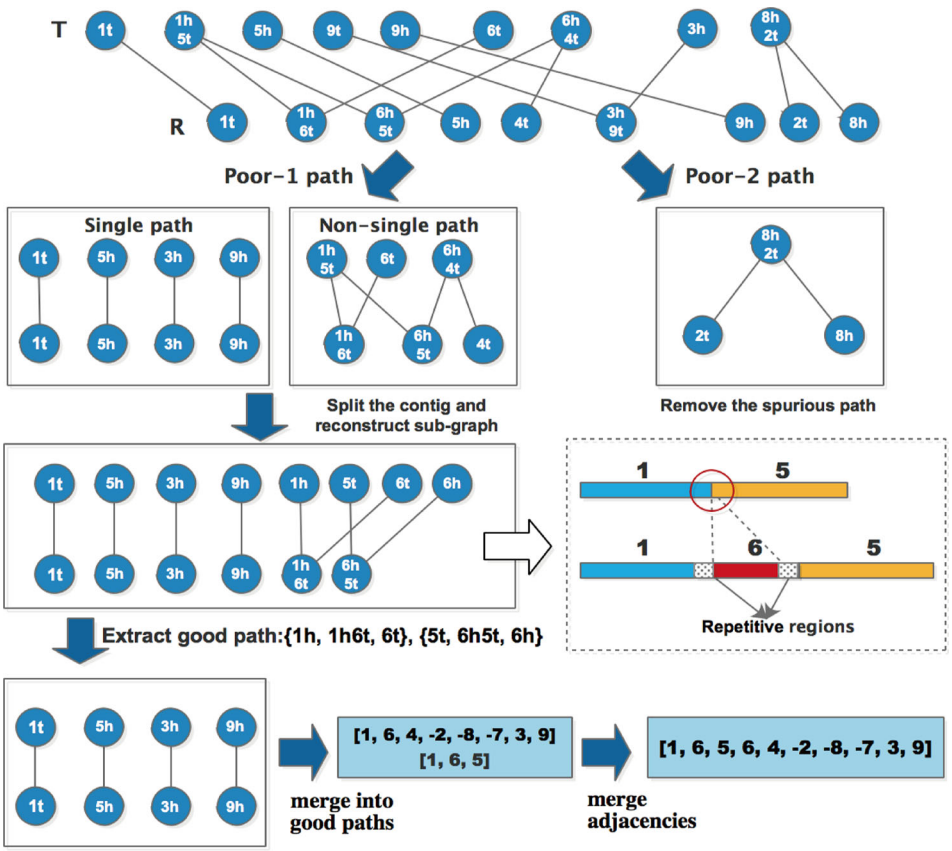


FIGURE 3 | Detail of optimization process.

EXPERIMENTAL RESULTS AND DISCUSSION

Datasets and Evaluation Metrics

In this study, we perform the experiments on four real bacterial genomes: M.abscessus, B.fragilis, R.sphaeroides and V.cholerae from GAGE-B (Genome Assembly Gold-standard Evaluation for Bacteria) (Tanja et al., 2013), GAGE-B evaluates the performance of multiple genome assemblers on a spectrum of bacterial genomes sequenced by the sequencing technologies of MiSeq and HiSeq. Here, we use the sequences generated by MiSeq technology, the average read length of these four species is 250 bp, the coverage is 100x, and the genome sizes are 5.1 Mb, 5.4 MB, 4.6 Mb, and 4.0 Mb, respectively. The

TABLE 2 | The experimental results of *M. abscessus*.

	Contigs num	Largest contig	Size	Genome fraction	N50	NGA50	MA
Velvet	203	226,629	5,136,825	98.965	48,155	41,485	54
ABySS	149	245,660	5,116,522	98.926	70,424	68,549	2
SOAPdenovo	91	286,460	5,133,667	99.139	131,561	113,272	19
(Velvet+ABySS)							
GAA	339	129,152	5,152,501	99.094	39,271	37,715	61
MIX	118	245,660	5,376,417	98.891	108,584	70,302	**18**
Metassembler	200	226,629	5,130,215	98.944	48,155	41,485	54
MAC	190	317,945	9,856,881	99.304	**163,219**	**90,766**	58
(Velvet+ABySS+SOAP)							
GAA	211	210,497	5,146,833	99.129	54,850	50,904	55
MIX	91	286,460	5,133,667	99.139	131,561	113,272	17
Metassembler	191	226,629	4,934,916	95.03	47,284	39,706	64
MAC	80	287,168	5,146,285	99.249	**141,537**	**131,561**	**10**

The bolded data indicates the highest value of N50 or NGA50 within each comparison.

TABLE 3 | The experimental results of *B. fragilis*.

	Contigs num	Largest contig	Size	Genome fraction	N50	NGA50	MA
Velvet	373	91,844	5,310,336	97.661	24,465	24,465	3
ABySS	87	430,487	5,380,960	98.451	130,570	130,570	2
SOAPdenovo *(Velvet+ABySS)*	79	606,530	5,341,631	98.226	246,346	246,346	0
GAA	2053	16,951	10,676,299	98.811	4,999	4,999	4
MIX	87	430,487	5,380,960	98.451	130,570	130,570	**2**
Metassembler	256	127,644	5,317,077	97.819	40,339	39,580	3
MAC *(Velvet+ABySS+SOAP)*	136	568,455	10,618,547	98.812	**270,064**	**144,965**	9
GAA	2933	429,861	15,592,962	98.896	6,079	6,075	4
MIX	55	700,546	6,089,165	98.554	353,741	380,728	9
Metassembler	194	215,440	5,317,760	97.819	57,802	57,596	3
MAC	42	1,195,331	5,355,147	98.306	**485,219**	**455,989**	**2**

The bolded data indicates the highest value of N50 or NGA50 within each comparison.

TABLE 4 | The experimental results of *R. sphaeroides* The highest value of N50 or NGA50 within each comparison.

	Contigs num	Largest contig	Size	Genome fraction	N50	NGA50	MA
Velvet	332	71,713	4,485,514	97.419	23,979	24,300	2
ABySS	382	71,578	4,503,182	97.76	21,441	21,441	1
SOAPdenovo *(Velvet+ABySS)*	354	115,051	4,527,360	97.98	33,491	33,491	1
GAA	1745	9,976	8,988,696	98.651	6,650	6,650	3
MIX	274	113,766	4,728,490	97.493	35,067	28,685	35
Metassembler	325	71,713	4,480,778	97.337	23,979	23,979	**2**
MAC *(Velvet+ABySS+SOAP)*	434	126,603	8,043,496	98.718	**53,057**	**52,641**	17
GAA	2683	13,133	13,487,438	99.281	7,589	7,571	4
MIX	237	171,915	4,982,251	98.446	51,508	41,915	22
Metassembler	323	71,713	4,477,669	97.269	23,979	23,979	**2**
MAC	122	173,958	4,574,809	98.282	**58,392**	**56,244**	7

The bolded data indicates the highest value of N50 or NGA50 within each comparison.

TABLE 5 | The experimental results of *V. cholerae*.

	Contigs num	Largest contig	Size	Genome fraction	N50	NGA50	MA
Velvet	156	246,346	3,944,260	97.563	92,036	63,574	14
ABySS	196	178,118	3,904,784	96.699	61,965	60,272	2
SOAPdenovo *(Velvet+ABySS)*	186	246,179	3,924,635	96.94	71,357	65,464	16
GAA	271	170,890	3,958,224	97.207	73,177	56,472	14
MIX	147	310,702	4,038,894	96.915	124,754	91,942	19
Metassembler	150	246,346	3,935,482	97.48	92,036	63,574	**13**
MAC *(Velvet+ABySS+SOAP)*	232	312,914	7,221,147	97.322	**174,216**	**163,176**	21
GAA	160	243,299	3,981,614	97.713	110,446	110,446	16
MIX	118	310,703	4,338,139	97.496	112,745	86,841	32
Metassembler	145	246,346	3,914,378	96.972	93,191	63,574	**13**
MAC	87	358,265	3,997,554	97.709	**167,523**	**110,538**	**13**

The bolded data indicates the highest value of N50 or NGA50 within each comparison.

insert sizes are 335 bp, 600 bp, 540 bp, and 335 bp, respectively. The detail information of raw data is listed in **Table S1** of Supplementary Material. All the assemblies and paired-end reads are available at the website of GAGE-B (http://ccb.jhu.edu/gage_b/).

The evaluation tool Quast (Gurevich et al., 2013) is used to estimate the contiguity and correctness of assemblies. For the metrics provided by Quast, N50 is the metric to evaluate contiguity without reference, and NGA50 could compare the assemblies to a reference genome to get more accurate and comprehensive evaluation. The number of misassemblies is an important metric to measure the correctness of assemblies. In most cases, the increase of N50 and NGA50 inevitably leads to more misassemblies. The contig sets generated by different tools

are evaluated in **Tables S2–S5** of Supplementary Material. The major objective of MAC is to increase the contiguity of assemblies, at the same time make sure the number of misassemblies reducing or at least not increasing.

Experimental Results

Although lots of assembly reconciliation tools have been developed, some of the tools have stopped updating, some need the reference of relative species, and some tools don't fit for merging relatively longer next-generation sequences which are in ~250bp. Therefore, we exclude these unsuitable tools, and only compare MAC with three assembly reconciliation tools: GAA, MIX and Metassembler. The assemblies are generated by Velvet, ABySS and SOAPdenovo. The contigs generated by Velvet are fragmental and with poor contiguity. ABySS could provide more reliable contigs which have less misassemblies. SOAPdenovo is a powerful tool which produces higher contiguity and correctness contigs. In the experiment, we test the merging performance of four reconciliation tools on the assemblies which have different features. The experiment results are shown in **Tables 2**–**5**. For each dataset, we take the experiments on two assemblies as input (Velvet+ABySS), and multiple assemblies as input (Velvet+ABySS+SOAPdenovo). "MA" in tables represent the numbers of misassemblies.

For the case of two assemblies as input, the number of misassemblies of four reconciliation tools have increased somewhat because the quality of input assemblies is relatively low. Even the metrics of N50/NGA50 have remained static or decreased for some tools, such as GAA and Metassembler. By comparison, MAC achieves significant growth in N50 and NGA50 compared to the original input assemblies and the merging results of other reconciliation tools in four datasets, although the number of misassemblies is basically flat.

For the case of three assemblies as input, the metrics of N50/NGA50 of four reconciliation tools have increased in various degrees, due to the addition of high quality assemblies generated by SOAPdenovo, while there is no obvious change in the number of misassemblies for GAA, MIX, and Metassembler. However, MAC not only achieves the obvious increase of N50 and NGA50, but also greatly reduces the number of misassemblies. Especially in the dataset of B.fragilis, the N50 and NAG50 of MAC are 485219 and 455989, which have the growth rate of 79.6% and 214.5%, respectively, compared to the case of two assemblies as input of MAC, and the growth rate of 96.9% and 81.7%, respectively, compared to the high quality input of SOAPdenovo. What's more, the number of misassemblies has decreased from 9 to 2, which is less than the number of velvet and equals to the number of ABySS.

From the results of **Tables 2**–**5**, MAC outperforms the other reconciliation tools, MAC is not only adapted to merge low quality assemblies to generate a more continuous one, but is also good at fusing different features between assemblies to further improve the contiguity of high quality assemblies, at the same time maintaining the correctness. In addition, we evaluated the computational costs of four tools, as shown in **Table S6** of supplementary material.

CONCLUSION

In this study, we have proposed a novel assembly reconciliation tool MAC. MAC classifies all the contigs into "U" and "T" by using alignment results and GC-content of paired-end reads, then identifies consensus blocks between assembly sets, through which unreliable fragments caused by uneven sequencing depth or sequencing errors could be filtered out. In addition, MAC utilizes adjacency algebraic model to fulfill the merging process. The adjacent graph is employed to identify good paths between consensus blocks, which could be used to generate some accurate fusions. Secondly, the classification result of contigs is used to optimize the processing steps of poor paths, through which repetitive regions could be eliminated by splitting contigs and reconstructing the adjacent graph. What's more, to solve the problem of unknown ranking of input assemblies, MAC designs a scoring function to evaluate the overall quality of assembly sets. The experimental results from four real species of GAGE-B illustrate that MAC performs better than other reconciliation tools.

DATA AVAILABILITY STATEMENT

The datasets of GAGE-B for this study can be found in https://ccb.jhu.edu/gage_b/. The source code of MAC is available at https://github.com/bioinfomaticsCSU/MAC.

AUTHOR CONTRIBUTIONS

LT, ML, F-XW, YP, and JW conceived the original study. LT carried out the analysis of sequencing data and developed the bioinformatics tool. ML contributed to designing the algorithm structure. F-XW, YP, and JW contributed to optimizing the performance of tool. All authors contributed to drafting the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China [No. 61732009, No. 61772557, and No. 61420106009].

ACKNOWLEDGMENTS

The Fourth CCF Bioinformatics Conference (CBC 2019).

SUPPLEMENTARY MATERIALS

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01396/full#supplementary-material

REFERENCES

Alkan, C., Sajjadian, S., and Eichler, E. E. (2010). Limitations of next-generation genome sequence assembly. *Nat. Methods* 8 (1), 61.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., et al. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2 (1), 10.

Casagrande, A., Fabbro, C. D., Scalabrin, S., and Policriti, A. (2009). "GAM: Genomic Assemblies Merger: A Graph Based Method to Integrate Different Assemblies", 2009 IEEE International Conference on Bioinformatics and Biomedicine. 18. (IEEE Computer Society), 321–326. doi: 10.1109/BIBM.2009.28.

Chen, K. T., Liu, C. L., Huang, S. H., Shen, H. T., Shieh, Y. K., Chiu, H. T., et al. (2018). CSAR: a contig scaffolding tool using algebraic rearrangements. *Bioinformatics*. 34 (1), 109–111. doi: 10.1093/bioinformatics/btx543

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4 (4), 265. doi: 10.1038/nnano.2009.12

Diguistini, S., Liao, N. Y., Platt, D., Robertson, G., Seidel, M., Chan, S. K., et al (2009). De novo genome sequence assembly of a filamentous fungus using sanger, 454 and illumina sequence data. *Genome Biol.* 10 (9), R94. doi: 10.1186/gb-2009-10-9-r94

Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., et al (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* 21 (12), 2224–2241. doi: 10.1101/gr.126599.111

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time dna sequencing from single polymerase molecules. *Science* 323 (5910), 133–138. doi: 10.1126/science.1162986

Feijã£O, P., and Meidanis, J. (2013). Extending the algebraic formalism for genome rearrangements to include linear chromosomes. *IEEE/ACM Trans.Comput. Biol. Bioinf.* 10 (4), 819–831. doi: 10.1109/TCBB.2012.161

Francesco, V., Alberto, P., and Federica, C. (2011). e-RGA: enhanced reference guided assembly of complex genomes. *Embnet J.* 17 (1), 46–54. doi: 10.14806/ej.17.1.208

Gnerre, S., and Jaffe, D. B. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. proc natl acad sci usa. *Proc. Nat. Acad. Sci. U. S .A.* 108 (4), 1513–1518. doi: 10.1073/pnas.1017351108

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). Quast: quality assessment tool for genome assemblies. *Bioinformatics* 29 (8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Bio.l* 10 (3), 1–13. doi: 10.1186/gb-2009-10-3-r32

Kent, W. J. (2002). Blat–the blast-like alignment tool. *Genome Res.* 12 (4), 656–664. doi:10.1101/gr.229202

Kolmogorov, M., Raney, B., Paten, B., and Pham, S. (2014). Ragout—a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* 30 (12), 302–309.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., et al (2004). Versatile and open software for comparing large genomes. *Genome Biol.* 5 (2), R12. doi: 10.1186/gb-2004-5-2-r12

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20 (2), 265–272. doi: 10.1101/gr.097261.109

Li, M., Tang, L., Wu, F. X., Pan, Y., and Wang, J. (2018a). SCOP: a novel scaffolding algorithm based on contig classification and optimization. *Bioinformatics* 35 (7), 1142–1150. doi: 10.1093/bioinformatics/bty773

Li, M., Tang, L., Liao, Z., Luo, J., Wu, F. X., Pan, Y., et al (2018b). A novel scaffolding algorithm based on contig error correction and path extension. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16 (3), 764–773. doi: 10.1109/TCBB.2018.2858267

Li, H. (2015). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 32 (14), 2103. doi: 10.1093/bioinformatics/btw152

Liao, X., Li, M., Luo, J., Zou, Y., Wu, F., Pan, Y., et al. (2018). Improving de novo assembly based on reads classification. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 14 (8). doi: 10.1109/TCBB.2018.2861380

Liao, X., Li, M., Luo, J., Zou, Y., Wu, F., Pan, Y., et al. (2019a).EPGA-SC: a framework for de novo assembly of single-cell sequencing reads. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2945761

Liao, X., Li, M., Zou, Y., Wu, F., Pan, Y., and Wang, J. (2019b). An efficient trimming algorithm based on multi-feature fusion scoring model for NGS data. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2897558

Lin, S. H., and Liao, Y. C. (2013). Cisa: contig integrator for sequence assembly of bacterial genomes. *Plos One* 8 (3), e60843. doi: 10.1371/journal.pone.0060843

Lu, C. L. (2015). An efficient algorithm for the contig ordering problem under algebraic rearrangement distance. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.* 22 (11), 975. doi: 10.1089/cmb.2015.0073

Luo, J., Wang, J., Li, W., Zhang, Z., Wu, F. X., Li, M., et al. (2015). EPGA2: memory-efficient de novo assembler. *Bioinformatics* 31 (24), 3988–3990. doi: 10.1093/bioinformatics/btv487

Luo, J., Wang, J., Zhang, Z., Li, M., and Wu, F. X. (2017). BOSS: a novel scaffolding algorithm based on an optimized scaffold graph. *Bioinformatics* 33 (2), 169–176. doi: 10.1093/bioinformatics/btw597

Nijkamp, J., Winterbach, W., Van, d. B. M., Daran, J. M., Reinders, M., and De, R. D. (2010). Integrating genome assemblies with maia. *Bioinformatics* 26 (18), i433–i439. doi: 10.1093/bioinformatics/btq366

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al (2012). GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 22 (3), 557–567. doi: 10.1101/gr.131383.111

Sankoff, D., Nadeau, J. H., Sankoff, D., and Nadeau, J. H. (2000). Comparative genomics: empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families. *Kluwer Acad. Publishers* , 213–223.doi: 10.1007/978-94-011-4309-7

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Res.* 20 (9), 1165–1173. doi: 10.1101/gr.101360.109

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M.İnanç Birol (2009). Abyss: a parallel assembler for short read sequence data. *Genome Res.* 19 (6), 1117. doi: 10.1101/gr.089532.108

Soueidan, H., Maurier, F., Groppi, A., Sirand-Pugnet, P., Tardy, F., Citti, C., et al. (2013). Finishing bacterial genome assemblies with mix. *Bmc Bioinf* 14 (S15), S16. doi: 10.1186/1471-2105-14-S15-S16

Tanja, M., Stephan, P., Stefan, C., Liu, X., Su, Q., Daniela, P., et al. (2013). GAGE-b: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 29 (14), 1718–1725. doi: 10.1093/bioinformatics/btt273

Vicedomini, R., Vezzi, F., Scalabrin, S., Arvestad, L., and Policriti, A. (2013). Gamngs: genomic assemblies merger for next generation sequencing. *Bmc Bioinf.* 14 (7), 1–18. doi: 10.1186/1471-2105-14-S7-S6

Wences, A. H., and Schatz, M. C. (2015). Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* 16 (1), 207. doi: 10.1186/s13059-015-0764-4

Wu, B., Wang, J., Luo, J., Li, M., Wu, F., and Pan, Y. (2017). MEC: Misassembly error correction in contigs using a combination of paired-end reads and GC-contents, in: IEEE International Conference on Bioinformatics and Biomedicine IEEE Computer Society. 216–221. doi: 10.1109/TCBB.2018.2876855

Yao, G., Ye, L., Gao, H., Minx, P., Warren, W. C., and Weinstock, G. M. (2012). Graph accordance of next-generation sequence assemblies. *Bioinformatics* 28 (1), 13–16. doi: 10.1093/bioinformatics/btr588

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res* 18 (5), 821–829. doi: 10.1101/gr.074492.107

Zimin, A. V., Smith, D. R., Sutton, G., and Yorke, J. A. (2005). Assembly reconciliation. *Bioinformatics* 24 (1), 42–45.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 31 January 2020
doi: 10.3389/fgene.2019.01346

Probing lncRNA–Protein Interactions: Data Repositories, Models, and Algorithms

Lihong Peng[1†], Fuxing Liu[1†], Jialiang Yang[2], Xiaojun Liu[1], Yajie Meng[3], Xiaojun Deng[1], Cheng Peng[1], Geng Tian[2] and Liqian Zhou[1*]*

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] Department of Sciences, Genesis (Beijing) Co. Ltd., Beijing, China, [3] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

Identifying lncRNA–protein interactions (LPIs) is vital to understanding various key biological processes. Wet experiments found a few LPIs, but experimental methods are costly and time-consuming. Therefore, computational methods are increasingly exploited to capture LPI candidates. We introduced relevant data repositories, focused on two types of LPI prediction models: network-based methods and machine learning-based methods. Machine learning-based methods contain matrix factorization-based techniques and ensemble learning-based techniques. To detect the performance of computational methods, we compared parts of LPI prediction models on Leave-One-Out cross-validation (LOOCV) and fivefold cross-validation. The results show that SFPEL-LPI obtained the best performance of AUC. Although computational models have efficiently unraveled some LPI candidates, there are many limitations involved. We discussed future directions to further boost LPI predictive performance.

Keywords: lncRNA–protein interaction, computational method, network-based method, machine learning-based method, data repositories

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Guohua Huang,
Shaoyang University, China
Guoxian Yu,
Southwest University, China

****Correspondence:***
Geng Tian
tiang@genesis.com
Liqian Zhou
zhoulq11@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *27 September 2019*
Accepted: *09 December 2019*
Published: *31 January 2020*

Citation:
Peng L, Liu F, Yang J, Liu X, Meng Y, Deng X, Peng C, Tian G and Zhou L (2020) Probing lncRNA–Protein Interactions: Data Repositories, Models, and Algorithms. Front. Genet. 10:1346. doi: 10.3389/fgene.2019.01346

INTRODUCTION

Long non-coding RNAs (lncRNAs) are transcripts with greater than 200 nucleotides but lack protein coding capacity (Sanchez Calle et al., 2018). lncRNAs are closely associated with various key biological processes, such as cell cycle regulation, immune response, and embryonic stem cell pluripotency (Liu et al., 2018; Agirre et al., 2019; Li et al., 2019b). More importantly, lncRNAs play an important role in understanding pathogenesis of various diseases, especially tumors (Chen et al., 2016a; Fu et al., 2017; Jiang et al., 2018; He et al., 2018a; Dallner et al., 2019). Although lncRNAs play a spectrum of regulatory roles across different cellular pathways, understanding about their regulatory mechanisms is very limited (Munschauer et al., 2018).

Recently, one broad theme is that lncRNAs can drive the assembly of RNA–protein complexes by facilitating the regulation of gene expression (Rinn and Chang, 2012; Chen and Yan, 2013; Hentze et al., 2018; Munschauer et al., 2018; Nozawa and Gilbert, 2019). lncRNAs achieve their specific functions by interacting with multiple proteins and thus regulating multiple cellular processes (Zhang et al., 2018c; Pyfrom et al., 2019). Studies reported that lncRNAs can activate post-transcriptional gene regulation, splicing, and translation by binding to proteins (Zhang et al,. 2018c; Li et al., 2019a) Therefore, identifying possible lncRNA–protein interactions (LPIs) is

essential for unraveling lncRNA-related activities (Qian et al., 2018; Zhang et al., 2018c; Zhao et al., 2018c). Wet experiments validated parts of LPIs, but experimental methods remain costly and time-consuming. Therefore, different computational models are explored to infer potential LPIs (Pan et al., 2016; Cheng et al., 2018; Zhang et al., 2018c; Zhao et al., 2018c). There exist numerous unexplored lncRNAs and proteins in public databases, which makes it possible to efficiently identify their underlying associations.

In this study, we introduced relevant repositories, summarized computational models and algorithms for LPI prediction, discussed their advantages and weaknesses by comparison, and presented further directions for boosting LPI prediction performance. We focused on two categories of computational models: network-based methods and machine learning-based methods. The machine learning-based methods contain matrix factorization-based methods and ensemble learning-based methods.

RELEVANT REPOSITORIES

There are abundant repositories related to LPI prediction. These repositories provide diverse information for efficiently uncovering potential LPIs.

Noncode

The NONCODE database (Zhao et al., 2015) (http://www.noncode.org/) is an interactive database aiming to collect the most complete annotation for ncRNAs, especially lncRNAs. The latest NONCODE database (current version v5.0) contains lncRNA information from 17 species including human, mouse, cow, rat, chimp, gorilla, orangutan, rhesus, opossum, platypus, chicken, zebrafish, fruit fly, *Caenorhabditis elegans*, yeast, Arabidopsis, and pig. There are 548,640 lncRNAs in the latest version. There are 172,216 and 131,697 lncRNAs from human and mouse, respectively. More importantly, NONCODE has introduced some important features including conservation annotation, lncRNA–disease associations, and an interface to select credible datasets.

NPInter

The NPInter database (Hao et al., 2016) (http://www.bioinfo.org.cn/NPInter/contact.htm) provides abundant association data that are experimentally verified. For example, the database contains information on interactions between noncoding RNAs (ncRNAs) and biomolecules including proteins, mRNAs, miRNAs, and genomic DNAs. The database contains 491,416 interactions in 188 tissues/cell lines from 68 types of experimental technology.

RAID

The RAID database (Yi et al., 2016) (http://www.rna-society.org/raid/) includes more than 40,668 lncRNA-associated RNA–protein interactions and more than 34,790 lncRNA-associated RNA–RNA interactions.

starBase

The starBase database (Li et al., 2013) (http://starbase.sysu.edu.cn/) contains more than 1,100,000 miRNA–ncRNA (CLIP) interactions, 117,000 RNA-binding protein (RBP)–ncRNA interactions, and 32,000 miRNA–ncRNA interactions. In addition, it provides more than 10,800 RNA-seq data and 10,500 miRNA-seq data from 32 cancer types and 3,236,000 mutations from 366 disease types.

VirBase

The ViRBase database (Li et al., 2014) (http://www.rna-society.org/virbase) integrates experimental and predictive association information from manual literature curation and other resources based on one common framework from 119 species, especially ncRNA-associated virus–virus, host–host, host–virus, and virus–host interactions.

POSTAR2

The POSTAR2 database (Zhu et al., 2018) (http://lulab.life.tsinghua.edu.cn/postar2/index.php) provides various post-transcriptional regulation data based on CLIP-seq, Ribo-seq, RNA-seq, and other high-throughput sequencing information from six species: yeast, Arabidopsis, fly, worm, mouse, and human. It hosts about 40 million RBP binding sites validated by CLIP-seq experiments. It provides three modules: the "RBP" module, "RNA" module, and "Translatome" module. The "RBP" module contains RBP binding sites and their annotations and functions. The "RNA" module is composed of a few sub-modules, including "disease," "variation," "crosstalk," and "binding sites," and is applied to annotate the RBP binding sites.

ChIPBase

The ChIPBase database (Zhou et al., 2016) (http://rna.sysu.edu.cn/chipbase/) is used to identify transcription factor binding sites and motifs, and decode transcriptional regulatory networks of miRNA, lncRNAs, and other ncRNAs from ChIP-seq data. It provides about 10,200 curated peak datasets from 10 species: human, mouse, fruit fly, worm, *Arabidopsis thaliana*, yeast, rat, zebrafish, *Xenopus tropicalis*, and chicken.

LNCipedia

The LNCipedia database (Volders et al., 2018) (https://lncipedia.org) is a comprehensive database. Its central work is to merge redundant transcripts from different data sources and group the transcripts into genes, thus producing a highly consistent database. The latest update of lncRNA (LNCipedia 5) contains information about annotation and sequence for 1,555 human lncRNAs from 2,482 lncRNA publications. This information originates from Ensembl (Cunningham et al., 2018), RefSeq (Rajput et al., 2018), and FANTOM CAT (Hon et al., 2017).

lncRNA2target

The lncRNA2Target database (Cheng et al., 2018) (http://123.59.132.21/lncrna2target) contains a comprehensive repository of lncRNA target genes to provide information about target genes regulated by lncRNAs. The latest version provides a special web

interface in which users can search the targets for a particular lncRNA or the lncRNAs for a particular gene.

lncRNAdb

The lncRNAdb database (Quek et al., 2014s) (http://lncrnadb.org) is a comprehensive database in compliance with the International Nucleotide Sequence Database Collaboration. It provides 287 eukaryotic lncRNAs and an interface enabling users to access sequence data, expression information, and the literature. The latest update of lncRNAdb integrated nucleotide sequence information, Illumina Body Atlas expression profiles, and a BLAST search tool.

lncRNASNP2

The lncRNASNP2 database (Miao et al., 2017) (http://bioinfo.life.hust.edu.cn/lncRNASNP2) provides 7,260,238 single nucleotide polymorphisms (SNPs) on 141,353 human lncRNA transcripts, and 3,921,448 SNPs on 117,405 mouse lncRNA transcripts. More importantly, it contains abundant information about mutations in lncRNAs and their impacts on lncRNA structure and function. It also provides online tools for analyzing new variants in lncRNA.

LbcRNAwiki

The lncRNAWiki database (Ma et al., 2014) (http://lncrna.big.ac.cn) integrated various human lncRNAs from different resources. It makes existing lncRNAs able to be updated, edited, and curated by diverse users. More importantly, any user can add newly uncovered lncRNAs.

Lnc2Cancer

The Lnc2Cancer database (Gao et al., 2018) (http://www.bio-bigdata.net/lnc2cancer) provides lncRNA–cancer associations supported by experiments. It contains 4,989 associations between 165 human cancer subtypes and 1,614 human lncRNAs, 366 experimentally validated circulating-related lncRNA-cancer associations, 593 drug-resistance-related lncRNA-cancer associations, and 1,928 prognosis-related lncRNA–cancer associations, and abundant lncRNA regulatory mechanisms in cancers including 211, 1139, 225, and 319 lncRNAs regulated by variant, miRNA, transcription factor, and methylation, respectively.

LncRNAdisease

The lncRNADisease database (Bao et al., 2018) (http://www.rnanut.net/lncrnadisease/) integrated experimentally validated circular RNA–disease associations, and regulatory mechanisms among mRNA, miRNA, and ncRNA. Particularly, it contains more than 200, 000 lncRNA–disease associations. In addition, it gives confidence scores for all ncRNA–disease associations and maps each disease to disease ontology and medical subject headings.

MNDR

The MNDR database (Cui et al., 2017) (http://www.rna-society.org/mndr/) integrates more than 260,000 ncRNA–disease associations. These associations are supported by 10 experiments and 4 predictive algorithms. The experimental repositories include Lnc2Cancer (Gao et al., 2018), dbDEMC (Yang et al,. 2016), LncRNADisease (Bao et al., 2018), MNDR (Wang et al., 2013), HMDD (Huang et al., 2018b), NSDNA (Wang et al., 2016a), LincSNP (Ning et al., 2016), miRCancer (Xie et al., 2013), PhenomiR (Ruepp et al., 2012), and miR2Disease (Jiang et al., 2008). The four prediction algorithms are LDAP (Lan et al., 2016), miRDP (Mørk et al., 2013) LncDisease (Wang et al., 2016b), and PBMDA (You et al., 2017). It provides 8,824 experimental lncRNA–disease, 70,381 experimental miRNA–disease, 118 experimental piRNA–disease, and 67 experimental snoRNA–disease associations across 6 mammalsix (*Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Pan troglodyte*, *Rattus norvegicus*, and *Sus scrofa*). In addition, it provides 153,508 predicted lncRNA–disease associations and 28,144 predicted miRNA–disease associations for *H. sapiens*. MNDR contains 19,575, 110, 4,150, and 23 non-redundant lncRNA–disease, piRNA–disease, miRNA–disease, and snoRNA–disease interactions, respectively, associated with 1,416 disease.

UniProt

The UniProt database (Consortium et al., 2018) (http://www.uniprot.org/) is an important database providing protein sequences and annotations. It provides 80 million sequences and is a useful tool. Users can calculate a new proteome identifier to find a particular assembly for a species or subspecies. It also provides an effective measurement for computing an annotation score for all entries.

METHODS

Most computational methods contain two procedures: data extraction and model selection. In the first part, computational methods usually extract LPIs related to human lncRNA, lncRNA sequences, and protein sequences from NPInter (Hao et al., 2016), NONCODE (Zhao et al., 2015), and UniProt (Consortium et al., 2018), respectively. Computational methods filter LPIs by removing lncRNAs/proteins only interacting with one protein/lncRNA. In the second procedure, computational methods design various models to uncover potential LPIs. These models can be roughly classified into two categories: network-based methods and machine learning-based methods.

Data Representation

Computational methods utilize an lncRNA set $l = \{l_1, l_2, l_3 \ldots. l_n\}$, a protein set, $P = \{p_1, p_2, p_3 \ldots. p_m\}$, and an LPI matrix $Y_{n\times m}$, where $y_{ij} = 1$ if there is an association between an lncRNA l_i and a protein p_j; otherwise, $y_{ij} = 0$.

Network-Based Methods

Network-based methods obtain better performance by effectively integrating related biological information and network propagation algorithms into a unified framework.

LPIHN

Li et al. (2015) developed an LPI prediction method combing a heterogeneous network model and random walk with restart, LPIHN. LPIHN can be broken down into four steps:

Step 1 Extracting known ncRNA–protein associations from the Npinter 2.0 database (Hao et al., 2016) and filtering the ncRNAs and their associated proteins based on organism and type of ncRNAs. LPIHN then selects lncRNAs from filtered ncRNAs based on the human lncRNA dataset provided by the NONCODE database (Zhao et al., 2015).

Step 2 Obtaining lncRNA expression profiles from the NONOCODE 4.0 database (Zhao et al., 2015). Given the expression profiles of two lncRNAs E_1 and E_2, LPIHN calculates lncRNA expression similarity based on the Pearson correlation coefficient:

$$SL(i,j) = \left| \frac{\mathrm{cov}(E_1, E_2)}{\sigma_{e_1} \sigma_{e_2}} \right| \tag{1}$$

where $cov(E_1, E_2)$ is the covariance of E_1 and E_2, and σ_{e_1} and σ_{e_2} are the standard deviations of E_1 and E_2, respectively.

Step 3 Extracting protein–protein interactions (PPIs) from STRING 9.1 (Szklarczyk et al., 2016) and obtaining 804 PPIs and the corresponding score matrix *SP*. *SP* is normalized as follows:

$$SP^{*}_{ij} = \frac{SP_{ij}}{\sqrt{M(i,i)M(j,j)}} \tag{2}$$

where *M* is a diagonal matrix, and $M(i, i)$ is the sum of row *i* in *SP*.

Step 4 Propagating the random walk to score for unknown lncRNA–protein pairs based on the following iterative equation:

$$Y_{t+1} = (1-\delta) W^T Y_t + \delta\, Y_0 \tag{3}$$

The details are shown as **Figure 1**.

LPLNP

Zhang et al. (2018b) proposed a linear neighborhood propagation-based method, LPLNP, to probe potential LPIs. LPLNP found novel LPIs through the following steps.

Step 1 Extracting 4,158 LPIs between 27 proteins and 990 lncRNAs from NPInter (Hao et al., 2016) and NONCODE (Zhao et al., 2015) by filtering unreliable lncRNA sequences and removing lncRNAs/proteins only interacting with one protein/lncRNA.

Step 2 Obtaining three types of features for lncRNAs (interaction profile, expression profile, and sequence composition) and two types of features for proteins [interaction profile and CTD (composition, transition, and destruction)].

Step 3 Computing linear neighborhood similarity and regularized linear neighborhood similarity between lncRNA/proteins by Eqs. (4) and (5), respectively:

$$\begin{aligned} &\epsilon_i = \| X_i - \sum_{i_j : X_{i_j} \in N(X_i)} w_{ii_j} X_{i_j} \|^2 \\ &s.t. \sum_{i_j : X_{i_j} \in N(X_i)} w_{ii_j} = 1, w_{ii_j} \geq 0 \end{aligned} \tag{4}$$

where X_i denoted the feature vector of the *i*th lncRNA, and $N(X_i)$ is *K* nearest neighbors of X_i.

$$\begin{aligned} &\epsilon_i = w_i^T (G^i + \lambda I) w_i \\ &s.t. \sum_{i_j : X_{i_j} \in N(X_i)} w_{ii_j} = 1, w_{ii_j} \geq 0 \end{aligned} \tag{5}$$

where $G_{i_j i_k} = (X_i - X_{i_j})^T (X_i - X_{i_j})$.

Step 4 Computing the interaction probabilities for unobserved lncRNA–protein pairs:

$$Y = (1-\alpha)(I - \alpha W)^{-1} Y^0 \tag{6}$$

The details are shown in **Figure 2**.

LPI-BNPRA

Zhao et al. (2018a) developed a novel LPI prediction model based on a bipartite network projection recommended technique, LPI-BNPRA. LPI-BNPRA can be broken down into five steps.

Step 1 Extracting 4,158 high-confidence LPIs between 990 lncRNAs and 27 proteins from NPInter (Hao et al., 2016) and NONCODE (Zhao et al., 2015) by filtering unreliable lncRNA sequences and removing lncRNAs/proteins only associated with one protein/lncRNA.

Step 2 Calculating lncRNA–lncRNA similarity based on the Smith–Waterman technique:

$$LSM(l_i, l_j) = \frac{sw(l_i, l_j)}{\max\ (sw(l_i, l_i), sw(l_j, l_j))} \tag{7}$$

where $sw(l_i, l_j)$ denotes the Smith–Waterman score between two lncRNAs l_i and l_j.

Step 3 Calculating the protein–protein similarity matrix based on the Smith–Waterman technique:

$$PSM(p_i, p_j) = \frac{sw(p_i, p_j)}{\max\left(sw(p_i, p_i), sw(p_j, p_j)\right)} \tag{8}$$

where $sw(p_i, p_j)$ denotes the Smith–Waterman score between two proteins p_i and p_j.

Step 4 For a given lncRNA l_j, computing its bias ratings of lncRNAs for a protein p_i with the agglomerative hierarchical clustering and associated measurement of minimum variance method:

$$r(p_i, l_j) = \frac{n_{cr}}{T(p_i)} \tag{9}$$

where n_{cr} is the number of lncRNAs in the cluster *cr* including l_j, and Tp_i is the number of all lncRNAs interacting with p_i.

Step 5 Finding LPI candidates based on the recommended bipartite network projection technique and bias ratings of every lncRNA for proteins:

$$R_{fin}(l_j) = \sum_{i=1}^{n} R_{fin}(p_i, l_j) \tag{10}$$

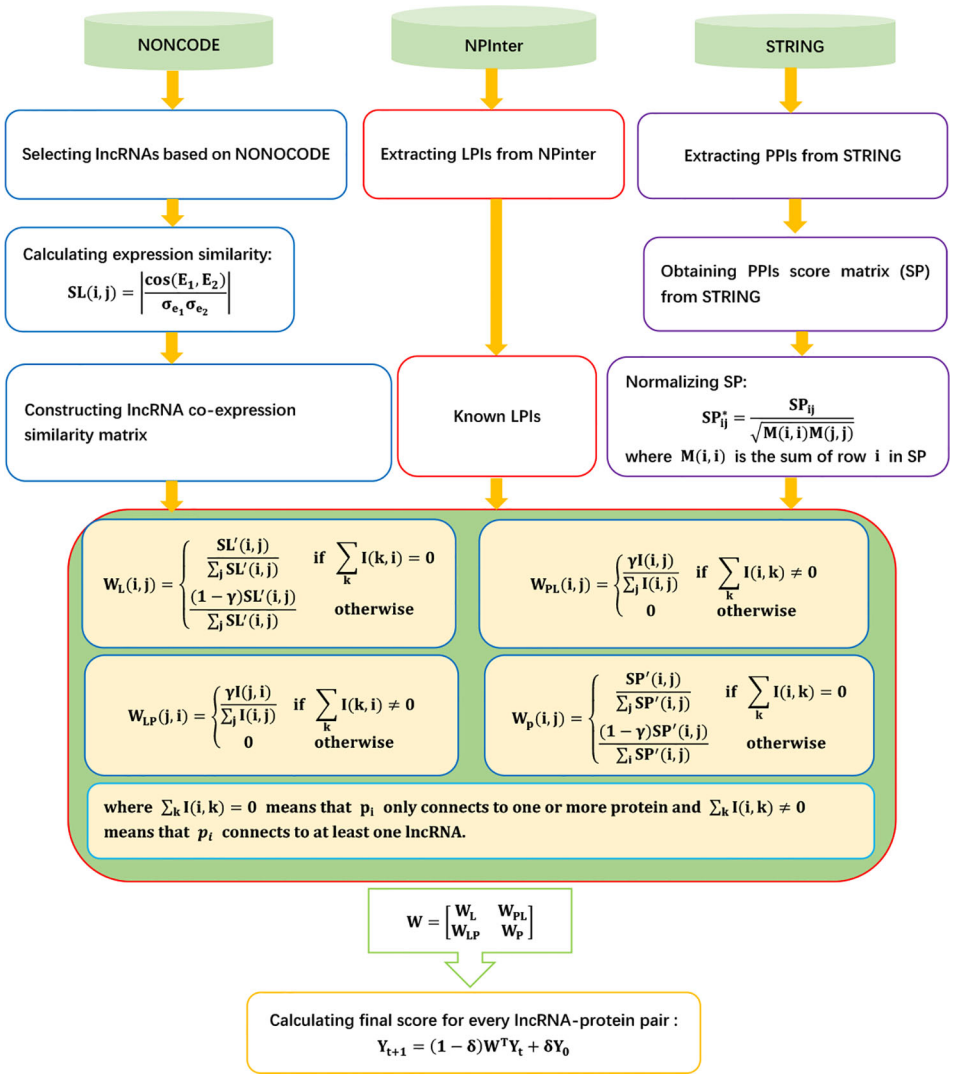


FIGURE 1 | Flowchart of LPI prediction method based on heterogeneous network model and random walk with restart.

where

$$R_{fin}(p_i,l_j)=\frac{r(p_i,l_j)}{\sum_{k=1}^{n}r(p_k,l_j)}\times R(p_i) \quad (11)$$

$$R(p_i)=\sum_{j=1}^{m}R(p_i,l_j) \quad (12)$$

$$R(p_i,l_j)=\frac{r_{ini}(p_i,l_j)}{\sum_{k=1}^{n}r_{ini}(p_k,l_j)}\times R_{ini}(l_j) \quad (13)$$

$$R_{ini}(l_j)=r_{ini}(p_i,l_j) \quad (14)$$

$$r_{ini}(p_i,l_j)=\frac{r(p_i,l_j)}{r_{ave}(p_i,l_j)} \quad (15)$$

$$r(p_i,l_j)=\frac{n_{cr}}{T(p_i)} \quad (16)$$

$$r_{ave}(p_i,l_j)=\frac{\sum_{j=1}^{m}r(p_i,l_j)}{T(p_i)} \quad (17)$$

The details are shown in **Figure 3**.

LPISNFHS

Zheng et al. (2017) presented a new LPI identification method, LPISNFH. LPISNFHS fused multiple protein–protein similarity networks, the similarity network fusion (SNF)

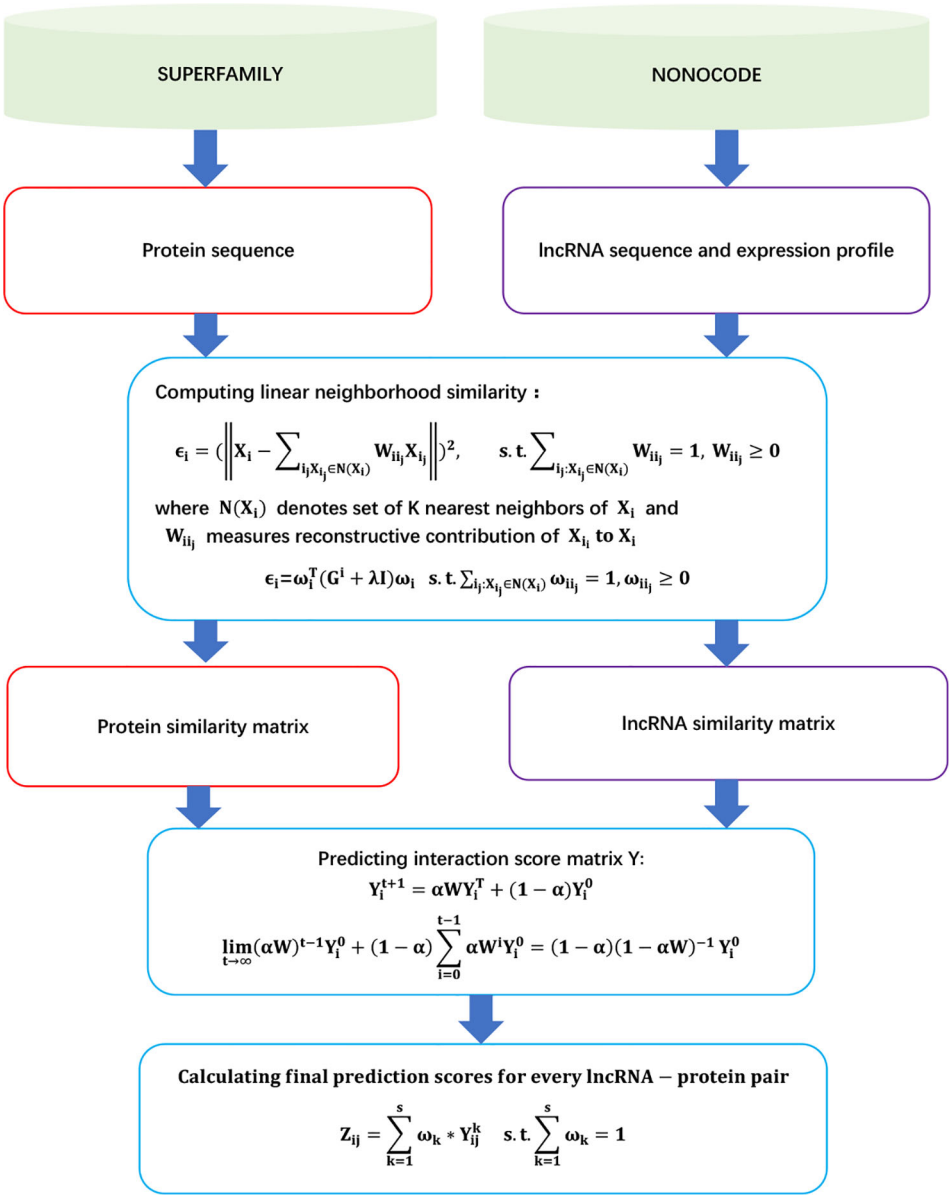


FIGURE 2 | Flowchart of linear neighborhood propagation-based LPI prediction method.

technique, HeteSim algorithm, and known LPI network into a unified framework. LPISNFH can be broken down into three steps.

Step 1 Obtaining 4,467 LPIs between 1,050 unique lncRNAs and 84 unique proteins from NPInter (Hao et al., 2016) and NONCODE (Zhao et al., 2015) by manually filtering LPIs not involving lncRNAs and removing the lncRNAs only associated with one protein.

Step 2 Constructing a protein–protein similarity network. LPISNFHS fused the sequence similarity, functional annotation semantic similarity (Go), domain similarity, and STRING similarity into a unified protein–protein similarity network based on the SNF technique.

Step 3 Inferring novel LPIs by combining the HeteSim algorithm and heterogeneous LPI network.

LPI-IBNRA

Xie et al. (2019) developed a LPI prediction model, LPI-IBNRA. LPI-IBNRA integrated lncRNA–protein interactions, protein–protein interactions, and similarity matrix for proteins and lncRNAs, and improved bipartite network recommender algorithm. LPI-IBNRA can be broken down into seven steps.

Step 1 Obtaining 4,796 LPIs between 1,105 lncRNAs and 26 proteins from NPInter (Hao et al., 2016) and NONCODE (Zhao et al., 2015) after filtering lncRNAs and proteins that have only one association.

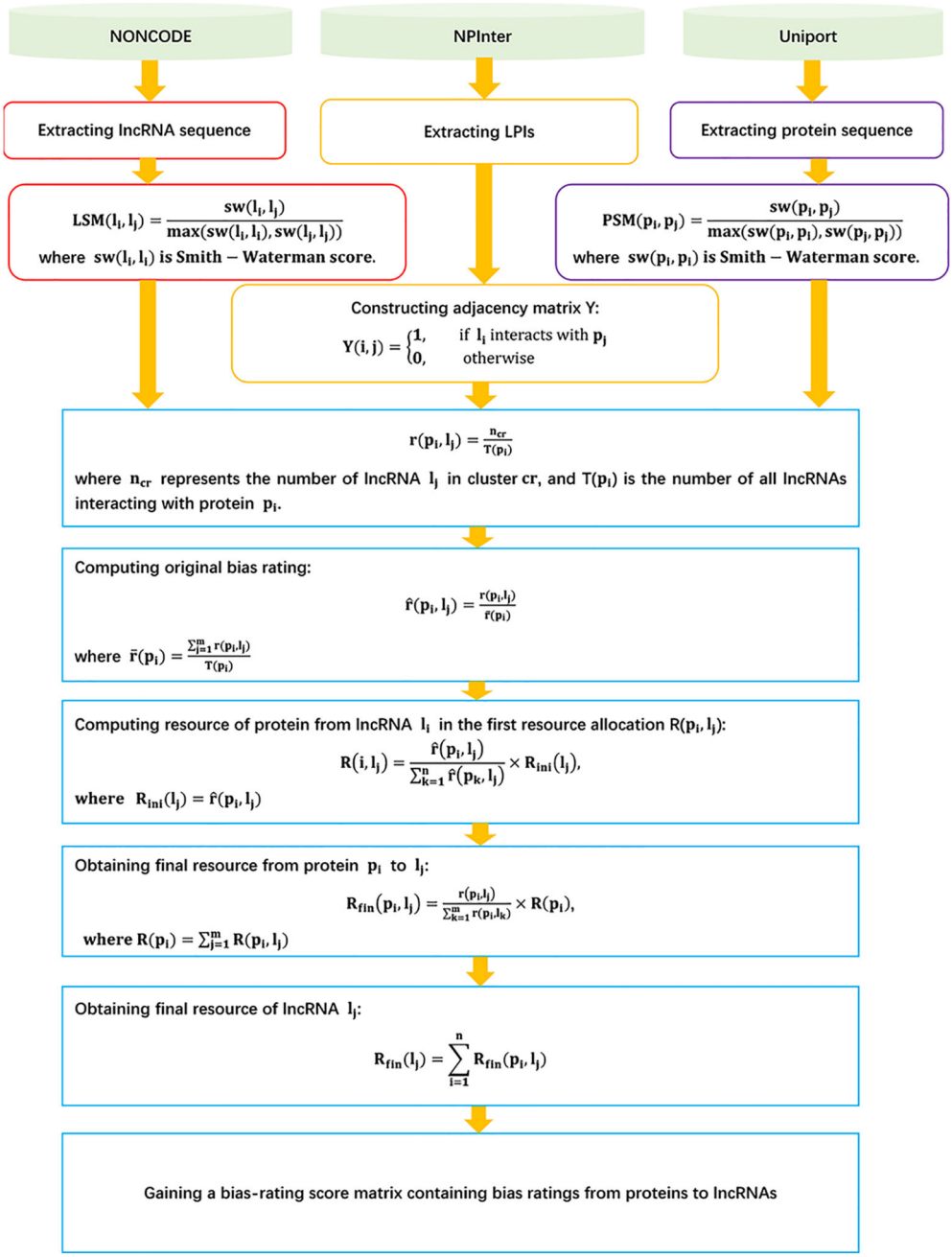


FIGURE 3 | Flowchart of LPI prediction model based on the recommended bipartite network projection technique.

Step 2 Computing lncRNA similarity matrix sim^L based on lncRNA expression similarity and Gaussian interaction profile (GIP) kernel similarity, and protein similarity matrix sim^P based on protein interaction similarity and GIP kernel similarity.

Step 3 Computing the score between protein p_i and lncRNA l_j based on protein similarity and lncRNA similarity by Eqs. (18) and (19), respectively.

$$S^P(p_i, l_j) = \begin{cases} \dfrac{\sum_{k=1}^{np} sim^P(p_i, p_k) I(p_k, l_j)}{\sum_{k=1}^{np} sim^P(p_i, p_k)} & if \quad I(p_i, l_j) = 1 \\ 0 & otherwise \end{cases} \tag{18}$$

$$S^L(p_i,l_j)=\begin{cases}\dfrac{\sum_{k=1}^{nl} I_{(p_i,l_k)} sim^L(l_k,l_j)}{\sum_{k=1}^{np} sim^L(l_k,l_j)} & if \quad I(p_i,l_j)=1\\ 0 & otherwise\end{cases} \tag{19}$$

Step 4 Obtaining the initialized association score matrix as follows:

$$S_{ini}=\gamma S^P+(1-\gamma)S^L \tag{20}$$

Step 5 Computing the first-round scores of the lncRNA l_k over all proteins:

$$s_1(l_k)=\sum_{j=1}^{np}\frac{S_{ini}(p_j,l_k)s_0(p_j)}{d(p_j)} \tag{21}$$

Step 6 Computing the second-round scores of the protein p_i over all lncRNAs:

$$s_2(p_i)=\sum_{k=1}^{nl}\frac{S_{ini}(p_i,l_k)}{d(l_k)}\sum_{k=1}^{np}\frac{S_{ini}(p_j,l_k)s_0(p_j)}{d(p_j)} \tag{22}$$

Step 7 Computing the final association score matrix:

$$S'_{fin}=W'S_{ini} \tag{23}$$

where $W'=W+\alpha W^2$ and $\alpha\in(-1,0)$.

The details are shown in **Figure 4**.

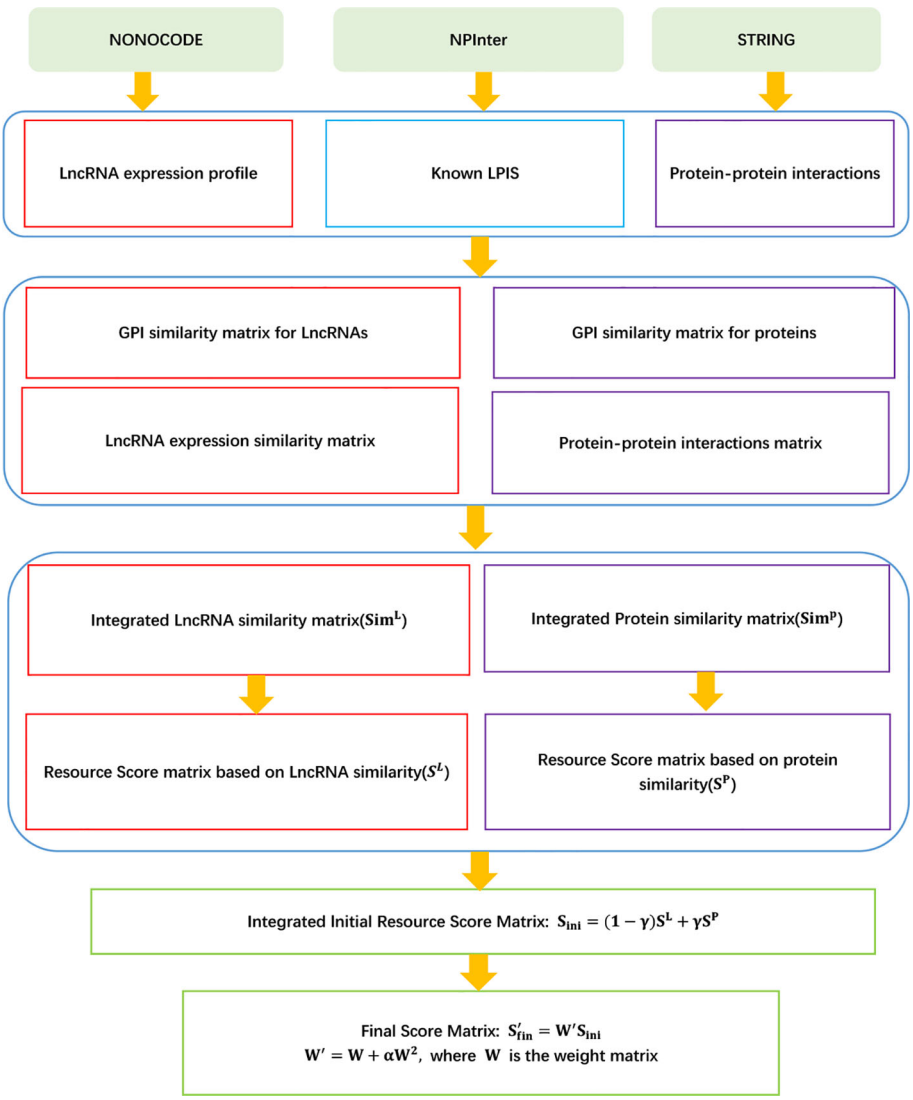


FIGURE 4 | Flowchart of LPI prediction method based on improved bipartite network recommender algorithm.

LPBNI

Ge et al. (2016) proposed an lncRNA–protein bipartite network inference method, LPBNI, to find potential LPIs. LPBNI can be broken down into five steps.

Step 1 Extracting data. LPBNI first downloads 7,576 ncRNA–protein associations from NPInter 2.0 (Hao et al., 2016) with the restricted type of "NONCODE" and organism "*Homo sapiens*." LPBNI then selects 2,380 lncRNAs based on a human lncRNA dataset provided by the NONCODE database (Zhao et al., 2015). Finally, LPBNI extracts 4,870 LPIs between 2,380 lncRNAs and 106 proteins.

Step 2 Utilizing the LPI network to construct a bipartite graph G (L, P, Y).

Step 3 Propagating known biological information in G. For a lncRNA l_j, S_L (l_j) denotes the score on l_j after the first step of propagation:

$$S_L(l_j) = \sum_{i=1}^{m} \frac{a_{ij} S_0(i)}{d(p_i)}, j \in \{1, 2, 3 \ldots . n\} \tag{24}$$

where S_0 (i) = s_{ij}, $i \in \{i, 2, \ldots, m\}$ denotes the original information of P for a given lncRNA l_j. s_{ij} = 1 if p_i associates with l_j; otherwise, $s_{ij} = 0$. $d(p_i) = \sum_{j=1}^{n} a_{ij}$ denotes the number of lncRNAs associated with p_i.

Step 4 Propagating all information in L back to P. $S_F(p_i)$ represents the final information on protein p_i to denote the associated score between p_i and l_j:

$$S_F(i) = \sum_{j=1}^{n} \frac{a_{ij} S_L(l_j)}{d(l_j)} = \sum_{j=1}^{n} \frac{a_{ij}}{d(l_j)} \sum_{k=1}^{m} \frac{a_{kj} S_0(k)}{d(p_k)} \tag{25}$$

where $d(l_i) = \sum_{i=0}^{m} a_{ij}$ is the number of proteins interacting with l_j.

Step 5 Computing the final associated score S_F after the above two-step information propagation yields

$$\vec{S}_F = W\overrightarrow{S_0} \tag{26}$$

where $\vec{S}_0$ denotes the column vector of S_0, $S_F(i) = \sum_{k=1}^{m} w_{ik} S_0(k)$, where $wij = \frac{1}{d(pi)} \sum_{j=1}^{n} \frac{a_{ij} a_{kj}}{d(lj)}$.

The details are shown in **Figure 5**.

ACCBN

Zhu et al. (2019) exploited an ant-colony-clustering-based bipartite network method for revealing potential LPIs, ACCBN. The model can be roughly broken down into three steps.

Step 1 Describing lncRNA interaction profiles and protein interaction profiles as row vectors and column vectors based on the LPI network, respectively.

Step 2 Calculating the probability that two entities x_i and x_j belong to the same cluster based on the ant colony clustering method:

$$p_{ij}(t) = \frac{\left[T_{ij}(t)\right]^{\alpha} \left[\eta_{ij}(t)\right]^{\beta}}{\sum_{j=1}^{k} \left[T_{ij}(t)\right]^{\alpha} \left[\eta_{ij}(t)\right]^{\beta}} \tag{27}$$

where

$$\eta_{ij} = \frac{1}{d_{ij}} \tag{28}$$

$$d_{ij} = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^2\right)^{\frac{1}{2}} \tag{29}$$

$$T_{ij}(t+1) = (1-\rho) T_{ij}(t) + \Delta T_{ij}(t) \tag{30}$$

$$T_{ij}(t) = \begin{cases} 1 & d_{ij} \le r \\ 0 & d_{ij} > r \end{cases} \tag{31}$$

$$\Delta T_{ij}(t) = \frac{Q}{d(x_i, c_j)} \tag{32}$$

where r is the cluster radius, c_j is the cluster center of the jth cluster, and $\alpha \in (0, 5)$, $\beta \in (0, 5)$, $\rho \in (0.1, 0.99)$, and $Q \in (1, 10000)$.

Step 3 Applying lncRNA–protein bipartite network to identify LPI candidates. Given a protein p_k, its association scores with all lncRNAs at the tth iteration P_k^t can be computed as follows:

$$P_k^t = \rho W P_k^{t-1} + (1-\rho) M(:, k) \tag{33}$$

where W is a similarity matrix.

The association scores for all proteins $\{p_1, p_2, \ldots, pm\}$ can be represented as follows:

$$P^t = \rho W P^{t-1} + (1-\rho) M \tag{34}$$

Machine Learning-Based Methods

Machine learning-based LPI prediction methods utilize machine learning-based models and algorithms to uncover potential LPIs. This type of method can be roughly classified into two categories: matrix factorization-based methods and ensemble learning-based methods.

Matrix Factorization-Based Models

Matrix factorization is exploited in recommendation systems and has been widely applied to bioinformatics (Shi et al., 2018; Zhang et al., 2018a; Zhao et al., 2018b; Cantini et al., 2019). Matrix factorization-based LPI prediction techniques transformed the problem of LPI identification into a recommender task, and adopted the matrix factorization model to capture unobserved LPIs. Given an LPI matrix Y and two nonnegative matrices $W \in \mathfrak{R}^{k \times n}$ and $H \in \mathfrak{R}^{k \times m}$ the problem of predicting LPIs can be formulated as the following objective function:

$$\min_{W,H} \| Y - W^T H \|_F^2 \quad \text{s.t.} \quad W \ge 0, H \ge 0 \tag{35}$$

A few LPI identification methods have been designed based on matrix factorization method.

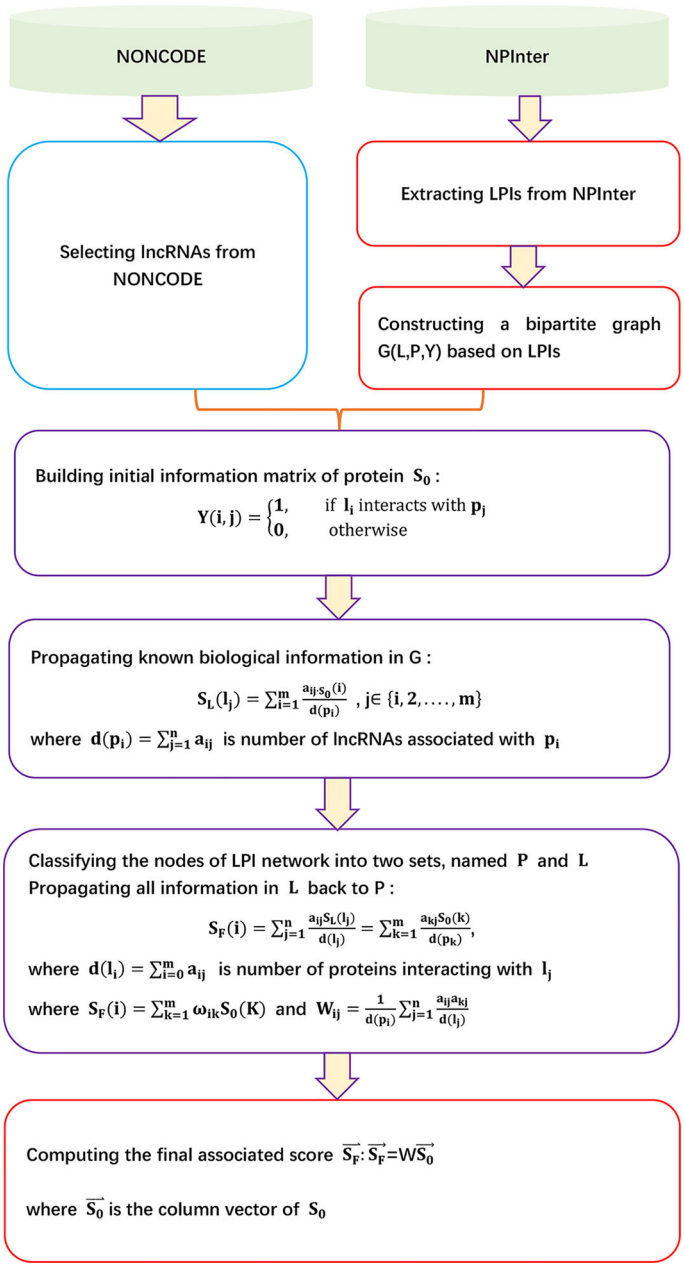


FIGURE 5 | Flowchart of lncRNA–protein bipartite network inference method.

LPGNMF

Zhang et al. (2018a) designed a graph regularized nonnegative matrix factorization-based (NMF) method to predict potential LPIs, LPGNMF. LPGNMF consists of three steps.

Step 1 Extracting LPI information based on data provided by NONCODE (Zhao et al., 2015), NPInter (Hao et al., 2016), and UniProt (Consortium et al., 2018). Obtaining 9,484 LPIs between 50 proteins and 2,190 lncRNAs after filtering and removing lncRNAs/proteins only interacting with one protein/lncRNA.

Step 2 Computing lncRNA similarity and protein similarity.

LPGNMF computes the lncRNA expression profile similarity $S^l(i, j)$:

Given the expression profiles of two lncRNAs E_1 and E_2, LPIHN calculates lncRNA expression similarity based on the Pearson correlation coefficient:

$$S^l(i,j)=\left|\frac{\mathrm{cov}(E_1,E_2)}{\sigma_{e_1}\sigma_{e_2}}\right| \tag{36}$$

where $cov(E_1, E_2)$ is the covariance of E_1 and E_2, and σ_{e_1} and σ_{e_2} are the standard deviations of E_1 and E_2, respectively.

LPGNMF computes the weight matrix based on lncRNA similarity:

$$M_{ij}^{l} = \begin{cases} 1 & i \in N(l_j) \quad \& \quad j \in N(l_i) \\ 0 & i \notin N(l_j) \quad \& \quad j \notin N(l_i) \\ & 0.5 \quad otherwise \end{cases} \tag{37}$$

Here, $N(l_i)$ and $N(l_j)$ denote the p nearest neighbors of l_i and l_j.

LPGNMF then calculates the sparse similarity matrix of lncRNAs $S^{l\star}$:

$$S_{ij}^{l\star} = M_{ij}^{l} S_{ij}^{l} \tag{38}$$

Similarly, LPGNMF calculates the sparse similarity matrix of proteins $S^{p\star}$.

Step 3 Building the following optimization model based on the graph regularized nonnegative matrix factorization method:

$$\begin{aligned} &\min_{W,H} \| Y - W^T H \|_F^2 + \lambda_p \sum_{i,j=1}^{n} \| w_i - w_j \|^2 S_{ij}^{p\star} \\ &+ \lambda_l \sum_{i,j=1}^{m} \| h_i - h_j \|^2 S_{ij}^{l\star} + \beta_1 \sum_{i,j=1}^{n} \| W(:,i) \|_1^2 \\ &+ \beta_2 \sum_{i,j=1}^{m} \| H(:,i) \|_1^2 s.t. \quad W \\ &\geq 0, H \geq 0 \end{aligned} \tag{39}$$

The details are shown in **Figure 6**.

LPI-NRLMF

Liu et al. (2017) designed a novel LPI identification model based on neighborhood regularized logistic matrix factorization, LPI-NRLMF. LPI-NRLMF can be roughly broken down into three steps.

Step 1 Extracting the lncRNA sequence, protein sequence, and LPIs based on data provided by NONCODE (Zhao et al., 2015), NPInter (Hao et al., 2016), and UniProt (Consortium et al., 2018); and obtaining 4,158 LPIs between 27 proteins and 990 lncRNAs.

Step 2 Computing lncRNA sequence similarity matrix *LSM* and protein sequence similarity matrix *PSM* based on the Smith–Waterman algorithm:

$$LSM(l_i, l_j) = \frac{sw(l_i, l_j)}{max \ (sw(l_i, l_i), sw(l_j, l_j))} \tag{40}$$

$$PSM(p_i, p_j) = \frac{sw(p_i, p_j)}{max \ (sw(p_i, p_i), sw(p_j, p_j))} \tag{41}$$

Step 3 Defining neighborhood information for lncRNAs and obtaining the adjacency matrix A of lncRNAs:

$$a_{iu} = \begin{cases} s_{iu}^{l} & if \quad l_u \in N(l_i) \\ 0 & otherwise \end{cases} \tag{42}$$

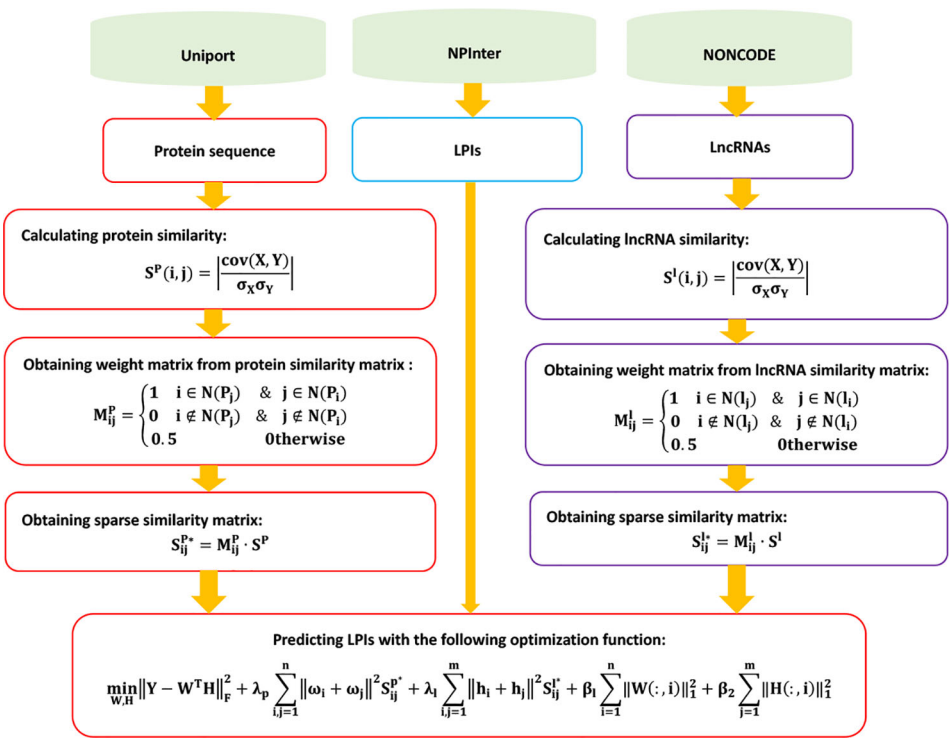


FIGURE 6 | Flowchart of LPI prediction method based on graph regularized nonnegative matrix factorization.

Similarly, LPI-NRLMF computes the adjacency matrix B of proteins.

Step 4 Computing associated scores S_N for unknown lncRNA–protein pairs based on the neighborhood regularized logistic matrix factorization model:

$$p_{ij} = \frac{exp\ (u_i v_j^T)}{1 + exp\ (u_i v_j^T)} \tag{43}$$

Here, $u_i \in \Re^{1xr}$ and $v_j \in \Re^{1xr}$ can be computed by the following neighborhood regularized logistic matrix factorization model:

$$\min_{U,V} \sum_{i=1}^{m}\sum_{j=1}^{n}(1 + cy_{ij} - y_{ij}) \ln[1 + \exp(u_i v_j^T)] - cy_{ij} u_i v_j^T + \frac{1}{2} tr[U^T(\lambda_l I + \alpha L^l)U] + \frac{1}{2} tr[V^T(\lambda_p I + \beta L^p)V] \tag{44}$$

where $L^l = (D_i^l + D_u^l) - (A + A^T)$, $D_i^l = \sum_{u=1}^{m} a_{iu}, D_u^l = \sum_{i=1}^{m} a_{iu}$. Similarly, L_P can be computed. $U \in \Re^{m\ x\ r}$ and $V \in \Re^{1\ x\ r}$ can be calculated by dividing L.

The details are shown in **Figure 7**.

IRWNRLPI

Zhao et al. (2018b) fused the random walk into LPI-NRLMF and exploited a novel LPI prediction model based on LPI-NRLMF, IRWNRLPI. IRWNRLPI is a semi-supervised learning-based model and does not require negative samples. IRWNRLPI contains the following five steps.

Step 1 Extracting the lncRNA sequence, protein sequence, and LPIs from NONCODE (Zhao et al., 2015), NPInter (Hao et al., 2016), and UniProt (Consortium et al., 2018); and obtaining 4,158 LPIs between 27 proteins and 990 lncRNAs.

Step 2 Computing the lncRNA sequence similarity matrix LS and protein sequence similarity matrix PS based on the Smith–Waterman algorithm:

$$LS(l_i, l_j) = \frac{sw(l_i, l_j)}{max\ (sw(l_i, l_i), sw(l_j, l_j))} \tag{45}$$

$$PS(p_i, p_j) = \frac{sw(p_i, p_j)}{max\ (sw(p_i, p_i), sw(p_j, p_j))} \tag{46}$$

Step 3 Building a random walk model to compute associated scores S_R for unknown lncRNA–protein pairs:

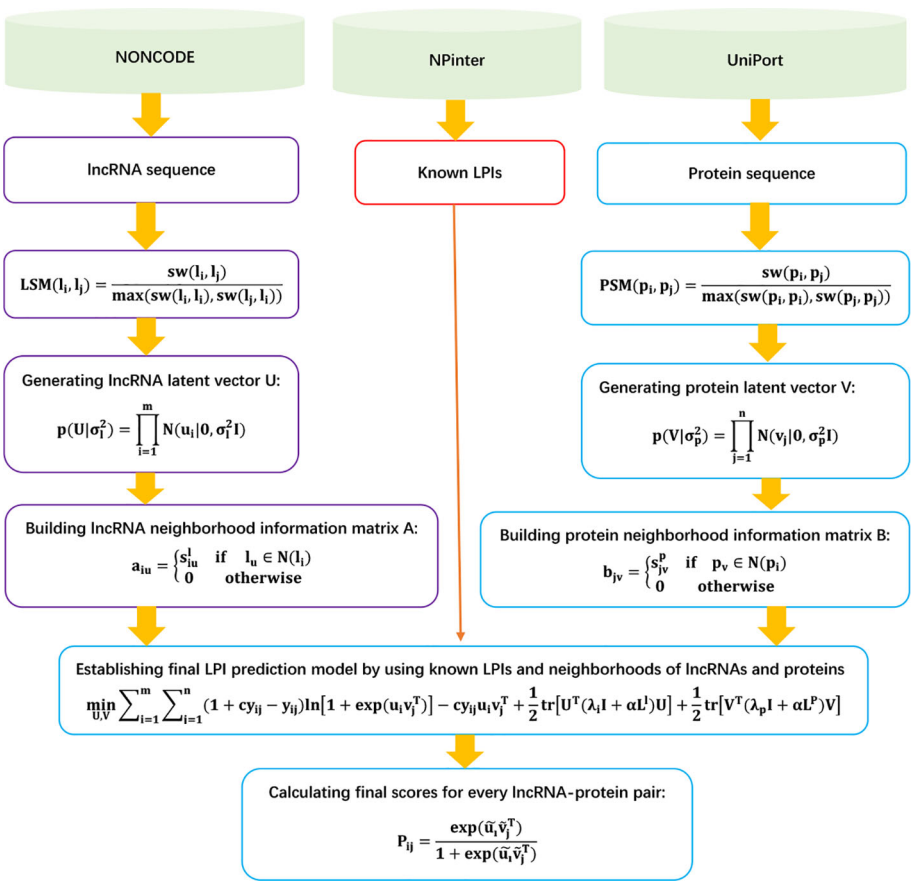


FIGURE 7 | Flowchart of LPI prediction model based on neighborhood regularized logistic matrix factorization.

$$S(t+1) = r_Q L_Q^T S(t) + p_Q(1-r_Q)X + r_U L_U^T S(t) + p_U(1-r_U)X \tag{47}$$

where r_{ij} represents the extent of association between a neighbor v_j and a protein p for a given node v_i. $L(l_{ij})_{M\ x\ M}$ is computed by $l_{ij} = r_{ij}/\sum_{j=1}^{N} r_{ij}$. IRWNRLPI divides L into two arrays of L_U and L_Q.

Step 4 Computing associated scores S_N for unknown lncRNA–protein pairs based on the neighborhood regularized logistic matrix factorization model:

$$p_{ij} = \frac{exp\ (u_i v_j^T)}{1 + exp\ (u_i v_j^T)} \tag{48}$$

$u_i \in \Re^{1\ x\ r}$ and $v_j \in \Re^{1\ x\ r}$ can be computed by the following neighborhood regularized logistic matrix factorization model:

$$\min_{U,V} \sum_{i=1}^{m}\sum_{j=1}^{n}(1 + cy_{ij} - y_{ij}) \ln[1 + \exp(u_i v_j^T)] - cy_{ij} u_i v_j^T + \frac{1}{2} tr[U^T(\lambda_l I + \alpha L^l)U] + \frac{1}{2} tr[V^T(\lambda_p I + \beta L^p)V] \tag{49}$$

where $U \in \Re^{m\ x\ r}$ and $V \in \Re^{n\ x\ r}$.

Step 5 Computing the final associated scores for unknown lncRNA–protein pairs:

$$S = \frac{S_R + S_N}{2} \tag{50}$$

The details are shown in **Figure 8**.

LPI-KTASLP

Shen et al. (2019) designed a kernel target alignment-based semi-supervised model, LPI-KTASLP, to find novel LPIs. LPI-KTASLP utilizes matrix factorization and an approximation technique. LPI-KTASLP can be roughly broken down into three steps.

Step 1 Computing lncRNA kernels and protein kernels from four levels.

Level 1 GIP kernel:

The GIP kernels between two lncRNAs and two proteins are defined as follows, respectively:

$$K_{GIP}^{lnc}(l_i, l_j) = \exp(-\gamma_{lnc} \| Y_{l_i} - Y_{l_j} \|^2) \tag{51}$$

$$K_{GIP}^{pro}(p_i, p_j) = \exp(-\gamma_{pro} \| Y_{p_i} - Y_{p_j} \|^2) \tag{52}$$

Level 2 Sequence kernel:

The sequence kernels of two lncRNAs and two proteins are defined as follows, respectively:

$$K_{SW}^{lnc}(l_i, l_k) = \frac{SW(S_{l_i}, S_{l_k})}{\sqrt{SW(S_{l_i}, S_{l_i})}\sqrt{SW(S_{l_k}, S_{l_k})}} \tag{53}$$

$$K_{SW}^{pro}(P_i, P_k) = \frac{SW(S_{p_i}, S_{p_k})}{\sqrt{SW(S_{p_i}, S_{p_i})}\sqrt{SW(S_{p_k}, S_{p_k})}} \tag{54}$$

where $SW(.,.)$ is the Smith–Waterman score, and S represents the sequence information of a lncRNA/protein.

Level 3 Sequence feature kernel:

Constructing radial basis function kernels K_{SF}^{lnc} and K_{SF}^{pro} for lncRNAs and proteins based on the conjoint triad and pseudo position-specific score matrix, respectively.

Level 4 lncRNA expression kernel:

Calculating the expression kernel of lncRNA K_{EXP}^{lnc} based on the expression profiles of lncRNAs provided by the NONCODE database (Zhao et al., 2015).

Step 2 Fusing the above kernels to generate the optimal kernel based on kernel target alignment:

$$K_{lnc}^{\star} = \sum_{a=1}^{4} w_a^{lnc} K_a^{lnc}, K_a^{lnc} \in \Re^{n\times n} \tag{55}$$

$$K_{pro}^{\star} = \sum_{a=1}^{3} w_a^{pro} K_a^{pro}, K_a^{pro} \in \Re^{m\times m} \tag{56}$$

Step 3 Constructing the following model to compute interaction probabilities for unobserved lncRNA–protein pairs based on matrix factorization, low-rank approximation, and eigen decomposition:

$$Y^{\star} = \frac{1}{1+3\delta} Y + \frac{1}{1+3\delta^2} V_{lnc}\left(D \odot \left(V_{lnc}^T F V_{pro}\right)\right) V_{pro}^T \tag{57}$$

The details are shown in **Figure 9**.

Ensemble-Based Methods

Ensemble learning methods are widely applied to LPI prediction. HLPI-Ensemble (Hu et al., 2018) and SFPEL-LPI (Zhang et al., 2018c) are two state-of-the-art ensemble-based LPI prediction methods.

HLPI-Ensemble

Hu et al. (2018) developed the HLPI-Ensemble method for human LPI identification. HLPI-Ensemble consists of two major processes: benchmark dataset construction and HLPI-Ensemble model construction.

In the first process, HLPI-Ensemble downloads lncRNA sequences, protein sequences, and LPIs from NONCODE (Zhao et al., 2015), UniProt (Consortium et al., 2018), and NPinter (Hao et al., 2016). HLPI-Ensemble then extracts 82 features of lncRNAs and 1,516 features of proteins based on Kmer, DAC, and PC-PseDNC-General.

In the second process, HLPI-Ensemble utilizes the ensemble technique and generates three ensemble learning frameworks, HLPI-SVM, HLPI-XGB, and HLPI-RF. These three frameworks are based on support vector machines (SVMs), extreme gradient boosting (XGB), and random forests (RFs), respectively. The details are shown in **Figure 10**.

SFPEL-LPI

Zhang et al. (2018c) exploited a sequence-based feature projection ensemble learning framework, SFPEL-LPI, to uncover novel LPIs. SFPEL-LPI integrated $\ell_{1,2}$-norm regularization, ensemble graph Laplacian regularization, and

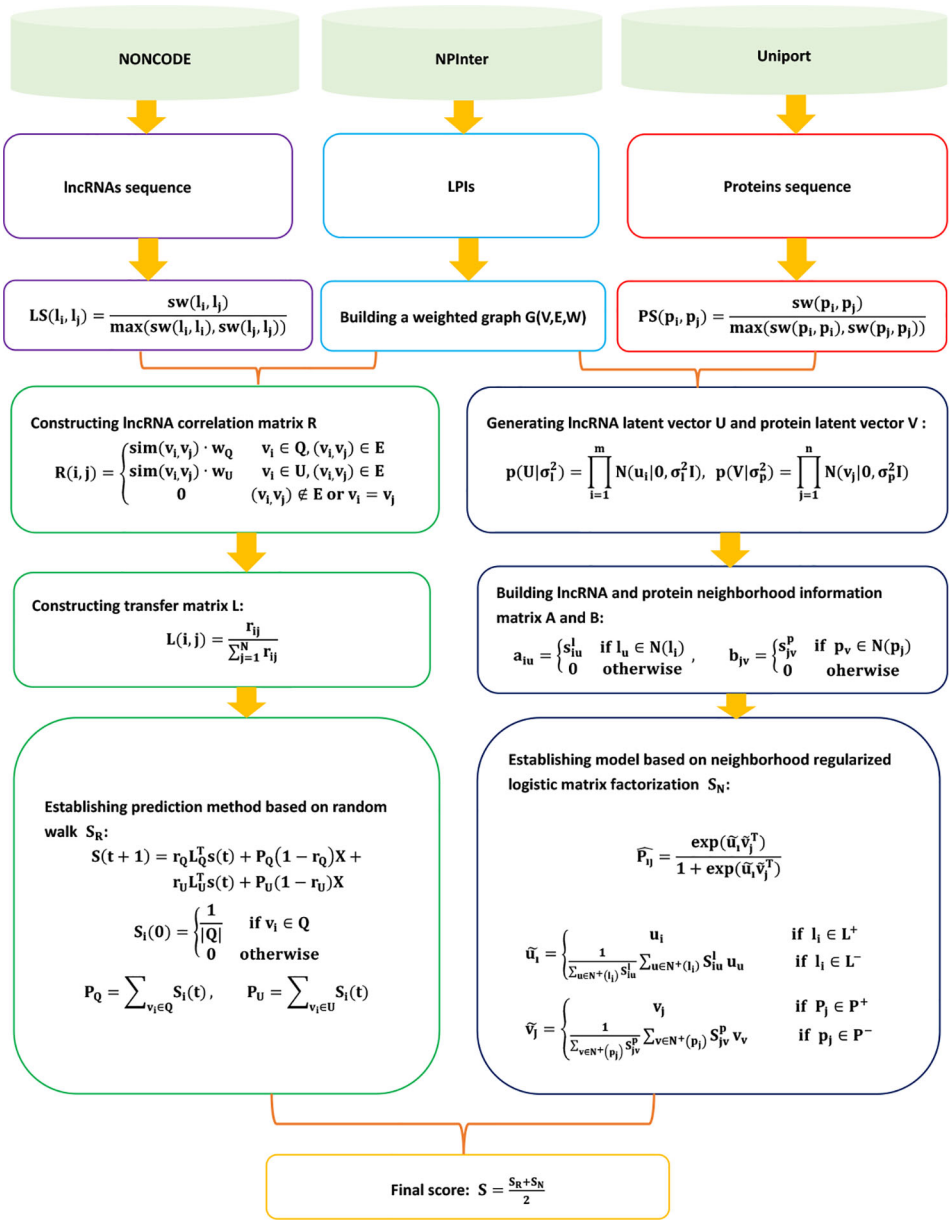


FIGURE 8 | Flowchart of LPI prediction model based on the random walk and neighborhood regularized logistic matrix factorization.

various biological information into a unified framework. It can be roughly broken down into five steps.

Step 1 Downloading LPIs, lncRNA sequences, and protein sequences from NPInter (Hao et al., 2016), NONCODE (Zhao et al., 2015), and SUMPERFAMILY (Pandurangan et al., 2018), respectively.

Step 2 Describing lncRNA and protein features based on sequence information and known LPIs.

SFPEL-LPI describes lncRNA features based on parallel correlation pseudo dinucleotide composition (PSEDNC). Given the occurrence frequency of different dinucleotides and the physicochemical properties of every dinucleotide, the PseDNC feature vector for an RNA sequence L can be represented as

$$L = [d_1, d_2, \ldots\ldots d_{16}, d_{16+1}, \ldots d_{16+\tau}] \tag{58}$$

where

$$d_k = \begin{cases} \frac{f_k}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\tau} \theta_j} & 1 \le k \le 16 \\ \frac{w\theta_{k-16}}{\sum_{i=1}^{16} f_i + w \sum_{j=1}^{\tau} \theta_j} & 17 \le k \le 16 + \tau \end{cases} \tag{59}$$

In addition, SFPEL-LPI represents the interaction profile of an lncRNA as a row vector of the LPI matrix Y: $IP_{L_i} = Y(i,:)$.

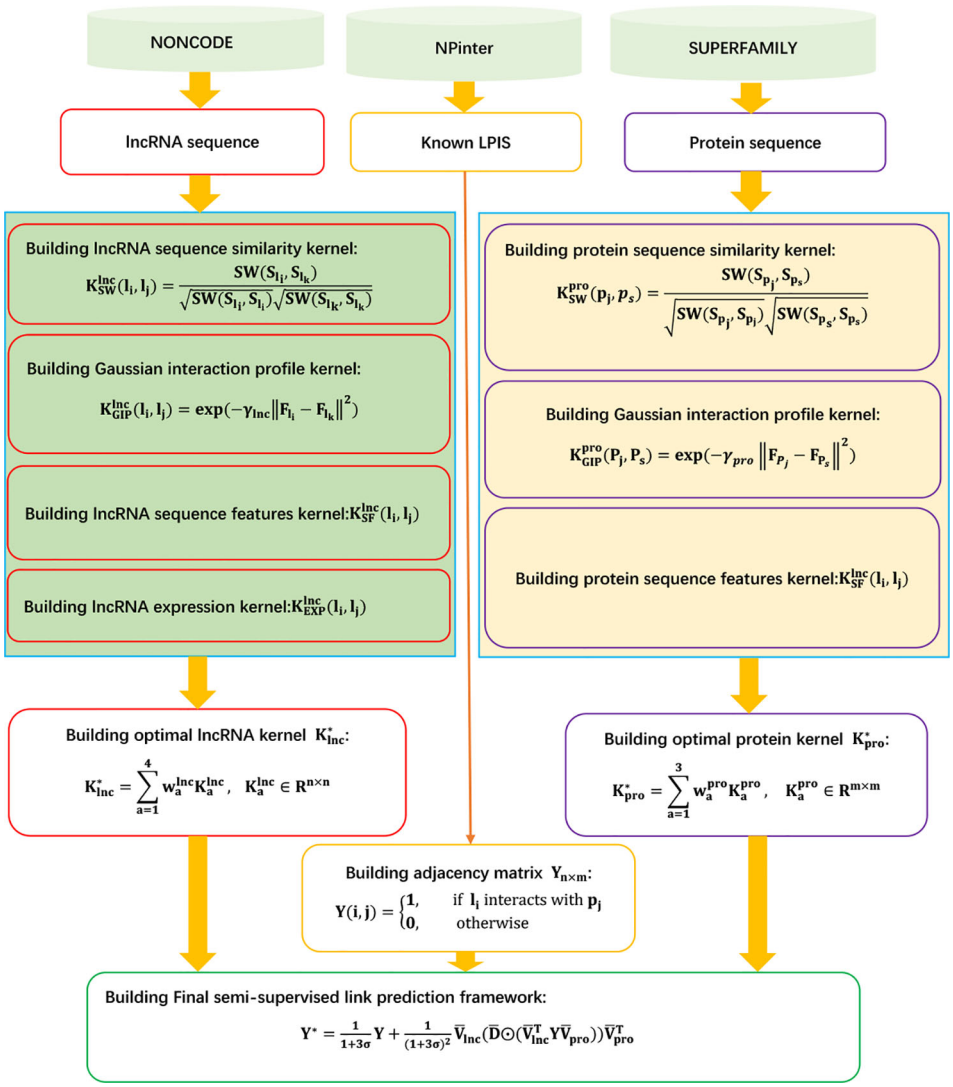


FIGURE 9 | Flowchart of kernel target alignment-based semi-supervised model for LPI prediction.

SFPEL-LPI describes protein features based on the parallel correlation pseudo amino acid composition (PseAAC):

$$P = [c_1, c_2, \ldots\ldots c_{20}, c_{20+1}, \ldots, c_{20+\tau}] \tag{60}$$

where

$$c_k = \begin{cases} \frac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\tau}\theta_j} & 1 \le k \le 20 \\ \frac{w\theta_{k-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{\tau}\theta_j} & 20 \le k \le 20+\tau \end{cases} \tag{61}$$

Similarly, the interaction profile of a protein can be defined as a column vector of the LPI matrix Y: $IP_{p_i} = Y(:, i)$.

Therefore, a features for lncRNAs/proteins can be represented as feature matrix: $\{X_i\}_{i=1}^{a}$.

Step 4 Computing lncRNA similarity and protein similarity.

SFPEL-LPI first computes the linear neighborhood similarity of lncRNAs based on PseDNC and IP.

SFPEL-LPI then computes the Smith–Waterman subgraph similarity (SWSS) of lncRNAs:

$$SWSS(L_i, L_j) = \sum_{P_{01}\in A(L_i)}\sum_{P_{02}\in A(L_j)} \frac{SW(P_{o1}, P_{o2})}{n1 \times n2} \tag{62}$$

Similarly, the PseAAC similarity, IP similarity, and SWSS similarity of proteins can be computed.

Therefore, b types of similarities of lncRNAs/proteins can be represented as b similarity matrices $\{W_i\}_{i=1}^{b}$.

Step 5 Computing the association scores for novel lncRNAs/proteins based on Eqs. (63) and (64).

$$R_l = \sum_{i=1}^{u} \theta_{li} X_{li} G_{li}^T \tag{63}$$

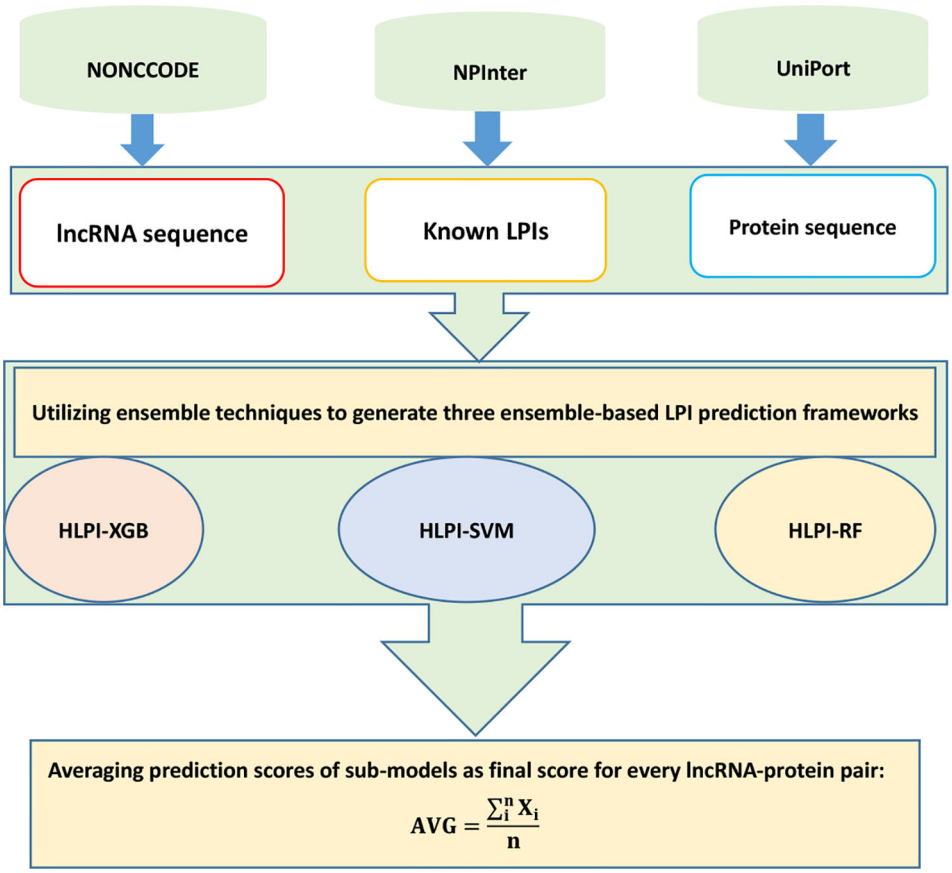


FIGURE 10 | Flowchart of ensemble-based LPI identification method.

$$R_p = \sum_{i=1}^{v} \theta_{pi} X_{pi} G_{pi}^T \tag{64}$$

G_i, R, and θ can be obtained by solving the following optimization model:

$$\begin{aligned} &\min_{G_i,R,\theta} \| R - Y \|_F^2 + \mu \sum_{i=1}^{a} \| X_i G_i^T - R \|_F^2 + \sum_{i=1}^{b} \theta_i^{\eta} tr(R^T (D_i - W_i) R) \\ &+ \lambda \sum_{i=1}^{a} \| G_i \|_{1,2}^2 \\ &s.t. \qquad G_i \geq 0, \sum_{i=1}^{b} \theta_i = 1 \end{aligned} \tag{65}$$

The details are shown in **Figure 11**.

Other Methods

There are several methods used to predict possible LPIs except for matrix factorization-based methods and ensemble learning-based methods, for example, Fisher's linear discriminant-based LPI prediction method (IncPro) (Lu et al., 2013), eigenvalue transformation-based semi-supervised model (LPI-ETSLP) (Hu et al., 2017), and kernel ridge regression model based on fast kernel learning(LPI-FKLKRR) (Shen et al., 2018).

lncPRO

Lu et al. (2013) explored a Fisher's linear discriminant-based LPI prediction method, lncPro. lncPro found new LPI through executing the following four steps.

Step 1 Downloading complexes data from the PDB database.

Step 2 Encoding sequence information into numerical feature vectors for lncRNAs and proteins based on the secondary structure, the Van der Waals' propensities, and the hydrogen-bonding propensities.

Step 3 Transforming the feature vectors to unify the dimension based on the Fourier series:

$$X_k' = \sqrt{\frac{2}{L}} \sum_{i=0}^{L} X_i \cos\left[\frac{\pi}{L}\left(n + \frac{1}{2}\right)\left(k + \frac{1}{2}\right)\right] \tag{66}$$

$$k = 0, 1, \ldots, 9$$

where L is the length of feature vector of lncRNAs/proteins.

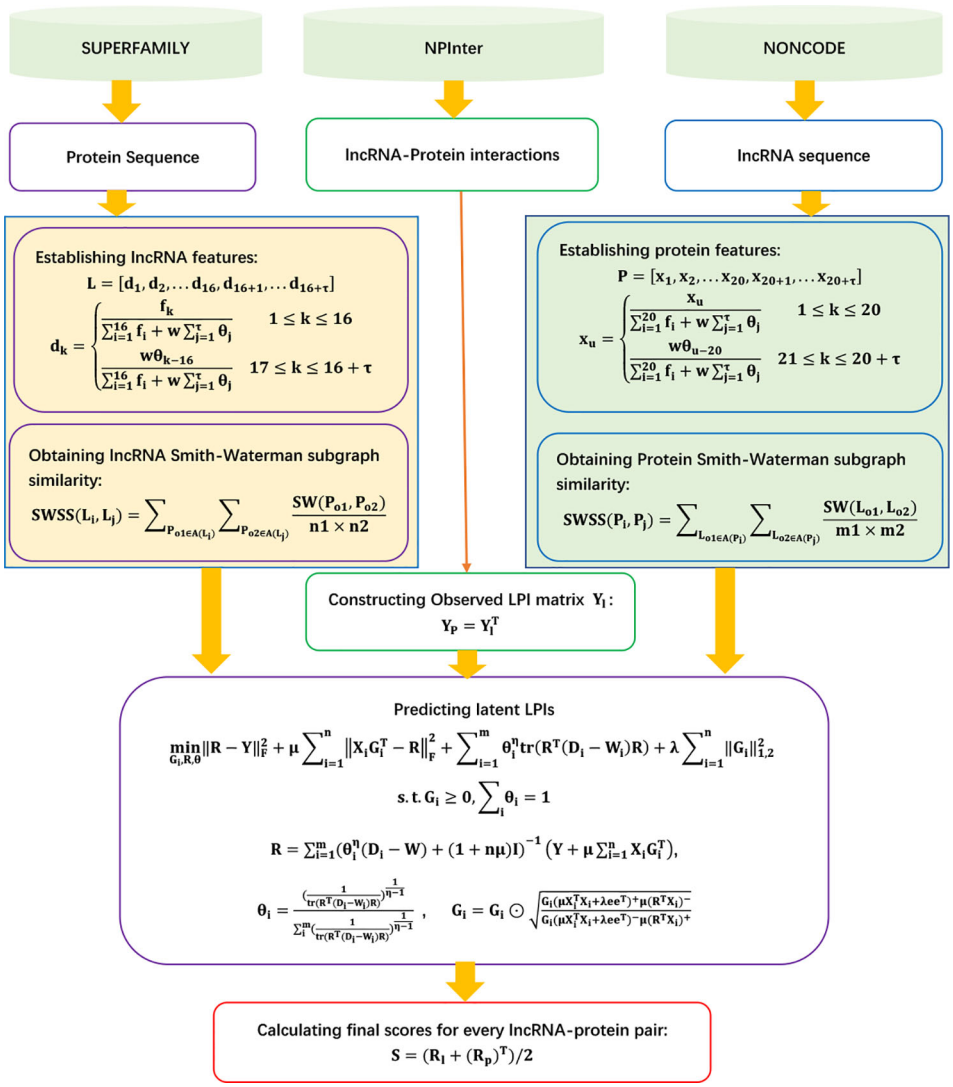


FIGURE 11 | Flowchart of LPI prediction method based on sequence feature projection ensemble learning framework.

Step 4 Calculating the final score matrix $< p|M|r >$ for the RNA feature vector r and a protein feature vector p based on Fisher's linear discriminant method:

$$< p|M|r >= M_1 p_1 r_1 + M_2 p_1 r_2 + M_3 p_2 r_1 + M_4 p_2 r_2 \tag{67}$$

LPI-ETSLP

Hu et al. (2017) presented an eigenvalue transformation-based semi-supervised model, LPI-ESTLP, to uncover the underlying LPIs. LPI-ESTLP can be broken down into three steps.

Step 1 Downloading lncRNA sequences, protein sequences, and LPIs from NONCODE (Zhao et al., 2015), UniProt (Consortium et al., 2018), and NPInter (Hao et al., 2016); and extracting 4,158 LPIs between 27 proteins and 990 lncRNAs after preprocessing.

Step 2 Computing the lncRNA sequence similarity matrix *LSM* and protein sequence similarity matrix *PSM* based on the Smith–Waterman algorithm:

$$LSM(l(i), l(j)) = \frac{sw(l(i), l(j))}{\max\left(sw(l(i), l(i)), sw(l(j), l(j))\right)} \tag{68}$$

$$PSM(p(i), p(j)) = \frac{sw(p(i), p(j))}{\max\left(sw(p(i), p(i)), sw(p(j), p(j))\right)} \tag{69}$$

Step 3 Calculating the score matrix based on the following objective function:

$$\bar{Y} = \frac{\bar{Y}_l + \bar{Y}_p}{2} \tag{70}$$

where

$$\bar{Y}_l = (\sigma L_l + I)^{-1} Y$$
$$\bar{Y}_p = (\sigma L_p + I)^{-1} Y \quad (71)$$

and $L_l = I - LSM$ and $L_p = I - PSM$ denote the Laplacian matrices of lncRNAs and proteins, respectively.

LPI-ETSLP can obtain the final scores between unobserved lncRNA–protein pairs by integrating eigenvalue transformation into Eq. 70:

$$Y = \frac{1}{2}(V_l U_l V_l^T + V_p^T U_p V_p) \quad (72)$$

where $\bar{U}_l$ is a diagonal matrix with $[\bar{U}_l]_{ii} = (1 + \sigma(1 - \lambda_{l_i}^{\alpha}))^{-1}$. $L_l = I - D_l^{-0.5} K_l D_l^{-0.5}$ and the eigen decomposition of K_l can be expressed as $K_l = V_l U_l V_l$. Similarly, $K_p = V_p U_p V_p$ and U_p can be defined.

The details are shown in **Figure 12**.

LPI-FKLKRR

Shen et al. (2018) developed an LPI prediction algorithm, LPI-FKLKRR, combining a kernel ridge regression model based on fast kernel learning. LPI-FKLKRR can be broken into six steps:

Step 1 Computing lncRNA GIP, sequence feature, sequence similarity, and lncRNA expression kernels K_{GIP}^{lnc}, K_{SW}^{lnc}, K_{SF}^{lnc}, and K_{EXP}^{lnc}.

Step 2 Computing protein GIP, sequence features, protein sequence similarity, and protein GO kernel K_{GIP}^{pro}, K_{SW}^{pro}, K_{SF}^{pro}, K_{GO}^{pro}.

Step 3 Generating the optimal lncRNA and protein kernels with fast kernel learning:

$$K_{lnc} = \sum_{a=1}^{4} w_a^{lnc} K_a^{lnc}, K_a^{lnc} \in \Re^{m \times m} \quad (73)$$

$$K_{pro} = \sum_{a=1}^{4} w_a^{pro} K_a^{pro}, K_a^{pro} \in \Re^{m \times m}$$

where w_a^{lnc} and w_a^{pro} represent each element in w_{lnc} and w_{pro}, respectively; K_a^{lnc} and K_a^{pro} denote the corresponding normalized similarity matrices in lncRNA and protein spaces, respectively.

Step 4 Constructing the optimization model to compute the optimal solution for w^{lnc} or w^{pro}:

$$\min_w w^T(A + \lambda I)w - 2b^T w$$
$$s.t. \sum_a^J w_a = 1 \quad (74)$$
$$A_{u,v} = tr(K_u^T K_v)$$

where w denotes the optimal solution w_{lnc} or w_{pro}, K_u and K_v denote two different kernel matrices, and $tr(\cdot)$ denotes the trace function.

Step 5 Computing lncRNA–protein association score matrix:

$$F^* = K_{lnc}(K_{lnc} + \lambda_\ell I)^{-1} F (K_{pro} + \lambda_p I)^{-1} K_{pro} \quad (75)$$

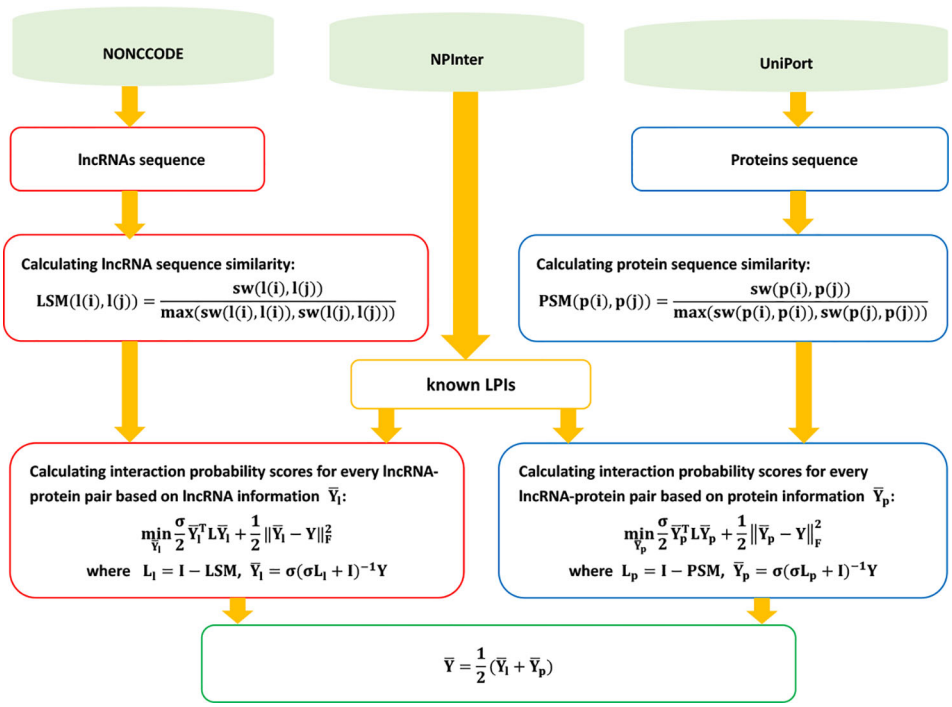


FIGURE 12 | Flowchart of eigenvalue transformation-based semi-supervised model.

Step 6 Producing the optimal F^* by adjusting the parameters λ_ℓ and λ_p.

The details are shown in **Figure 13**.

DISCUSSION

lncRNAs play important regulatory roles in diverse biological processes, such as protein modification, DNA methylation, and chromosome (Weber et al., 2018; Huang et al., 2018a; He et al., 2018b; Zhao et al., 2018c). However, the regulatory mechanism remains unknown (Esteller, 2011; Jiang et al., 2018; Agirre et al., 2019). Studies reported that identifying protein molecules binding specific lncRNAs help to probe the mechanism of lncRNAs (Lu et al., 2013; Ge et al., 2016; Chen et al., 2018). Therefore, identifying possible LPIs has an important role in understanding lncRNA-related activities (Lu et al., 2013; Pan et al., 2016; Peng et al., 2017; Zhang et al., 2018c).

However, experimental methods are expensive and time-consuming. For limited existing knowledge, computational methods become vital as a silver-bullet solution to capture LPIs on a large scale, which contributes to prioritize LPI candidates and deploys further experimental validation (Chen et al., 2018).

In this study, databases involved in LPI identification are summarized. More importantly, the components of state-of-the-art computational models for LPI prediction, such as network-based methods and machine learning-based methods, are introduced. Particularly, machine learning-based models can be broken into matrix factorization-based methods and ensemble learning-based methods. To consider the performance of LPI prediction methods, we compared nine models (IRWNRLPI, LPBNI, LPGNMF, LPI-BNPRA, LPI-ETSLP, LPIHN, LPI-NRLMF, LPLNP, and SFPEL-LPI) on leave-one-out cross-validation (LOOCV). These nine models are conducted on the datasets provided by the corresponding papers. Parameters are set as the values recommended by the corresponding studies. **Table 1** shows the comparison results based on AUC, precision, accuracy, and F1. In **Table 1**, SFPEL-LPI obtained the best performances of AUC and accuracy; LPGNMF obtained the best performances of precision and F1. The results demonstrated that SFPEL-LPI can correctly predict LPIs with a relative high proportion. LPGNMF can better identify potential LPIs when taking into account the proportion of correctly predicted LPIs and successfully predicted LPIs.

To further detect the performance of SFPEL-LPI, we compared it with four representative LPI prediction methods, LPBNI, LPI-ETSLP, LPIHN, and LPLNP, on fivefold cross-validation. The experiments were conducted on the same dataset, i.e., LPIs, lncRNA sequences, and protein sequences are from NPInter (Hao et al., 2016), NONCODE (Zhao et al., 2015), and SUMPERFAMILY (Pandurangan et al., 2018), respectively. The details are shown in **Table 2**. The results demonstrate that SFPEL-LPI obtained the best performance of AUC and can better identify possible LPIs.

In general, network-based methods have become one type of effective tool in possible LPI identification by utilizing LPI network, lncRNA similarity network, and protein similarity matrix. Although network-based methods efficiently discovered unknown LPIs and obtained promising results from the perspective of propagation (Li et al., 2015; Ge et al., 2016; Zheng et al., 2017; Zhao et al., 2018b), this type of method has some weaknesses.

1. Parts of computational methods tested their performances only on one database, which may result in biased predictions because of the sparse nature of LPI data (Li et al., 2015). More importantly, the lack of known LPIs limits the further research of LPI prediction in a larger network (Ge et al., 2016).
2. It is important to unravel potential LPIs for lncRNAs/proteins without any associated information (we represent these lncRNAs/proteins as new lncRNAs/proteins); however, most network-based models fail to capture LPI candidates (Zhang et al., 2018b).
3. Current network-based methods tend to be biased to the lncRNAs/proteins with more known associated proteins. Some lncRNAs/proteins interact with multiple proteins/lncRNAs and others interact with a few or even only one protein/lncRNA in an LPI network. The unbalanced nature of degree distributions in the LPI network may affect prediction performance. Increasing resistance based on the random walk may improve predictive accuracy for LPI prediction models (Li et al., 2015).
4. Parts of methods compute lncRNA similarities based on the expression profile and may produce incomplete coverage of the lncRNA similarity network when adding LPI datasets. This problem may be solved by increasing appropriate data including LPIs (Li et al., 2015).
5. Network-based methods can be applied to an LPI network in which there exists at least one link between two nodes. Especially for a bipartite network, network-based methods require that each node in the network has at least two linkages. However, the LPI network is usually composed of a few isolated subnetworks, and most of the existing network-based models fail to identify the LPIs between the lncRNAs in one subnetwork and the proteins in another (Ge et al., 2016).
6. Most current network-based methods utilized local network information and showed better performance; however, many previous computational biology studies showed that global network information contributes to capturing the associations between two entities, such as LPIs (Karuza et al., 2016; Meng et al., 2016; Shi et al., 2017).
7. Biology finally aims at providing personalized medicine for cancer patients, and it is a key issue to predict relevant drugs/targets for a certain disease by integrating multiple heterogeneous networks and constructing multiple-partite biological networks, such as protein–lncRNA–disease association networks and drug–protein–lncRNA–disease networks. However, current network-based methods are still not applied to this type of prediction (Yao et al., 2016; Yang et al., 2017; Bester et al., 2018; Lu et al., 2018; Ping et al., 2018; Fan et al., 2019).

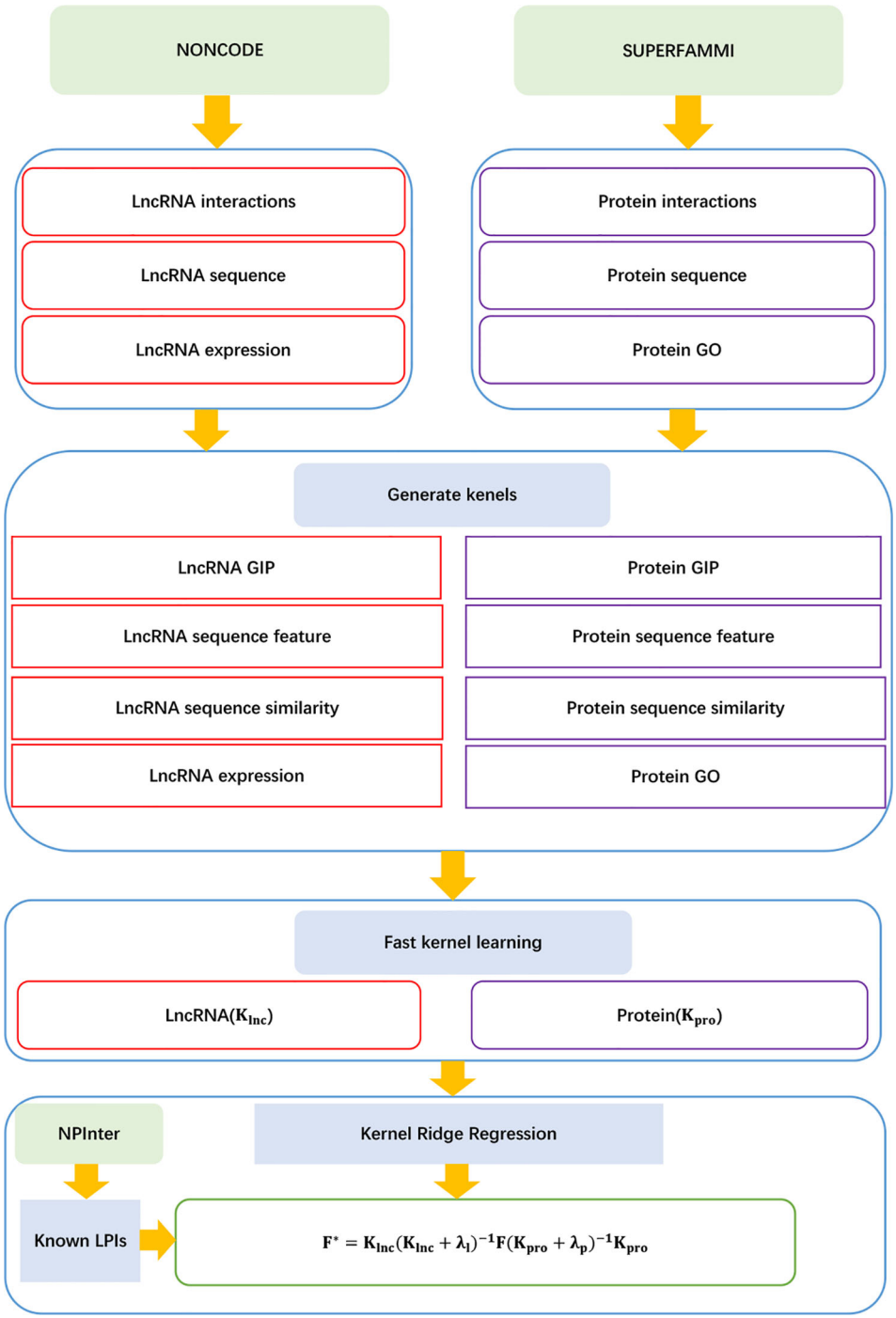


FIGURE 13 | Flowchart of LPI prediction method based on fast kernel learning with kernel ridge regression.

In summary, machine learning-based LPI prediction methods have some limitations.

1. There are no non-LPIs (negative samples) with experimental validation; therefore, most supervised learning-based LPI prediction models can only randomly select unknown lncRNA–protein pairs as negative LPIs. However, this part of randomly selected negative LPIs may contain true LPIs (positive samples) as well, which significantly influences the predictive performance (Liu et al., 2017; Zhao et al., 2018a; Zhao et al., 2018b; Zhang et al., 2018c; Shen et al., 2019). Although semi-supervised learning-based models utilized

TABLE 1 | Performance of LPI prediction methods on LOOCV.

Methods	AUC	precision	accuracy	F1
IRWNRLPI	0.9150	0.7178	0.9009	0.6516
LPBNI	0.8586	0.9681	0.9581	0.3868
LPGNMF	0.8520	**1**	0.7854	**0.6871**
LPI-BNPRA	0.8754	0.6540	0.8799	0.5564
LPI-ETSLP	0.8876	0.5932	0.8834	0.5978
LPIHN	0.8030	0.3713	0.9581	0.3868
LPI-NRLMF	0.9025	0.6129	0.8804	0.6197
LPLNP	0.9594	0.1153	0.9592	0.1621
SFPEL-LPI	**0.9735**	0.0016	**0.9731**	0.0033

These bolded texts represent that the corresponding method is the best among comparison methods.

unlabeled information to decrease the limitations of negative LPI selection, it still has the same disadvantage as classifier combination (Liu et al., 2017; Zhang et al., 2018a; Shen et al., 2019).

2. Some machine learning-based methods constructed two different classifiers, based on lncRNAs and proteins, respectively. The final results are an average of the performances of two predictive models. This type of model will produce biased results (Zhao et al., 2018b).
3. Many lncRNAs/proteins do not have known association information with any proteins/lncRNAs, and we represent them as new lncRNAs/proteins. Most current predictive models are unable to capture possible proteins/lncRNAs for new lncRNAs/proteins (Zhang et al., 2018c).
4. The proposed methods rely heavily on known LPI data; however, the current number of known LPIs is still very low. Therefore, most machine learning-based models are trained using RNA–protein interaction information instead of LPI data. This results in limited predictive performances (Liu et al., 2017; Zhao et al., 2018a). With the increase in experimentally validated LPIs, the prediction performances of models will improve (Zhao et al., 2018b).
5. The better performances of existing machine learning methods rely severely on data called features (Goodfellow et al., 2016). Current computational methods utilize various lncRNA features and protein features. However, identifying more appropriate features for a given task is still a challenge (Liu et al., 2017; Min et al., 2017). More importantly, these features are not available for all proteins or lncRNAs (Liu et al., 2017; Zhang et al., 2018c).
6. Most experimental data are provided by the NPInter database. NPInter is a relatively abundant database for lncRNA and protein data, but it only provides gene–protein interaction data corresponding to relevant lncRNAs instead of direct LPIs. Gene–protein interactions were directly applied to machine learning-based methods to find possible ncRNA–protein associations and did not discover true LPIs (Liu et al., 2017; Zhao et al., 2018a; Zhao et al., 2018b).
7. Most current computational models for LPI interaction prediction are measured based on cross-validation. Park and Marcotte (2012) used a proteochemometrics model (Wikberg and Mutulis, 2008) for drug–protein interaction prediction and observed that the paired nature of input samples has significant implications on the cross-validation of these pair-input methods. That is to say, there are significant cross-validation differences between input sample and out-of-sample interactions (Park and Marcotte, 2012). For drug–target interaction identification problems, the paired feature of input samples may produce a natural partition of test pairs, and thus the pair-input methods may obtain significantly distinct prediction accuracies for different test classes (Chen et al., 2015). The same situation applies to LPI prediction, which is still a pair-input computational identification problem.

TABLE 2 | Performance of LPI prediction methods on fivefold cross-validation.

Methods	AUC	Precision	Accuracy	F1
LPBNI	0.84177	0.2898	0.9431	0.3336
LPI-ETSLP	0.8876	**0.5932**	0.8834	**0.5978**
LPIHN	0.8531	0.4139	0.9581	0.3868
LPLNP	0.9104	0.4102	**0.9646**	0.4520
SFPEL-LPI	**0.9200**	0.4490	0.9600	0.4702

These bolded texts represent that the corresponding method is the best among comparison methods.

CONCLUSION AND FURTHER RESEARCH

There are a few LPIs and numerous unknown lncRNA–protein pairs not validated by experimental methods in the existing databases. In addition, similar lncRNAs tend to interact with similar proteins, and vice versa (Xiao et al., 2017; Zhang et al., 2018a). Therefore, LPI data have a sparse, low-rank, and unbalanced nature (Li et al., 2015; Zhang et al., 2018a; Shen et al., 2019). With the development of experimental technology, more LPIs will be confirmed, and thus the prediction accuracy of computational models will increase. In this section, we present some suggestions for further research based on the nature of LPI data.

Fusing Comprehensive LPI Datasets

Parts of computational methods tested their performances only on one database, which may result in biased predictions because of the sparse nature of LPI data (Li et al., 2015). More importantly, existing computational models utilize various biological information from proteins and lncRNAs, for example, physicochemical properties including hydrogen bonding, secondary structure, and van der Waals propensities (Belluci et al., 2011; Xiao et al., 2017). It is important to utilize diverse biological features to improve the performances of LPI prediction models. However, these features are not available for all proteins or lncRNAs, and thus computational methods cannot capture LPI candidates when information is unavailable (Zhang et al., 2018c). Therefore, exploring advanced data fusion methods to integrate more available data sources may further boost the performance of LPI identification.

Focusing on the drawbacks of current network-based LPI identification methods, future research can begin with

integrating more heterogeneous networks, such as protein–protein interaction network (Zhang et al., 2019a), lncRNA–miRNA interaction network (Zeng et al., 2016; Huang et al., 2018c; Zhao et al., 2019), lncRNA–mRNA interaction network (Alaei et al., 2019), lncRNA–disease association network (Fu et al., 2017; Wang et al., 2019), and lncRNA–miRNA–mRNA regulatory network (Chen et al., 2018; Zhang et al., 2019b). However, how to address the data conflict problems while integrating diverse LPI data from different repositories is a challenge.

Although there are not currently data conflict solutions for LPI prediction, we can find some clues by other problems in the area of bioinformatics. For example, Liu et al. (2015) set a confidence level for each DTI and gave a higher score to a DTI from a more reliable data repository. For example, the STITCH database assigns a score with a range [0, 1,000] to each DTI based on four types of different sources: model prediction, text mining, manually curated databases, and experimental validation. Particularly, Liu et al. (2015) gave DTIs from Matador and DrugBank the highest values (1,000) because DTIs from these two databases are reported by biochemical experiments and relevant studies. Lou et al. (2017) exploited another type of data fusion from a multiple-views perspective. This involved five steps: screening relevant information from different data sources; removing isolated nodes without edges in the networks; fusing various types of nodes and edges and building a heterogeneous network; constructing multiple similarity networks to boost the network heterogeneity; and excluding homologous nodes from the constructed heterogeneous networks to further reduce the possible redundancy of associated information. Inspired by these two methods, we can fuse diverse heterogeneous data to improve performance in future research. More importantly, new exploited network-based methods should be implemented on a constructed heterogeneous network rather than a single network.

Screening Credible Negative Samples

There are some known LPIs (positive samples) and abundant unknown lncRNA–protein pairs in existing LPI data resources. More importantly, there are no experimentally validated non-LPIs, and thus most supervised learning-based models have no other choice but to randomly screen negative LPIs from unlabeled lncRNA–protein pairs or even regarded all unlabeled lncRNA–protein pairs as negative samples (Liu et al., 2017; Zhao et al., 2018b). However, the randomly screened negative LPIs may contain positive LPIs as well, and thus there are severe biases in supervised learning-based techniques. Therefore, exploiting an efficient model to select high-quality negative samples is a challenging task for boosting LPI prediction accuracy.

Cheng et al. (2017) designed a FInding Reliable nEgative samples method (FIRE) to select negative RNA–protein interactions. FIRE was based on the following assumption: given a known RNA–protein interaction between an RNA *i* and a protein *j*, for an RNA *k*, the more differences between *i* and *k*, the less possibility that *k* interacts with *j*, and vice versa. FIRE screened negative RNA–protein interactions through the following steps: computing the protein similarity matrix, building a positive sample set based on known interaction information, scoring an unknown RNA–protein pair not included in positive sample set based on protein similarities, generating *m* negative samples by sorting these RNA–protein pairs *via* their scores in increasing order, and selecting the top-*m* RNA–protein pairs. Similarly, we may generate negative LPIs based on lncRNA–lncRNA similarities, protein–protein similarities, and the above assumption.

Positive-unlabeled (PU) learning (de Campos et al., 2018; Sansone et al., 2018; Yang et al., 2018) is applied to various situations. In PU learning, a supervised learning-based method is designed to learn a classification model from a positive sample set and an unlabeled dataset from an unknown class. Yang et al. (2018) designed an adaptive sampling framework with class label noise based on PU learning and introduced two new bioinformatic applications: identifying kinase-substrates and identifying transcription factor target genes. Therefore, PU learning may be one strong way to solve the problem of lacking negative LPIs.

Deep Learning

Existing computational methods have utilized different lncRNA features and protein features. For example, Bellucci et al. (2011) integrated three types of physicochemical properties, including hydrogen bonding, secondary structure, and van der Waals propensities; meanwhile, Lu et al. (2013) used six types of RNA secondary structures (besides physicochemical properties), which were provided by Bellucci et al. (2011). Therefore, designing more powerful models to integrate relevant biological features is a key issue. However, features are typically exploited by human biomedical engineers, and determining which features are more suitable for LPI prediction remains difficult. More importantly, encoding vectors that are too short may restrict the prediction accuracy of classification model. More importantly, most computational models only used sequence information but did not consider structure information (Peng et al., 2019).

Deep learning-based computational models composed of multiple processing layers require very little engineering knowledge and can efficiently extract features from raw data and construct high-level representations (Wei et al., 2018; Peng et al., 2019). These types of models have been applied to diverse analysis problems, and have obtained better performance due to the excellent power of feature learning (Jurtz et al., 2017; Min et al., 2017; Peng et al., 2019). Therefore, it is valuable and feasible to exploit deep learning-based methods to highly and effectively represent biological features for relevant entities in bioinformatics (Min et al., 2017; Zhang et al., 2018d; Peng et al., 2019; Zeng et al., 2019), such as information relevant to LPI prediction (Xiao et al., 2017; Shen et al., 2019; Zhu et al., 2019). More importantly, although deep learning demonstrated promising performance, it is not a silver bullet in LPI prediction. There still exist many challenges in LPI identification, such as the imbalanced nature of LPI data, limited LPI data, appropriate architecture selection, hyper parameter selection, and interpretation of learning results (Min et al., 2017). Therefore, solving these problems is the key to promoting deep learning-based LPI prediction models in future research.

Particularly, deep learning can be combined with PU learning and improve the performance of computational models (Bepler et al., 2018; Pati et al., 2018). For example, Bepler et al., 2018 designed the first particle-picking framework, Topaz. Topaz combined a convolutional neural network with a generalized-expectation-binomial-based objective function. The convolutional neural network was used to train classification models using only positive and unlabeled samples. Meanwhile, the generalized-expectation-binomial-based objective function was used to learn model parameters based on positive and unlabeled samples. Topaz utilized convolutional neural network classifiers to fit labeled particles (samples) and the remaining unlabeled samples based on the minibatched stochastic gradient decent method. Deep learning methods based on PU learning provide valuable insight and may be a starting point for deep learning applied to LPI prediction in future research.

Capturing LPI Candidates for New LncRNAs/Proteins

Network-based methods can be applied to an LPI network that has least one link between two nodes. For a bipartite network especially, network-based methods require that each node in the network has at least two linkages. That is to say, network-based methods cannot discover possible proteins for any lncRNA–protein pair without any known reachable paths in the LPI network (Ge et al., 2016; Zhang et al., 2018c). These lncRNAs/proteins without any interaction information are regarded as new lncRNAs/proteins (Zhang et al,. 2018c).

Given a known LPI dataset, we aim to predict (S1) LPIs between known lncRNAs and known proteins; (S2) LPIs between new lncRNAs and known proteins; (S3) LPIs between known lncRNAs and new proteins; and (S4) LPIs between new lncRNAs and new proteins. S1 has the most abundant association information, S2 and S3 have less data, and S4 has the least data. Computational models appropriate for S2 can still be applied to S3, and vice versa.

To the best of our knowledge, SFPEL-LPI provided by Zhang et al. (2018c) may be one of the rare computational methods for predicting possible LPIs for new lncRNAs/proteins. Although few computational models can be applied to the last three situations, some methods have been designed to solve similar problems in other areas in bioinformatics, and thus provide some clues for LPI prediction. For example, Shi et al. (2015) enhanced the similarity measures and introduced the concept of a "super-target" to capture the missing interactions for new drugs/targets. Furthermore, Chen et al. (2016b) exploited a miRNA–disease association prediction model based on within and between scores (WBSMDA) to uncover possible miRNA–disease associations for new miRNAs/diseases. These solutions provide clues for capturing LPI candidates for new lncRNAs/proteins.

Cross-Validation

Inspired by the evaluation methods proposed by Park and Marcotte (2012) and Chen et al. (2015), the test samples of LPIs could be categorized into four different groups: C1 is composed of the test samples sharing both lncRNAs and proteins with the training samples; C2 is composed of the test samples sharing only lncRNA with the training samples; C3 is composed of the test samples sharing only proteins with the training samples; and C4 is composed of the test samples sharing neither lncRNAs nor proteins with the training samples (Chen et al. (2015)). Therefore, it is vital to give cross-validation results under the above four independent test classes for LPI prediction.

AUTHOR CONTRIBUTIONS

LP and FL contributed equally to this work. LP, FL, XD, CP, and LZ introduced LPI data repositories and computational models. LP and FL wrote the paper. XL and YM revised original draft. LP, JY, GT, and LZ discussed the computational models and gave conclusion and further research. All authors read and approved the final manuscript.

FUNDING

This research was funded by the Natural Science Foundation of China (Grant 61803151), the Natural Science Foundation of Hunan province (Grant 2018JJ2461, 2018JJ3570), and the Project of Scientific Research Fund of Hunan Provincial Education Department (Grant 17A052).

ACKNOWLEDGMENTS

We would like to thank all authors of the cited references.

REFERENCES

Agirre, X., Meydan, C., Jiang, Y., Garate, L., Doane, A. S., Li, Z., et al. (2019). Long non-coding rnas discriminate the stages and gene regulatory states of human humoral immune response. *Nat. Commun.* 10, 821. doi: 10.1038/s41467-019-08679-z

Alaei, S., Sadeghi, B., Najafi, A., and Masoudi-Nejad, A. (2019). Lncrna and mrna integration network reconstruction reveals novel key regulators in esophageal squamous-cell carcinoma. *Genomics* 111, 76–89. doi: 10.1016/j.ygeno.2018.01.003

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2018). Lncrnadisease 2.0: an updated database of long non-coding rna-associated diseases. *Nucleic Acids Res.* 47, D1034–D103D, 1037. doi: 10.1093/nar/gky905

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods (Nature Publishing Group)* 8 (6), 444. doi: 10.1038/nmeth.1611

Bepler, T., Morin, A., Noble, A. J., Brasch, J., Shapiro, L., and Berger, B. (2018). "Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs," in *Research in computational molecular biology: Annual International Conference, RECOM: proceedings. RECOMB (Conference: 2005-) (NIH Public Access)* (Nature Publishing Group), vol. 10812, p 245–247.

Bester, A. C., Lee, J. D., Chavez, A., Lee, Y.-R., Nachmani, D., Vora, S., et al. (2018). An integrated genome-wide crispra approach to functionalize lncrnas in drug resistance. *Cell* 173, 649–664. doi: 10.1016/j.cell.2018.03.052

Cantini, L., Kairov, U., De Reyniès, A., Barillot, E., Radvanyi, F., and Zinovyev, A. (2019). Assessing reproducibility of matrix factorization methods in

independent transcriptomes. *Bioinformatics*. doi: 10.1093/bioinformatics/btz225/5426054

Chen, X., and Yan, G.-Y. (2013). Novel human lncrna-disease association inference based on lncrna expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Chen, X., Yan, C. C., Zhang, X., Zhang, X., Dai, F., Yin, J., et al. (2015). Drug–target interaction prediction: databases, web servers and computational models. *Briefings In Bioinf.* 17, 696–712. doi: 10.1093/bib/bbv066

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016a). Long non-coding rnas and complex diseases: from experimental results to computational models. *Briefings In Bioinf.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., Yan, C. C., Zhang, X., You, Z.-H., Deng, L., Liu, Y., et al. (2016b). Wbsmda: within and between score for mirna-disease association prediction. *Sci. Rep.* 6, 21106. doi: 10.1038/srep21106

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2018). Computational models for lncrna function prediction and functional similarity calculation. *Briefings In Funct. Genomics* 18, 58–82. doi: 10.1093/bfgp/ely031

Cheng, Z., Huang, K., Wang, Y., Liu, H., Guan, J., and Zhou, S. (2017). Selecting high-quality negative samples for effectively predicting protein-rna interactions. *BMC Syst. Biol.* 11, 9. doi: 10.1186/s12918-017-0390-8

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018). Lncrna2target v2. 0: a comprehensive database for target genes of lncrnas in human and mouse. *Nucleic Acids Res.* 47, D140–D144. doi: 10.1093/nar/gky1051

Consortium, U., Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., et al. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46, 2699. doi: 10.1093/nar/gky092

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2017). Mndr v2. 0: an updated resource of ncrna–disease associations in mammals. *Nucleic Acids Res.* 46, D371–D374. doi: 10.1093/nar/gkx1025

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., et al. (2018). Ensembl 2019. *Nucleic Acids Res.* 47, D745–D751. doi: 10.1093/nar/gky1113

Dallner, O. S., Marinis, J. M., Lu, Y.-H., Birsoy, K., Werner, E., Fayzikhodjaeva, G., et al. (2019). Dysregulation of a long noncoding rna reduces leptin leading to a leptin-responsive form of obesity. *Nat. Med.* 1, 507–516. doi: 10.1038/s41591-019-0370-1

de Campos, L. M., Fernández-Luna, J. M., Huete, J. F., and Redondo-Expósito, L. (2018). Positive unlabeled learning for building recommender systems in a parliamentary setting. *Inf. Sci.* 433, 221–232. doi: 10.1016/j.ins.2017.12.046

Esteller, M. (2011). Non-coding rnas in human disease. *Nat. Rev. Genet.* 12, 861. doi: 10.1038/nrg3074

Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., Zhu, K., and Lu, S. (2019). Prediction of lncrna-disease associations by integrating diverse heterogeneous information sources with rwr algorithm and positive pointwise mutual information. *BMC Bioinf.* 20, 87. doi: 10.1186/s12859-019-2675-y

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2017). Matrix factorization-based data fusion for the prediction of lncrna–disease associations. *Bioinformatics* 34, 1529–1537. doi: 10.1093/bioinformatics/btx794

Gao, Y., Wang, P., Wang, Y., Ma, X., Zhi, H., Zhou, D., et al. (2018). Lnc2cancer v2. 0: updated database of experimentally supported long non-coding rnas in human cancers. *Nucleic Acids Res.* 47, D1028–D102D, 1033. doi: 10.1093/nar/gky1096

Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding rna–protein interactions. *Genomics Proteomics Bioinf.* 14, 62–71. doi: 10.1016/j.gpb.2016.01.004

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning* (Cambridge, Massachusetts, USA: MIT press).

Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., et al. (2016). Npinter v3. 0: an upgraded database of noncoding rna-associated interactions. *Database* 2016, 1–9. doi: 10.1093/database/baw057

He, W., Ju, Y., Zeng, X., Liu, X., and Zou, Q. (2018a). Sc-ncdnapred: a sequence-based predictor for identifying non-coding dna in saccharomyces cerevisiae. *Front. In Microbiol.* 9, 2174. doi: 10.3389/fmicb.2018.02174

He, Y., Zuo, Q., Edwards, J., Zhao, K., Lei, J., Cai, W., et al. (2018b). Dna methylation and regulatory elements during chicken germline stem cell differentiation. *Stem Cell Rep.* 10, 1793–1806. doi: 10.1016/j.stemcr.2018.03.018

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of rna-binding proteins. *Nat. Rev. Mol. Cell Biol.* 19, 327. doi: 10.1038/nrm.2017.130

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J., Gough, J., et al. (2017). An atlas of human long non-coding rnas with accurate 5' ends. *Nature* 543, 199. doi: 10.1038/nature21374

Hu, H., Zhu, C., Ai, H., Zhang, L., Zhao, J., Zhao, Q., et al. (2017). Lpi-etslp: lncrna–protein interaction prediction using eigenvalue transformation-based semi-supervised link prediction. *Mol. Biosyst.* 13, 1781–1787. doi: 10.1039/c7mb00290d

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi-ensemble: Prediction of human lncrna-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Huang, X., Zhou, X., Hu, Q., Sun, B., Deng, M., Qi, X., et al. (2018a). Advances in esophageal cancer: a new perspective on pathogenesis associated with long non-coding rnas. *Cancer Lett.* 413, 94–101. doi: 10.1016/j.canlet.2017.10.046

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2018b). Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. *Nucleic Acids Res.* 47, D1013–D101D, 1017. doi: 10.1093/nar/gky1010

Huang, Z.-A., Huang, Y.-A., You, Z.-H., Zhu, Z., and Sun, Y. (2018c). Novel link prediction for large-scale mirna-lncrna interaction network in a bipartite graph. *BMC Med. Genomics* 11, 113. doi: 10.1186/s12920-018-0429-8

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2008). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104. doi: 10.1093/nar/gkn714

Jiang, C., Ding, N., Li, J., Jin, X., Li, L., Pan, T., et al. (2018). Landscape of the long non-coding rna transcriptome in human heart. *Brief. Bioinform.* 20 (5), 1812–1825. doi: 10.1093/bib/bby052

Jurtz, V. I., Johansen, A. R., Nielsen, M., Almagro Armenteros, J. J., Nielsen, H., Sønderby, C. K., et al. (2017). An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33, 3685–3690. doi: 10.1093/bioinformatics/btx531

Karuza, E. A., Thompson-Schill, S. L., and Bassett, D. S. (2016). Local patterns to global architectures: influences of network topology on human learning. *Trends In Cogn. Sci.* 20, 629–640. doi: 10.1016/j.tics.2016.06.003

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F.-X., Pan, Y., et al. (2016). Ldap: a web server for lncrna-disease association prediction. *Bioinformatics* 33, 458–460. doi: 10.1093/bioinformatics/btw639

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2013). starbase v2. 0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Li, Y., Wang, C., Miao, Z., Bi, X., Wu, D., Jin, N., et al. (2014). Virbase: a resource for virus–host ncRNA-associated interactions. *Nucleic Acids Res.* (Oxford University Press) 43, D578–D582. doi: 10.1093/nar/gku903

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting long noncoding rna and protein interactions using heterogeneous network model. *BioMed. Res. Int.* 2015, 1–11. doi: 10.1155/2015/671950

Li, Y., Egranov, S. D., Yang, L., and Lin, C. (2019a). Molecular mechanisms of long noncoding rnas-mediated cancer metastasis. *Genes Chromosomes Cancer* 58, 200–207. doi: 10.1002/gcc.22691

Li, Y.-P., Duan, F.-F., Zhao, Y.-T., Gu, K.-L., Liao, L.-Q., Su, H.-B., et al. (2019b). A trim71 binding long noncoding rna trincr1 represses fgf/erk signaling in embryonic stem cells. *Nat. Commun.* 10, 1368. doi: 10.1002/gcc.22691

Liu, H., Sun, J., Guan, J., Zheng, J., and Zhou, S. (2015). Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 31, i221–i229. doi: 10.1093/bioinformatics/btv256

Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., et al. (2017). Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* 8, 103975. doi: 10.18632/oncotarget.21934

Liu, M., Zhang, H., Li, Y., Wang, R., Li, Y., Zhang, H., et al. (2018). Hotair, a long noncoding rna, is a marker of abnormal cell cycle regulation in lung cancer. *Cancer Sci.* 109, 2717. doi: 10.1111/cas.13745

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., et al. (2013). Computational prediction of associations between long non-coding rnas and proteins. *BMC Genomics* 14, 651. doi: 10.1186/1471-2164-14-651

Lu, C., Yang, M., Luo, F., Wu, F.-X., Li, M., Pan, Y., et al. (2018). Prediction of lncrna–disease associations based on inductive matrix completion. *Bioinformatics* 34, 3357–3364. doi: 10.1093/bioinformatics/bty327

Luo, Y., Zhao, X., Zhou, J., Yang, J., Zhang, Y., Kuang, W., et al. (2017). A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* 8, 573. doi: 10.1038/s41467-017-00680-8

Ma, L., Li, A., Zou, D., Xu, X., Xia, L., Yu, J., et al. (2014). Lncrnawiki: harnessing community knowledge in collaborative curation of human long non-coding rnas. *Nucleic Acids Res.* 43, D187–D192. doi: 10.1093/nar/gku1167

Meng, L., Striegel, A., and Milenković, T. (2016). Local versus global biological network alignment. *Bioinformatics* 32, 3155–3164. doi: 10.1093/bioinformatics/btw348

Miao, Y.-R., Liu, W., Zhang, Q., and Guo, A.-Y. (2017). lncrnasnp2: an updated database of functional snps and mutations in human and mouse lncrnas. *Nucleic Acids Res.* 46, D276–D280. doi: 10.1093/nar/gkx1004

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Briefings In Bioinf.* 18, 851–869. doi: 10.1093/bib/bbw068

Mørk, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J., and Jensen, L. J. (2013). Protein-driven inference of mirna–disease associations. *Bioinformatics* 30, 392–397. doi: 10.1093/bioinformatics/btt677

Munschauer, M., Nguyen, C. T., Sirokman, K., Hartigan, C. R., Hogstrom, L., Engreitz, J. M., et al. (2018). The norad lncrna assembles a topoisomerase complex critical for genome stability. *Nature* 561, 132. doi: 10.1038/s41586-018-0453-z

Ning, S., Yue, M., Wang, P., Liu, Y., Zhi, H., Zhang, Y., et al. (2016). Lincsnp 2.0: an updated database for linking disease-associated SNPs to human long non-coding rnas and their TFBS. *Nucleic Acids Res.* gkw945 45 (D1), D74–D78. doi: 10.1093/nar/gkw945

Nozawa, R.-S., and Gilbert, N. (2019). Rna: Nuclear glue for folding the genome. *Trends In Cell Biol.* 29 (3), 201–211. doi: 10.1016/j.tcb.2018.12.003

Pan, X., Fan, Y.-X., Yan, J., and Shen, H.-B. (2016). Ipminer: hidden ncrna-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics* 17, 582. doi: 10.1186/s12864-016-2931-8

Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B., and Gough, J. (2018). The superfamily 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res.* 47, D490–D494. doi: 10.1093/nar/gky1130

Park, Y., and Marcotte, E. M. (2012). Flaws in evaluation schemes for pair-input computational predictions. *Nat. Methods* 9, 1134. doi: 10.1038/nmeth.2259

Pati, P., Andani, S., Pediaditis, M., Viana, M. P., Ruschoff, J. H., Wild, P., et al. (2018). "Deep positive-unlabeled learning for region of interest localization in breast tissue images," in *Medical Imaging 2018: Digital Pathology (International Society for Optics and Photonics)* (SPIE), vol. 10581, p 1058107.

Peng, W., Li, M., Chen, L., and Wang, L. (2017). Predicting protein functions by using unbalanced random walk algorithm on three biological networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 14, 360–369. doi: 10.1109/TCBB.2015.2394314

Peng, C., Han, S., Zhang, H., and Li, Y. (2019). Rpiter: A hierarchical deep learning framework for ncrna–protein interaction prediction. *Int. J. Mol. Sci.* 20, 1070. doi: 10.3390/ijms20051070

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2018). A novel method for lncrna-disease association prediction based on an lncrna-disease association network. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16, 688–693. doi: 10.1039/c9mo00092e

Pyfrom, S. C., Luo, H., and Payton, J. E. (2019). Plaidoh: a novel method for functional prediction of long non-coding rnas identifies cancer-specific lncrna activities. *BMC Genomics* 20, 137. doi: 10.1186/s12864-019-5497-4

Qian, X., Zhao, J., Yeung, P. Y., Zhang, Q. C., and Kwok, C. K. (2018). Revealing lncrna structures and interactions by sequencing-based approaches. *Trends In Biochem. Sci.* 44 (1), 33–52. doi: 10.1016/j.tibs.2018.09.012

Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., et al. (2014). lncrnadb v2. 0: expanding the reference database for functional long noncoding rnas. *Nucleic Acids Res.* 43, D168–D173. doi: 10.1093/nar/gku988

Rajput, B., Pruitt, K. D., and Murphy, T. D. (2018). Refseq curation and annotation of stop codon recoding in vertebrates. *Nucleic Acids Res.* 47, 594–606. doi: 10.1093/nar/gky1234

Rinn, J. L., and Chang, H. Y. (2012). Genome regulation by long noncoding rnas. *Annu. Rev. Biochem.* 81, 145–166. doi: 10.1146/annurev-biochem-051410-092902

Ruepp, A., Kowarsch, A., and Theis, F. (2012). "Phenomir: micrornas in human diseases and biological processes," in *Next-Generation MicroRNA Expression Profiling Technology* (Totowa, NJ, USA: Humana Press), p 249–260.

Sanchez Calle, A., Kawamura, Y., Yamamoto, Y., Takeshita, F., and Ochiya, T. (2018). Emerging roles of long non-coding rna in cancer. *Cancer Sci.* 109, 2093–2100. doi: 10.1111/cas.13642

Sansone, E., De Natale, F. G., and Zhou, Z.-H. (2018). Efficient training for positive unlabeled learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 41 (1), 2584–2598. doi: 10.1109/TPAMI.2018.2860995

Shen, C., Ding, Y., Tang, J., and Guo, F. (2018). Multivariate information fusion with fast kernel learning to kernel ridge regression in predicting lncrna-protein interactions. *Front. In Genet.* 9, 716. doi: 10.3389/fgene.2018.00716

Shen, C., Ding, Y., Tang, J., Jiang, L., and Guo, F. (2019). Lpi-ktaslp: Prediction of lncrna-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/ACCESS.2019.2894225

Shi, J.-Y., Yiu, S.-M., Li, Y., Leung, H. C., and Chin, F. Y. (2015). Predicting drug–target interaction for new drugs using enhanced similarity measures and super-target clustering. *Methods* 83, 98–104. doi: 10.1016/j.ymeth.2015.04.036

Shi, C., Li, Y., Zhang, J., Sun, Y., and Philip, S. Y. (2017). A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.* 29, 17–37. doi: 10.1109/TKDE.2016.2598561

Shi, J.-Y., Zhang, A.-Q., Zhang, S.-W., Mao, K.-T., and Yiu, S.-M. (2018). A unified solution for different scenarios of predicting drug-target interactions *via* triple matrix factorization. *BMC Syst. Biol.* 12, 136. doi: 10.1186/s12918-018-0663-x

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The string database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* gkw937 45 (D1), D362–D368. doi: 10.1093/nar/gkw937

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., et al. (2018). Lncipedia 5: towards a reference set of human long non-coding rnas. *Nucleic Acids Res.* 47, D135–D139. doi: 10.1093/nar/gky1031

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., et al. (2013). Mammalian ncrna-disease repository: a global view of ncrna-mediated disease network. *Cell Death Dis.* 4, e765. doi: 10.1038/cddis.2013.292

Wang, J., Cao, Y., Zhang, H., Wang, T., Tian, Q., Lu, X., et al. (2016a). Nsdna: a manually curated database of experimentally supported ncrnas associated with nervous system diseases. *Nucleic Acids Res.* 45, D902–D907. doi: 10.1093/nar/gkw1038

Wang, J., Ma, R., Ma, W., Chen, J., Yang, J., Xi, Y., et al. (2016b). Lncdisease: a sequence based bioinformatics tool for predicting lncrna-disease associations. *Nucleic Acids Res.* 44, e90–e90. doi: 10.1093/nar/gkw093

Wang, Y., Yu, G., Wang, J., Fu, G., Guo, M., and Domeniconi, C. (2019). Weighted matrix factorization on multi-relational data for lncrna-disease association prediction. *Methods*. doi: 10.1016/j.ymeth.2019.06.015

Weber, A., Schwarz, S. C., Tost, J., Trümbach, D., Winter, P., Busato, F., et al. (2018). Epigenome-wide dna methylation profiling in progressive supranuclear palsy reveals major changes at dlx1. *Nat. Commun.* 9, 2929. doi: 10.1038/s41467-018-05325-y

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217. doi: 10.1016/j.jpdc.2017.08.009

Wikberg, J. E., and Mutulis, F. (2008). Targeting melanocortin receptors: an approach to treat weight disorders and sexual dysfunction. *Nat. Rev. Drug Discovery* 7, 307. doi: 10.1038/nrd2331

Xiao, Y., Zhang, J., and Deng, L. (2017). Prediction of lncrna-protein interactions using hetesim scores based on heterogeneous networks. *Sci. Rep.* 7, 3664. doi: 10.1038/s41598-017-03986-1

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). mircancer: a microrna–cancer association database constructed by text mining on literature. *Bioinformatics* 29, 638–644. doi: 10.1093/bioinformatics/btt014

Xie, G., Wu, C., Sun, Y., Fan, Z., and Liu, J. (2019). Lpi-ibnra: Long non-coding rna-protein interaction prediction based on improved bipartite network recommender algorithm. *Front. In Genet.* 10, 343. doi: 10.3389/fgene.2019.00343

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2016). dbdemc 2.0: updated database of differentially expressed mirnas in human cancers. *Nucleic Acids Res.* 45, D812–D818. doi: 10.1093/nar/gkw1079

Yang, H., Shang, D., Xu, Y., Zhang, C., Feng, L., Sun, Z., et al. (2017). The lncrna connectivity map: using lncrna signatures to connect small molecules, lncrnas, and diseases. *Sci. Rep.* 7, 6655. doi: 10.1038/s41598-017-06897-3

Yang, P., Ormerod, J. T., Liu, W., Ma, C., Zomaya, A. Y., and Yang, J. Y. (2018). Adasampling for positive-unlabeled and label noise learning with bioinformatics applications. *IEEE Trans. cybernetics*, 49, 1–12. doi: 10.1109/TCYB.2018.2816984

Yao, B., Ma, F., Su, J., Wang, X., Zhao, X., and Yao, M. (2016). "Scale-free multiple-partite models towards information networks," in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (Hobeken, NJ, USA: IEEE press), p 549–554.

Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., et al. (2016). Raid v2. 0: an updated resource of rna-associated interactions across organisms. *Nucleic Acids Res.* 45, D115–D118. doi: 10.1093/nar/gkw1052

You, Z.-H., Huang, Z.-A., Zhu, Z., Yan, G.-Y., Li, Z.-W., Wen, Z., et al. (2017). Pbmda: a novel and effective path-based computational model for mirna-disease association prediction. *PloS Comput. Biol.* 13, e1005455. doi: 10.1371/journal.pcbi.1005455

Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks. *Briefings In Bioinf.* 17, 193–203. doi: 10.1093/bib/bbv033

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019). Predicting disease-associated circular rnas using deep forests combined with positive-unlabeled learning methods. *Briefings In Bioinf*. doi: 10.1093/bib/bbz080

Zhang, T., Wang, M., Xi, J., and Li, A. (2018a). Lpgnmf: Predicting long non-coding rna and protein interaction using graph regularized nonnegative matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinf*. doi: 10.1109/TCBB.2018.2861009

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018b). The linear neighborhood propagation method for predicting long non-coding rna–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.jpdc.2017.08.009

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018c). Sfpel-lpi: Sequence-based feature projection ensemble learning for predicting lncrna-protein interactions. *PloS Comput. Biol.* 14, e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., et al. (2018d). Deep learning in omics: a survey and guideline. *Briefings In Funct. Genomics* 18, 41–57. doi: 10.1093/bfgp/ely030

Zhang, L., Yu, G., Xia, D., and Wang, J. (2019a). Protein-protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi: 10.1016/j.compbiolchem.2019.107147

Zhang, W., Li, Z., Guo, W., Yang, W., and Huang, F. (2019b). A fast linear neighborhood similarity-based network link inference method to predict microrna-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinf*. doi: 10.1109/TCBB.2019.2931546

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., et al. (2015). Noncode 2016: an informative and valuable data source of long non-coding rnas. *Nucleic Acids Res.* 44, D203–D208. doi: 10.1093/nar/gkv1252

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018a). The bipartite network projection-recommended algorithm for predicting long non-coding rna-protein interactions. *Mol. Ther.-Nucleic Acids* 13, 464–471. doi: 10.1016/j.omtn.2018.09.020

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018b). Irwnrlpi: integrating random walk and neighborhood regularized logistic matrix factorization for lncrna-protein interaction prediction. *Front. In Genet.* 9, 239. doi: 10.3389/fgene.2018.00239

Zhao, X.-Y., Xiong, X., Liu, T., Mi, L., Peng, X., Rui, C., et al. (2018c). Long noncoding rna licensing of obesity-linked hepatic lipogenesis and nafld pathogenesis. *Nat. Commun.* 9, 2986. doi: 10.1038/s41467-018-05383-2

Zhao, X., Tang, D.-Y., Zuo, X., Zhang, T.-D., and Wang, C. (2019). Identification of lncrna–mirna–mrna regulatory network associated with epithelial ovarian cancer cisplatin-resistant. *J. Cell. Physiol.* 234 (11), 19886–19894. doi: 10.1002/jcp.28587

Zheng, X., Wang, Y., Tian, K., Zhou, J., Guan, J., Luo, L., et al. (2017). Fusing multiple protein-protein similarity networks to effectively predict lncrna-protein interactions. *BMC Bioinf.* 18, 420. doi: 10.1186/s12859-017-1819-1

Zhou, K.-R., Liu, S., Sun, W.-J., Zheng, L.-L., Zhou, H., Yang, J.-H., et al. (2016). Chipbase v2. 0: decoding transcriptional regulatory networks of non-coding rnas and protein-coding genes from chip-seq data. *Nucleic Acids Res.* gkw965 45 (D1), D43–D50. doi: 10.1093/nar/gkw965

Zhu, Y., Xu, G., Yang, Y. T., Xu, Z., Chen, X., Shi, B., et al. (2018). Postar2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* 47, D203–D211. doi: 10.1093/nar/gky830

Zhu, R., Li, G., Liu, J.-X., Dai, L.-Y., and Guo, Y. (2019). Accbn: ant-colony-clustering-based bipartite network method for predicting long non-coding rna–protein interactions. *BMC Bioinf.* 20, 16. doi: 10.1186/s12859-018-2586-3

Conflict of Interest: Authors GT and JY were employed by the company Geneis (Beijing) Co. Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 05 February 2020
doi: 10.3389/fgene.2019.01401

Evaluation of Pathway Activation for a Single Sample Toward Inflammatory Bowel Disease Classification

Xingyi Li[1], Min Li[1], Ruiqing Zheng[1], Xiang Chen[1], Ju Xiang[1,2], Fang-Xiang Wu[3] and Jianxin Wang[1]*

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] Neuroscience Research Center & Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [3] Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Meng Zhou,
Wenzhou Medical University, China
Xiangrong Liu,
Xiamen University, China
Zhi-Ping Liu,
Shandong University, China

***Correspondence:**
Min Li
limin@mail.csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 29 September 2019
Accepted: 23 December 2019
Published: 05 February 2020

Citation:
Li X, Li M, Zheng R, Chen X, Xiang J, Wu F-X and Wang J (2020) Evaluation of Pathway Activation for a Single Sample Toward Inflammatory Bowel Disease Classification. Front. Genet. 10:1401. doi: 10.3389/fgene.2019.01401

Since similar complex diseases are much alike in clinical symptoms, patients are easily misdiagnosed and mistreated. It is crucial to accurately predict the disease status and identify markers with high sensitivity and specificity for classifying similar complex diseases. Many approaches incorporating network information have been put forward to predict outcomes, but they are not robust because of their low reproducibility. Several pathway-based methods are robust and functionally interpretable. However, few methods characterize the disease-specific states of single samples from the perspective of pathways. In this study, we propose a novel framework, Pathway Activation for Single Sample (PASS), which utilizes the pathway information in a single sample way to better recognize the differences between two similar complex diseases. PASS can mainly be divided into two parts: for each pathway, the extent of perturbation of edges and the statistic difference of genes caused by a single disease sample are quantified; then, a novel method, named as an AUCpath, is applied to evaluate the pathway activation for single samples from the perspective of genes and their interactions. We have applied PASS to two main types of inflammatory bowel disease (IBD) and widely verified the characteristics of PASS. For a new patient, PASS features can be used as the indicators or potential pathway biomarkers to precisely diagnose complex diseases, discover significant features with interpretability and explore changes in the biological mechanisms of diseases.

Keywords: similar complex diseases, pathway activation, single sample, inflammatory bowel disease, pathway biomarkers

INTRODUCTION

Complex diseases threaten human health and life quality. Similar complex diseases make the early diagnosis of patients more difficult due to similar clinical symptoms. Therefore, mining effective biological information to accurately discriminate between similar complex diseases has become the most important research area of biomedicine. In the previous research, several methods based on a

single biological network, such as the metabolic network, regulatory network, or protein–protein interaction (PPI) network, have been put forward to aid in disease prediction, diagnosis, prognosis, and so on (Winter et al., 2012; Cun and Fröhlich, 2013). Nevertheless, these methods are not robust because of the low reproducibility (Yousefi and Dougherty, 2012; Amar et al., 2015; Choi et al., 2017) that results from the cellular heterogeneity within tissues, the heterogeneity of samples, and errors of measuring technologies.

Since genes generally take effect synergistically by forming functional modules, inferring features related to disease classification at the functional level can effectively ameliorate the adverse effects of heterogeneity and obtain more reproducible markers. Some methods utilize Gene Ontology (Ashburner et al., 2000) to differentiate disease states (Zhang et al., 2017) while others integrate pathway information. Pathways reflect biological processes within cells, such as metabolism, signaling, and growth cycles, and markers identified based on pathway information can thus maintain functional interpretability (Haider et al., 2018). Moreover, the occurrence and progression of complex diseases, such as inflammatory bowel disease (IBD), are often related to the dysregulation of significant pathways. Discovering the involved pathways and quantifying their disorders are of great significance in understanding complex diseases (Bild et al., 2006; Thomas et al., 2008; Markert et al., 2011; Drier et al., 2013).

A series of methods for disease classification integrate pathway information from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) or Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). Several works extract significant features from the genes along pathways to distinguish diseases (Huang et al., 2003; Bild et al., 2006; Lee et al., 2008a; Young and Craft, 2016). Although these works can combine pathway information to classify diseases effectively, they only regard a pathway as a set of genes and ignore the edge information between genes, which may lead to the loss of important information related to diseases. To overcome this problem, some methods for analyzing the intrinsic structures of pathways and integrating topological characteristics of pathways have been proposed (Liu et al., 2013; Han et al., 2017). These existing algorithms can effectively utilize the topological information of pathways to predict disease status. Nevertheless, none of them assesses condition-specific states for each patient from a pathway perspective, but this is essential to revealing the molecular mechanisms of complex diseases at the system level.

By analyzing the high-dimensional information of expression data and the differential distribution (i.e., volcano distribution) of a single patient against a given number of normal samples (Liu et al., 2016), we propose a novel framework to classify two similar complex diseases by evaluating the pathway activation based on single sample analysis. Our method consists of two steps: (1) a fully connected network for each pathway is constructed and the perturbation of each edge in the network caused by the introduction of each disease sample is evaluated. For all genes in the pathways, the statistical difference of gene expression between a single disease sample and normal samples is evaluated; (2) a novel method named as AUCpath is introduced to evaluate the pathway activation for single sample (PASS) of each pathway from both node and edge aspects, which converts the high-dimensional, small-sample gene expression matrix into a PASS matrix. Finally, a random forest classifier based on PASS features is built to examine the classification performance.

We applied PASS to classify ulcerative colitis (UC) and Crohn's disease (CD) (Ananthakrishnan, 2015). UC and CD have many common clinical features, such as abdominal pain, diarrhea, recurrent episodes, and so on. They are therefore collectively referred to as IBD. IBD is a special kind of intestinal inflammatory disease caused by common factors such as genetics, environmental triggers, immunoregulatory defects, and microbial exposure (Hanauer, 2006). Currently, there is no gold standard for discriminating UC and CD, but the responses and effects after medication of these two complex diseases are not the same (Akobeng et al., 2016; Baumgart and Sandborn, 2007), and this has motivated many attempts to understand the differences in the molecular characteristics between these two similar complex diseases at the tissue level (Lawrance et al., 2001; Burczynski et al., 2006; Wu et al., 2007). The improved understanding of the differential mechanisms of UC and CD from a molecular perspective can improve the diagnostic accuracy and have the potential to improve the therapeutic effect and the success rate of clinical trials.

We compare our method with seven network-based, GO-based, and pathway-based methods, respectively, and obtain prominent performance against these methods. In addition, our experimental results showed that our method can elucidate the molecular mechanism of UC and CD and has the potential to identify biomarkers with functional interpretability.

MATERIALS AND METHODS

Dataset and Preprocessing

We downloaded two pediatric datasets and three adult datasets from the Gene Expression Omnibus (GEO) (Edgar et al., 2002), namely GSE9686 (Carey et al., 2007), GSE3365 (Burczynski et al., 2006), GSE36807 (Montero-Meléndez et al., 2013), GSE71730 (Gurram et al., 2016), and GSE16879 (Arijs et al., 2009). All of them contain UC, DC, and normal samples.

In order to maintain the consistency of data and reduce the impact of noise, we selected data from the same anatomical site and patients under the same conditions. We excluded samples of CD patients during treatment for GSE9686 and samples of Crohn's ileitis for GSE16879. We mapped probes to gene ID using files provided by the corresponding platforms, discarded probes corresponding to multiple genes, and chose the median when multiple probes were mapped to the same gene to eliminate the influence of measurement errors. Only genes detected in all datasets can be used for the downstream analysis. As a result, there were 11242 genes included in all five datasets. **Table 1** summarizes the above datasets.

TABLE 1 | Summary of the gene expression datasets.

Name	Healthy	UC	CD	Total genes	Type of samples	Reference	URL
GSE9686	8	5	11	15747	Pediatric samples	(Carey et al., 2007)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9686
GSE3365	42	26	59	12432	Adult samples	(Burczynski et al., 2006)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3365
GSE36807	7	15	13	20486	Adult samples	(Montero-Meléndez et al., 2013)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36807
GSE71730	10	15	22	20486	Pediatric samples	(Gurram et al., 2016)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71730
GSE16879	6	24	19	20486	Adult samples	(Arijs et al., 2009)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16879

From the KEGG database, all human pathways were downloaded using the KEGGgraph package (Zhang and Wiemann, 2009). A total of 294 pathways were extracted. Each pathway consisted of a set of genes and their interactions; genes were represented by nodes, and interactions were edges in the KEGG human pathways. Genes that were not present in the expression profiles and their corresponding interactions were discarded. Considering the following analysis, pathways containing only one edge were not included. Finally, 291 pathways were retained, and these contained 3926 genes in total.

Pathway Activation for Single Sample

Pathway-based features are more robust while maintaining biological interpretability and tend to be small in number, which can prevent overfitting. In this study, we introduced a new method, called PASS, to evaluate the state of each known pathway. PASS defined the state of a pathway from the aspect of genes and regulatory links. Although it was difficult to analyze the regulatory links in the pathway for each patient, the sample-specific network (SSN) analysis provided a feasible and effective way to mine the different regulatory patterns for each patient.

In this study, we first constructed a fully connected network for each pathway. For each dataset, we analyzed the condition-specific state for each disease sample based on the pathway and thus assessed the PASS features. The schematic diagram of our framework is shown in **Figure 1**.

Statistical Difference of Edges Between Single Disease Sample and Normal Samples

For each fully connected network, we used a group of n healthy samples to calculate the Pearson correlation coefficient (PCC) of each pair of genes as background value of the corresponding edge, denoted as PCC_n. PCC_n is defined as follows:

$$PCC_n(x_1, x_2) = \frac{E(x_1x_2) - E(x_1)E(x_2)}{\sqrt{E(x_1^2) - E^2(x_1)}\sqrt{E(x_2^2) - E^2(x_2)}} \quad (1)$$

where x_1 and x_2 are the expression profiles of a pair of genes that correspond to an edge, and E represents the operator of mathematical expectation.

Next, a single disease sample was added to the set of the normal samples, and the new PCC was calculated and denoted as PCC_{n+1}. After that, the difference between background and interference values for the edges in each fully connected network could be quantified, which is represented as ΔPCC_n (equal to PCC_{n+1}–PCC_n). The difference was derived from the influence of the newly added disease sample, thus it can reflect the specific characteristics of this single sample. Statistically, ΔPCC_n obeys the volcano distribution. Therefore, the significance of ΔPCC_n can be estimated by the hypothesis test Z-test. Z-value is calculated as follows:

$$Z = \frac{\Delta PCC_n}{(1 - PCC_n^2)/(n-1)} \quad (2)$$

Statistical Difference of Gene Expressions Between Single Disease Sample and Normal Samples

The statistical difference of genes between single disease sample and normal samples in the expression level was calculated by fold change:

$$FC(x_i) = \frac{b}{\bar{a}} \quad (3)$$

where b represents the expression value of gene x_i in the individual disease sample and $\bar{a}$ is the mean of expression values of gene x_i over the n healthy samples.

Pathway Activation for a Single Sample

Based on the single sample analysis, we used AUCpath to estimate the activation of a pathway, which can evaluate the enrichment of an attention set as an area under the receiving operating characteristic curve (AUC) according to the ranking of all objects in a fully connected network. There were two sets, called the attention set and the background set. The attention set contained the subset of objects we considered as important, while the background set contained all the possible objects except important objects. We described the states of pathways from the aspect of genes and regulatory links.

From the perspective of edges, the input was the Z-value of all edges in each fully connected network, and the output was the activation of each pathway. The scoring approach was divided into two steps. First, the edges that exist in the pathway were regarded as an attention set (i.e., positive label), and the artificially added edges (in the step of the construction of fully connected network) were considered as the background set (i.e., negative label). Then, all edges in each fully connected network were ranked in ascending order of their Z-values. Second, AUC

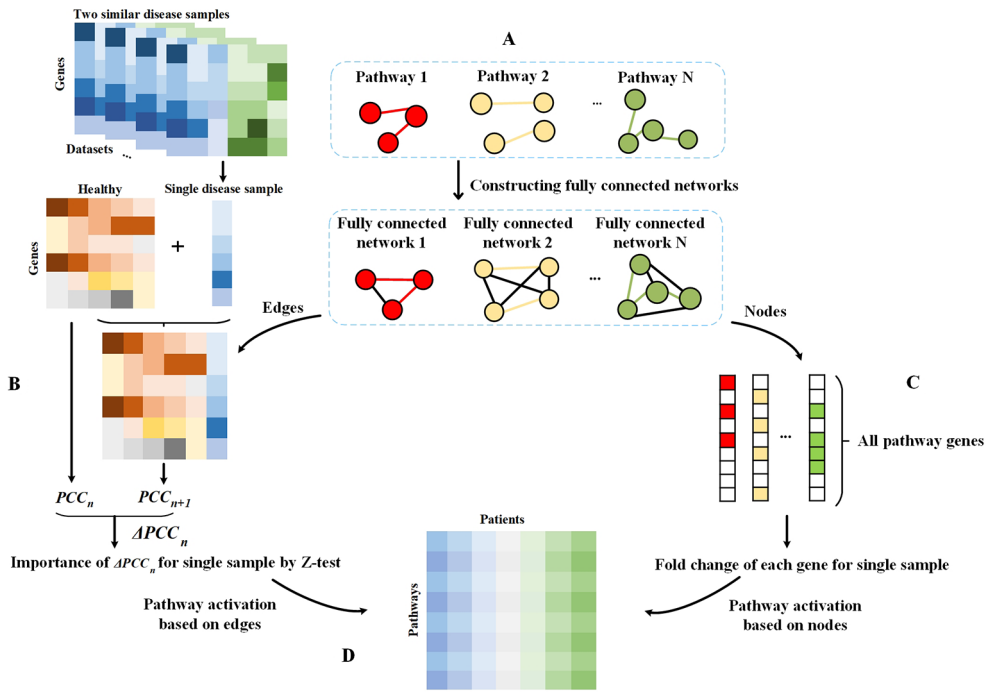


FIGURE 1 | Schematic diagram of the framework. **(A)** Fully connected networks are derived from pathways. The colored edges represent the real interactions in pathways and are regarded as attention sets, and the black edges are artificially added to construct fully connected networks and regarded as the background sets. **(B)** Single sample theory for the evaluation of differential value of each edge in fully connected networks. **(C)** Obtaining all genes in pathways. For a pathway, genes on this pathway are considered as the attention set and others are treated as the background set. The fold change value of each gene in each disease sample relative to the normal sample is evaluated for subsequent analysis. **(D)** PASS expression matrix. For each pathway, the AUCpath is used to evaluate the enrichment of edges in a pathway as an AUC according to the ranking of all edges in a fully connected network, whereby all edges were ranked according to their Z-scores. For each pathway, the AUCpath is used to assess the enrichment of genes in the pathway as an AUC based on the ranking of all pathway genes, whereby all genes were ranked according to the gene expression data.

was applied to evaluate whether edges in a pathway are enriched in the top ranking, and we thus regarded the AUC value as the quantitative indicator of pathway activation. It is defined as follows:

$$AUCpath = \frac{\sum_{i \in importantSubset} rank_i - \frac{m(1+m)}{2}}{m \times n} \quad (4)$$

where $rank_i$ represents the ranked position of the i-th edge of the attention set, m represents the number of edges in the attention set, and n is the number of edges in the background set.

Besides, considering that genes were also critical for mining effective information, we calculated the pathway activation from the perspective of genes. We first obtained all genes in pathways. For each pathway, genes on it were regarded as an attention set, and other genes were considered as the background set. Then, we assessed the enrichment of genes in the attention set as AUC based on the ranking of all genes, whereby all genes were ranked in ascending order according to their fold change between a single disease sample and normal samples.

After the evaluation of pathway activation from both nodes and edges, we obtained a matrix with PASS scores for pathways and patients.

RESULTS AND DISCUSSION

Stronger Effectiveness of PASS Compared to the Representative Feature Engineering Methods

We built a comprehensive scheme to demonstrate the performance of our approach for distinguishing two similar diseases as well as compare them with other state-of-the-art feature engineering methods. We selected seven representative methods from three aspects: network-based, GO-based and pathway-based methods, that is, NetRank (Winter et al., 2012), stSVM (Cun and Fröhlich, 2013), comparative network stratification (CNS) (Zhang et al., 2017), principal component analysis (PCA) (Young and Craft, 2016), normal tissue centroid (NTC) (Young and Craft, 2016), gene expression deviation (GED) (Young and Craft, 2016), and probabilistic pathway score (PROPS) (Han et al., 2017). For a better comparison, we downloaded the PPI network from STRING database (http://string-db.org/) for NetRank, stSVM and CNS, and collected biological processes (BP) terms of Gene Ontology (GO) (http://www.geneontology.org/) for CNS.

NetRank (Winter et al., 2012) is a modification of PageRank. For a given gene, NetRank identifies the rank of a gene according to the rank of its neighbors in a PPI network.

stSVM (Cun and Fröhlich, 2013) is a feature selection method which smooths the marginal statistic for differential expression genes by random walk kernel.

CNS (Zhang et al., 2017) is a framework that captures functional features for discriminating the disease states. Genes that are enriched by the same function (GO term) are aggregated through a flux balance model, and functional modules that maximize the distinction between UC and CD are then obtained.

For genes on each pathway, PCA (Young and Craft, 2016) compresses gene expression data and extracts principal components for the classification of disease status. For the hyperspace formed by genes on a particular pathway, NTC (Young and Craft, 2016) treats each disease sample as a point in the hyperspace and computes the Euclidean distance between the coordinates of disease samples and healthy samples. GED (Young and Craft, 2016) firstly uses the Kolmogorov–Smirnov test to capture genes that have the different distribution in normal and disease samples, and scores of those genes are then calculated based on the expression deviation in normal and disease samples. According to the scores, GED gives two features to each pathway, one for over-expression and one for under-expression. PROPS (Han et al., 2017) regards each pathway as a Gaussian Bayesian model. For each gene, after calculating the parameters in the model through normal samples, probabilistic pathway scores can be obtained using the loglikelihood values.

Improved Discrimination of PASS Evaluated by Classification Performance Analysis

We used the random forest classifier to verify the classification results and applied three-fold cross-validation considering the small sample size of several datasets. For unbiased evaluation, we repeated these experiments for a total of 500 times for the entire datasets. The results of eight methods are shown as ROC curves and AUC corresponding to the ROC in **Figure 2** and **Table 2**, respectively. Although the AUC of PROPS on GSE3365 somewhat exceeded PASS, and the AUC of PCA on GSE16879 was equal to PASS, our method was more stable and more prominent than the other seven methods on the five datasets.

Analysis of Differential Pathways With Significance According to PASS

In order to validate the effectiveness of PASS features, we analyzed the differential pathways according to the PASS index. The p-value was calculated using two-sample t-test for the five datasets. **Supplementary Figure S1** shows the quantitative distribution of p-value of differential pathways based on the PASS scores for the five datasets. The pathway activation we defined can acquire lots of differential features with significance in two similar diseases, which indicates that the PASS index can widen the gap between UC and CD.

We analyzed pathways that were differentially expressed (p-value < 0.05) on all the datasets (**Supplementary Table S1**). The majority of differential pathways have been shown to be related to IBD as reported in the literature (**Table 3**). These pathways not only demonstrate the metabolic and immune abnormalities of IBD, but they also reveal the pathogenesis of IBD from specific perspectives. Furthermore, the expression of genes in differential pathways related to IBD can reflect the changes in the course of disease. For the differential pathways associated with IBD, we analyzed the up-regulation and down-regulation of differentially expressed genes with significance in UC and normal samples, CD and normal samples. **Figure 3** shows the Venn diagrams of *Epstein-Barr virus infection* pathway, and others are shown in Supplementary **Figures S2–S10**. Most genes have the same regulatory relationship in UC and CD, but a small number of genes have different expressions. This also verifies that these two

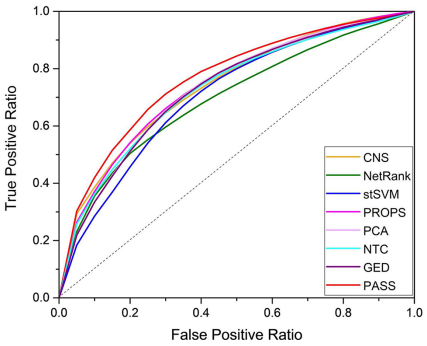


FIGURE 2 | Aggregate ROC curves.

TABLE 2 | Classification performance comparison on independent datasets.

Methods	GSE9686	GSE3365	GSE36807	GSE71730	GSE16879
PASS	0.94	0.77	0.78	0.74	0.72
NetRank	0.88	0.75	0.65	0.69	0.56
stSVM	0.88	0.72	0.75	0.71	0.55
CNS	0.91	0.75	0.75	0.70	0.69
PCA	0.91	0.66	0.73	0.69	0.72
NTC	0.89	0.72	0.75	0.67	0.68
GED	0.88	0.70	0.73	0.67	0.70
PROPS	0.88	0.78	0.70	0.73	0.67

TABLE 3 | Differential pathways related to IBD.

Entry	Name	Reference
hsa05169	Epstein-Barr virus infection	(Yanai et al., 1999)
hsa00190	Oxidative phosphorylation	(Soderholm et al., 2000; Söderholm et al., 2002)
hsa00531	Glycosaminoglycan degradation	(Lee et al., 2008b)
hsa00730	Thiamine metabolism	(Mehanna et al., 2008)
hsa00860	Porphyrin and chlorophyll metabolism	(Jansson et al., 2009)
hsa04012	ErbB signaling pathway	(Ando et al., 2013)
hsa04340	Hedgehog signaling pathway	(Ghorpade et al., 2013)
hsa04920	Adipocytokine signaling pathway	(Karmiris et al., 2006)
hsa00062	Fatty acid elongation	(Belluzzi et al., 2000)
hsa00020	Citrate cycle (TCA cycle)	(Schicho et al., 2012)

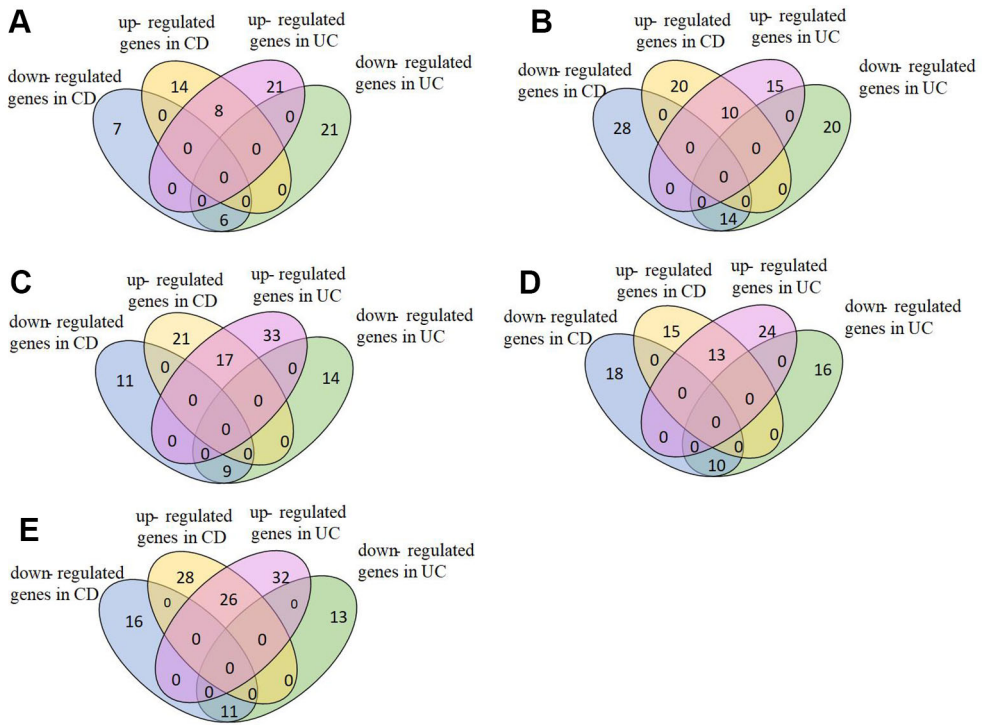


FIGURE 3 | Expression of genes in Epstein-Barr virus infection pathway. **(A)**GSE9686,**(B)**GSE3365, **(C)**GSE36807, **(D)**GSE71730, **(E)**GSE16879.

types of diseases are very similar, but there are differences between them.

Furthermore, we have visualized samples using the two principal components of our PASS features and overlaid the classification results from PASS model (**Figure 4**). The CD samples misclassified as UC and the UC samples misclassified as CD are mainly concentrated in the overlapping regions of the two types of diseases. However, some UC samples are more like CD samples, while some CD samples resemble UC samples, which leads to the misclassification of samples.

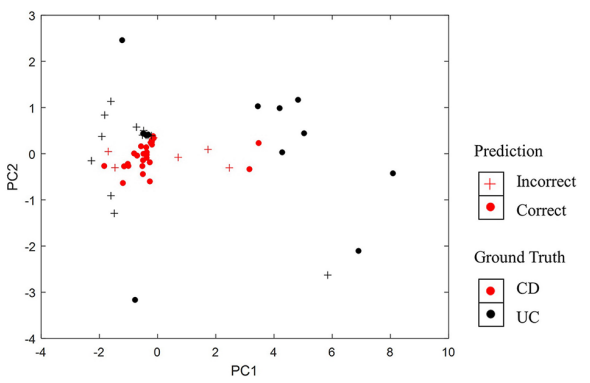


FIGURE 4 | Visualization of classification results using the two principal components of PASS features.

Enrichment of Known Disease-Associated Genes

After choosing a p-value < 0.01 as the threshold of statistical significance, we obtained the significant differential pathways. Next, we analyzed the enrichment of the known disease-associated genes (DAGs) in differential expression pathways. DAGs relevant to UC and CD were collected from DisGeNET (Piñero et al., 2016), and a hypergeometric test was used to calculate the p-value of the enrichment of DAGs:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (5)$$

where N is the number of genes in all pathways, M is the number of DAGs, n is the number of genes in the differential pathways, and m is the number of DAGs enriched in the differential pathways.

For convenience, we transformed p-value to $-log_{10}(p\text{–}value)$. We compared the statistical significance of the enrichment of DAGs in the significant differential pathways identified by PASS index with other pathway-based indexes (**Figure 5**). It shows that, with the exception of being outperformed by PROPS in GSE36807, the differential pathways obtained from PASS values have the statistical significance of the enrichment of DAGs and

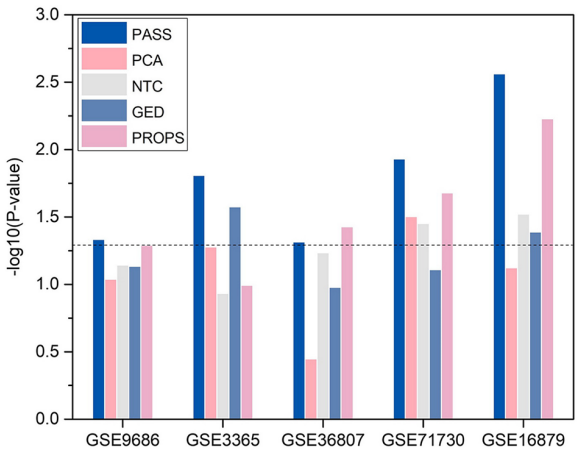


FIGURE 5 | Enrichment of known disease-associated genes in the significant differential pathways.

have lower p-values than other methods in all datasets. This indicates that the PASS index has the ability to identify differential features enriched by DAGs.

CONCLUSION

Complex diseases are not determined by a single gene, but by the combination of multiple genes, multiple factors, genetics, and the environment, similar complex diseases are more difficult to diagnose due to similar symptoms. In this study, we have presented PASS as a novel framework for classifying two main types of IBD from a single disease sample rather than a population of patients. For each pathway, we evaluated the difference between each patient and healthy sample from the perspective of genes and their interactions and calculated the pathway activation of individual samples. From the edge aspect, we constructed a fully connected network for each pathway, where edges in the pathway were regarded as the attention sets and artificially added edges were used as the background sets. Subsequently, we calculated the extent of perturbation of each edge based on single sample theory. From the node perspective, we collected all genes on all pathways. For each pathway, nodes on it were the attention set and others were the background set. Then, we evaluated the statistic difference of each node between single patient and healthy samples. Hereafter, we evaluated the pathway activation of each patient by computing the enrichment of attention set as an AUC according to the ranking of all genes or edges in the fully connected network.

We applied our method to UC and CD, which are two similar complex diseases of IBD. We compared PASS with seven state-of-the-art approaches (NetRank, stSVM, CNS, PCA, NTC, GED, and PROPS) on five IBD datasets. The results show that our PASS had the more discriminative power and was more stable than other seven methods. Besides, the PASS index can capture more differential expressed pathways with biological interpretability, which indicates that our PASS feature can widen the gap between UC and CD and aid researchers in comprehending the pathogenesis of these two similar complex diseases.

Our method can be applied to the classification of two similar diseases and has improved classification accuracy compared to seven state-of-the-art methods. However, due to the complexity and difficulty of similar complex diseases, there is still a space for improvement in the discriminative power. The performance of the PASS method relies on the all human pathway data and the topology of pathways, and more complete pathway information can better reveal the biological processes within cells and the statistic difference between a single disease sample and healthy samples calculated by our method can be also more accurate. With the rapid development of human interaction databases, we believe that the completer and more accurate pathway information could help to further improve the diagnosis of UC and CD.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at Gene Expression Omnibus (GSE9686, GSE3365, GSE36807, GSE71730, GSE16879).

AUTHOR CONTRIBUTIONS

XL and RZ conceived and designed the experiments. XL and XC performed the experiments and analyzed the data. XL wrote the paper. ML, JX, F-XW, and JW supervised the experiments and reviewed the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (61832019, 61702054), the 111 Project (No. B18059), the Hunan Provincial Innovation Foundation For Postgraduate (CX20190123), and the Hunan Provincial Natural Science Foundation of China (Grant No. 2018JJ3568).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01401/full#supplementary-material

REFERENCES

Akobeng, A. K., Zhang, D., Gordon, M., and MacDonald, J. K. (2016). Oral 5-aminosalicylic acid for maintenance of medically-induced remission in Crohn's disease. *Cochrane Database Syst. Rev.* 9, CD003715. doi: 10.1002/14651858.CD003715.pub3

Amar, D., Hait, T., Izraeli, S., and Shamir, R. (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic Acids Res.* 43, 7779–7789. doi: 10.1093/nar/gkv810

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat. Rev. Gastroenterol. Hepatol.* 12, 205. doi: 10.1038/nrgastro.2015.34

Ando, Y., Yang, G.-X., Kenny, T. P., Kawata, K., Zhang, W., Huang, W., et al. (2013). Overexpression of microRNA-21 is associated with elevated pro-inflammatory cytokines in dominant-negative TGF-β receptor type II mouse. *J. Autoimmun.* 41, 111–119. doi: 10.1016/j.jaut.2012.12.013

Arijs, I., De Hertogh, G., Lemaire, K., Quintens, R., Van Lommel, L., Van Steen, K., et al. (2009). Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PloS One* 4, e7984. doi: 10.1371/journal.pone.0007984

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25. doi: 10.1038/75556

Baumgart, D. C., and Sandborn, W. J. (2007). Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* 369, 1641–1657. doi: 10.1016/S0140-6736(07)60751-X

Belluzzi, A., Boschi, S., Brignola, C., Munarini, A., Cariani, G., and Miglio, F. (2000). Polyunsaturated fatty acids and inflammatory bowel disease. *Am. J. Clin. Nutr.* 71, 339s–342s. doi: 10.1093/ajcn/71.1.339s

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353. doi: 10.1038/nature04296

Burczynski, M. E., Peterson, R. L., Twine, N. C., Zuberek, K. A., Brodeur, B. J., Casciotti, L., et al. (2006). Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *J. Mol. Diagn.* 8, 51–61. doi: 10.2353/jmoldx.2006.050079

Carey, R., Jurickova, I., Ballard, E., Bonkowski, E., Han, X., Xu, H., et al. (2007). Activation of an IL-6: STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm. Bowel Dis.* 14, 446–457. doi: 10.1002/ibd.20342

Choi, J., Park, S., Yoon, Y., and Ahn, J. (2017). Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers. *Bioinformatics* 33, 3619–3626. doi: 10.1093/bioinformatics/btx487

Cun, Y., and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PloS One* 8, e73074. doi: 10.1371/journal.pone.0073074

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* 110, 6388–6393. doi: 10.1073/pnas.1219651110

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Ghorpade, D. S., Sinha, A. Y., Holla, S., Singh, V., and Balaji, K. N. (2013). NOD2-nitric oxide-responsive microRNA-146a activates Sonic hedgehog signaling to orchestrate inflammatory responses in murine model of inflammatory bowel disease. *J. Biol. Chem.* 288, 33037–33048. doi: 10.1074/jbc.M113.492496

Gurram, B., Salzman, N., Kaldunski, M., Jia, S., Li, B., Stephens, M., et al. (2016). Plasma-induced signatures reveal an extracellular milieu possessing an immunoregulatory bias in treatment-naive paediatric inflammatory bowel disease. *Clin. Exp. Immunol.* 184, 36–49. doi: 10.1111/cei.12753

Haider, S., Yao, C. Q., Sabine, V. S., Grzadkowski, M., Stimper, V., Starmans, M. H., et al. (2018). Pathway-based subnetworks enable cross-disease biomarker discovery. *Nat. Commun.* 9, 4746. doi: 10.1038/s41467-018-07021-3

Han, L., Maciejewski, M., Brockel, C., Gordon, W., Snapper, S. B., Korzenik, J. R., et al. (2017). A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 34, 985–993. doi: 10.1093/bioinformatics/btx651

Hanauer, S. B. (2006). Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. *Inflamm. Bowel Dis.* 12, S3–S9. doi: 10.1097/01.MIB.0000195385.19268.68

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., et al. (2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* 34, 226. doi: 10.1038/ng1167

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., et al. (2009). Metabolomics reveals metabolic biomarkers of Crohn's disease. *PloS One* 4, e6386. doi: 10.1371/journal.pone.0006386

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Karmiris, K., Koutroubakis, I. E., Xidakis, C., Polychronaki, M., Voudouri, T., and Kouroumalis, E. A. (2006). Circulating levels of leptin, adiponectin, resistin, and ghrelin in inflammatory bowel disease. *Inflamm. Bowel Dis.* 12, 100–105. doi: 10.1097/01.MIB.0000200345.38837.46

Lawrance, I. C., Fiocchi, C., and Chakravarti, S. (2001). Ulcerative colitis and Crohn's disease: distinctive gene expression profiles and novel susceptibility candidate genes. *Hum. Mol. Genet.* 10, 445–456. doi: 10.1093/hmg/10.5.445

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008a). Inferring pathway activity toward precise disease classification. *PloS Comput. Biol.* 4, e1000217. doi: 10.1371/journal.pcbi.1000217

Lee, H.-S., Han, S.-Y., Bae, E.-A., Huh, C.-S., Ahn, Y.-T., Lee, J.-H., et al. (2008b). Lactic acid bacteria inhibit proinflammatory cytokine expression and bacterial glycosaminoglycan degradation activity in dextran sulfate sodium-induced colitic mice. *Int. Immunopharmacol.* 8, 574–580. doi: 10.1016/j.intimp.2008.01.009

Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., et al. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* 29, 2169–2177. doi: 10.1093/bioinformatics/btt373

Liu, X., Wang, Y., Ji, H., Aihara, K., and Chen, L. (2016). Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res.* 44, e164–e164. doi: 10.1093/nar/gkw772

Markert, E. K., Mizuno, H., Vazquez, A., and Levine, A. J. (2011). Molecular classification of prostate cancer using curated expression signatures. *Proc. Natl. Acad. Sci.* 108, 21276–21281. doi: 10.1073/pnas.1117029108

Mehanna, H. M., Moledina, J., and Travis, J. (2008). Refeeding syndrome: what it is, and how to prevent and treat it. *Bmj* 336, 1495–1498. doi: 10.1136/bmj.a301

Montero-Meléndez, T., Llor, X., García-Planella, E., Perretti, M., and Suárez, A. (2013). Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PloS One* 8, e76235. doi: 10.1371/journal.pone.0076235

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Söderholm, J. D., Olaison, G., Peterson, K., Franzen, L., Lindmark, T., Wirén, M., et al. (2002). Augmented increase in tight junction permeability by luminal stimuli in the non-inflamed ileum of Crohn's disease. *Gut* 50, 307–313. doi: 10.1136/gut.50.3.307

Schicho, R., Shaykhutdinov, R., Ngo, J., Nazyrova, A., Schneider, C., Panaccione, R., et al. (2012). Quantitative metabolomic profiling of serum, plasma, and urine by 1H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. *J. Proteome Res.* 11, 3344–3357. doi: 10.1021/pr300139q

Soderholm, J. D., Wiren, M., Franzen, L. E., Perdue, M. H., and Olaison, G. (2000). Topical phase effects of acetylsalicylic acid on human small bowel epithelium: Inhibition of oxidative phosphorylation and increased tight junction permeability. *Gastroenterology* 118, A811. doi: 10.1016/S0016-5085(00)85386-X

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Thomas, D. C., Baurley, J. W., Brown, E. E., Figueiredo, J. C., Goldstein, A., Hazra, A., et al. (2008). Approaches to complex pathways in molecular epidemiology: summary of a special conference of the American Association for Cancer Research. *Cancer Res.* 68, 10028–10030. doi: 10.1158/0008-5472.CAN-08-1690

Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knösel, T., et al. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PloS Comput. Biol.* 8, e1002511. doi: 10.1371/journal.pcbi.1002511

Wu, F., Dassopoulos, T., Cope, L., Maitra, A., Brant, S. R., Harris, M. L., et al. (2007). Genome-wide gene expression differences in Crohn's disease and

ulcerative colitis from endoscopic pinch biopsies: insights into distinctive pathogenesis. *Inflamm. Bowel Dis.* 13, 807–821. doi: 10.1002/ibd.20110

Yanai, H., Shimizu, N., Nagasaki, S., Mitani, N., and Okita, K. (1999). Epstein-Barr virus infection of the colon with inflammatory bowel disease. *Am. J. Gastroenterol.* 94, 1582. doi: 10.1111/j.1572-0241.1999.01148.x

Young, M. R., and Craft, D. L. (2016). Pathway-informed classification system (PICS) for cancer analysis using gene expression data. *Cancer Inf.* 15, 151–161. CIN.S40088. doi: 10.4137/CIN.S40088

Yousefi, M. R., and Dougherty, E. R. (2012). Performance reproducibility index for classification. *Bioinformatics* 28, 2824–2833. doi: 10.1093/bioinformatics/bts509

Zhang, J. D., and Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* 25, 1470–1471. doi: 10.1093/bioinformatics/btp167

Zhang, C., Liu, J., Shi, Q., Zeng, T., and Chen, L. (2017). Comparative network stratification analysis for identifying functional interpretable network biomarkers. *BMC Bioinf.* 18, 48. doi: 10.1186/s12859-017-1462-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 February 2020
doi: 10.3389/fgene.2020.00083

Integrative Analysis for Identifying Co-Modules of Microbe-Disease Data by Matrix Tri-Factorization With Phylogenetic Information

Yuanyuan Ma[1]*, *Guoying Liu*[1], *Yingjun Ma*[2] and *Qianjun Chen*[2,3]

[1] School of Computer and Information Engineering, Anyang Normal University, Anyang, China, [2] School of Computer, Central China Normal University, Wuhan, China, [3] School of Life Science, Hubei University, Wuhan, China

Microbe-disease association relationship mining is drawing more and more attention due to its potential in capturing disease-related microbes. Hence, it is essential to develop new tools or algorithms to study the complex pathogenic mechanism of microbe-related diseases. However, previous research studies mainly focused on the paradigm of "one disease, one microbe," rarely investigated the cooperation and associations between microbes, diseases or microbe-disease co-modules from system level. In this study, we propose a novel two-level module identifying algorithm (MDNMF) based on nonnegative matrix tri-factorization which integrates two similarity matrices (disease and microbe similarity matrices) and one microbe-disease association matrix into the objective of MDNMF. MDNMF can identify the modules from different levels and reveal the connections between these modules. In order to improve the efficiency and effectiveness of MDNMF, we also introduce human symptoms-disease network and microbial phylogenetic distance into this model. Furthermore, we applied it to HMDAD dataset and compared it with two NMF-based methods to demonstrate its effectiveness. The experimental results show that MDNMF can obtain better performance in terms of enrichment index (EI) and the number of significantly enriched taxon sets. This demonstrates the potential of MDNMF in capturing microbial modules that have significantly biological function implications.

Keywords: microbe-disease association, matrix factorization, phylogenetic distance, human microbiome, co-modules

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Wei-Hua Chen,
Huazhong University of Science and Technology, China
Wen Zhang,
Huazhong Agricultural University, China

***Correspondence:**
Yuanyuan Ma
chonghua_1983@126.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *22 August 2019*
Accepted: *24 January 2020*
Published: *21 February 2020*

Citation:
Ma Y, Liu G, Ma Y and Chen Q (2020) Integrative Analysis for Identifying Co-Modules of Microbe-Disease Data by Matrix Tri-Factorization With Phylogenetic Information. Front. Genet. 11:83. doi: 10.3389/fgene.2020.00083

INTRODUCTION

With the development of high-throughput sequencing technology, such as 16S ribosomal RNA (16S rRNA), more and more microbes were identified. Nearly 10^{14} bacterial cells are existed in human internal gut and provide a wide variety of gene products which induce diverse metabolic activities (Micah et al., 2007; Shah et al., 2016). The dynamic balance of human microbiome composition is essential to maintain good health. Once such balance is broken, many closely related human disease and disorders may be caused (Medzhitov, 2007; Thiele et al., 2013), such as colorectal cancer (CRC)

(Boleij et al., 2014), obesity (Turnbaugh et al., 2009), inflammatory bowel disease (IBD) (Qin et al., 2010), bacterial vaginosis (Fredricks et al., 2005), and so on. For example, Jorth et al. have reported that gene expression profiles of periodontitis-related microbial communities have highly conserved changes, relative to healthy samples (Jorth et al., 2014). It means that microbiome composition changes in oral cavity could be associated with pathogenesis of periodontitis. Furthermore, Socransky et al. have found that subgingival plaque is connected with several major microbial taxon including *Fusobacterium*, *Prevotella*, and so on (Socransky et al., 1998). Chen et al. have observed that the colonization with *Helicobacter pylori* has negative correlation with the symptom of allergy (pollens and molds), especially in the childhood (Chen and Blaser, 2007; Blaser, 2014). All these reveal the potential association between pathogenic microorganisms and complex human diseases.

Considering the key role of microbes in health, many important projects including the Human Microbiome Plan (HMP) (Gevers et al., 2012), the Earth Microbiome Project (EMP) (Gilbert et al., 2010), Metagenomics of the Human Intestinal Tract (MetaHIT) (Ehrlich and Consortium, 2011) were launched to investigate the relationships between microbiota and diseases. Moreover, some related databases and tools have been developed to analyze the increasing information for disease-related microbes. A human microbe-disease association database, called HMDAD (Ma et al., 2016a), manually collected 483 microbe-disease association entries from previously published literatures. These databases provide a possibility for microbe-disease association relationship prediction by computational approaches. Zhang et al. proposed bidirection similarity integration method (BDSILP) for predicting microbe-disease associations by integrating the disease-disease semantic similarity and the microbe-microbe functional similarity. Wang et al. proposed a semisupervised computational model called LRLSHMDA to predict large-scale microbe-disease association (Wang et al., 2017). Huang et al. combined neighbor-based collaborative filtering and graph-based model into a unified objective function to predict microbe-disease relationship (Huang et al., 2017). He et al. integrated symptom-based disease similarity network into graph regularized nonnegative matrix factorization models (GRNMF), meanwhile utilizing neighbor information to boost the performance of GRNMF (He et al., 2018). Zhang et al. utilized the advantages of ensemble learning to improve the performance of association prediction, which provided a new way for mining microbe-disease relationship (Zhang et al., 2018a; Zhang et al., 2019). All these efforts pave the way for further understanding complex regulatory mechanisms by means of which disease-related microbiota get involved.

However, cellular system is complicatedly organized and biological functions are mainly performed in a highly modular manner (Barabasi and Oltvai, 2004; Chen and Zhang, 2018). In microbial ecosystems, microbes often cooperate with each other to finish some biochemical activities. For example, *ammonifiers* decompose nitrogen-containing organic compounds to release ammonia. *Nitrous acid bacteria* (also known as *ammonia oxidizing bacteria*) oxidize ammonia to nitrous acid. Then, *nitric acid bacteria* (also known as *nitrous acid oxidizing bacteria*) oxidize nitrous acid to nitric acid. These two types of bacteria can obtain the energy needed for growth from the above oxidation process. Therefore, the mutualism relationship among *ammonifier*, *nitrous acid bacteria*, and *nitric acid bacteria* forces them to form a tight biological community. Guo et al. studied the contributions of high-order metabolic interactions to the activity of four-species microbial community and demonstrated that the interactions between pairwise species play an important role in predicting the complex cellular network behavior (Guo and Boedicker, 2016). Although knowledge about microbe-disease associations could provide helpful insights into understanding complex disease mechanisms (Huang et al., 2017; He et al., 2018), the "one-disease, many microbes" models ignore interactions within microbial community composed of several species.

Recently, multilayer interaction and modular organization have attracted more and more attentions. Several studies proposed co-module discovery methods to identify combinatorial patterns using pairwise gene expression and drug response data (Kutalik et al., 2008; Chen and Zhang, 2016). In addition, Chen et al. proposed a new method based nonnegative matrix factorization (NMF) to reveal drug-gene module connections from different molecular levels (Chen and Zhang, 2018). Cai et al. proposed a new network-guided sparse binary matching model to jointly analyze the gene-drug patterns hidden in the pharmacological and genomic datasets with the additional prior information of genes and drugs (Cai et al., 2018). Chen et al. also proposed a higher order graph matching with multiple network constraints (gene network and drug network) to identify co-modules from different multiple data sources (Chen et al., 2018).

All these have made great progresses to study the coordinate regulatory mechanisms between two or more biological molecular networks from a systematic view. However, as far as we know, less work focuses on microbe-disease co-modules discovering. Previous studies mainly aimed to microbe-disease association prediction, and did not reveal within-module interactions (microbe-microbe, disease-disease) from the same level and cross-module interactions (microbe-disease) from multiple molecular levels.

To this end, we design a new algorithm based on NMF to construct the two-level microbe-disease module network by Gaussian profile kernel similarity (MDNMF). In order to improve efficiency and effectiveness of the proposed algorithm, we introduce human symptoms-disease network (Zhou et al., 2014) and microbial phylogenetic distance into this model, which makes functionally similar microbes (diseases with similar symptoms) tend to appear in the same microbial module (disease module). We applied MDNMF to HMDAD dataset and compared it with two classical NMF methods to demonstrate its effectiveness. The experimental results show that the majority of identified microbial modules have significant functional implications [significantly enriched in taxon sets that

refer to groups of microbes that has something in common (Dhariwal et al., 2017)]. **Figure 1** gives the illustrative example of MDNMF.

The contribution of this paper lies in (1) an efficient two-level module discovering algorithm (MDNMF) has been proposed to reveal microbe-microbe, disease-disease and microbe-disease modules association. (2) The phylogenetic distance of disease-related microbes is introduced into the proposed MDNMF model to make phylogenetically close microbes tend to intertwine in the development of similar disease. To our knowledge, this is the first attempt to link microbial phylogenetic relatedness to NMF-based module identification. (3) The proposed MDNMF algorithm is easily extended to other multiple-level molecular network application, for example, virus-host co-modules, microbe-drug co-modules discovering, and so on. The rest of this paper is organized as: in the next section, we give a brief overview of NMF and MDNMF. And then, followed by the experimental results and the conclusions are provided in the last section.

MATERIALS AND METHODS

Dataset

The dataset is downloaded from the Human Microbe-Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad) (Ma et al., 2016a). It contains 483 microbe-disease associations, which cover 292 microbes and 39 diseases. By 16S RNA sequencing techniques, most microbe names was recorded at the genus level. Based on these known microbe-disease relation, an adjacency matrix $X \in \mathbb{R}^{292\times 39}$ can be constructed where $X_{ij}=1$ if microbe i is related to disease j, and vice versa.

The NMF Model

NMF and its variants have been widely applied to various fields including bioinformatics (Ma et al., 2016b; Ma et al., 2017; Chen and Zhang, 2018). In NMF, given an original data matrix $X \in \mathbb{R}^{n\times m}$, we seek to find two low-rank matrices $W \in \mathbb{R}^{n\times k}$ (also called basis matrix) and $H \in \mathbb{R}^{k\times m}$ (coefficient matrix) to approximate X, such that $X \approx WH$, where $k << \min(m,n)$. Here, data X can be represented as the linear additional combination of basis vectors. We can obtain such a decomposition by solving the following least squares problem:

$$\min_{W,H\geq 0} \| X - WH \|_F^2, \tag{1}$$

where $||\bullet||_F$ denotes Frobenius norm.

Gaussian Interaction Profile Kernel Similarity for Microbes

Based on the hypothesis that functionally similar microbes could be associated with more common human diseases, Gaussian kernel interaction profiles can be used to calculate the inferred microbe similarity (Wang et al., 2017; He et al., 2018). Given microbe-disease association matrix X, the *ith* row of X indicates the interaction profiles between microbe m_i and all the diseases. For any two microbes m_i and m_j, their similarity can be computed as follows:

$$MS(m_i, m_j) = \exp\left(-\gamma_m \| X_{i,\star} - X_{j,\star} \|^2\right), \tag{2}$$

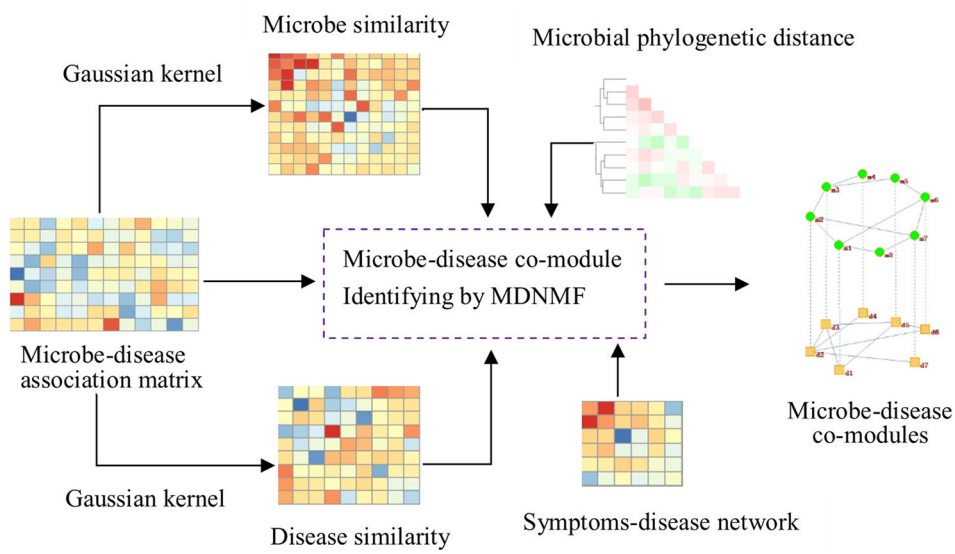


FIGURE 1 | Illustrative example of MDNMF. First, based on Gaussian kernel function we can obtain microbe and disease similarity matrices from the original microbe-disease association matrix. Then, these three matrices are served as the input of MDNMF. Simultaneously, in order to improve the accuracy of module finding and biological interpretability of modules identified by MDNMF, human symptoms-disease network and microbial phylogenetic distance are also introduced into the model. At last, microbe-disease co-modules from different levels can be obtained.

where $X_{i,*}$ denotes the *ith* row of matrix X. γ_m is bandwidth parameter that needs to be normalized based on a novel bandwidth parameter γ'_m and the interaction profile for each microbe, i.e., the *ith* row of X:

$$\gamma_m = \gamma'_m / \left(\frac{1}{n_m} \sum_{i=1}^{n_m} \| X_{i,*} \|^2 \right). \quad (3)$$

Here, n_m is the number of microbes related to all diseases (here, n_m=292). γ'_m was set as 1 according to the previous study (Wang et al., 2017). In this way, microbe similarity matrix *MS* can be constructed, the element of *MS* indicates the similarity score between two arbitrary microbes.

Gaussian Interaction Profile Kernel Similarity for Diseases

Similarly, Gaussian kernel based disease similarity matrix can be inferred as follows:

$$DS(d_i, d_j) = \exp\left(-\gamma_d \| X_{*,i} - X_{*,j} \|^2\right) \quad (4)$$

$$\gamma_d = \gamma'_d / \left(\frac{1}{n_d} \sum_{i=1}^{n_d} \| X_{*,i} \|^2 \right), \quad (5)$$

where $X_{*,i}$ denotes the *ith* column of X, n_d is the number of diseases related to all microbes (n_d=39), γ'_d was also assigned to 1.

Phylogenetic Distance for Disease-Related Microbes

Gaussian interaction profiles kernel similarity reflects the intertwining between microbes in term of microbe-disease association relationship. However, functionally similarity could not be explained only by disease relatedness, homology and phylogenetic correlation should be considered as side information to make the connected microbes in the microbe-disease association matrix likely to be placed in the same co-modules.

We searched 91 nucleotide sequences of disease-related microbes from NCBI, and imported them into MEGA to compute the phylogenetic distance between pairwise sequences by Kimura 2-parameter model. Other parameters are set in default. Thus, we can obtain the final microbial phylogenetic distance matrix M_{phy} which is used to enforce microbe members within identified modules likely to be near in phylogeny.

In order to demonstrate the role of phylogenetic information in identifying disease-related microbe modules, we extract the top 10 largest and smallest phylogenetic distance pairs as illustrative examples to further analyze whether closely related taxa tend to associate with the same disease, or similar diseases. For each microbe-microbe phylogenetic distance pair, we compute the Jaccard coefficient (JC) between two microbe-related disease profiles (rows of microbe-disease association matrix). The results shows that top 10 microbe pairs which are closely related in genetic have the largest JCs in terms of disease profile similarities. Similarly, we also compute the disease similarities between phylogenetically distant microbes and find that 9 in 10 microbe pairs have the smallest JCs. This suggests that closely related taxa tend to associate with the same disease or similar diseases, and phylogenetically distant taxa usually have distinct disease profiles.

The MDNMF Algorithm

Besides the typical NMF as *Dataset* described, tri-factor NMF (tri-NMF, $X \approx FSG$) is also an important matrix factorization method for clustering (Ding et al., 2006). In tri-NMF, factorized matrices *F*,*G* provide an approach to perform biclustering of X, respectively. Factorized matrix *s* not only provides an additional degree of freedom to enforce the reconstruct error tiny, but also implicitly denotes the relationship between clusters (Ding et al., 2005). In particular, given the symmetric similarity matrix A, we can decompose it into $A \approx HS_H{}^T$. The similarity matrix reflects the intrinsic connection patterns within its original data matrix (Van Dam et al., 2017). In this paper, we propose a novel algorithm MDNMF to simultaneously factorize two similarity matrices (microbe similarity matrix *MS*, disease similarity matrix *DS*) and one microbe-disease association matrix X. The objective function is formulated as follows:

$$\begin{aligned} \min_{H_1,\, H_2,\, S_1,\, S_2} & \| MS - H_1 S_1 H_1^T \|_F^2 + \lambda_1 \| X - H_1 H_2^T \|_F^2 + \lambda_2 \\ & \| DS - H_2 S_2 H_2^T \|_F^2 \\ & \text{s.t.} \quad H_1,\, H_2,\, S_1,\, S_2 \geq 0. \end{aligned} \quad (6)$$

where $MS \in \mathbb{R}^{n_m \times n_m}$, $DS \in \mathbb{R}^{n_d \times n_d}$ are microbe-microbe and disease-disease similarity matrices, respectively. $H_1 \in \mathbb{R}^{n_m \times k}$, $H_2 \in \mathbb{R}^{n_d \times k}$ are cluster indication matrices, $S_1 \in \mathbb{R}^{k \times k}$, $S_2 \in \mathbb{R}^{k \times k}$ are the symmetric matrices. Here, k is the number of clusters, and λ_1, λ_2 are the parameters to balance the weights of three terms in Eq.6. The second term $\| X - H_1 H_2^T \|_F^2$ establishes the one-to-one relationships between identified microbe modules and disease modules. Moreover, it can be regarded as a tri-NMF $\| X - H_1 I H_2^T \|_F^2$, here I is the identity matrix which enforce the *ith* module identified by microbe clustering indication matrix H_1 is only bound up with the *ith* module by H_2. The other two terms respectively identify one type of modules at individual levels and reveal the module associations within them *via* S_1 and S_2.

In order to further improve the performance of the proposed algorithm, we introduce symptoms-based disease similarity network and microbial phylogenetic distance into MDNMF. The symptoms-based disease similarity was previously studied based on co-occurrence of disease/symptom terms (Zhou et al., 2014). Here, we use DS_{sym} to denote symptoms-based disease similarity matrix. The objective function of MDNMF (Eq.6) can be rewritten as follows:

$$\begin{aligned} \min_{H_1,\, H_2,\, S_1,\, S_2} & \| MS - H_1 S_1 H_1^T \|_F^2 + \lambda_1 \| X - H_1 H_2^T \|_F^2 + \lambda_2 \\ & \| DS - H_2 S_2 H_2^T \|_F^2 + \mu \left(tr\left(H_1^T L_1 H_1\right) + tr\left(H_2^T L_2 H_2\right) \right) \\ & \text{s.t.} \quad H_1,\, H_2,\, S_1,\, S_2 \geq 0. \end{aligned} \quad (7)$$

Where L_1=D_1-MS_{phy}, L_2=D_2-DS_{symp} are Laplacian matrices, $(D_1)_i = \sum_j (MS_{phy})_{ij}$,$(D_2)_i = \sum_j (DS_{symp})_{ij}$ are degree matrices, respectively. MS_{phy}=1-$M_{phy}\mu$ is the regularization parameter and the whole last term in Eq.7 is used to exert a penalty for violating the prior cognition about microbial phylogeny and disease phenotype associations.

Note that disease symptoms dataset collected from PubMed literatures contains diseases and symptoms terms. The association between symptoms and diseases are quantified using term co-occurrence (just like in the field of information retrieval, if the document and keyword simultaneously appear, the corresponding position of the word-document matrix is set to the frequency of co-occurrence). And then, each disease can be represented by a vector of symptoms. At last, the cosine similarity function is used to quantify the similarity between two diseases. The link weight between two diseases quantifies the similarity of their respective symptoms. Thus, these two disease similarities based on microbes and human symptoms are different essentially in that HMDAD dataset describes the binary relationships between microbes and diseases, however, disease symptoms dataset describes the co-occurrence relationships between symptoms and diseases. Integrating them into the objective of MDNMF will simultaneously take account of the diffusion and propagation of the information from different source.

We used the multiplicative update rules to solve MDNMF problem and can find a local minimal solution by alternately updating matrices H_1, H_2, S_1, S_2.

(1) Fix H_1,H_2, S_2 and update S_1 with

$$(S_1)_{ij} \leftarrow (S_1)_{ij} \frac{\left(H_1^T M S H_1\right)_{ij}}{\left(H_1^T H_1 S_1 H_1^T H_1\right)_{ij}} \tag{8}$$

(2) Fix H_1, H_2, S_1 and update S_2 with

$$(S_2)_{ij} \leftarrow (S_2)_{ij} \frac{\left(H_2^T D S H_2\right)_{ij}}{\left(H_2^T H_2 S_2 H_2^T H_2\right)_{ij}} \tag{9}$$

(3) Fix S_1, S_2, H_2 and update H_1 with

$$(H_1)_{ij} \leftarrow (H_1)_{ij} \frac{\left(2MSH_1S_1 + \lambda_1 X H_2 + \mu D_1 H_1\right)_{ij}}{\left(2H_1 S_1 H_1^T H_1 S_1 + \lambda_1 H_1 H_2^T H_2 + \mu M S_{phy} H_1\right)_{ij}} \tag{10}$$

(4) Fix S_1, S_2, H_1 and update H_2 with

$$(H_2)_{ij} \leftarrow (H_2)_{ij} \frac{\left(2\lambda_2 DSH_2S_2 + \lambda_1 X^T H_1 + \mu D_2 H_2\right)_{ij}}{\left(2\lambda_2 H_2 S_2 H_2^T H_2 S_2 + \lambda_1 H_2 H_1^T H_1 + \mu D S_{symp} H_2\right)_{ij}} \tag{11}$$

Determination of Modules

In fact, the same microbe may play different roles in the development of diseases. Therefore, the idea of soft clustering is more suitable to model the function associations among microbes. The factorized matrices H_1, H_2 can be used to identify two types of modules, respectively. The elements with relatively large values of each column of H_1 (H_2) is assigned to the members of corresponding module. We calculate the threshold for each feature (each row $h^1_{i,*}$ of H_1 ($h^2_{i,*}$ of H_2)) with

$$Th(f) = \mu(f) + t\sigma(f), \tag{12}$$

where $\mu(f) = \frac{1}{k}\sum_k h_{fk}$, $\sigma(f) = \sqrt{\frac{1}{n-1}\sum_k (h_{fk} - \mu(f))^2}$, t is a given threshold. Based on this rule, we determined the *ith* module members if the entries of h^*_{fi} are larger than Th (f). In *Experimental Results and Discussion* section, we set t=1.5 for two clustering indication matrices H_1 and H_2 to identify modules with proper resolution.

Determination of Module Links

Given the symmetric similarity matrix A, tri-NMF factorizes it to be $A \approx HSH^T = \sum_{i=1}^k \sum_{j=1}^k s_{ij} h_i h_j^T$. Here, h_i denotes the *ith* column of H, s_{ij} is the corresponding element of s. The latent clustering indication vector h_i can reconstruct the original similarity matrix A, and s_{ij} can be viewed as the weight of $h_i h_j^T$. It means that the larger s_{ij} is, the stronger the connection between the modules identified by h_i and h_j is. Therefore, the diagonal elements of s can be used to evaluate the quality of clustering, and the off-diagonal elements can be used to establish the possible connections between different modules.

Functional Enrichment Analysis for Co-Modules

We use MicrobiomeAnalyst (Dhariwal et al., 2017) tools to conduct functional enrichment analysis for microbe modules, and select the significantly enriched taxon set terms if P-value < 0.005 and FDR < 0.05 (hypergeometric tests). Because MicrobiomeAnalyst provides 229 taxon sets associated with host-intrinsic factors such as diseases. For microbe-disease co-modules we define the enrichment indices between significantly enriched taxon set terms and diseases within the same co-module to evaluate the performance of different algorithms. The enrichment index (EI) is formulated as follows:

$$EI = \frac{|\{\textit{significantly} \text{ enriched } \textit{taxon} \text{ set}\} \cap \{\textit{diseases}\}|}{|\{\textit{significantly} \text{ enriched } \textit{taxon} \text{ set}\} \cup \{\textit{diseases}\}|}, \tag{13}$$

where $|\{$*significantly* enriched *taxon set*$\}|$ denotes the number of significantly enriched taxon sets, $|\{$diseases$\}|$ denotes the number of diseases which is related to microbes within the same co-module. Generally speaking, higher EI_s indicates good clustering quality of identified co-modules.

EXPERIMENTAL RESULTS AND DISCUSSION

Results and Comparison

We compared MDNMF with typical NMF and NetNMF (Chen and Zhang, 2018) (without considering microbial phylogenetic

information and symptoms-based disease similarity) by applying them to HMDAD dataset. Since NMF-based algorithms cannot guarantee a global optimal solution, we run 50 times with different initializations and selected the factorization with minimal objective function value as the downstream analysis.

We adopted *EI* (as described in *Functional Enrichment Analysis for Co-Modules*) and the number of significantly enriched microbe taxon set (TS_{sig}) as metrics to evaluate the performance of different algorithms. Other taxon sets (*OTS*=|{*significantly* enriched *taxon set*}|=|*identified disease-related* taxon sets|) indicate the significantly enriched taxon sets that are not considered by *EI*. To some extent, the number of other taxon sets reflects the identified ability of different methods in potential microbe function modules discovering. Extensive comparison experiments are conducted and the results are shown in **Table 1**.

TABLE 1 | The performance of three co-model discovering algorithms in term of *EI* and TS_{sig}.

	(#) identified co-modules	*EI*	(#) TS_{sig}	*OTS*
NMF	12	0.08676	39	29
NetNMF	13	0.11563	49	36
MDNMF	14	0.30182	62	48

**(P-value < 0.005 and FDR < 0.05). # represents the number of identified co-modules or significantly enriched taxon sets.*

As **Table 1** shown, compared with other two NMF-based algorithms, MDNMF achieves the best performance in terms of *EI* and TS_{sig}, indicating that MDNMF could potentially discover the meaningful function modules as much as possible by introducing symptoms-based disease network and microbe phylogenetic distance.

Comparison of All the Significantly Enriched Taxon Sets of Modules Identified by MDNMF, NMF, and NetNMF

To demonstrate the effectiveness of MDNMF, we compared the microbe modules identified by these three approaches in terms of biologically functional enrichment. We performed microbe taxon set enrichment analysis for these three groups of modules and reserved the taxon set (*TS*) terms (FDR < 0.05, hypergeometric test) which are significantly enriched by two modules derived of MDNMF and NetNMF (or NMF). Then, for each *TS* term, we calculated enrichment scores (-log10(*p-value*)) and took the highest scores among all modules as the final score of this *TS* for each method. Note that the co-modules identified by MDNMF cover about 20 microbes and 3 diseases on average. There is only one co-module which contains no

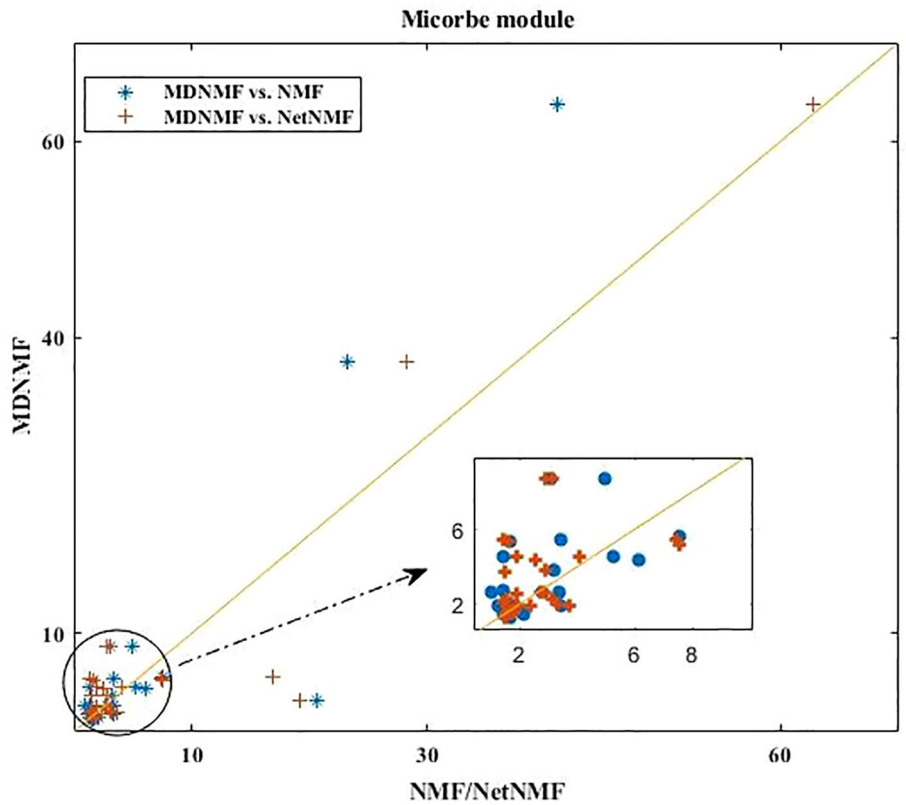


FIGURE 2 | Comparison of all the enriched *TS* terms of microbe modules detected by MDNMF, NMF, and NetNMF using HMDAD dataset.

diseases. This is consistent with the average size of each microbe or disease module (see *Parameter Analysis*).

Applying MDNMF to HMDAD dataset, many *TS* terms are above the diagonal line (see **Figure 2**). Specifically, the enriched *TS* terms obtained by MDNMF have more significant Q-value (FDR < 0.05) than those of NMF and NetNMF. For microbe modules, 58.33% (MDNMF versus NMF, P < 0.005 and FDR < 0.05, hypergeometric test) and 47.06% (MDNMF versus NetNMF, P < 0.005 and FDR < 0.05, hypergeometric test) *TS* terms are above the central diagonal line, respectively.

As **Figure 2** shown, compared to NetNMF, microbe modules identified by MDNMF had lower significance for 52.94% modules. One of the possible reasons is that when selecting microbes, NetNMF just concerns the relationships among microbes from the original microbe-disease association matrix, whereas MDNMF has to take their phylogenetic relationships into account. This kind of extra constrains of MDNMF might affect the selected microbe subsets and their enriched functions. Despite that, MDNMF still identified more significantly enriched taxon sets than NetNMF (62 vs. 49, **Table 1**).

Parameter Analysis

In MDNMF, there are three parameters:λ_1, λ_2' and μ. We set $\lambda_1 = \frac{n_m}{n_d}$, $\lambda_1 = \frac{n_m^2}{n_d^2}$ according to the previous study (Chen and Zhang, 2018). When applying these three NMF-based algorithms to HMDAD data, the reduced dimension k is needed to be pre-determined. Here, we selected k=15 from the candidate set {10,15,20}, and μ=0.001 from {0.001,0.01,0.1}, respectively. Under this setting, the number of identified microbe modules with significantly enriched taxon sets terms is highest (hypergeometric tests, P-value < 0.005 and FDR < 0.05). Mode selection is demonstrated in **Figure 3**.

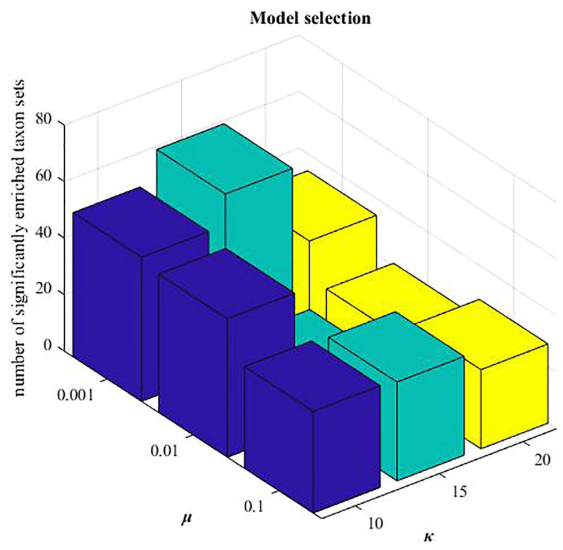


FIGURE 3 | Model selection of parameters: μ and k.

Case Studies

To further validate the performance of MDNMF, we select several microbe-disease co-modules identified by MDNMF to analyze their biological functions and inner connections. In total, 60% microbe modules are enriched in at least one *TS* term. In these identified microbe-disease co-modules, the diseases caused

TABLE 2 | The identified microbe-disease co-modules by MDNMF.

Co-module_id	Disease module	Microbe module	Taxon sets (matched disease, descending order by FDR)	Associated co-module
9	*Bacterial Vaginosis* Clostridium difficile infection (CDI) *Ileal Crohn's disease(CD)* *Irritable bowel syndrome (IBS)* *Liver cirrhosis* Necrotizing Enterocolitis Periodontal *Type 1 diabetes*	*Actinobacteria* *Bacteroidaceae* *Bacteroides* *Bacteroides uniformis* *Bacteroidetes* *Firmicutes* *Fusobacteria* *Fusobacterium* *Haemophilus* *Lachnospiraceae* *Lactobacillus* *Prevotella* *Proteobacteria* *Streptococcus* *Veillonella*	*Liver Cirrhosis* Chronic Obstructive Pulmonary Disease *Bacterial Vaginosis (increase)* Asthma Colorectal Carcinoma Resistance to Immune Checkpoint Inhibitors (increase) *Type I Diabetes* *Diarrhea Irritable Bowel Syndrome (IBS)* Parkinsons (increased) Third Trimester (vs First Trimester, increase) *Crohn's Disease*	10,4,7

** Colors indicate different diseases or enriched taxon sets.*

TABLE 3 | The detailed information of identified microbe-disease co-module 4.

Co-module_id	Disease module	Microbe module	Taxon sets (matched disease, descending order by FDR)	Associated co-module
4	*Allergic sensitization* *Constipation* *IBS* COPD *Cystic fibrosis* *Eczema* *IBD* New-onset untreated rheumatoid arthrits *Psoriasis* Rheumatoid arthrits *Ulcerative colitis*	*Acinetobacter* *Bacteroides ovatus* *Bacteroides vulgatus* *Burkholderia* *Clostridium coccoides* *Clostridium difficile* *Clostridium leptum* *Dietzia maris* *Escherichia coli* *Lysobacter*	*Cystic Fibrosis* *Atopic dermatitis* Aging (decrease) Dandruff *Crohn's Disease (increase)* Head and neck squamous cell carcinoma (increase)	9,7

**Colors indicate different diseases or enriched taxon sets.*

by microbes also exist in their matched disease modules. **Tables 2** and **3** show two of the identified microbe-disease co-modules and the associations between different disease (microbe) modules (according to S_2). As *The MDNMF Algorithm* shown, in tri-factor NMF $X \approx HS_H{}^T$, the matrix S has a special meaning. To see this, let us assume that $H^TH=I$. Setting the derivative $\partial \min \| X - HS_{H^T} \|^2 / \partial S$ to be 0, we can obtain:

$$S = H^T XH, \text{ or } S_{lk} = h_l^T X h_k = \frac{\sum_{i \in} C_l \sum_{j \in} C_k x_{ij}}{\sqrt{n_l n_k}}. \quad (14)$$

S indicates proper normalized within-cluster sum of weights ($l = k$) and between-cluster sum of weights ($l \neq k$). Therefore, S provides a good representation for the clustering quality. If the clusters are separated well, respectively the diagonal elements of S will be much larger than the off-diagonal elements. We conduct extensive experiments, and find that some off-diagonal elements are large, for example co-modules 4 and 9. According to Eq.14, this case may reflect a close connection between these two modules. The connections can provide some insights to further understand the relationships between microbe and disease, disease and disease, and microbe and microbe.

As **Table 2** shown, in co-module 9, 5 of 8 diseases (62.5%, same color from disease module and taxon sets columns indicates matched or associated disease) are in accord with significantly enriched microbe *TS* terms (FDR < 0.05). Besides, several *TS* such as "Chronic Obstructive Pulmonary Disease," "Asthma," "Colorectal Carcinoma," "Resistance to Immune Checkpoint Inhibitors (increase)" which have no matched diseases are also identified. This could provide potential associations among diseases or microbes. **Figure 4** shows top biological terms enriched in the microbe module 9.

In order to demonstrate that MDNMF can indeed cluster similar diseases to the same co-module, we retrieval each disease existed in co-module 9 from the MeSH website (https://meshb.nlm.nih.gov) and find that most of the diseases belong to the same MeSH disease category. For example, Ileal Crohn's disease (CD), Irritable bowel syndrome (IBS), Liver cirrhosis and Necrotizing enterocolitis are clustered together and they are all divided into the same MeSH disease category C06 (Digestive System Diseases). Interestingly, Clostridium infections and Bacterial vaginosis which belong to C01 (Bacterial Infections and Mycoses) are also divided into the co-module. A detailed analysis of these related diseases may yield novel insights into the more and more widely recognized the associations between microbes and human diseases.

Based on the factorized matrix s_2, we identified the connections among microbe modules 9 and 4, 7, 10. For example, microbe modules 9 and 4 share the "Crohn's Disease" and "Head and neck squamous cell carcinoma" microbe sets, but focus opposite aspects. In microbe module 9, the enriched microbe *TS* term "Crohn's Disease" is decreased, but is increased in module 4. These two microbe modules may afford us an opportunity to further investigate the complicated pathogenic mechanism in system level.

Without loss of generality, we also analyzed another microbe-disease co-module 4, the detailed information is shown in **Table 3**.

From **Table 3**, we can see that 7 of 10 diseases (70%, same color from the "disease module" and "taxon sets" columns indicates matched or associated disease) are in accord with significantly enriched microbe *TS* terms (FDR < 0.05). Especially, for enriched microbe *TS* term "Atopic dermatitis," three diseases ("Allergic sensitization," "Eczema," and "Psoriasis") in matched disease module are associated with it. This demonstrates the ability of

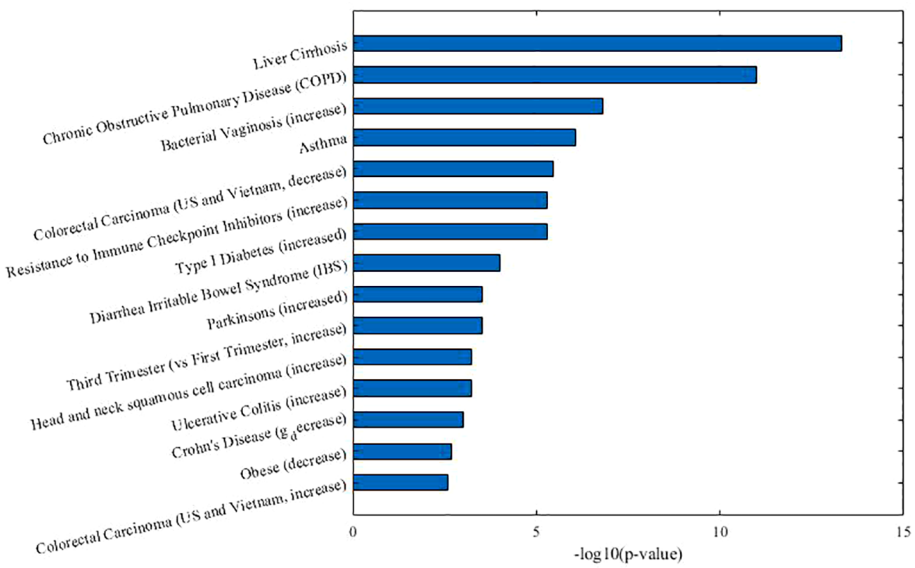


FIGURE 4 | Top enriched biological terms in microbe module 9.

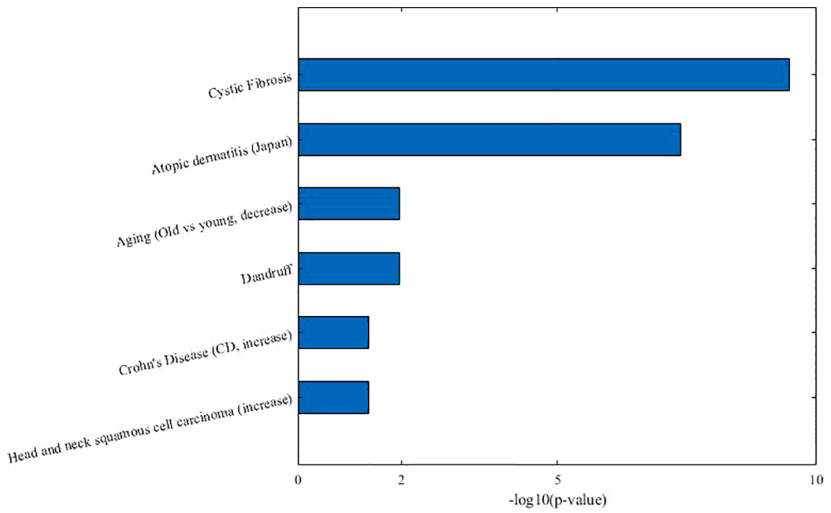


FIGURE 5 | Top enriched biological terms in microbe module 4.

the proposed MDNMF algorithm in finding correlation among diseases and microbes. **Figure 5** shows top biological terms enriched in microbe module 4.

Similarly, we retrieval each disease member in co-module 4 from the MeSH website and find that a few similar diseases belong to the same MeSH disease category. For example, Eczema, Psoriasis, Rheumatoid arthritis, and New-onset untreated rheumatoid arthritis are all from the same MeSH disease category C17 (Skin and Connective Tissue Diseases). In addition, we also find that Chronic Obstructive Pulmonary Disease (COPD), Cystic Fibrosis, Allergic sensitization, and Intestinal diseases (IBS, Irritable bowel disease, and Ulcerative colitis) have also been clustered together. Several diseases belong to two or more MeSH categories, which indicates the pathological connections between the human genetic susceptibility to infectious diseases and inflammatory diseases.

Based on factorized matrix s_2, we can find that co-module 4 has more links to co-module 7($s_{4,7}$=2.72). Matched disease modules 4 and 7 own the similar disease members, such as "Allergic sensitization" (from module 4) and "Asthma" (from module 7) induced by "Atopic dermatitis." Besides, two corresponding microbe modules 4 and 7 share *TS* term "Aging."

Note that in **Tables 2** and **3** some related diseases and microbes are divided into different co-modules. One possible of reasons is that the connection weight between these co-modules is large, MDNMF as a soft clustering approach, cannot well separately these related microbes or disease. In the future, we will design more robust threshold selecting method to assign each diseases or microbes to accurate modules.

In summary, for the identified module pairs by MDNMF, especially for microbe modules, some of them share a few biological functions (*TS*), but also have their special roles. Simultaneously, some associations between microbe modules, disease modules can be also detected by MDNMF.

CONCLUSIONS

The association between microbes and human diseases has been verified by more and more researches. However, previous studies mainly focused on detecting the relationship such as "one microbe, one disease," rarely analyzed the pathogenesis of microbial-related complex diseases from a modular perspective. In this paper, we propose a novel microbe-disease co-module detecting algorithm MDNMF to construct a two-level module network by integrating two similarity matrices (microbe-microbe, disease-disease similarity matrices) and one microbe-disease bipartite network. Using the identified individual modules from different levels (microbe, disease levels) and their links, we are able to find a few disease-related microbes (taxon sets) which provide an opportunity to further understand the microbe high-order relationship and their potential functions.

Meanwhile, in order to improve the accuracy of module finding and biological interpretability of modules identified by MDNMF, we introduce human symptoms-disease network and microbial phylogenetic distance into the model. Compared with other two NMF-based approaches, MDNMF can achieve better performance in terms of *EI* and the number of significantly enriched taxon sets. The proposed MDNMF is also easily extended to other multiple-level molecular network application, for example, virus-host co-modules, microbe-drug co-modules discovering, and so on.

DATA AVAILABILITY STATEMENT

The data and MDNMF codes analyzed during the study are available in the GitHub repository, https://github.com/chonghua-1983/MDNMF.

AUTHOR CONTRIBUTIONS

YuM wrote the manuscript and developed the algorithms. YuM and GL developed the concept for the structure and content of the manuscript. YiM wrote the code used in the paper. QC critically revised the manuscript in final. All authors reviewed and approved the final version of the manuscript.

FUNDING

This work has been accepted by CBC2019, supported by the National Natural Science Foundation of China (No.61532008), Key Research Projects of Henan Higher Education Institutions (No. 20B520002), The Key Technology R & D Program of Henan Province (202102310561).

REFERENCES

Barabasi, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113. doi: 10.1038/nrg1272

Blaser, M. (2014). *Missing microbes* (United Kingdom and Commonwealth: Oneworld Publications).

Boleij, A., Hechenbleikner, E. M., Goodwin, A. C., Badani, R., Stein, E. M., Lazarev, M. G., et al. (2014). The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* 60, 208–215. doi: 10.1093/cid/ciu787

Cai, J., Cai, H., Chen, J., and Yang, X. (2018). Identifying "Many-to-Many" relationships between gene-expression data and drug-response data *via* sparse binary matching. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2018.2849708

Chen, Y., and Blaser, M. J. (2007). Inverse associations of Helicobacter pylori with asthma and allergy. *Arch. Internal Med.* 167, 821–827. doi: 10.1001/archinte.167.8.821

Chen, J., and Zhang, S. (2016). Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. *Bioinformatics* 32, 1724–1732. doi: 10.1093/bioinformatics/btw059

Chen, J., and Zhang, S. (2018). Discovery of two-level modular organization from matched genomic data *via* joint matrix tri-factorization. *Nucleic Acids Res.* 46, 5967–5976. doi: 10.1093/nar/gky440

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2018). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35, 602–610. doi: 10.1093/bioinformatics/bty662

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45, W180–W188. doi: 10.1093/nar/gkx295

Ding, C., He, X., and Simon, H. D. (2005). "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM International Conference on Data Mining* (SIAM), USA, 606–610.

Ding, C., Li, T., Peng, W., and Park, H. (2006). "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (ACM), New York, NY , 126–135.

Ehrlich, S. D., and Consortium, M. (2011). "MetaHIT: The European Union Project on metagenomics of the human intestinal tract," in *Metagenomics of the human body.* (New York, NY: Springer), 307–316.

Fredricks, D. N., Fiedler, T. L., and Marrazzo, J. M. (2005). Molecular identification of bacteria associated with bacterial vaginosis. *New Engl. J. Med.* 353, 1899–1911. doi: 10.1056/NEJMoa043802

Gevers, D., Knight, R., Petrosino, J. F., Huang, K., Mcguire, A. L., Birren, B. W., et al. (2012). The Human microbiome project: a community resource for the healthy human microbiome. *PloS Biol.* 10, e1001377. doi: 10.1371/journal.pbio.1001377

Gilbert, J. A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C. T., Brown, C. T., et al. (2010). Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Standards In Genomic Sci.* 3, 243–248. doi: 10.4056/sigs.1433550

Guo, X., and Boedicker, J. Q. (2016). The contribution of high-order metabolic interactions to the global activity of a four-species microbial community. *PloS Comput. Biol.* 12, e1005079. doi: 10.1371/journal.pcbi.1005079

He, B.-S., Peng, L.-H., and Li, Z. (2018). Human microbe-disease association prediction with graph regularized non-negative matrix factorization. *Front. Microbiol.* 9. doi: 10.3389/fmicb.2018.02560

Huang, Y.-A., You, Z.-H., Chen, X., Huang, Z.-A., Zhang, S., and Yan, G.-Y. (2017). Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Trans. Med.* 15, 209. doi: 10.1186/s12967-017-1304-7

Jorth, P., Turner, K. H., Gumus, P., Nizam, N., Buduneli, N., and Whiteley, M. (2014). Metatranscriptomics of the human oral microbiome during health and disease. *MBio* 5, e01012–e01014. doi: 10.1128/mBio.01012-14

Kutalik, Z., Beckmann, J. S., and Bergmann, S. (2008). A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.* 26, 531–539. doi: 10.1038/nbt1397

Ma, W., Zhang, L., Zeng, P., Huang, C., Li, J., Geng, B., et al. (2016a). An analysis of human microbe–disease associations. *Brief. Bioinf.* 18, 85–97. doi: 10.1093/bib/bbw005

Ma, Y., Hu, X., He, T., and Jiang, X. (2016b). Hessian regularization based symmetric nonnegative matrix factorization for clustering gene expression and microbiome data. *Methods* 111, 80–84. doi: 10.1016/j.ymeth.2016.06.017

Ma, Y., Hu, X., He, T., and Jiang, X. (2017). Clustering and integrating of heterogeneous microbiome data by joint symmetric nonnegative matrix factorization with laplacian regularization. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2017.2756628

Medzhitov, R. (2007). Recognition of microorganisms and activation of the immune response. *Nature* 449, 819–826. doi: 10.1038/nature06246

Micah, H., Claire, F.-L., and Rob, K. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* 449, 804–810. doi: 10.1038/nature06244

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Shah, P., Fritz, J. V., Glaab, E., Desai, M. S., Greenhalgh, K., Frachet, A., et al. (2016). A microfluidics-based *in vitro* model of the gastrointestinal human–microbe interface. *Nat. Commun.* 7, 11535. doi: 10.1038/ncomms11535

Socransky, S., Haffajee, A., Cugini, M., Smith, C., and Kent, R.Jr. (1998). Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* 25, 134–144. doi: 10.1111/j.1600-051X.1998.tb02419.x

Thiele, I., Heinken, A., and Fleming, R. M. (2013). A systems biology approach to studying the role of microbes in human health. *Curr. Opin. Biotechnol.* 24, 4–12. doi: 10.1016/j.copbio.2012.10.001

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Van Dam, S., Vosa, U., Van Der Graaf, A., Franke, L., and De Magalhaes, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinf.* 19, 575–592. doi: 10.1093/bib/bbw139

Wang, F., Huang, Z.-A., Chen, X., Zhu, Z., Wen, Z., Zhao, J., et al. (2017). LRLSHMDA: laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* 7, 7601. doi: 10.1038/s41598-017-08127-2

Zhang, W., Yang, W., Lu, X., Huang, F., and Luo, F. (2018a). The bi-direction similarity integration method for predicting microbe-disease associations. *IEEE Access* 6, 38052–38061. doi: 10.1109/ACCESS.2018.2851751

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018b). SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PloS Comput. Biol.* 14, e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., et al. (2019). SFLLN: A sparse feature learning ensemble method with linear neighborhood regularization for predicting drug–drug interactions. *Inf. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhou, X., Menche, J., Barabási, A.-L., and Sharma, A. (2014). Human symptoms–disease network. *Nat. Commun.* 5, 4212. doi: 10.1038/ncomms5212

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 February 2020
doi: 10.3389/fgene.2020.00160

Predicting Stage-Specific Recurrent Aberrations From Somatic Copy Number Dataset

Chaima Aouiche[1†], *Bolin Chen*[1,2,3*†] *and Xuequn Shang*[1,2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [2] Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Xi'an, China, [3] Centre for Multidisciplinary Convergence Computing, School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Exploring the evolution process of cancers and its related complex molecular mechanisms at the genomic level through pathological staging angle is particularly important for providing novel therapeutic strategies most relevant to every cancer patient diagnosed at each stage. This is because the genomic level involving copy number variation (CNV) has been recognized as a critical genetic variation, which has a large influence on the progression of a variety of complex diseases. Great efforts have been devoted to the identification of recurrent aberrations, single genes and individual static pathways related to cancer progression. However, we still have little knowledge about the most important aberrant genes related to the pathology stages and their interconnected pathways from genomic profiles. In this study, we propose an identification framework that allows determining cancer-stages specific patterns dynamically. Firstly, a two-stage GAIA method is employed to identify stage-specific aberrant copy number variants segments. Secondly, stage-specific cancer genes fully located within the aberrant segments are then identified according to the reference annotation dataset. Thirdly, a pathway evolution network is constructed based on the impacted pathways functions and their overlapped genes. The involved significant functions and evolution paths uncovered by this network enabled investigation of the real progression of cancers, and thus facilitated the determination of appropriate clinical settings that will help to assess risk in cancer patients. Those findings at individual levels can be integrated to identify robust biomarkers in cancer progressions.

Keywords: cancer evolution, somatic copy number alteration, aberrant genes, pathological stages, pathway interaction network

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Binhua Tang,
Hohai University, China
Andrew Dellinger,
Elon University, United States

***Correspondence:**
Bolin Chen
blchen@nwpu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *23 October 2019*
Accepted: *11 February 2020*
Published: *26 February 2020*

Citation:
Aouiche C, Chen B and Shang X (2020) Predicting Stage-Specific Recurrent Aberrations From Somatic Copy Number Dataset. Front. Genet. 11:160. doi: 10.3389/fgene.2020.00160

1. INTRODUCTION

Somatic copy number alterations (SCNAs) are one of the prevalent forms of genetic variations which play important roles in the progression of numerous diseases, such as cancers (Zack et al., 2013; Heitzer et al., 2016). SCNAs have much clinical relevance compared to other genetic alterations, and they can be good markers of cancer genome aggressiveness (Heitzer et al., 2016). Hence, the identification of specific signatures from CNAs will shed light on elucidating the complex mechanisms behind cancers evolution, and therefore lead to a promotive development in cancer treatment strategies (Lowe et al., 1994; Tsao et al., 2005; Kim et al., 2008; Cheang et al., 2009).

The evolution of cancers involves many complex and dynamic cellular processes that can be precisely described through pathological stages, which are often divided into several stages, from the initial stage to the later deleterious stage. Where cancers at early appearance (stage I or II) are typically viewed as treatable; however, many more aggressive and active therapies would be needed as they developed to harmful stages (stage III or IV). Thus, there was a critical need toward the extraction of reliable biomarkers characterizing the dynamics associated with these stages, including (1) stage-specific recurrent SCNAs, (2) their related aberrant genes, and (3) their enriched dysfunctional pathways (Chen et al., 2009, 2010; Lee et al., 2016; Liang et al., 2016; Wang et al., 2016; Nibourel et al., 2017; Zhu et al., 2017).

Recent developments on high-throughput genomic technologies have generated diverse tumor datasets with various clinical/pathological stages, conditions or tissues, for which CNAs and other omics-data have been collected. They provide effective ways to identify different biological patterns including individual genes, pathways, specific loci and individual chromosomal regions. However, the majority of these proposed ways completely ignore the topology and the interaction between these patterns, as well as their specificity along with the pathology stages. Since specific genes and pathways extracted from these stages across different regions will often act together in complex systems (Karczewski and Snyder, 2018; Ma et al., 2018), whose dynamic events are the results of multiple complex interactions that help to extract useful dynamic cellular functions, and that can well illustrate the progression and metastasis of cancers.

Fortunately, the usage of biological networks/pathways has turned out to be an effective method to describe the details of the dynamic changes and functional mechanisms associated with the individual stages of cancers, where individual nodes represent biological entities, i.e., genes or pathways, and each edge corresponds to an interaction between a pair of nodes. Those biological networks include but not limited to cellular pathways, gene regulation networks (Vaquerizas et al., 2009), protein-protein interaction networks (Schwikowski et al., 2000), and many disease related networks (Menche et al., 2015). Such networks can be efficiently used to investigate the dynamic biological activity behind cancers evolution.

A suite of well-established algorithms has also been proposed at the chromosome level to accurately detect recurrent SCNAs (Morganella et al., 2011), to investigate multiple cancer stages (Xia et al., 2004), or to use gene expressions to analyze the evolution processes of cancers.

To further extend the study to individual cancer stages, we propose an analysis framework to elucidate the dynamic evolution processes of cancers. Firstly, the recurrent aberrations associated with cancer-specific stages were discerned through (a) the identification of occurring sequential changes moving from stage I to stage IV and (b) the determination of correlations between higher frequency of CNA and the higher aggressive stage. Secondly, the stage-specific cancer related genes were carefully detected via the obtained CNV information. Thirdly, the stage-specific pathways were extracted and a pathway interaction network was generated by connecting functional pathways in adjacent stages. The remainder of the paper includes three sections: section 2 discusses the data sources and the methodology used in the identification framework, section 3 reports the results, and section 4 provides the conclusion of the study.

TABLE 1 | The clinical and CNV datasets information from Broad Firehose TCGA project.

Pathology stages	Clinical samples	CNV samples
Pathology_t_1	9	1,255
Pathology_t_2	46	9,232
Pathology_t_3	145	32,293
Pathology_t_4	19	4,360

2. MATERIALS AND METHODS

2.1. Data Collection and Grouping

Clinical and Somatic copy number alteration (by SNP 6.0 array) datasets on Level3 colorectal cancer (COADREAD) were downloaded from the Broad GDAC Firehose[1].

Somatic copy number alteration (SCNA) minus germline SCNA was produced using GISTIC 2.0 and then divided into four groups based on the available clinical information of the same group of clinical patients. From clinical data, we take only the patients with available "pathology t stage" information, which defines the diagnosis stage of individual samples (t_1, t_2, t_3, and t_4). For the sample collection, we count the number of patients in the four t stages. Those individual samples with pathological information were aligned to the corresponding SCNA samples to get their copy number information for our following analysis.

Finally, 219 samples ($t_1 = 9$, $t_2 = 46$, $t_3 = 145$, and $t_4 = 19$) retained from clinical data were mapped to 47,140 samples from SCNA data ($t_1 = 1{,}255$, $t_2 = 9{,}232$, $t_3 = 32{,}293$, and $t_4 = 4{,}360$), respectively, and used to conduct our subsequent analysis. These details are shown in **Table 1**.

In addition, for recurrent CNAs identification from pre-computed GISTIC 2.0 SCNA data, GAIA (Morganella et al., 2011) with FDR $Q < 0.10$ was applied separately for each pathology stage using ten iterations. For genomic SCNA gains and losses plotting, an R script was used with a cut-off also specified at FDR $Q < 0.10$. For the genes annotation of the recurrent SCNA regions, the biomaRt (Durinck et al., 2005) and GenomicRanges (Lawrence et al., 2013) packages available through Bioconductor of R Studio were considered.

For the network construction, pathways were extracted from the Reactome database[2]. Since pathways with a smaller number of genes may lack significant biological knowledge, we collected, in this study, a set of pathways by filtering those with five genes. We ended up with 447 impacted pathways.

[1] http://firebrowse.org/
[2] http://www.reactome.org

A

	P1	P2	P3	P4	P5	P6	P7
Sample 1	0	+	0	-	-	-	0
Sample 2	0	-	+	-	-	-	-
Sample 3	0	+	+	+	-	-	-

B

	P1	P2	P3	P4	P5	P6	P7
Sample 1	0	1	0	0	0	0	0
Sample 2	0	0	1	0	0	0	0
Sample 3	0	1	1	1	0	0	0

C

	P1	P2	P3	P4	P5	P6	P7
Sample 1	0	0	0	1	1	1	0
Sample 2	0	1	0	1	1	1	1
Sample 3	0	0	0	0	1	1	1

FIGURE 1 | GAIA illustrative example. **(A)** Represent an example of matrix A, where + denotes gain, − denotes loss and 0 denotes no alteration. **(A)** Contain two homogeneous regions from probes P4 to P6 for samples S1 and S2 and from probes P5 to P7 for samples S2 and S3. **(B,C)** Show the matrices AL and AD determined of the matrix in **(A)**.

2.2. Stage-Specific Related Recurrent Somatic Copy Number Alteration Regions Identification

To identify the recurrent SCNA for the series of the pathological stages separately, a two-stage GAIA (genomic analysis of important aberrations) method (Morganella et al., 2011) was performed to determine the most significant recurrent CNA for the four pathology stages. In particular, this method follows two main steps: Significance testing and Homogeneous peel-off, to identify the most significant independent regions where a discrete representation of data is mainly considered.

Based on that, we first build a CNV matrix of regions using probes meta file from GISTIC 2.0 (available at[3]). Then, we define the recurrent CNA by FDR $Q < 0.10$ using ten iterations. Finally, we generate the genomic plots of the four stages using a GAIA plot function in R Studio, with the cut-off set also to FDR $Q < 0.10$.

Suppose there is a set of N samples (patients) and M observed probes, the data can be arranged as an $N \times M$ dimension matrix A. As an illustrative example (**Figure 1**), A can represent a chromosome of seven observed probes and three samples. The matrix A can be split into two matrices AL and AD where each element $a\_ij \in AL(AD)$ $i = 1, \ldots, N$ and $j = 1, \ldots, M$ can be denoted either by 1 as a gain (or loss) found in the j_th marker of the i_th sample, or by 0 otherwise as shown in **Figures 1B,C**, which represents the matrices AL and AD determined from the matrix A reported in **Figure 1A**. Three major steps can be applied to this matrix (gain or loss interest) to identify the significant peaks and omit the spurious peaks in a region based on q-values configuration, h-values calculation and multiple iterations. More details are described here and depicted in **Figure 2**.

First, a permutation test is performed on every individual marker to compute the probability distribution, so that we can estimate the statistical significance of the observed data.

Second, in order to define the homogeneous regions, we focus on the state of every paired adjacent markers (j and $j + 1$) rather than a single marker, and we calculate the degree of homogeneity between them. Given a matrix H of size $(N \times M - 1)$, with an element H_{ij} that has the value of 0 for maximum homogeneity, or the value of 0.5 for a medium homogeneity, or the value of 1 for a minimum homogeneity. From this matrix, we can obtain overall information on the homogeneity of the dataset based on the (h-value) that can be computed as follow:

$$h_j = \frac{1}{N} \sum_{i=1}^{N} H_{ij}, \quad j = 1, \ldots, M - 1 \tag{1}$$

Third, an iterative peel-off procedure is carried out on the matrix H by expanding the left and right boundaries of the region until the following conditions are satisfied. The left boundary

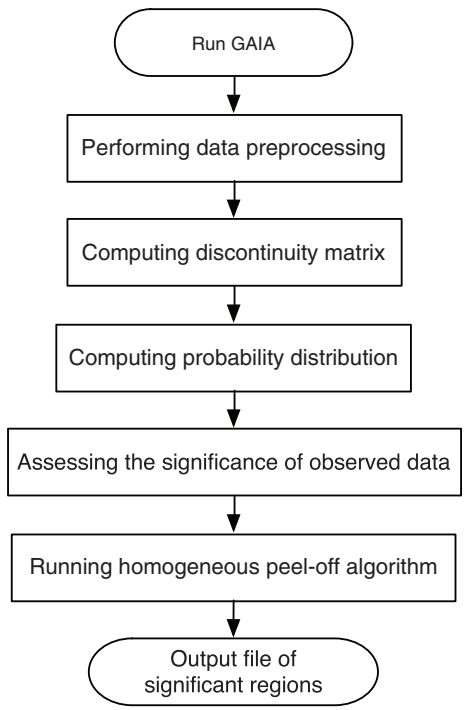


FIGURE 2 | The flow chart of the GAIA method implementation steps.

[3]ftp://ftp.broadinstitute.org/pub/GISTIC2.0/hg19support/

expanded if:

$$q_{l-1} \leq q_{thr} \quad \text{AND} \quad h_{l-1} \leq h_{thr} \tag{2}$$

and the right boundary expanded if:

$$q_{r+1} \leq q_{thr} \quad \text{AND} \quad h_r \leq h_{thr} \tag{3}$$

where l and r denote the left and the right boundary of the peak with minimum q-value, with $1 \leq l, r \leq M$, while h_thr represents a significance threshold value for homogeneity measurement. This value can take 0, 1, or values between 0 and 1.

Remarkably, large recurrent SCNAs have been produced in this study at different chromosome positions moving from pathology_t_1 to pathology_t_4. More details are shown in **Figures 3–6**, respectively, which summarize the frequencies of the four pathology stages.

2.3. Stage-Specific Related Aberrant Genes Identification

The second essential step allowing a comprehensive elucidation of the cancer evolution process after SCNA regions identification is to identify the corresponding signature genes for individual stages. Therefore, the aberrant recurrent regions obtained previously at every pathology stage were then annotated to retrieve the genes that were significantly amplified or deleted. Using the reference annotation dataset of genes of biomaRt (Durinck et al., 2005), the final set of genes at *cut-off* = 0.10 with the precise co-ordinates regions from human genes in which it was found to have CNA, have been obtained. Further details are shown in **Table 2**, which lists the total number of genes selected in the four pathology stages.

2.4. Stage-Related Pathways Extraction

After obtaining the deviant amplified or deleted genes at every pathology stage, the given genes were aligned to pathways on the basis of the biological pathways in the Reactome database from which a total of 3,305 was collected. The pathways found include clusters of pathways from different pathologies: 396 pathways from pathology_t_1, 895 pathways from pathology_t_2, 1,218 pathways from pathology_t_3, and 796 pathways from pathology_t_4.

As long as a single gene can be assigned to different pathways, and the latter would consist of a different number of genes, we set the study sample to every pathology's pathways consisting of genes whose size is >5. This is due to the fact that pathways with fewer genes would have limited biological content (Ahn et al., 2014). Therefore, a total of 656 pathways ($t_1 = 5, t_2 = 110, t_3 = 447, t_4 = 94$) was collected (**Table 2**). Finally, duplicated pathways were omitted, and only pathways that occurred in at least two pathological stages were extracted and considered as our stage-specific pathways to be further analyzed.

2.5. Pathway Evolution Network Construction

After identifying the signature genes for each stage and after extracting and integrating their specific Reactome pathways, they are pooled together, their terms are unified, and their official annotated pathway descriptions are obtained from the database. Next, a pathway interaction network related to SCNA is constructed where each node represents a biological specific pathway, and if the two pathways share common genes, then they are connected.

To clearly illustrate the dynamic evolution process through this pathway network, specific colors were used to evince the pathways that get evolved between the four individual stages, and the width of edges is applied to indicate the strength of associations between them. The width was calculated using an overlap score defined as:

$$W = \frac{k^2}{p * q}. \tag{4}$$

where k represents the number of the overlapping genes between a pair of pathway P_i and pathway P_j, p and q stand for the total numbers of genes in P_i and P_j, respectively.

3. RESULTS AND DISCUSSIONS

3.1. Stage-Related Recurrent Genome-Wide SCNAs Frequencies

The recurrent CNAs from four pathology stages were identified by investigating the sequential changes from pathology_t_1 to pathology_t_4 according to their different frequencies. This is based on the assumption that higher frequency of CNA will correlate with higher cancer stages. In fact, large genomic differences in recurrent SCNAs were observed in each pathology stage. **Figures 3–6** represent the genome-wide amplifications and deletions of the four pathology stages, which generated with cut-off defined at FDR $Q < 0.10$. To be more specific, there were no significant segments in stage 1, but for stage 2, stage 3, and stage 4, the most of their regions were significantly amplified or deleted.

Moreover, more aberrant chromosomes get involved in these three stages. The frequency of aberrant segments were higher in stage 2 than in stage 1, and it kept increasing in stage 3. For example, stage 1 involved only three abnormal chromosomes with very low frequency. However, stage 2 and stage 3 involved more abnormal chromosomes segments with higher frequencies of amplifications or deletions. A clear evolution process of cancer could be observed by connecting those major chromosomal abnormalities stage-by-stage.

3.2. The Number of Stage-Specific Related Genes

The amplified and deleted genes which fully located within the aberrant regions of the four pathological stages were detected by using the biomaRt and the GenomicRanges packages in R (**Table 2**), wherein a total of 423, 3,265, 8,500, and 2,244 genes were identified as representative signature genes in stage 1, 2, 3, and 4, respectively. All of these potential candidate

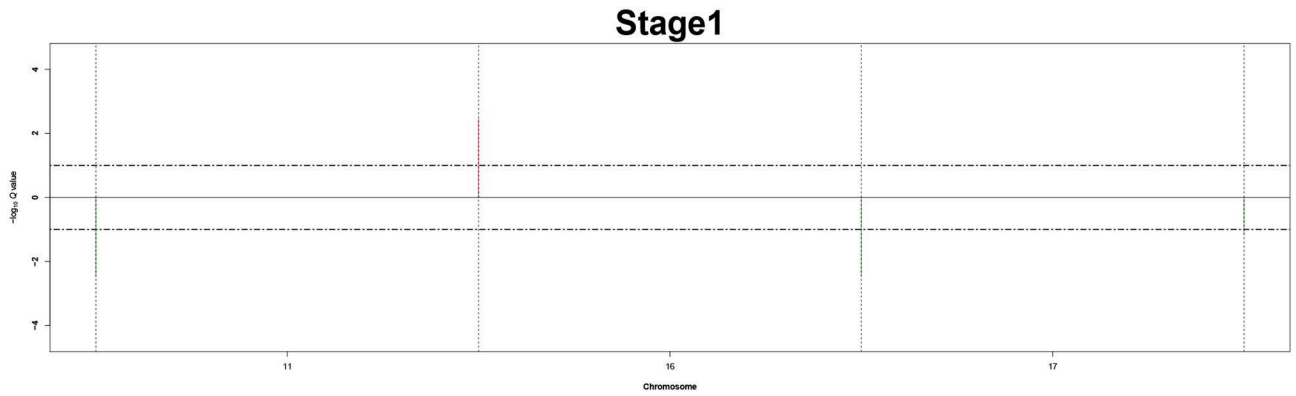


FIGURE 3 | Recurrent genome-wide SCNAs in stage 1. Genome-wide amplifications (red blocks) and deletions (green blocks) in stage 1.

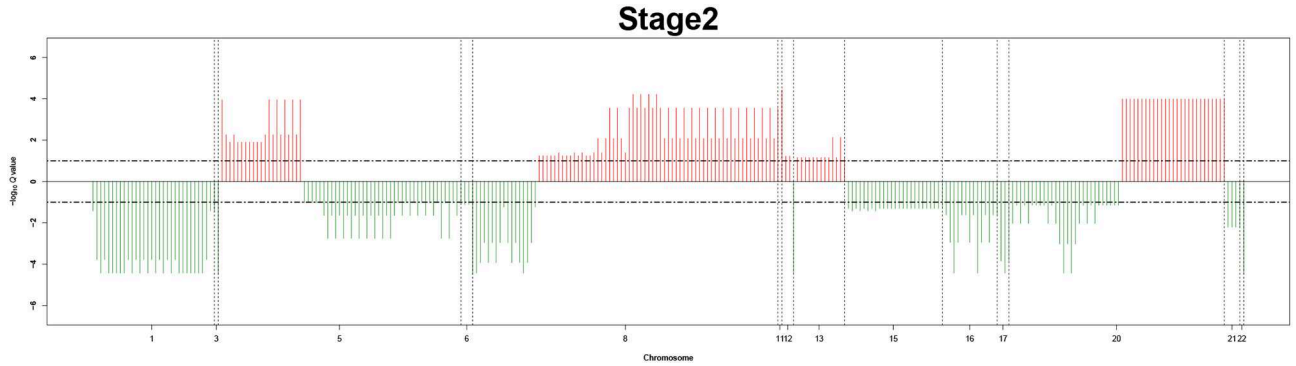


FIGURE 4 | Recurrent genome-wide SCNAs in stage 2. Genome-wide amplifications (red blocks) and deletions (green blocks) in stage 2.

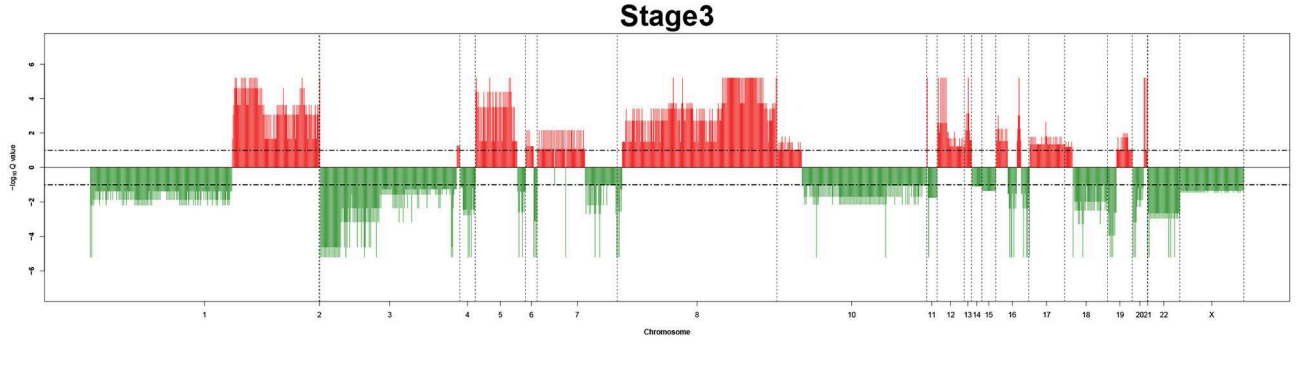


FIGURE 5 | Recurrent genome-wide SCNAs in stage 3. Genome-wide amplifications (red blocks) and deletions (green blocks) in stage 3.

genes were carried out for pathway network generation and functions interpretation, due to their ability to effectively explore cancer progression.

3.3. Dynamic Pathway Interaction Network Generation and Visualization

The evolution network was generated by considering the enriched pathways as nodes, and the overlapping genes in two corresponding pathways as edges. The network contains 50 nodes and 339 edges. Different colors (pink, orange, green, yellow) were used to showcase how these pathways evolved across the four pathologies adjacent stages, whereas the width of edges indicated the strength of their connections. The network was then visualized by Cytoscape software, where the different significant evolution paths are shown. These further details are depicted in **Figure 7**.

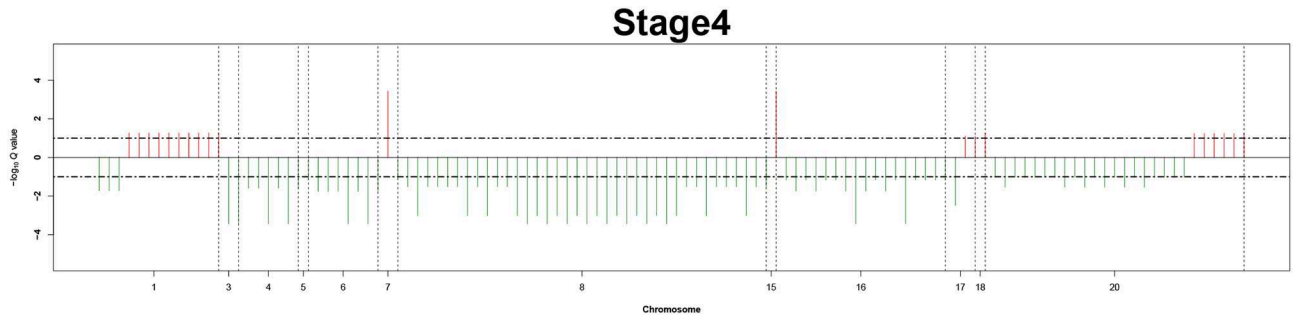


FIGURE 6 | Recurrent genome-wide SCNAs in stage 4. Genome-wide amplifications (red blocks) and deletions (green blocks) in stage 4.

TABLE 2 | The number of aberrant genes and enriched pathways detected at each pathology stage.

Pathology stages	Defined # of genes	# Of aligned pathways
Pathology_t_1	423	5
Pathology_t_2	3,265	110
Pathology_t_3	8,500	447
Pathology_t_4	2,244	94

TABLE 3 | The pathway enrichment of both the amplified and deleted genes from each pathology stage.

Pathway
DNA repair
Transport of small molecules
Developmental biology
Programmed cell death
Cell-cell communication
Hemostasis
Post-translational protein modification
Cellular responses to external stimuli

3.4. Stage-Related SCNAs Pathways Specific Functions Interpretation

The substantial analysis in this study confirmed the efficacy of the proposed framework. The detected genes were first enriched in many important pathways and these pathways, in turn, were strongly related to many critical cellular functions, such as cell cycle, disease, gene expression (Transcription), immune system, neuronal system, signal transduction, and metabolism of proteins and RNA. Some extra extremely enriched pathways obtained from both the amplified genes and deleted genes are shown in **Table 3**.

Interestingly, most of these functions were related to the immune system. This preliminary investigation can be clearly seen from the evolution network depicted in **Figure 7**. In this network, most of the pathways-related immune system were strongly related to each other with thicker edges. Furthermore, since the pathways enriched from the deleted genes (2,630 pathways) were higher than those of the amplified genes (2,069), the genes annotated in them were probably dynamically changed with the four pathological stages as can be observed from the evolution paths of the constructed evolution network. This dynamic change may lead to decrease the immunity in colorectal cancer and thus to homeostasis perturbance. Therefore, increasing the immunity activities across the stages will be effective and beneficial for many cancer types.

Moreover, signal transduction and cell cycle were also highlighted here. These functions are invariably perturbed in cancer since they are essential in regulating, activating multiple cellular process and signaling molecules. They can induce cell proliferation, differentiation, and survival of various cancers (Cao et al., 2014).

These functions were also involved in diverse human and animal diseases, and they provide useful information to understand the initiation and progression of many complex diseases.

4. CONCLUSION

Complex diseases evolution process is too difficult to be inferred by single genes, individual pathways or even a type of genomic data. However, understanding this evolution mechanism at a single level can be leveraged to identify more robust biomarkers and valid biological functions when integrating it with other genomic levels.

CNAs hold a very important role in cancers. Therefore, finding the recurrent CNA from cancer specific stages is a promising task for identifying their essential driver events. We have proposed to investigate the key indicators associated with cancer progressions by: (1) identifying the sequential changes/chromosomal abnormalities related to these stages, (2) defining their significant key genes, and (3) generating an evolution network rather than gene networks.

We have also used an interesting rCNA-algorithm that has the ability to identify many significant recurrent regions, due to its powerful homogeneous peel-off and its parameter setting that is very straightforward.

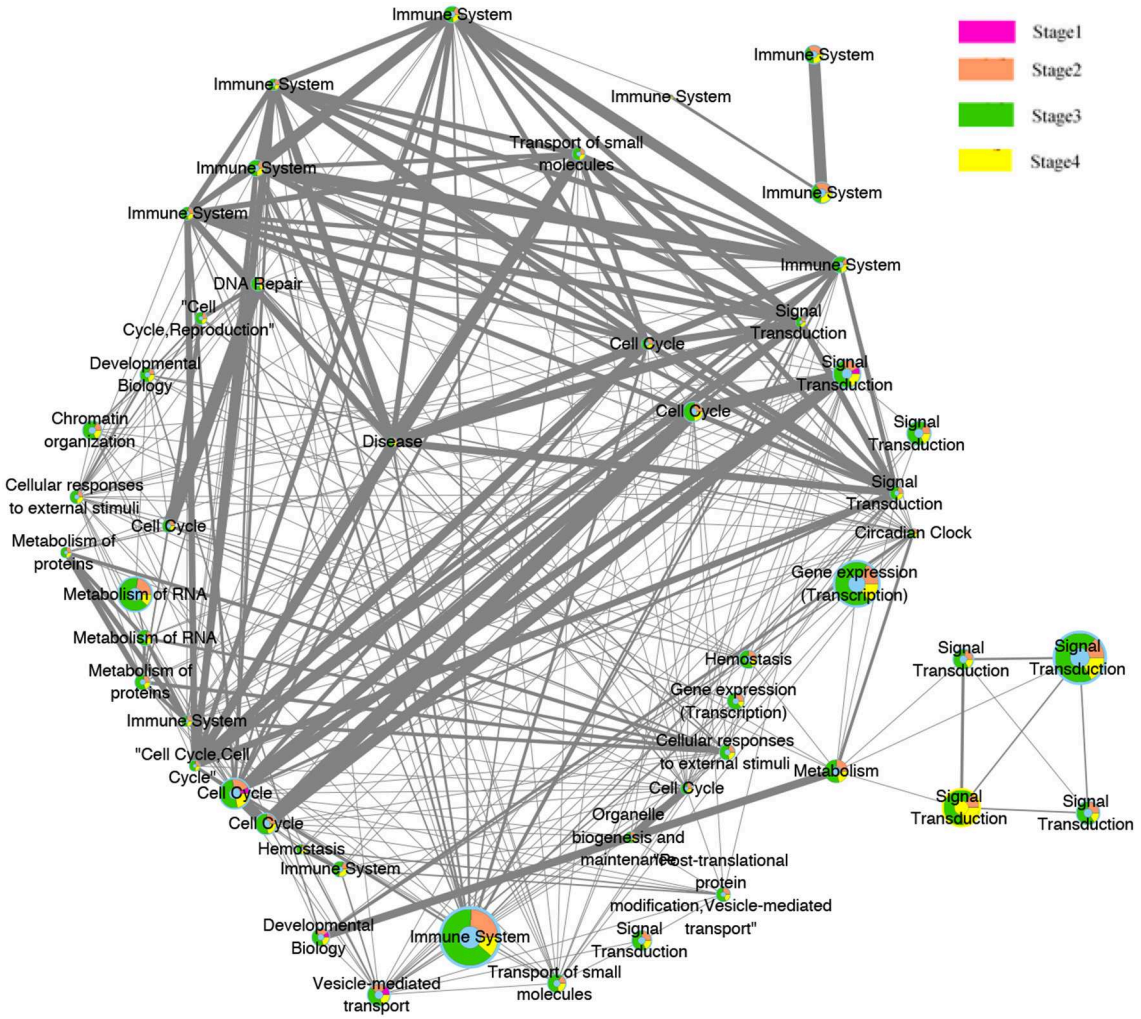


FIGURE 7 | Pathway functions interaction network. Node illustrate the biological pathways function. Edges illustrate the relationships between the functions at the adjacent stages. The size of the node is proportional to the number of genes in the pathway, The thicker the edges, the more overlapped genes between pathways of the adjacent stages. The color of pink, orange, green, and yellow inside the nodes indicate the pathways functions belongs to the four stages.

These critical factors identified from this valid alternative method enabled us to identify the differences between the molecular portraits of the different pathological stages, and improved our understanding of the pathogenesis and underlying molecular mechanism related to cancer initiation and progression. Moreover, the aberrant candidate genes and pathways characterized every pathology stage identified here could give us a clue to specific therapeutic targets for treatment of cancers.

In summary, such findings at a single level will help decide which types of omics data and methodologies will be better integrated to improve clinical research endpoints, and therefore get insights into the serious issues driving complex diseases. Furthermore, an interesting work would be to not only compare CNA events between cancer stages, but to also link these to somatic mutations in CIN (chromosomal instability) signature genes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: (1) FireBrowse: http://firebrowse.org, (2) Reactome database: http://www.reactome.org, (3) CNV data: ftp://ftp.broadinstitute.org/pub/GISTIC2.0/hg19support/.

AUTHOR CONTRIBUTIONS

BC initialized this study. CA and BC discussed many times to finalize the work plan. XS gave suggestions many times to modify this study. CA conducted the numerical experiments and drafted the manuscript. All authors read the manuscript and revised it, and agreed with the final version.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61972320, 61772426, 61702161, 61702420, 61702421, and 61602386, the Fundamental Research Funds for the Central Universities under Grant No. 3102019DX1003, the Key Research and Development and Promotion Program of Henan Province of China under Grant 182102210213, the Key Research Fund for Higher Education of Henan Province of China under Grant 18A520003, and the Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

ACKNOWLEDGMENTS

This paper has been reviewed and accepted by the Fourth CCF Bioinformatics Conference (CBC 2019).

REFERENCES

Ahn, T., Lee, E., Huh, N., and Park, T. (2014). Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics* 30, i422–i429. doi: 10.1093/bioinformatics/btu449

Cao, X. Q., Lu, H. S., Zhang, L., Chen, L. L., and Gan, M. F. (2014). Mekk3 and survivin expression in cervical cancer: association with clinicopathological factors and prognosis. *Asian Pac. J. Cancer Prev.* 15, 5271–5276. doi: 10.7314/APJCP.2014.15.13.5271

Cheang, M. C., Chia, S. K., Voduc, D., Gao, D., Leung, S., Snider, J., et al. (2009). Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J. Natl. Cancer Inst.* 101, 736–750. doi: 10.1093/jnci/djp082

Chen, L., Wang, R., Li, C., and Aihara, K. (2010). *Modeling Biomolecular Networks in Cells: Structures and Dynamics*. London: Springer-Verlag.

Chen, L., Wang, R. S., and Zhang, X. S. (2009). *Biomolecular Networks: Methods and Applications in Systems Biology*. Hoboken, NJ: John Wiley & Sons.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., et al. (2005). Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439–3440. doi: 10.1093/bioinformatics/bti525

Heitzer, E., Ulz, P., Geigl, J. B., and Speicher, M. R. (2016). Non-invasive detection of genome-wide somatic copy number alterations by liquid biopsies. *Mol. Oncol.* 10, 494–502. doi: 10.1016/j.molonc.2015.12.004

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19:299. doi: 10.1038/nrg.2018.4

Kim, E. S., Hirsh, V., Mok, T., Socinski, M. A., Gervais, R., Wu, Y. L., et al. (2008). Gefitinib versus docetaxel in previously treated non-small-cell lung cancer (interest): a randomised phase III trial. *Lancet* 372, 1809–1818. doi: 10.1016/S0140-6736(08)61758-4

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013). Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9:e1003118. doi: 10.1371/journal.pcbi.1003118

Lee, J. H., Zhao, X. M., Yoon, I., Lee, J. Y., Kwon, N. H., Wang, Y. Y., et al. (2016). Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. *Cell Discov.* 2:16025. doi: 10.1038/celldisc.2016.25

Liang, L., Fang, J. Y., and Xu, J. (2016). Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene* 35:1475. doi: 10.1038/onc.2015.209

Lowe, S. W., Bodis, S., McClatchey, A., Remington, L., Ruley, H. E., Fisher, D. E., et al. (1994). p53 status and the efficacy of cancer therapy *in vivo*. *Science* 266, 807–810. doi: 10.1126/science.7973635

Ma, X., Sun, P. G., and Zhang, Z. Y. (2018). An integrative framework for protein interaction network and methylation data to discover epigenetic modules. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1855–1866. doi: 10.1109/TCBB.2018.2831666

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Morganella, S., Pagnotta, S. M., and Ceccarelli, M. (2011). Finding recurrent copy number alterations preserving within-sample homogeneity. *Bioinformatics* 27, 2949–2956. doi: 10.1093/bioinformatics/btr488

Nibourel, O., Guihard, S., Roumier, C., Pottier, N., Terre, C., Paquet, A., et al. (2017). Copy-number analysis identified new prognostic marker in acute myeloid leukemia. *Leukemia* 31:555. doi: 10.1038/leu.2016.265

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18:1257. doi: 10.1038/82360

Tsao, M. S., Sakurada, A., Cutz, J. C., Zhu, C.-Q., Kamel Reid, S., Squire, J., et al. (2005). Erlotinib in lung cancer molecular and clinical predictors of outcome. *N. Engl. J. Med.* 353, 133–144. doi: 10.1056/NEJMoa050736

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10:252. doi: 10.1038/nrg2538

Wang, H., Liang, L., Fang, J. Y., and Xu, J. (2016). Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. *Oncogene* 35:2011. doi: 10.1038/onc.2015.304

Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., et al. (2004). Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.* 73, 1051–1087. doi: 10.1146/annurev.biochem.73.011303.073950

Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., et al. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* 45:1134. doi: 10.1038/ng.2760

Zhu, G., Yang, H., Chen, X., Wu, J., Zhang, Y., and Zhao, X. M. (2017). Cstea: a webserver for the cell state transition expression atlas. *Nucleic Acids Res.* 45, W103–W108. doi: 10.1093/nar/gkx402

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 February 2020
doi: 10.3389/fgene.2020.00105

Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms

Haihe Shi * *and Xuchu Zhang*

School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

In recent years, there has been an explosive increase in the amount of bioinformatics data produced, but data are not information. The purpose of bioinformatics research is to obtain information with biological significance from large amounts of data. Multiple sequence alignment is widely used in sequence homology detection, protein secondary and tertiary structure prediction, phylogenetic tree analysis, and other fields. Existing research mainly focuses on the specific steps of the algorithm or on specific problems, and there is a lack of high-level abstract domain algorithm frameworks. As a result, multiple sequence alignment algorithms are complex, redundant, and difficult to understand, and it is not easy for users to select the appropriate algorithm, which may lead to computing errors. Here, through in-depth study and analysis of the heuristic multiple sequence alignment algorithm (HMSAA) domain, a domain-feature model and an interactive model of HMSAA components have been established according to the generative programming method. With the support of the PAR (partition and recur) platform, the HMSAA algorithm component library is formalized and a specific alignment algorithm is assembled, thus improving the reliability of algorithm assembly. This work provides a valuable theoretical reference for the applications of other biological sequence analysis algorithms.

Keywords: heuristic multiple sequence alignment algorithms, feature model, generative programming, component interaction model, partition and recur platform

OPEN ACCESS

Edited by:
Yanjie Wei,
Shenzhen Institutes of Advanced Technology (CAS), China

Reviewed by:
Pu-Feng Du,
Tianjin University, China
Wang-Ren Qiu,
Jingdezhen Ceramic Institute, China
Weiguo Liu,
Shandong University, China

***Correspondence:**
Haihe Shi
haiheshi@jxnu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 November 2019*
Accepted: *29 January 2020*
Published: *27 February 2020*

Citation:
Shi H and Zhang X (2020) Component-Based Design and Assembly of Heuristic Multiple Sequence Alignment Algorithms. Front. Genet. 11:105. doi: 10.3389/fgene.2020.00105

INTRODUCTION

Since the beginning of the 21st century, with the development of high-throughput sequencing technology, gene sequencing has become much cheaper and more efficient, enabling the development of various genome projects. Since the implementation of the Human Genome Project (Collins et al., 1998), the amount of bioinformatics data being produced has grown explosively, with genome sequencing data doubling every 4–5 months. At the same time, bioinformatics (Zhang, 2000), a new interdisciplinary subject, has developed rapidly. Bioinformatics covers all aspects of the acquisition, processing, storage, distribution, analysis, and interpretation of biological information. It integrates tools from mathematics, computer science, and biology to clarify and understand the biological significance of large amounts of data (Hogeweg and Searls, 2011). One of the major problems faced by bioinformatics today is how to process the

data generated by genetic engineering. Data are not information; data need to be mined using systematic scientific methods to find biologically relevant information.

Sequence alignment is a fundamental method to study biological sequence data in bioinformatics (Mount, 2005). The theoretical basis of sequence alignment is the chemistry in biology, that is, if the similarity between two biological sequences reaches a threshold, it is considered that they have similar functions and structures as well as evolutionary relationships. By comparing an unknown biological sequence with a known functional structure, and identifying similar regions between them, the homology between the species can be judged, and the biological information contained in the unknown sequence can be revealed. According to the number of sequences to be aligned, sequence alignment can be divided into pairwise and multiple sequence alignment. The standard solution for pairwise sequence alignment is to use a dynamic programming algorithm to find the optimal solution. The classical algorithm is the Needleman-Wunsch (Needleman and Wunsch, 1970) algorithm, which is used to solve the global pairwise sequence alignment problem; the more biologically significant local alignment problem can be solved by the Smith-Waterman (Smith and Waterman, 1981) algorithm. Also, the heuristic-based BLAST (Altschul et al., 1990) algorithm is widely used in similarity sequence searches of gene databases. Theoretically, the dynamic programming approach to pairwise sequence alignment can be used for multiple sequence alignment problems. A two-dimensional dynamic programming matrix is extended to the three-dimensional or multi-dimensional case, where the dimension of the matrix reflects the number of sequences to be compared. This method is only suitable for multiple sequence alignments with few dimensions, otherwise it will be a great challenge with respect to computer resources. It has been proved that the multiple sequence alignment problem based on the SP (sum of pairs) metric is NP (Wang and Jiang, 1994), and multiple sequence alignment uses a heuristic algorithm. Here, we mainly focus on the heuristic multiple sequence alignment algorithm (HMSAA) domain.

HMSAAs include progressive alignment (Feng and Doolittle, 1987) and iterative alignment (Wang and Li, 2004); this paper mainly considers the progressive alignment method. The progressive multiple sequence alignment algorithm was proposed by Feng and Doolittle in 1987. Thompson and Higgins implemented the progressive multiple sequence alignment algorithm and proposed the ClustalW (Thompson et al., 1994) algorithm. Subsequently, Notredame et al. (2000) proposed the T-Coffee (tree-based consistency objective function for alignment evaluation) algorithm; the latter two algorithms are the most commonly used progressive multiple sequence alignment algorithms. The HAlign (Zou et al., 2015) algorithm is a progressive alignment algorithm based on central star alignment. Clustal Omega (Sievers et al., 2011) is a completely rewritten and revised version of the widely used Clustal series of programs for multiple sequence alignment. The main improvement over ClustalW algorithm is the use of the mBed algorithm to generate guide trees of any size and the use of HHalign Package based on the idea of hidden Markov model in the last step of Profile alignment. The main disadvantage of the progressive multiple sequence alignment algorithm is its principle of "once vacant, always vacant." The errors generated in the alignment will always affect the sequence alignment process, which may lead to a suboptimal result and reduce the accuracy of the algorithm. The basic idea of the progressive alignment algorithm is that there is an evolutionary relationship between the multiple sequences that are aligned; after determining the evolutionary order of the sequences, they are gradually aligned along the evolutionary order until all sequences are aligned. This means that before proceeding to the progressive alignment, it is necessary to find the evolutionary relationship between the sequences. At present, optimization of the progressive alignment algorithm usually focuses on the step of confirming the evolutionary relationship (Zhang et al., 2005; Huo and Xiao, 2007). In order to speed up sequence alignment when the scale of the alignment is large, parallel computing may be combined with progressive alignment (Hung et al., 2015). The basic idea of iterative alignment is first to improve the multiple sequence alignment based on an algorithm that can generate alignments, through a series of iterations, until the alignment results no longer improve or have reached the maximum number of iterations. This paper mainly considers the combination of iterative alignment and progressive alignment. Such algorithms, which include MultAlin (Corpet, 1988) and Muscle (Edgar, 2004), have improved robustness and wider application scope.

At present, most research on sequence alignment algorithms focuses on the optimization of specific steps of a particular algorithm. The optimization effect on different sequences will be different, and the diversity and complexity of sequence alignment algorithms may make it difficult for users to select an algorithm appropriate to the characteristics of a given sequence, resulting in unnecessary computing errors in practice. On the other hand, it may be difficult for users to understand the structure of a sequence alignment algorithm, which may affect its correct use and to some extent affect the accuracy of the sequence analysis. The specificity and low abstraction of a sequence alignment algorithm reduce its reusability and maintainability. Therefore, it is necessary to study sequence alignment algorithms at the domain level. Concerns on algorithm families will be helpful for extracting the commonality and variability of different algorithms and for the formal development of sequence alignment algorithms.

In this work, the generative programming method is used to design an abstract generic algorithm component library, after which a specific alignment algorithm for the HMSAA domain is assembled, thus improving the reliability and reusability of the algorithms. First, domain analysis of HMSAA is carried out, the common domain features and variability features are identified, and a domain feature model of HMSAA is established. Furthermore, relationships among features are analyzed and an interaction model of algorithm components is designed and constructed. Finally, using a generic abstract programming language, Apla, the domain components are formally implemented and a high abstract component library is built on top of Apla.

RELATED METHODOLOGY AND TECHNOLOGY

Generative Programming

Software reuse is considered to be one of the solutions to the "software crisis." High-quality software reuse can improve the efficiency and quality of software development and ultimately result in the construction of an industrialization pipeline to develop software. Generative programming (Czarnecki and Eisenecker, 2000) is the use of components and the creation of software products in an automated manner. Implementation consists of two steps. First, the current software development model is transformed into the development of the software system family. Then, a generator is used to automatically assemble the components. Through domain analysis of the software system family, generative programming constructs a domain model of the system family and further develops the domain design and domain based on this model. New software development in the same field is based on the established domain model, and the reusable components are selected for assembly and implementation. It is not the development of software.

A domain model based on generative programming includes a problem space, a solution space, and domain-specific configuration knowledge for mapping between the two. The problem space is used to represent the requirements of the customization system, and is mainly for use by application programmers and customers. The solution space includes the implementation components required for the system family implementation and the combination, dependencies, and interactions among implementation components. Domain-specific configuration knowledge is mainly used to separate the problem space and solution space, which not only reduces the redundancy and coupling of the implementation components but also improves their composability and reusability. The composition of such a generative domain model is shown in **Figure 1**.

Domain Modeling

Domain modeling requires the identification and modeling of key concepts (Lee et al., 2002). Feature engineering (Turner et al., 1998) considers features to be first-order entities that traverse the software life cycle and span the problem space and solution space, and reduce the difference in demand awareness between users and software developers through features. Features in FODA (feature-oriented domain analysis) (Kang et al., 1990) are considered to be user-visible, significant, and distinctive aspects, qualities, characteristics, *etc*., in a software system. Features are the domain knowledge accumulated by users and experts from long-term practice in a domain. Feature modeling is an activity that models the commonality and variability of features and the relationships among them. Zhang and Mei (2003) proposed a feature-oriented domain modeling (FODM) method that considered the features of services, functions, behavioral characteristics, *etc*. This was based on service analysis activities, functional analysis activities, and behavioral characteristics analysis in combination with domain terminology analysis, commonality and variability analysis, interactive process analysis, and quality demand analysis concurrently, with continuous retrospective refinement to finally obtain the feature model. The domain modeling process is illustrated in **Figure 2**.

Partition and Recur Method

PAR (Xue, 1993; Xue, 1997; Xue, 1998; Shi and Xue, 2009; Xue, 2016) (partition and recur) is a formal development method based on partition and recursion, containing an algorithm design language (Radl; recurrence-based algorithm design language), an abstract generic programming language (Apla), and a unified algorithm design and proof method, as well as a series of generation systems (the PAR platform).

The Apla language can be used to directly write programs using abstract data types and abstract procedures. It has the advantages of concise and rigorous mathematical language, and

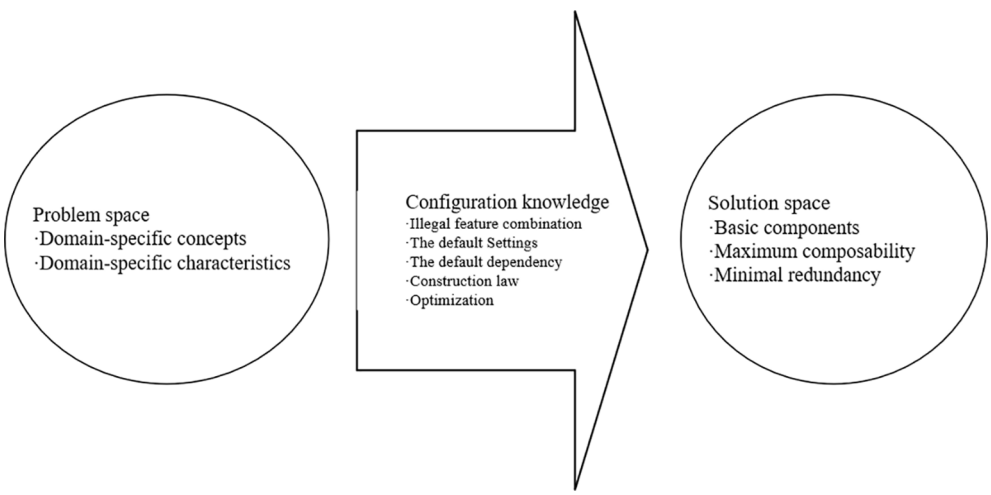


FIGURE 1 | Composition of generative domain model.

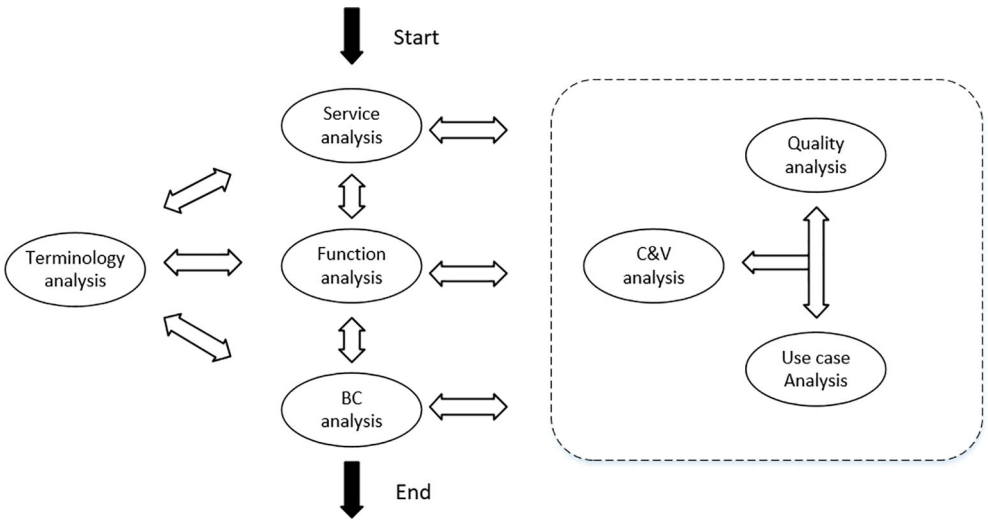


FIGURE 2 | Feature modeling process.

its high level of abstraction makes it suitable for describing abstract algorithmic programs.

The generic programming mechanisms supported by Apla include type parameterization, subroutine parameterization, and user-defined generic abstract data types (ADT). 1) Apla introduces the keyword *sometype* to define the type variable, the type parameter, the parameter return value type of the procedure function, and the basic type of the combined data type. The type is used as a parameter to implement the generalization of the program. 2) Apla subroutine parameterization includes procedure parameterization and function parameterization. In a subroutine, the keywords *proc* and *func* are used to declare procedure parameters and function parameters, and the procedure or function is used as a parameter list. 3) As well as the predefined ADT in Apla, users can create custom ADT to make the language more flexible and the program description more powerful. These custom operations include the definition and the implementation of ADT. The ADT definition contains the operation name, the operation type, the parameters of the operation, *etc*. The ADT implementation gives the specific implementation methods of these operations, and *define*, *ADT*, *enddef*, *implement*, *endimp*, and other keywords are used to describe the custom ADT. In addition, the PAR platform supports the transformation of Apla into an executable high-level programming language such as C++ or Java.

HEURISTIC MULTIPLE SEQUENCE ALIGNMENT ALGORITHM MODELING

In this section, the FODM method is used to construct the feature model according to the service, function, and behavior characteristics in the HMSAA domain. Heuristic multiple sequence alignment operations are core services in the domain. The sequence legality check (*seq_check*), heuristic alignment mode selection (*heur_mode*), pairwise sequence alignment operation (*psa*), distance matrix (*dist_matrix*), result output (*result_op*), progressive alignment (*prog_align*), and iterative alignment (*iter_align*) are the main functions in the domain. Progressive alignment and iterative alignment are sub-functions of heuristic alignment mode selection. Sequence legality check, heuristic alignment mode selection, and alignment result output are mandatory, where as function, pairwise sequence alignment operation, and distance matrix are optional. For progressive alignments, the progressive alignment mode *(prog_align_mode)* is a behavioral feature that has the following three values: the phylogenetic tree (*phy_tree*), the extended library (*expan_lib*), and the center alignment. For pairwise sequence alignment operations, the pairwise sequence alignment mode (*psa_mode*) is a behavioral feature that has two values, fast alignment (*k-mer*) and dynamic programming alignment (*dp*). According to the above analysis, a feature model was constructed for the domain, as shown in **Figure 3**.

DESIGN AND IMPLEMENTATION OF HEURISTIC MULTIPLE SEQUENCE ALIGNMENT ALGORITHM COMPONENTS

Interaction of Algorithm Components in Heuristic Multiple Sequence Alignment Algorithm Domain

According to the feature model described in the previous section, in order to achieve a complete library of algorithm components, it is necessary to further analyze the interaction modes among different algorithm components. The interactions of algorithm components involve constraints and dependencies between

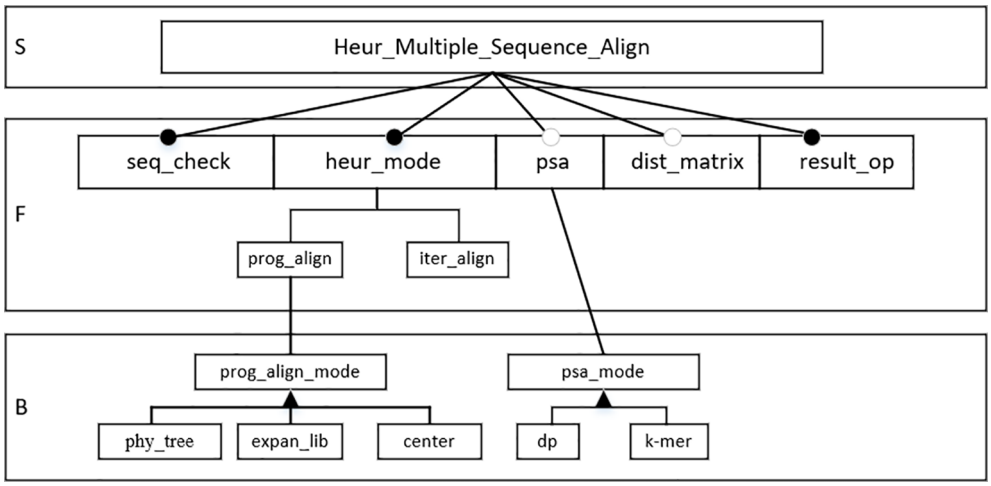


FIGURE 3 | Feature model of heuristic multiple sequence alignment algorithm (HMSAA).

features. Therefore, this section describes an interaction model of algorithm components in the HMSAA domain according to their interaction modes.

Through the establishment of the HMSAA feature model, it can be concluded that the algorithm consists of four main process features, i.e., heuristic alignment mode selection, progressive alignment, iterative alignment, and result output. In addition, the input of the algorithms in this domain consists of sequences of biological information, including DNA, RNA, and protein sequences. Before the implementation of the algorithm, the legality of the sequence information needs to be checked, for example, a DNA sequence can only contain four letters, *A*, *T*, *C*, and *G*. The main components in this domain are sequence legality checking, heuristic alignment mode selection, progressive alignment, iterative alignment, and result output. Other features and data structures in the feature model are used

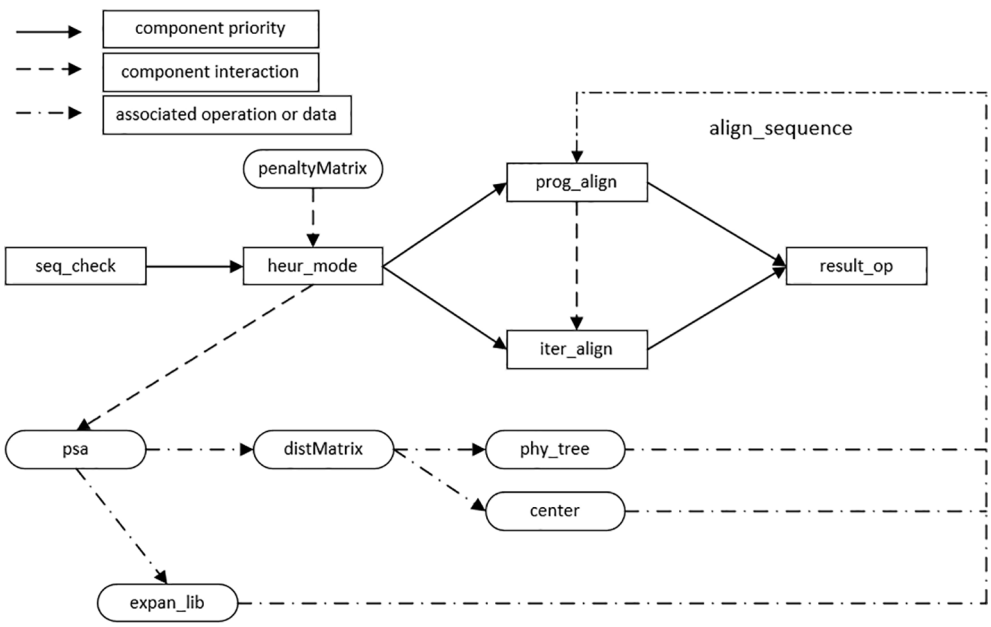


FIGURE 4 | Algorithm components interaction model.

as auxiliary components, and the interaction model of components is established according to the dependencies between them, as shown in **Figure 4**.

The nodes connected by solid lines are the basic components that must be included in the HMSAA domain, namely, the three mandatory features and two sub-features selected by the heuristic mode. The solid line arrow denotes the execution priority of the component from high to low along the direction of the arrow.

Dotted and underlined arrows represent the data, structures, and associated operations required in the assembly of the algorithm components. For example, the establishment of an extended library requires the information of alignment in *psa*. The dotted arrow indicates the interaction between the two components during the execution of the algorithm. For example, before an iterative alignment, a progressive alignment should be carried out. By setting the number of iterations and the iterative method, the result of the progressive alignment is iterated until there is no further change or the maximum number of iterations has been reached.

In summary, the above interaction model includes most mainstream HMSAAs, including the progressive alignment algorithm based on tree and central alignment, the progressive alignment algorithm based on a compatible optimization objective function to form an expanded library, and the multiple sequence alignment algorithm combining iteration and progress. Here we just outline simple formal specification description of two components below for examples.

1 seq_check component

```
|[in a[][]: Array[][]; out bl: Boolean]|
AQ: bio-sequences.
AR: if bl is true, the legality check is passed;
false is the opposite.
```

2 prog_align component

```
|[in heur_mode: ADT; phy_tree: ADT; a[][]:
Array[][] out b[][]: Array[][]]|
AQ: phy_tree component, the sequence to be
aligned, heur_mode component.
AR: alignment of multiple sequences.
```

Here *in* and *out* in the front of pre-condition *AQ* are two key words defined in PAR platform and are used to denote the input and output respectively; array, *Boolean*, etc., are the predefined types in PAR platform, and AR stands for post-condition of algorithm.

Apla Formal Implementation

In this section, we make use of the advantages of Apla, including high-level abstraction, strong support for ADT, and easy correctness verification, and formally implement the HMSAA model. Here, only the implementation of the tree-based progressive alignment algorithm is illustrated.

1 seq_check component

Check whether the sequence group meets the biological definition. For example, the character set of the DNA sequence is {A, T, C, G}.

```
procedure seq_check(a[]:array[String]);
```

2 Penalty component

We designed the penalty model as an ADT, using an affine penalty model, where *sometype* is a keyword in the Apla language that defines the type variable. *GapOpen*, *GapExtend*, and *score* represent the penalty points of open vacancy, extended vacancy, and non-vacancy, respectively.

```
define ADT penaltyMatrix(sometypeelem);
GapOpen : Integer;
GapExtend : Integer;
score:array[array[Integer]];
enddef.
```

3 heur_mode component

The *Heur_mode* component is defined as an ADT that selects the operation mode of multiple sequence alignment and defines the data structure and information required for alignment. The *setPenaltyMatrix*, *setGapOpen*, and *setGapExtend* functions are to set penalty matrix, open vacancy penalty, and extend vacancy penalty, respectively. The generic procedure *tree_prog_align* sets the alignment mode to one designated by the user. The useLib means to select the expan-lib component, useFullPW denotes the use of the conventional dynamic programming pairwise sequence alignment algorithm, useIter represents the iteration, and treeAlgorithm is the algorithm to generate the phylogenetic tree.

```
define ADTheur_mode(sometypeelem);
function setPenaltyMatrix(pm:
penaltyMatrix):Array[Array[Integer]];
functionsetGapOpen(gapOpen: Integer):
Integer;
function setGapExtend(gapExtend:
Integer): Integer;
procedureprog_align(useLib: bool;
useFullPW: bool; useIter: bool;
treeAlgorithm : String)
enddef.
```

4 dist_matrix component

The *dist_matrix* is defined as an ADT that calculates the distance matrix element and returns it using the score of the pairwise alignment, and the pairwise sequence alignment operation is defined as the generic parameter. The function *getDist* is used to get the data from the distance matrix. Proc psa is described in detail in reference (Shi and Zhou, 2019).

```
define ADT dist_matrix (sometypeelem);
   function calDistMat (proc psa(…):Array
[Array[Integer]];//.
   function getDist (ii: Integer; jj:
```

```
    Integer): Integer;
    ......
    enddef;
```

5 phy_tree component

Phylotree data structure is defined as an ADT, facilitating subsequent operations on the tree. The parameters *treeMess*, *left*, and *right*, respectively, represent the information of the tree node and the left and right subtrees. The *phylotree* component is defined as an ADT that generates a phylogenetic tree using the data in the distance matrix. The ADT contains the generic procedure *generateTree* and takes *selAlgorithm* as its generic parameter. The generic procedure can generate phylogenetic trees through different algorithms. The function *calWeight* calculates the weight of each sequence when calculating the score of multiple sequences alignment, the generic procedure *getStepsForMSA* is used to generate the sequence of subsequent multiple sequence alignments, and the generic procedure *readTree* is used to read information from the generated phylogenetic tree.

```
define ADT phyloTree (sometypeelem);
    treeMess: Array[Array[Integer]];
    left: Integer[];
    right: Integer[];
enddef.
define ADT phy_tree (sometypeelemMatrix);
    procedure generateTree(distMat:
elemMatrix; seqName: String[];
funcselAlgorithm():String;treeName :
String; result: Boolean);
    function calWeight (firstSeq: Integer;
lastSeq: Integer; seqsW: Array[Integer]):
Array[Integer];
    procedure readTree (seqName: String[];
treeName: String; firstSeq: Integer;
lastSeq: Integer);
    procedure getStepsForMSA
(procreadTree; distMat: elemMatrix;
result: Boolean);
    ......
enddef.
```

6 prog_align component

The *prog_align* component is defined as an ADT that includes the generic procedure *multiSeqAlign*, which performs progressive alignment according to the alignment order obtained from the phylogenetic tree and the sequence weight.

```
define ADT prog_align (sometypeelem);
  procedure multiSeqAlign (seqs: Array
[String]; steps: elem; seqName: String[];
seqW:Array[Integer]; start: Integer);
  ......
enddef.
```

7 result_op component

The *result_op* component is defined as an ADT. It is composed of two generic procedures, *multiAlign_op* and *phyloTree_op*. The *multiAlign_op* procedure annotates the results of multiple sequence alignments and outputs them; *pathAlignOutput* is the path of the output file. The p*hyloTree_op* procedure outputs the phylogenetic tree; here, *pathTreeOutput* is the path of the output file.

```
define ADT result_op(sometypeelem)
  procedure multiAlign_op
(pathAlignOutput: String; seqs: Array
[String]; seqName: String[];
sometypeprog_align);
  procedure phyloTree_op (pathTreeOutput:
String; seqName: String[];
sometypephyTree; sometypedistMat);
enddef.
```

ASSEMBLY OF CLUSTAL ALGORITHM

In this section, a phylogenetic tree-based progressive alignment algorithm, *clustalW*, is assembled on top of the HMSAA component library introduced in previous section. The Apla program is as follows.

```
program clustalW;
  const/* input sequences*/
  var
  seqs, seqsName: Array[String];//Seqs is
  the sequence to be aligned
     //seqsName is the identification name
  of the sequence
  const pathTreeOutput, pathAlignOutput:
  String;
  /*omit the initialization of pairwise
  sequence alignment*/
  ADT pm: new penaltyMatrix ();
  ADT psa: new psa (……);
  ADT distM: new dist_matrix (psa);
  ADT phyloTree: new phylotree ();
  ADT tree: new phy_tree (phyloTree;
  distM);
  ADT msa: new prog_align (tree);
  ADT mode: new heur_mode (pm);
  var
  clustalw: mode; gapOpen, gapExtend:
  Integer; penalty: Array[Array
  [Integer]];
  begin
  clustalw.setPenaltyMatrix (penalty);
  clustalw.setGapOpen (gapOpen);
  clustalw.setGapExtend (gapExtend);
  end;
  ADT resultOp: new result_op
  ();//instantiate and initialize the
  required components
```

```
    procedure heur_multiple_sequence_align
    (clustalw; psa; distM; tree; msa;
    resultOp);
    //heuristic multiple sequence alignment
    operations
    var
    NJTree: String; result: Boolean
    begin
      check (seqs);
clustalw.prog_align (false; true; false;
    "NJTree");
  if(clustalw.getUseLib = false)→
   if(clustalw.getUseFullPW = true) →
   distM.calDistMat (psa);
   tree.generateTree
         (clustalw.getTreeAlgorithm;
         seqsName; clustalw; distM;
         result);
      msa.multiSeqAlign (seqs; seqsName;
      tree; 0);
      resultOp (msa; tree; pathTreeOutput;
      pathAlignOutput);
end.
```

EXPERIMENTS

As the Apla language cannot run directly, in this section we make use of the PAR platform to transform the Apla algorithm components into the corresponding C++ components.

ADT algorithm components in Apla containing only data members are transformed into *struct* data types in C++, such as *penaltyMatrix* and *phyloTree*. The results are as follows.

```
struct penaltyMatrix
{
    int gapOpen;
    int gapExtend;
    vector < vector < int>> score;
};
struct phyloTree
{
    vector < vector < int> > treeMess;
    vector < double > leftBranch;
    vector < double > rightBranch;
};
```

ADT components containing data members and member functions are transformed into classes in C++, such as *dist_matrix* and *phy_tree*. The function body code is long, and so part of it is omitted here. The partial result of the transformation is shown in **Figure 5**.

Generic procedures and functions defined in Apla are converted into separate class member functions in C++ to reduce coupling between components. In particular, the calling functions are converted into indicator functions in C++, and the generic parameter is converted into the pointer parameter to implement the polymorphism of the Apla program. After converting each component into C++, the Apla code for the heuristic multiple sequence alignment operation is converted into the main function executed in C++; finally, the *clustalW* algorithm program is run through manual assembly of the components, as shown in **Figure 6**.

To test the program, we used four pieces of DNA data, *Cyprinus carpio* (common carp) alpha-globin, *Homo sapiens* (human) alpha globin, *Mus musculus* (house mouse) alpha-globin, *Capra hircus* (goat) alpha-globin. The alignment results between our algorithm and the other two Clustal algorithms are shown in **Figure 7**. Due to the different selected pairwise

```
class dist_matrix
{
public:
    dist_matrix(int seqs_num){...}
    ~dist_matrix() {...};
    vector<vector<double>> dist;
    void calDistMat(int (Psa::* psa)(Score_matrix_mani&,
        const string&, const string&),
        vector<string>* seqs, Score_matrix_mani** matrix) {...};
    double getDist(double i , double j) {...};

};


class phy_tree
{
public:
    phy_tree() {...};
    ~phy_tree() {...};
    void generateTree(dist_matrix* dist,vector<string> seqsName, string algorithm;
                      string* phylipName, bool* result) {...};
    void readTree(vector<string> seqsName, string treeFileName, int first, int last) {...};
    void calWeight(vector<int>* seqWeights,vector<string> seqsName, int first, int last) {...};
    vector<vector<int>> getStepsForMSA(int first, int seq_num, string* treeFileName,
                      dist_matrix* dist) {...};
};
```

FIGURE 5 | Result of ADT transformation.

```
int main()
{
    const char* path = "D:\\VS workspace\\Hmsaa\\Hmsaa\\1.fa.txt";
    const char* pathAlignOutput = "D:\\VS workspace\\Hmsaa\\Hmsaa\\1.fa.dnd";
    const char* pathTreeOutput = "D:\\VS workspace\\Hmsaa\\Hmsaa\\1.fa.aln";
    heur_mode hm;
    hm.prog_align(false, true, false, "NJTree");
    result_op output;  Psa psa;  phy_tree tree;  treeNode* t; int *next;
    int first, last;  bool result = false; string treeFileName;
    vector<vector<int>> steps; vector<vector<int>>* stepsPtr; vector<int>* seqWeights;
    dist_matrix distM(seq_num);  prog_align msa;
    Multiple mul_seq = readSeq().readFasta(path);
    vector<string> mul_seq_seqs = mul_seq.seqs;
    vector<string> mul_seq_names = mul_seq.names;
    int seqs_num = mul_seq_seqs.size();
    check().checkDna(mul_seq_seqs);
    if(h.getUseLib()==false)
    {
      if(h.getUseFullPW())
      {
        psa = new DPPairwise(&mul_seq_seqs, seqs_num);
        distM.calDistMat(&Psa::psa, &mul_seq_seqs, psa->get_matrix());
        tree.generateTree(&distM, seqsName, hm.getTreeAlgorithm(), &treeFileName, &result);
        tree.calWeight(seqWeights, mul_seq_names, first, last, &treeFileName, &distM);
        steps = tree.getStepsForMSA(first, seq_num, &treeFileName, stepsPtr, &distM);
        msa.multiSeqAlign(mul_seq_seqs, mul_seq_names, &seqWeight, &steps, 0);
        output.multiAlign_op(pathAlignOutput, mul_seq_seqs, mul_seq_names, &msa);
        output.phyloTree_op(pathTreeOutput, mul_seq_names, &tree, &distM);
      }
      else{......}
    }
    else{......}
}
```

FIGURE 6 | C++ assembly process of clustalW.

alignment parameters and types, the structure of the phylogenetic tree is different from that of the ClutsalW algorithm, and the sequence of alignment has also changed, but the results remain biologically significant.

SUMMARY AND FUTURE WORK

As a key topic in bioinformatics research, sequence alignment algorithm and its applications have attracted extensive research attention. However, there has been no work considering it as a domain for high-level abstraction to improve the reliability and the productivity of the algorithms, and to reduce the probability of suboptimal solutions, errors of the algorithm, etc.

Generative programming and the composition of a generative domain model are first briefly presented in the paper, and the FODM method is described. The feature model can be obtained by taking the characteristics of the service, function, and behavior characteristics of the domain into account, and carrying out a series of feature analysis activities.

By using generative programming and feature modeling, the HMSAA domain has been analyzed, resulting in the following three algorithm classes: progressive alignment algorithms based on tree and central alignment; progressive alignment algorithms based on the compatible optimization objective function to form an expansion library; and multiple sequence alignment algorithms based on a combination of iteration and increment. Through analysis of this domain, general and variable features have been extracted and mapped to components, and an HMSAA feature model has been established. Moreover, an interaction model of HMSAA domain components has been designed based on the relationships among features and formally implemented using the generic abstract programming language Apla in support of the PAR platform. An HMSAA component library has been established, the reliability of which can be guaranteed owing to the ease of verification with the Apla language.

It is expected that the formal components could be automatically or semi-automatically assembled to generate a specific problem-solving algorithm, thus reducing the errors resulting from manual algorithm selection for multiple sequence alignment, and improving the algorithm efficiency, which will enable assembly of a new, more efficient, multiple sequence alignment algorithm. Furthermore, the high-level abstraction of generic components, such as *generateTree*, provides a diversity of algorithm components assembly as well as a good demonstration of the connections between algorithm features, thus improving the understandability and ease-of-use of algorithms.

Next, we will release our codes in GitHub. Future work also include developing a user-friendly visual interface to facilitate component assembly. Users will be able to generate different sequence alignment algorithms by selecting different components *via* the interface and use XML files to describe the composition and constraint relations among components, without any change to the component library. We are encouraged by the success of algorithm assembly on the PAR platform.

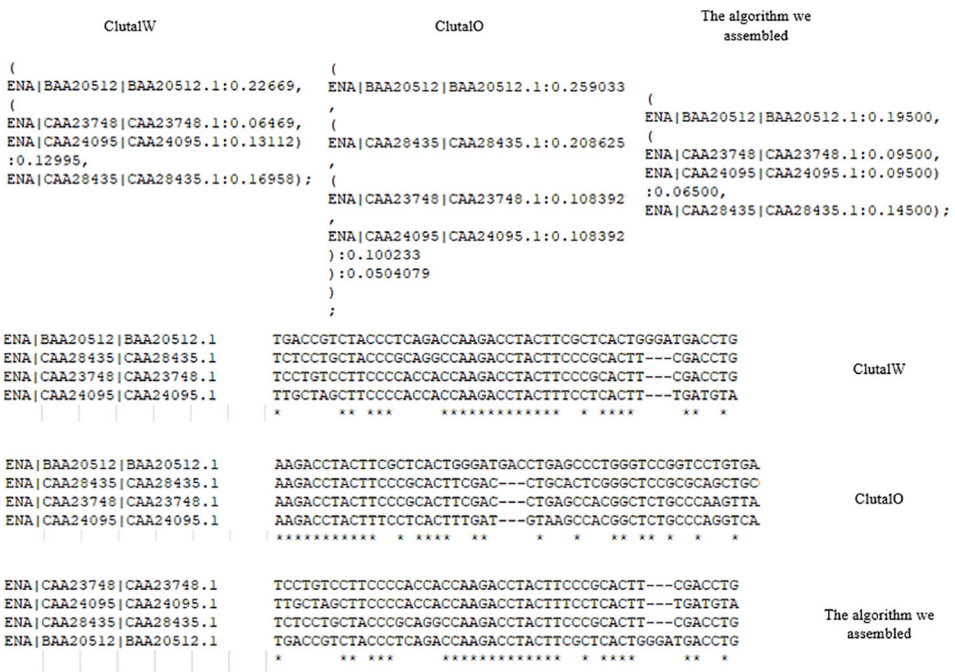


FIGURE 7 | Alignment results.

The methodology and techniques for HMSAA are not only applicable to multiple sequence alignment algorithms but also have theoretical reference significance and practical application value for other biological sequence analysis algorithms, such as the assembly algorithm based on DeBruijn graph structure used in the process of gene assembly (Li et al., 2010; Peng et al., 2012).We are currently applying some of these ideas to more problems in the domain of biological sequence analysis, to implement automatic or semi-automatic assembly of an algorithm component library based on the PAR platform. We hope to report on this work in the near future.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

AUTHOR CONTRIBUTIONS

HS instructed the whole research work and revised the paper. XZ did the codes work and the experiments. All authors read and approved the final manuscript and are agree to be accountable for all aspects of the work.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos.61662035, 61762049, and 61862033.

ACKNOWLEDGMENTS

We thank the reviewers of CBC2019 for their helpful comments.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. R. (1998). New goals for the U.S. Human genome project: 1998-2003. *Science* 282 (5389), 682–689. doi: 10.1126/science.282.5389.682

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* 16 (22), 10881–10890. doi: 10.1093/nar/16.22.10881

Czarnecki, K., and Eisenecker, U. (2000). *Generative programming: methods, tools, and applications* (New York: ACM Press/Addison-Wesley Publishing Co.).

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32 (5), 1792–1797. doi: 10.1093/nar/gkh340

Feng, D. F., and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisiteto correct phylogenetic trees. *J. Mol. Evol.* 25 (4), 351–360. doi: 10.1007/bf02603120

Hogeweg, P., and Searls, D. B. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput. Biol.* 7 (3), e1002021. doi: 10.1371/journal.pcbi.1002021

Hung, C. L., Lin, Y. S., Lin, C. Y., Chung, Y. C., and Chung, Y. F. (2015). CUDA Clustalw: an efficient parallel algorithm for progressive multiple sequence alignment on multi-GPUs. *Comput. Biol. Chem.* 58, 62–68. doi: 10.1016/j.compbiolchem.2015.05.004

Huo, H. W., and Xiao, Z. W. (2007). A multiple alignment approach for DNA sequences based on the maximum weighted path algorithms. *J. Softw.* 18 (2), 185–195. doi: 10.7666/d.y858982

Kang, K., Cohen, S., Hess, J., Nowak, W., and Peterson, S. (1990). "Feature-oriented domain analysis (FODA) feasibility study," in *Technical Report CMU/SEI-90-TR-21*. Ed. P. A. Pittsburgh (Software Engineering Institute, Carnegie Mellon University).

Lee, K., Kang, K. C., and Lee, J. (2002). "Concepts and guidelines of feature modeling for product line software engineering," in *International Conference on Software Reuse* (Berlin, Heidelberg: Springer), 62–77. doi: 10.1007/3-540-46020-9_5

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20 (2), 265–272. doi: 10.1101/gr.097261.109

Mount, D. W. (2005). *Bioinformatics Sequence and Genome Analysis*. (New York: Cold Spring Harbor Laboratory Press). doi: 10.1242/jcs.00197

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48 (3), 443–453. doi: 10.1016/0022-2836(70)90057-4

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-Coffee: anovel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302 (1), 205–217. doi: 10.1006/jmbi.2000.4042

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28 (11), 1420–1428. doi: 10.1093/bioinformatics/bts174

Shi, H. H., and Xue, J. Y. (2009). PAR-based formal development of algorithms. *Chin. J. Comput.* 32 (5), 982–991. doi: 10.3724/SP.J.1016.2009.00982

Shi, H. H., and Zhou, W. X. (2019). Design and implementation of pairwise sequence alignment algorithm components based on dynamic programming. *J. Comput. Res. Dev.* 56 (9), 1907–1917. doi: 10.7544/issn1000-1239.2019.20180835

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539. doi: 10.1038/msb.2011.75

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147 (1), 195–197. doi: 10.1016/0022-2836(81)90087-5

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (22), 4673–4680. doi: 10.1093/nar/22.22.4673

Turner, C. R., Fuggetta, A., Lavazza, L., and Wolf, A. L. (1998). "Feature engineering," in *Proceedings of International Workshop on Software Specification and Design*. (Ise-Shima, Japan), 162–164. doi: 10.1109/IWSSD.1998.667935

Wang, L., and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1 (4), 337–348. doi: 10.1089/cmb.1994.1.337

Wang, Y., and Li, K. B. (2004). An adaptive and iterative algorithm for refining multiple sequence alignment. *Comput. Biol. Chem.* 28 (2), 141–148. doi: 10.1016/j.compbiolchem.2004.02.001

Xue, J. Y. (1993). Two new strategies for developing loop invariants and their applications. *J. Comput. Sci. Tech.* 8 (2), 147–154. doi: 10.1007/BF02939477

Xue, J. Y. (1997). A unified approach for developing efficient algorithmic programs. *J. Comput. Sci. Tech.* 12 (4), 314–329. doi: 10.1007/BF02943151

Xue, J. Y. (1998). Formal derivation of graph algorithmic programs using partition-and-recur. *J. Comput. Sci. Tech.* 13 (6), 553–561. doi: 10.1007/bf02946498

Xue, J. Y. (2016). "Genericity in PAR platform," in *International Workshop on Structured Object-Oriented Formal Language and Method* (Cham: Springer), 3–14. doi: 10.1007/978-3-319-31220-0_1

Zhang, W., and Mei, H. (2003). A feature-oriented domain model and its modeling process. *J. Softw.* 14 (8), 1345–1356. doi: 10.13328/j.cnki.jos.2003.08.001

Zhang, J., Guo, M. Z., and Wang, Y. D. (2005). A heuristic algorithm for multiple sequence alignment base on progressive multiple alignment. *China J. Bioinf.* 3 (4), 171–174. doi: 10.3969/j.issn.1672-5565.2005.04.008

Zhang, Z. T. (2000). The current status and the prospect of bioinformatics. *World Sci. Tech. R. D.* 22 (6), 17–20. doi: 10.3969/j.issn.1006-6055.2000.06.004

Zou, Q., Hu, Q., Guo, M., and Wang, G. (2015). HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics* 31 (15), 2475–2481. doi: 10.1093/bioinformatics/btv177

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 February 2020
doi: 10.3389/fgene.2020.00075

DSPLMF: A Method for Cancer Drug Sensitivity Prediction Using a Novel Regularization Approach in Logistic Matrix Factorization

Akram Emdadi[1] *and Changiz Eslahchi*[1,2]*

[1] Department of Computer Sciences, Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran, [2] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

The ability to predict the drug response for cancer disease based on genomics information is an essential problem in modern oncology, leading to personalized treatment. By predicting accurate anticancer responses, oncologists achieve a complete understanding of the effective treatment for each patient. In this paper, we present DSPLMF (**D**rug **S**ensitivity **P**rediction using **L**ogistic **M**atrix **F**actorization) approach based on Recommender Systems. DSPLMF focuses on discovering effective features of cell lines and drugs for computing the probability of the cell lines are sensitive to drugs by logistic matrix factorization approach. Since similar cell lines and similar drugs may have similar drug responses and incorporating similarities between cell lines and drugs can potentially improve the drug response prediction, gene expression profile, copy number alteration, and single-nucleotide mutation information are used for cell line similarity and chemical structures of drugs are used for drug similarity. Evaluation of the proposed method on CCLE and GDSC datasets and comparison with some of the state-of-the-art methods indicates that the result of DSPLMF is significantly more accurate and more efficient than these methods. To demonstrate the ability of the proposed method, the obtained latent vectors are used to identify subtypes of cancer of the cell line and the predicted IC50 values are used to depict drug-pathway associations. The source code of DSPLMF method is available in https://github.com/emdadi/DSPLMF.

Keywords: cancer, drug response, recommender system, matrix factorization, personalized treatment

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Jing Lu,
Walmart Labs, United States
Mufeng Hu,
AbbVie, United States

***Correspondence:**
Changiz Eslahchi
Ch-Eslahchi@sbu.ac.ir

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *11 November 2019*
Accepted: *23 January 2020*
Published: *27 February 2020*

Citation:
Emdadi A and Eslahchi C (2020) DSPLMF: A Method for Cancer Drug Sensitivity Prediction Using a Novel Regularization Approach in Logistic Matrix Factorization. Front. Genet. 11:75. doi: 10.3389/fgene.2020.00075

INTRODUCTION

Cancer is a genetic disease that results when cellular changes and accumulation of different types of mutations cause the uncontrolled growth and division of cells. There are more than 200 different types of cancer, having a significant global impact on public health. Since cancer is a disease of genetic complexity and diversity, the drug response for different patients can be different. The main reason for this occurrence is the difference in the molecular and genetic information of individuals, such as gene expression data, the type of mutation in the genome and copy number alteration

information. These findings and achievements have recently made a significant challenge in the prediction of drug response for an individual patient in the research of precision medicine.

High-throughput drug screening technologies on several panels of cancer cell lines have been provided. For instance, two recent consortiums Genomics of Drug Sensitivity in Cancer (GDSC) Yang et al. (2012) and Cancer Cell Line Encyclopedia (CCLE) Barretina et al. (2012) have collected around 1,000 cell lines and their pharmacological profiles for several cancer drugs. The IC50 measure (minimal concentration of drug that induced 50% cell line death) is usually used as a sensitivity measure. To facilitate and speed up drug discovery and prediction process, many methods have been developed in these fields by researches from numerous domains such as computational biology, machine learning, and data mining approaches.

In the challenge of the DREAM project, the performance of 44 drug response prediction algorithms was considered for breast cancer cell lines. The introduced algorithms were evaluated using the weighted probabilistic c-index (WPC-index) and resampled Spearman correlation Costello et al. (2014). Various machine learning methods have been proposed in this area. Barretina et al. proposed a method for predicting drug response based on naive Bayes classifier that selected importance features by two steps. First, they used Wilcoxon Sum Rank Test and Fisher Exact Test to select the 30 top features and then they applied naive Bayes classifier for drug response prediction Barretina et al. (2012). SVM-RFE method is a wrapper that used SVM classifier and recursive feature selection method Dong et al. (2015). FSelector method used *k*-nearest neighbor (KNN) algorithm based on selected features that are achieved by information entropy Soufan et al. (2015). Suphavilai et al. (2018) proposed the CaDRReS method as a predictor cancer drug response model based on the recommender system and learning projections for drugs and cell lines into a latent space. AutoBorutaRF was presented by Xu et al., based on feature selection for classification of anticancer drug responses. The method first built a subset of essential features, then used Boruta algorithms Kursa et al. (2010) to select some features for applying Random-Forest classifier to predict drug response Lu et al. (2019).

In this paper, we modeled the cancer drug sensitivity problem based on "Recommender Systems" approach. A logistic matrix factorization algorithm was used for predicting drug cancer response. By applying the proposed model to GDSC and CCLE datasets, we proved that DSPLMF is of excellent prediction accuracy.

MATERIALS AND METHOD

Datasets

The performance of drug response prediction algorithms was evaluated on two benchmark datasets, including GDSC and CCLE. The datasets were downloaded by using R package PharmacoGx Smirnov et al. (2015). In these datasets, there are several types of information such as IC50 values according to the set of cell lines and drugs and some other information such as gene expression profile, copy number alteration, and single-nucleotide mutation that used in the model designing for more efficiency. Since in these datasets some of the above information is missing, the method of compensating for missing values given by Lu et al. (2019) is used. The missing value for a cell line can belong to response value, copy number alteration, and single-nucleotide mutation features. The cell lines with more than 50% missing value were removed from the dataset and for remaining, the missing values were predicted from the known values of *k*-nearest cell lines. At the end, 555 cell lines and 98 drugs remain without any missing value for GDSC and 363 cell lines and 24 drugs for CCLE datasets.

Method

The main idea of the model DSPLMF is to construct a classification model for predicting how a cell line responds to a drug. Since drug response can be divided into two classes "sensitivity" and "resistance," there are many ways for the purpose of classification based on IC50 values. By considering the histograms of IC50, we observed some histograms are normal-like, and others have skewness. Also, it can be supposed that the labels of classes should be determined by the data of individual drugs. For normal-like histograms, median, and mean are the same. If the histogram is skewed right, the mean is greater than the median, and if the histogram is skewed left, the mean is smaller than the median. We chose medium because we wanted to set a single, universal standard threshold for all drugs. So, the strategy introduced by Li et al. (2015) was used and the median of IC50 values were applied as a threshold for classification. The "sensitivity" or class with label 1 was assigned to a cell line if its IC50 is smaller than the median of cell lines for an individual drug and "resistance" or class with label 0 to a cell line was assigned, otherwise. DSPLMF method has four main steps as follows.

In the first step, by converting the model to a classification problem, a 0,1 observation matrix was achieved, as cell lines and drugs are rows and columns of the matrix, respectively. Then, a logistic matrix factorization method for constructing the latent vectors for each cell line and drug is applied. In the second step, for improving the prediction accuracy of the model, the similarity information for cell lines and drugs are used. In the third step, a model is applied to learn to predict the probability that a new cell line would sensitive to a drug. Subsequently, with applying the threshold to predicted probabilities of the cell line-drug pairs, we classified each pair to sensitive or resistance class. In the next section, first the similarity matrices used in the model, were introduced and then the details of each step are explained in the following steps. The main scheme of DSPLMF algorithm is represented in **Figure 1**.

Similarity Matrix

Cell Line Similarity

In this part, the four similarities between each pair of cell lines based on the information of gene expression, single-nucleotide mutation, copy number alteration, and IC50 values were defined.

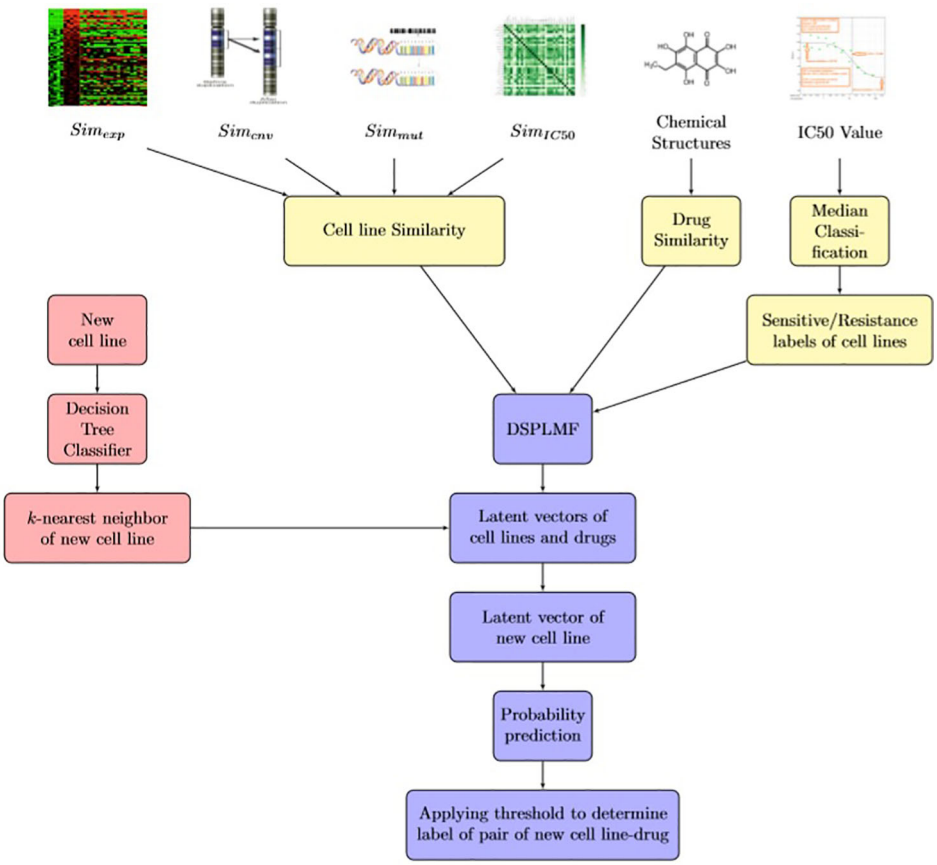


FIGURE 1 | Scheme of DSPLMF algorithm. First, similarities between each pair of cell lines are constructed based on the information of gene expression, single-nucleotide mutation, copy number alteration, and IC50 values. Also, similarity between each pair of drugs is defined based on chemical substructure and the median of IC50 values are applied as a threshold for classification. Using DSPLMF model, the latent vectors for each cell line and drug are achieved. For each new cell line, decision tree classifier is applied to find its *t*-most nearest neighbors and the probabilities that this cell line is sensitive to drugs are estimated based on the latent vectors of its neighbors. Eventually, a threshold is applied on probabilities to assign sensitive or resistance class to each new cell line-drug pair.

- **Gene expression Similarity,** $\mathrm{Sim_{exp}}$ Gene expression information is an auxiliary feature for similarity between cell lines. Let e_i denoted the gene expression vector of cell line c_i in cancerous conditions. For pair of cell lines c_i and c_j, $Sim_{exp}(c_i, c_j)$ is defined as the Pearson correlation between the vectors e_i and e_j and the gene expression similarity matrix between cell lines considered as $Sim_{exp} = [Sim_{exp}(c_i, c_j)]_{n \times n}$, where n is the numbers of cell lines. Each entry of these metrics is in [−1,1]. The numbers of considered genes for two datasets GDSC and CCLE for similarity measure are 11,712 and 19,389, respectively. So the length of vector e_i is 11,712 and 19,389 for GDSC and CCLE dataset, respectively.Q[SpecialChar] Verify that all the equations and special characters are displayed correctly.
- **Single-nucleotide mutation Similarity**, $\mathrm{Sim_{mut}}$ Let zero-one vectors m_i indicate that whether a mutation occurred in the set of genes for cell line c_i or not. $Sim_{mut}(c_i, c_j)$ is defined as the Jaccard similarity between the vectors m_j and m_j and the single-nucleotide mutation similarity matrix between cell lines considered as $Sim_{mut} = [Sim_{mut}(c_i, c_j)]_{n \times n}$. Each entry of these metrics is in [0, 1]. The mutation information of 54 genes are accessible for cell lines in GDSC dataset and 1667 genes for cell lines in CCLE dataset, respectively.
- **Copy number alteration Similarity,** $\mathrm{Sim_{cnv}}$ Let v_i denoted the copy number alteration vector for cell line c_i. $Sim_{cnv}(c_i, c_j)$ is defined as the Pearson correlation between the vectors v_i and v_j and the copy number alteration similarity matrix between cell lines considered as $Sim_{cnv} = [Sim_{cnv}(c_i, c_j)]_{n \times n}$. Each entry of these metrics is in [−1, 1]. The information of copy number alteration of 24,959 and 24,960 genes for two GDSC and CCLE datasets are accessible, respectively.
- **IC50 value Similarity,** $\mathrm{Sim_{IC50}}$ Moreover, the similarity between cell lines proposed by Liu et al. (2018) based on the correlation between their response IC50 values was used. Let IC_i denoted the vector of IC50 values of drugs in cell line c_i. $Sim_{IC50}(c_i, c_j)$ is defined as the Pearson correlation between the vectors IC_i and IC_j and the similarity based on IC50 matrix between cell lines considered as $Sim_{IC50} = [Sim_{IC50}(c_i, c_j)]_{n \times n}$ and each element of these metrics in [−1, 1].

To aggregate these similarities to a single matrix, $Sim_{total} = [SC_{ij}]_{n \times n}$, the following formula is used:

$$Sim_{total} = \frac{\lambda Sim_{exp} + \gamma Sim_{cnv} + \phi Sim_{mut} + \psi Sim_{IC50}}{\lambda + \gamma + \phi + \psi} \quad (1)$$

where γ, λ, ϕ and ψ are parameters that represent the importance of each of the matrix and tuned in the model. The numbers of considered genes for two datasets GDSC and CCLE for Sim_{exp} are 11,712 and 19,389, respectively. The mutation information of 54 genes is accessible for cell lines in GDSC dataset and 1,667 genes for cell lines in CCLE *dataset*. The information of copy number alteration of 24,959 and 24,960 genes for two GDSC and CCLE datasets are accessible, respectively. Since three matrices Sim_{exp}, Sim_{cnv}, and Sim_{mut} have been constructed by different sets of genes (the number of common genes between them is about 50%), there is not an additive relation between them. In general, an absolute correlation coefficient of >0.7 among two or more predictors indicates the presence of collinearity. But as **Table 1** shows, all correlation coefficients between similarity matrices are very low, so there is not collinearity between matrices and they can be linearly combined.

Drug Similarity, Sim_{drug}

Since it is expected that similar drugs have the same effect on cell lines, drug similarity information for predicting drug response was used in the proposed method. A drug can be represented as a binary feature vector, by using drug substructures, drug transporters, drug targets, drug enzymes, drug pathways, drug indications, or drug side effects information. Since there is only information about chemical substructures, for each drug we have a zero-one vector of size 881, where 881 is the number of known chemical substructures of a drug. In this vector one indicates the presence of a substructure of drug and zero otherwise. We downloaded the substructure for each drug from PubChem. The PubChem system generates a binary substructure fingerprint for chemical structures. These fingerprints are used by PubChem for similarity neighboring and similarity searching. Let V_{d_i} and V_{d_j} are the vectors correspond to the drugs d_i and d_j. Similarity (di, dj) is considered as Jaccard similarity between these two vectors. We construct the matrix $Sim_{drug} = [SD_{ij}]_{m \times m}$ as similarity matrix between each pair of drugs.

Logistic Matrix Factorization

Assume the set of cell lines is denoted by C = {c_1, c_2, …, c_n } and the set of drugs is denoted by D = {d_1, d_2, …, d_m }, where n and m are the numbers of cell lines and the numbers of drugs, respectively. The relationship between cell lines and drugs are represented by a binary matrix $Q = [q_{ij}]_{n \times m}$, where each element $q_{ij} \in \{0, 1\}$. If a cell line is c_i sensitive to a drug d_j, $q_{ij} = 1$ and otherwise $q_{ij} = 0$. The probability of sensitivity of a cell line to a drug is defined by a logistic function as follows:

$$p_{ij} = \frac{exp\ (u_i v_j^T + \beta_i^c + \beta_j^d)}{1 + exp\ (u_i v_j^T + \beta_i^c + \beta_j^d)} \quad (2)$$

where u_i nd v_j are the latent vectors of size L corresponding to i-th cell line and j-th drug, respectively and the latent vectors of all cell lines and all drugs are denoted by U and V, respectively. On the other hands, the non-negative values β_i^c and β_j^d are the bias parameters according to cell line i and drug j, respectively. Moreover, we denoted $\beta^c \in \mathbb{R}^{n \times 1}$ and $\beta^d \in \mathbb{R}^{m \times 1}$ as bias vectors for cell lines and drugs, respectively. Bias parameters are considered because some cell lines respond significantly to many drugs and there are cell lines that respond to few drugs. Similarly for some drugs, there are many cell lines that respond to them, and there are drugs that most cell lines do not respond to significantly. Thus, by applying these parameters, we try to reduce bias. The vectors $\beta^c = (\beta_1^c, \ldots, \beta_n^c)$and $\beta^d = (\beta_1^d, \ldots, \beta_m^d)$ considered as bias vector of the model.

In this model, all the data in the training set are assumed to be independent. So the probability that matrix Q occurred, considering the latent and bias vectors, can be computed as:

$$p(Q|U, V, \beta^c, \beta^d) = \left(\prod_{1 \le i \le n, 1 \le j \le m, q_{ij}=1} [p_{ij}^{q_{ij}} (1 - p_{ij})^{(1-q_{ij})}]^r \right) \times \left(\prod_{1 \le i \le n, 1 \le j \le m, q_{ij}=0} p_{ij}^{q_{ij}} (1 - p_{ij})^{(1-q_{ij})} \right). \quad (3)$$

When $q_{ij} = 1$ then both $r\ (1 - q_{ij})$ and $1 - q_{ij}$ are zero. Similarly, when $q_{ij} = 0$, $rq_{ij} = q_{ij} = 0$. So, formula 3 is rewritten as follows:

$$p(Q|U, V, \beta^c, \beta^d) = \left(\prod_{1 \le i \le n, 1 \le j \le m, q_{ij}=1} p_{ij}^{rq_{ij}} (1 - p_{ij})^{(1-q_{ij})} \right) \times \left(\prod_{1 \le i \le n, 1 \le j \le m, q_{ij}=0} p_{ij}^{rq_{ij}} (1 - p_{ij})^{(1-q_{ij})} \right). \quad (4)$$

Finally, the above probability is shown as follows:

$$p(Q|U, V, \beta^c, \beta^d) = \prod_{i=1}^{n} \prod_{j=1}^{m} p_{ij}^{rq_{ij}} (1 - p_{ij})^{(1-q_{ij})}. \quad (5)$$

Where ($r \ngeq 1$) is used to control the importance levels of observed interactions. In some classification problems with two classes (0 and 1), lack of information make us to assign label zero to some objects. But, it may be that the real label of these objects are one. So, the members of class one are highly trusted, while some members assign to class zero because of lack of information. As an example, in drug-target prediction or drug-drug interaction prediction models, the observed interacting drug-target pairs or drug-drug pairs have been experimentally verified; thus, they are

TABLE 1 | Correlation coefficient between four matrices Sim_{exp}, Sim_{cnv}, Sim_{mut}, and Sim_{IC50}.

Correlation Coefficient	Sim_{exp}	Sim_{cnv}	Sim_{mut}	Sim_{IC50}
Sim_{exp}	1.0	0.24	–0.11	0.19
Sim_{cnv}	0.24	1.0	0.14	0.015
Sim_{mut}	–0.11	0.14	1.0	–0.06
Sim_{IC50}	0.19	0.015	–0.06	1.0

more trustworthy and important than the unknown pairs. Toward more accurate modeling for these prediction models, the authors can assign higher importance levels to the interaction pairs than unknown pairs. This importance weighting strategy (considering $r > 1$) has been demonstrated to be effective for personalized recommendations. On the other hand, in DSPLMF model, both classes (sensitivity and resistance) have the same importance and validity. So, we set r to be one.

We also deposited zero-mean spherical Gaussian priors on latent vectors of cell lines and drugs as:

$$p(U|\sigma_c^2) = \prod_{i=1}^{n} \mathcal{N}(u_i|0, \sigma_c^2 I) \tag{6}$$

$$p(V|\sigma_d^2) = \prod_{j=1}^{m} \mathcal{N}(v_j|0, \sigma_d^2 I) \tag{7}$$

where I denotes the identity matrix and σ_c^2 and σ_d^2 are parameters for controlling the variances of prior distributions of cell lines and drugs. Based on Bayesian theorem we have:

$$p(M|Q) = \frac{p(Q|M)p(M)}{p(Q)} . \tag{8}$$

Since U, V, β^c, β^d are the parameters in the model M, Bayesian theorem is as follows:

$$p(U, V, \beta^c, \beta^d|Q) = \frac{p(Q|U, V, \beta^c, \beta^d)p(U|\sigma_c^2)p(V|_{\sigma}^{d}2)}{p(Q)} . \tag{9}$$

So we can conclude the following relation:

$$p(U, V, \beta^c, \beta^d|Q) \propto p(Q|U, V, \beta^c, \beta^d)p(U|\sigma_c^2)p(V|\sigma_d^2) . \tag{10}$$

According to the Bayesian theorem and equations 5, 6, and 7, the log of the posterior distribution is estimated as follows:

$$\begin{aligned} \log\, p(U, V, \beta^c, \beta^d|Q, \sigma_c^2, \sigma_d^2) = \sum_{i=1}^{n}\sum_{j=1}^{m}[rq_{ij}(u_i v_j^T + \beta_i^c + \beta_j^d) - \\ \left(1 + rq_{ij} - q_{ij}\right) \log\, (1 + \exp\, (u_i v_j^T + \beta_i^c + \beta_j^d))] - \\ \frac{\lambda_c}{2}\sum_{i=1}^{n}||u_i||_2^2 - \frac{\lambda_d}{2}\sum_{j=1}^{m}||v_j||_2^2 + T \end{aligned} \tag{11}$$

In formula 11, regarding how Bayesian theorem is applied to classification problems, we could convert the direct proportional relation between the left hand side and the numerator of the fraction of equation 10 to equalized, by adding constant term T to the formula. Where T is independent of the model parameters Hand et al. (1999). $\lambda_c = \frac{1}{\sigma_c^2}$, $\lambda_d = \frac{1}{\sigma_d^2}$. The parameters of the model can be learned by maximizing the above formula, which is equivalent to minimizing the following objective function:

$$\begin{aligned} \min_{U,V,\beta^c,\beta^d} \sum_{i=1}^{n}\sum_{j=1}^{m}[(1 + rq_{ij} - q_{ij}) \log\, (1 + \exp\, (u_i v_j^T + \beta_i^c + \beta_j^d)) - \\ rq_{ij}(u_i v_j^T + \beta_i^c + \beta_j^d)] + \frac{\lambda_c}{2}||U||_F^2 + \frac{\lambda_d}{2}||V||_F^2 \end{aligned} \tag{12}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of matrix.

For regularization the objective function 12, for each cell line c_i, we choose the set $N_k(c_i)$ that denotes the k-most similar cell lines to c (except c_i) using Sim_{total} matrix. We constructed adjacency matrix $A = [a_{ij}]_{nxn}$ that represents cell line neighborhood information as follow:

$$a_{ij} = \begin{cases} SC_{ij} & c_j \in N_k(c_i) \\ 0 & otherwise \end{cases} . \tag{13}$$

A is an $n \times n$ matrix, which for the row corresponding to cell line c_i, the entries of columns corresponding to the k-most similar cell lines of c_i are obtained from their similarities, Sim_{total} matrix, and the other elements of this row are zero.

Similarly, for a drug d_i, the set $N_k(di)$ denotes the k-most similar drugs to d_i (except d_i) using Sim_{drug} matrix. The adjacency matrix B to describe the drug neighborhood information is denoted by $B = [b_{ij}]_{m\times m}$, where;

$$b_{ij} = \begin{cases} SD_{ij} & d_j \in N_k(d_i) \\ 0 & otherwise \end{cases} . \tag{14}$$

B is an $m \times m$ matrix, which for the row corresponding to drug d_i, the entries of columns corresponding to the k-most similar drugs of d_i are obtained from their similarities, Sim_{drug} matrix, and the other elements of this row are zero.

To illustrate the data structure of these similarity matrices, as an example, for $k = 5$ and 24 drugs in CCLE dataset, the similarity matrix B is denoted in **Figure 2A**. **Figure 2B**, shows the graph corresponding to this matrix. As it can be seen from **Figure 2A**, each row i of the matrix has five nonzero elements corresponding to the five-most similar drugs of d_i in Sim_{drug} matrix, and the other elements are zero. In **Figure 2B**, the degree of each node is five and the red edges denote the neighbors of the nutlin-3. 5-most similar drugs to Nutlin-3 based on sim drug matrix are AEW541, AZD0530, Lapatinib, crizotinib, and sorafenib.

To minimize the distance between feature vector corresponding to cell line i and vectors of its nearest neighbors in latent space, we minimize two objective functions in formulas 15, 16 as follows:

$$\begin{aligned} \frac{\alpha}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij}||u_i - u_j||_F^2) \\ = \frac{\alpha}{2}[\sum_{i=1}^{n}(\sum_{j=1}^{n}a_{ij})u_i u_i^T + \sum_{j=1}^{n}(\sum_{i=1}^{n}a_{ij})u_j u_j^T] - \frac{\alpha}{2}tr(U^T A U) - \\ \frac{\alpha}{2}tr(U^T A^T U) = \frac{\alpha}{2}tr(U^T H^c U) \end{aligned} \tag{15}$$

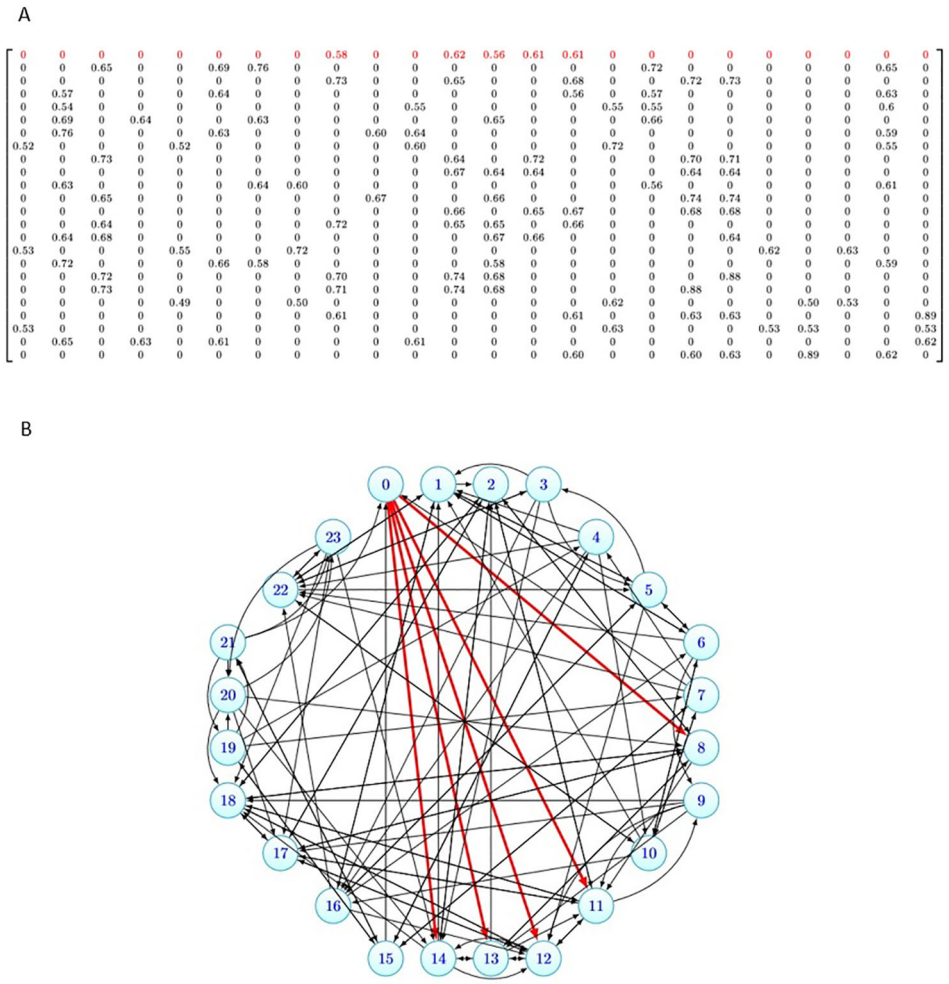


FIGURE 2 | Data structure of the similarity matrix for 24 drugs in Cancer Cell Line Encyclopedia (CCLE) dataset. **(A)** The similarity matrix $B_{24 \times 24}$. **(B)** The graph corresponding to the similarity matrix $B_{24 \times 24}$.

$$\frac{\beta}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(b_{ij}||v_i - v_j||_F^2)$$

$$= \frac{\beta}{2}\left[\sum_{i=1}^{m}(\sum_{j=1}^{m}b_{ij})v_i v_i^T + \sum_{j=1}^{m}(\sum_{i=1}^{m}b_{ij})v_j v_j^T\right] - \frac{\beta}{2}tr\left(V^T B V\right) - \tag{16}$$

$$\frac{\beta}{2}tr\left(V^T B^T V\right) = \frac{\beta}{2}tr\left(V^T H^d V\right)$$

Where tr(.) is the trace of a matrix, $H^c = (E^c + \tilde{E}^c) - (A + A^T)$, E^c and $\tilde{E}^c$ are two diagonal matrices and their diagonal elements are $E^c_{ii} = \sum_{j=1}^{n}(a_{ij})$ and $\tilde{E}^c_{jj} = \Sigma_{i=1}^{n}(a_{ij})$, $H^d = (E^d + \tilde{E}^d) - (B + B^T)$. E^d and $\tilde{E}^d$ are two diagonal matrices and their diagonal elements are $E^d_{ii} = \sum_{j=1}^{m}(b_{ij})$ and $E^d_{jj} = \sum_{i=1}^{m}(b_{ij})$. α and β are two parameters for weighting the similarity between cell lines and drugs, respectively.

The values of two matrices A and B show the similarity of the cell lines to each other and the similarity of the drugs to each other, respectively. Using the calculation of Frobenius norm multiplied by the elements in A and B is because of we would like more similar cell lines (drugs), have closer latent vectors in the latent space. But, the parameters α and β determine the effectiveness of these two matrices A and B in the objective function. By these strategies and tuning the parameters α and β, we determine the impact of cell line similarity and drug similarity in DSPLMF method.

By plugging two equations 15, 16 into formula 12, we will have the following:

$$\min_{U,V,\beta^c,\beta^d} \sum_{i=1}^{n}\sum_{j=1}^{m}(1 + rq_{ij} - q_{ij})\log\left(1 + \exp\left(u_i v_j^T + \beta_i^c + \beta_j^d\right)\right) -$$

$$rq_{ij}(u_i v_j^T + \beta_i^c + \beta_j^d) + \frac{\lambda_c}{2}||U||_F^2 + \frac{\alpha}{2}tr(U^T H^c U) + \frac{\lambda_d}{2}||V||_F^2 +$$

$$\frac{\beta}{2}tr(V^T H^d V) \tag{17}$$

Finally, we upgrade the formula 17 as follows:

$$\min_{U,V,\beta^c,\beta^d} \sum_{i=1}^{n}\sum_{j=1}^{m}(1+rq_{ij}-q_{ij})\log\,(1+\exp\,(u_i v_j^T+\beta_i^c+\beta_j^d))-r\,q_{ij}(u_i v_j^T+\beta_i^c+\beta_j^d)+\frac{1}{2}tr[U^T(\lambda_c I+\alpha H^c)U]+\frac{1}{2}tr[V^T(\lambda_d I+\beta H^d)V] \quad (18)$$

By this function, we try to predict the latent vectors of cell lines and drugs, where the similar cell lines or drugs have closer latent vectors to their KNNs.

For optimization the above function, the alternating gradient descent method was used. In each iteration of this algorithm, first U and β_i^c are fixed to compute V and β_j^d and then V and β_j^d are fixed to compute U and β_i^c. Besides, to accelerate the convergence, the AdaGrad algorithm was applied and the details of this algorithm are deposited in the **Supplementary File 3 (Data Sheet 3)**. The objective function in formula 18 is denoted by Y and the partial gradients of biases and latent vectors are calculated as follow:

$$\frac{\partial Y}{\partial u_i}=\sum_{j=1}^{m}\frac{v_j^T(1+rq_{ij}-q_{ij})(exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}{(1+exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}-rq_{ij}v_j^T+(\lambda_c u_i+\alpha H_{ij}^c u_i)$$

$$\frac{\partial Y}{\partial v_j}=\sum_{i=1}^{n}\frac{u_i(1+rq_{ij}-q_{ij})(\exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}{(1+exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}-rq_{ij}u_i+(\lambda_d v_j+\beta H_{ij}^d v_j)$$

$$\frac{\partial Y}{\partial \beta_i^c}=\sum_{j=1}^{m}\frac{(1+rq_{ij}-q_{ij})(\exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}{(1+\exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}-rq_{ij}$$

$$\frac{\partial Y}{\partial \beta_j^d}=\sum_{i=1}^{n}\frac{(1+rq_{ij}-q_{ij})(\exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}{(1+\exp\;\;(u_i v_j^T+\beta_i^c+\beta_j^d))}-rq_{ij} \quad (19)$$

Once the latent matrices U and V and the biases β_i^c and β_j^d have been learned, the probability of sensitivity cell line i to drug j can be estimated by logistic function in formula 2. Since in our model, the importance of the positive observations and negative observations are the same, we set $r = 1$ in this logistic function.

Prediction

When a new cell line is given, its information of IC50 of the drugs is unknown and Sim_{IC50} matrix values cannot be calculated, while it must be calculated to predict the latent vectors of this new cell line. In this section, we introduced a classification model for predicting t-most nearest neighbors by using the similarity values between cell lines which are obtained from gene expression profile, copy number alteration and single-nucleotide mutation information. The purpose of this model is to find t-most nearest neighbors for the new cell line and then to estimate the latent vector for this new cell line based on average of latent vectors of its neighbors. After obtaining the latent vector, we can predict the IC50 values across all drugs for the new cell line. For training the model, 10-fold cross validation technique is used on cell line dataset, so the dataset was partitioned into 10 equal-sized subsets, nine subsets were used as the train set for learning this classification model. A single subset was used as the test set to predict the t-most nearest neighbors for each cell line of this set.

In this classification model, the amounts of Sim_{IC50} matrix of train set were converted to 0 or 1. To do this, the values of each row of the matrix are sorted in descending order and then t-largest values are set to 1 and remaining values are set to 0. Among the methods available for classification, we chose "Decision Tree Classifier" method. It is one of the predictive modeling approaches that used tree models to predict the value of a target variable based on several input features. Where leaves represent class labels and branches denote conjunctions of features that lead to those class labels. Learned trees can be represented as sets of if-then rules. Decision tree classifier is a heuristic and nonbacktracking search through the space of all possible decision trees. The main idea of decision tree classification is recursively partition data into subgroups. The functionality of decision tree classification is as follows: Polat and Güneş (2007)

- Choosing an attribute and formulating a logical attribute test.
- Branching on each test result, transferring subset of examples (training information) to the appropriate child node to satisfy that result.
- Running each child's node recursively.
- The end rule indicates when a leaf node is to be declared.

For decision tree classifier, the three features of train set, Sim_{exp}, Sim_{cnv}, and Sim_{mut}, are considered as input and 0 or 1 value of each pair (c_i, c_j) are considered as output and then as the classifier train. If the number of predicted nearest neighbors for a cell line was less than t, we considered them as nearest neighbors for this cell line. If this number was greater than t, t neighbors were selected randomly. Finally, u_i was estimated as the average of latent vectors of neighbors of the new cell line c_i.

When the latent vector of the new cell line is predicted, the probabilities that this cell line is sensitive to drugs are estimated. Eventually, a threshold on probabilities to assign sensitive or resistance class to each cell line-drug pair is applied. So if the predicted value is lower than this threshold for a cell line-drug pair, the resistance class is assigned to it; otherwise, it is labeled as a sensitive class.

RESULT

We empirically evaluate our proposed approach and compare it against some of the state-of-the-art methods. This section first describe evaluation criteria and then demonstrate the performance of DSPLMF method.

Evaluation Criteria

To evaluation the performance of DSPLMF method, the 10-fold cross-validation Was performed and this process was repeated 30 times. The mean of following criteria was obtained in the 30

times and it was used as the final criteria to evaluate the predictive performance of the methods.

$$\begin{aligned} Accuracy &= \frac{TP + TN}{TP + FP + TN + FN} \\ Recall &= \frac{TP}{TP + FN} \\ Precision &= \frac{TP}{TP + FP} \\ Specificity &= \frac{TN}{TN + FP} \\ F_1Score &= \frac{2TP}{2TP + FP + FN} \\ MCC &= \frac{TP*TN - FP*FN}{\sqrt{(TP + FP)(TP + FN)(FP + TN)(FN + TN)}} \end{aligned} \tag{20}$$

where *TP* or true positive prediction is the number of cell lines labeled with sensitivity and predicted as sensitivity. *TN* or true negative is the number of cell lines labeled with resistance and predicted as resistance. *FP* or false positive is the number of cell lines labeled with resistance and predicted as sensitivity. *FN* or false negative is the number of cell lines labeled with sensitivity and predicted as resistance.

In addition to the above metrics, we used area under the receiver operating characteristic curve (*AUC*), which is one of the most important evaluation metrics for checking the performance of any classification model. This metric was calculated for the methods.

Comparison With the State-of-the-Art-Methods

To demonstrate the effectiveness of our method, we compared the predictive performance of the proposed model against the state-of-the-art-methods such as naive Bayes Barretina et al. (2012), SVM-RFE Dong et al. (2015), FSelector Soufan et al. (2015), CaDRReS Suphavilai et al. (2018), AutoBorutaRF Lu et al. (2019), and the AutoHidden method, which is constructed based on the hidden layer of the autoencoder in AutoBorutaRF method as features Lu et al. (2019).

All the methods mentioned above are classification models except the CaDRReS, since this method predicted IC50 values as output, a threshold was applied for its output. So if the value predicted for a cell line-drug pair is smaller than this threshold, the resistance class was assigned to it; otherwise, it was labeled with sensitive class. The median of the IC50 values was chosen as the best threshold for this algorithm. The results of the mentioned methods on two datasets GDSC and CCLE are shown in **Tables 2** and **3**, and the bold number represents the best result. The results of **Table 2** show that the value of *Accuracy* criterion by DSPLMF has increased by 0.03 compared to the result of the best algorithm, AutoBorutaRF. Furthermore, the value of *Recall*, F_1Score, *MCC*, and *AUC* criteria have increased by 0.10, 0.05, 0.06, and 0.05 compared to the best algorithm. Only in the case of the *Specificity* criterion, the naive Bayes method performs significantly better than the other methods. The reason is that this method has predicted zero class data for most of the data, and by looking at the result of other criteria, such as *Accuracy*, *Recall*, and F_1Score for this method, we can see that this method does not predict sensitive class data very well. The results of **Table 3** are the same as those in the previous table, except that the best result for the *AUC* criterion belongs to the AutoBorutaRF method, demonstrating the effectiveness of this method. The best result for the *Specificity* criterion belongs to the AutoHidden method; the low performance of other criteria indicates that this method is weak in predicting sensitive data. In general, the results of these two tables show that the DSPLMF significantly outperforms other methods. Thus, it is evident our method able to find much

TABLE 2 | Prediction performance of the different algorithms based on seven criteria on Genomics of Drug Sensitivity in Cancer (GDSC) dataset.

Method	*Accuracy*	*Recall*	*Precision*	*Specificity*	F_1Score	*MCC*	*AUC*
DSPLMF	**0.682**	**0.750**	**0.671**	0.615	**0.702**	**0.373**	**0.760**
CaDRReS	0.541	0.540	0.547	0.546	0.549	0.110	0.510
AutoBorutaRF	0.653	0.652	0.646	0.654	0.650	0.310	0.711
naive Bayes	0.610	0.424	0.590	**0.796**	0.494	0.247	0.679
SVM-RFE	0.594	0.579	0.589	0.609	0.585	0.191	0.515
FSelector	0.606	0.617	0.593	0.595	0.606	0.215	0.647
AutoHidden	0.578	0.557	0.571	0.598	0.565	0.158	0.609

The 10-fold cross validation is applied on the evaluation metrics and the mean value of them is used as criteria for comparison.

TABLE 3 | Prediction performance of the different algorithms based on seven criteria on Cancer Cell Line Encyclopedia (CCLE) dataset.

Method	*Accuracy*	*Recall*	*Precision*	*Specificity*	F_1Score	*MCC*	*AUC*
DSPLMF	**0.770**	**0.723**	**0.636**	0.772	**0.677**	**0.481**	0.776
CaDRReS	0.671	0.353	0.493	0.830	0.412	0.202	0.501
AutoBorutaRF	0.763	0.656	0.594	0.813	0.624	0.452	**0.821**
naive Bayes	0.683	0.332	0.406	0.919	0.366	0.275	0.779
SVM-RFE	0.728	0.428	0.631	0.812	0.523	0.296	0.551
FSelector	0.743	0.506	0.630	0.805	0.563	0.353	0.737
AutoHidden	0.697	0.133	0.201	**0.950**	0.356	0.219	0.706

The 10-fold cross validation is applied on the evaluation metrics and the mean value of them is used as criteria for comparison.

more useful features for drug response prediction rather than other methods. Overall, DSPLMF improvement on the GDSC dataset is stronger.

Performance of the Novel Regularization Approach

To evaluate the improvement of the logistic matrix factorization method by applying the novel regularization approach, we compared the predictive performance of the DSPLMF model against the logistic matrix factorization method without the novel regularization approach. In this model, the classification method for predicting t-most nearest neighbors for each new cell line by using the similarity values between cell lines which are obtained from gene expression profile, copy number alteration and single-nucleotide mutation information, is not applied. The result of the above algorithm based on seven criteria on GDSC and CCLE datasets is calculated, and the 10-fold cross-validation is applied on the evaluation metrics, and the mean value of them is used as criteria for comparison. The results of **Tables 2** and **4** show that the value of *Accuracy* criterion by DSPLMF on GDSC dataset has increased by 0.10 compared to the result of the logistic matrix factorization method without the novel regularization approach. Furthermore, the value of *Recall*, *Precision*, *Specificity*, F_1*Score*, *MCC*, and *AUC* criteria have increased by 0.04, 0.10, 0.17, 0.07, 0.21, and 0.14 compared to this algorithm. The results of **Tables 3** and **4** show that the value of *Accuracy* criterion by DSPLMF on CCLE dataset has increased by 0.10 compared to the result of the logistic matrix factorization method without the novel regularization approach. Furthermore, the value of *Recall*, *Precision*, *Specificity*, F_1*Score*, *MCC*, and *AUC* criteria has increased by 0.05, 0.11, 0.10, 0.09, 0.16, and 0.10 compared to this algorithm. So, using of the classification method for predicting t-most nearest neighbors of each new cell line in logistic matrix factorization algorithm, will increase the performance by 10%.

TABLE 4 | Prediction performance of the logistic matrix factorization method without the novel regularization approach based on seven criteria on Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) datasets.

Dataset	*Accuracy*	*Recall*	*Precision*	*Specificity*	F_1*Score*	*MCC*	*AUC*
GDSC	0.580	0.713	0.571	0.442	0.630	0.168	0.626
CCLE	0.672	0.673	0.523	0.670	0.582	0.328	0.671

The 10-fold cross validation is applied on the evaluation metrics and the mean value of them is used as criteria for comparison.

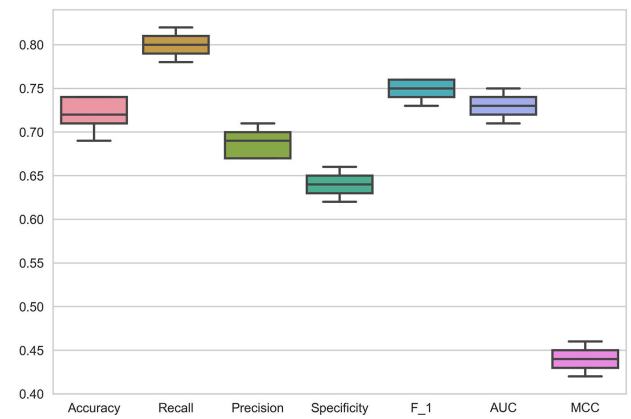


FIGURE 3 | Box Plots of seven criteria on haematopoietic cell lines in Genomics of Drug Sensitivity in Cancer (GDSC) dataset to show the prediction performance of DSPLMF method.

Specific Tissue of Cell Line Type

The data in the GDSC dataset is related to different cancers. To demonstrate the performance of DSPLMF method on cancer tissue type, 73 hematopoietic cell lines and 98 drugs from GDSC dataset are considered. This specific type of cell lines are used to train the proposed model and predicted responses for the drugs based on this tissue type. **Figure 3** shows the results of all mentioned criteria on these cell lines for the DSPLMF method using 30 times 10-fold cross-validation. The mean of these values are shown in **Table 5**. As the table shows, if the algorithm is specifically run on a particular type of cancer, it would be expected to yield better results than when considering different types of cancer. These results indicate that DSPLMF can also achieve consistent performance on a specific type of cancer.

Correlation Between Predicted and Observed Responses Values

For further evaluation and to demonstrate the performance of the proposed algorithm, the scatter plots of observed versus predicted responses values for four drugs in CCLE are illustrated in **Figure 4**. The values predicted by our model are probabilities that cell lines are sensitive to the drugs. For calculation correlation between predicted and observed responses values, the values $(u_i v_j^T + \beta_i^c + \beta_j^d)$ in Formula 2 as the predicted IC50 values for cell line c_i and drug d_j were used. As the plots indicate, there is a high correlation between observed and predicted response values. The scatter plots of all 24 drugs in the CCLE dataset are illustrated in the **Supplementary File 2 (Data Sheet 2: Figures S1–S4)**.

Learning Hyperparameters

For tuning hyperparameters, GDSC dataset has been used, and the obtained hyperparameters are considered for both datasets. The 10-fold cross-validation procedure is applied on GDSC and hyperparameters are chosen empirically by maximizing the summing up of the Accuracy, Recall, Precision, Specificity, F_1Score, and MCC criteria. For each set of hyperparameters, the whole 10-fold process is repeated 30 times and the average value of the above summing has been calculated. Since the search space of hyperparameters values is large, a grid-search procedure for choosing the hyperparameters was applied.

The dimension of latent space, L, was selected between 1 and 98, the number of drugs. The number of KNNs for building $N_k(c_i)$ in equation 13 and the number of t-nearest neighbors in prediction section, were selected from 1 to 50 by step 2. The impact factors of nearest neighbors α and β in equations 15 and 16 were picked from $\{2^{-5}, 2^{-4},\ldots, 2^{2}\}$ and the variance controlling parameters, λ_c and λ_d, were chosen from$\{2^{-5}, 2^{-4},\ldots, 2^{1}\}$. The γ, λ, ϕ and ψ parameters

TABLE 5 | Prediction performance of DSPLMF method on haematopoietic cell lines based on seven criteria on Genomics of Drug Sensitivity in Cancer (GDSC) dataset.

Method	*Accuracy*	*Recall*	*Precision*	*Specificity*	*F_1Score*	*MCC*	*AUC*
DSPLMF	0.721	0.800	0.690	0.645	0.750	0.441	0.730

The 10-fold cross validation is applied on the evaluation metrics and the mean value of them is used as criteria for comparison.

represent the importance of each similarity measure between cell lines in formula 1 and were selected from 1 to 10. Threshold parameter applied on equation 2 for determining the label of the class for each new cell line c_i, and was picked from 0.1 to 1 by step 0.1, and the best accuracy of the result is obtained by threshold=0.6.

In **Table 6**, the learned hyperparameters using GDSC dataset is shown. For both datasets, these tuned hyperparameters are used to design the model, except to L, that is calculated for CCLE dataset separately and for this dataset it is set as 23.

DISCUSSION

Cell Line Subtypes in Latent Space

We used 555 cell lines from different cancerous tissue types in GDSC dataset. For representing the higher similarity between latent vectors $\tilde{u}_i$ of the cell lines from the same tissue type rather than the cell lines from different tissue types, the t-SNE plot for some tissue types of cancer cell lines is shown in **Figure 5**. Top five most frequent tissue types including, breast, central nervous system, hematopoietic and lymphoid tissue, COREAD, and lung cancer were considered. As it can be seen from **Figure 5** (A), the embedded latent vectors of the cell lines with the same tissue type are located closer than the cell lines with diverse tissue types. This suggests that the proposed method assigned more similar latent vector to cell lines with the same tissue type. In the following, we consider an example of some latent vectors and the similarities between them: Let v_1, v_2 and v_3 are three latent vectors obtained DSPLMF method of length 95 corresponding to Breast cancer cell line $BT - 20$, Breast cancer cell line $BT - 549$ and hematopoietic cancer cell line $CA46$, respectively. v_1 = [0.01, 0.23, −0.14,…, 0.12]$_{1x95}$, v_2 = [0.17,0.67, −0.1,…,0.34]$_{1x95}$ and v_3 = [0.89, −0.9, 0.55,…, −0.17]$_{1x95}$. Similarity (v_1,v_2) = 0.78, Similarity (v_1,v_2) = 0.13 and Similarity(v_2,v_3) = 0.04. As the results show, two vectors belonging to the same tissue types are more similar than two vectors that belong to two different tissue types. Also, in the t-SNE plot, these two vectors belonging to the same tissue types are closer than two vectors that belong to two different tissue types.

In **Figure 5B**, the latent vectors of different subtypes of lung cancer were considered. These different subtypes are: adenocarcinoma, large cell, squamous cell, and small cell carcinoma. In this figure, the closeness of vectors

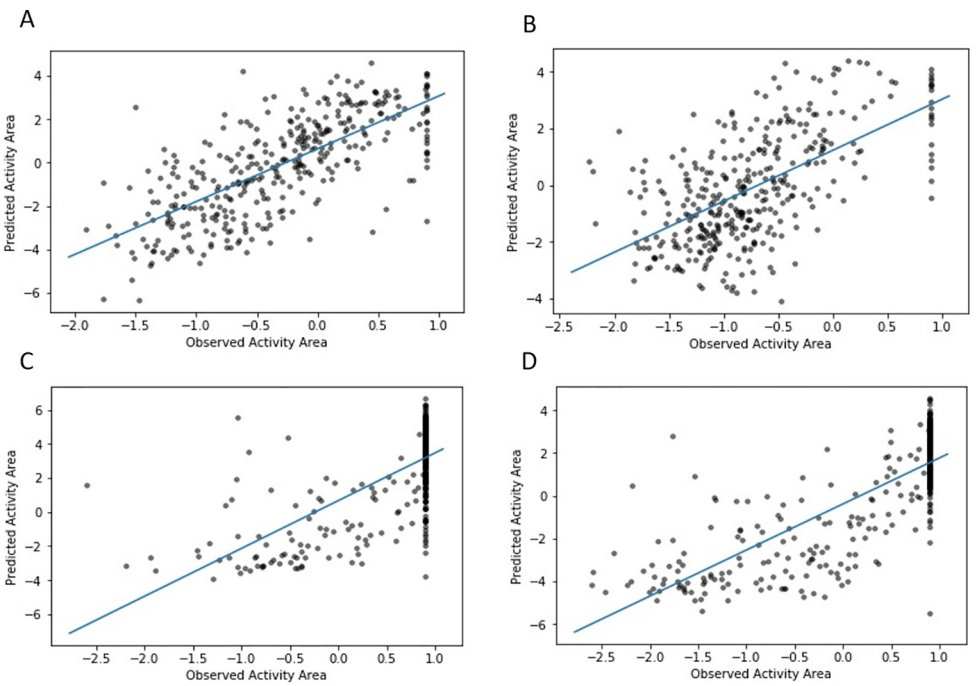


FIGURE 4 | Correlations between observed and predicted activity areas using DSPLMF method for CCLE cell lines across four drugs. **(A)** shows the scatter plot of observed and predicted drug responses for Topotecan with 0.71 as Pearson Correlation. **(B)** shows the scatter plot of observed and predicted drug responses for 17-AAG with 0.60 as Pearson Correlation. **(C)** shows the scatter plot of observed and predicted drug responses for AZD6244 with 0.68 as Pearson Correlation. **(D)** shows the scatter plot of observed and predicted drug responses for PD-032590 with 0.79 as Pearson Correlation.

TABLE 6 | Learned hyperparameters of DSPLMF method based on Genomics of Drug Sensitivity in Cancer (GDSC) dataset.

Hyperparameters	L	k	t	λ_c	λ_d	α	β	λ	γ	ϕ	ψ	*Threshold*
value	95	20	20	0.6	0.6	0.5	0.1	1	1	1	3	0.6

corresponding to cell line of the same subtype in this cancer justifies the efficiency of obtained latent vectors.

Investigation Drug-Pathway Association

For inferring drug-pathway associations, the heatmap of Pearson correlation between predicted drug responses and pathway activity scores similar to Suphavilai et al. (2018) is used. We considered 50 Biocarta pathway gene sets from MSigDB Liberzon et al. (2011), and pathway activity scores for CCLE cell lines were calculated as follows:

Let PW is a pathway and G(PW) = $\{g_1, g_2, \ldots, g_r\}$ is the set of genes corresponding to pathway PW. Let fold-change value of g_i in cell line c_j is x_{ij}, which is obtained by:

$$x_{ij} = Log_2(expression\ \ intensity\ \ of\ \ g_i\ \ in\ \ cell\ \ line\ \ c_j) - median(Log_2(expression\ \ intensity\ \ of\ \ g_i\ \ in\ \ all\ \ cell\ \ lines)) \quad (21)$$

Pathway activity score of pathway PW for cell line c_j, $PAS_j(PW)$ was calculated by formula 22.

$$PAS_j(PW) = \sum_{i=1}^{r} x_{ij} \quad (22)$$

Pathway activity score of PW for all cell line, $PAS(PW)$, are considered as the vector $PAS(PW) = [PAS_1(PW), PAS_2(PW), \ldots, PAS_n(PW)]$, where n is the numbers of cell lines. Also, the predicted drug responses by DSPLMF for each drug were considered as the vector $IC50_{predicted} = [IC_1, IC_2, \ldots, IC_n]$.

Then, the association between drug d_j and pathway PW is computed by the Pearson correlation between $IC50_{predicted}$ for drug d_j and $PAS(PW)$. A positive correlation indicates that a pathway plays a role in drug resistance and negative correlation demonstrated that a pathway is important in drug sensitivity. The result of the Pearson correlation of 30 pathway gene sets and 24 drugs of CCLE dataset is shown in **Figure 6** and the result of 20 other pathways is represented in the **Supplementary File 1 (Data Sheet 1)**. In this figure, the blue is represented the assistance and the red is represented the resistance case. Below, we investigated several instances that indicates consistency between the result of calculated Pearson correlation and previous studies and researches.

- The activation score of the HDAC (Histone deacetylases) pathway is negatively correlated (assistant association) with predicted IC50 value of some drugs such as Panobinostat. These observations were consistent with two studies, showing that the Panobinostat can inactive HDAC pathway De Marinis et al. (2013); Yee and Raje (2018).
- We observed the RELA (Acetylation and Deacetylation of RelA in The Nucleus) pathway had an assistant association with the 17-AAG (HSP90 inhibitor) drug. The RELA gene is one member of the NF-kB family and two important roles of the RELA are the transcriptional regulation and NF-kB signed transduction. Since the 17-AAG drug affects the NF-kB activity, it also affects the RELA gene and RELA pathway Thangjam et al. (2014).
- The activation score of the *EGFR – SMRTE* pathway was negatively correlated with predicted IC50 value of four EGFR

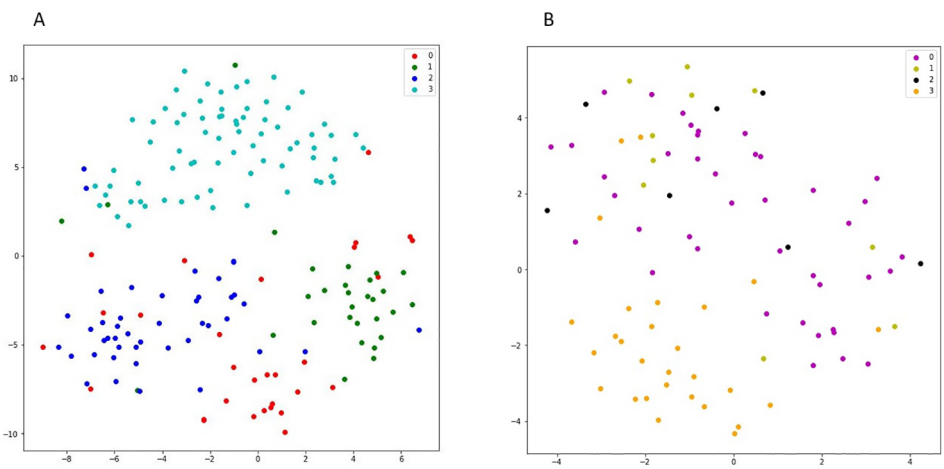


FIGURE 5 | (A) shows the t-SNE plot of latent space corresponding to four different cancer subtypes in GDSC dataset. In this figure, red points show the latent vectors of breast cancer and green, dark blue, and light blue points show the latent vectors of COREAD, central nervous system, and haematopoietic and lymphiod tissue, respectively. **(B)** shows the t-SNE plot of latent space corresponding to different lung cancer subtypes in Genomics of Drug Sensitivity in Cancer (GDSC) dataset. In this figure, purple points show the latent vectors of adenocarcinoma and light green, black and orange points show the latent vectors of large cell, squamous cell, and small cell carcinoma, respectively.

inhibitors drugs, namely, Lapatinib, Erlotinib, Vandetanib, and AZD0530. These observations matched the previous study that denoted the amplification of the EGFR gene is correlated with a high response to EGFR inhibitors Normanno et al. (2006). Moreover, the predicted IC50 values of the Crizotinib (ALK-inhibitor) were positively correlated with the activity score of this pathway and this issue was confirmed in the previous studies Sasaki et al. (2011).

- The MTA3 (Downregulated of MTA-3 in ER-negative Breast Tumors) pathway was associated (positively correlated) with two predicted IC50 vectors belong to L-685458(gamma-secretase) and PD-0332991(CDK4/6) drugs. Therefore, the cell lines with inactivated MTA3 pathway tend to sensitive to these two drugs Suphavilai et al. (2018).
- The VEGF-Hypoxia-Angiogenesis (VEGF) pathway was assistance associated with two RAF inhibitors drugs, namely, PLX4720 and RAF265 drugs that were verified in the previous researches. One of these studies considered inducing the VEGF expression by Raf promotes angiogenesis and blocking *RAF/MEK/ERK* pathway by RAF inhibitors McCubrey et al. (2007). Moreover, the activity

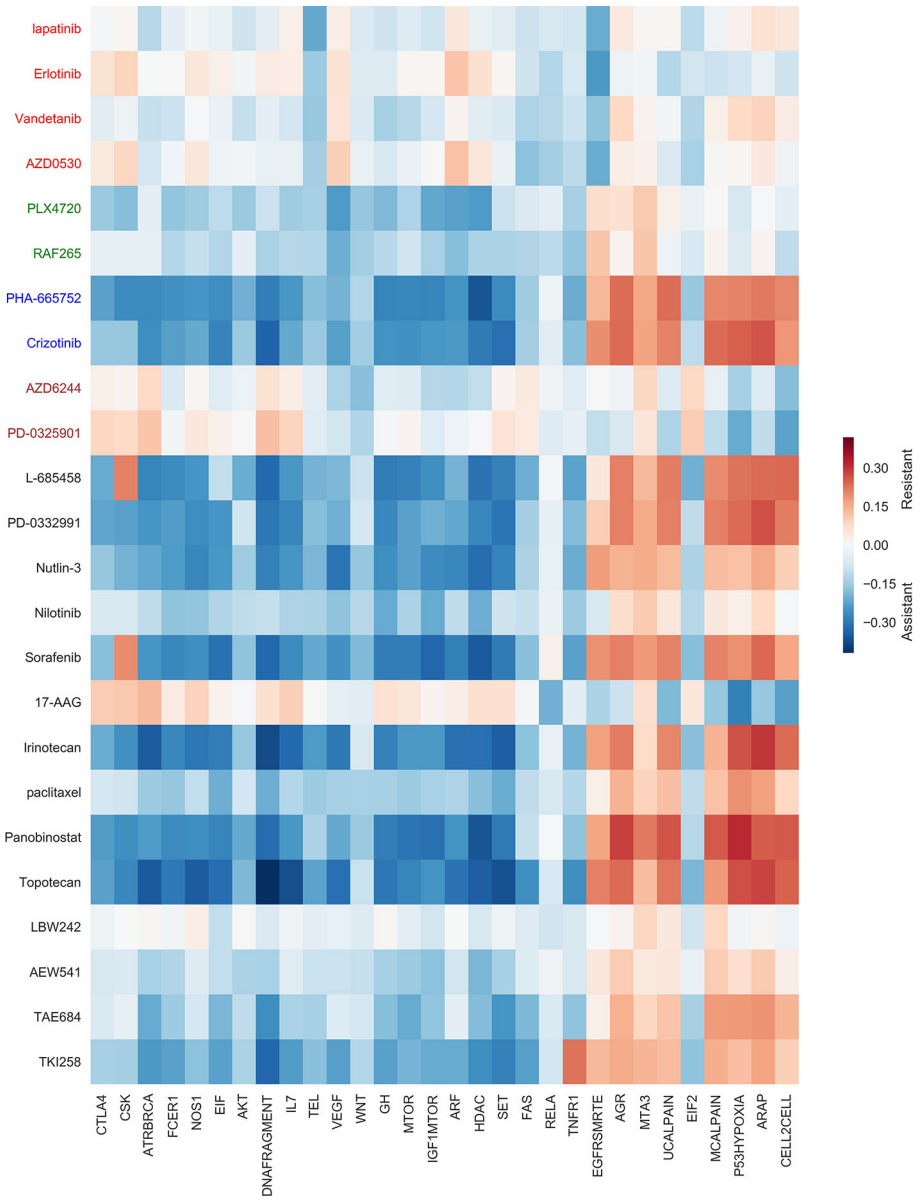


FIGURE 6 | Drug-pathway association based on Cancer Cell Line Encyclopedia (CCLE) dataset. For visualization, 30 Biocarta pathways across 24 drugs were selected. Negative and positive correlations between pathway activity and drug sensitivity scores are denoted as being "assistant" and "resistant" associations, respectively. The blue color is represented the assistance and the red color is represented the resistance.

score of the VEGF pathway was negatively correlated with Sorafenib drug Liu et al. (2006).
- The activity score of the mTOR Signaling Pathway that is a central regulator of metabolism and physiology was negatively correlated with predicted Ic50 vector of some drugs such as Panobinostat. Various preclinical studies have been performed to combine panobinostat with several drugs as mTOR inhibitor Singh et al. (2016).
- It has been shown that c-met inhibitor drugs such as PHA-665752 and Crizotinib can inhibit WNT pathway activity in tumour cells. We observed the activity score of this pathway was negatively correlated with predicted IC50 vectors of these drugs Tuynman et al. (2008); Zhang et al. (2018).
- The assistant association was observed between $L-685458$ drug and IGF-1 MTOR pathways. These observations were also reported by Shih et al Shih and Wang (2007).
- We observed that the MEK inhibitors such as *AZD*6244 and $PD-0325901$ were positively correlated with activity scores for the EIF2 pathway. Therefore, as mentioned in the previous researches, the cell lines with inactivated EIF2 p athway were sensitive to these drugs Quevedo et al. (2000); Liberzon et al. (2011).

Conclusion

In this work, we introduce a novel method for cancer drug sensitivity prediction based on a recommender system approach. A logistic matrix factorization is applied to predict the extent to which a cell line is sensitive to a drug. The advantage of this method is to obtain latent features of cell lines and drugs for better prediction performance. Since the similarity information of cell lines and drugs can improve higher predictive power, some information such as gene expression profile, copy number alteration and single-nucleotide mutation data for cell lines and Chemical structures of drugs are used.

To demonstrate the validity of DSPLMF method for identifying drug response 10-fold cross validation on CCLE and GDSC datasets are performed. The comparison of DSPLMF with six other the state-of-the-art prediction methods showed that DSPLMF outperformed other methods. The results indicated that the proposed method was able to uncover much more effective features than the other methods for drug response prediction.

DATA AVAILABILITY STATEMENT

The source code of proposed method and *Datasets* folder for GDSC and CCLE datasets as input data are available in https://github.com/emdadi/DSPLMF and **Supplementary File 4 (Data Sheet 4)**.

AUTHOR CONTRIBUTIONS

AE designed the algorithm, performed the experiments, and wrote the main manuscript text and the programming codes. CE conducted the experiments and analyzed the results. All authors reviewed the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00075/full#supplementary-material

SUPPLEMENTARY FILE 1 (DATA SHEET 1) | Results of drug pathway association on CCLE dataset.

SUPPLEMENTARY FILE 2 (FIGURES S1–S4) | The scatter plots of all 24 drugs in the CCLE dataset.

SUPPLEMENTARY FILE 3 (DATA SHEET 3) | AdaGrad Algorithm.

SUPPLEMENTARY FILE 4 (DATA SHEET 4) | Implementation Codes.

REFERENCES

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603. doi: 10.1038/nature11003

Costello, J. C., Heiser, L. M., Georgii, E., Gönen, M., Menden, M. P., Wang, N. J., et al. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* 32, 1202. doi: 10.1038/nbt.2877

De Marinis, F., Atmaca, A., Tiseo, M., Giuffreda, L., Rossi, A., Gebbia, V., et al. (2013). A phase ii study of the histone deacetylase inhibitor panobinostat (lbh589) in pretreated patients with small-cell lung cancer. *J. Thoracic. Oncol.* 8, 1091–1094. doi: 10.1097/JTO.0b013e318293d88c

Hand, D., Kok, J. N., and Berthold, M. R. (1999). *Advances in Intelligent Data Analysis: Third International Symposium, IDA-99 Amsterdam, The Netherlands, August 9-11, 1999 Proceedings* (Verlag Berlin Heidelberg: Springer Science & Business Media).

Kursa, M. B., and Rudnicki, W. R.. (2010). Feature selection with the boruta package. *J. Stat. Softw.* 36, 1–13. doi: 10.18637/jss.v036.i11

Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., et al. (2015). Development of a drug-response modeling framework to identify cell line derived translational biomarkers that can predict treatment outcome to erlotinib or sorafenib. *PloS One* 10, e0130700. doi: 10.1371/journal.pone.0130700

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Liu, H., Zhao, Y., Zhang, L., and Chen, X. (2018). Anti-cancer drug response prediction using neighbor-based collaborative filtering with global effect removal. *Mol. Therapy-Nucleic Acids* 13, 303–311. doi: 10.1016/j.omtn.2018.09.011

Liu, L., Cao, Y., Chen, C., Zhang, X., McNabola, A., Wilkie, D., et al. (2006). Sorafenib blocks the raf/mek/erk pathway, inhibits tumor angiogenesis, and induces tumor cell apoptosis in hepatocellular carcinoma model plc/prf/5. *Cancer Res.* 66, 11851–11858. doi: 10.1158/0008-5472.CAN-06-1377

Lu, X., Gu, H., Wang, Y., Wang, J., and Qin, P. (2019). Autoencoder based feature selection method for classification of anticancer drug response. *Front. In Genet.* 10, 233. doi: 10.3389/fgene.2019.00233

McCubrey, J. A., Steelman, L. S., Chappell, W. H., Abrams, S. L., Wong, E. W., Chang, F., et al. (2007). Roles of the raf/mek/erk pathway in cell growth,

malignant transformation and drug resistance. *Biochim. Biophys. Acta (BBA)-Mol. Cell Res.* 1773, 1263–1284. doi: 10.1016/j.bbamcr.2006.10.001

Normanno, N., De Luca, A., Bianco, C., Strizzi, L., Mancino, M., Maiello, M. R., et al. (2006). Epidermal growth factor receptor (egfr) signaling in cancer. *Gene* 366, 2–16. doi: 10.1016/j.gene.2005.10.018

Polat, K., and Güneş, S. (2007). Classification of epileptiform eeg using a hybrid system based on decision tree classifier and fast fourier transform. *Appl. Math. Comput.* 187, 1017–1026. doi: 10.1016/j.amc.2006.09.022

Quevedo, C., Alcázar, A., and Salinas, M. (2000). Two different signal transduction pathways are implicated in the regulation of initiation factor 2b activity in insulin-like growth factor-1-stimulated neuronal cells. *J. Biol. Chem.* 275, 19192–19197. doi: 10.1074/jbc.M000238200

Sasaki, T., Koivunen, J., Ogino, A., Yanagita, M., Nikiforow, S., Zheng, W., et al. (2011). A novel alk secondary mutation and egfr signaling cause resistance to alk kinase inhibitors. *Cancer Res.* 71, 6051–6060. doi: 10.1158/0008-5472.CAN-11-1340

Shih, I.-M., and Wang, T.-L. (2007). Notch signaling, γ-secretase inhibitors, and cancer therapy. *Cancer Res.* 67, 1879–1882. doi: 10.1158/0008-5472.CAN-06-3958

Singh, A., Patel, V. K., Jain, D. K., Patel, P., and Rajak, H. (2016). Panobinostat as pan-deacetylase inhibitor for the treatment of pancreatic cancer: recent progress and future prospects. *Oncol. Ther.* 4, 73–89. doi: 10.1007/s40487-016-0023-1

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., et al. (2015). Pharmacogx: an r package for analysis of large pharmacogenomic datasets. *Bioinformatics* 32, 1244–1246. doi: 10.1093/bioinformatics/btv723

Soufan, O., Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2015). Dwfs: a wrapper feature selection tool based on a parallel genetic algorithm. *PloS One* 10, e0117988. doi: 10.1371/journal.pone.0117988

Suphavilai, C., Bertrand, D., and Nagarajan, N. (2018). Predicting cancer drug response using a recommender system. *Bioinformatics* 34, 3907–3914. doi: 10.1093/bioinformatics/bty452

Thangjam, G. S., Dimitropoulou, C., Joshi, A. D., Barabutis, N., Shaw, M. C., Kovalenkov, Y., et al. (2014). Novel mechanism of attenuation of lps-induced nf-κb activation by the heat shock protein 90 inhibitor, 17-n-allylamino-17-demethoxygeldanamycin, in human lung microvascular endothelial cells. *Am. J. Respiratory Cell Mol. Biol.* 50, 942–952. doi: 10.1165/rcmb.2013-0214OC

Tuynman, J. B., Vermeulen, L., Boon, E. M., Kemper, K., Zwinderman, A. H., Peppelenbosch, M. P., et al. (2008). Cyclooxygenase-2 inhibition inhibits c-met kinase activity and wnt activity in colon cancer. *Cancer Res.* 68, 1213–1220. doi: 10.1158/0008-5472.CAN-07-5172

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2012). Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

Yee, A. J., and Raje, N. S. (2018). Panobinostat and multiple myeloma in 2018. *Oncol.* 23, 516–517. doi: 10.1634/theoncologist.2017-0644

Zhang, Y., Xia, M., Jin, K., Wang, S., Wei, H., Fan, C., et al. (2018). Function of the c-met receptor tyrosine kinase in carcinogenesis and associated therapeutic opportunities. *Mol. Cancer* 17, 45. doi: 10.1186/s12943-018-0796-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 03 March 2020
doi: 10.3389/fgene.2020.00155

Genetic Variants Detection Based on Weighted Sparse Group Lasso

Kai Che[1], *Xi Chen*[1], *Maozu Guo*[1,2,3*], *Chunyu Wang*[1] *and Xiaoyan Liu*[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [3] Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China

Identification of genetic variants associated with complex traits is a critical step for improving plant resistance and breeding. Although the majority of existing methods for variants detection have good predictive performance in the average case, they can not precisely identify the variants present in a small number of target genes. In this paper, we propose a weighted sparse group lasso (WSGL) method to select both common and low-frequency variants in groups. Under the biologically realistic assumption that complex traits are influenced by a few single loci in a small number of genes, our method involves a sparse group lasso approach to simultaneously select associated groups along with the loci within each group. To increase the probability of selecting out low-frequency variants, biological prior information is introduced in the model by re-weighting lasso regularization based on weights calculated from input data. Experimental results from both simulation and real data of single nucleotide polymorphisms (SNPs) associated with *Arabidopsis* flowering traits demonstrate the superiority of WSGL over other competitive approaches for genetic variants detection.

Keywords: genome-wide association studies, genetic variants, single nucleotide polymorphisms, minimum allele frequency, sparse group lasso

OPEN ACCESS

Edited by:
Fa Zhang,
Institute of Computing Technology (CAS), China

Reviewed by:
Qi Ren,
Tianjin University, China
Hui Ding,
University of Electronic Science and Technology of China, China
Qin Ma,
The Ohio State University, United States

****Correspondence:***
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *19 November 2019*
Accepted: *10 February 2020*
Published: *03 March 2020*

Citation:
Che K, Chen X, Guo M, Wang C and Liu X (2020) Genetic Variants Detection Based on Weighted Sparse Group Lasso. Front. Genet. 11:155. doi: 10.3389/fgene.2020.00155

INTRODUCTION

Since completion of the sequencing-based structural genome project, the focus of life science research has gradually shifted from determining the composition of DNA sequences to elucidating the function of identified genes. However, the greatest challenge of functional genomics is to determine the risk genes associated with complex diseases or traits among the huge amount of DNA sequences. Approximately, 90% of all gene fragments in any two individuals of almost all organisms are identical; thus, the fragments affecting individual characteristics, diseases, or traits only appear in a small range of sequences (Tenaillon et al., 2001; Reich et al., 2002). Polygenic recombination or mutation can cause individual differences in genome sequences, resulting in genetic polymorphism. Single nucleotide polymorphisms (SNPs) are the most common form of such genetic variation. Therefore, identification and characterization of SNPs help to discover the underlying causes of various diseases or variable traits and to develop new therapeutic strategies and targets for drug development or crop improvement.

The goal of genome-wide association studies (GWAS) is to elucidate the relationship between millions of SNPs and complex traits (Klein et al., 2005). A single-locus association approach is

typically used in GWAS; however, the "polygenic theory" proposes that complex traits are controlled by the action of multiple SNPs together rather than by individual genes or variants (Dudbridge, 2016). Since the number of SNPs far exceeds the number of samples in a multi-loci association study, the "curse of dimensionality" becomes the main challenge of this type of analysis (Waddell et al., 2005). Many machine-learning algorithms have been widely used to overcome this limitation and facilitate investigating the association between traits with SNPs. Based on current approaches, association studies can be divided into two main categories: one based on feature selection (FS) and the other based on statistical machine learning with regularizing penalty.

FS is the process of selecting the most effective features among a set of original features so as to reduce the dimensionality of the dataset. There are two types of FS methods: the wrapper method as a dependent classifier (Hall and Smith, 1999), and the filter method as an independent classifier (Liu and Setiono, 1996). Typically, the wrapper and filter approaches are combined as the final selected method. When applying FS methods to GWAS, the SNPs are treated as the features, phenotypes are the labels, and the candidate SNPs are then selected according to their associations with phenotypes. Numerous FS methods have been applied in genetic association studies (Evans, 2010; Batnyam et al., 2013; Anekboon et al., 2014; Alzubi et al., 2017; An et al., 2017; Setiawan et al., 2018; Tsamardinos et al., 2019). For example, Evans (2010) combined two filter FS methods with classification methods in a machine-learning approach, and obtained strong association results. To further improve the accuracy of the selected SNPs, Batnyam et al. (2013) applied four popular FS approaches (Robnik-Šikonja and Kononenko, 2003; Liang et al., 2008; Seo and Oh, 2012; Lee et al., 2013) to select novel SNPs, which were then used to generate artificial features by applying a feature fusion method. Finally, the artificial features were classified by traditional classifiers. As an alternative combinational algorithm, Anekboon et al. (2014) proposed a correlation-based FS method as a filter to first select a portion of the SNPs, followed by a wrapper phase to sequentially feed each of these SNPs into k-nearest neighbor, artificial neural network, and Ridge regression classifiers. Alzubi et al. (2017) developed a hybrid FS method by combining conditional mutual information maximization and support vector machine-recursive feature elimination (SVM-RFE). An et al. (2017) used a hierarchical feature and sample selection framework to gradually select informative features and discard ambiguous samples in multiple steps to improve the classifier learning. Setiawan et al. (2018) firstly employed random forest algorithm to reduce the search space, then selected associated SNPs by sequential forward floating selection. Tsamardinos et al. (2019) applied p-values of conditional independence tests and meta-analysis techniques to select features, and made use of parallel techonology to increase the computing speed. Current methods based on FS have sufficient ability for selecting a candidate feature set. Nevertheless, it is important to use available biological information as prior knowledge in biocomputing. Since FS methods can only reflect the dataset itself, they are not suitable to screen features based on prior biological knowledge.

Regression models with penalty can also be used for GWAS. With this approach, the SNPs correspond to the independent variables, and phenotypes are mapped to dependent variables in the regression model. Since the number of SNPs typically far exceeds the number of samples, it is necessary to regularize the sparsity of coefficients in the regression model. As a representative example, the well-established lasso method proposed by Tibshirani (1996) can learn a sparse weight vector by penalizing the weight vector with a 1-norm loss while shrinking less important coefficients to zeros. Owing to this property, lasso and its extensions have been widely applied in the detection of genetic variants (Cao et al., 2014; Arbet et al., 2017; Tamba et al., 2017; Cherlin et al., 2018; Wang et al., 2019). For example, Cao et al. (2014) incorporated prior information in lasso to further increase the selection accuracy. Arbet et al. (2017) imposed a permutation method on lasso to improve the performance of the algorithm. Tamba et al. (2017) first reduced the number of SNPs to a moderate size, then used expectation maximization Bayesian lasso to detect the quantitative trait nucleotide (QTN). Cherlin et al. (2018) used lasso to explore the association between phenotype and SNP data and achieved good prediction. Wang et al. (2019) promoted a precision lasso that utilized regularization governed by the covariance and inverse covariance matrices of explanatory variables to increase sparse variable selection. However, SNPs (features) are generally found in groups, whereas lasso does not encourage sparsity between groups. Yuan and Lin, (2006) proposed the group lasso (GL) method, which sets a regularization of the sum of the ℓ_2 norm onto groups that encourages only a few groups to be selected. The GL approach has also been successfully applied in GWAS (Li et al., 2015; Lim and Hastie, 2015; Gossmann et al., 2017; Du et al., 2018). Gossmann et al. (2017) extended sorted L1 penalized estimation (SLOPE) in the spirit of Group LASSO to handle group structures between the predictor variables. Du et al. (2018) proposed the SCCA with truncated L1 penalized and GL to improve the performance and effectiveness of discovering SNPs or QTs in imaging genetics. However, once a group is chosen, all of its comprising features are also selected, which is not compliant with the actual biological situation in which SNPs are distributed sparsely across the genome in only a few groups. Simon et al. (2013) developed sparse GL (SGL) that uses the ℓ_2 penalty to select only a subset of the groups and the ℓ_1 penalty to select only a subset of the variables within the group. Indeed, SGL has been widely applied in detecting genetic variants (Rao et al., 2015; Li et al., 2017; Samal et al., 2017; Guo et al., 2019). Samal et al. (2017) proposed a method based on SGL to identify phenotype associated extreme currents decomposed from metabolic networks data. Combined SGL with group-level graph structure, which takes advantages of gene-level priors to penalty the nucleotide-level sparsity to identify the risk SNPs. Guo et al. (2019) proposed a method that combined SGL and linear mixed model (LMM) for multivariate associations of quantitative traits, and it obtained a good power. Despite this

improvement, the limitation of this method is that SGL selects sparse features within a group, but gives the same penalty for all features within the group. Consequently, this approach can easily result in swiping out low-frequency features that may play an important role in influencing phenotypes. To overcome this obstacle, it is important to assign different penalties to different features. Ideally, candidate SNPs should have a smaller penalty weight while others would have a larger penalty weight. In this way, candidate SNPs will stand out among the data more readily. To achieve this goal, we here propose a novel approach termed weighted SGL (WSGL) by introducing biological prior information for more accurate genetic variants detection. Specifically, we compute the minimum allele frequency (MAF) among a dataset of SNPs and use those values to reweight as the ℓ_1 penalty of each SNP site, which can increase the chance of retaining low-frequency variants without loss of information. To validate this approach, we compared the performance of our model with simulation and real data against the three mainstream models discussed above.

MATERIALS AND METHODS

Materials

Simulation Data

We used *Arabidopsis thaliana* data from Atwell et al. (2010), downloaded from https://github.com/Gregor-Mendel-Institute/atpolydb for the simulation. We used a quality control protocol on the original data. The SNPs are eliminated by the standard that Minor Allele Frequency (MAF) is < 0.01, the missing rate is > 0.05, or the allele frequencies are not in Hardy-Weinberg (P < 0.0001). After data preprocessing, we chose 200 genes on chromosome 1 covering a total of 1,993 SNPs. Twenty of these SNPs were chosen as the associated variants.

Real Data

The genotype information was the same as that obtained from the simulation data. Ten phenotypes were selected among the 107 reported. First, from chromosome 1 to 5, we chose the first 1,000 genes, which were sorted according to sequence length, including 49,962 SNPs. Second, we selected 19 genes containing 367 SNPs, which have been verified to be associated with flowering time in *Arabidopsis*. Thus, a total of 50,329 SNPs were analyzed in our experiments.

Statistical Model and Methods

We first give a problem statement, followed by a brief overview of lasso and its extension for application in a genetic association study. Finally, we describe our new WSGL method.

Let $X = (x_1, x_2,\ldots, x_n)^T$ denote the $n \times p$ genotype matrix, where n is the number of samples and p is the number of genotypes. Let $Y = (y_1, y_2,\ldots, y_n)^T$ represent the $n \times 1$ phenotype vector, containing the phenotype values of the n samples. We then establish a linear model between X and y:

$$Y = X\beta + \epsilon \tag{1}$$

where $\beta = (\beta_1, \beta_2,\ldots, \beta_n)^T$ is a $p \times 1$ regression coefficients vector, and $\epsilon \sim N(0, 1)$.

Lasso and Its Extension for Association Mapping

Tibshirani (1996) proposed the popular lasso estimator,

$$\min_{\beta} \frac{1}{2} \| y - X\beta \|_2^2 + \lambda \| \beta \|_1 \tag{2}$$

where β is the regression coefficients vector, and x, corresponding to the nonzero estimated coefficients in β, represents the candidate SNPs. $||\beta||_1$ is the ℓ_1 penalty item. λ is a regularization parameter, and its size determines the sparsity. When $\lambda = 0$, the lasso estimator is equivalent to ordinary least-squares regression.

However, the lasso applies to the situation in which the variables are independent of each other. For the situation in which the variables can be divided into m groups, Yuan and Lin (2006) proposed the GL estimator,

$$\min_{\beta} \frac{1}{2} \| y - \sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + \lambda \sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 \tag{3}$$

where m is the group of variables, the first part is OLS, the second part is the sum of the ℓ_2 penalty of the coefficients of each group, and λ is the regularization parameter. If the size of the group is 1, it will degenerate to the standard lasso.

The GL can generate a sparse in groups; however, the variables in a group are not sparse. To solve this problem, Simon et al. (2013) proposed the SGL,

$$\min_{\beta} \frac{1}{2n} \| y - \sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 + \alpha\lambda \| \beta \|_1 \tag{4}$$

where λ still controls the overall penalty and α determines the ratio between ℓ_1 and ℓ_2. When $\alpha = 1$, it will be transformed into lasso, whereas when $\alpha = 0$, it will be GL. SGL can either select the variables in a group-by-group manner, or screen the individual variables in the remaining groups.

Our Method

With respect to the genetic association problem, the variables in a group have different effects on the independent variable. However, the SGL uses the same penalty coefficients for all variables, regardless of the relative importance among SNPs in the screened groups.

To tackle this problem, we introduce the prior information ω in the model to improve the statistical power, and propose the WSGL,

$$\min_{\beta} \frac{1}{2n} \| y - \sum_{l=1}^{m} X^{(l)} \beta^{(l)} \|_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l} \| \beta^{(l)} \|_2 + \alpha\lambda \| \omega\beta \|_1 \tag{5}$$

The objective function in (5) is clearly convex; therefore, the optimal solution can be achieved by subgradient equations. Let $\hat{\beta}$ be the optimal solution of WSGL. For group $k = (1, 2,\ldots, m)$, the solution $\hat{\beta}^{(k)}$ satisfies

$$\frac{1}{n} X^{(k)T} \left(y - \sum_{l=1}^{m} X^{(l)} \hat{\beta}^{(l)} \right) = \sqrt{p_k}(1-\alpha)\lambda\mu^{(k)} + \alpha\lambda\omega^{(k)} v^{(k)} \tag{6}$$

where $\mu^{(k)}$ and $v^{(k)}$ are subgradients of $\| \hat{\beta}^{(k)} \|_2$ and $\| \hat{\beta}^{(k)} \|_1$, respectively. According to Simon et al. (2013), $\mu^{(k)} = \hat{\beta}^{(k)} / \| \hat{\beta}^{(k)} \|_2$ if $\beta^{(k)} \neq 0$; otherwise, $\| \mu^{(k)} \|_2 \leq 1$. $v_j^{(k)} = sign(\hat{\beta}_j^{(k)})$ when $\hat{\beta}_j^{(k)} \neq 0$; otherwise, $\| v_j^{(k)} \|_2 \leq 0$.

Following the analysis in Simon et al. (2013), the condition for $\hat{\beta}^{(k)} = 0$ is

$$\| S\left(X^{(k)T}\gamma_{(-k)}/n, \alpha\lambda\omega^{(k)}\right) \|_2 \leq \sqrt{p_k}(1-\alpha)\lambda \tag{7}$$

where $\gamma_{(-k)} = y - \sum_{l \neq k} X^l \hat{\beta}^{(l)}$ is the partial residual of y, and S is defined as $(S(a,b))_j = sign(a_j)(|a_j| - b_j)_+$.

If $\hat{\beta}^{(k)} \neq 0$, the subgradient condition for $\beta_i^{(k)}$ becomes

$$\frac{1}{n} X_i^{(k)T}\left(y - \sum_{l=1}^{m} X^{(l)}\hat{\beta}^{(l)}\right) = \sqrt{p_k}(1-\alpha)\lambda \frac{\hat{\beta}_i^{(k)}}{\| \hat{\beta}^{(k)} \|_2} + \alpha\lambda\omega_i^{(k)} v_i^{(k)} \tag{8}$$

This is satisfied for $\hat{\beta}^{(k)} = 0$, if $|X_i^{(k)T}\gamma_{(-k,i)}| \leq n\alpha\lambda\omega$, where $\gamma_{(-k,i)} = \gamma_{(-k)} - \sum_{j \neq i} X_j^{(k)} \beta^{(k)}$ is the partial residual of y.

When $\beta_i^{(k)} \neq 0$, we can get

$$\hat{\beta}_i^{(k)} = \frac{S\left(X_i^{(k)T}\gamma_{(-k,i)}/n, \alpha\lambda\omega\right)}{X_i^{(k)T}X_i^{(k)}/n + (1-\alpha)\lambda / \| \hat{\beta}^{(k)} \|_2} \tag{9}$$

For each locus, MAF indicates, to some degree, its rareness. The MAF of low-frequency variants is usually small, so the associated low-frequency variants are more susceptible to sparsity regularization than other common variants. With normal sparse group lasso, the pressure of being zeroed out on each locus within the same group is equally high. In this case, those low-frequency variants are more likely to be excluded during the process. So selection of an appropriate weight can help to filter out more accurate candidate low-frequency variants.

There are several approaches for deciding the weights. For example, a small penalty can be assigned to the loci in known susceptibility genes to ensure including them into the model. Alternatively, the weights can be dependent on the MAF. For a dataset including both low-frequency and common variants, low-frequency markers are assigned smaller weights to compensate for their low frequencies. Here, we assign each locus a weight as follows: $weight = 2\sqrt{MAF(1-MAF)}$. Each weight ω_i is calculated in advance, which contains genotypes and biological explanations. The importance of the ith variable can be adjusted by the weight ω_i. Thus, to choose a locus, we can give it a relatively small penalty weight. Conversely, a larger weight can be assigned to exclude a locus. If $\omega_i = 1$, our model will be transformed to the SGL. Moreover, it is important to select an optimal regularization parameter λ, as a larger λ will generate a sparser result. For the present model, we chose cross-validation to select the optimal λ.

A brief algorithmic description of our method is shown in **Algorithm 1**. Let n represent the number of samples and p be the number of genotypes. The time complexity of subgradient step in each iteration is $O(np)$. In real data, p is usually supposed to be large, resulting in comparatively high time complexity. Therefore, in genome-wide association analysis, we suggest to analyze chromosomes individually for huge genome.

ALGORITHM 1 | Parameter estimation for weighted sparse group lasso.

Input: Genotype X, phenotype y ratio α, regularization hyperparameter λ

Output: Estimated $\hat{\beta}$

```
calculate ω = 2√(MAF(1 − MAF));
while not converge do
    for k from 1 to number_of_groups do
        for i from 1 to length_of_groups (k) do
            update β̂_i^(k) using equation (9);
return β̂;
```

Performance Measurements

For performance evaluation of the new model, we treat the loci detection as a binary classification under class imbalance, in which associated loci are assigned the label 1, and all others are assigned the label 0. The testing frequency of each locus is then regarded as the predicted probability for label 1. The receiver operating characteristic (ROC) curve and the area under the precision-recall curve (AUPR) are typically used for performance assessments. The ROC curve is plotted based on the sensitivity and specificity, whereas AUPR is generated based on the precision and recall. In our problem, the number of variants is significantly lower than the number of all loci, resulting in an imbalanced dataset. In the ROC curve, the false positive rate cannot descend greatly when the true negative is huge. However, the AUPR is sensitive to false positive. Considering these factors, we chose the AUPR as the performance metric for this purpose.

RESULTS AND DISCUSSION

Experiments on Simulation Data

For assessing the performance of WSGL in selecting candidate SNPs associated with a trait of interest, its performance was compared with lasso, GL, and SGL. Two parameters needed to be controlled in this experiment: α, which is the proportion of ℓ_1 and ℓ_2 loss in SGL, and λ, which is the coefficient of the entire regularization term and influences the sparsity. We set α to 0.95. Based on the results of cross-validation, λ was set to 0.09.

Figure 1 shows the results of the four methods with the simulation data, which clearly exhibits the superior performance of WSGL. The AUPR of WSGL is 0.652, which outperformed lasso by 23.6%, GL by 50.8%, and SGL by 24.4%. Lasso uses ℓ_1 to guarantee the sparsity of selected SNPs, but does not consider the group information; therefore, the candidate SNPs may be selected from all groups equally. Although GL imposes group information on the model, it still lacks sparsity constraint within the group, which does not correspond with the biological assumption that only a small number of candidate SNPs are contained in a small number of groups. SGL considers the sparsity between and within groups, but can still easily exclude

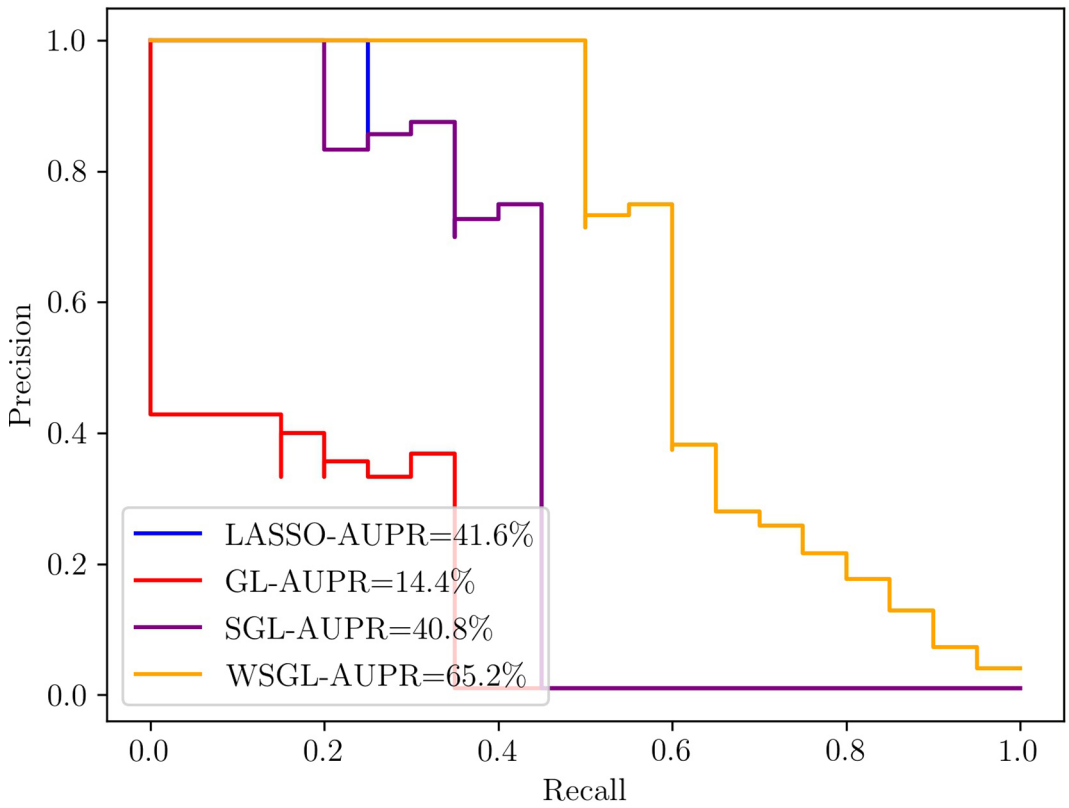


FIGURE 1 | Precision-recall (PR) curves of WSGL and the other methods.

important SNPs with a lower MAF. By introducing biological information to adjust the penalty of SNPs in the selected groups, WGSL places less weight on the low-frequency variants and thus increases their chance of being kept out. Despite its simplicity, the simulation results demonstrated the effectiveness of this approach for screening out important SNPs.

To further compare the performance of the four algorithms, we computed their AUPR values by fixing α at 0.95 and varying λ from 0.01 to 0.1 by steps of 0.01. As shown in **Figure 2**, with smaller λ, the model shows lower sparsity. When λ is 0.01 or 0.02, the model will include more SNPs, which may include more non-candidate SNPs that would cause a high false positive rate. Conversely, as λ increases, the number of selected SNPs decreases, which might result in the loss of some candidate SNPs, leading to a low TP rate. However, WSGL will include more candidate low-frequency loci by introducing prior knowledge to adjust the weight. Accordingly, WSGL keeps the highest position starting from $\lambda = 0.03$. When λ increases from 0.02 to 0.05, the AUPR of WSGL increases significantly from 58% to 64.5%, whereas the AUPR of lasso decreases from 59.2% to 53.2%, and that of SGL decreases from 59.9% to 56.1%. Surprisingly, the AUPR of GL decreases even more sharply from 51.1% to 34.1%. When λ reaches 0.05, the AUPR of WSGL tends to be stable, and the peak of 65.2% occurs at $\lambda = 0.09$. The AUPR of both lasso and SGL gradually decreases, and finally drops to around 40%. When λ is 0.1, the AUPR of GL drops to 13.3%. These results were consistent with our expectation that the performance of WSGL would be the best, SGL would perform better than lasso, and GL would show the worst performance overall.

Experiments on Real Data

To verify the ability of WSGL to detect candidate SNPs, we compared the performance of the four models using *Arabidopsis* flowering time data with known genetic associations. The dataset included 10 different phenotypes, FT10, FT16, FT22, LD, LDV, SD, SDV, LN10, LN16, and LN22, and the descriptions of the 10 phenotypes are shown in **Table 1**. We analyzed the associated number of genes covered by 100 SNPs with top probabilities of being target loci.

As shown in **Table 2**, WSGL could link more candidate genes with phenotypes FT10, FT16, FT22, LD, SD, and SDV. In particular, WSGL demonstrated excellent performance for FT10, not only by selecting less groups but also by including less SNPs within each group, and the ratio of candidate genes was 23.08%. By contrast, the ratios of candidate genes were 4.65%, 9.09%, and 5.13% for lasso, GL, and SGL, respectively. For phenotypes FT16, FT22, LD, SD, and SDV, WSGL still achieved the best detection performance. However, unexpectedly, the GL model obtained better results for the first four phenotypes. We consider that this may be due to the specific distribution of loci in the dataset. In cases for which most or all of

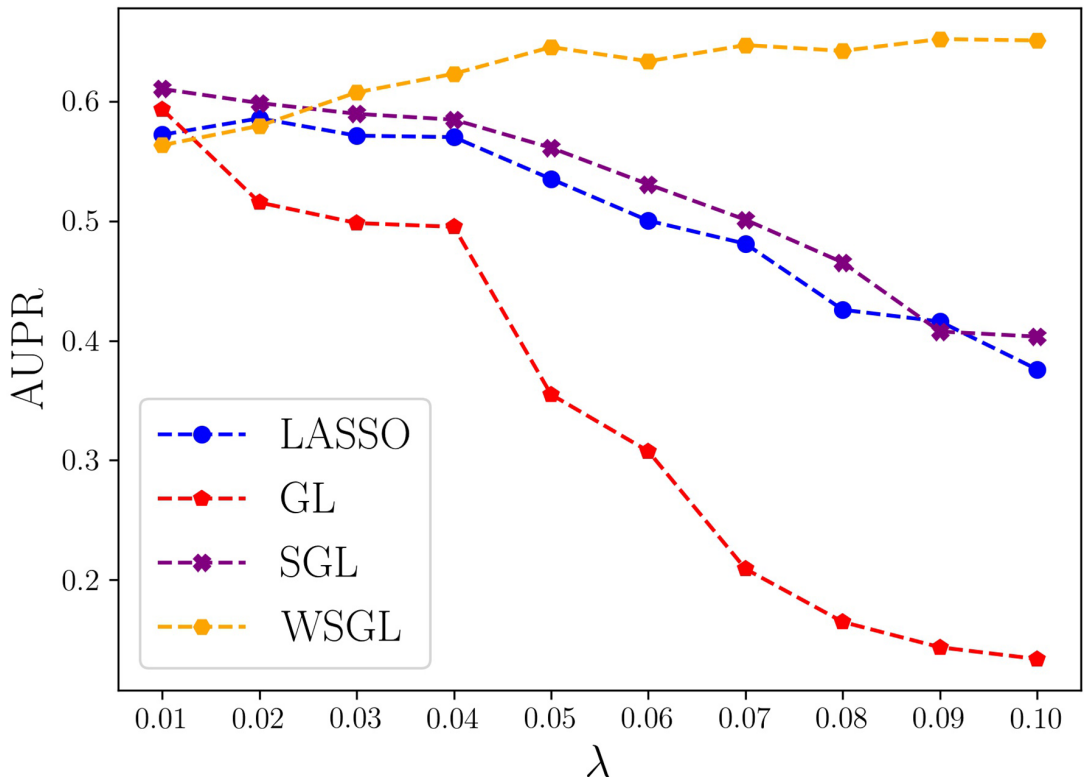


FIGURE 2 | Precision-recall (PR) curves of WSGL and the other methods for varying λ.

TABLE 1 | Description of the 10 flowering related phenotypes in *A.thaliana* in real data application.

Phenotype	Accessions	Phenotype description	Growths conditions	Phenotype scoring
LD	167	Days to flowering time (FT) under Long Day (LD) and Short Days (SD) +/– vernalization	18°C 16-h daylight	Number of days following stratification to opening of the first flower. The experiment was stopped at 200d, and accessions that had not flowered at the point were assigned a value of 200.
LDV	168		18°C 16-h daylight, vernalized (5wks 4)	
SD	162		18°C 16-h daylight	
SDV	159		18°C 16-h daylight, vernalized (5wks 4)	
FT10	194		10°C 16-h daylight	
FT16	193		16°C 17-h daylight	Plants were checked bi-weekly for presence of first buds, and the average flowering time and average leaf number of four plants of the same accession at each temperature were collected.
FT22	193	Flowering time (FT) and leaf number at flowering time (LN)	22 °C 18-h daylight	
LN10	177		10 °C 19-h daylight	
LN16	176		16°C 20-h daylight	
LN22	176		22°C 21-h daylight	

TABLE 2 | Summary of four methods associations found in real data.

Phenotype	Method	Number of genes covered by top 100 SNPs	Number of genes in the 19 genes	Ratio of candidate genes
FT10	Lasso	86	4	4.65%
	GL	66	6	9.09%
	SGL	78	4	5.13%
	WSGL	26	6	23.08%
FT16	Lasso	76	8	10.53%
	GL	62	8	12.9%
	SGL	64	7	10.94%
	WSGL	67	10	14.93%
FT22	Lasso	78	7	8.79%
	GL	72	7	9.72%
	SGL	77	6	7.79%
	WSGL	71	9	12.68%
LD	Lasso	81	9	11.11%
	GL	67	9	13.43%
	SGL	73	11	15.07%
	WSGL	74	12	16.22%
LDV	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
SD	Lasso	78	5	6.41%
	GL	70	5	7.14%
	SGL	79	6	7.59%
	WSGL	77	6	7.79%
SDV	Lasso	84	1	1.19%
	GL	66	1	1.52%
	SGL	78	2	2.56%
	WSGL	72	2	2.78%
LN10	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
LN16	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–
LN22	Lasso	6	6	–
	GL	6	6	–
	SGL	6	6	–
	WSGL	6	6	–

the candidate objects are located in only one group, GL will apparently show a good result. By contrast, all four methods could link all six genes with LDV, LN10, LN16, and LN22. This surprising result may reflect the strong association between the selected SNPs and these phenotypes, which is highly discriminable. Nevertheless, this assessment demonstrated that our new weighted method achieves the best performance overall, highlighting the importance of considering prior biological information for selection of candidate SNPs.

CONCLUSION

We proposed a method named weighted sparse group lasso (WSGL) to improve the detection of genetic variants. WSGL incorporates the ℓ_1 penalty, ℓ_2 penalty, and prior biological knowledge into a single linear regression model, and then uses SGL to either select or clear out all SNPs in a group potentially associated with a phenotype of interest. To screen candidate low-frequency variants, we introduced the MAF as the weight to re-scale each element for calculating ℓ_1 loss. In addition, WSGL can detect meaningful associations with more accuracy compared to available methods, which conforms with the general assumption that complex traits are affected by a few SNPs in a few genes. Experiments with both simulation and real data of SNPs related to the flowering time of *A. thaliana* demonstrated the effectiveness of our approach.

DATA AVAILABILITY STATEMENT

We used *Arabidopsis thaliana* data from Atwell et al. (2010), downloaded from https://github.com/Gregor-Mendel-Institute/atpolydb for the simulation.

AUTHOR CONTRIBUTIONS

Conceptualization: KC. Formal analysis: KC. Funding acquisition: MG, CW, and XL. Methodology: KC. Validation: KC and XC. Writing—original draft: KC and XC. Writing—review and editing, KC, MG, CW, and XL.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61571163, 61532014, 61671189, 61872114, and 61871020) and the National Key Research and Development Plan of China (Grant No. 2016YFC0901902).

REFERENCES

Alzubi, R., Ramzan, N., Alzoubi, H., and Amira, A. (2017). A hybrid feature selection method for complex diseases snps. *IEEE Access* 6, 1292–, 1301. doi: 10.1109/ACCESS.2017.2778268

An, L., Adeli, E., Liu, M., Zhang, J., Lee, S.-W., and Shen, D. (2017). A hierarchical feature and sample selection framework and its application for alzheimer's disease diagnosis. *Sci. Rep.* 7, 45269. doi: 10.1038/srep45269

Anekboon, K., Lursinsap, C., Phimoltares, S., Fucharoen, S., and Tongsima, S. (2014). Extracting predictive snps in crohn's disease using a vacillating genetic algorithm and a neural classifier in case–control association studies. *Comput. Biol. Med.* 44, 57–65. doi: 10.1016/j.compbiomed.2013.09.017

Arbet, J., McGue, M., Chatterjee, S., and Basu, S. (2017). Resampling-based tests for lasso in genome-wide association studies. *BMC Genet.* 18, 70. doi: 10.1186/s12863-017-0533-3

Atwell, S., Huang, Y. S., Vilhjálmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in arabidopsis thaliana inbred lines. *Nature* 465, 627. doi: 10.1038/nature08800

Batnyam, N., Gantulga, A., and Oh, S. (2013). "An efficient classification for single nucleotide polymorphism (snp) dataset," in *Computer and Information Science* (Berlin, Germany: Springer), 171–185. doi: 10.1007/978-3-319-00804-2_13

Cao, S., Qin, H., Deng, H.-W., and Wang, Y.-P. (2014). A unified sparse representation for sequence variant identification for complex traits. *Genet. Epidemiol.* 38, 671–679. doi: 10.1002/gepi.21849

Cherlin, S., Howey, R. A., and Cordell, H. J. (2018).Using penalized regression to predict phenotype from snp data, in: *BMC Proc. (BioMed Central)* 12 223–228. doi: 10.1186/s12919-018-0149-2

Du, L., Liu, K., Zhang, T., Yao, X., Yan, J., Risacher, S. L., et al. (2018). A novel scca approach via truncated l1-norm and truncated group lasso for brain imaging genetics. *Bioinformatics* 34, 278–285. doi: 10.1093/bioinformatics/btx594

Dudbridge, F. (2016). Polygenic epidemiology. *Genet. Epidemiol.* 40, 268–272. doi: 10.1002/gepi.21966

Evans, D. T. (2010). *A SNP microarray analysis pipeline using machine learning techniques. Ph.D. thesis* (Athens, OH, USA: Ohio University).

Gossmann, A., Cao, S., Brzyski, D., Zhao, L.-J., Deng, H.-W., and Wang, Y.-P. (2017). A sparse regression method for group-wise feature selection with false discovery rate control. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15, 1066–1078. doi: 10.1109/TCBB.2017.2780106

Guo, Y., Wu, C., Guo, M., Zou, Q., Liu, X., and Keinan, A. (2019). Combining sparse group lasso and linear mixed model improves power for finding genetic variants underlying quantitative traits. *Front. Genet.* 10, 271. doi: 10.3389/fgene.2019.00271

Hall, M. A., and Smith, L. A. (1999). Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, in: *FLAIRS conference.*, Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference; 1999 March 1-5. (Orlando, Florida, USA: DBLP) 1999 235–239.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J.-Y., Sackler, R. S., Haynes, C., et al. (2005). Complement factor h polymorphism in age-related macular degeneration. *Science* 308, 385–389. doi: 10.1126/science.1109557

Lee, J., Batnyam, N., and Oh, S. (2013). Rfs: Efficient feature selection method based on r-value. *Comput. Biol. Med.* 43, 91–99. doi: 10.1016/j.compbiomed.2012.11.010

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *Ann. Appl. Stat.* 9, 640. doi: 10.1214/15-AOAS808

Li, J., Dong, W., and Meng, D. (2017). Grouped gene selection of cancer via adaptive sparse group lasso based on conditional mutual information. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 15, 2028–2038. doi: 10.1109/TCBB.2017.2761871

Liang, J., Yang, S., and Winstanley, A. (2008). Invariant optimal feature selection: A distance discriminant and feature ranking based solution. *Pattern Recognition* 41, 1429–1439. doi: 10.1016/j.patcog.2007.10.018

Lim, M., and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graphical Stat* 24, 627–654. doi: 10.1080/10618600.2014.938812

Liu, H., and Setiono, R. (1996). "A probabilistic approach to feature selection-a filter solution," in *ICML (Citeseer)*, vol. 96. , 319–327.

Rao, N., Nowak, R., Cox, C., and Rogers, T. (2015). Classification with the sparse group lasso. *IEEE Trans. Signal Process.* 64, 448–463. doi: 10.1109/TSP.2015.2488586

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., et al. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* 32, 135. doi: 10.1038/ng947

Robnik-Šikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.* 53, 23–69. doi: 10.1023/A:1025667309714

Samal, S., Radulescu, O., Weber, A., and Fröhlich, H. (2017). Linking metabolic network features to phenotypes using sparse group lasso. *Bioinf. (Oxf. Engl.)* 33, 3445–3453. doi: 10.1093/bioinformatics/btx427

Seo, M., and Oh, S. (2012). Cbfs: High performance feature selection algorithm based on feature clearness. *PloS One* 7, e40419. doi: 10.1371/journal.pone.0040419

Setiawan, D., Kusuma, W. A., and Wigena, A. H. (2018). Snp selection using variable ranking and sequential forward floating selection with two optimality criteria. *J. Eng. Sci. Technol. Rev.* 11. doi: 10.25103/jestr.115.09

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graphical Stat* 22, 231–245. doi: 10.1080/10618600.2012.681250

Tamba, C. L., Ni, Y.-L., and Zhang, Y.-M. (2017). Iterative sure independence screening em-bayesian lasso algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13, e1005357. doi: 10.1371/journal.pcbi.1005357

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of dna sequence polymorphism along chromosome 1 of maize (zea mays ssp. mays l.). *Proc. Natl. Acad. Sci.* 98, 9161–9166. doi: 10.1073/pnas.151244298

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.: Ser. B. (Methodol.)* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tsamardinos, I., Borboudakis, G., Katsogridakis, P., Pratikakis, P., and Christophides, V. (2019). A greedy feature selection algorithm for big data of high dimensionality. *Mach. Learn.* 108, 149–202. doi: 10.1007/s10994-018-5748-7

Waddell, M., Page, D., and Shaughnessy, J.Jr. (2005). "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in *Proceedings of the 5th International Workshop on Bioinformatics* (ACM), 21–28.

Wang, H., Lengerich, B. J., Aragam, B., and Xing, E. P. (2019). Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* 35, 1181–1187. doi: 10.1093/bioinformatics/bty750

Yuan, M., and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. (Stat. Methodol.)* 68, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 10 March 2020
doi: 10.3389/fbioe.2020.00080

AgeGuess, a Methylomic Prediction Model for Human Ages

*Xiaoqian Gao[1], Shuai Liu[1], Haoqiu Song[1,2], Xin Feng[1], Meiyu Duan[1], Lan Huang[3] and Fengfeng Zhou[1]**

[1] *BioKnow Health Informatics Laboratory Key Laboratory of Symbolic Computation and Knowledge Engineering, College of Computer Science and Technology, Ministry of Education, Jilin University, Changchun, China,* [2] *College of Computer Science, Hubei University of Technology, Wuhan, China,* [3] *Key Laboratory of Symbolic Computation and Knowledge Engineering, College of Computer Science and Technology, Ministry of Education, Jilin University, Changchun, China*

Aging was a biological process under regulations from both inherited genetic factors and various molecular modifications within cells during the lifespan. Multiple studies demonstrated that the chronological age may be accurately predicted using the methylomic data. This study proposed a three-step feature selection algorithm AgeGuess for the age regression problem. AgeGuess selected 107 methylomic features as the gender-independent age biomarkers and the Support Vector Regressor (SVR) model using these biomarkers achieved 2.0267 in the mean absolute deviation (MAD) compared with the real chronological ages. Another regression algorithm Ridge achieved a slightly better MAD 1.9859 using the same biomarkers. The gender-independent age prediction models may be further improved by establishing two gender-specific models. And it's interesting to observe that there were only two methylation biomarkers shared by the two gender-specific biomarker sets and these two biomarkers were within the two known age-associated biomarker genes CALB1 and KLF14.

Keywords: age prediction, methylomic biomarker, regression, support vector regressor, ridge

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Zengyou He,
Dalian University of Technology, China
Man Kit Lei,
University of Georgia, United States
Chunhou Zheng,
Anhui University, China

****Correspondence:***
Fengfeng Zhou
FengfengZhou@gmail.com;
ffzhou@jlu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *09 November 2019*
Accepted: *29 January 2020*
Published: *10 March 2020*

Citation:
Gao X, Liu S, Song H, Feng X, Duan M, Huang L and Zhou F (2020) AgeGuess, a Methylomic Prediction Model for Human Ages. Front. Bioeng. Biotechnol. 8:80. doi: 10.3389/fbioe.2020.00080

INTRODUCTION

Aging is a ubiquitous phenomenon in almost all the multi-cellular organisms (Horn and Schweppe, 2015). It is also a challenging issue concerned by citizens in many countries (Baltes and Smith, 2003; Banister et al., 2012). Evidences were accumulating about that aging is a biological process strictly regulated by epigenetic modifications rather than random events (Fraga and Esteller, 2007; Martino et al., 2011; Schellenberg et al., 2011; Pal and Tyler, 2016). So it's technically reasonable to estimate an individual's biological age through the biomarkers like telomere length (Saeed et al., 2012; Barrett et al., 2013), age-dependent changes in T cell DNA (Zubakov et al., 2010; Ou et al., 2012), and RNA biomarkers (Alvarez and Ballantyne, 2006), etc. Recent studies also demonstrated that DNA methylation levels at certain CpG residues were linearly associated with the biological ages, and may serve well as age biomarkers (Zubakov et al., 2016).

DNA methylation has been implicated to be involved in various aging-associated biological processes (Jones et al., 2015; Field et al., 2018). DNA methylation is a biological process of selectively adding a methyl group to a cytosine to form 5-cytosine facilitated by a DNA methyltransferase (Moore et al., 2013). This epigenetic modification plays an essential role in transcriptional regulation and other biological processes (Vaillancourt et al., 2017; Suzuki et al., 2019). Quite a few age prediction models were proposed based on the methylation biomarkers. Besides clinical

application, these models can also be used in forensic investigation (Vidaki and Kayser, 2018; Alsaleh and Haddrill, 2019). Blood and other liquids are one of the most important biological evidences found in the crime scene, so it's necessary to use the whole blood to establish an accurate age prediction model.

The major challenge is finding a subset of methylation features with a good age prediction performance using the methylomic datasets. About half a million methylation features may be generated for one sample by the popular array-based methylome profiling technologies like Illumina HumanMethylation450 BeadChip (450K) (Fernandez-Jimenez et al., 2019). The feature number is much larger than the sample number, and a step of feature selection has to be conducted to avoid the model over-fitting (Feng et al., 2018).

The existing methylome-based age prediction studies explored different feature selection algorithms to find the best age-associated biomarkers. Horvath used the elastic net algorithm to select 353 methylomic features to predict the human ages and the mean absolute error of the predicted age was about 3.6 years (Horvath, 2013). Yi et al. detected three age-related gene fragments from the blood samples of 40 volunteers and used the CpG locus of these fragments to train the age-regression model with a prediction difference of 4 years compared with the real ages (Yi et al., 2015). Hong et al. proposed a linear regression-based age prediction model, which achieved 94.5% in correlation and 3.13 years in the mean absolute deviation (Small et al., 2011) from the chronological ages (Hong et al., 2017). Another study investigated this forensic problem by selecting 23 methylomic features and established a multi-variate regression model with an age prediction deviation of about 4.6 years (Vidaki et al., 2017).

Feature selection algorithm has been utilized in many biomedical research areas. Various biomedical high-throughput data producing technologies were rapidly invented and developed and may produce as many as millions of features per sample (Diao and Vidyashankar, 2013; Ye et al., 2017; Ceglia et al., 2018). But the number of samples collected in a study was usually limited by the difficulty of patient recruitment and the cost of generating the data. So a biomedical big data project usually had a much larger number of features than the number of samples. A feature selection algorithm may significantly reduce the model complexity and the possibility of over-fitting (Le et al., 2017; Ma and Fan, 2017). Feature selection was not only widely used in the bioinformatics problems of genes (Tian et al., 2019), proteins (Liu et al., 2019), and metabolism system (Grissa et al., 2016), but also played an important role in the analysis of biomedical images (Pan et al., 2019) and time series data (Li et al., 2017).

This study proposed a three-step feature selection algorithm, AgeGuess, to find the best age prediction biomarkers using the methylomic profiles. The metrics Maximal Information Coefficient (MIC) was a sensitive correlation measurement (Reshef et al., 2011) and was utilized to remove those methylomic features with small MIC association with ages. The remaining features were recursively eliminated based on the evaluation of a support vector regressor. The last step removed the features iteratively based on an exhaustive screening. Our experimental data demonstrated an improved prediction performance of chronological ages. Gender information was also evaluated in further optimizing the age prediction models.

MATERIALS AND METHODS

Dataset Summary

This study used the methylomic dataset GSE40279, which was publicly available from the database Gene Expression Omnibus (GEO) (Clough and Barrett, 2016). The dataset GSE40279 was profiled using the methylomic platform Illumina HumanMethylation450 BeadChip (accession GPL13534) (Alsaleh and Haddrill, 2019). There were 656 samples with chronological ages in this dataset, and each sample was profiled for 485,577 methylomic resides (Alsaleh and Haddrill, 2019). The methylome was generated using the human whole blood samples, obtained from 426 Caucasians and 230 Hispanics individuals with chronological ages 19–101. As similar to the existing study (Hannum et al., 2013), sex chromosomes were excluded from analysis in this study. So there were 473,034 CpG features left for further analysis.

Feature Selection Algorithm AgeGuess

Not all of these half-million methylomic features were associated with the aging process and all the existing studies selected a subset of features for building their age prediction models (Horvath, 2013; Yi et al., 2015; Hong et al., 2017; Vidaki et al., 2017). So this study proposed a feature selection algorithm AgeGuess to find a feature subset with the best age prediction performance.

Single-step feature selection algorithm may be roughly grouped as two major types, i.e., filters and wrappers (Suto et al., 2016). A filter evaluated each feature's association with the class labels with the assumption of inter-feature independence and can be easily scaled to a large number of features (Guyon and Elisseeff, 2003; Solorio-Fernández et al., 2016). A wrapper screened a feature subset by a heuristic rule for its classification performance of a user-defined classifier. A wrapper usually outperforms a filter in accuracy with the cost of a high computational complexity (Guyon and Elisseeff, 2003; Solorio-Fernández et al., 2016). In order to fully utilize the advantages of both filters and wrappers, a multi-step feature selection algorithm may significantly reduce the number of features in the first step. Then more sophisticated and slow algorithms may be utilized. The following algorithm AgeGuess was designed based on this rule for the chronological ages.

Firstly, AgeGuess selected 10,000 methylomic features that were highly correlated with the sample label, i.e., chronological age. There were 473,034 methylomic features for each sample in this dataset, and not all these features contributed to the age prediction. The metrics Maximum Information Coefficient (MIC) demonstrated a very sensitive power in detecting linear and non-linear correlations between two variables (Reshef et al., 2011). This study calculated the MIC correlation of each methylated features with the chronological ages, and kept the 10,000 features with the largest MIC values for further analysis.

Then the Recursive Feature Elimination (RFE) strategy was utilized to remove un-related features. The RFE strategy relied on the feature ranking and iteratively removed the k least-ranked

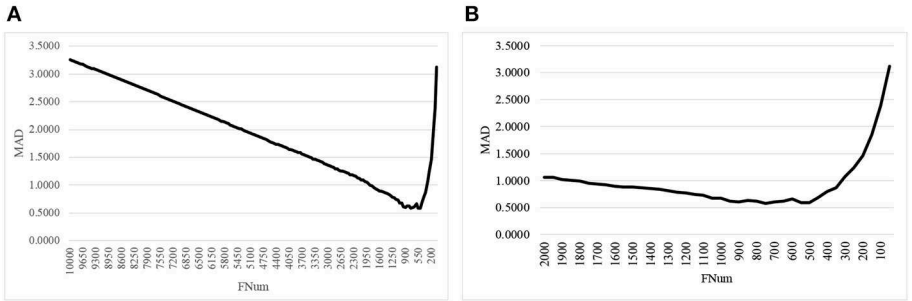


FIGURE 1 | The line plot of the regression metrics MAD of AgeGuess. **(A)** RFE strategy to removed 50 features in each iteration on [10000, 50] and **(B)** The scale was zoomed to [2000, 50]. The horizontal axis was the number of features remained for building the classification model.

features. The investigated problem in this study was a regression model, and the Support Vector Regressor (SVR) was used to calculate the metrics to rank the features. The trained SVR model produced a weight vector Feature Importance, and the features were sorted by the descendent order of the weights. This procedure was conducted iteratively until all the features were removed. The feature subset with the best regression performance was returned.

One more redundancy-removal step was conducted to further refine the feature subset obtained in the above step. The iterative exclusion of the feature with the least performance decrease was carried out, which was the same as the backFS strategy in the other studies (Feng et al., 2019; Zhang et al., 2019). The performance was calculated by the 10-fold cross validation strategy.

A good feature selection algorithm tended to select fewer features and to achieve a higher prediction performance. But these two performance metrics usually cannot achieved simultaneously. So this study defined the integrated evaluation index (EI) as the optimization goal. EI was defined as (MAD+FNum/100), where MAD was the mean absolute deviation and FNum was the number of features selected by the feature selection algorithm. This regression performance metrics suggested one more selected feature increased the overall performance by 0.01. And the metrics EI was used to optimize the above-mentioned backFS strategy.

Performance Evaluation Metrics

This study investigated the age prediction problem using the 656 samples from the platform GEO. Multiple regression performance metrics were used to evaluate how the generated regression model performed. The metrics Mean Absolute Deviation (Small et al., 2011) was the averaged absolute error value between the predicted age and the chronological age (Pan et al., 2019). The Mean Squared Error (MSE) and the squared root version of MSE (RMSE) were another two widely used regression performance metrics (Liu et al., 2019; Thompson et al., 2019). The metrics Goodness of Fit (R2) quantitatively evaluated how well the regression model fitted the data (Chong et al., 2017). These regression metrics were implemented in the package scikit-learn version 0.19.1 of Python version 3.6.4.

RESULTS

Optimizing the Proposed Algorithm AgeGuess

The proposed feature selection algorithm AgeGuess selected 10,000 out of the 473,034 methylomic features with the largest MIC coefficients (Reshef et al., 2011) with the chronological ages. AgeGuess hypothesized that the contributions of the excluded features may be neglected since their MIC coefficients with the chronological ages were small.

The second step of AgeGuess utilized the RFE framework to iteratively remove the features, as shown in **Figure 1**. Due to the number of remaining features was still very large, this study set $k = 50$, i.e., 50 features with the least Feature Importance weights calculated by the trained SVR model were removed in each iteration. **Figure 1A** illustrated that the majority of the 10,000 methylation features didn't contribute to the age prediction performance. And there was a "valley" smaller than 1,500 features in the line plot in **Figure 1A**. So **Figure 1B** zoomed in the line plot within the range [2000, 50]. The data showed that the small MAD value was achieved between 900 and 500. And the minimum value MAD = 0.5809 was achieved with 750 features.

The proposed algorithm AgeGuess further removed the redundancies in the methylated features by the function backFS (Feng et al., 2019; Zhang et al., 2019). The 750 methylation features chosen in the above step was iteratively evaluated and one feature was removed per iteration if its removal generated the least contribution to the age prediction performance metrics EI. **Figure 2A** illustrated that the valley was around 100 features in the horizontal axis. The plot was further zoomed-in for the number of features between 50 and 150, as shown in **Figure 2B**. The age regression metrics EI reached the minimum 3.0316 when 107 features were selected.

The SVR regression model was trained using the 107 methylation features, and was evaluated by the following regression performance metrics. **Figure 3** illustrated that the RealAge and the PredAge were very close to each other. The prediction performance was averaged over the 10-fold cross validations, and 10 random rusns were averaged to generate the final results. The Mean Absolute Deviation (Small et al., 2011) was 2.0267 years. AgeGuess's model achieved the other two

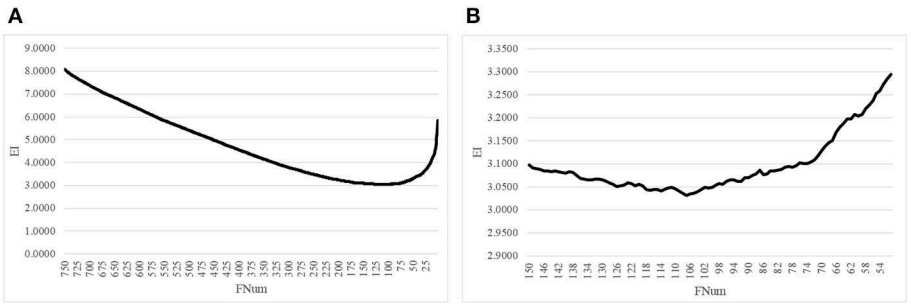


FIGURE 2 | Iterative elimination of redundant features by backFS. **(A)** The line plot for the numbers of features no more than 750 features. **(B)** The zoomed-in plot for the numbers of features between [50, 150]. The horizontal axis was the number of features. And the vertical axis was the regression performance metrics EI.

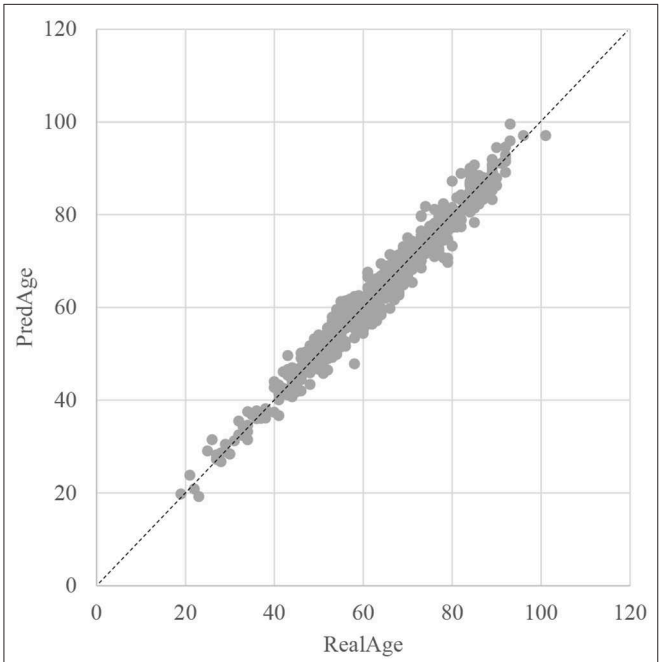


FIGURE 3 | Dot plot between the real chronological age and the predicted age of these samples. The horizontal axis was the chronological age of a sample (RealAge) and the vertical axis was this sample's age averaged over the 10-fold cross validation (PredAge). The regressor was SVR. The perfect prediction of age was represented by the gapped line y=x.

metrics RMSE and R2 were 1.6149 and 0.9672, respectively. The regression coefficients of the methylomic features were given in **Supplementary Table 1**.

Comparison With Other Commonly Used Feature Selection Algorithms

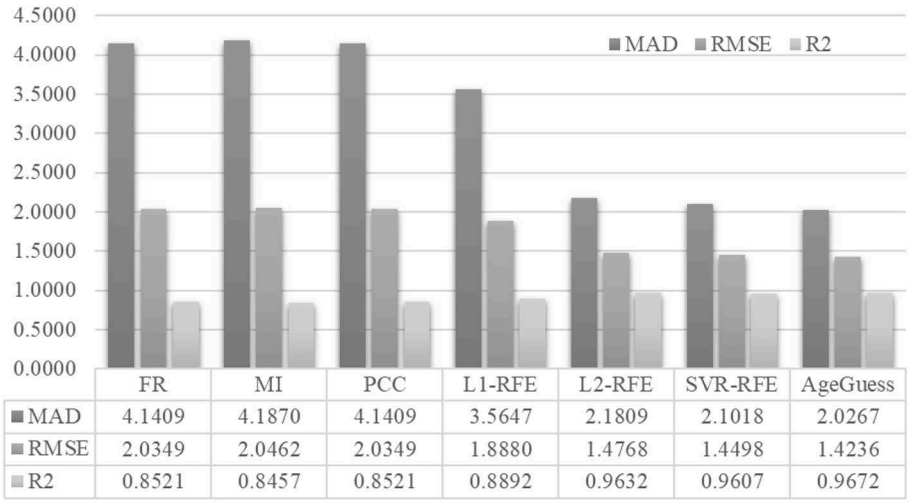
This study compared the proposed AgeGuess with the existing feature selection algorithms. Three filter algorithms were evaluated, i.e., the uni-variate F-Regression (FR), Mutual Information (MI), and Pearson Correlation Coefficient (PCC). Filter algorithms returned an ordered list of all the features and the same number of features as AgeGuess was used for a fair comparison. Three recursive feature elimination (RFE) algorithms were also compared with AgeGuess, i.e., L1-RFE, L2-RFE, and SVR-RFE. An RFE algorithm eliminated a feature if its removal induced the least regression performance loss. And the regression performances of the above three RFE algorithms were calculated by the L1-regularized, L2-regularized and Support-Vector-based regressors, respectively. The number of selected features was an importance factor of a feature selection algorithm. So we also set the number of features selected by these RFE algorithms to the same as AgeGuess.

Figure 4 demonstrated that AgeGuess outperformed the existing feature selection algorithms in all the three regression performance metrics. AgeGuess achieved 2.0267 in MAD, which was 2.1142 smaller than that of FR and 2.1603 smaller than that of MI. A larger R2 value suggested that a regressor performed better. AgeGuess achieved the best R2 and outperformed the next best algorithm L2-RFE by 0.0040 in R2. The smaller RMSE was the better. And AgeGuess outperformed the next best algorithm SVR-RFE by 0.0262 in RMSE.

We also compared our best model with the existing age prediction models and AgeGuess performed the best on estimating the chronological ages. Weidner et al. used 102 methylation features from the same dataset as this study to establish their age predictor, which achieved 4.12 in MAD, 5.34 in RMSE and 0.87 in R2 (Weidner et al., 2014). Another study also used the same dataset as this study and detected 41 methylomic features as the age biomarkers. They built the age predictor achieving 10.69 in MAD (Sarac et al., 2017). The same features from the study (Shadrina et al., 2018) were used to train the regressor as in this study and the age predictor only achieved 9.9017 in MAD, 12.1120 in RMSE and 0.0521 in R2, respectively.

Gender Specificity of Age Prediction

The literature provided different ideas on the correlations between aging and gender variations. Hannum et al. proposed that aging was impacted by various factors and utilized the information of gender and body mass index (BMI) together with the methylomic features in building an age predictor (Hannum et al., 2013). Their model achieved 3.9 years in the age prediction errors and 96% in the correlations of the predicted ages with the chronological ages. Their data suggested that gender was a



	FR	MI	PCC	L1-RFE	L2-RFE	SVR-RFE	AgeGuess
MAD	4.1409	4.1870	4.1409	3.5647	2.1809	2.1018	2.0267
RMSE	2.0349	2.0462	2.0349	1.8880	1.4768	1.4498	1.4236
R2	0.8521	0.8457	0.8521	0.8892	0.9632	0.9607	0.9672

FIGURE 4 | Performance comparison of AgeGuess with six existing feature selection algorithms. The regression performance metrics MAD, RMSE, and R2 were Mean Absolute Error, squared root of mean squared error, and the Goodness of Fit (R2), respectively.

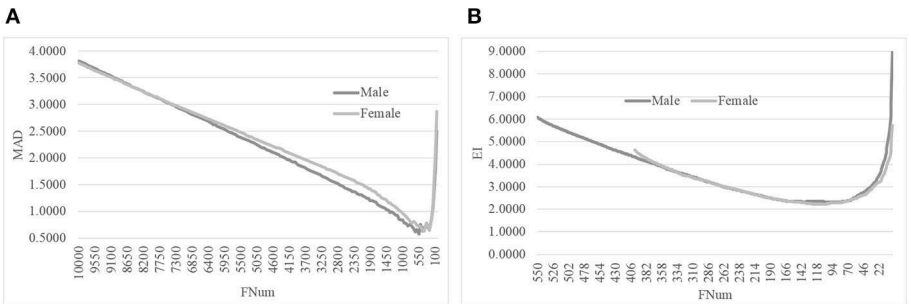


FIGURE 5 | Line plots of AgeGuess's steps 2 and 3. **(A)** The second step of AgeGuess screened features using SVR-RFE. **(B)** The third step of AgeGuess further eliminated redundant features by backFS.

significant factor to the aging rate. But professor Steve Horvath hypothesized that an age-dependent CpG signatures may be defined independent of genders and his group built a gender-independent age predictor achieving 3.6 years in the metrics median error.

We evaluated this hypothesis with the gender-specific models using the same feature selection algorithm on the same dataset, as shown in **Figure 5**. The original dataset was split into the dsMale and dsFemale datasets, and the same feature selection procedure AgeGuess was carried out on these two datasets. **Figure 5A** suggested that AgeGuess achieved 0.5783 and 0.6287 in MAD for the datasets dsMale and dsFemale, respectively. **Figure 5B** demonstrated that the last step of AgeGuess further refined the gender-specific models to achieve 2.2954 and 2.2148 in EI, respectively. So the Male and Female models outperformed the model using the dataset dsMale∪dsFemale by at least 0.6605 in MAD. And the gender-specific models used the similar numbers of features compared with the original model using the dataset combined from both dsMale and dsFemale.

The SVR regression model trained on the dataset dsMale achieved 1.5072 in MAD, 1.3804 in RMSE and 0.9832 in R2. The three performance metrics of the model trained on dsFemale were 1.1669, 1.2112, and 0.9881, respectively. So both gender-specific models outperformed the best model trained over dsFemale∪dsMale, which achieved 2.0267 in MAD, 1.6149 in RMSE and 0.9672 in R2. The dot plots in **Figure 6** illustrated how well gender-specific age prediction models achieved on estimating the chronological ages. The regression coefficients of the methylomic features for the two gender-specific models were given in **Supplementary Tables 2, 3**.

Evaluating AgeGuess on Another Methylomic Dataset on the EPIC BeadChip

A new methylation probing array, the Infinium MethylationEPIC (EPIC array), was recently launched and provided 868564 methylomic features, which was almost two times as that of the Illumina 450 k array. The EPIC array shared about 94% of the probes in the 450 k array (McEwen et al., 2018; Alsaleh and Haddrill, 2019).

AgeGuess was applied to an independent dataset GSE116339 generated on the EPIC arrays (Curtis et al., 2019). This dataset

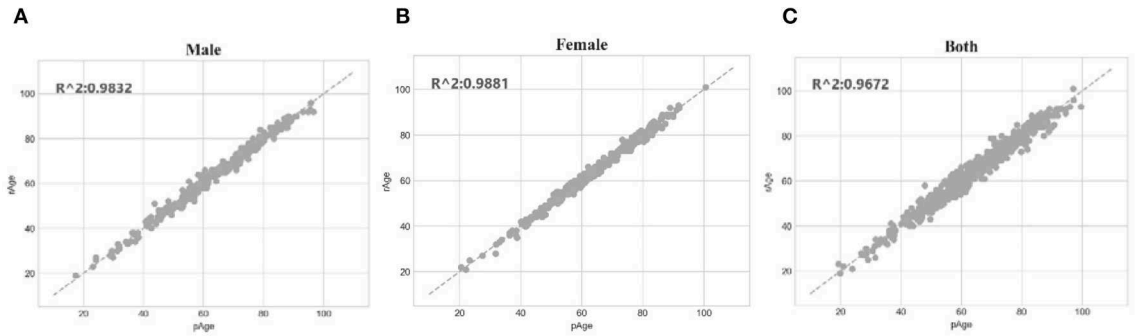


FIGURE 6 | Gender-specific age prediction performances. **(A)** Dot plot for male samples. **(B)** Dot plot for female samples. **(C)** Dot plot for the dataset combined both male and female samples. The perfect prediction of age was represented by the gapped line y=x.

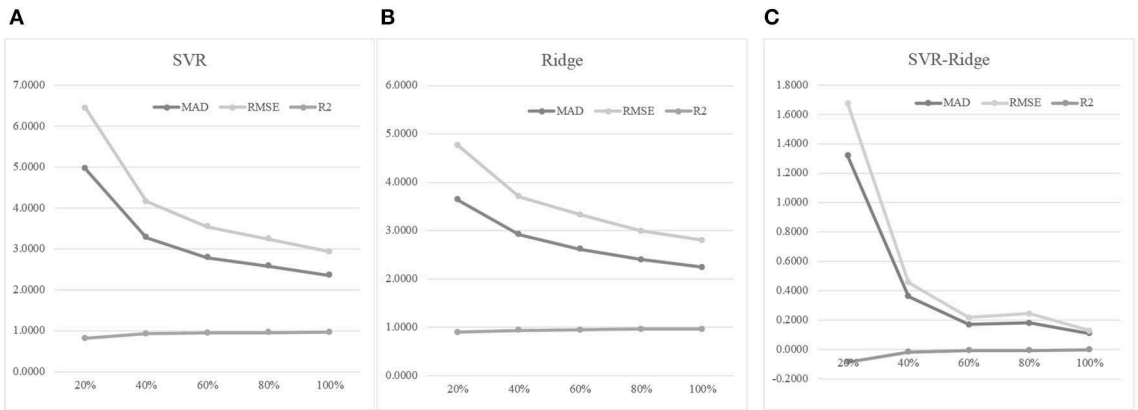


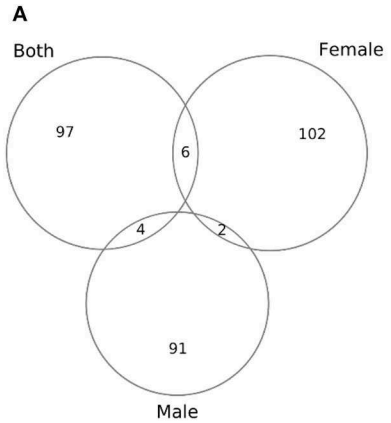
FIGURE 7 | The training dataset size was important for the age prediction performance. **(A)** The regression model was trained using the regressor SVR. **(B)** The regression model was trained using the regressor Ridge. **(C)** The performance metrics of SVR minus those of Ridge. The horizontal axis was the percentage of the training dataset used for training the model. The three regression performance metrics MAD, RMSE and R2 were calculated.

was publicly available from the database Gene Expression Omnibus (Clough and Barrett, 2016) and provided the methylomes of 679 whole blood samples with the chronological ages (Curtis et al., 2019). AgeGuess finally selected 388 CpG features to establish the age prediction model. Two hundred fourteen of these 388 features were shared with the 450 k array and the other 174 features were EPIC-specific. The Mean Absolute Deviation (MAD) of this model was 2.4780, while the other two metrics RMSE and R2 were 1.8101 and 0.9319, respectively. So the EPIC array-based model performed slightly worse in the metrics MAD than the model based on the 450 k array. And it also used more than three times of features than the 450 k array-based model. The experimental data suggested that the EPIC array may need the 6% of the 450 k array-specific methylomic features to precisely describe the aging process.

Impact of Training Dataset Sizes on Age Prediction Performances

An experiment series was carried out to evaluate how different numbers of training samples may impact the age prediction performances, as shown in **Figure 7**. Firstly, 30% of the whole dataset was randomly selected as the test dataset. Then we randomly selected 20, 40, 60, 80, and 100% of the remaining samples to train the regression models, and tested the model prediction performances on the test dataset. **Figure 7A** suggested that more training samples did improve the regression model's performances. The 40% model improved the 20% model by 33.94% in MAD, but the 60% model only achieved a 14.94% improvement in MAD compared with the 40% model. And even smaller improvements were achieved when more training samples were added. Similar patterns were observed for the other two regression performance metrics RMSE and R2.

Another regression algorithm Ridge was evaluated for its age prediction performances using the same features, as shown in **Figures 7B,C**. The Ridge-based age prediction models also demonstrated a similar pattern on different numbers of training samples, as shown in **Figure 7B**. After 60% of samples in the training dataset was used to train the model, more training samples didn't facilitate a major model improvement. We calculated the metrics differences between SVR and Ridge, as shown in **Figure 7C**. A small value of MAD or RMSE suggested



B

	BothModel	MaleModel	FemaleModel
Both	2.0267	3.6709	3.0453
Male	2.3385	1.5072	4.9130
Female	2.2405	5.6581	1.1669

FIGURE 8 | Gender-specific methylomic biomarkers for age prediction. **(A)** Venn plot of the three sets of methylomic biomarkers. The dataset "Both" included both male and female samples. The two datasets "Male" and "Female" consisted of male and female samples, respectively. **(B)** Each column gave the metrics MAD values of the age regression SVR model trained by the biomarker set denoted on the first row. BothModel, MaleModel and FemaleModel denoted the sets of biomarkers detected using the datasets Both, Male, and Female.

a good age prediction model, and **Figure 7C** illustrated that the MAD or RMSE values of Ridge were always smaller than those of SVR. And a large R2 value suggested a good regression model. **Figure 7C** illustrated that Ridge was always larger than SVR in the performance metrics R2. So the regression algorithm Ridge outperformed SVR in all the three regression performance metrics MAD, RMSE, and R2.

The Biological Relevance of Age Biomarkers to the Aging Process

Figure 8 illustrated that there were little overlaps between the gender-specific methylomic biomarkers, and there were no methylomic biomarkers shared among the three sets of biomarkers BothModel/MaleModel/FemaleModel. The data suggested that there existed differences in aging biomarkers between males and females. Even the aging biomarkers of the BothModel performed worse on the individual genders (datasets dsMale and dsFemale). And the cross-gender validation demonstrated much worse age regression performances, as shown in **Figure 8B**.

Some of the gender-specific age methylomic biomarkers were known to have gender-biased expression patterns (Gershoni and Pietrokovski, 2017). There were two female-biased age methylomic biomarkers were cg06419846 (gene CD248) and cg25371036 (gene AMOTL1), which were from the chromosome 11 (Gershoni and Pietrokovski, 2017). CD248 was observed to be hypermethylated during aging and suggested the impaired T cell functionality in the aged adults (Tserel et al., 2015). AMOTL1 (Angiomotin Like 1) was also differentially expressed in different age groups of females, which was verified by the quantitative real-time PCR (qRT-PCR) (Pelissier et al., 2014).

Some of the male-specific age methylomic biomarkers in this study were also supported by the literature. Both of the two biomarkers cg25478614 (gene SST) and cg04084157 (gene VGF) were observed to exhibit male-biased expression patterns (Gershoni and Pietrokovski, 2017). The gene SST received hypermethylation to decline its expressions gradually with age (McKinney et al., 2015). The SST+ neurons may also be impacted with chronic exposures to different photoperiods and resulted in behavioral alternations (Pritchard et al., 2019). The gene VGF encoded the Nerve Growth Factor Inducible protein and gradually increased its expressions in the T lymphocytes when the host age increases (Busse et al., 2014).

These gender-specific biomarker genes were screened by the online GO (Gene Ontology) analysis system DAVID version 6.8 (Huang da et al., 2009a,b). The biomarker genes were input as the foreground and the species Homo sapiens was chosen as the background. The enriched terms with $P \leq 0.05$ in the functional annotation chart were collected for further analysis, as shown in **Supplementary Table 4**. **Figure 8A** suggested that the three datasets dsBoth, dsFemale and dsMale shared very few biomarkers. **Supplementary Table 4** further supported the observation with that only one GO term (biological process "regulation of catalytic activity") was shared by two datasets dsBoth and dsMale. The top two ranked terms in the female biomarkers were two molecular function terms "RNA polymerase II transcription factor activity, ligand-activated sequence-specific DNA binding" and "RNA polymerase II core promoter proximal region sequence-specific DNA binding." The female-specific aging associated RNA polymerase II activities were supported by the experimental evidences observed from the female rat brain (Shults et al., 2015) and the female rat liver (Spindler et al., 1991). While we focused on the aging biomarkers from the dataset dsBoth, the top-ranked enriched GO term was the biological process "homophilic cell adhesion via plasma membrane adhesion molecules," as shown in **Supplementary Table 4**. It is well-known that the growth hormone was actively involved in the aging process and some of the state-of-the-art results were reviewed in Allshouse et al. (2018) and Bartke (2019).

DISCUSSION

The aging process was impacted by both inherited genetic and environmental factors. Multiple studies demonstrated that the methylomic biomarkers served as a rich information source for predicting the chronological ages (Hong et al., 2017; Shadrina et al., 2018). Most of the existing studies selected their age biomarkers based on these biomarkers' biological relevance to the aging process (Zubakov et al., 2016) or statistically correlations with the chronological ages (Shadrina et al., 2018).

This study hypothesized that the chronological age may be more accurately predicted using delicately chosen methylomic biomarkers. A three-step feature selection algorithm AgeGuess was proposed and evaluated for the age regression problem based on the methylomic features. The SVR model using the AgeGuess-selected methylomic biomarkers outperformed the existing age prediction models. Our experimental data suggested that another regression algorithm Ridge achieved a slightly better age regression performance compared with the SVR model. So the AgeGuess-selected features represented important age biomarkers independent of regression algorithms.

This study further investigated whether the age process was gender-specific. The proposed algorithm AgeGuess selected 97 methylomic biomarkers for the male samples, and 110 biomarkers for the females. But there were only two methylomic biomarkers cg26290632 (gene CALB1) and cg07955995 (gene KLF14) selected by AgeGuess in both the male and females samples. Both CALB1 (Loerch et al., 2008) and KLF14 (Small et al., 2011) were known age-related biomarkers. CALB1 demonstrated robustly down-regulated expression across rhesus monkeys and humans (Loerch et al., 2008; Pabba et al., 2017). While KLF14 served as a master regulator of many genes and its altered methylation patterns were associated with the aging process (Spolnicka et al., 2018). But both of these two genes didn't demonstrate gender-specific patterns. So these two genes may be robust age biomarkers without gender-bias. Some of the gender-specific age methylomic biomarkers were also supported by the literature.

The age prediction models proposed in this study may need further validated by various tissue samples. Gene expression patterns differred across tissues, so did patterns of DNA methylation (Decato et al., 2017; Zhou et al., 2017; Slieker et al., 2018). Only whole blood methylation samples were used in this study. Considering the influence factors such as tissues and environments, the age prediction models in this study may have reduced prediction capabilities for forensic samples other than whole blood. In addition, Hannum et al., demonstrated that some electronic health record (EHR) data like BMI may be integrated with the methylomic data to achieve a better age prediction (Hannum et al., 2013). So more types of biomedical data of the participants may further improve the proposed models.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The EPIC array dataset can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116339.

AUTHOR CONTRIBUTIONS

FZ and XG conceived and designed the project and polished the manuscript. XG, SL, XF, and MD wrote the code and conducted the experiments. XG and HS worked on the manuscript revision according to the reviewers' comments. XG and LH discussed the experimental results and drafted the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This research was funded by the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), the Education Department of Jilin Province (JJKH20180145KJ), and the startup grant of the Jilin University. This research was also partially supported by the Bioknow MedAI Institute (BMCPP-2018-001), the High Performance Computing Center of Jilin University, and the Fundamental Research Funds for the Central Universities, JLU.

ACKNOWLEDGMENTS

Insightful comments from the three reviewers were greatly appreciated.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00080/full#supplementary-material

REFERENCES

Allshouse, A., Pavlovic, J., and Santoro, N. (2018). Menstrual cycle hormone changes associated with reproductive aging and how they may relate to symptoms. *Obstet. Gynecol. Clin. North Am.* 45, 613–628. doi: 10.1016/j.ogc.2018.07.004

Alsaleh, H., and Haddrill, P. R. (2019). Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC(R) BeadChip. *Forensic Sci. Int.* 303:109944. doi: 10.1016/j.forsciint.2019.109944

Alvarez, M., and Ballantyne, J. (2006). The identification of newborns using messenger RNA profiling analysis. *Anal. Biochem.* 357, 21–34. doi: 10.1016/j.ab.2006.06.024

Baltes, P. B., and Smith, J. (2003). New frontiers in the future of aging: From successful aging of the young old to the dilemmas of the fourth age. *Gerontology* 49, 123–135. doi: 10.1159/000067946

Banister, J., Bloom, D. E., and Rosenberg, L. (2012). "Population aging and economic growth in China," in *The Chinese Economy*, ed D. E. Bloom (Boston, MA: Springer), 114–149. doi: 10.1057/9781137034298_7

Barrett, E. L., Burke, T. A., Hammers, M., Komdeur, J., and Richardson, D. S. (2013). Telomere length and dynamics predict mortality in a wild longitudinal study. *Mol. Ecol.* 22, 249–259. doi: 10.1111/mec.12110

Bartke, A. (2019). Growth hormone and aging: updated review. *World J. Mens Health* 37, 19–30. doi: 10.5534/wjmh.180018

Busse, S., Steiner, J., Micheel, J., Dobrowolny, H., Mawrin, C., Krause, T. J., et al. (2014). Age-related increase of VGF-expression in T lymphocytes. *Aging* 6, 440–453. doi: 10.18632/aging.100656

Ceglia, N., Liu, Y., Chen, S., Agostinelli, F., Eckel-Mahan, K., Sassone-Corsi, P., et al. (2018). CircadiOmics: circadian omic web portal. *Nucleic Acids Res.* 46, W157–W162. doi: 10.1093/nar/gky441

Chong, A. Y., Doyle, B. J., Jansen, S., Ponosh, S., Cisonni, J., and Sun, Z. (2017). Blood flow velocity prediction in aorto-iliac stent grafts using computational fluid dynamics and Taguchi method. *Comput Biol. Med.* 84, 235–246. doi: 10.1016/j.compbiomed.2017.03.015

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Curtis, S. W., Cobb, D. O., Kilaru, V., Terrell, M. L., Kennedy, E. M., Marder, M. E., et al. (2019). Exposure to polybrominated biphenyl (PBB) associates with genome-wide DNA methylation differences in peripheral blood. *Epigenetics* 14, 52–66. doi: 10.1080/15592294.2019.1565590

Decato, B. E., Lopez-Tello, J., Sferruzzi-Perri, A. N., Smith, A. D., and Dean, M. D. (2017). DNA methylation divergence and tissue specialization in the developing mouse placenta. *Mol. Biol. Evol.* 34, 1702–1712. doi: 10.1093/molbev/msx112

Diao, G., and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics* 194, 781–783. doi: 10.1534/genetics.113.150896

Feng, L., Peng, F., Li, S., Jiang, L., Sun, H., Ji, A., et al. (2018). Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males. *Forensic Sci. Int. Genet.* 35, 38–45. doi: 10.1016/j.fsigen.2018.03.009

Feng, X., Zhang, R., Liu, M., Liu, Q., Li, F., Yan, Z., et al. (2019). An accurate regression of developmental stages for breast cancer based on transcriptomic biomarkers. *Biomark Med.* 13, 5–15. doi: 10.2217/bmm-2018-0305

Fernandez-Jimenez, N., Allard, C., Bouchard, L., Perron, P., Bustamante, M., Bilbao, J. R., et al. (2019). Comparison of Illumina 450K and EPIC arrays in placental DNA methylation. *Epigenetics* 14, 1177–1182. doi: 10.1080/15592294.2019.1634975

Field, A. E., Robertson, N. A., Wang, T., Havas, A., Ideker, T., and Adams, P. D. (2018). DNA methylation clocks in aging: categories, causes, and consequences. *Mol Cell* 71, 882–895. doi: 10.1016/j.molcel.2018.08.008

Fraga, M. F., and Esteller, M. (2007). Epigenetics and aging: the targets and the marks. *Trends Genet.* 23, 413–418. doi: 10.1016/j.tig.2007.05.008

Gershoni, M., and Pietrokovski, S. (2017). The landscape of sex-differential transcriptome and its consequent selection in human adults. *BMC Biol.* 15:7. doi: 10.1186/s12915-017-0352-z

Grissa, D., Petera, M., Brandolini, M., Napoli, A., Comte, B., and Pujos-Guillot, E. (2016). Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Front. Mol. Biosci.* 3:30. doi: 10.3389/fmolb.2016.00030

Guyon, I., and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res. Special Issue Variable Feat. Select.* 3, 1157–1182.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* 49, 359–367. doi: 10.1016/j.molcel.2012.10.016

Hong, S. R., Jung, S. E., Lee, E. H., Shin, K. J., Yang, W. I., and Lee, H. Y. (2017). DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. *Forensic Sci. Int. Genet.* 29, 118–125. doi: 10.1016/j.fsigen.2017.04.006

Horn, V., and Schweppe, C. (2015). "Introduction: transnational aging: current insights and future challenges," in *Transnational Aging*, eds V. Horn and C. Schweppe (New York, NY: Routledge), 13–28. doi: 10.4324/9781315756394

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* 14:3156. doi: 10.1186/gb-2013-14-10-r115

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jones, M. J., Goodman, S. J., and Kobor, M. S. (2015). DNA methylation and healthy human aging. *Aging Cell* 14, 924–932. doi: 10.1111/acel.12349

Le, T. T., Simmons, W. K., Misaki, M., Bodurka, J., White, B. C., Savitz, J., et al. (2017). Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests. *Bioinformatics* 33, 2906–2913. doi: 10.1093/bioinformatics/btx298

Li, R., Potter, T., Huang, W., and Zhang, Y. (2017). Enhancing performance of a hybrid EEG-fNIRS system using channel selection and early temporal features. *Front. Hum. Neurosci.* 11:462. doi: 10.3389/fnhum.2017.00462

Liu, J., Sun, D., Chen, L., Fang, Z., Song, W., Guo, D., et al. (2019). Radiomics analysis of dynamic contrast-enhanced magnetic resonance imaging for the prediction of sentinel lymph node metastasis in breast cancer. *Front. Oncol.* 9:980. doi: 10.3389/fonc.2019.00980

Loerch, P. M., Lu, T., Dakin, K. A., Vann, J. M., Isaacs, A., Geula, C., et al. (2008). Evolution of the aging brain transcriptome and synaptic regulation. *PLoS ONE* 3:e3329. doi: 10.1371/journal.pone.0003329

Ma, L., and Fan, S. (2017). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinform.* 18:169. doi: 10.1186/s12859-017-1578-z

Martino, D. J., Tulic, M. K., Gordon, L., Hodder, M., Richman, T. R., Metcalfe, J., et al. (2011). Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans. *Epigenetics* 6, 1085–1094. doi: 10.4161/epi.6.9.16401

McEwen, L. M., Jones, M. J., Lin, D. T. S., Edgar, R. D., Husquin, L. T., MacIsaac, J. L., et al. (2018). Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array. *Clin Epigenet.* 10:123. doi: 10.1186/s13148-018-0556-2

McKinney, B. C., Lin, C. W., Oh, H., Tseng, G. C., Lewis, D. A., and Sibille, E. (2015). Hypermethylation of BDNF and SST genes in the orbital frontal cortex of older individuals: a putative mechanism for declining gene expression with age. *Neuropsychopharmacology* 40, 2604–2613. doi: 10.1038/npp.2015.107

Moore, L. D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology* 38, 23–38. doi: 10.1038/npp.2012.112

Ou, X. L., Gao, J., Wang, H., Wang, H. S., Lu, H. L., and Sun, H. Y. (2012). Predicting human age with bloodstains by sjTREC quantification. *PLoS ONE* 7:e42412. doi: 10.1371/journal.pone.0042412

Pabba, M., Scifo, E., Kapadia, F., Nikolova, Y. S., Ma, T., Mechawar, N., et al. (2017). Resilient protein co-expression network in male orbitofrontal cortex layer 2/3 during human aging. *Neurobiol. Aging* 58, 180–190. doi: 10.1016/j.neurobiolaging.2017.06.023

Pal, S., and Tyler, J. K. (2016). Epigenetics and aging. *Sci. Adv.* 2:e1600584. doi: 10.1126/sciadv.1600584

Pan, X., Liu, B., Wen, X., Liu, Y., Zhang, X., Li, S., et al. (2019). D-GPM: a deep learning method for gene promoter methylation inference. *Genes* 10:807. doi: 10.3390/genes10100807

Pelissier, F. A., Garbe, J. C., Ananthanarayanan, B., Miyano, M., Lin, C., Jokela, T., et al. (2014). Age-related dysfunction in mechanotransduction impairs differentiation of human mammary epithelial progenitors. *Cell Rep.* 7, 1926–1939. doi: 10.1016/j.celrep.2014.05.021

Pritchard, R., Chen, H., Romoli, B., Spitzer, N. C., and Dulcis, D. (2019). Photoperiod-induced neurotransmitter plasticity declines with aging: an epigenetic regulation? *J. Comp. Neurol.* 582, 199–210. doi: 10.1101/563213

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting novel associations in large data sets. *Science* 334, 1518–1524. doi: 10.1126/science.1205438

Saeed, M., Berlin, R. M., and Cruz, T. D. (2012). Exploring the utility of genetic markers for predicting biological age. *Leg. Med.* 14, 279–285. doi: 10.1016/j.legalmed.2012.05.003

Sarac, F., Seker, H., and Bouridane, A. (2017). Exploration of unsupervised feature selection methods to predict chronological age of individuals by utilising CpG dinucleotics from whole blood. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2017, 3652–3655. doi: 10.1109/EMBC.2017.8037649

Schellenberg, A., Lin, Q., Schuler, H., Koch, C. M., Joussen, S., Denecke, B., et al. (2011). Replicative senescence of mesenchymal stem cells causes DNA-methylation changes which correlate with repressive histone marks. *Aging* 3, 873–888. doi: 10.18632/aging.100391

Shadrina, A., Tsepilov, Y., Sokolova, E., Smetanina, M., Voronina, E., Pakhomov, E., et al. (2018). Genome-wide association study in ethnic Russians suggests an association of the MHC class III genomic region with the risk of primary varicose veins. *Gene* 659, 93–99. doi: 10.1016/j.gene.2018.03.039

Shults, C. L., Pinceti, E., Rao, Y. S., and Pak, T. R. (2015). Aging and loss of circulating 17β-estradiol alters the alternative splicing of ERβ in the female rat brain. *Endocrinology* 156, 4187–4199. doi: 10.1210/en.2015-1514

Slieker, R. C., Relton, C. L., Gaunt, T. R., Slagboom, P. E., and Heijmans, B. T. (2018). Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenet. Chromatin* 11:25. doi: 10.1186/s13072-018-0191-3

Small, K. S., Hedman, A. K., Grundberg, E., Nica, A. C., Thorleifsson, G., Kong, A., et al. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* 43, 561–564. doi: 10.1038/ng1011-1040c

Solorio-Fernández, S., Carrasco-Ochoa, J. A., and Martínez-Trinidad, J. F. (2016). A new hybrid filter–wrapper feature selection method for clustering based on ranking. *Neurocomputing* 214, 866–880. doi: 10.1016/j.neucom.2016.07.026

Spindler, S. R., Grizzle, J. M., Walford, R. L., and Mote, P. L. (1991). Aging and restriction of dietary calories increases insulin receptor mRNA, and aging increases glucocorticoid receptor mRNA in the liver of female C3B10RF1 mice. *J. Gerontol.* 46, B233–237. doi: 10.1093/geronj/46.6.B233

Spolnicka, M., Pospiech, E., Adamczyk, J. G., Freire-Aradas, A., Peplonska, B., Zbiec-Piekarska, R., et al. (2018). Modified aging of elite athletes revealed by analysis of epigenetic age markers. *Aging* 10, 241–252. doi: 10.18632/aging.101385

Suto, J., Oniga, S., and Sitar, P. P. (2016). "Comparison of wrapper and filter feature selection algorithms on human activity recognition," in *2016 6th International Conference on Computers Communications and Control* (Oradea: ICCCC), 124–129. doi: 10.1109/ICCCC.2016.7496749

Suzuki, T., Yamazaki, H., Honda, K., Ryo, E., Kaneko, A., Ota, Y., et al. (2019). Altered DNA methylation is associated with aberrant stemness gene expression in earlystage HNSCC. *Int. J. Oncol.* 55, 915–924. doi: 10.3892/ijo.2019.4857

Thompson, N. R., Katzan, I. L., Honomichl, R. D., and Lapin, B. R. (2019). PROMIS global health item nonresponse: is it better to impute missing item responses before computing T-scores? *Qual. Life Res.* 29, 537–546. doi: 10.1007/s11136-019-02327-1

Tian, Q., Zou, J., Fang, Y., Yu, Z., Tang, J., Song, Y., et al. (2019). A hybrid ensemble approach for identifying robust differentially methylated loci in pan-cancers. *Front. Genet.* 10:774. doi: 10.3389/fgene.2019.00774

Tserel, L., Kolde, R., Limbach, M., Tretyakov, K., Kasela, S., Kisand, K., et al. (2015). Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. *Sci. Rep.* 5:13107. doi: 10.1038/srep13107

Vaillancourt, K., Ernst, C., Mash, D., and Turecki, G. (2017). DNA methylation dynamics and cocaine in the brain: progress and prospects. *Genes* 8:138. doi: 10.3390/genes8050138

Vidaki, A., Ballard, D., Aliferi, A., Miller, T. H., Barron, L. P., and Syndercombe Court, D. (2017). DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Sci. Int. Genet.* 28, 225–236. doi: 10.1016/j.fsigen.2017.02.009

Vidaki, A., and Kayser, M. (2018). Recent progress, methods and perspectives in forensic epigenetics. *Forensic Sci. Int. Genet.* 37, 180–195. doi: 10.1016/j.fsigen.2018.08.008

Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P., et al. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol.* 15:R24. doi: 10.1186/gb-2014-15-2-r24

Ye, Y., Zhang, R., Zheng, W., Liu, S., and Zhou, F. (2017). RIFS: a randomly restarted incremental feature selection algorithm. *Sci. Rep.* 7:13013. doi: 10.1038/s41598-017-13259-6

Yi, S. H., Jia, Y. S., Mei, K., Yang, R. Z., and Huang, D. X. (2015). Age-related DNA methylation changes for forensic age-prediction. *Int. J. Legal Med.* 129, 237–244. doi: 10.1007/s00414-014-1100-3

Zhang, Y., Chen, C., Duan, M., Liu, S., Huang, L., and Zhou, F. (2019). BioDog, biomarker detection for improving identification power of breast cancer histologic grade in methylomics. *Epigenomics* 11:1717–1732. doi: 10.2217/epi-2019-0230

Zhou, J., Sears, R. L., Xing, X., Zhang, B., Li, D., Rockweiler, N. B., et al. (2017). Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation. *BMC Genomics* 18:724. doi: 10.1186/s12864-017-4115-6

Zubakov, D., Liu, F., Kokmeijer, I., Choi, Y., and van Meurs, J. B. J. (2016). Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Forensic Sci. Int. Genet.* 24, 33–43. doi: 10.1016/j.fsigen.2016.05.014

Zubakov, D., Liu, F., van Zelm, M. C., Vermeulen, J., Oostra, B. A., van Duijn, C. M., et al. (2010). Estimating human age from T-cell DNA rearrangements. *Curr. Biol.* 20, R970–971. doi: 10.1016/j.cub.2010.10.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fgene.2020.00391

Development of an Early Prediction Model for Subarachnoid Hemorrhage With Genetic and Signaling Pathway Analysis

Wanjing Lei[1], *Han Zeng*[2], *Hua Feng*[3,4], *Xufang Ru*[3,4], *Qiang Li*[3,4], *Ming Xiao*[1], *Huiru Zheng*[5], *Yujie Chen*[3,4]* *and Le Zhang*[1,2]*

[1] College of Computer Science, Sichuan University, Chengdu, China, [2] College of Computer and Information Science, Southwest University, Chongqing, China, [3] Department of Neurosurgery, Southwest Hospital, Third Military Medical University, Chongqing, China, [4] State Key Laboratory of Trauma, Burn and Combined Injury, Third Military Medical University, Chongqing, China, [5] School of Computing, Ulster University, Coleraine, United Kingdom

Subarachnoid hemorrhage (SAH) is devastating disease with high mortality, high disability rate, and poor clinical prognosis. It has drawn great attentions in both basic and clinical medicine. Therefore, it is necessary to explore the therapeutic drugs and effective targets for early prediction of SAH. Firstly, we demonstrate that LCN2 can effectively intervene or treat SAH from the perspective of cell signaling pathway. Next, three potential genes that we explored have been validated by manually reviewed experimental evidences. Finally, we turn out that the SAH early ensemble learning predictive model performs better than the classical LR, SVM, and Naïve-Bayes models.

Keywords: bioinformatics, genomics, big data, artificial intelligence, genetics

OPEN ACCESS

Edited by:
Yi Zhao,
Beijing University of Chinese Medicine, China

Reviewed by:
Ping Luo,
University Health Network, Canada
Zhi-Ping Liu,
Shandong University, China
Sheng Chen,
Zhejiang University, China

***Correspondence:**
Yujie Chen
yujiechen6886@foxmail.com
Le Zhang
zhangle06@scu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *06 October 2019*
Accepted: *30 March 2020*
Published: *21 April 2020*

Citation:
Lei W, Zeng H, Feng H, Ru X, Li Q, Xiao M, Zheng H, Chen Y and Zhang L (2020) Development of an Early Prediction Model for Subarachnoid Hemorrhage With Genetic and Signaling Pathway Analysis. Front. Genet. 11:391. doi: 10.3389/fgene.2020.00391

INTRODUCTION

Subarachnoid hemorrhage (SAH) is the fastest developing and most critical hemorrhagic cerebrovascular disease, accounting for 5% of cerebrovascular diseases (Macdonald, 2014), and is associated with high rates of mortality and disability and poor clinical prognosis (Suarez et al., 2006). Although there have been significant advances in diagnostic methods, surgery, and endovascular techniques in recent years, the mortality rate of SAH remains as high as 15% (Macdonald et al., 2008).

Recent research has shown that early brain injury (EBI) may be the main cause of poor prognosis in SAH patients. Therefore, current SAH studies focus on exploring therapeutic drugs and targets for reduction of EBI after SAH and the early prediction of SAH (Sozen et al., 2011).

Lipocalin 2 (LCN2) is an acute secretory protein that regulates the pathophysiological processes of various organ systems in mammals and participates in the intrinsic immune protection of the central nervous system (CNS) (Flo et al., 2004; Ferreira et al., 2015). Studies of acute white matter injury in a mouse SAH model and the role of LCN2 in injury (Egashira et al., 2014) indicate that LCN2 plays an important part in SAH-induced white matter injury. Since above evidences suggest that LCN2 is closely related to SAH, we propose our first research question: is specific intervention for LCN2 (Warszawska et al., 2013) a promising SAH treatment strategy?

On the other hand, most previous studies (Chu et al., 2011; Ni et al., 2011; Zhang et al., 2017a) have only explored biomarkers for SAH prediction and treatment in a narrow molecular range, rather than taking a genome-wide approach. We propose our second research question: could we use a genome-wide approach to find potential biomarkers for SAH based on the effects of LCN2 treatment?

Previous studies have usually predicted SAH based on diagnostic imaging (Frontera et al., 2006; Ramos et al., 2019) and clinical automation data (Roederer et al., 2014), which may not provide enough predictive power. Thus, we propose our third research question: could we use key genes to build a more powerful early prediction model for SAH?

In this paper, we propose a new research plan to answer the above three research questions. First, we use SAH intervention experiments to screen out candidate genes that are susceptible to LCN2, then employ Fisher's exact test (Xie et al., 2011; Li et al., 2017; Xia et al., 2017; Zhang et al., 2019b) to choose signaling pathways from among the candidates under different experimental conditions. Second, we use E-Bayes (Carlin and Louis, 2010), SVM-RFE (Duan et al., 2005), SPCA (Zou et al., 2006), and statistical tests (Zhang et al., 2016, 2018, 2019b,d, 2020; Xiao et al., 2019) to investigate key genes from experimental data by considering both SAH and LCN2 as factors. Third, we integrate the logistic regression (LR), support vector machine (SVM), and Naive-Bayes algorithms (Xia et al., 2017; Zhang et al., 2017a, 2019a) into an ensemble learning model (Gao et al., 2017; Zhang et al., 2019b) to build a model for early SAH prediction.

First, manual review of the experimental evidence (Osuka et al., 2006; Majdalawieh et al., 2007; Hanafy et al., 2010; Hao et al., 2014; Kwon et al., 2015; Yu et al., 2018) demonstrates that we could intervene or treat SAH by targeting LCN2 from a cell signaling pathway perspective. Next, we explore three key genes that are sensitive to both SAH and LCN2 treatment, again using manual review of the experimental evidence (Huang et al., 2016; Sabo et al., 2017; Yu et al., 2018) to cross-validate the relationships between SAH and these key genes. Finally, we show that our SAH early prediction ensemble-learning model outperforms the classical LR, Naive-Bayes, and SVM models. In summary, we consider that this work provides a novel strategy for the future study of clinical treatment of SAH and related diseases.

MATERIALS AND METHODS

Experimental Configuration

All experimental procedures were approved by the Ethics Committee of Southwest Hospital and were performed in accordance with the guidelines of the National Institutes of Health Guide for the Care and Use of Laboratory Animals.

Intervention Experiment for SAH

The original chip data for this experiment were provided by the Department of Neurosurgery, Southwest Hospital, PLA Military Medical University. SAH and sham-operated models were established; details are given in the **Supplementary Material**. Each experimental group included five mice, and the white matter area of the cerebral cortex was taken for gene chip testing. A total of 10 original chip samples were obtained from the SAH intervention experiments; these were divided equally into two groups as follows.

(1) SAH disease group: brain tissue in the white matter region of the cerebral cortex of SAH mice.

(2) Control group normal-1: brain tissue in the white matter region of the cerebral cortex of normal mice.

The chip was an Affymetrix GeneChip Mouse Gene 1.0 ST Array. Raw data included sample RNA extraction (white matter brain cells from the SAH model and from normal mice), sample RNA quality detection (total RNA>1 ug), cDNA synthesis, sense strand cDNA fragmentation, biotin labeling, chip hybridization, chip elution, and chip scanning. The raw data are available at http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8407.

We then carried out mass analysis and used the R Bioconductor package to perform quality control for each original chip (the SAH disease group and the control group normal-1). In the output gray scale image (**Figure S1**) for each chip sample, each chip name and the four corner patterns were very clear, and the contrast between light and dark was moderate.

The right panel of **Figure 1A** shows the Relative Log Expression (RLE) boxplot for these 10 chips. The center of each sample was close to the position RLE = 0. This indicates that the expression levels of most genes in the sample were consistent. In addition, **Figure S2** describes a normalized unscaled standard errors (NUSE) detection (Marta and Marc, 2014). Since **Figure S2** shows that the center of each sample is close to the position NUSE = 1, we consider that the samples are too stable to have obvious batch effect. Then, we used Robust Multi-chip Analysis (RMA) (Irizarry et al., 2003) for data preprocessing, including background and perfect match probes (PM) correction, normalization, and summarization, to obtain the probe expression data matrix (**Table S1**). Finally, clustering analysis (Liu et al., 2019; Xiao et al., 2019; Zhang et al., 2019c; Wu and Zhang, 2020) (**Figure S3**) shows that the major differences between the chip of each group comes from SAH.

Intervention Experiment for LCN2

Here, in order to interfere with the expression of LCN2, 2 μL of specific short interfering RNAs (siRNAs) was delivered into the lateral ventricle with a Hamilton syringe. The injection was performed 48 h before SAH and three groups were used, as described below. We detail the procedures in the **Supplementary Material**.

(1) SAH-siRNA-LCN2: the SAH model was established and treated with intrathecal injection of LCN2 siRNA, and two samples were taken on the first and third days after surgery.

(2) SAH-siRNA-NC: the SAH model was established and treated with intrathecal NC siRNA, and two samples were taken on the first and third days after surgery, which helped us to remove the interference factors associated with the siRNA vector.

(3) Control group normal-2: the brain tissue of the white matter region of the cerebral cortex without any treatment.

The total number of samples in all experiments was 25 (**Table 1**). RNA sequencing was performed on the samples and the raw data are available at https://www.ncbi.nlm.nih.gov/sra/PRJNA575372.

Workflow of the Study

The workflow of the study is illustrated in **Figure 1**. First, we designed the intervention experiment for SAH detailed in section "Intervention Experiment for SAH", which allowed us to obtain the differential genes under different experimental conditions. Based

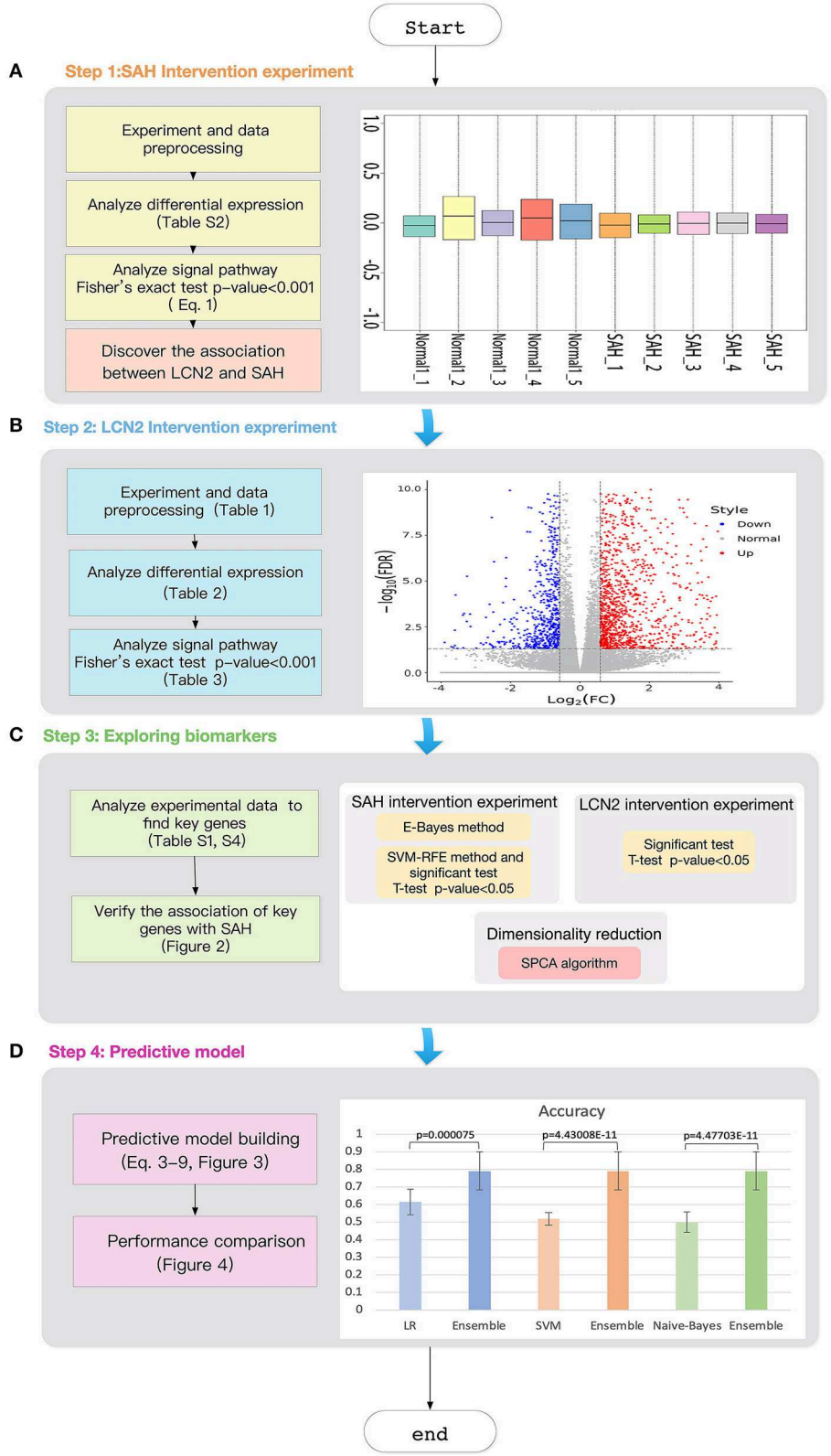


FIGURE 1 | Workflow of the study. **(A)** SAH intervention experimental chip RLE box line diagram; the abscissa is log_2 (Median value of sample expression) and the ordinate represents each chip; **(B)** The volcano map of the comparison group SAH-siRNA-NC (1 day) vs normal-2. The abscissa is $\log_2(Fold\ change)$ and the ordinate is $-log_{10}(FDR)$; The red point is the up-regulated gene, the blue point is the down-regulated gene, and the non-dispersive point is the non-differentiated gene; **(C)** Key gene screening workflow; **(D)** The accuracy for ensemble learning, LR, SVM and Naive-Bayes.

TABLE 1 | Experimental sample description after LCN2 intervention experiment.

Sample	Number of samples	Description
SAH-siRNA-LCN2(1day)	5	Mouse (SAH) brain cells, Intrathecal injection of LCN2 siRNA for 1 day
SAH-siRNA-LCN2(3day)	5	Mouse (SAH) brain cells, Intrathecal injection of LCN2 siRNA for 3 day
SAH-siRNA-NC(1day)	5	Mouse (SAH) brain cells, Intrathecal injection of blank siRNA for 1 day
SAH-siRNA-NC(3day)	5	Mouse (SAH) brain cells, Intrathecal injection of blank siRNA for 3 day
Normal-2	5	Mouse (normal) brain cells, blank control group-2

TABLE 2 | Differential expressed genes for different experimental group.

Experimental group	Total number of genes	Up-regulation of genes	Down-regulation of genes
SAH-siRNA-LCN2(1day) VS normal-2	25342	1541	634
SAH-siRNA-LCN2 (3day) VS normal-2	25055	1264	451
SAH-siRNA-NC(1day) VS normal-2	25384	1159	556
SAH-siRNA-NC(3day) VS normal-2	25564	1297	409
SAH-siRNA- LCN2 (1day) VS SAH-siRNA-NC(1day)	25293	99	14
SAH-siRNA- LCN2 (3day) VS SAH-siRNA-NC(3day)	25251	5	18

on these differential genes, we could identify the key signaling pathways.

As targeting LCN2 could result in changes in these related signaling pathways (causing remission or promotion of SAH), we consider that LCN2 plays an important part in the entire biological cell process for SAH.

Next, we used an intervention experiment for LCN2 to obtain gene expression levels for diseased and normal mouse brain cells at different time points. Then, we employed commonly used dimensional reduction algorithms to explore three key genes under the impact of both SAH and LCN2 treatment.

Finally, we used these three key genes as classifiers to develop an ensemble learning model for early SAH prediction, the predictive power of which was much better than that of the classic LR, Naive-Bayes, and SVM models.

RESULTS

Signaling Pathway Analysis

Differentially Expressed Gene Selection

We used E-Bayes, one of the most commonly used methods for differential expression analysis (Edwards et al., 2005), to screen the differential genes by setting *Fold change* $\geq$ 1.5 and p-value < 0.05. **Table S2** lists 2942 differentially expressed genes, accounting for 10.16% of the total number of genes (28,944). Among them, there were 1016 and 1926 genes with upregulated and downregulated expression (**Figure S4**), respectively.

Pathway Analysis

We used Equation 1 and the data in **Table S3** to explore related signaling pathways by carrying out Fisher's exact test (Xia et al., 2017) using Kobas 3.0 (Wu et al., 2006; Xie et al., 2011; Ai and Kong, 2018) for the differentially expressed genes from **Table S2**.

$$p_F(n_f, n, N_f, N) = 2 * \sum_{x=1}^{n_f} \frac{\binom{n}{x}\binom{N-n}{N_f-x}}{\binom{N}{N_f}} \quad (1)$$

Here, N is the number of genes in the sample and n is the number of genes contained in the pathway. N_f is the number of differentially expressed genes and n_f is the number of differentially expressed genes included in the pathway.

The Fisher's exact test assumes $H_0: p_1 = p_2$; the alternative hypothesis is $H_1: p_1 \neq p_2$. $p_1$is the probability that the differentially expressed gene will fall in the pathway, and p_2 is the probability that the non-differentiated gene does not fall in the pathway. The p-value (p_F) of Fisher's exact test was obtained by Equation 1.

Table S2 lists 70 signaling pathways for which the p-value was less than 0.001. LCN2 is a protein involved in MAPK signaling pathways that protects the CNS as part of the innate immune system (Warszawska et al., 2013). Previous studies have shown that LCN2 activates phosphorylation of p38 MAPK, which phosphorylates the Ser168 and Ser170 sites of NFATc4 and inhibits nuclear translocation of NFATc4 (Olabisi et al., 2008). NFATc4 is a key factor in remyelination and closely related to SAH, indicating that white matter damage after SAH is associated with remyelination (Kao et al., 2009; Guo et al., 2017).

Therefore, we hypothesize that LCN2 could promote the phosphorylation of transcription factor NFATc4 and inhibit its nuclear transcription by activating p38 MAPK, thereby preventing remyelination and causing white matter damage after SAH.

LCN2 Intervention Experimental Results Analysis

To prove our hypothesis, we designed a LCN2 intervention experiment (**Figure 1B**) to test whether LCN2 could affect SAH from the perspective of the differential expressed genes and the related signaling pathways.

First, we used the DESeq2 (Varet et al., 2016) method to select differentially expressed genes from SAH-siRNA-LCN2 and normal-2, SAH-siRNA-NC and normal-2, and SAH-siRNA-LCN2 and SAH-siRNA-NC groups on days 1 and 3, respectively (**Table 1**). The results are shown in **Table 2**, **Table S4**, and **Figure S5**.

TABLE 3 | Cross-validated SAH related signaling pathway.

Experimental group	Important pathways related to SAH
SAH-siRNA-LCN2 (1day) VS normal-2	PI3K-Akt (Hao et al., 2014), Jak-STAT (Osuka et al., 2006), p53 (Yu et al., 2018), TNF (Hanafy et al., 2010), Toll-like receptor (Kwon et al., 2015), NF-kappaβ (Majdalawieh et al., 2007)
SAH-siRNA-LCN2 (3day) VS normal-2	PI3K-Akt (Hao et al., 2014), Jak-STAT (Osuka et al., 2006), p53 (Yu et al., 2018), TNF (Hanafy et al., 2010), Toll-like receptor (Kwon et al., 2015), NF-kappaβ (Majdalawieh et al., 2007)
SAH-siRNA-NC (1day) VS normal-2	PI3K-Akt (Hao et al., 2014), Jak-STAT (Osuka et al., 2006), TNF (Hanafy et al., 2010), Toll-like receptor (Kwon et al., 2015), NF-kappaβ (Majdalawieh et al., 2007)
SAH-siRNA-NC (3day) VS normal-2	PI3K-Akt (Hao et al., 2014), Jak-STAT (Osuka et al., 2006), TNF (Hanafy et al., 2010), Toll-like receptor (Kwon et al., 2015), NF-kappaβ (Majdalawieh et al., 2007)
SAH-siRNA- LCN2 (1day) VS SAH-siRNA-NC (1day)	TNF (Hanafy et al., 2010), Toll-like receptor (Kwon et al., 2015)
SAH-siRNA- LCN2 (3day) VS SAH-siRNA-NC (3day)	Transcriptional misregulation in cancer (Lee and Young, 2013)

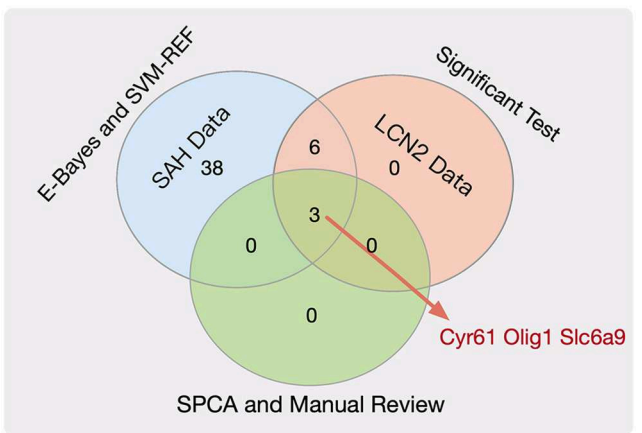


FIGURE 2 | Venn plot for the key genes.

Next, we used Kobas 3.0 (Wu et al., 2006; Xie et al., 2011; Ai and Kong, 2018) to carry out Fisher's exact test for the differential genes in **Table 2**, to identify related signaling pathways (**Table S5**). Next, we used the manually reviewed evidence (Osuka et al., 2006; Majdalawieh et al., 2007; Hanafy et al., 2010; Hao et al., 2014; Kwon et al., 2015; Yu et al., 2018) to cross-validate the SAH-related signaling pathways in **Table S5**. **Table 3** lists the cross-validated SAH-related signaling pathways.

As shown in **Table 3**, all the experimental groups had SAH-related signaling pathways except the transcriptional misregulation in cancer signaling pathway (Lee and Young, 2013) in the SAH-siRNA-LCN2 (3 day) vs. SAH-siRNA-NC (3 day) experimental group. However, as one of the proteins from this pathway, Gzmb (**Table S5**), is closely associated with post-ischemic brain cell death (Chaitanya et al., 2010), we consider that it could be a new target for secondary brain injury inhibition (Armstrong et al., 2017). Therefore, we conclude that specific intervention for LCN2 is a promising SAH treatment strategy.

Feature Selection

After demonstrating the impact of LCN2 on SAH, we chose potential biomarkers for SAH using a genome-wide approach. **Figure 1C** shows the workflow used to choose key genes that were not only related to both SAH and LCN2 but were also insensitive to treatment at different time points. **Figure 1C** shows the following three modules.

(1) SAH intervention experiment module

Owing to the large number of differential genes (**Table S2**), it was necessary to further narrow down the scope of the screening. First, we used the E-Bayes method (Edwards et al., 2005) to filter the probe expression data matrix (**Table S1**) by the E-Bayes function of R's limma package (Smyth et al., 2005). The differential probes were obtained by setting the filter parameters to *Fold change* ≥ 2 and p-value < 0.05.

Second, we used SVM-RFE (Duan et al., 2005) (Equation 2) to rank the genes in the probe expression data matrix, and then carried out the t-test and F-test (Zhang et al., 2017b) for the top 100 genes.

$$\begin{cases} DJ(i) = (1/2)\alpha^T H\alpha - (1/2)\alpha^T H(-i)\alpha \\ H = y_i y_j K(x_i, x_j) \end{cases} \quad (2)$$

where y_i and y_j represent the classification labels of probes x_i and x_j, respectively; $K(x_i, x_j)$ is the kernel function, $i, j = 1, 2, \ldots, n$; α is obtained by training the SVM classifier; $DJ(i)$ is the sort function; and H is the matrix.

We then combined the results of these two methods to obtain the significant probes for both the E-Bayes and SVM-RFE methods.

Finally, we used the transcription cluster annotation file (version: MoGene-1_0-st-v1) downloaded from the Affy (Gautier et al., 2004) website to extract the gene ID for these probes, resulting in 47 key genes (**Table S6**).

(2) LCN2 intervention experiment module

We performed t-tests and F-tests (Zhang et al., 2017b) for the key genes (**Table S6**) in the SAH-siRNA-LCN2 (1 day) vs. normal-2 and SAH siRNA-LCN2 (3 day) vs. normal-2 groups (**Table S4**).

There were 15 and 13 statistically significantly differential genes for the SAH-siRNA-LCN2 (1 day) vs. normal-2 group (**Table S7**) and the SAH-siRNA-LCN2 (3 day) vs. normal-2 group (**Table S8**), respectively. Taking the intersection of the results from these two experimental groups gave nine key genes, Tk1, Cyr61, Nupr1, Dcn, Lum, Olig1, Pcolce2, Slc6a9, and Kcnt2, which were sensitive to both SAH and LCN2 intervention, regardless of treatment, at different time points.

(3) Dimensional reduction module

Next, we employed the SPCA algorithm (Zou et al., 2006; Li et al., 2017) to perform dimensional reduction for the nine

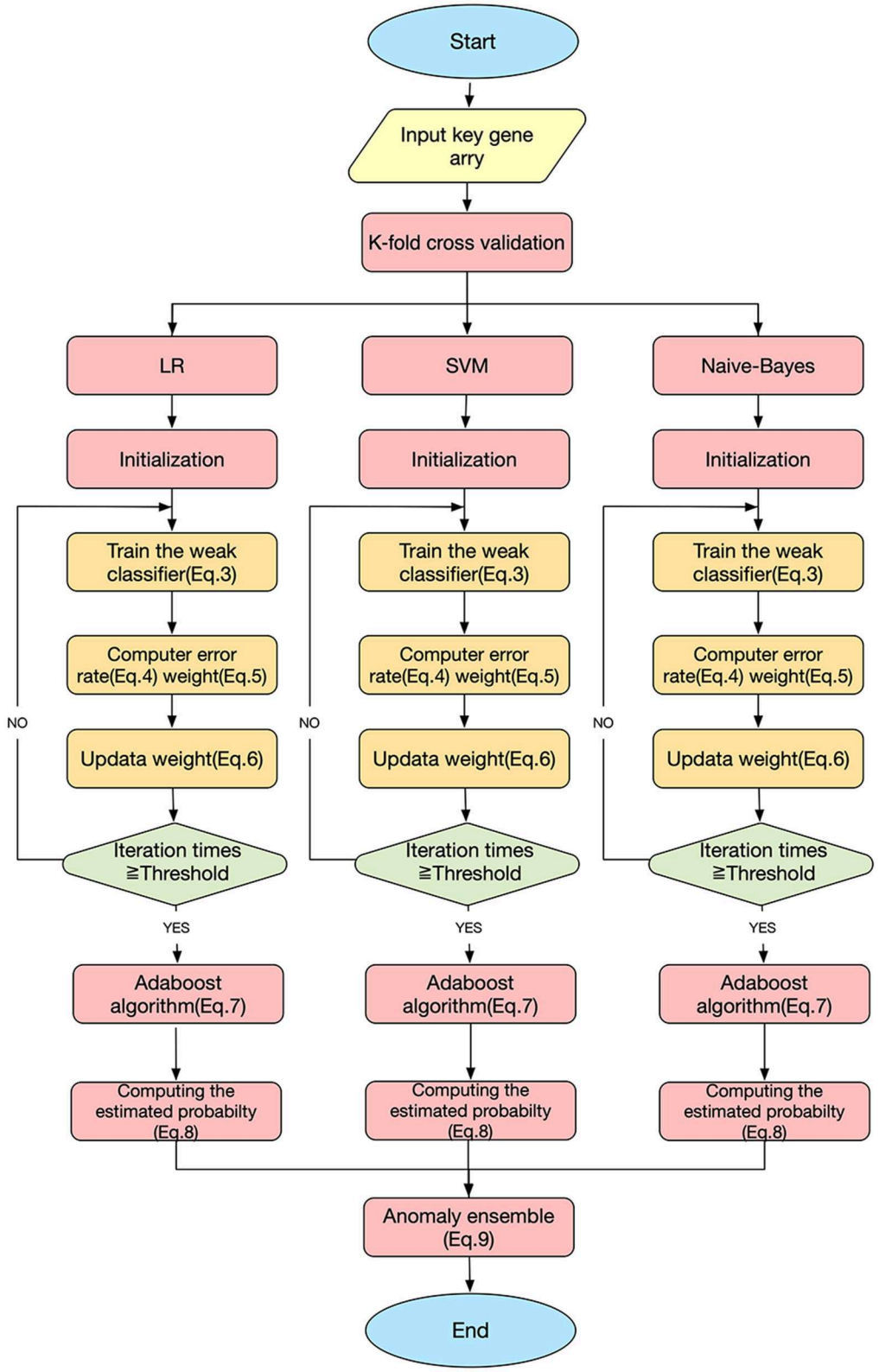


FIGURE 3 | SAH predictive ensemble learning model.

key genes. This resulted in five candidate genes (Tk1, Cyr61, Olig1, Slc6a9, and Pcolce2). However, manual review of the experimental evidence indicated that only Cyr61 (Yu et al., 2018), Olig1 (Sabo et al., 2017), and Slc6a9 (Huang et al., 2016) were closely related to SAH, cerebral hemorrhage, and brain injury. Therefore, we considered these three genes (**Figure 2**, **Table S9**) to be potential biomarkers for SAH.

Ensemble Learning Model

Early SAH Prediction Model

This study used three classification algorithms, LR (Hosmer et al., 2013), SVM (Suykens and Vandewalle, 1999), and Naive-Bayes (Wang et al., 2007) to develop the SAH prediction model, using the selected key genes as the respective classifiers. These three classic methods were then integrated into a novel ensemble learning model to improve the predictive accuracy.

Figure 3 shows the workflow of the SAH prediction model, based on our previous studies (Li et al., 2017; Xia et al., 2017; Zhang et al., 2019b). The key equations of the model are as follows.

$$D_t(i) = \frac{1}{n} \tag{3}$$

$$\varepsilon_t = \frac{number\ of\ incorrectly\ classified\ samples}{total\ number\ of\ samples} \tag{4}$$

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \tag{5}$$

$$D_{t+1}(i) = \frac{D_t(i)}{sum(D)} \begin{cases} \exp(-\alpha_t), & if\ h_t(x_i) = y_i \\ \exp(\alpha_t), & if\ h_t(x_i) \neq y_i \end{cases} \tag{6}$$

$$H_m(x) = sign\sum_{t=0}^{T}\alpha_t h_t(x) \tag{7}$$

$$E_{H_m,} = \sum_{m=1}^{3} P_{H_m} \tag{8}$$

$$Y(x) = \begin{cases} 1 & E_{H_m} \geq 0.5 \\ 0 & E_{H_m} < 0.5 \end{cases} \tag{9}$$

Here, $D_t(i)$ is the weight distribution, t is the iteration time, i is the index of the sample, and n is the number of the sample. ε_t and α_t are the error rate and weight of each weak classifier h_t, respectively. For a sample set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, x_n are the samples and $y_n \in \{0, 1\}$ are the labels; $y_i = 0$ indicates that x_i is not an SAH patient, and $y_i = 1$ indicates that x_i is an SAH patient. H_m is the homomorphic integration for each weak classifier h_t; m is the index of the weak classifier, m$= 1, 2, 3$; T is the threshold of the iteration time; P_{H_m} is the predictive probability of disease; and E_{H_m} is the estimated probability of the model H_m. $Y(x)$ is the result of the final classifier obtained by a voting method (Dietterich, 2000).

Predictive Performance Comparison

Figure 4A compares the classification performance for the LR, Naive-Bayes, SVM, and ensemble learning models, based on four commonly used classification measurements (**Table S10**) (Zhang et al., 2019b). The numerical values used in **Figure 4A** are listed in **Table S11**; these demonstrate that the ensemble learning method outperforms the other three methods with respect to accuracy, precision, sensitivity and specificity. The ROC chart plotted in **Figure 4B** compares the classification effects of LR, Naive-Bayes, SVM, and ensemble learning models. The classification effect of ensemble learning models is also superior to the other three.

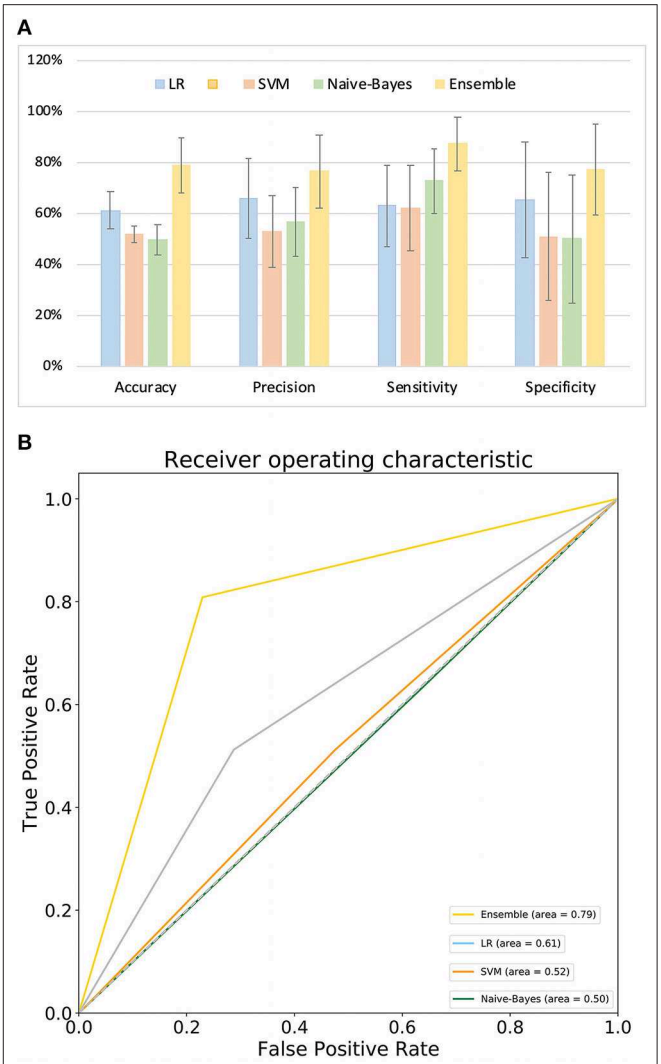


FIGURE 4 | Model performance. **(A)** Comparison of classification performance of LR, SVM, Naive-Bayes, and ensemble learning model; **(B)** ROC chart plotted for LR, SVM, Naive-Bayes, and ensemble learning model.

DISCUSSION

This study aimed to interrogate the potential therapeutic targets of SAH and use them as classifiers to develop a model for early prediction of SAH.

To achieve this aim, we proposed the following three scientific questions. First, is specific intervention involving LCN2 a promising SAH treatment strategy? Second, could we

choose potential biomarkers for SAH at a genome-wide level by considering the effects of LCN2? Third, could we use key genes to build an SAH early prediction model with strong predictive power?

Regarding the first question, as the manually reviewed experimental evidence (Osuka et al., 2006; Majdalawieh et al., 2007; Hanafy et al., 2010; Hao et al., 2014; Kwon et al., 2015; Yu et al., 2018) and the results in **Table 3** all indicate that LCN2-related signaling pathways play an important part in the pathogenesis SAH, we propose that LCN2 could promote or alleviate SAH-related diseases, and could also be used to treat SAH in the future.

To answer the second question, we used mathematical algorithms to explore five potential gene biomarkers (Tk1, Cyr61, Olig1, Slc6a9, and Pcolce2), considering the impact of both SAH and LCN2 treatment at different time points, and also used the manually reviewed experimental evidence to demonstrate that Cyr61 (Yu et al., 2018), Olig1 (Sabo et al., 2017), and Slc6a9 (Huang et al., 2016) were closely related to SAH. Although Tk1 and Pcolce2 have not been reported to be associated with SAH, we will investigate their connections in future work.

Regarding the third question, although this study represents significant progress in SAH prediction, it had several drawbacks. For example, the SAH intervention experiment sample size was too small for us to demonstrate high predictive accuracy for the model. In future work, we will integrate more recent bioinformatics research algorithms (Zhang et al., 2016, 2017a, 2018, 2019a,d; Gao et al., 2017; Zhang and Zhang, 2017) and data into the system to overcome the problems.

In summary, this study analyzed the impact of LCN2 on SAH and explored the key biomarkers of SAH under LCN2 treatment at different time points. An ensemble learning model was developed to predict SAH occurrence. The results demonstrate that LCN2 (Warszawska et al., 2013) can effectively intervene in or treat SAH from a cell signaling pathway perspective. Also, three key genes were identified and validated by manual review of the experimental evidence (Huang et al., 2016; Sabo et al., 2017; Yu et al., 2018). Finally, the results showed that the ensemble learning model performed better for early SAH prediction than the classical LR, SVM, and Naive-Bayes models.

DATA AVAILABILITY STATEMENT

The raw data supporting the results of this article can be found in ArrayExpress (accession ID: E-MTAB-8407) and BioProject (accession ID: PRJNA575372).

ETHICS STATEMENT

The animal study was reviewed and approved by the Ethics Committee of Southwest Hospital.

AUTHOR CONTRIBUTIONS

LZ and YC conceived the study and developed the model. HZe and WL performed the simulations for the model. WL and HZe wrote the manuscript. MX and HZh performed the analysis for the model. HF, XR, and QL contributed to acquisition of data. All authors read and approved the final manuscript.

FUNDING

This work has been supported in part by the National Science and Technology Major Innovation Program (No. 2018ZX10201002) and supported by the National Natural Science Foundation of China (No. 61372138), State Key Laboratory of Trauma, Burn and Combined Injury (No. SKLRCJF01), and Chongqing Talent Program (No. 4139Z2391).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00391/full#supplementary-material

REFERENCES

Ai, C., and Kong, L. (2018). CGPS: a machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. *J. Genet. Genomics* 45, 489–504. doi: 10.1016/j.jgg.2018.08.002

Armstrong, C. W., Bosio, E., Neil, C., Brown, S. G., Hankey, G. J., and Fatovich, D. M. (2017). Distinct inflammatory responses differentiate cerebral infarct from transient ischaemic attack. *J. Clin. Neurosci.* 35, 97–103. doi: 10.1016/j.jocn.2016.09.011

Carlin, B. P., and Louis, T. A. (2010). *Bayes and Empirical Bayes Methods for Data Analysis*. New York, NY: Chapman and Hall/CRC.

Chaitanya, G., Schwaninger, M., Alexander, J., and Babu, P. P. (2010). Granzyme-b is involved in mediating post-ischemic neuronal death during focal cerebral ischemia in rat model. *Neuroscience* 165, 1203–1216. doi: 10.1016/j.neuroscience.2009.10.067

Chu, S., Feng, D., Ma, Y., Zhang, H., Zhu, Z. A., Li, Z., et al. (2011). Expression of HGF and VEGF in the cerebral tissue of adult rats with chronic hydrocephalus after subarachnoid hemorrhage. *Mol. Med. Rep.* 4, 785–791. doi: 10.3892/mmr.2011.500

Dietterich, T. G. (2000). "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems* (Berlin, Heidelberg: Springer), 1–15.

Duan, K-B., Rajapakse, J. C., Wang, H., and Azuaje, F. (2005). Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Trans. Nanobiosci.* 4, 228–234. doi: 10.1109/TNB.2005.853657

Edwards, J. W., Page, G. P., Gadbury, G., Heo, M., Kayo, T., Weindruch, R., et al. (2005). Empirical bayes estimation of gene-specific effects in micro-array research. *Funct. Integr. Genomics* 5, 32–39. doi: 10.1007/s10142-004-0123-0

Egashira, Y., Hua, Y., Keep, R. F., and Xi, G. (2014). Acute white matter injury after experimental subarachnoid hemorrhage: potential role of lipocalin 2. *Stroke* 45, 2141–2143. doi: 10.1161/STROKEAHA.114.005307

Ferreira, A. C., Mesquita, S. D., Sousa, J. C., Correianeves, M., Sousa, N., Palha, J. A., et al. (2015). From the periphery to the brain: Lipocalin-2, a friend or foe? *Progr. Neurobiol.* 131, 120–136. doi: 10.1016/j.pneurobio.2015.06.005

Flo, T. H., Smith, K. D., Sato, S., Rodriguez, D., Holmes, M. A., Strong, R. K., et al. (2004). Lipocalin 2 mediates an innate immune response to bacterial infection by sequestrating iron. *Nature* 432, 917–921. doi: 10.1038/nature03104

Frontera, J. A., Claassen, J., Schmidt, J. M., Wartenberg, K. E., Temes, R., Connolly, E. S., et al. (2006). Prediction of symptomatic vasospasm after

subarachnoid hemorrhage: the modified fisher scale. *Neurosurgery* 59, 21–27. doi: 10.1227/01.NEU.0000218821.34014.1B

Gao, H., Yin, Z., Cao, Z., and Zhang, L. (2017). Developing an agent-based drug model to investigate the synergistic effects of drug combinations. *Molecules* 22:2209. doi: 10.3390/molecules22122209

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). Affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Guo, D., Wilkinson, D. A., Thompson, B. G., Pandey, A. S., Keep, R. F., Xi, G., et al. (2017). MRI characterization in the acute phase of experimental subarachnoid hemorrhage. *Transl. Stroke Res.* 8, 234–243. doi: 10.1007/s12975-016-0511-5

Hanafy, K. A., Grobelny, B., Fernandez, L., Kurtz, P., Connolly, E., Mayer, S. A., et al. (2010). Brain interstitial fluid TNF-α after subarachnoid hemorrhage. *J. Neurol. Sci.* 291, 69–73. doi: 10.1016/j.jns.2009.12.023

Hao, X-K., Wu, W., Wang, C-X., Xie, G-B., Li, T., Wu, H-M., et al. (2014). Ghrelin alleviates early brain injury after subarachnoid hemorrhage via the PI3K/Akt signaling pathway. *Brain Res.* 1587, 15–22. doi: 10.1016/j.brainres.2014.08.069

Hosmer, D. W. Jr., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons. doi: 10.1002/9781118548387

Huang, B., Xie, Q., Lu, X., Qian, T., Li, S., Zhu, R., et al. (2016). GlyT1 inhibitor NFPS exerts neuroprotection via GlyR alpha1 subunit in the rat model of transient focal cerebral ischaemia and reperfusion. *Cell. Physiol. Biochem.* 38, 1952–1962. doi: 10.1159/000445556

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Kao, S-C., Wu, H., Xie, J., Chang, C-P., Ranish, J. A., Graef, I. A., et al. (2009). Calcineurin/NFAT signaling is required for neuregulin-regulated schwann cell differentiation. *Science* 323, 651–654. doi: 10.1126/science.1166562

Kwon, M., Woo, S., Kurland, D., Yoon, S., Palmer, A., Banerjee, U., et al. (2015). Methemoglobin is an endogenous toll-like receptor 4 ligand—relevance to subarachnoid hemorrhage. *Int. J. Mol. Sci.* 16, 5028–5046. doi: 10.3390/ijms16035028

Lee, T. I., and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* 152, 1237–1251. doi: 10.1016/j.cell.2013.02.014

Li, T., Cheng, Z., and Zhang, L. (2017). Developing a novel parameter estimation method for agent-based model in immune system simulation under the framework of history matching: a case study on influenza A virus infection. *Int. J. Mol. Sci.* 18:E2592. doi: 10.3390/ijms18122592

Liu, G-D., Li, Y-C., Zhang, W., and Zhang, L. (2019). A brief review of artificial intelligence applications and algorithms for psychiatric disorders. *Engineering*. doi: 10.1016/j.eng.2019.06.008. [Epub ahead of print].

Macdonald, R. L. (2014). Delayed neurological deterioration after subarachnoid haemorrhage. *Nat. Rev. Neurol.* 10:44. doi: 10.1038/nrneurol.2013.246

Macdonald, R. L., Kassell, N. F., Mayer, S., Ruefenacht, D., Schmiedek, P., Weidauer, S., et al. (2008). Clazosentan to overcome neurological ischemia and infarction occurring after subarachnoid hemorrhage (CONSCIOUS-1) randomized, double-blind, placebo-controlled phase 2 dose-finding trial. *Stroke* 39, 3015–3021. doi: 10.1161/STROKEAHA.108.519942

Majdalawieh, A., Zhang, L., and Ro, H. S. (2007). Adipocyte enhancer-binding protein-1 promotes macrophage inflammatory responsiveness by up-regulating NF-kappaB via IkappaBalpha negative regulation. *Mol. Biol. Cell* 18, 930–942. doi: 10.1091/mbc.e06-03-0217

Marta, R., and Marc, R. R. (2014). IQRray, a new method for affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics* 32:2565. doi: 10.1093/bioinformatics/btw374

Ni, W., Gu, Y., Song, D. L., Leng, B., Li, P., and Mao, Y. (2011). The relationship between IL-6 in CSF and occurrence of vasospasm after subarachnoid hemorrhage. *Acta Neurochir. Suppl.* 110(Pt. 1), 203–208. doi: 10.1007/978-3-7091-0353-1_35

Olabisi, O. A., Soto-Nieves, N., Nieves, E., Yang, T. T., Yang, X., Raymond, Y., et al. (2008). Regulation of transcription factor NFAT by ADP-ribosylation. *Mol. Cell. Biol.* 28, 2860–2871. doi: 10.1128/MCB.01746-07

Osuka, K., Watanabe, Y., Yamauchi, K., Nakazawa, A., Usuda, N., Tokuda, M., et al. (2006). Activation of the JAK-STAT signaling pathway in the rat basilar artery after subarachnoid hemorrhage. *Brain Res.* 1072, 1–7. doi: 10.1016/j.brainres.2005.12.003

Ramos, L. A., Van Der Steen, W. E., Barros, R. S., Majoie, C. B., Van Den Berg, R., Verbaan, D., et al. (2019). Machine learning improves prediction of delayed cerebral ischemia in patients with subarachnoid hemorrhage. *J. Neurointerv. Surg.* 11, 497–502. doi: 10.1136/neurintsurg-2018-014258

Roederer, A., Holmes, J. H., Smith, M. J., Lee, I., and Park, S. (2014). Prediction of significant vasospasm in aneurysmal subarachnoid hemorrhage using automated data. *Neurocrit. Care* 21, 444–450. doi: 10.1007/s12028-014-9976-9

Sabo, J. K., Heine, V., Silbereis, J. C., Schirmer, L., Levison, S. W., and Rowitch, D. H. (2017). Olig1 is required for noggin-induced neonatal myelin repair. *Ann. Neurol.* 81, 560–571. doi: 10.1002/ana.24907

Smyth, G. K., Ritchie, M., Thorne, N., and Wettenhall, J. (2005). LIMMA: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (New York, NY: Springer), 397–420.

Sozen, T., Tsuchiyama, R., Hasegawa, Y., Suzuki, H., Jadhav, V., Nishizawa, S., et al. (2011). Immunological response in early brain injury after SAH. *Acta Neurochir. Suppl.* 110(Pt 1), 57–61. doi: 10.1007/978-3-7091-0353-1_10

Suarez, J. I., Tarr, R. W., and Selman, W. R. (2006). Aneurysmal subarachnoid hemorrhage. *N. Engl. J. Med.* 354, 387–396. doi: 10.1056/NEJMra052732

Suykens, J. A., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Varet, H., Brillet-Guéguen, L., Coppée, J-Y., and Dillies, M-A. (2016). SARTools: a DESeq2-and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PLoS ONE* 11:e0157022. doi: 10.1371/journal.pone.0157022

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi: 10.1128/AEM.00062-07

Warszawska, J., Gawish, R., Sharif, O., Sigel, S., Doninger, B., Lakovits, K., et al. (2013). Lipocalin 2 deactivates macrophages and worsens pneumococcal pneumonia outcomes. *J. Clin. Invest.* 123, 3363–3372. doi: 10.1172/JCI67911

Wu, J., Mao, X., Cai, T., Luo, J., and Wei, L. (2006). KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.* 34, 720–724. doi: 10.1093/nar/gkl167

Wu, W., and Zhang, L. (2020). Exploring the dynamics and interplay of human papillomavirus and cervical tumorigenesis by integrating biological data into a mathematical model. *BMC Bioinformatics*.

Xia, Y., Yang, C., Hu, N., Yang, Z., He, X., Li, T., et al. (2017). Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model. *BMC Genomics* 18:950. doi: 10.1186/s12864-016-3256-3

Xiao, M., Yang, X., Yu, J., and Zhang, L. (2019). CGIDLA: Developing the Web Server for CpG Island related Density and LAUPs (Lineage-associated Underrepresented Permutations) Study. *IEEE/ACM Transac. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2019.2935971. [Epub ahead of print].

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* 39, W316–W322. doi: 10.1093/nar/gkr483

Yu, S., Zeng, Y. J., and Sun, X. C. (2018). Neuroprotective effects of p53/microRNA-22 regulate inflammation and apoptosis in subarachnoid hemorrhage. *Int. J. Mol. Med.* 41, 2406–2412. doi: 10.3892/ijmm.2018.3392

Zhang, L., Bai, W., Yuan, N., and Du, Z. (2019a). Comprehensively benchmarking applications for detecting copy number variation. *PLoS Comput. Biol.* 15:e1007069. doi: 10.1371/journal.pcbi.1007069

Zhang, L., Li, J., Yin, K., Jiang, Z., Li, T., Hu, R., et al. (2019b). Computed tomography angiography-based analysis of high-risk intracerebral haemorrhage patients by employing a mathematical model. *BMC Bioinformatics* 20:193. doi: 10.1186/s12859-019-2741-5

Zhang, L., Liu, G., Kong, M., Li, T., Wu, D., Zhou, X., et al. (2019d). Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. *Bioinformatics*. 26:btz542. doi: 10.1093/bioinformatics/btz542

Zhang, L., Liu, Y., Wang, M., Wu, Z., Li, N., Zhang, J., et al. (2017a). EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. *J. Mol. Cell* 9, 477–488. doi: 10.1093/jmcb/mjx056

Zhang, L., Qiao, M., Gao, H., Hu, B., Tan, H., Zhou, X., et al. (2016). Investigation of mechanism of bone regeneration in a porous biodegradable calcium

phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. *Nanoscale* 8, 14877–14887. doi: 10.1039/C6NR01637E

Zhang, L., Xiao, M., Zhou, J., and Yu, J. (2018). Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a jellyfish-based LAUPs analysis application (JBLA). *Bioinformatics* 34, 3624–3630. doi: 10.1093/bioinformatics/bty392

Zhang, L., and Zhang, S. (2017). Using game theory to investigate the epigenetic control mechanisms of embryo development: comment on: "Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition" by Qian Wang et al. *Phys. Life Rev.* 20, 140–142. doi: 10.1016/j.plrev.2017.01.007

Zhang, L., Zheng, C. Q., Li, T., Xing, L., Zeng, H., Li, T. T., et al. (2017b). Building up a robust risk mathematical platform to predict colorectal cancer. *Complexity* 2017:8917258. doi: 10.1155/2017/8917258

Zhang, L., Li, P., Zhao, J., Yang, X., Li, Z., and Yu, J. (2019c). *Research Progress on the Three-Dimensional Structure of Genome*. SCIENTIA SINICA Vitae.

Zhang, L., Zichun, D., Jun, Y., and Ming, X. (2020). *CpG-island-based annotation and analysis of human housekeeping genes. Brief. Bioinformatics*. bbz134. doi: 10.1093/bib/bbz134

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comp. Graph. Stat.* 15, 265–286. doi: 10.1198/106186006X113430

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

METHODS
published: 28 April 2020
doi: 10.3389/fgene.2020.00407

An Adaptive Sparse Subspace Clustering for Cell Type Identification

Ruiqing Zheng[1], *Zhenlan Liang*[1], *Xiang Chen*[1], *Yu Tian*[1], *Chen Cao*[2] *and Min Li*[1]*

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] Departments of Biochemistry & Molecular Biology and Medical Genetics, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB, Canada

The rapid development of single-cell transcriptome sequencing technology has provided us with a cell-level perspective to study biological problems. Identification of cell types is one of the fundamental issues in computational analysis of single-cell data. Due to the large amount of noise from single-cell technologies and high dimension of expression profiles, traditional clustering methods are not so applicable to solve it. To address the problem, we have designed an adaptive sparse subspace clustering method, called AdaptiveSSC, to identify cell types. AdaptiveSSC is based on the assumption that the expression of cells with the same type lies in the same subspace; one cell can be expressed as a linear combination of the other cells. Moreover, it uses a data-driven adaptive sparse constraint to construct the similarity matrix. The comparison results of 10 scRNA-seq datasets show that AdaptiveSSC outperforms original subspace clustering and other state-of-art methods in most cases. Moreover, the learned similarity matrix can also be integrated with a modified t-SNE to obtain an improved visualization result.

Keywords: single cell RNA-seq, subspace clustering, adaptive sparse strategy, similarity learning, visualization

OPEN ACCESS

Edited by:
Hongmin Cai,
South China University of Technology, China

Reviewed by:
Chunhou Zheng,
Anhui University, China
Tao Zeng,
Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, China

***Correspondence:**
Min Li
limin@mail.csu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *23 October 2019*
Accepted: *31 March 2020*
Published: *28 April 2020*

Citation:
Zheng R, Liang Z, Chen X, Tian Y, Cao C and Li M (2020) An Adaptive Sparse Subspace Clustering for Cell Type Identification. Front. Genet. 11:407. doi: 10.3389/fgene.2020.00407

1. INTRODUCTION

Cells are the basic functional unit all organisms are made of and play significant roles in the different stages of life. Through various DNA and RNA sequencing data, researchers have a comprehensive and deep understanding of cell biology. However, traditional sequencing data is obtained from bulks of cells, and these are composed of the mixed effect of numerous cells and ignore cell heterogeneity. These bulk-seq data will lead to deviations in downstream analysis if a specific type of cell is expected. Recently, single-cell sequencing techniques have developed rapidly and make up the defect of bulk sequencing data. Although the single-cell sequencing technique cannot capture all cell information, it provides a great opportunity to reveal the characteristics of an individual cell.

The fundamental step of analyzing the single-cell data is to identify the cell types. Utilizing single-cell RNA-seq (scRNA-seq) data to obtain the cell clusters is one of the most efficient methods available. The amount of clustering methods on the basis of scRNA-seq data have been proposed. A group of methods are focused on calculating more accurate and robust similarity scores between cells. SNN-cliq (Xu and Su, 2015) constructed the distance matrix and counted the number of common neighbor cells for each pair of cells as the similarity scores and then incorporated these within a clique-based clustering method. Seurat (V3.0) was inspired by an SNN-cliq and applied the SNN graph with a louvain algorithm (Butler et al., 2018; Stuart et al., 2019). Seurat is one of the most widely used methods. SIMLR (Wang et al., 2017) and SC3 (Kiselev et al., 2017) adopted multiple similarity metrics from different aspects. In SIMILR, we could learn the inherent similarity matrix from a different resolution of Gaussian kernels, while SC3 combined multiple sub-clustering results

together to build up a consensus matrix. Random forest (Pouyan and Kostka, 2018) was another way to calculate the similarity. The correlation coefficient has been proven to be effective when estimating the pairwise similarity of cells, and a high-order correlation coefficient was also applied in the scRNA-seq data analysis (Jiang et al., 2018; Tang et al., 2019). Compared to the methods based on pair-wise distance or correlation measurement, SinNLRR (Zheng et al., 2019b) considered the subspace characteristics of cells' expression and assumed the low rank and non-negative properties of the similarity matrix. Besides, several methods, including nonnegative matrix factorization (NMF) (Shao and Höfer, 2017; Zhu et al., 2017), imputation, and dimensionality reduction-based methods (Yau et al., 2016; Lin et al., 2017), have been used widely in assessing cellular heterogeneity. In the other aspect, the increasing number of well-learned scRNA-seq datasets also drives the appearance of supervised methods. These methods depended on labeled training datasets or some prior biological knowledge, such as gene markers (Wagner and Yanai, 2018; Pliner et al., 2019). According to the latest study (Abdelaal et al., 2019), most of the supervised methods are sensitive to prior knowledge, dataset complexity, or input features. Moreover, this kind of method has a fixed resolution and cannot find the detailed subtypes from a rough cell group. In this study, we have focused on the unsupervised clustering methods to identify the cell types. Inspired by previous methods, calculating the distance or similarity matrix of cells is a critical step. To recognize more accurate similarities of cells from high dimensional expression profiles, we have proposed an adaptive sparse subspace clustering method called AdaptiveSSC. AdaptiveSSC follows the subspace assumption and remains the nearest neighbors of a cell by a data-driven adaptive sparse constraint. The derived similarity matrix is used to obtain the clustering result and visualization. AdaptiveSSC obtains an improved performance on multiple experimental datasets.

2. MATERIALS AND METHODS

The pipeline of AdaptiveSSC is shown in **Figure 1**. Taking the scRNA-seq expression matrix as the input, AdaptiveSSC constructs the sparse cell-to-cell similarity matrix by keeping the most similar cells for each cell before then applying it to spectral clustering and modified t-distributed stochastic neighbor embedding (t-SNE) to obtain cell groups and the visualization result.

2.1. Data Pre-processing

The quantified scRNA-seq data contain thousands of genes, and the sparsity of gene expression is usually high. Therefore, AdaptiveSSC filters the genes expressed in <10% of the cells (the maximum number is 100), which are not regarded as informative genes. AdaptiveSSC investigates the linear effect of other cells on the target cell. To remove the scale of cells' expression, the L_2 normalization is carried on the original gene expression matrix.

$$X_{ij} = G_{ij} / \sqrt{\sum_{k=1}^{M} G_{kj}^2} \tag{1}$$

where G is the original expression matrix with M genes and N cells. The normalized matrix X is used in the following calculation.

2.2. Adaptive Sparse Subspace Clustering

Most clustering methods depend on the calculation of the similarity or distance matrix. The most popular similarity measurements include Euclidean distance, Pearson or Spearman correlations, and cosine similarity, which are all based on a pairwise estimation. The scRNA-seq data usually contains thousands of genes; however, only a part of a gene determines the cell type, which corresponds to a low-dimensional manifold surface. According to the common strategy in manifold learning, only the local measurement of similarity or distance is reliable, so previous scRNA-seq clustering methods (Xu and Su, 2015; Wang et al., 2017) usually apply k-nearest neighbors (KNN) to keep the locality. However, the KNN is used arbitrary to select the same number of neighbors for each cell, and the selection of k would have a great influence on the final result in some situations. In order to overcome these shortcomings, we propose an adaptive sparse subspace clustering method, which we have called AdaptiveSSC.

AdaptiveSSC is developed from sparse subspace clustering (SSC) methods. SSC is proposed to solve the motion segmentation and face clustering problems (Elhamifar and Vidal, 2013). SSC assumes that the feature vector of a sample can be expressed as the linear combination of other samples in the same subspace or type. Based on the assumption, the expression of a cell $X_i = c_1X_1 + c_2X_2 + \cdots + c_{i-1}X_{i-1} + c_{i+1}X_{i+1} + \cdots + c_NX_N$ and c_k is the subspace coefficient denoting the similarity score between cells. If the cell i and k are the same type, $c_k > 0$, otherwise it is 0. By adding l_1 term, the most similar cells lying in the same subspace are retained. Extending it to all cells, the calculation of the subspace coefficient matrix is defined as Equation (2):

$$min|C|_1 \qquad s.t, \qquad X = XC \quad \text{and} \quad diag(C) = 0 \tag{2}$$

where X is the normalized expression matrix. C is the coefficient matrix and C_{ij} denotes similarity between cell i and j. $|\cdot|_1$ denotes l_1 norm. The larger values in C mean the more similar cells. The relaxation formula of the optimization problem is shown:

$$min\frac{1}{2}||X - XC||_F^2 + \lambda|J|_1 \qquad s.t, \quad diag(C) = 0 \quad \text{and} \quad C - J = 0 \tag{3}$$

where $||\cdot||_F^2$ means the Fresenius norm and λ is the l_1 penalty factor, which controls the sparsity of the coefficient matrix. J is an auxiliary matrix.

In the Equation (3), the coefficient matrix C is sensitive to the selection of the l_1 penalty factor. Another problem is that the same penalty factor for all coefficients will lead to the loss

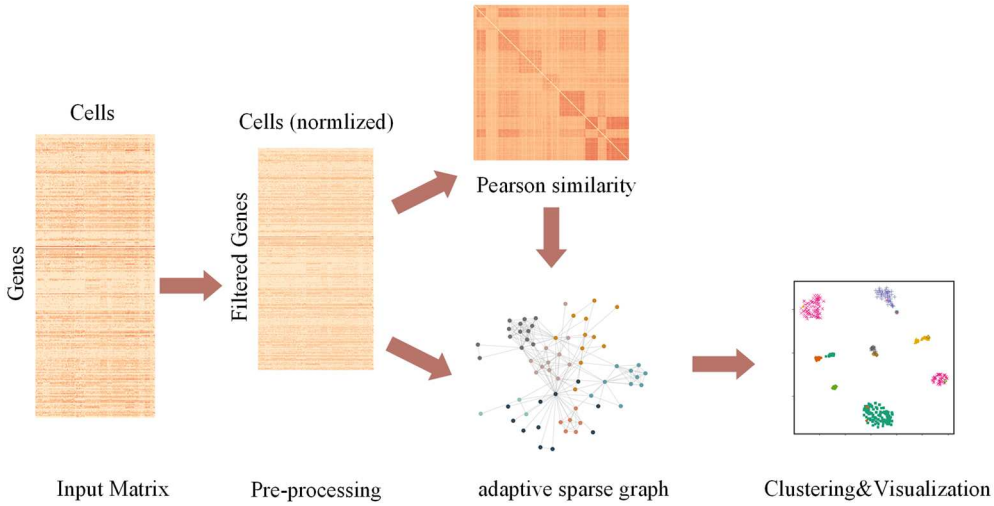


FIGURE 1 | The pipeline of AdaptiveSSC to identify and visualize cell types from scRNA-seq data.

of consistency between estimation and variable selection (Zou, 2006). Therefore, we have introduced a data-driven adaptive strategy to solve these problems. As a Pearson correlation has been proven to be effective when measuring the similarity in previous studies (Kiselev et al., 2017; Wang et al., 2017), we utilized it to adjust the penalty factor for each coefficient. If the correlation of two cells is high, the penalty factor is decreased and vice versa. The modified optimization problem is therefore defined:

$$min\frac{1}{2}||X-XC||_F^2+\lambda|\frac{J}{W}|_1\ s.t,\quad diag(C)=0\quad \text{and}\quad C-J=0$$
$$\text{where, } W_{ij}=\begin{cases} pearson(X_i,X_j) & \text{If } pearson(X_i,X_j)>0 \\ 0 & \text{otherwise}\end{cases} \tag{4}$$

where $\frac{J}{W}$ means element division of matrix J and W. We set the negative value of the Pearson correlation to 0. Because only the trend of the expression of two cells are positively correlated, we regard them as similar cells. Some zero values in W would lead to zero values in J during the optimization.

Alternating direction method of multipliers (ADMM) (Boyd et al., 2011) is an efficient method to solve Equation (4). According to ADMM, the augmented Lagrangian formula is defined:

$$\iota_{\gamma,\lambda}(C,J,Y)=\frac{1}{2}||X-XC||_F^2+\lambda|\frac{J}{W}|_1+tr(Y^T(C-J))+\frac{1}{2\gamma}||C-J||^2\quad \text{and}\quad diag(C)=0 \tag{5}$$

where Y is a dual variable, γ is an augmented Lagrangian penalty parameter, and tr means the trace of the matrix. ADMM updates C, Y, or J by fixing others. In iteration $k+1$, the optimized form of C^{k+1}, J^{k+1}, and Y^{k+1} is shown in Equations (6–8):

$$C^{k+1}=\left(X^TX+\frac{1}{\gamma}I\right)^{-1}\left(X^TX+\frac{1}{\gamma}\left(J^k-Y^k\right)\right)$$
$$C^{k+1}=C^{k+1}-diag\left(C^{k+1}\right) \tag{6}$$

$$J^{k+1}=threshold_{\frac{\lambda}{W},\gamma}\left(C^{k+1}+Y^k\right)$$
$$=sign\left(C^{k+1}+Y^k\right)\cdot max\left(|C^{k+1}+Y^k|-\frac{\lambda}{\gamma W},0\right) \tag{7}$$
$$J^{k+1}=J^{k+1}-diag\left(J^{k+1}\right)$$

$$Y^{k+1}=Y^k+\frac{1}{\gamma}\left(C^{k+1}-J^{k+1}\right) \tag{8}$$

where $sign()$ means the sign function. The convergence of ADMM mainly includes primal residuals and dual residuals. On the basis of updating process, the penalty parameter γ affects the speed of convergence. In AdaptiveSSC, we apply a balance strategy (Boyd et al., 2011) between primal residuals and dual residuals to adjust γ. The setting of γ is shown:

$$\gamma_{k+1}=\begin{cases}\gamma_k/2, & \text{when } ||r^k||_2>\mu||s^k||_2, \\ 2\gamma_k, & \text{when } ||s^k||_2>\mu||r^k||_2, \\ \gamma_k, & \text{others.}\end{cases} \tag{9}$$

where $r^k=C^k-J^k$ is the primal residual and $s^k=\frac{1}{\gamma}\left(J^k-J^{k-1}\right)$ is the dual residual. The μ is set to 50 as default. To reduce the computational complexity, γ is updated by 10 iterations. When $max(abs(C-J))<0.0001$ or the number of iteration is larger

TABLE 1 | Single cell RNA-seq datasets.

Datasets	Cell number	Gene number	Techniques
Darmanis (Darmanis et al., 2015)	420	22,085	SMARTer
Kolod (Kolodziejczyk et al., 2015)	704	10,685	Smart-Seq2
Treutlein (Treutlein et al., 2014)	80	959	SMARTer
Yan (Yan et al., 2013)	90	20,214	Tang et al., 2011
Ting (Ting et al., 2014)	114	14,405	Single CTC RNA-Seq
Engel (Engel et al., 2016)	203	23,337	Smart-seq2
Kumar (Kumar et al., 2014)	361	11,497	SMARTer
Vento (Vento-Tormo et al., 2018)	5,418	33,693	Smart-seq2
Baron (Baron et al., 2016)	8,569	20,125	inDrop
Shekhar (Shekhar et al., 2016)	26,830	13,166	Drop-seq

than 200, this update process is finished. To keep the symmetry of the similarity matrix, the final similarity matrix $S = C^T + C$.

Finally, the spectral clustering (SC) (Von Luxburg, 2007) is applied on the learned similarity matrix. The SC is based on the point of graph cut and utilizes the characteristic of the corresponding Laplacian matrix to divide the graph into several clusters. In AdaptiveSSC, we use the normalized Laplacian matrix $L^{norm} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, where D is the degree matrix, to obtain its k eigenvectors corresponding to the smallest k eigenvalues. Then, k-means is used to obtain the final clusters.

3. RESULTS AND DISCUSSION

3.1. scRNA-seq Datasets

We collected 10 scRNA-seq datasets to evaluate the performance of AdaptiveSSC. These datasets are based on different single-cell techniques or protocols, such as Smart-seq, SMARTer, and Drop-seq based methods. Meanwhile, the scale of these datasets ranges from the tens to the tens of thousands. The variety of the datasets could indicate the generalization ability of AdaptiveSSC comprehensively. The details of these datasets are shown in **Table 1**. All datasets contain the real cell types from the original researches.

3.2. Evaluation Metrics

In order to compare the performance of different clustering methods, we selected two popular metrics: normalized mutual information (NMI) and adjusted rand index (ARI). Both NMI and ARI can quantify the consistency between the clustering results and the real labels. The definition of NMI and ARI is shown:

$$NMI\,(T,P) = \frac{I(T,P)}{\left[H(T) + H(P)\right]} \tag{10}$$

$$ARI\,(T,P) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right]/\binom{n}{2}} \tag{11}$$

Where T and P mean the real labels and clustering labels, respectively. In Equation (11), n_{ij} denotes the number of cells belonging to i group in real labels and j group in clustering labels; n_i denotes the number of cells belonging to the i group in real labels, while n_j denotes the number of cells belonging to the j group in clustering labels.

3.3. Parameter Analysis

Although the adaptive strategy is used in AdaptiveSSC, there are still some hyperparameters to be set. The most important hyperparameter is the l_1 penalty factor λ. By the adaptive adjustment, the learned similarity matrix is not so sensitive to it. We evaluated the NMI and ARI of AdaptiveSSC on eight small datasets (smaller than 5,000 cells) with λ ranging from 0.01 to 0.19 and the interval set to 0.02. The results for eight small datasets are shown in **Figure 2**. Based on the result, when the λ was in the 0.01–0.05, both NMI and ARI were in the best range and were more stable. Therefore, we used $\lambda = 0.03$ as a default in AdaptiveSSC. During the experiment, we also found the optimal λ was not consistent for big datasets (in Baron is 0.01 and in Shekhar and Vento is 0.007). We recommend that users select the proper λ by grid searching with the following rule. If the corresponding sparsity of C is between 0.02 and 0.05, the λ should be selected. In Baron and Shekhar, we selected the corresponding λ with the sparsity of C is 0.03.

3.4. Comparison Analysis of Clustering Methods

To validate the effectiveness of AdaptiveSSC, we selected seven competitive methods, including SIMLR (Wang et al., 2017), MPSSC (Park and Zhao, 2018), SNN-cliq (Xu and Su, 2015), RAFSIL (Pouyan and Kostka, 2018), Seurat(V3.0) (Butler et al., 2018; Stuart et al., 2019), SinNLRR (Zheng et al., 2019b), and sparse subspace clustering (SSC) (Elhamifar and Vidal, 2013). All these methods are based on the construction of similarity matrix. SNN-cliq and Seurat recalculate the similarities based on their shared neighbors. SIMILR and MPSSC focus on the different resolution of Gaussian kernels, while RAFSIL applies random forest. SinNLRR is based on the subspace assumption with low rank constraint. The original SSC was selected as the baseline method. The results of NMI and ARI on 10 datasets are shown

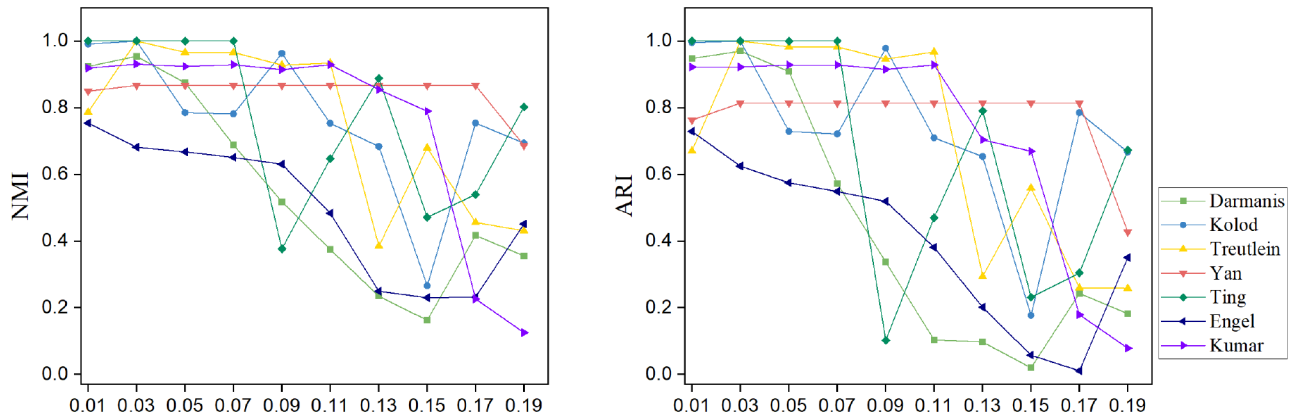


FIGURE 2 | The corresponding NMI and ARI with different values of λ on eight datasets.

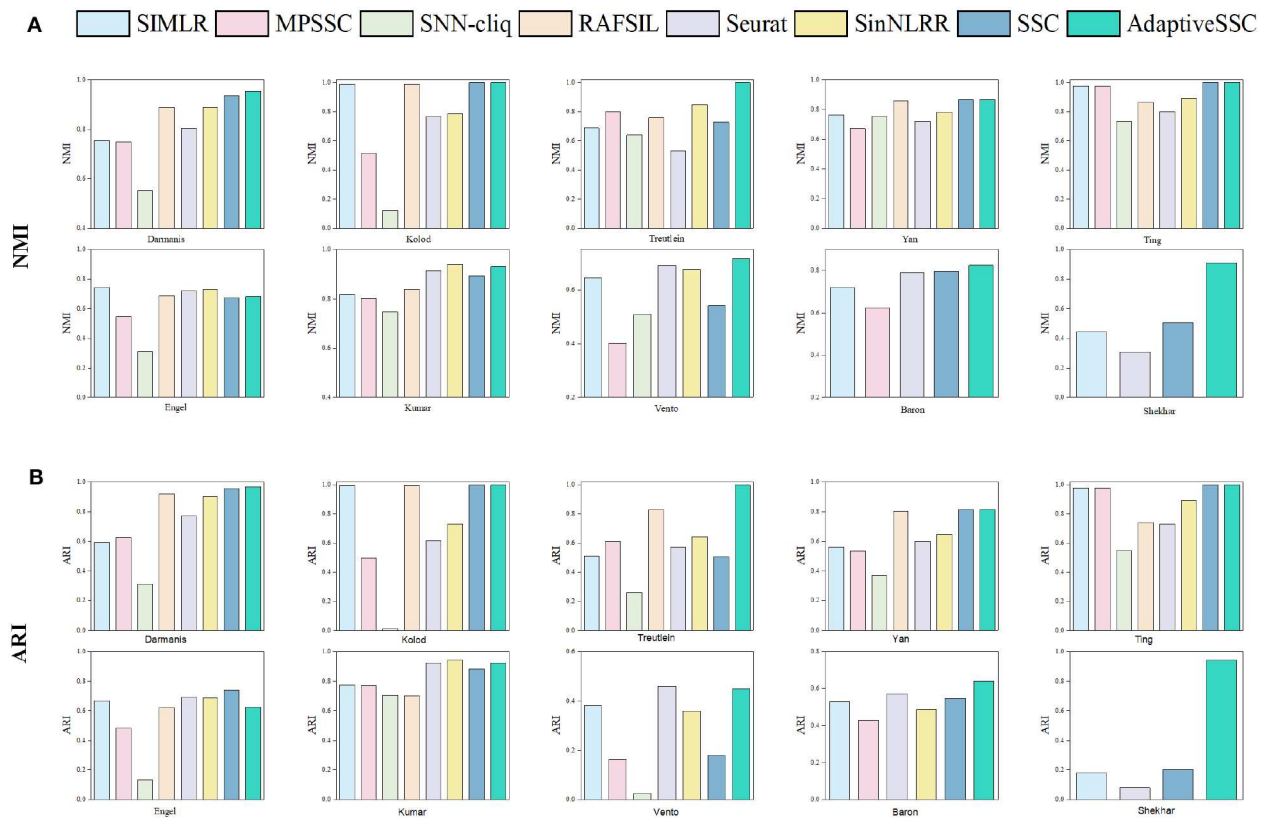


FIGURE 3 | The corresponding **(A)** NMI and **(B)** ARI of SIMLR, MPSSC, SNN-cliq, RAFSIL, Seurat, SinNLRR, SSC, and AdaptiveSSC on 10 datasets.

in **Figure 3**. Compared to SSC, AdaptiveSSC improved NMI and ARI in six datasets. Especially in Treutelin, Kumar, Vento, and Shekhar, AdaptiveSSC exhibited a significant improvement, more so than SSC, which means the adaptive penalty factor leads to the more accurate similarity matrix. In Kolod and Ting, AdaptiveSSC achieved the same performance with SSC. Overall, AdaptiveSCC exhibited a better performance than SSC in most cases. Besides, AdaptiveSSC achieved the best (or a tie for first place) performance in seven datasets upon NMI and eight datasets upon ARI compared with other six state-of-the-art methods. It is worth noting that only AdaptiveSSC obtains the perfect result on Treutelin. The results in Baron and Shekhar also verify AdaptiveSSC's effectiveness in large datasets. Estimation of the number of cell types is another important aspect in

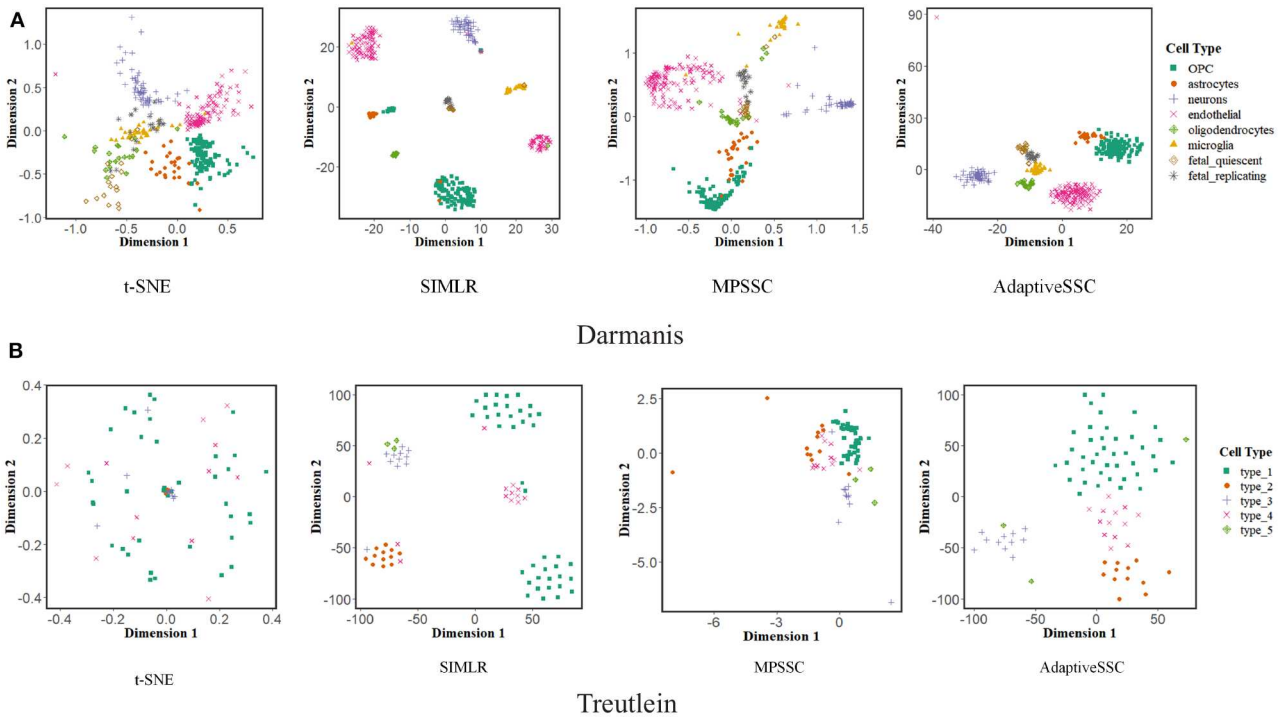


FIGURE 4 | The visualization of t-SNE, SIMLR, MPSSC, and AdaptiveSSC on **(A)** Darmanis and **(B)** Treutelin.

application. In AdaptiveSSC, we also used *eigengap* to determine the number of clusters, which was popular in previous studies. The results can be found in the **Supplementary Material**. As shown in the results, none of the methods predict the correct number of clusters in all datasets. However, AdaptiveSSC obtains the correct number of clusters in three datasets and gets the closest number in five datasets, which is a better selection overall. Moreover, we select five different scale datasets to evaluate the computational efficiency of these methods. The running time can be found in the **Supplementary Material**. AdaptiveSSC has a faster speed than SSC but is still time-consuming in large datasets compared with SIMLR and Seurat. All the experiments run on the server with 24 cores and 512 GB memory. The methods with running time more than 36 h are excluded, such as RAFSIL, SNN-cliq, and SinNLRR in large scale datasets, and MPSSC gets out of memory error on Shekhar.

3.5. Comparison Analysis of Visualization

Visualization of scRNA-seq is another important issue. Previous study (Wang et al., 2017) proposed a modified t-distributed stochastic neighbor embedding (t-SNE) to validate the performance of learned similarity. We also adopted this evaluation to AdaptiveSSC and generate 2D-embedding images on Darmanis and Treutelin with the learned similarity matrix of t-SNE, SIMLR, MPSSC, and AdaptiveSSC, respectively. The result is shown in **Figure 4**. The points with the same color mean they have the same cell type. Compared to other methods, AdaptiveSSC could group the same cells together and exhibits good silhouettes. Although SIMILR and MPSSC contain more dense parts, they divide cells with same type into different cliques, which are usually far away from each other. This will give the researchers a misconception that they are belong to exactly different types. Therefore, AdaptiveSSC has a better performance and potential in the visualization of scRNA-seq data.

3.6. Discussion and Conclusion

The identification of cell types is a fundamental problem is scRNA-seq data analysis. In recent years, a lot of clustering methods have been proposed to solve it. However, most of these methods do not exhibit a good generalization on different datasets. In this study, we proposed a subspace clustering with an adaptive sparse constraint, called AdaptiveSSC. AdaptiveSSC regards the expression of a cell can be expressed as a linear combination of other cell's expression from the same type. A data-driven adaptive sparse strategy is applied to keep the locality of cells in the original dimension and decrease the sensitivity to the penalty factor. Eight scRNA-seq datasets were used to evaluate the performance of AdaptiveSSC. By comparing with SSC, AdaptiveSSC improves the clustering results significantly in some cases, which indicates the effectiveness of our strategy. Moreover, six state-of-the-art methods were selected as comparison. From the NMI and ARI, AdaptiveSSC achieves the best performance in most of datasets. Finally, we integrated the learned similarity with modified t-SNE further, which also shows the powerful potential of AdaptiveSSC in visualization.

However, the computational efficiency of AdaptiveSSC is still low for large datasets and should be improved in the future. Some strategies used in the fast clustering method could be considered to make AdaptiveSSC more efficient (Ren et al., 2019). Moreover, AdaptiveSSC explores the cell heterogeneity from a gene level, but it is also important to study the different biological functions of cells. Regulatory modules (Aibar et al., 2017) have been proved effective when showing the functional heterogeneity of cells. It is possible to identify the cell type from the whole gene regulatory network perspective (Li et al., 2017; Zheng et al., 2018, 2019a). Besides, motivated by previous studies (Lan et al., 2018; Chen et al., 2019; Shi et al., 2019), multi-view learning and integrating with prior knowledge are promising directions to improve the accuracy of clustering and give a higher resolution of cell types.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/zrq0123/AdaptiveSSC.

AUTHOR CONTRIBUTIONS

RZ and CC designed the methodology. RZ, ZL, XC, and YT run the comparison experiments on datasets. RZ and ML wrote the paper. All authors revised and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61732009) and the Fundamental Research Funds for the Central Universities of Central South University (Nos. 2018zzts028, 2019zzts592).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00407/full#supplementary-material

REFERENCES

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M. J., et al. (2019). A comparison of automatic cell identification methods for single-cell RNA-sequencing data. *Genome Biol.* 20:194. doi: 10.1186/s13059-019-1795-z

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., et al. (2017). Scenic: single-cell regulatory network inference and clustering. *Nat. Methods* 14, 1083–1086. doi: 10.1038/nmeth.4463

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.* 3, 346–360. doi: 10.1016/j.cels.2016.08.011

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36:411. doi: 10.1038/nbt.4096

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.-P. P., et al. (2019). ILDMSF: Inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE/ACM Trans. Comput. Biol. Bioinformatics.* doi: 10.1109/TCBB.2019.2936476

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7285–7290. doi: 10.1073/pnas.1507125112

Elhamifar, E., and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2765–2781. doi: 10.1109/TPAMI.2013.57

Engel, I., Seumois, G., Chavez, L., Samaniego-Castruita, D., White, B., Chawla, A., et al. (2016). Innate-like functions of natural killer t cell subsets result from highly divergent gene programs. *Nat. Immunol.* 17, 728–739. doi: 10.1038/ni.3437

Jiang, H., Sohn, L. L., Huang, H., and Chen, L. (2018). Single cell clustering based on cell-pair differentiability correlation and variance analysis. *Bioinformatics* 34, 3684–3694. doi: 10.1093/bioinformatics/bty390

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-Seq data. *Nat. Methods* 14:483. doi: 10.1038/nmeth.4236

Kolodziejczyk, A. A., Kim, J. K., Tsang, J. C., Ilicic, T., Henriksson, J., Natarajan, K. N., et al. (2015). Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17, 471–485. doi: 10.1016/j.stem.2015.09.011

Kumar, R. M., Cahan, P., Shalek, A. K., Satija, R., DaleyKeyser, A. J., Li, H., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature* 516, 56–61. doi: 10.1038/nature13920

Lan, W., Wang, J., Li, M., Liu, J., Wu, F.-X., and Pan, Y. (2018). Predicting microRNA-disease associations based on improved microRNA and disease similarities. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15, 1774–1782. doi: 10.1109/TCBB.2016.2586190

Li, M., Zheng, R., Li, Y., Wu, F.-X., and Wang, J. (2017). MGT-SM: a method for constructing cellular signal transduction networks. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 16, 417–424. doi: 10.1109/TCBB.2017.2705143

Lin, P., Troup, M., and Ho, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* 18:59. doi: 10.1186/s13059-017-1188-0

Park, S., and Zhao, H. (2018). Spectral clustering based on learning similarity matrix. *Bioinformatics* 34, 2069–2076. doi: 10.1093/bioinformatics/bty050

Pliner, H. A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* 16, 983–986. doi: 10.1038/s41592-019-0535-3

Pouyan, M. B., and Kostka, D. (2018). Random forest based similarity learning for single cell RNA sequencing data. *Bioinformatics* 34, i79–i88. doi: 10.1093/bioinformatics/bty260

Ren, X., Zheng, L., and Zhang, Z. (2019). SSCC: a novel computational framework for rapid and accurate clustering large-scale single cell RNA-seq data. *Genomics Proteomics Bioinformatics* 17, 201–210. doi: 10.1016/j.gpb.2018.10.003

Shao, C., and Höfer, T. (2017). Robust classification of single-cell transcriptome data by nonnegative matrix factorization. *Bioinformatics* 33, 235–242. doi: 10.1093/bioinformatics/btw607

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* 166, 1308–1323. doi: 10.1016/j.cell.2016.07.054

Shi, Q., Zhang, C., Hu, B., and Zeng, T. (2019). Multi-view subspace clustering analysis for aggregating multiple heterogeneous omics data. *Front. Genet.* 10:744. doi: 10.3389/fgene.2019.00744

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. doi: 10.1016/j.cell.2019.05.031

Tang, F., Lao, K., and Surani, M. A. (2011). Development and applications of single-cell transcriptome analysis. *Nat. Methods* 8, S6–S11. doi: 10.1038/nmeth.1557

Tang, H., Zeng, T., and Chen, L. (2019). High-order correlation integration for single-cell or bulk RNA-seq data analysis. *Front. Genet.* 10:371. doi: 10.3389/fgene.2019.00371

Ting, D. T., Wittner, B. S., Ligorio, M., Jordan, N. V., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., et al. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* 509, 371–375. doi: 10.1038/nature 13173

Vento-Tormo, R., Efremova, M., Botting, R. A., Turco, M. Y., Vento-Tormo, M., Meyer, K. B., et al. (2018). Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* 563, 347–353. doi: 10.1038/s41586-018-0698-6

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. doi: 10.1007/s11222-007-9033-z

Wagner, F., and Yanai, I. (2018). Moana: A robust and scalable cell type classification framework for single-cell RNA-seq data. *bioRxiv [preprint]* 456129. doi: 10.1101/456129

Wang, B., Zhu, J., Pierson, E., Ramazzotti, D., and Batzoglou, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* 14:414. doi: 10.1038/nmeth.4207

Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. *Mol. Biol.* 20, 1131–1139. doi: 10.1038/nsmb.2660

Yau, C., and Zurauskiene, J. (2016). pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics* 17:140. doi: 10.1186/s12859-016-0984-y

Zheng, R., Li, M., Chen, X., Wu, F.-X., Pan, Y., and Wang, J. (2018). Bixgboost: a scalable, flexible boosting-based method for reconstructing gene regulatory networks. *Bioinformatics* 35, 1893–1900. doi: 10.1093/bioinformatics/bty908

Zheng, R., Li, M., Chen, X., Zhao, S., Wu, F., Pan, Y., et al. (2019a). An ensemble method to reconstruct gene regulatory networks based on multivariate adaptive regression splines. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2019.2900614

Zheng, R., Li, M., Liang, Z., Wu, F.-X., Pan, Y., and Wang, J. (2019b). SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics* 35, 3642–3650. doi: 10.1093/bioinformatics/btz139

Zhu, X., Ching, T., Pan, X., Weissman, S. M., and Garmire, L. (2017). Detecting heterogeneity in single-cell RNA-seq data by non-negative matrix factorization. *PeerJ* 5:e2888. doi: 10.7717/peerj.2888

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101, 1418–1429. doi: 10.1198/016214506000000735

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 13 May 2020
doi: 10.3389/fbioe.2020.00328

Protein Network Studies on PCOS Biomarkers With S100A8, Druggability Assessment, and RNA Aptamer Designing to Control Its Cyst Migration Effect

Subramaniyan Manibalan[1], Ayyachamy Shobana[1], Manickam Kiruthika[1], Anant Achary[1], Madasamy Swathi[1], Renganathan Venkatalakshmi[2], Kandasamy Thirukumaran[2], K. Suhasini[2] and Sharon Roopathy[2]*

[1] Centre for Research, Kamaraj College of Engineering and Technology, Madurai, India, [2] Department of Biotechnology, Kamaraj College of Engineering and Technology, Madurai, India

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Ankush Sharma,
University of Oslo, Norway
Qing Li,
Huntsman Cancer Institute, University of Utah, United States

***Correspondence:**
Subramaniyan Manibalan
manibalanbt@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: 04 October 2019
Accepted: 25 March 2020
Published: 13 May 2020

Citation:
Manibalan S, Shobana A, Kiruthika M, Achary A, Swathi M, Venkatalakshmi R, Thirukumaran K, Suhasini K and Roopathy S (2020) Protein Network Studies on PCOS Biomarkers With S100A8, Druggability Assessment, and RNA Aptamer Designing to Control Its Cyst Migration Effect. Front. Bioeng. Biotechnol. 8:328. doi: 10.3389/fbioe.2020.00328

The prevalence of polycystic ovary syndrome (PCOS) has been gradually increasing among adult females worldwide. Laparoscopy drilling on ovary is the only available temporary solution with a high incidence of reoccurrence. S100A8 with S100A9 complex is believed to facilitate the cyst migration in PCOS condition. The high evident protein interaction network studies between PCOS biomarkers, cancer invasion markers, and the interactors of S100A8 confirm that this protein has strong interaction with other selective PCOS biomarkers, which may be associative in the immature cyst invasion process. Through the network studies, intensive structural and pathway analysis, S100A8 is identified as a targetable protein. In this research, the non-SELEX *in silico* method is adapted to construct RNA Library based on the consensus DNA sequence of Glucocorticoid Response Element (GRE) and screened the best nucleotide fragments which are bound within the active sites of the target protein. Selected sequences are joined as a single strand and screened the one which competitively binds with minimal energy. *In vitro* follow-up of this computational research, the designed RNA aptamer was used to infect the MCF7 cell line through Lipofectamine 2000 mediated delivery to study the anti-cell migration effect. Wound Scratch assay confirms that the synthesized 18-mer oligo has significant inhibition activity toward tumor cell migration at the cellular level.

Keywords: network analysis, druggability, RNA aptamer, lim method, pcos targets, protein network

INTRODUCTION

Nucleotide aptamers are successfully explored as better therapeutics to treat diseases and disorders. Time-consuming low-throughput procedures have been in practice to design the aptamers *in vitro* (Ghavami et al., 2009). Therefore, *in silico* non-SELEX approach is the better choice to perform the selection of aptamers, which involves the construction of an oligonucleotide library without amplification and binding them with suitable target protein unlike SELEX (Berezovski et al., 2006; Tseng et al., 2011). Designing the RNA aptamer for the validated biomarker helps us to normalize the disease state at the genetic level. Hence, delivering a well-designed aptamer against response elements (REs) can control the strange translation of the target gene. REs are the critical

elements involved in the activation of target gene regulation. Inhibiting biomarkers of specific pathophysiological conditions at the molecular level is a better choice to oversee the disease (Strimbu and Tavel, 2010). Target validation is one of the necessary procedures in drug discovery protocol. Since the exact cause of polycystic ovary syndrome (PCOS) is imprecise (Sir-Petermann et al., 2002), it is tedious to identify the best target clinically. Assay on endometrial cell migration is one of the diagnostic tools to identify the complications of this syndrome, and metformin has a proven attenuating effect on the invasion of endometrial cells of diseased women (Tan et al., 2011). Previously, researchers have found that 500 biomarkers are prevalent in PCOS (Dai and Lu, 2012). In this research work, we focused on S100A8 protein which is one of the important biomarkers in PCOS. Protein–protein interaction network (PPIN) is used to identify the associative proteins and its pathways in PCOS. Additionally, druggable properties of S100A8 were studied through pocket analysis. Besides, the aptamer library for specific RE of S100A8 was constructed by a non-SELEX fragment approach. The best aptamer sequence was screened through quality assessments, such as affinity and stability parameters.

MATERIALS AND METHODS

Network Profile of S100A8 in Polycystic Ovary Syndrome and Enrichment Analysis

Interactors of S100A8 are obtained from BioGRID$^{3.5}$, a dataset repository (Oughtred et al., 2019), and the molecular interaction network was constructed in STRING Database (Szklarczyk et al., 2019). Biomarkers specific to PCOS and cancer cell invasion are retrieved from the recent research articles (Daan et al., 2016; Lu et al., 2017; Gerashchenko et al., 2019) and are used to construct another network. Both the networks were merged to find the first shell interactors of S100A8. Cytoscape 3.7.2 is employed to merge the networks and find proteins which are associated with S100A8. Pathways of S100A8 and its clusters are identified by using ClueGO, a Cytoscape application for clustering the functional network by terms or pathways (Bindea et al., 2009). Molecular functions of Gene Ontology (GO), Reactome Pathway Database (Croft et al., 2011), and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways are used as resources for enrichment analysis. *P*-value 0.005 is set as a significance to select the clusters of S100A8.

Structure and Druggability Studies on S100A8

Druggability analysis is used to predict the receptiveness and stability of drug target. Physiochemical and geometric properties such as number of pockets, druggable score, and pocket volume determine the efficiency of the target candidate. DoGSite Scorer is used for binding site prediction analysis and druggability assessment, which is based on heavy-atom coordinates employing support vector machines (SVMs) (Volkamer et al., 2012). Pocket volume, lipophilic character, and pocket enclosures were accounted for simple score calculation to suggest the competence of targetability. Three-dimensional structure of the target was retrieved from PDB (ID: 5HLV) and used for the druggable screening.

Glucocorticoid Response Elements for S100A8

REs are the inducers of the receptor and ligand interaction which results in the expression or activation of a particular protein. Since the aptamers are crucial elements in the control of target expression so we decide to design RNA oligomer against specific PCOS targets. Glucocorticoid RE (GRE) (Hsu et al., 2005), hypoxia RE (HRE) (Rees et al., 2001), antioxidant RE (ARE) (Nioi et al., 2003), and interferon gamma (INF-γ) RE (IRE) (Yang et al., 1990) are identified as the influencing REs of the S100A8 gene. Specifically, GREs have the proficiency to inhibit S100A8 through the downregulation of leukocyte transmigration. Glucocorticoids also induce the expression of inhibition factor for macrophage migration, which ultimately downregulates the cyst inflammation. The earlier research report shows that GRE consists of two half-sites with three spacer bases; the consensus pseudo palindromic sequence of GRE is 5′ C<u>AGAACA</u>**TCA**TGTTCTGA 3′ (Weikum et al., 2017).

Nucleotide Fragment Library Construction

RNA Composer utilizes the Dot-Bracket format notation of the secondary structure sequence to model the RNA, and the 3D element of modeled RNA was chosen from RNA Frabase (Biesiada et al., 2016). RE is a sequence, which binds with the receptor and plays a crucial role in expression, so the RNA analog library of specific RE was created to mimic the inhibitory action. The consensus sequence was segregated as fragments in such a way that six nucleotides at a stretch were taken per fragment (**Figure 4A**) for analog library construction. Resulted library sequences were later utilized for binding studies with the target by RNA-Lim method and recognized the various conformations of fragments bound in the active sites of the protein (Hall et al., 2015). Fragments which bound on active sites are selected to design the high précised aptamer model. Diversity in the exhibited conformations of ssRNA–protein complexes was meticulously sampled to construct a fragment library. MC-Fold | MC-Sym pipeline was employed to obtain the secondary and tertiary structures of the constructed aptamer (Parisien and Major, 2008). The proposed mechanism for PCOS control through the aptamer binding on S100A8 is illustrated in **Figure 1**. Refinement on docking results of their chain-forming poses was done in Molecular Operating Environment (MOE) suit (Ahirwar et al., 2016). The fragment-based approach was adopted for competent docking with S100A8 (Ahirwar et al., 2016); this unusual method has numerous advantages over conventional rigid based docking.

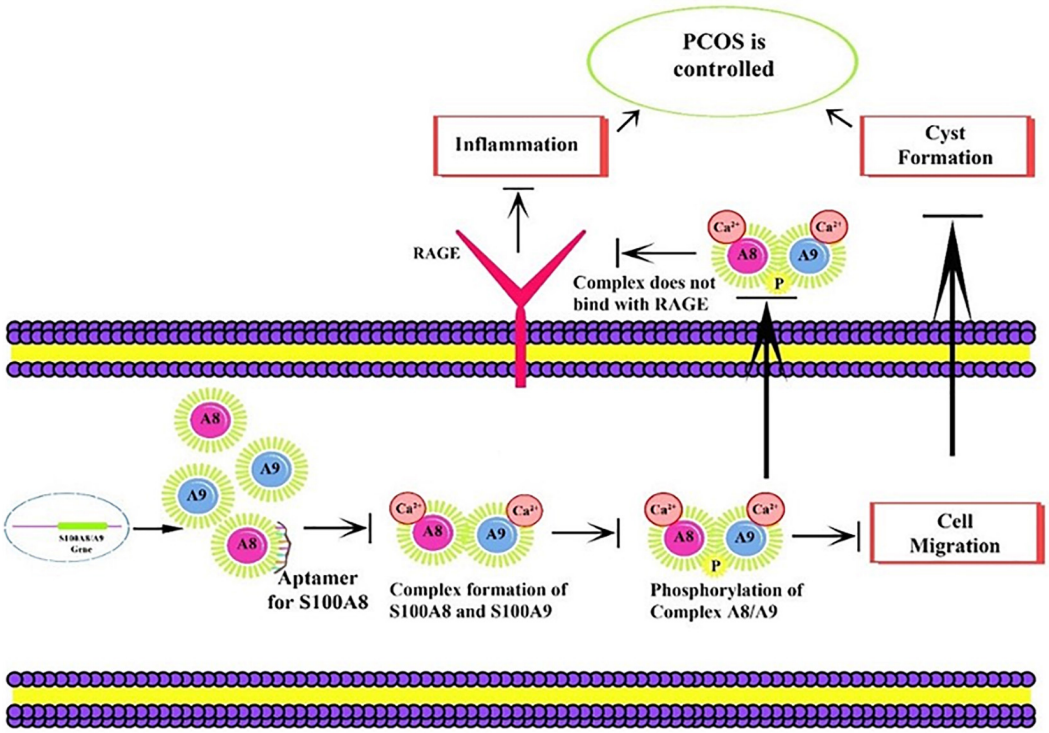


FIGURE 1 | Proposed mechanism of aptamer inhibition. Aptamer binds to the S100A8 and prevents the intracellular initial complexation with S100A9. Due to aptamer binding, extracellular receptor of advanced glycation end products (RAGE) complexed inflammatory sequences are also prevented.

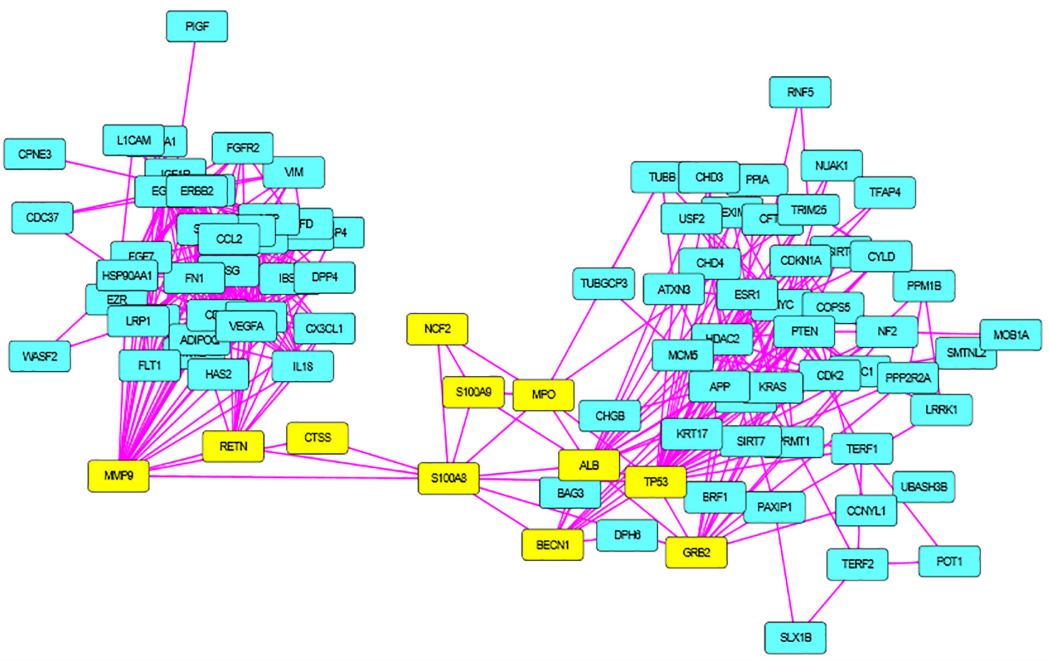


FIGURE 2 | Merged network of Polycystic Ovary Syndrome (PCOS). Combined network of S100A8 integrators with PCOS biomarkers; yellow colored nodes are the first shell interacting proteins with S100A8 that are suspected for cyst migration activity.

Affinity and Stability Studies of the Designed Aptamer

PatchDock tool is efficiently used to study the binding properties of designed aptamers with target protein (Schneidman-Duhovny et al., 2005).Based on the global binding energy, FireDock is employed to decipher the docked result by flexible refinements rather than the rigidity of protein and also it optimizes the side-chain residues, which minimizes the rigid body conformation of the interactive protein (Mashiach et al., 2008). Previously, it was reported that the stability of RNA will be analyzed by the inverted repeats which form stable hairpin loops (Ahirwar et al., 2016). Oligoanalyzer is an inclusive oligonucleotide scrutinizer employed to check the hairpin loop and stability of the designed aptamer (Owczarzy et al., 2008).

Anti-migration by Scratch Wound Healing Assay on Cell Lines

MCF-7 cells were seeded into a 24-well tissue culture plate containing antibiotic-free minimal essential medium (MEM) and incubated for 24 h at 37°C with 5% CO_2. Sterile microtip was used to a make scratch on the 80% confluence monolayer (Camorani et al., 2014). The culture medium was immediately replaced with fresh medium to remove the dislodged cells. Then, 20 nmol/l of the designed aptamer with the transfecting agent, Lipofectamine 2000 (Invitrogen), was dissolved in dimethyl sulfoxide (DMSO) for timeline studies (Zhou et al., 2008). Cell migration of both sample and control were monitored and compared to study the aptamer effect on cell migration.

RESULTS AND DISCUSSION

S100A8 Network With Polycystic Ovary Syndrome Biomarkers

S100A8 interacts with 74 proteins (**Supplementary File S1**). The high confidential STRING network of S100A8 interactors has 55 nodes with 181 edges. The second network with biomarkers has 98 nodes with 431 edges; among 98, two proteins (SLC35D2, MORC4) are unconnected with the main network. Interactor network and biomarker network details are given in **Supplementary Material**. Merged network with connected nodes of the interactors and the biomarkers shows 96 nodes with 430 edges (interactions) that are illustrated in **Figure 2**. Immediate interacting proteins with s100A8 are shown in yellow. By the network analysis, we found that S100A8 directly interact with 10 proteins (**Table 1**).

A total of 246 ontology processes were found within the given significant *p*-value, among them, S100A8 is found in 88 different processes. Particularly, 10 ovulation and maturation-related GO terms with 14 proteins are identified in the enrichment analysis and are listed in **Table 2**. Apart from S100A8, RETN and S100A9 are found in both networks and also in enriched biological terms with high distribution. ESR1, GDF9, PDGFRA, and LEP are the other proteins found in a greater number of pathways comparatively in the selected terms. Distribution of proteins is given in the graph of **Figure 3A**.

TABLE 1 | Proteins associated with S100A8 in the network.

Protein ID	Name	Polycystic Ovary Syndrome (PCOS) relative function
MMP9	Matrix Metallopeptidase 9	Proteolytic activity on the extracellular matrix (ECM) and involved in leukocyte migration
RETN	Resistin	Promotes chemotaxis in myeloid cells
CTSS	Cathepsin S	Endo protease involved in the removal of unwanted proteins
S100A9	S100 Calcium Binding Protein A9	Potent amplifier of inflammation as well as in cancer development and tumor spread
NCF2	Neutrophil Cytosolic Factor 2	Involved in superoxide generation
MPO	Myeloperoxidase	Produce hypochlorous acid and other toxic intermediates which enhance PMN microbicidal activity
ALB	Albumin	Regulates blood plasma colloid osmotic pressure and acts as a carrier protein for hormones, fatty acids, metabolites, and exogenous drugs
BECN1	Beclin 1	Mediates vesicle-trafficking processes, tumorigenesis, neurodegeneration, and apoptosis
TP53	Tumor Protein P53	Prevents CDK7 kinase activity when associated to CAK complex in response to DNA damage, thus stopping cell cycle progression
GRB2	Growth Factor Receptor Bound Protein 2	Adapter protein involved in the Ras signaling pathway

Target Compatibility Evaluation

S100A8 is involved in seven GO functions which are positively regulated cyst formation and cancer cell migration. Additionally, S100A8 poses 18 pockets, among them, nine are druggable (score $>$ 0.3) and four shows better cutoff scores (**Table 3**). Targets with low specificity on small-molecule were identified as poorly druggable targets (Barelier et al., 2010). Here we have found four high scored druggable pockets in the selected S100A8 (**Figure 3C**). Structural features and active sites of S100A8 are shown in **Figures 3B,D**, respectively. Due to the positive druggable results, it is considered as a notable target to control PCOS. Considerably, calcium-binding protein (S100A8) acts as a ligand for receptor of advanced glycation end products (RAGE) which is involved in many inflammatory and oncogenic pathways. There is evidence that S100A8 has a growth-promoting effect, and it helps cells to acquire cell migration activity through the RAGE binding pathway (Ghavami et al., 2008). S100A8 causes uteroplacental perfusion deficiency which leads to embryo abortion that supports the competence of our target selection (Sir-Petermann et al., 2002). Structural analysis shows that S100A8 has two helix loop helix Ca2+ binding domains known as EF-hands and exists as a complex with S100A9. Calprotectin is present in 1q21 locus of chromosome 1 in humans and has a molecular

TABLE 2 | Enriched terms of S100A8 in Polycystic Ovary Syndrome (PCOS).

GO term	Group *p*-value	No. of proteins	Associated proteins found
Ovulation cycle	2.11E-13	9	EGFR, ESR1, GDF9, HAS2, LEP, PDGFRA, RETN, S100A8, S100A9
Female sex differentiation	6.47E-21	12	CTNNA1, ESR1, GDF9, ICAM1, LEP, MYC, PDGFRA, RBP4, RETN, S100A8, S100A9, VEGFA
Ovarian follicle development	6.47E-21	7	CTNNA1, ESR1, ICAM1, MYC, S100A8, S100A9, VEGFA
Ovulation cycle	1.38E-24	9	EGFR, ESR1, GDF9, HAS2, LEP, PDGFRA, RETN, S100A8, S100A9
Ovulation cycle process	1.38E-24	7	ESR1, GDF9, LEP, PDGFRA, RETN, S100A8, S100A9
Development of primary female sexual characteristics	1.38E-24	11	CTNNA1, ESR1, GDF9, ICAM1, LEP, MYC, PDGFRA, RETN, S100A8, S100A9, VEGFA
Gonad development	1.38E-24	11	CTNNA1, ESR1, GDF9, ICAM1, LEP, MYC, PDGFRA, RETN, S100A8, S100A9, VEGFA
Ovarian follicle development	1.38E-24	7	CTNNA1, ESR1, ICAM1, MYC, S100A8, S100A9, VEGFA
Female gonad development	1.38E-24	11	CTNNA1, ESR1, GDF9, ICAM1, LEP, MYC, PDGFRA, RETN, S100A8, S100A9, VEGFA
Regulation of female gonad development	1.38E-24	4	GDF9, RETN, S100A8, S100A9

weight of 10–12 kDa. During tumor development, chromosomal rearrangements take place in the locus of the S100A8 gene and majorly contribute to the cyst formation in PCOS. Also, serum calgranulin (S100A8 and S100A9) levels are higher in women with PCOS than normal women (Dai and Lu, 2012). This evidently shows that binding of S100A8 with RAGE facilitates the p38 mitogen-activated protein (MAP) kinase signaling through calcium phosphorylation which also governs cyst migration.

Construction of RNA Analog Library Using Glucocorticoid Response Element

The fragment-based approach of aptamer docking yielded better interaction with S100A8. By the RNA-Lim method,

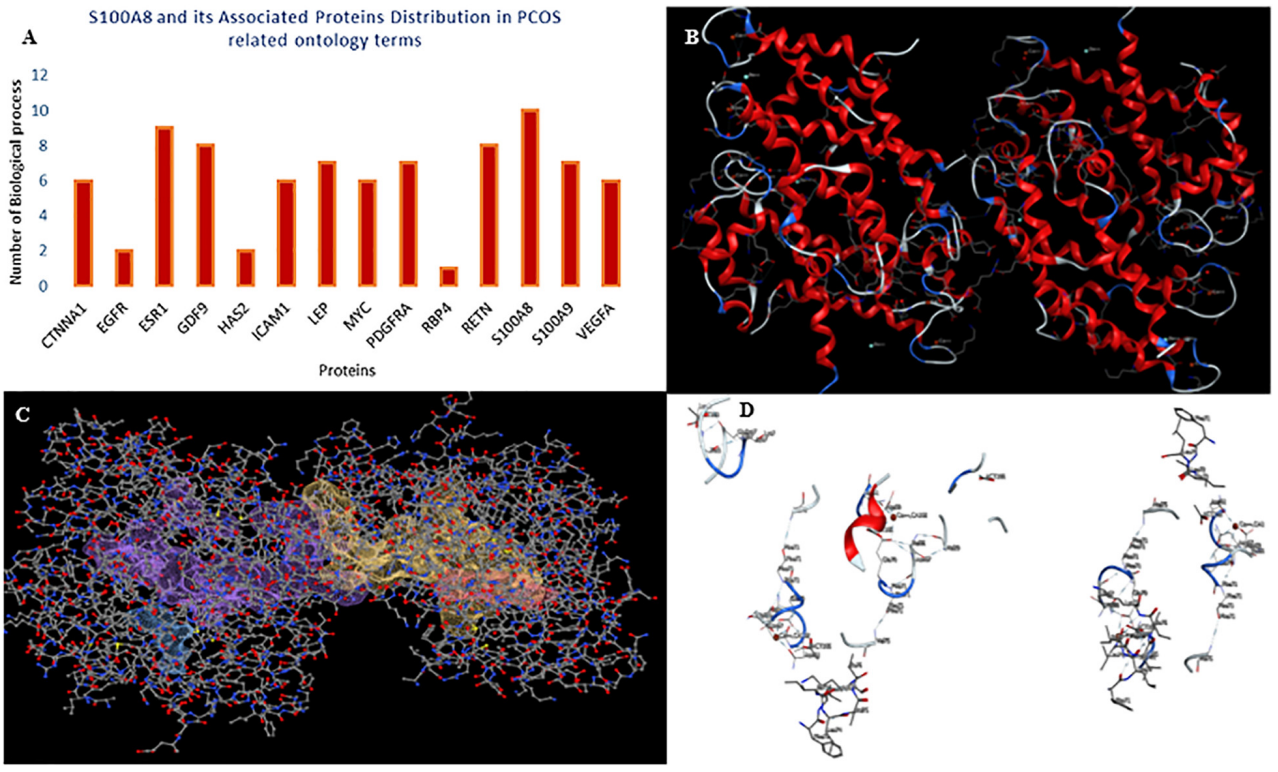


FIGURE 3 | Gene ontology enrichment and structural properties of S100A8. **(A)** Distribution of proteins in polycystic ovary syndrome (PCOS)-related terms in enrichments. **(B)** Homo dimer crystal structure of S100A8—global symmetric view. **(C)** High druggable pockets of S100A8—shown in filled purple, sky blue, yellow, and orange colors. **(D)** Active sites of S100A8.

TABLE 3 | Druggability assessment of S100A8 protein.

Pocket ID	Volume A^2	Surface A^2	Drug score	Simple score
P_0	2,694.56	2,620.71	0.81	0.61
P_1	2,652.05	2,822.45	0.81	0.64
P_3	209.65	195.14	0.66	0
P_4	180.41	168.05	0.6	0
P_2	252.39	481.15	0.5	0.14
P_6	166.91	273.52	0.37	0
P_7	165.23	357.95	0.37	0.04
P_5	173.21	283.31	0.35	0
P_8	137.56	211.73	0.33	0
P_9	130.58	204.55	0.28	0
P_10	127.43	196.55	0.27	0
P_11	120.91	239.32	0.27	0
P_14	109.89	180.88	0.26	0
P_12	117.99	208.39	0.25	0
P_13	116.19	209.36	0.22	0
P_15	109.89	259.05	0.16	0
P_16	106.85	216.05	0.15	0
P_17	100.78	249.25	0.14	0

18 fragments with the consensus sequence of GRE were constructed and used for binding analysis (**Figure 4A**). Frag6, Frag9, and Frag10 showed better interaction (**Table 4**) in the active domain of target with minimal global binding energy. Among the three possible conformations, sequence 1 (**Figure 4B**) shows better thermal stability and lowest energy than the other two sequences. Optimal structure with a folding simulation at physiological pH shows there are three nucleotides at positions 4–7 that make intramolecular base pairing for loop structure (**Figure 4C**). Energy minimized aptamers are significantly stable, and the aptamers with a binding energy of ≥-40 are optimal in the therapeutical aspect (Pagano et al., 2008). Oligo fragments selected are by their binding ability on the active sites of the target. The compiled 18-mer binds effectively than the fragments. Stability comparison among the newly constructed aptamer sequences is stated in **Table 5**. Among the three, Apt1 has high stability with a melting temperature of 41.8°C, and also the simulation studies confirmed that it requires the minimum free energy (–27.93 kcal/mol) for hairpin loop formation. In addition, the Apt1 fragment poses low molecular weight (5,327.4 g/mol) comparatively. Aptamers in practice are available in the range of 15–81 nucleotide length with higher molecular weights (Shigdar et al., 2013), but here, the designed is 18-mer with lower molecular weight so the plasma clearance may be faster.

Interaction, Inhibition, and Stability Studies

Primarily, RAGE being the receptor for S100A8 was docked to confirm for its binding ability in the domain, which may affect the binding of the designed aptamer. As a result

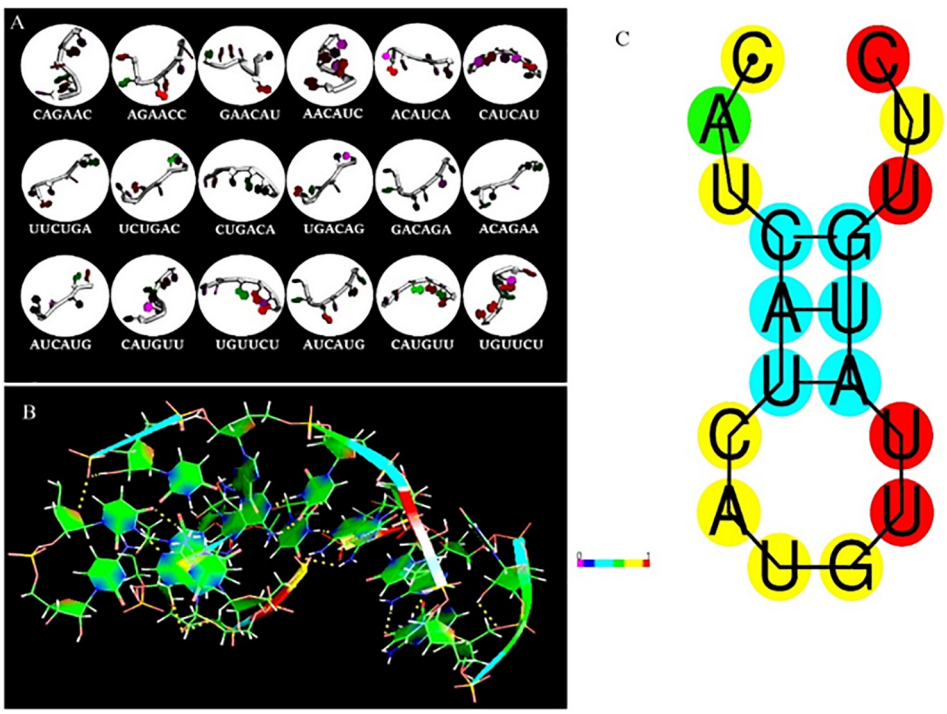


FIGURE 4 | RNA library and structures. **(A)** Library construction by RNA-Lim method using the consensus sequence of glucocorticoid response element (GRE). **(B)** 3D structure of potential aptamer candidate. **(C)** Sequence and physiological structure of aptamer after optimal simulation and folding, and the color of bases indicates their energy levels.

TABLE 4 | Binding energies of RNA analog fragments with S100A8.

Fragments	Predicted ΔG (kcal/mol)	Fragments	Predicted ΔG (kcal/mol)
Frag 1	–16.24	Frag 10	–30.51
Frag 2	–8.55	Frag 11	–14.34
Frag 3	–16.52	Frag 12	–15.66
Frag 4	–51.85	Frag 13	–15.17
Frag 5	–18.46	Frag 14	–11.21
Frag 6	–31.71	Frag 15	–18.70
Frag 7	–16.54	Frag 16	–10.68
Frag 8	–13.17	Frag 17	–21.76
Frag 9	–38.76	Frag 18	–23.71

of protein–protein docking analysis, Arg 114 residue at A domain of RAGE interacting with Gln 44 residue at H domain of S100A8 is found as the most active interaction. To test the comparison of binding interactions, RAGE was docked with GRE, which resulted in a binding energy of –24.38, comparatively higher than its binding with the designed aptamers (–46.33) that is shown in **Table 6**; this infers that the designed nucleotide aptamer also binds efficiently at the S100A8 binding domain of RAGE (**Figures 5B,C**). In parallel, interactions of S100A8 with GRE and S100A8 with the designed aptamer were inspected to find the competency, which was found as –22.11 and –45.32 energy levels, respectively. The designed aptamer binds efficiently in the active dimer of the target (**Figure 5A**).

Anti-cell Migration Assay on MCF-7 Cell Line

Within 4 h of scratch, development of closure was seen in the control (which does not have aptamer), the wounded area has turned into a normal layer when compared with the initial image of well. In the aptamer well, there is no cell migration observed even after the fourth hour of incubation, it was confirmed in the images of 0 and 4 h of wounded well (**Figure 6**).

CONCLUSION

From the network analysis, S100A8 is identified as a targetable protein to control PCOS. The druggable property of the target was validated by topological measures. S100A8 acts as a ligand for RAGE to promote cell migration in cancers and PCOS conditions. GRE inhibits S100A8 by competitive binding at the minimal level through a feedback mechanism. Additionally, S100A9 and resistin were also found along with

TABLE 6 | Docking results.

Protein	Target	Binding site	Global energy
RAGE	S100A8	Arg A 114 → Gln H44	–25.75
RAGE	GRE	Arg B 203 → U_{14}, Arg B 228 → G_{17}	–24.38
S100A8	GRE	Asn D 61 → C_6, Ala B 1 → U_{13}	–22.11
RAGE	Aptamer	• Try B 118 → A_{13}, Arg B 216 → A_{13} • Arg B 218 → G_{10}, Asn B 25 → U_{16} • Gln B 24 → U_{16}	–46.33
S100A8	Aptamer	• Lys B 36 → U_9, Lys F 48 → U_{17} • Ser H 86 → A_{13}, Asp C 32 → G_{10} • Lys B 18 → G_{10}, Lys B 21 → G_{10}	–45.32

S100A8 as associative proteins. We adopted a computational method to develop an RNA aptamer and designed 18

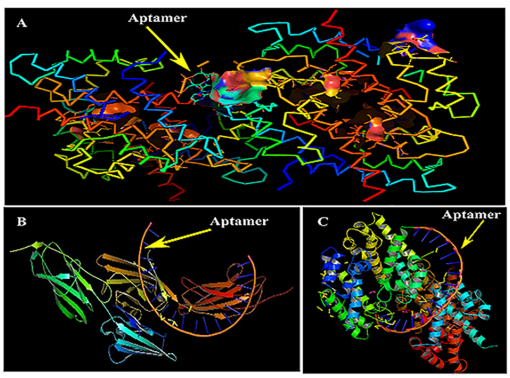


FIGURE 5 | Mechanism of inhibition and docking with RNA aptamer. **(A)** Aptamer bases binding at the active sites of S100A8. **(B)** Binding pose of aptamer on receptor of advanced glycation end products (RAGE) protein. **(C)** Optimal binding pose of aptamer on the S100A8.

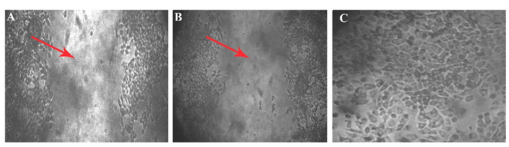


FIGURE 6 | Wound scratch assay. Cell migration studies by wound healing assay on MCF7 cell line. Leitz labovert FS inverted microscope was used to view the cell migration. **(A)** A 4 mm scratch with a sterile tip at 0 h, and the wound is pointed at with an arrowhead in the 10 × magnified illustration. **(B)** Ten times magnified view of aptamer incubated (4 h) cell line. No cell migration was observed after 4 h of aptamer incubation. **(C)** Fifty times magnification of scratch in control plate (cell migration seen) after the fourth hour.

TABLE 5 | Aptamer stability comparison.

	Aptamers	GC content (%)	Tm (°C)	Molecular weight (g/mol)	ΔGmax (kcal/mol)
Apt1	CAUCAUCAUGUUAUGUUC	33.3	41.8	5327.4	–27.93
Apt2	AACAUCACAGAAGACAGA	38.9	37	5504.7	–28.3
Apt3	CUGACAACAUCAAUCAUG	38.9	36.6	5395.5	–29.04

oligos based on the consensus sequences of GRE, which binds to both RAGE and S100A8. In addition to the computational studies, the cell line studies proved the anti-migration activity of the designed aptamer at minimal dose delivery with Lipofectamine 2000. The newly designed 18mer effectively stopped the cancer cell migration through dual action, and it is identified as a potential therapeutic to control PCOS and cancers.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

All authors listed have made substantial, direct, and intellectual contribution to the work and approved it for publication.

ACKNOWLEDGMENTS

We thank the scientific committee of Kamaraj College of Engineering and Technology for their encouragement with best project award in TECHNOVISION'17. Our sincere thanks to Tamil Nadu State Council for Science and Technology (TNSCST), Directorate of Technical Education, Chennai, for their partial financial support through student project scheme. I dedicate this manuscript as a gift to my student A. Shobana for her recent birthday. This manuscript has been released as a pre-print at bioRxiv (https://www.biorxiv.org/content/10.1101/603357v1), doi: https://doi.org/10.1101/603357 entitled "*Identification of target candidate in polycystic ovarian syndrome and in vitro evaluation of therapeutic activity of the designed RNA aptamer*," and this is a revised version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00328/full#supplementary-material

REFERENCES

Ahirwar, R., Smita, N., Shikha, A., Srinivasan, R., Souvik, M., and Pradip, N. (2016). In silico selection of an aptamer to estrogen receptor alpha using computational docking employing estrogen response elements as aptamer-alike molecules. *Sci. Rep.* 6:21285. doi: 10.1038/srep21285

Barelier, S., Pons, J., Gehring, K., Lancelin, J.-M., and Krimm, I. (2010). Ligand specificity in fragment-based drug design. *J. Med. Chem.* 53, 5256–5266. doi: 10.1021/jm100496j

Berezovski, M. V., Musheev, M. U., Drabovich, A. P., Jitkova, J. V., and Krylov, S. N. (2006). Non-SELEX: selection of aptamers without intermediate amplification of candidate oligonucleotides. *Nat. Protoc.* 1:1359. doi: 10.1038/nprot.2006.200

Biesiada, M., Pachulska-Wieczorek, K., Adamiak, R. W., and Purzycka, K. J. (2016). RNAComposer and RNA 3D structure prediction for nanotechnology. *Methods* 103, 120–127. doi: 10.1016/j.ymeth.2016.03.010

Bindea, G., Bernhard, M., Hubert, H., Pornpimol, C., Marie, T., Amos, K., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Camorani, S., Esposito, C. L., Rienzo, A., Catuogno, S., Iaboni, M., Condorelli, G., et al. (2014). Inhibition of receptor signaling and of glioblastoma-derived tumor growth by a novel PDGFRbeta aptamer. *Mol. Ther.* 22, 828–841. doi: 10.1038/mt.2013.300

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39, D691–D697. doi: 10.1093/nar/gkq1018

Daan, N. M., Koster, M. P., de Wilde, M. A., Dalmeijer, G. W., Evelein, A. M., Fauser, B. C., et al. (2016). Biomarker profiles in women with PCOS and PCOS Offspring; a pilot study. *PLoS One* 11:e0165033. doi: 10.1371/journal.pone.0165033

Dai, G., and Lu, G. (2012). Different protein expression patterns associated with polycystic ovary syndrome in human follicular fluid during controlled ovarian hyperstimulation. *Reprod Fertil. Dev.* 24, 893–904. doi: 10.1071/RD11201

Gerashchenko, T. S., Nikita, M. N., Nadezhda, V. K., Sofia, Y. Z., Marina, V. Z., and Nadezhda, V. (2019). Markers of cancer cell invasion: are they good enough? *J. Clin. Med.* 8:1092. doi: 10.3390/jcm8081092

Ghavami, S., Rashedi, I., Dattilo, B. M., Eshraghi, M., Chazin, W. J., Hashemi, M., et al. (2008). S100A8/A9 at low concentration promotes tumor cell growth via RAGE ligation and MAP kinase-dependent pathway. *J. Leukoc. Biol.* 83, 1484–1492. doi: 10.1189/jlb.0607397

Ghavami, S., Seth, C., Mohammad, H., Mehdi, E., Walter, J. C., Andrew, J. H., et al. (2009). S100A8/A9: a janus-faced molecule in cancer therapy and tumorgenesis. *Eur. J. Pharmacol.* 625, 73–83. doi: 10.1016/j.ejphar.2009.08.044

Hall, D., Li, S., Yamashita, K., Azuma, R., Carver, J. A., and Standley, D. M. (2015). RNA-LIM: a novel procedure for analyzing protein/single-stranded RNA propensity data with concomitant estimation of interface structure. *Anal. Biochem.* 472, 52–61. doi: 10.1016/j.ab.2014.11.004

Hsu, K., Passey, R. J., Endoh, Y., Rahimi, F., Youssef, P. P., Yen, T., et al. (2005). Regulation of S100A8 by glucocorticoids. *J. Immunol.* 174, 2318–2326. doi: 10.4049/jimmunol.174.4.2318

Lu, C., Xiaoqin, L., Lin, W., Ning, J., Jun, Y., Xiaobo, Z., et al. (2017). Integrated analyses for genetic markers of polycystic ovary syndrome with 9 case-control studies of gene expression profiles. *Oncotarget* 8, 3170–3180. doi: 10.18632/oncotarget.13881

Mashiach, E., Schneidman-Duhovny, D., Andrusier, N., Nussinov, R., and Wolfson, H. J. (2008). FireDock: a web server for fast interaction refinement in molecular docking. *Nucleic Acids Res.* 36, W229–W232. doi: 10.1093/nar/gkn186

Nioi, P., McMahon, M., Itoh, K., Yamamoto, M., and Hayes, J. D. (2003). Identification of a novel Nrf2-regulated antioxidant response element (ARE) in the mouse NAD(P)H:quinone oxidoreductase 1 gene: reassessment of the ARE consensus sequence. *Biochem. J.* 374(Pt 2), 337–348. doi: 10.1042/bj20030754

Oughtred, R., Chris, S., Bobby-Joe, B., Jennifer, R., Lorrie, B., Christie, C., et al. (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* 47, D529–D541. doi: 10.1093/nar/gky1079

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., et al. (2008). IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res.* 36, W163–W169. doi: 10.1093/nar/gkn198

Pagano, B., Martino, L., Randazzo, A., and Giancola, C. (2008). Stability and binding properties of a modified thrombin binding aptamer. *Biophys. J.* 94, 562–569. doi: 10.1529/biophysj.107.117382

Parisien, M., and Major, F. (2008). The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452, 51–55. doi: 10.1038/nature06684

Rees, B. B., Bowman, J. A., and Schulte, P. M. (2001). Structure and sequence conservation of a putative hypoxia response element in the lactate dehydrogenase-B gene of Fundulus. *Biol. Bull.* 200, 247–251. doi: 10.2307/1543505

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33, W363–W367.

Shigdar, S., Qiao, L., Zhou, S.-F., Xiang, D., Wang, T., Li, Y., et al. (2013). RNA aptamers targeting cancer stem cell marker CD133. *Cancer Lett.* 330, 84–95. doi: 10.1016/j.canlet.2012.11.032

Sir-Petermann, T., Maliqueo, M., Angel, B., Lara, H. E., Pérez-Bravo, F., Recabarren, S. E., et al. (2002). Maternal serum androgens in pregnant women with polycystic ovarian syndrome: possible implications in prenatal androgenization. *Hum. Reprod.* 17, 2573–2579. doi: 10.1093/humrep/17.10.2573

Strimbu, K., and Tavel, J. A. (2010). What are biomarkers? *Curr. Opin. HIV AIDS* 5, 463–466. doi: 10.1097/COH.0b013e32833ed177

Szklarczyk, D., Annika, L. G., David, L., Alexander, J., Stefan, W., Jaime, H., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tan, B. K., Raghu, A., Jing, C., Hendrik, L., Louis, J. S. C., Harpal, S. R., et al. (2011). Metformin treatment exerts antiinvasive and antimetastatic effects in human endometrial carcinoma cells. *J. Clin. Endocrinol. Metab.* 96, 808–816. doi: 10.1210/jc.2010-1803

Tseng, C. Y., Ashrafuzzaman, M., Mane, J. Y., Kapty, J., Mercer, J. R., and Tuszynski, J. A. (2011). Entropic fragment-based approach to aptamer design. *Chem. Biol. Drug Des.* 78, 1–13. doi: 10.1111/j.1747-0285.2011.01125.x

Volkamer, A., Kuhn, D., Rippmann, F., and Rarey, M. (2012). DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessment. *Bioinformatics* 28, 2074–2075. doi: 10.1093/bioinformatics/bts310

Weikum, E. R., Knuesel, M. T., Ortlund, E. A., and Yamamoto, K. R. (2017). Glucocorticoid receptor control of transcription: precision and plasticity via allostery. *Nat. Rev. Mol. Cell Biol.* 18, 159–174. doi: 10.1038/nrm.2016.152

Yang, Z., Sugawara, M., Ponath, P. D., Wessendorf, L., Banerji, J., Li, Y., et al. (1990). Interferon gamma response region in the promoter of the human DPA gene. *Proc. Natl. Acad. Sci. U.S.A.* 87, 9226–9230. doi: 10.1073/pnas.87.23.9226

Zhou, J., Li, H., Li, S., Zaia, J., and Rossi, J. J. (2008). Novel dual inhibitory function aptamer-siRNA delivery system for HIV-1 therapy. *Mol. Ther.* 16, 1481–1489. doi: 10.1038/mt.2008.92

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 October 2020
doi: 10.3389/fgene.2020.507038

JS-MA: A Jensen-Shannon Divergence Based Method for Mapping Genome-Wide Associations on Multiple Diseases

*Xuan Guo**

Department of Computer Science and Engineering, University of North Texas, Denton, TX, United States

Taking advantage of the high-throughput genotyping technology of Single Nucleotide Polymorphism (SNP), Genome-Wide Association Studies (GWASs) have been successfully implemented for defining the relative role of genes and the environment in disease risk, assisting in enabling preventative and precision medicine. However, current multi-locus-based methods are insufficient in terms of computational cost and discrimination power to detect statistically significant interactions with different genetic effects on multifarious diseases. Statistical tests for multi-locus interactions (≥2 SNPs) raise huge analytical challenges because computational cost increases exponentially as the growth of the cardinality of SNPs in an interaction module. In this paper, we develop a simple, fast, and powerful method, named JS-MA, based on Jensen-Shannon divergence and agglomerative hierarchical clustering, to detect the genome-wide multi-locus interactions associated with multiple diseases. From the systematical simulation, JS-MA is more powerful and efficient compared with the state-of-the-art association mapping tools. JS-MA was applied to the real GWAS datasets for two common diseases, i.e., Rheumatoid Arthritis and Type 1 Diabetes. The results showed that JS-MA not only confirmed recently reported, biologically meaningful associations, but also identified novel multi-locus interactions. Therefore, we believe that JS-MA is suitable and efficient for a full-scale analysis of multi-disease-related interactions in the large GWASs.

Keywords: GWAS, Jensen-Shannon divergence, clustering, epistasis, genetic factors

OPEN ACCESS

Edited by:
Fa Zhang,
Chinese Academy of Sciences, China

Reviewed by:
Xiaodan Fan,
The Chinese University of Hong Kong, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Xuan Guo
xuan.guo@unt.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *24 October 2019*
Accepted: *21 September 2020*
Published: *30 October 2020*

Citation:
Guo X (2020) JS-MA: A Jensen-Shannon Divergence Based Method for Mapping Genome-Wide Associations on Multiple Diseases. Front. Genet. 11:507038. doi: 10.3389/fgene.2020.507038

1. INTRODUCTION

Genome-wide association studies (GWASs) have been proved to be a powerful tool to identify the genetic susceptibility of associations between a trait of interests using statistical tests (Sabaa et al., 2013). Recent studies have confirmed that single nucleotide polymorphisms (SNPs) are associated with a variety of common diseases (Peter and Hunter, 2009). The current primary research paradigm in GWASs is dominated by analyzing the susceptibility of single SNP to one disease at a time. One SNP might only explain a small part of causal genetic effects for multiple complex diseases (He and Lin, 2011). The word, epistasis, is defined generally as the interaction among different genes (Cordell, 2002). Many studies have demonstrated that epistasis is an important contributor to genetic variation in complex diseases. Most common diseases, such as obesity (Cordell, 2009), cancer (Ritchie et al., 2001), diabetes (Wang et al., 2012), and heart disease (Nelson et al., 2001), are complex traits, which result from a joint effect of various genetic

variants, environmental factors, or their interactions. It is of great interest for us to identify the genetic risk factors for complex diseases, so as to understand disease mechanisms, develop effective treatments, and improve public health. The cost of genomic technologies is falling exponentially over time. For instance, the Human Genome Project took 13 years and cost $2.7 billion in the early twenty-first century, whereas now we can sequence a genome with $1,000 and less than a week. The availability of large-scale genotyping technology with its rapid improvement makes the cost of genome-wide analyses widely decrease, and a great number of large-scale genetic association studies are initiated. Complex diseases do not show the "simple" inheritance pattern observed in Mendelian diseases, where alterations in a single gene or a unique locus are causal for a phenotype. In complex disease, multiple genes are involved, each with low-penetrance that each gene modestly increases the probability of disease and does not ultimately determine disease status. These factors often render the traditional genetic dissection approaches, such as linkage analysis, ineffective tools to study complex diseases. In this article, we consider epistatic interactions as the statistically significant associations of d-SNP modules ($d \geq 2$) with multiple phenotypes (Wang et al., 2011).

The problem of detecting high-order genome-wide epistatic interaction for case-control data has attracted more research interests recently. Generally, there are two challenges in mapping genome-wide associations for multiple diseases on a large GWAS dataset (Guo et al., 2014a): the first is arose from the heavy computational burden, i.e., the number of association patterns increases exponentially as the order of interaction goes up. For example, there are around 6.25×10^{11} statistical tests required to detect pairwise interactions for a moderate dataset with ∼500,000 SNPs. The second challenge is that existing approaches do not have enough statistical powers to report significant high-order multi-locus interaction on multiple diseases. Because of the huge number of hypotheses and the limited sample size, a large proportion of significant associations are expected to be false positives. In recent, many computational algorithms have been proposed to overcome the above difficulties. They can be broadly classified into three categories (Xie et al., 2012): exhaustive search, stepwise search, and heuristics approach. The naive solution to tack the problem is exhaustive search using statistical tests, like χ^2 test, exact likelihood ratio test or entropy-based test, for all SNP modules (Wan et al., 2010c; Liu et al., 2011; Yung et al., 2011). In order to minimize the huge computation requests, stepwise search strategies select a subset of SNPs or their combinations based on some low-order measurement tests, then extend them to higher-order interactions if it is statistically possible (Marchini1 et al., 2005; Li, 2008). Heuristic methods adopt machine learning or stochastic procedures to search the space of interactions rather than explicitly enumerating all combinations of SNPs (Zhang and Liu, 2007; Wan et al., 2010b). More details about the popular GWAS mapping tools can be found in recent surveys (Guo et al., 2014b; Niel et al., 2015; Visscher et al., 2017; Wen et al., 2017).

To the best of our knowledge, most epistasis detecting tools are only capable of identifying interactions on the data of GWAS with two groups, i.e., case-control studies. These tools are incompetent to discover genetic factors with diverse effects on multiple diseases. Moreover, using a limited number of case samples may lose the benefit of alleviating deficiency of statistical powers by pooling different disease samples together. Recently, Guo et al. developed a Bayesian inference based method, named DAM, to detect multi-locus epistatic interactions on multiple diseases (Guo et al., 2015, 2017). From our experiments, DAM took 3 days to finish the analyzing a real GWAS dataset using a desktop computer and only reported a few significant epistatic interactions. In this manuscript, we present a heuristic method, named JS-AM, based on Jensen-Shannon divergence and agglomerative hierarchical clustering to select a set of candidate SNPs that potentially have effects on multiple phenotypic traits (Guo, 2015). A stepwise interaction evaluation is engaged in JS-MA to further determining the association types. Systematic experiments on both simulated and real GWAS datasets demonstrate that JS-AM is feasible for identifying multi-locus interaction using GWAS datasets and enriches some novel, significant high-order epistatic interactions with various effects on multiple diseases.

2. MATERIALS AND METHODS

2.1. Notation

For a GWAS dataset, let L denote the total number of groups, including $L-1$ case groups and one control group. Each group has N_l samples with $l \in \{1, 2, \ldots, L\}$. Let N be the total count of samples from these L groups, and M be the number of diallelic SNP markers. In general, the major alleles are represented by uppercase letters (e.g., A, B,...) and the minor alleles are represented by lowercase letters (e.g., a, b). We use $\{0, 1, 2\}$ to represent $\{AA, Aa, aa\}$. We use X to indicate the SNP set, where x_i indicates the i-th SNP. Let $g_{x_i,\ldots,x_j}$ be the combination of genotypes giving a list of SNPs $\{x_i, \ldots, x_j\}$. The probability distribution of $g_{x_i,\ldots,x_j}$ is denoted as $p_{g_{x_i,\ldots,x_j}}$, or p_g for simplicity.

Different from the most existing methods that deal with one case and one control groups, we have two or more cases. The number of partitions of L groups is known as the Bell number (Guo et al., 2015). The SNPs can be assigned to be associated with one or more cases either with the same or different effects. Here, we call the assignment based on association as trait-association types, or AT in short. An example about five association types for a three-group dataset is shown in **Figure 1**. In this example, each AT includes 2 SNPs. There are three different probability distributions of genotype combinations, which are labeled by color white, gray, and black. SNPs 1 & 2 are related to case 1, and we call this type effect as AT1. Similarly, we call the trait-association types for SNPs 3 & 4 and SNPs 5 & 6 are AT2 and AT3, respectively. For SNPs 7 & 8, the genotype combinations display different effects on two cases, and we label it as AT4. For the last two SNPs, they are not related to any case, i.e., following the same probability distribution among three groups, and we call it AT5. In general, the number association types is increasing as the number of phenotype groups increases, which is controlled by the Bell number. We use Ψ to denote the set of association types that have

		AT 1		AT 2		AT 3		AT 4		AT 5	
	SNP ID	1	2	3	4	5	6	7	8	9	10
Case 1		2	1	0	0	2	1	1	0	0	0
		2	1	1	1	2	1	1	0	1	1
		2	1	2	2	2	1	1	0	2	2
		2	1	0	0	2	1	1	0	0	0
		2	1	0	0	2	1	1	0	0	0
Case 2		0	0	2	1	2	1	2	1	0	0
		1	1	2	1	2	1	2	1	1	1
		2	2	2	1	2	1	2	1	2	2
		0	0	2	1	2	1	2	1	0	0
		0	0	2	1	2	1	2	1	0	0
Control		0	0	0	0	0	0	0	0	0	0
		1	1	1	1	1	1	1	1	1	1
		2	2	2	2	2	2	2	2	2	2
		0	0	0	0	0	0	0	0	0	0
		0	0	0	0	0	0	0	0	0	0

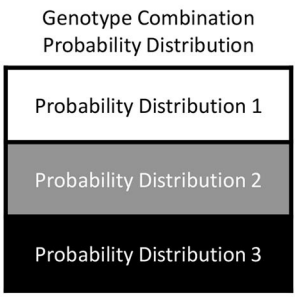


FIGURE 1 | The illustration for five association types by giving three groups. Ten SNPs of AT 1, 2, 3, 4, and 5 are associated with the phenotype traits with interactions between each pair of them.

different probability distribution between the case and control groups. Given L groups, we denote the number of all pairwise combinations as $|H| = L(L-1)/2$ and the combination set as $H = \{h_1, \ldots, h_{|H|}\}$. The probability distributions of genotype data in h_i combination are denoted as $p^{(h_i)}$ and $q^{(h_i)}$ for the first and second groups, respectively.

2.2. Jensen-Shannon Divergence

We used a distance measurement based on the Jensen-Shannon divergence (JS) for measuring the similarity between two SNPs. JS is a popular distance measurement based on Kullback-Leibler divergence (Lin, 1991), which evaluates the similarity between two probability distributions. Given two distributions, p and q, both with g categories, the Kullback-Leibler divergence is defined as follows:

$$\mathbb{KL}\left(p \parallel q\right) = \sum_{i=1}^{g} p_g log \frac{p_g}{q_g} \tag{1}$$

The KL divergence is not a distance because it is not symmetric. One symmetric version of KL divergence is JS, defined as:

$$JS\left(p, q\right) = 0.5\mathbb{KL}\left(p \parallel \frac{p+q}{2}\right) + 0.5\mathbb{KL}\left(q \parallel \frac{p+q}{2}\right) \tag{2}$$

where $\frac{p+q}{2}$ is the pointwise mean of p and q. Here, for a genotype g, $\frac{p+q}{2}$ is equal to the average of p_g and q_g. Given a pairwise group combination h_k and two SNPs, x_i and x_j, we denote the probability distributions of the genotype combination of x_i and x_j as p^{h_k} for the first group and q^{h_k} for the second group. Based on JS, we define the distance between two SNPs, x_i and x_j as follows:

$$Dist(x_i, x_j) = \frac{\sum_{h_k \in H} JS\left(p^{h_k}, q^{h_k}\right)}{|H|} \tag{3}$$

If these two SNPs are associated to any cases, the distribution of genotype combinations in case groups should be the same as the one in control. And $Dist(x_i, x_j)$ should be a very small value toward 0; otherwise, $Dist(x_i, x_j)$ is a large value toward 1.

2.3. Clustering

Our goal is to find a list of SNP modules containing $d (d \geq 2)$ SNPs, which have large JS dissimilarity between any two groups. It is computationally expensive to examine all d SNP combinations when $d \geq 3$ given millions of SNPs in one dataset. In order to diminish the time complexity, we use agglomerative hierarchical clustering to group SNPs into clusters so that SNPs jointly affecting a trait go into separate clusters. More specifically, the complete-linkage clustering criterion was used to determine the distance between sets of SNPs. The distance from an SNP, x_i, to a cluster, C, is defined as

$$Dist(x_i, C) = \max_{x_j \in C} Dist(x_i, x_j) \tag{4}$$

The distance between two clusters is defined as

$$Dist(C_i, C_j) = \max_{a \in C_i, b \in C_j} Dist(a, b) \tag{5}$$

In the implementation of JS-MA, we used the nearest-neighbor chain algorithm (Murtagh, 1983; Müllner, 2011). Compared to the greedy algorithm that repeatedly forms a new cluster by merging the closest pair of clusters, the nearest-neighbor chain algorithm runs faster by merging pairs of clusters in a different order. In brief, the nearest neighbor chain algorithm grows a chain of clusters, where the newly added cluster is the nearest neighbor of the previous one, and stops growing when reaching a pair of clusters that are mutual nearest neighbors. For our complete-linkage clustering criterion, the nearest neighbor chain algorithm can be guaranteed to generate the same hierarchical clustering as the greedy algorithm (Murtagh, 1983; Müllner, 2011). The time complexity of the nearest-neighbor chain algorithm is $O(M^2)$, where M is the number of SNPs. In our setting, we will stop the chain growing once the number of clusters reaches the expected number. Here, the number of clusters is a user-defined parameter. It can be set to the largest, expected size of epistatic modules. In our simulation, we set the number of clusters to two and three for 2- and 3-locus models, respectively. In the real data experiments, we set the number of clusters to ten. Once the clustering is done, top f SNPs from every cluster are selected for further interaction testing. Here, f is a user-defined number. An SNP will be picked if it shows a high dissimilarity measured by JS with other SNPs between any two groups. Every SNP is ranked based on the following score.

$$Score(x) = \sum_{x \notin C_i,} Dist(x, C_i) \tag{6}$$

2.4. Stepwise Evaluation of Interaction

We apply the χ^2 statistic and the conditional χ^2 test similar to the ones in (Guo et al., 2015) to measure the statistical significance for a SNP module. Let $\mathbb{A} = (x_1, x_2, \ldots, x_d : T)$ denote an SNP module $\mathbb{A}$ with d SNPs of association type T. We use $\chi^2(x_1, x_2, \ldots, x_d : T)$ to denote the χ^2 statistic of $\mathbb{A}$ and $\chi^2(x_1, x_2, \ldots, x_d | x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}} : T)$ as the conditional χ^2 statistic given a subset $\mathbb{A}' = (x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}})$ with d' SNPs. The χ^2 statistic is calculated as

$$\chi^2(x_1, x_2, \ldots, x_d : T) = \sum_{i=1}^{|S_T|} \sum_{s=1}^{3^d} \frac{(n_{i,s} - e_{i,s})^2}{e_{i,s}} \tag{7}$$

where $n_{i,s}$ is the frequency of s-th genotype combination in i-th disjoint set for the association type T, $e_{i,s}$ is the corresponding expected frequency, and S_T denotes all the disjoint sets for L groups. The degrees of freedom for Equation (7) is $(|S_T| - 1) \cdot (3^d - 1)$. The conditional χ^2 statistic is defined as follows

$$\chi^2(x_1, \ldots, x_d | x_{c_1}, \ldots, x_{c_{d'}} : T) = \sum_{\iota=1}^{3^{d'}} \sum_{i=1}^{|S_T|} \sum_{s=1}^{3^{d-d'}} \frac{(n_{i,s}^{(\iota)} - e_{i,s}^{(\iota)})^2}{e_{i,s}^{(\iota)}} \tag{8}$$

where we calculate χ^2 statistic for $\mathbb{A} - \mathbb{A}'$ separately for each genotype combination in $\mathbb{A}'$. The degrees of freedom for Equation (8) is $3^{d'} \cdot (|S_T| - 1) \cdot (3^{d-d'} - 1)$. We treat SNPs as redundant SNPs when they are conditional independent given a subset of the SNP module. To avoid the redundant SNPs, we are looking for compact epistatic interactions, which is defined as follows:

Definition 1. *An SNP module $\mathbb{A} = (x_1, x_2, \ldots, x_d)$ is considered as a significant, compact interaction given a significant level α_d, if it meets the following two conditions:*
(1) The p*-value of $\chi^2(x_1, \ldots, x_d) \leq \alpha_d$, where the* p*-value of $\chi^2(x_1, \ldots, x_d) = \min_T \chi^2(x_1, \ldots, x_d : T)$;*
(2) The p*-value of $\chi^2(x_1, \ldots, x_d | x_{c_1}, \ldots, x_{c_{d'}}) \leq \alpha_d$, for $\forall \mathbb{A}' = (x_{c_1}, x_{c_2}, \ldots, x_{c_{d'}})$, given the association type $= \arg\min_T \chi^2(x_1, \ldots, x_d : T)$.*

Based on the Definition 1, we develop a stepwise algorithm to search for d-locus significant compact interactions. We assume that one SNP can only participate in one significant interaction and is only associated with one association type. We first search all modules with only one SNP based on Definition 1. Then we recursively enlarge the SNP module size by one at a time until it reaches a user pre-set value d. We add all novel d-way interactions (i.e., none of the SNPs in the module has been reported earlier) that are significant to a list $\mathbb{L}$ after applying Bonferroni correction for $\Psi \cdot \binom{M}{d}$ tests. For the interactions whose subsets have been reported as significant before, we use the conditional independent test, and put the interaction in $\mathbb{L}$ if it is still significant after Bonferroni correction for $\Psi \cdot \binom{M}{d} \cdot \binom{d}{d'}$ tests. We also apply a distance constraint that the physical distance between two SNPs in a multi-locus module should be at least 1Mb when analyzing real data. This constraint is used to avoid associations that might be due to the linkage disequilibrium effect (Cordell, 2002).

2.5. Algorithm

The details of the JS-MA algorithm are shown in Algorithm 1 consisting of three steps: clustering, SNP ranking, and stepwise evaluation. In clustering, the nearest neighbor chain algorithm repeatedly follows a chain of clusters, where each cluster is has the smallest distance to the previous one, until the number of clusters reaching user-defined parameter. In the second step, all SNPs are ranked based on Equation (6) and inserted into a size-limited descending list to select promising SNPs. In the last step, the χ^2 and the conditional χ^2 statistics are used to search for the significant, compact epistatic interactions.

3. EXPERIMENTAL DESIGN

In this section, we introduce the simulation design, including the definitions of 10 two-locus, 6 three-locus multi-disease models and the power metric. The other start-of-the-art methods we used to compared with JS-MA, including BOOST (Wan et al., 2010a), DAM (Guo et al., 2015), SEE (Sun et al., 2019), and SNPRuler (Wan et al., 2010b). Note that BOOST and SEE are designed for detecting gene-gene interactions, i.e., interactions between two loci.

3.1. Data Simulation

To evaluate the performance of JS-MA, we perform extensive simulation experiments using 10 two-locus disease models (Model 1–10) and 6 three-locus models (Model 11–16) with three groups, including 2 case and 1 control groups. Since there are three phenotype groups, we could have five different association types (ATs 1–5). Note that AT1 and AT2 are equivalent if case 1 and case 2 are interchangeable, which is the case in our simulation.

The odds tables describing these 16 models are in the **Supplementary Material**. For the two-locus models, models 1–4 are the base models, and the rest are derived from the base ones by combining two models or letting one case group follow the same distribution as the control group. For the four two-locus base models, we took the same parameters as in Wan et al. (2010a) and Guo et al. (2014a). More specifically, we have h^2 = 0.03 for Model 1, h^2 = 0.02 for Models 2, 3, and 4 and $p(D)$ = 0.1 for all four models. Minor allele frequencies (*maf*) are set to three levels: {0.1, 0.2, 0.4}. For the three-locus models, models 11 to 13 are the base models, the rest are derived using the same way as for the two-locus models. We set h^2 = 0.03 and $p(D)$ = 0.1 for Model 11, 12, and 13. The solved parameters μ and θ under different settings are provided in the **Supplementary Material**. The genotypes of unassociated SNP are generated by the same procedure used in previous studies (Guo et al., 2014a) with *mafs* sampled from [0.05, 0.5].

As introduced in the section 2.1, AT1 indicates the loci having different effects on the first case group compared to the other groups. AT2 indicates the loci having different effects on the second case group compared to the other groups. AT3 indicates the loci showing an identical effect on both case groups but different from the control group. AT4 indicates the loci with distinct effects on each group. We generate 100 replicas for each model, as well as for each *maf*. Note that some models do not have mathematical solution for μ and θ when $maf = 0.1$ or $= 0.2$. In this case, the power metric value is missing for all methods. Each simulated replica contains $M = 1,000$ SNPs. The sample sizes of two case groups and one control group are set to (500, 500, 1, 000) or (1, 000, 1, 000, 2, 000).

3.2. Statistical Power

The measure of discrimination power is defined as the fraction of 100 replicas on which the ground-truth associations are the top one signification epistatic interactions.

4. RESULTS AND DISCUSSION

In this section, we first present the type 1 error rate of JS-MA under the null model. And then we show the experimental results on the simulated datasets. We also present the results of JS-MA on two real GWAS datasets from WTCCC (Zeggini et al., 2007), i.e., Rheumatoid Arthritis (RA) and Type 1 Diabetes (T1D). Note that among these five approaches, only JS-MA and DAM are able to label the association types that we defined in section 2.1, and the rest methods can only report the interactions without information about the phenotype(s) on which they have genetic effects.

4.1. Null Simulation to Test Type I Errors

We examined the type I error rate for interactions with different number of SNPs, i.e., $d = 2, 3, 4$. We generated 1,000 null datasets for six settings, respectively. Specifically, we fixed the number of SNP to 1,000 and vary the number of samples in each group. The first four settings contained the following numbers of samples: $N1 = \{200, 200, 400\}$, $N2 = \{400, 400, 800\}$, $N3 = \{800, 800, 1, 600\}$, and $N4 = \{1, 600, 1, 600, 3, 200\}$, where the first two numbers indicated the sizes of two case groups, and the last number was the control group size. For the last two settings, using $N4$, we increased the number of SNP to 2,000 and 4,000. All SNPs were generated independently, with *maf* uniformly distributed in [0.05, 0.5]. Note that we set the significance level to 0.1 and applied the Bonferroni correction for multiple hypothesis

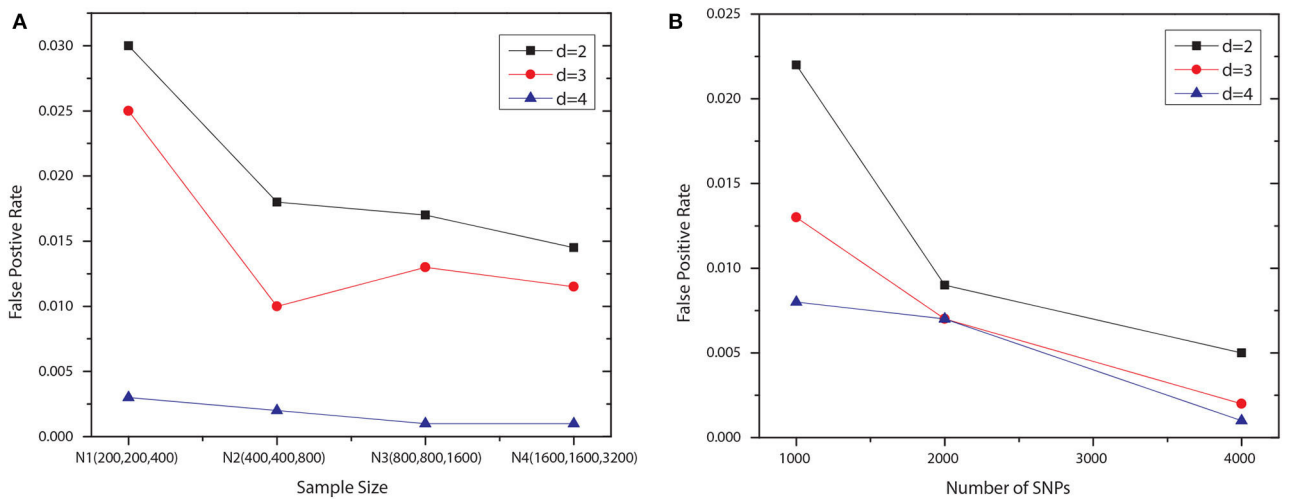


FIGURE 2 | False positive rates of JS-MA under null simulation. The plots in **(A,B)** show the false positive rates for different *d*s, sample sizes, and the numbers of SNP.

testing. The degree of freedom for Pearson's χ^2 test is $df = (|T| - 1)(|G| - 1)$, where $|T|$ denotes the number of disjoint set of groups for the association type $|T|$, and G is the set of genotypes given the SNP module. The degree of freedom for conditional χ^2 test is $|G'|(|T| - 1)(|G/G'| - 1)$, where G' is the set of genotypes given a subset of the SNP module, and G/G' denotes the set of genotypes for the rest SNPs. The results shown in **Figure 2** demonstrated that JS-MA can well control the type I error rate.

4.2. Simulation Experiments on Two-Locus Models

We tested the performance of JS-MA and four other methods on the datasets generated by two-locus models. The test results

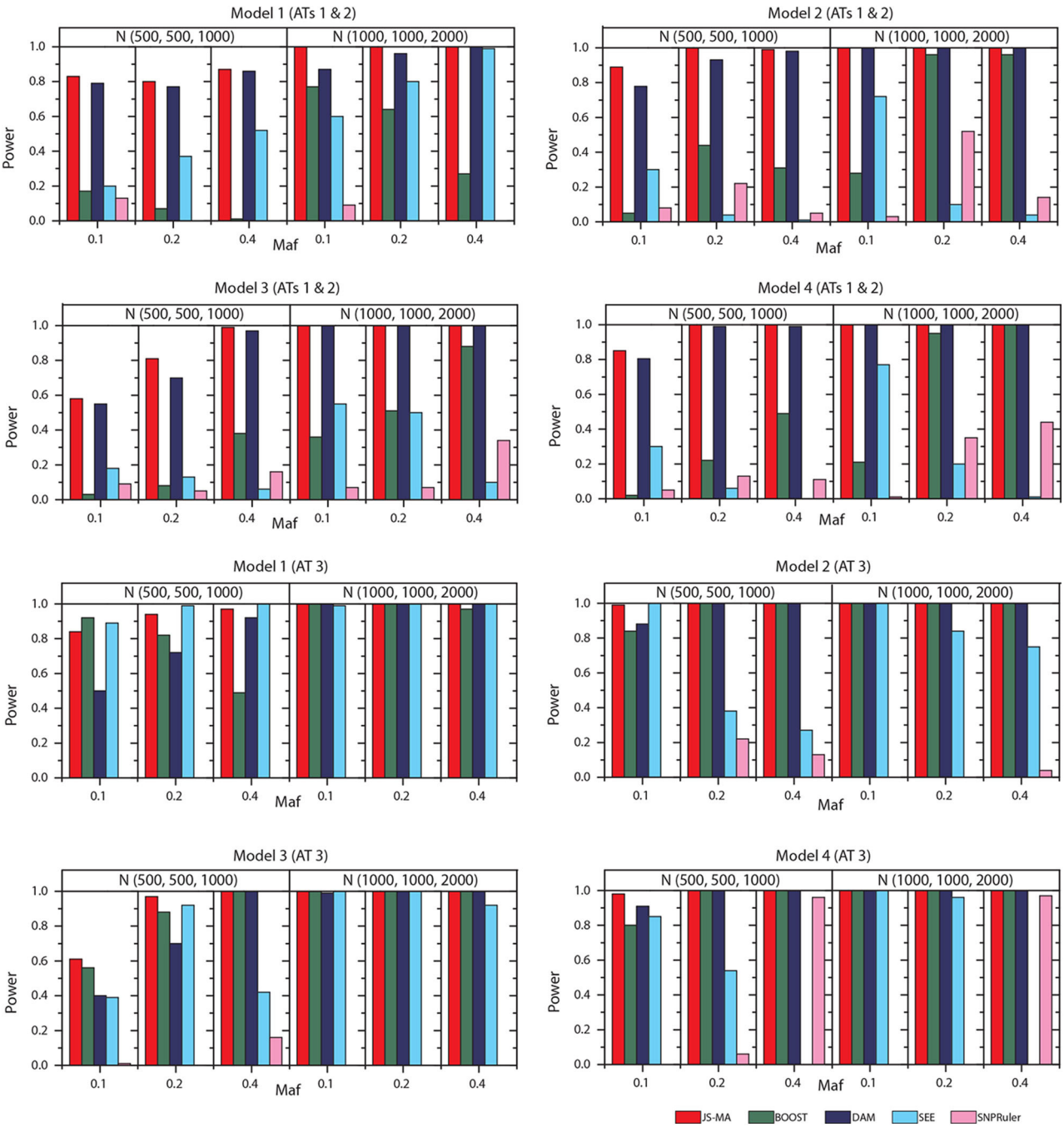


FIGURE 3 | Performance comparison between JS-AM, BOOST, DAM, SEE, and SNPRuler on the simulated two-locus models 1, 2, 3, and 4 for association types 1, 2, and 3.

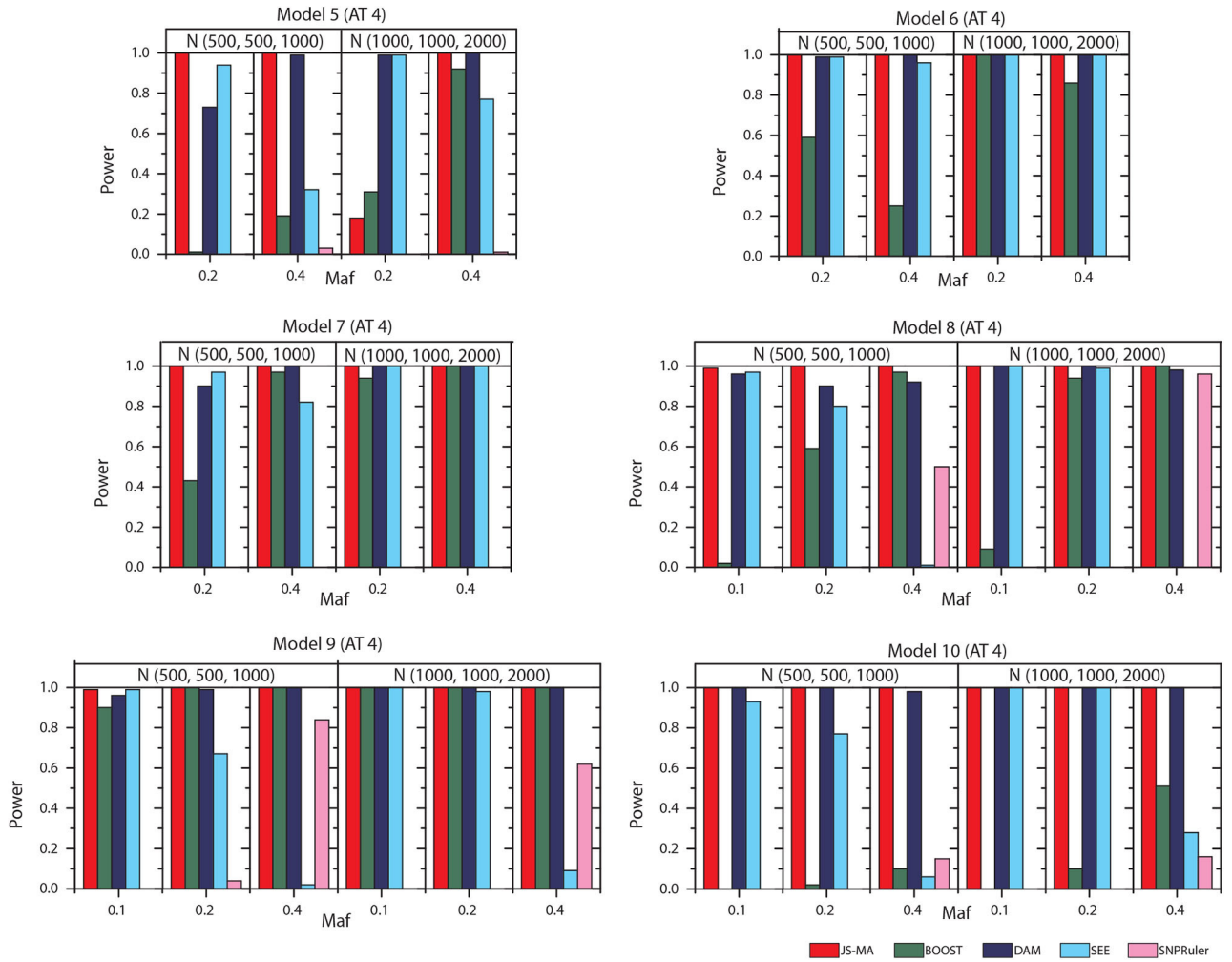


FIGURE 4 | Performance comparison between JS-AM, BOOST, DAM, SEE, and SNPRuler on the simulated two-locus models 5–10 for association type 4. Note that the models 5, 6, and 7 have no mathematical solution when *maf* = 0.1.

are illustrated in **Figures 3**, **4**. As we expected, the powers of all methods increased when the sample size increased from (500, 500, 1,000) to (1,000, 1,000, 2,000). For all models, the powers of JS-MA and SEE increased when the *maf* increased from 0.1 to 0.4. We do not observe a similar trend for BOOST, DAM, and SNPRuler. All models were more powerful for AT3 than ATs 1 and 2 because ATs 1 and 2 have some cases similar to controls, which makes it hard to locate the embedded interactions. Overall, the powers of JS-MA are higher compared to other methods except in a few cases where the power is comparable with others. For a more intuitive comparison, we adopt a concept, overall quality $q = 100 \times n_{correct}/n_{total}$ from (Guo et al., 2014a), where $n_{correct}$ is the number of datasets from which the method successfully detected the ground-truth interaction, and n_{total} is the total number of datasets. The overall quality of JS-MA, BOOST, DAM, SEE, and SNPRuler are 94, 50, 89, 51, and 11% for the sample size $(500, 500, 1,000)$, and 97, 78, 93, 71, and 13% for the sample size $(1,000, 1,000, 2,000)$, respectively. It showed that JS-MA achieved 3–5% better results than the second best.

4.3. Simulation Experiments on Three-Locus Models

The experimental results on models 11–16 are shown in **Figures 5**, **6**. In these experiments, BOOST and SEE were dropped because they cannot detect three-locus interactions. From **Figures 5**, **6** we can find that all three methods had nearly no power when the sample size is small. It is reasonable since a high-order interaction needs to have larger effect size for small sample size compared to large sample size. When the sample size was doubled, all three methods started to gain some power. Compared to the results from two-locus models, all the methods are not as powerful as before. In all settings, JS-MA is the most powerful approach. Using the same overall quality measurement introduced in the last section, JS-MA, DAM, and SNPRuler

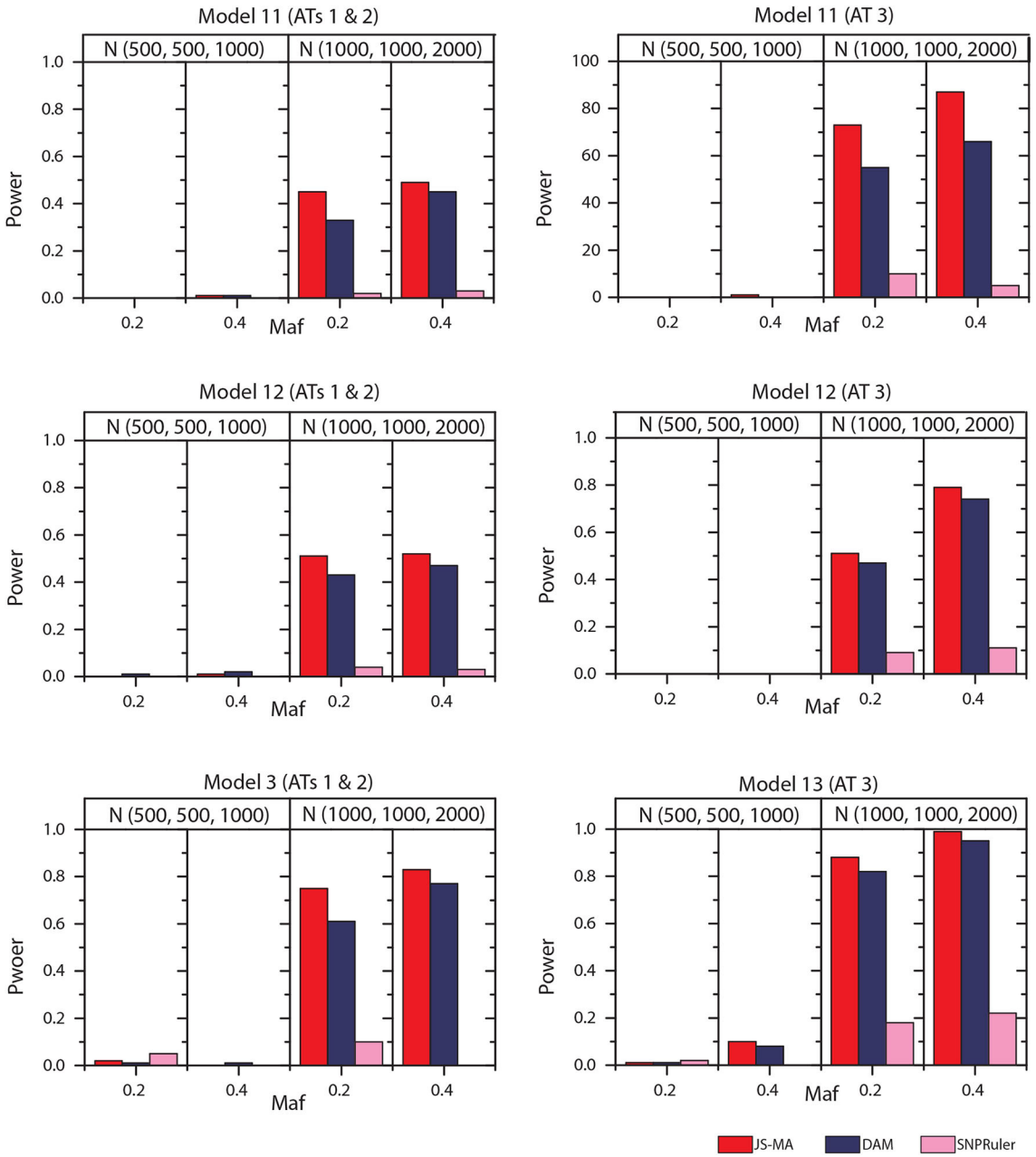


FIGURE 5 | Performance comparison between JS-AM, DAM, and SNPRuler on the simulated three-locus models 11, 12, and 13 for association types 1, 2, and 3.

reached 4, 3, and 1% for sample size (500, 500, 1,000), and 77, 70, and 9% for sample size (1,000, 1,000, 2,000), respectively.

4.4. Computation Efficiency

From a practical point of view, a challenging bottleneck of mapping multi-locus epistatic interactions in GWASs is the computational efficiency. Traditional tools for two-locus epistatic interaction detection usually take several days for a dataset with millions of SNPs using a standard desktop (Wan et al., 2010a). We measured the running time of JS-MA, BOOST, DAM, SEE, and SNPRuler on one computing node of an HPC system with a UNIX operating system, Intel Xeon E5-2699v4 Broadwell, and 128 GB memory. The results are shown in **Table 1**. Here, we set the target number of SNPs in an epistatic interaction to be two, and the rest of the parameters for each tool were left unchanged with default values. **Table 1** showed that JS-MA was faster than BOOST, DAM, and SNPRuler in most scenarios. The running time used by JS-MA did not increase as fast as SNPRuler and DAM did when the number of SNPs increased. Since SEE is a heuristic method, it used the least amount of time. However, its

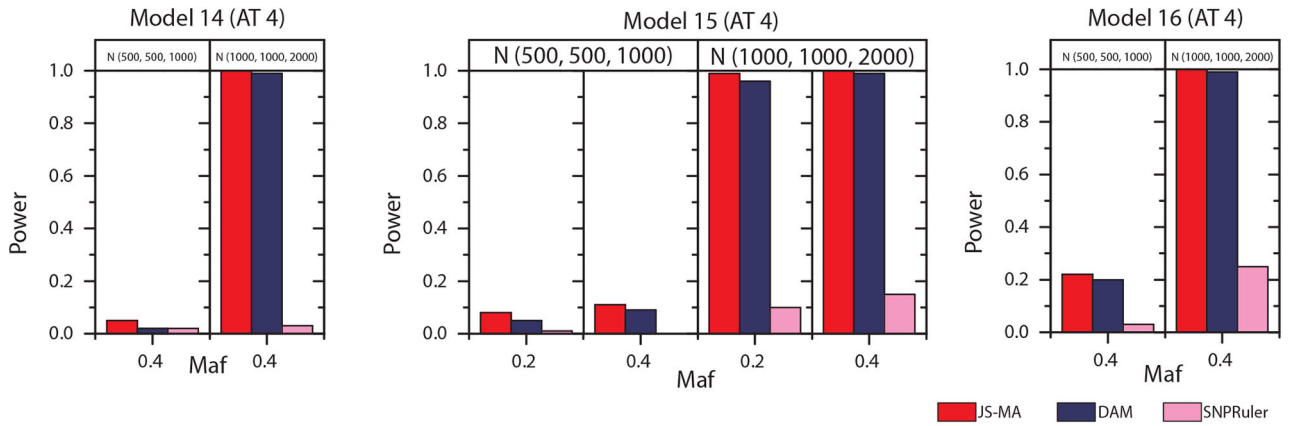


FIGURE 6 | Performance comparison between JS-AM, DAM, and SNPRuler on the simulated three-locus models 14, 15, and 16 for association type 4. Note that the models 14 and 16 have no mathematical solution when $maf = 0.2$.

TABLE 1 | Time comparison of JS-MA, BOOST, DAM, SEE, and SNPRuler (in seconds).

Data size	JS-MA	BOOST	DAM	SEE	SNPRuler
N = 6,000, M = 1,000	8	6	31	6	13
N = 6,000, M = 5,000	20	31	187	10	184
N = 6,000, M = 10,000	81	96	512	18	741

performance is not as good as the rest. We also measured the memory consumption for JS-MA. JS-MA used 10, 62, and 130 MB for 1,000, 5,000, and 10,000 SNPs, respectively. The majority of the consumed memory was used for storing the genotype data.

4.5. Experiments on The WTCCC Data

We employed JS-MA to analyze real data from the WTCCC Zeggini et al. (2007) for two common human diseases, i.e., Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D). There are 3999 cases and 3004 shared controls. We constructed a dataset with RA as case 1 and T1D as case 2. The procedure of quality control is the same as presented in Guo et al. (2014a). After the SNP filtration, the dataset contains 333,739 high-quality SNPs. By setting $f \times k = 100$ with $k = 10$ as the number of clusters, JS-MA finished the searching in 3 h using the same computing node, which was used in the computation time analysis. JS-MA reported some novel epistatic interactions. For example, (rs6679677, rs805301) was labeled as AT4, and its p-value is 6.2×10^{-120} from the χ^2 test. For this interaction, rs6679677, located on Chromosome 1, has been reported to be associated with both RA and T1D (Burton et al., 2007). The association between rs6679677 and T1D is due to a closely linked, potentially causal variant identified as rs2476601, which is also known as Arg620Trp (Smyth et al., 2008). Whereas, rs805301 is located inside gene BAG6 on Chromosome 6. BAG6 encodes a nuclear protein that forms a complex with E1A binding protein p300 and is required for the response to DNA damage. The SNP module (rs6679677, rs805301) shows different association effects on RA and T1D compared to the control group. Another interesting interaction is (rs200991, rs11171739) labeled as AT2, and its p-value is 6.7×10^{-26} from the χ^2 test. In this interaction, rs200991 is located on Chromosome 6 near the gene, HIST1H2BN, which encodes Histone H2B type 1-N. Histones play a central role in transcription regulation, DNA repair, DNA replication, and chromosomal stability. And rs11171739 has been reported to be associated with T1D (Burton et al., 2007). AT2 means the SNP module may not have a genetic effect on RA.

Algorithm 1: The JS-MA Algorithm.

Require: An $N \times (M + 1)$ matrix
Require: Number of clusters k, top f SNPs in a cluster
Read $N \times (M + 1)$ matrix file
Calculate the pairwise distance based on JS (Equation 3)
Initialize each SNP as a cluster
$n \leftarrow M$
while $n > k$ **do**
  Apply nearest neighbor chain algorithm
  $n- = 1$
end while
Initialize descending list $\mathbb{L}$ with length $f \times k$
for each SNP x **do**
  Calculate $Score(x)$
  Place x into $\mathbb{L}$ if $Score(x)$ is among top f SNPs
end for
Stepwise evaluate all possible SNP modules using SNPs in $\mathbb{L}$

JS-MA also reported some three-locus epistatic interactions. For instance, (rs6679677, rs377763, rs9273363) labeled as AT2 with p-value 1.3×10^{-116}. Both rs377763 and rs9273363 are located on Chromosome 6. rs377763 is near the downstream of gene NOTCH4, which is found to be associated with multiple sclerosis, a chronic inflammatory disease. rs9273363 is inside the gene HLA-DQA1, which plays a critical role in the immune system. The protein produced from the HLA-DQA1 gene binds to the protein produced from the MHC class II gene, HLA-DQB2.

Many studies have reported the MHC region on chromosome 6 with respect to infection, inflammation, autoimmunity, and transplant medicine (Lechler and Warrens, 2000; Wan et al., 2010a; Zhang et al., 2012). A four-locus interaction found by JS-MA is (rs10924239, rs17432869, rs7610077, rs11098422) labeled as AT4 with p-value 3.9×10^{-106}. rs10924239 is an intron variant of the gene KIF26B on Chromosome 1. KIF26B is essential for embryonic kidney development. rs17432869 is located on Chromosome 2 and inside gene LOC105373439, which is an RNA Gene and is affiliated with the ncRNA class. rs7610077 is located on Chromosome 3 and inside gene SNX4, which encodes a member of the sorting nexin family. rs11098422 is located on Chromosome 4 and inside gene NDST3, whose expression impacts the cardiovascular system. Validating the relationship between these SNP modules and RA and T1D is beyond the scope of this work. The significant enrichment of some genotype combinations from these SNP modules in both cases implies that they might interact and/or be associated with these two diseases.

5. CONCLUSION

The enormous number of SNPs genotyped in genome-wide case-control studies poses a significant computational challenge in the identification of gene-gene interactions. During the last few years, many computational and statistical tools are developed to find gene-gene interactions for the data containing only two traits, i.e., case-control groups. Here, we present a novel method, named "JS-MA," to address the computation and statistical power issues in multi-disease GWASs. We have successfully applied JS-MA to systematically simulated datasets and analyzed two real GWAS datasets. Our experimental results on both simulated and real data demonstrate that JS-MA is capable of detecting high-order epistatic interactions for multiple diseases at the genome-wide scale. It is worth mentioning that when JS-MA is used to analyze real data, quality control procedures are necessary because sequencing bias and genotyping bias could confound JS-MA by leading to false-positives. For example, the coverage bias caused by sequencing machines may have SNPs with low, uneven coverage. Thus, quality control is required to filter out unreliable SNPs.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

XG designed, implemented, and tested the proposed methods.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under award number R15LM013460. This work was part of the author's dissertation at Georgia State University. It was extended from the work presented as a pre-print at https://arxiv.org/abs/1811.07099.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.507038/full#supplementary-material

REFERENCES

Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678. doi: 10.1038/nature05911

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* 11, 2463–2468. doi: 10.1093/hmg/11.20.2463

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* 10, 392–404. doi: 10.1038/nrg2579

Guo, X. (2015). *Searching genome-wide disease association through SNP data* (Ph.D. thesis), Georgia State University, Atlanta, GA, United States.

Guo, X., Meng, Y., Yu, N., and Pan, Y. (2014a). Cloud computing for detecting high-order genome-wide epistatic interaction via dynamic clustering. *BMC Bioinformatics* 15:102. doi: 10.1186/1471-2105-15-102

Guo, X., Yu, N., Gu, F., Ding, X., Wang, J., and Pan, Y. (2014b). Genome-wide interaction-based association of human diseases-a survey. *Tsinghua Sci. Technol.* 19, 596–616. doi: 10.1109/TST.2014.6961029

Guo, X., Zhang, J., Cai, Z., Du, D.-Z., and Pan, Y. (2015). "Dam: a bayesian method for detecting genome-wide associations on multiple diseases," in *Bioinformatics Research and Applications*, eds R. Harrison, Y. Li, and I. Măndoiu (Norfolk, VA: Springer), 96–107. doi: 10.1007/978-3-319-19048-8_9

Guo, X., Zhang, J., Cai, Z., Du, D.-Z., and Pan, Y. (2017). Searching genome-wide multi-locus associations for multiple diseases based on bayesian inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 600–610. doi: 10.1109/TCBB.2016.2527648

He, Q., and Lin, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* 27, 1–8. doi: 10.1093/bioinformatics/btq600

Lechler, R., and Warrens, A. N. (2000). *HLA in Health and Disease* San Diego, CA: Academic Press.

Li, J. (2008). A novel strategy for detecting multiple loci in genome-wide association studies of complex diseases. *Int. J. Bioinform. Res. Appl.* 4, 150–163. doi: 10.1504/IJBRA.2008.018342

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory* 37, 145–151. doi: 10.1109/18.61115

Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., et al. (2011). Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genet.* 7:e1001338. doi: 10.1371/journal.pgen.1001338

Marchini,1, J., Donnelly,1, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 37, 413–417. doi: 10.1038/ng1537

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *arXiv* 1109.2378.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* 26, 354–359. doi: 10.1093/comjnl/26.4.354

Nelson, M., Kardia, S., Ferrell, R., and Sing, C. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res.* 11, 458–470. doi: 10.1101/gr.172901

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Front. Genet*. 6:285. doi: 10.3389/fgene.2015.00285

Peter, K., and Hunter, D. J. (2009). Genetic risk prediction: are we there yet? *N. Engl. J. Med*. 360, 1701–1703. doi: 10.1056/NEJMp0810107

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet*. 69, 138–147. doi: 10.1086/321276

Sabaa, H., Cai, Z., Wang, Y., Goebel, R., Moore, S., and Lin, G. (2013). Whole genome identity-by-descent determination. *J. Bioinform. Computa. Biol*. 11:1350002. doi: 10.1142/S0219720013500029

Smyth, D. J., Cooper, J. D., Howson, J. M., Walker, N. M., Plagnol, V., Stevens, H., et al. (2008). Ptpn22 trp620 explains the association of chromosome 1p13 with type 1 diabetes and shows a statistical interaction with hla class II genotypes. *Diabetes* 57, 1730–1737. doi: 10.2337/db07-1131

Sun, L., Liu, G., Su, L., and Wang, R. (2019). See: a novel multi-objective evolutionary algorithm for identifying snp epistasis in genome-wide association studies. *Biotechnol. Biotechnol. Equip*. 33, 529–547. doi: 10.1080/13102818.2019.1593052

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet*. 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., et al. (2010a). Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet*. 87, 325–340. doi: 10.1016/j.ajhg.2010.07.021

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., and Yu, W. (2010b). Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 26, 30–37. doi: 10.1093/bioinformatics/btp622

Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L. S., and Yu, W. (2010c). Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics* 26, 2517–2525. doi: 10.1093/bioinformatics/btq486

Wang, Y., Cai, Z., Stothard, P., Moore, S., Goebel, R., Wang, L., et al. (2012). Fast accurate missing snp genotype local imputation. *BMC Res. Notes* 5:404. doi: 10.1186/1756-0500-5-404

Wang, Y., Liu, G., Feng, M., and Wong, L. (2011). An empirical comparison of several recent epistatic interaction detection methods. *Bioinformatics* 27, 2936–2943. doi: 10.1093/bioinformatics/btr512

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., et al. (2017). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Brief. Bioinform*. 19, 700–712. doi: 10.1093/bib/bbw145

Xie, M., Li, J., and Jiang, T. (2012). Detecting genome-wide epistases based on the clustering of relatively frequent items. *Bioinformatics* 28, 5–12. doi: 10.1093/bioinformatics/btr603

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). Gboost: a gpu-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics* 27, 1309–1310. doi: 10.1093/bioinformatics/btr114

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., et al. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316, 1336–1341. doi: 10.1126/science.1142364

Zhang, J., Wu, Z., Gao, C., and Zhang, M. (2012). High-order interactions in rheumatoid arthritis detected by bayesian method using genome-wide association studies data. *Med. J.* 3, 56–66. doi: 10.3844/amjsp.2012.56.66

Zhang, Y., and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet*. 39, 1167–1173. doi: 10.1038/ng2110

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership