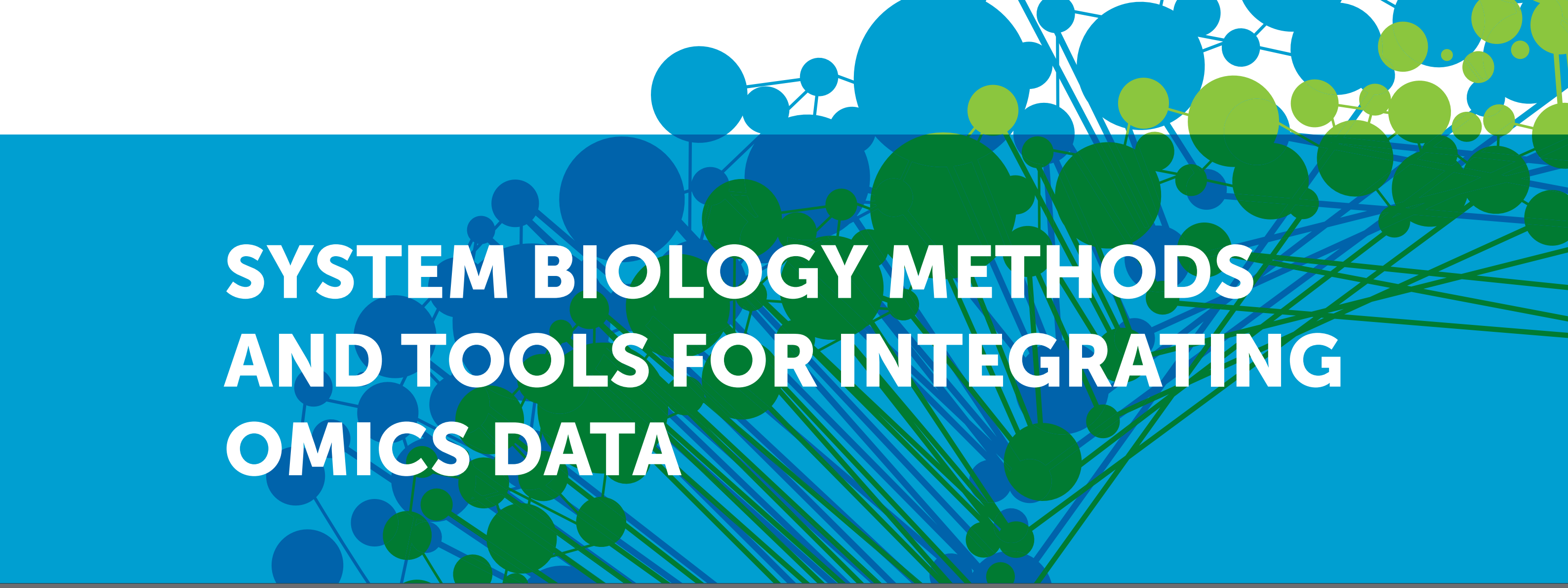


SYSTEM BIOLOGY METHODS AND TOOLS FOR INTEGRATING OMICS DATA

EDITED BY: Liang Cheng, Lei Deng and Mingxiang Teng
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-333-0
DOI 10.3389/978-2-88966-333-0

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: researchtopics@frontiersin.org

SYSTEM BIOLOGY METHODS AND TOOLS FOR INTEGRATING OMICS DATA

Topic Editors:
Liang Cheng, Harbin Medical University, China
Lei Deng, Central South University, China
Mingxiang Teng, Moffitt Cancer Center, United States

Citation: Cheng, L., Deng, L., Teng, M., eds. (2021). System Biology Methods and Tools for Integrating Omics Data. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-333-0

Table of Contents

frontiers in Genetics

EDITORIAL
published: 12 November 2020
doi: 10.3389/fgene.2020.563108

Editorial: System Biology Methods and Tools for Integrating Omics Data

Liang Cheng[1,2]*, *Lei Deng*[3]* and *Mingxiang Teng*[4]*

[1] NHC Key Laboratory of Molecular Probe and Targeted Theranostics, Harbin Medical University, Harbin, China, [2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [3] School of Computer Science and Technology, Central South University, Changsha, China, [4] Moffitt Cancer Center, Tampa, FL, United States

Keywords: OMICS data, data mining, machine learning, complex disease, system biology

OPEN ACCESS

Edited and reviewed by:
Simon Charles Heath,
Center for Genomic Regulation
(CRG), Spain

****Correspondence:***
Liang Cheng
chl198478@126.com
Lei Deng
leideng@csu.edu.cn
Mingxiang Teng
tengmx@gmail.com

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *17 May 2020*
Accepted: *06 October 2020*
Published: *12 November 2020*

Citation:
Cheng L, Deng L and Teng M (2020)
Editorial: System Biology Methods
and Tools for Integrating Omics Data.
Front. Genet. 11:563108.
doi: 10.3389/fgene.2020.563108

Editorial on the Research Topic

System Biology Methods and Tools for Integrating Omics Data

With the rapid evolution of sequencing technologies, it becomes more and more easy for researchers to analyze the expression level of molecules or variations in the genome, transcriptome, and proteome in wet labs. These technological innovations have advanced the life science community in terms of revealing disease risk factors such as gene variations or expressions, clinical phenotypes, etc. Accompanied by technological advances, significant amounts of sequencing data have been generated in the field to then be interpreted using novel data integration methods.

To this end, it is urgent to develop methods and tools to better utilize omics datasets in disease studies. One way would be to evaluate the associations between different diseases or sub-types by analyzing omics datasets across individual laboratories. e.g., LncRNAs biomarkers, associated with clinical sub-types and the prognosis of diffuse large B-cell lymphoma, were discovered and validated by re-annotating the probes and analyzing the data of multiple microarray platforms. Another way would be to reveal potential characteristics of diseases by integrating multi-level omics data. Gene targets of complex diseases could, for example, be predicted by integrating summary data from GWAS and eQTL studies. Integration of omics data by exploring computational tools is likely to be challenging for most biologists, as most tools require a certain level of computing knowledge one the part of the users to be operated optimally. It is consequently of great import to establish automated pipelines that combine these tools. In summary, the current challenge for understanding complex disease is to mine novel and precise characterization through the fusing of multi-level omics data using system biology approaches. Here, we organized a Research Topic on “System Biology Methods and Tools for Integrating Omics Data.” In total, 22 outstanding works were presented in this thematic issue, six of which have been highlighted as follows.

- Zhao et al. integrated GWAS and eQTL of brain data to identify SNPs and genes that are related to four types of strokes (ischemic stroke, large artery stroke, cardioembolic stroke, and small vessel stroke). They explored the genetic pathogenesis based on the loci, genes, gene expression, and phenotypes. There, 38 SNPs that affect expression of 14 genes were found to be associated with stroke. Among them, one gene was found for large artery stroke, six genes for cardioembolic stroke, and eight genes for small vessel stroke. To explore the effects of environmental factors on stroke, they further identified methylation susceptibility loci associated with stroke using mQTL. A total of 31 of the 38 eQTLs were also identified as mQTLs. In a short, this study explored the genetic pathogenesis of strokes.
- Zhou et al. carried out a comprehensive analysis of single-cell genomic copy number variations (CNVs) in VHL/PBRM1-negative Clear-cell renal cell carcinoma (ccRCC). Through functional enrichment analysis, they found that the amplified genes are significantly associated with cancer-related signaling transduction pathways. Besides, receptor protein tyrosine kinase (RTK) genes also showed widespread CNVs in cancer cells. In short, their studies indicated that the genomic CNVs in RTK genes and downstream signaling transduction pathways may be involved in VHL/PBRM1-negative ccRCC pathogenesis and progression.
- Hong and Wang designed a novel method, *Frin*, for studying genome evolutionary history. Phylogenetic tree and phylogenetic network are state-of-art ways for understanding the process of biological evolution. Since each taxon in a phylogenetic tree could have more than one parent, phylogenetic trees cannot capture the complexity of evolutionary information implicit in phylogeny. Hong and Wang presented a phylogenetic network-based method *Frin* to express genome evolutionary histories. Unlike the previous methods heavily relying on the order of input data, *Frin* unified the different input orders as the same dataset for different networks.
- Han et al. explored lncRNAs of Multiple Sclerosis (MS) by integrating the RNA-seq data from multiple studies. lncRNAs were deemed as important regulatory factors in MS pathogenesis. Current research has been limited by small sample sizes or heterogeneity among various tissues. RNA-seq has become a powerful approach to quantify the abundances of lncRNA transcripts. The authors collected MS-related RNA-seq data from a variety of previous studies, and integrated the data using an expression-based meta-analysis to identify differentially expressed lncRNA between MS patients and controls in all samples and sub-groups. Results showed that a potential important function of lncRNAs may be involved in the regulation of ribonucleoproteins and TNF cytokines receptors in MS.
- Gan et al. proposed a new approach, *TriPCE*, introducing a tri-clustering strategy to integrative pan-cancer epigenomic analysis. *TriPCE* can identify coherent patterns of various epigenetic modifications across different cancer types. To validate its capability, they applied *TriPCE* to analyze six important epigenetic marks among seven cancer types and identified significant cross-cancer epigenetic similarities. The results highlighted specific epigenetic patterns among the investigated cancers. The functional gene analysis further demonstrated strong relevance of studied gene sets with cancer development and revealed a consistent risk tendency among these investigated cancer types.
- Zeng et al. developed a hybrid deep neural network framework *4mcDeep-CBI*, aiming to identify 4mC sites. Preliminary extracted features were fed to the Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory network (BLSTM) to generate advanced features. Taking the advanced features as input, they designed an integrated algorithm to improve feature representation. Experimental results on a large new dataset showed that *4mcDeep-CBI* could achieve generally better performances when identifying 4mC sites compared to other state-of-art predictors.

Each study in the special issue was peer reviewed by two or three external reviewers. We would like to thank all the authors for contributing their work to our hot thematic issue and all the reviewers for their time and efforts. Finally, we would like to thank the Chief Editor and Editorial Office of Frontiers in Genetics for their support during the whole processes.

AUTHOR CONTRIBUTIONS

LC, LD, and MT conducted this topic issue and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The Tou-Yan Innovation Team Program of the Heilongjiang Province (2019-15); National Natural Science Foundation of China (61871160); Heilongjiang Province Postdoctoral Fund (LBH-TZ20); and Young Innovative Talents in Colleges and Universities of Heilongjiang Province (2018-69).

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: 29 May 2019
doi: 10.3389/fgene.2019.00474

Systems Chemical Genetics-Based Drug Discovery: Prioritizing Agents Targeting Multiple/Reliable Disease-Associated Genes as Drug Candidates

Yuan Quan[1†], *Zhi-Hui Luo*[2†], *Qing-Yong Yang*[1†], *Jiang Li*[1], *Qiang Zhu*[1], *Ye-Mao Liu*[1], *Bo-Min Lv*[1], *Ze-Jia Cui*[1], *Xuan Qin*[1], *Yan-Hua Xu*[3], *Li-Da Zhu*[1*] *and Hong-Yu Zhang*[1*]

[1] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China, [2] College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan, China, [3] Sci-meds Biopharmaceutical Co., Ltd., Wuhan, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Leyi Wei,
Tianjin University, China
Shikui Tu,
Shanghai Jiao Tong University, China
Zhenjia Wang,
University of Virginia, United States

Correspondence:
Li-Da Zhu
zhulinda@hotmail.com
Hong-Yu Zhang
zhy630@mail.hzau.edu.cn

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *07 March 2019*
Accepted: *01 May 2019*
Published: *29 May 2019*

Citation:
Quan Y, Luo Z-H, Yang Q-Y, Li J, Zhu Q, Liu Y-M, Lv B-M, Cui Z-J, Qin X, Xu Y-H, Zhu L-D and Zhang H-Y (2019) Systems Chemical Genetics-Based Drug Discovery: Prioritizing Agents Targeting Multiple/Reliable Disease-Associated Genes as Drug Candidates. Front. Genet. 10:474. doi: 10.3389/fgene.2019.00474

Genetic disease genes are considered a promising source of drug targets. Most diseases are caused by more than one pathogenic factor; thus, it is reasonable to consider that chemical agents targeting multiple disease genes are more likely to have desired activities. This is supported by a comprehensive analysis on the relationships between agent activity/druggability and target genetic characteristics. The therapeutic potential of agents increases steadily with increasing number of targeted disease genes, and can be further enhanced by strengthened genetic links between targets and diseases. By using the multi-label classification models for genetics-based drug activity prediction, we provide universal tools for prioritizing drug candidates. All of the documented data and the machine-learning prediction service are available at SCG-Drug (http://zhanglab.hzau.edu.cn/scgdrug).

Keywords: drug discovery, disease associated genes, drug targets, systems chemical genetics, machine learning

INTRODUCTION

Finding novel drugs or new uses for old drugs is one of the most important motivations of life sciences. Drug development is a costly process. The rich knowledge accumulated by modern life sciences is, thus, highly expected to reduce the attrition rate during drug development. From a chemical viewpoint, drugs exert therapeutic effects by inhibiting or activating one or more of the target genes/proteins associated with certain diseases. Therefore, gene-disease association information is crucial for drug discovery (Brinkman et al., 2006; Sanseau et al., 2012; Wang Z. Y. et al., 2012; Plenge et al., 2013; Okada et al., 2014; Nelson et al., 2015).

In life sciences, genetics is best dedicated to revealing gene-disease links. Thus, genetics makes great contributions to the pharmaceutical industry. For example, disease-associated genes identified by medical genetics constitute a promising source of drug targets (Brinkman et al., 2006; Sanseau et al., 2012; Wang Z. Y. et al., 2012; Plenge et al., 2013; Okada et al., 2014; Nelson et al., 2015). Moreover, the pathogenesis revealed by genetics is also of high value for drug discovery. If a disease arises from gain of function (GOF) mutation of a target gene, the corresponding drugs must be antagonists or inhibitors; while for a disease induced by loss of function (LOF) mutation of a gene, the targeted drugs must be agonists (Wang and Zhang, 2013).

Thousands of disease-associated genes have been identified by traditional Mendelian genetics and recently developed genome- and phenome-wide association studies (GWAS and PheWAS, respectively). However, nearly all studies attributed diseases to variations at a single genetic locus. Most diseases are caused by multiple pathogenic factors (Yildirim et al., 2007; Hopkins, 2008; Guney et al., 2016); thus, a majority of the identified links between diseases and single genetic variations are not strong enough to have therapeutic value. For example, only ∼5% of the drug-disease associations derived from PheWAS are supported by clinical evidence (Rastegar-Mojarad et al., 2015). Thus, to utilize the medical genetic information more efficiently in drug development, we should aim at multiple genes associated with certain diseases rather than a single pathogenic factor to identify potential drugs. To test this hypothesis, we retrieved the genes responsible for various disorders and collected the chemical agents targeting these genes. A comprehensive analysis on the relationships between agent activity/druggability and target genetic characteristics revealed that the agents targeting multiple pathogenic factors were more likely to show desired medicinal activities and to be clinically approved. The therapeutic potential of agents can be enhanced with the consolidation of genetic links between targets and diseases. These observations allowed us to predict agent activities using machine learning methods, which are definitely helpful to prioritize drug candidates.

RESULTS

Data Preparation and Validation

The information for agent-target interaction was obtained through retrieving Drug-Gene Interaction database (DGIdb) (Wagner et al., 2015), Therapeutic Target Database (TTD) (Qin et al., 2014), and DrugBank (Law et al., 2014). Only the clinically supported or approved activities of the agents were used in the present study, which were derived from DrugBank, TTD, and ClinicalTrials (Zarin et al., 2011; Law et al., 2014; Qin et al., 2014). The disease-associated gene information was derived from the following eight databases: Genetic Association Database (GAD) (Becker et al., 2004), Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2005), Clinvar (Landrum et al., 2014), Orphanet (http://www.orpha.net/consor/cgi-bin/index.php), DisGeNET (Piñero et al., 2015), INtegrated TaRget gEne PredItion (INTREPID) (Chen and Tian, 2016), GWASdb (Nelson et al., 2015), and The Human Gene Mutation Database (HGMD) (Wang X. et al., 2012) (**Figure 1**).

To facilitate the present analysis, a natural language processing tool MetaMap was used to convert disease terms of genes and indication annotations of agents to Unified Medical Language System (UMLS) concepts (Aronson, 2001), where the Medical Subject Headings (MeSH) thesaurus was selected as the vocabulary source of UMLS (Liu et al., 2014). Using the disease classes provided by pharmaprojects (Similarity threshold: 0.75) (Mcinnes et al., 2009), the chemical agents were indicated for treating 667 disease classes and the disorder-related genes were associated with 703 disease classes (**Figure 1**). All of the data are freely available at SCG-Drug (http://zhanglab.hzau.edu.cn/scgdrug).

Data validation was performed by the following analyses. First, we assessed the reliability of the gene-disease pairs by examining whether similar diseases cover similar gene sets. The disease similarity was measured using UMLS::similarity (Mcinnes et al., 2009); the disease gene set distance was calculated using the Tanimoto coefficient (see Methods). As shown in **Figure 2A**, a definite correlation exists between disease similarity and gene set distance. That is, if two diseases exhibit similar symptoms, then these diseases tend to involve similar genes, validating the identified gene-disease pairs. Then, we used a similar method to evaluate the quality of agent-disease pairs. A good correlation was observed between disease similarity and agent set distance (**Figure 2A**), supporting the reliability of agent-disease pairs. Therefore, one can infer the activities of agents through their target-associated genetic diseases, provided the agents and the targets are truly linked. As illustrated in **Figure 2B**, for the agents in TTD, DGIdb, and DrugBank, 4.1, 4.7, and 5.3% of their genetics-implicated activities are supported by clinical trials, respectively, comparable with the PheWAS-based activity prediction efficiency (Rastegar-Mojarad et al., 2015). However, if the agents were randomly assigned with targets (for 10,000 times), the clinically supported activities derived from genetic predictions are significantly rarer than those from real agent-target pairs (**Figure 2B**, $P < 10^{-4}$). This 10,000-permutation test validates the agent-target associations.

Dependence of Agent Activity/Druggability on Target Quantity

Based on the validated data, we can investigate how the agent activity/druggability depends on the target characteristics. As illustrated in **Figure 3**, for the agents targeting a single disease gene, 3.0% of genetics-derived activities are supported by clinical test and only 0.6% are clinically approved (**Table S1**). For agents targeting two disease-associated genes, 4.1% of genetics-implicated activities are clinically supported, and 1.5% have been introduced to the market (**Table S1**). The clinically active ratio of agents culminates to 26.7%, and the approval ratio is up to 11.4%, when the agents targeting tens of disorder genes. Together, the therapeutic potential of agents increases steadily with increasing number of targeted disease genes (**Figure 3**).

Drug action is usually considered a specific process. It is thus of apparent interest to investigate the molecular mechanisms underlying the promiscuity of the multi-target agents. Considering the fact that human genes generate a large number of paralogs during evolution, a primary explanation is that the multiple targets covered by the agents have similar sequences and functions. Indeed, the sequences for target pairs hit by the agents are more similar than those randomly selected from the target set ($P = 2.20 \times 10^{-16}$, Wilcoxon rank-sum test) (**Figure 4A**), where the needle program of EMBOSS package (Rice et al., 2000) was used to do pairwise alignments. Furthermore, it was found that the target pairs covered by the agents are significantly enriched with paralogs (4.72% (2,602 of 55,110), derived from Ensemble database), compared with the randomly combined target pairs (0.10% (4,029 of 3,955,078), *P*

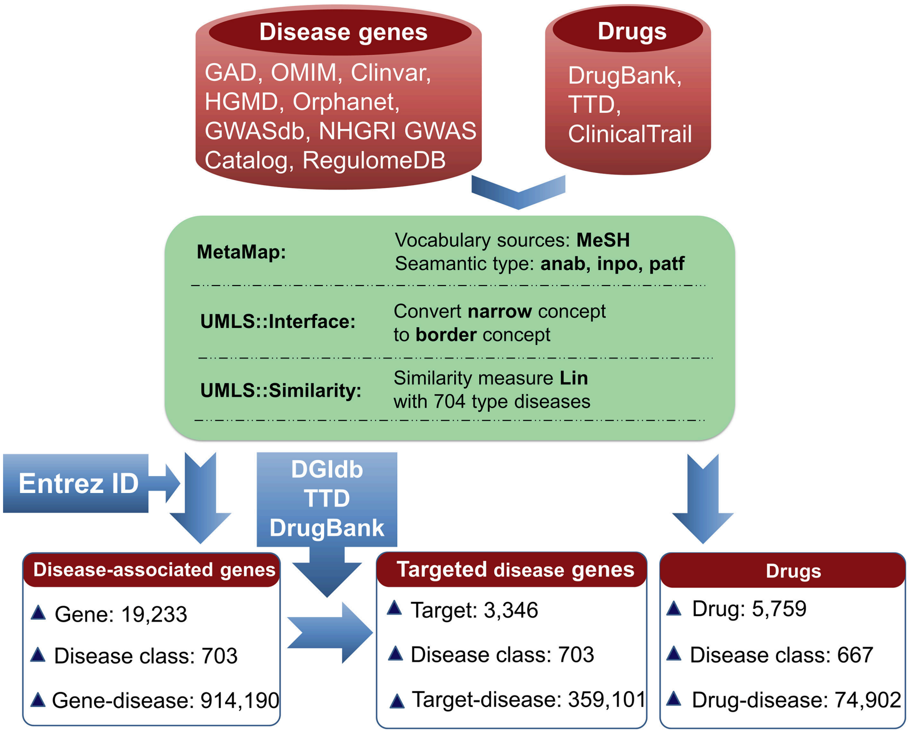


FIGURE 1 | Pipeline for data processing. Disease-associated genes were derived from eight databases. Agent activities were obtained from TTD, DrugBank, and ClinicalTrials. The disease terms of genes and the indication annotations of agents were uniformed to UMLS concepts using MetaMap. Using the disease classes provided by pharmaprojects (Similarity threshold: 0.75), 703 types of diseases for 19,233 genes were identified, resulting in 914,190 gene-disease pairs. Through searching DGIdb, TTD, and DrugBank, 3,346 genes were targeted by 14,558 agents. 3,346 targets were associated with 703 diseases, resulting in 359,101 gene-disease pairs; 5,759 agents were indicated for treating 667 diseases, resulting in 74,902 agent-disease pairs.

~ 0, hypergeometric test). Besides, the GO-based Czekanowski–Dice distances (Ovaska et al., 2008) of the gene pairs targeted by the agents are evidently smaller than those of randomly selected target pairs ($P = 2.20 \times 10^{-16}$, Wilcoxon rank-sum test) (**Figure 4B**). These observations not only support the evolutionary explanation to the molecular basis of multi-target drug action, but also provide useful clues to addressing the concerns about the side effects of promiscuous agents.

Despite the achievements of multi-target strategy for drug discovery, questions concerning security remain, as the tendency to act on multiple genes may increase the probability of inducing adverse effects. The present analyses indicate that these agents prefer to target genes with similar sequences and functions, namely paralogs, which means that the agent-targeting process is not so random that it will constrain the agent activities into a relatively narrow range. This is definitely beneficial to alleviate the side effects of multi-target agents and thus helpful to enhance their druggability.

Furthermore, we analyzed the chemical genetic data recorded in connectivity map (cMap) (Lamb et al., 2006). The cMap comprises 7,056 gene expression profiles for five human cell lines treated with 1,309 agents. Using the biclustering approach FABIA (factor analysis for bicluster acquisition), we have generated 49 gene modules for cMap data, establishing links between gene modules and chemical agents (Xiong et al., 2014). Therefore, each agent has a gene module profile, and the promiscuity of the agent increases with the increasing number of modules the agent covers. As shown in **Figure 5A**, with the increase of targets, the agents indeed cover more gene modules, supporting the opinion that multi-targeted agents have a higher risk of yielding unwanted effects. However, the druggability analysis indicated that with the increasing number of targets, the drug approval ratio does not decrease but rather increases slightly (**Figure 5B**). Moreover, if only disease-associated genes are considered, the drug approval ratio increases evidently with the increase of targeted gene number (**Figure 5C**). This observation strongly suggests that despite the enhanced risk in side effects, multi-targeted agents are still very promising in drug development.

Dependence of Agent Activity/Druggability on Target Quality

Besides the quantity of agent targets, their quality also influences the medicinal potential of agents in principle. Our prior study has revealed that the agents targeting "top genes" have higher therapeutic potential (Quan et al., 2018), where "top genes" were defined as those tightly associated with certain diseases. Four disease-gene databases, i.e., AlzGene (Bertram et al., 2007), SzGene (Allen et al., 2008), PDGene (Lill et al., 2012), and MSGene (Lill et al., 1994), provide "top genes" annotations for

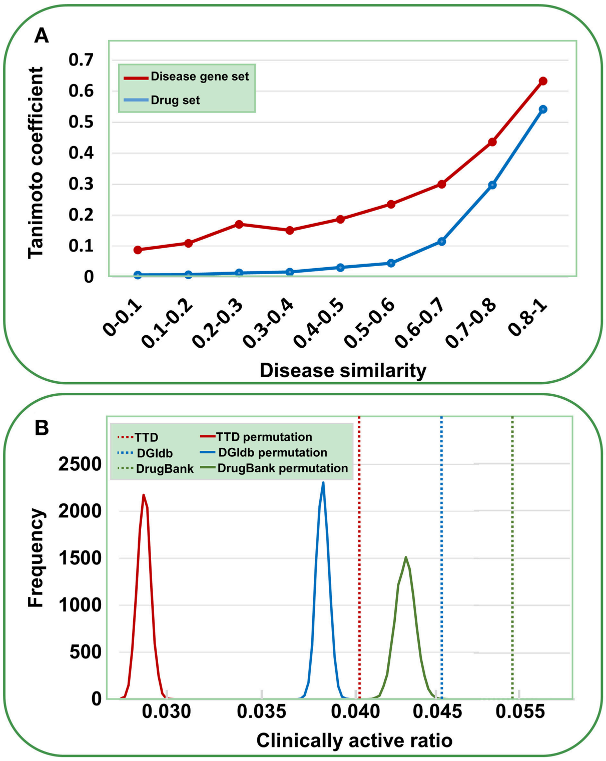


FIGURE 2 | Validation of gene-disease pairs, agent-disease pairs and agent-target pairs. **(A)** Correlations between disease similarity and disease gene set distance or drug set distance. The disease similarity was measured using the UMLS::similarity, and the disease gene set or drug set distance was characterized by Tanimoto coefficient. **(B)** Clinically active ratios of genetics-implicated agent activities. The red, brown, and green vertical dashed lines indicate the clinically active ratios derived from real agent- target pairs in TTD, DGIdb, and DrugBank, respectively. The curves show the clinically active ratio frequency distributions for 10,000 random permutations of agent-target pairs.

Alzheimer's disease, schizophrenia, Parkinson's disease, multiple sclerosis, respectively. From DGIdb, TTD and DrugBank, we retrieved 3,692 agents targeting the genes including "top genes" contained in these four databases (**Table S2**). As illustrated in **Figure 6**, multi-target agents exhibit higher medicinal potential than single-target counterparts, consistent with the above observations. Next, for the agents covering "top genes," their genetics-derived activities are more likely to be supported by clinical evidence and be clinically approved (**Figure 6** and **Table S2**), indicating the importance of target quality in genetics-based drug discovery.

However, only a few genetic databases contain quality information for disease genes. Considering the above finding that multi-target agents usually hit paralogs, we speculated that ohnolog genes, i.e., paralogs generated by whole genome duplication, may be used as "top genes" instead. Ohnolog genes have been recognized to significantly enrich disease genes, compared with other paralog genes, because of their strong dosage balance (Makino and Mclysaght, 2010; McLysaght et al., 2014; Xie et al., 2016; Sekine and Makino, 2017).

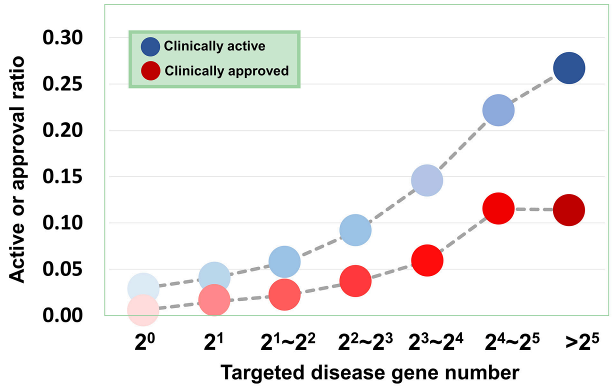


FIGURE 3 | Dependence of agent activity/druggability on target quantity. Therapeutic potential of agents increases with increasing number of targeted disease genes.

As illustrated in **Figure 7**, the agents covering disease-associated ohnolog genes indeed exhibit higher approved potential ($P < 1.09 \times 10^{-61}$, hypergeometric test), suggesting that disease-associated ohnolog genes can be regarded as "top genes" to some extent. This finding is very useful in establishing the machine-learning models for drug activity prediction (see below for details).

Target Quality Evaluation and Druggability Score of Disease Genes

Eight disease gene databases (including Clinvar, OMIM, HGMD, Orphanet, GWASdb, INTREPID, GAD, and DisGeNET) are used in the present study. The target quality of each database must be different, which stimulated our interest to do an evaluation by comparing the clinically supported ratio of genetics-implicated agent activities derived from eight databases. The results showed that target genes of Clinvar have the highest quality, in which 16.52% of genetics-based activity predictions are supported by clinical test. The target quality (measured by clinically active ratio) of other databases declines in the order: OMIM (15.01%), HGMD (14.09%), Orphanet (13.62%), GWASdb (10.53%), INTREPID (7.08%), GAD (5.75%), and DisGeNET (4.14%) (**Figure 8** and **Table S3**). This observation inspired us to propose a parameter for quantitatively measuring the druggability of disease genes. First, the genes derived from different databases were given different quality scores, with the highest-quality database (i.e., Clinvar) being assigned with the highest score (eight points), while the lowest (i.e., DisGeNET) with the lowest score (one point). Then, the scores were summed up for each disease gene to define its druggability (**see Methods**). The higher the score is, the more druggable the disease gene. Apparently, a gene may have different scores for different diseases.

This scoring system is validated by the following observations. First, for the disease genes with higher druggability scores, the genetics-implicated activities of agents are more possible to be clinically supported and approved (**Figure 9** and **Table S4**). Considering the correlation between gene druggability and pathogenicity (Plenge et al., 2013; Quan and Zhang, 2016),

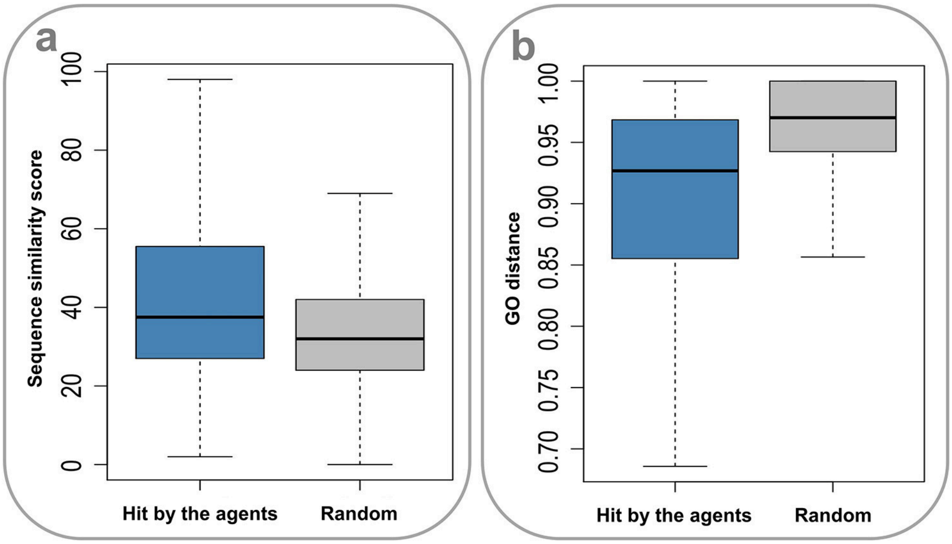


FIGURE 4 | Sequence similarity and GO distances of gene pairs targeted by the multi-target agents. **(A)** The sequences for target pairs hit by the agents are more similar than those randomly selected from the target set (*P* = 2.20 × 10–16, Wilcoxon rank-sum test). **(B)** The GO-based Czekanowski–Dice distances of the gene pairs targeted by the agents are evidently smaller than those of randomly selected target pairs (*P* = 2.20 × 10–16, Wilcoxon rank-sum test).

it is inferred that druggability score is also appropriate for characterizing gene-disease links. Indeed, the "top genes" derived from AlzGene, SzGene, PDGene, and MSGene, which are tightly connected with diseases, exhibit much higher druggability scores than other genes with the same pathogenic annotations ($P = 2.51 \times 10^{-52}$, Wilcoxon rank-sum test) (**Figure 10**). Therefore, each disease can be characterized by the corresponding scored genes, constituting a gene profile pertinent to the disease. Different diseases can be compared through calculating Spearman's rank correlation between their gene profiles. It is interesting to notice that the diseases exhibiting similar gene profiles display similar symptoms measured by UMLS::similarity (**Figure 11**), validating the scoring system in characterizing gene-disease links. Together, it is concluded that druggability score can be used to measure target quality and genetic links between genes and diseases, which is of great value in drug activity prediction by machine-learning models.

Agent Activity Prediction With Multi-Label Classification Model

The above analysis implied that it is possible to establish drug-activity prediction models based on the genetic information of drug targets. Since a drug is usually associated with multiple activities for diseases and a disease could be treated by multiple drugs, drug-activity prediction problem can be considered as a multi-label classification task. In this paper, we adopted a method of multi-label k-nearest neighbor (MLKNN) which can construct high-accuracy multi-label prediction models for drug-activity prediction (Zhang and Zhou, 2007; Wen et al., 2015).

First, we investigate a variety of features to represent the characters of druggability. Considering that various features may bring diverse information as well as noise, we adopt ensemble learning method to select suitable features to build the models (Lee and Soo, 2013; Yang et al., 2014; Zhang et al., 2015). Considering that agents targeting multiple disease genes, in particular "top disease genes" and genes with high druggability scores, tend to show high therapeutic potential (**Figures 3**, **6**, **7**, **9**), we rationally selected four parameters to build the models. The first parameter characterizes the overall score of genes responsible for certain diseases within drug targets, and the second parameter is the normalized average value of the overall score. The third and fourth parameters describe the absolute number and relative ratio of ohnologous disease genes (serving as "top genes") within drug targets, respectively (**see Methods**).

Representation of drug labels is a crucial step in multi-label learning. An agent-disease pair was regarded as a positive, if the drug hits one or more disease genes and is indicated for treating this disease. An agent-disease pair was regarded as a negative, if the drug targets one or more disease genes but is not annotated for controlling this disease. As a result, a total of 74,902 positives covering 5,759 agents and 667 diseases, and 3,778,517 negatives were selected.

Given a dataset of n drugs denoted as $\{(x_i, y_i)\}_{i=1}^{n}$, x_i and y_i are the *p*-dimensional feature vector and *q*-dimensional disease vector for the *i*th drug, respectively. Our goal is to build the functional relationship $Y = F(X) : 2^p \rightarrow 2^q$ between exploratory variables (feature vector) and target values (agent-activity vector) for multi-label learning.

First, four MLKNN models were constructed based on four features. Then, each model was evaluated by the internal 5-fold cross validation on the training data. As a result, five MLKNN models were built based on five internal folds and selected features. The final prediction result is the average and standard deviation scores of outputs by five MLKNN models. At last, we used the ensemble learning method to combine four features and generate high-accuracy prediction models **(see Methods**).

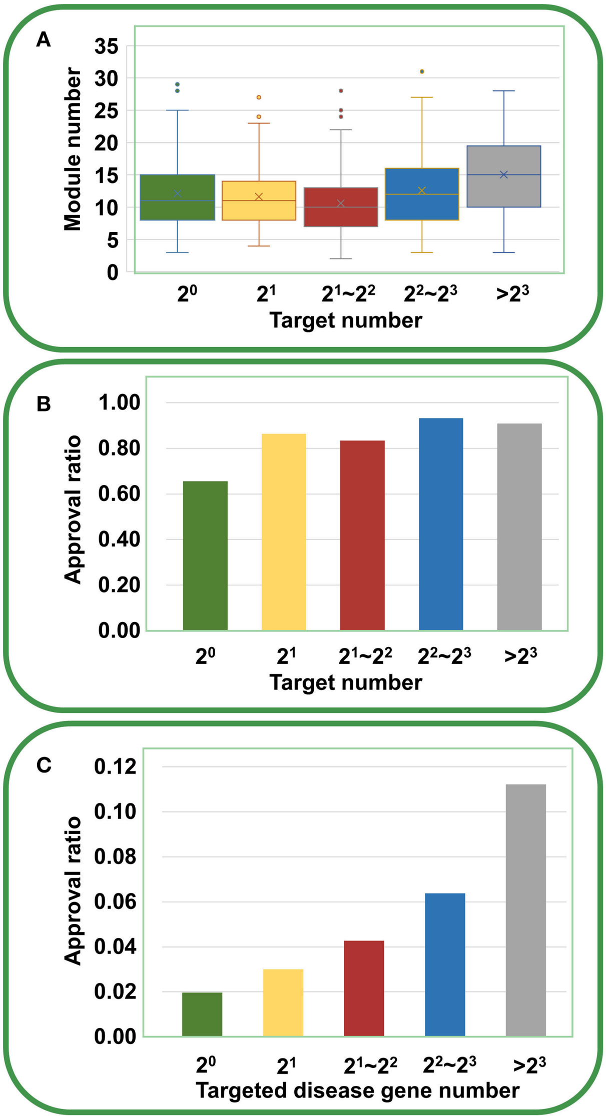


FIGURE 5 | Relationships between druggability and target number of agents derived from cMap. **(A)** With the increasing number of targets, the agents cover more gene modules (ANOVA: $P = 1.94 \times 10^{-9}$). **(B)** With the increasing number of targets, the drug approval ratio increases slightly. **(C)** If only disease-associated genes are considered, the drug approval ratio rises evidently with the increase of targeted gene number.

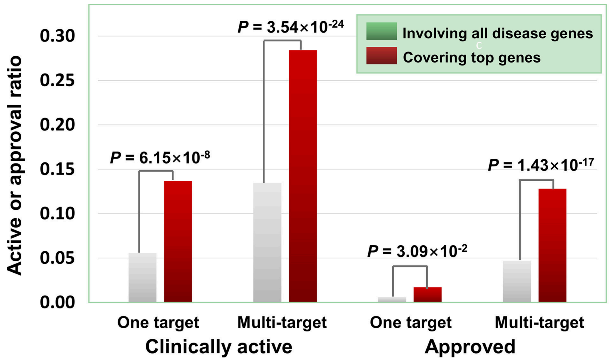


FIGURE 6 | Effects of top genes on the clinically active/approval ratio of agents. The top genes were derived from AlzGene, SZGene, PDGene, and MSGene. From DGIdb, TTD and DrugBank, we retrieved 3,692 agents targeting the genes contained in the four databases, of which 726 targeted at least one top gene. The results show that for the agents covering top genes, their genetics-implicated activities are more likely to be supported by clinical trials and to be clinically approved (*P*-values were calculated using the hypergeometric test).

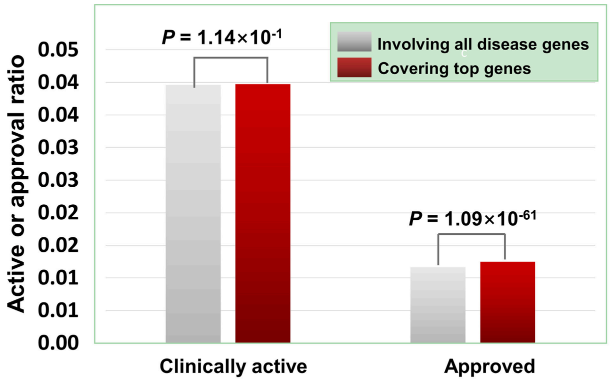


FIGURE 7 | Effects of disease-associated ohnolog genes on the clinically active/approval ratio of agents. A total of 7,294 ohnolog genes were obtained from Makino and Mclysaght's work31, in which 5,265 genes were disease-associated. Searching DGIdb, TTD and DrugBank revealed that 4,058 agents targeted 1,164 of the 5,265 ohnolog genes. The results show that for the agents covering disease-associated ohnolog genes, their genetics-derived activities are more likely to be supported by clinical evidence and be clinically approved (*P*-values were calculated using the hypergeometric test).

The performance of assembled classifier for agent-activity prediction is shown in **Figure 12**. For a 5-fold stratified cross-validation with a 1,000 repeat, MLKNN displays the best performance (**Table S5**). By inputting the 5,759 original agents and associated targets into the models (where the threshold of predictive value was set to 0.5), 11,649 activities were predicted. 67.01% of the predicted activities are supported by clinical trials, and 14.52% have been approved, which are much higher than the overall ratio of genetics-implicated clinical activity and approved indication (3.96 and 1.16%, respectively).

To examine for which kind of diseases the predictions are most relevant, we compared the clinically active/approval ratio of the predicted results for various diseases. It was found that, leukemia and lymphoma have the most predictions (**Table S6**). To demonstrate the usefulness of the present method, we tested the predicted anti-leukemia agents by cytotoxicity experiment. Using our models, 809 agents were predicted to have anti-leukemia potential, of which 550 (67.99%) have been validated by prior clinical tests. Thus, it is intriguing to examine the anti-leukemia potential of the rest 259 agents. 14 of 259 agents are commercially available, which were evaluated by K562 (chronic myeloid leukemia-derived cancer cell line) cytotoxicity assays. The results show that 10 agents (71.43%) can inhibit the growth of K562 efficiently (**Figure 13**) (**Table S7**), with

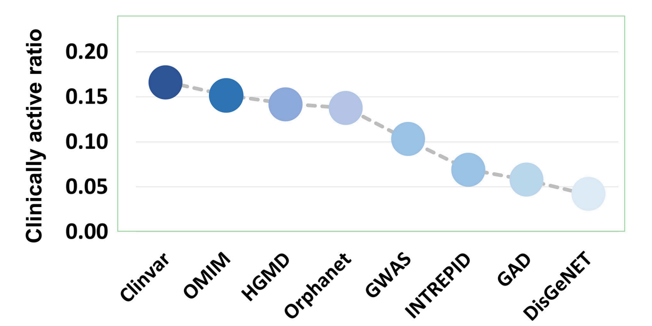


FIGURE 8 | Clinically active ratio of genetics-implicated agent indications derived from different disease gene databases.

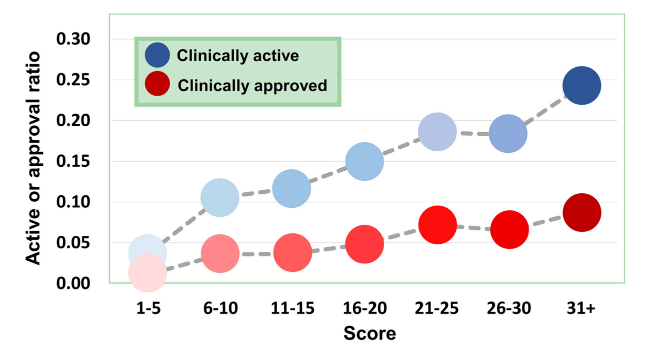


FIGURE 9 | Dependence of agent activity/druggability on target quality. With the increase of druggability scores of target genes, the therapeutic potential of corresponding agents also increases.

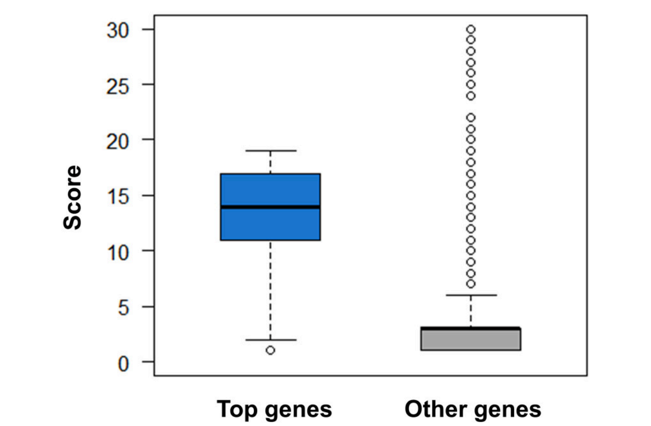


FIGURE 10 | Comparison of druggability scores for top genes derived from AlzGene, SzGene, PDGene, MSGene, and ordinary genes with the same pathogenic annotations. The top genes exhibit evidently higher scores than other genes ($P = 2.51 \times 10–52$, Wilcoxon rank-sum test).

IC50 values ranging from 0.106 (saracatinib) to 111.2 μM (veliparib) (**Table S7**).

To facilitate the use of the machine-learning prediction models, we developed a web server SCG-Drug (Systems Chemical Genetics-Drug, http://zhanglab.hzau.edu.cn/scgdrug) that allows a quick and intuitive access to the background information and predicted results. Currently, SCG-Drug contains 5,759 agents, 703 diseases and 19,233 genes derived from various databases. By inputting the target information of any agents into SCG-Drug, one can use the established machine-learning models to predict the potential activities of the agents. The SCG-Drug web interfaces allow users to explore medicinal information related to a given drug, disease or gene through four interfaces in "Analysis" page: "Drug", "Batch prediction," "Disease," and "Gene." The "Drug" interface allows users to submit a single drug to retrieve target genes and potential activities of the query drug. For example, when a user submits a single drug that was shown in the dropdowns, the drug will be searched in the database directly. If it is unable to find any matches for the search term, the user will be asked to input the corresponding target genes of the drug. Then, the system will call the prediction module. Alternatively, the system allows the user to upload a file on the "Batch prediction" interface, in which an agent and corresponding targets are in a single row and the terms in each row are separated by tabs, along with an email address to which the predicted activities of the agents will be sent. Offline prediction automatically starts, and the predicted results will be sent to the user via e-mail. The "Disease" interface allows users to obtain relevant disease genes with druggability score, and database source by querying standardized disease descriptions of MeSH. The "Gene" interface allows users to explore gene-related diseases (with druggability score) and drugs only by submitting a gene name or an Entrez ID, which have been documented in the server. In addition, users can obtain the information for documented drugs (with normalized indications) and targets/genes (with normalized disease descriptions) from "Download" page. The data and the machine-learning models will be updated regularly.

DISCUSSION

Selecting agents with desired activities and high druggability from an infinite chemical space is a fundamental task for drug development. Previous studies have revealed that genetic disease genes can provide valuable clues for drug activity prediction and druggability assessment (Brinkman et al., 2006; Sanseau et al., 2012; Wang Z. Y. et al., 2012; Plenge et al., 2013; Wang and Zhang, 2013; Okada et al., 2014; Nelson et al., 2015). However, these studies are limited to single-drug-single-target paradigm. Because most complex diseases are caused by multiple pathogenic factors, it is reasonable to speculate that targeting multiple disorder factors will better navigate the drug space. In this study, by a comprehensive analysis, we clearly indicate that aiming at multiple disease genes is helpful to prioritize drug candidates with promising activities and high druggability. Additionally, the strengthened genetic links between target genes and diseases are helpful to improve the medicinal potential of drug candidates. The drug-gene interaction information is expected to be rapidly accumulated through emerging techniques in chemical biology. However, the identification of reliable

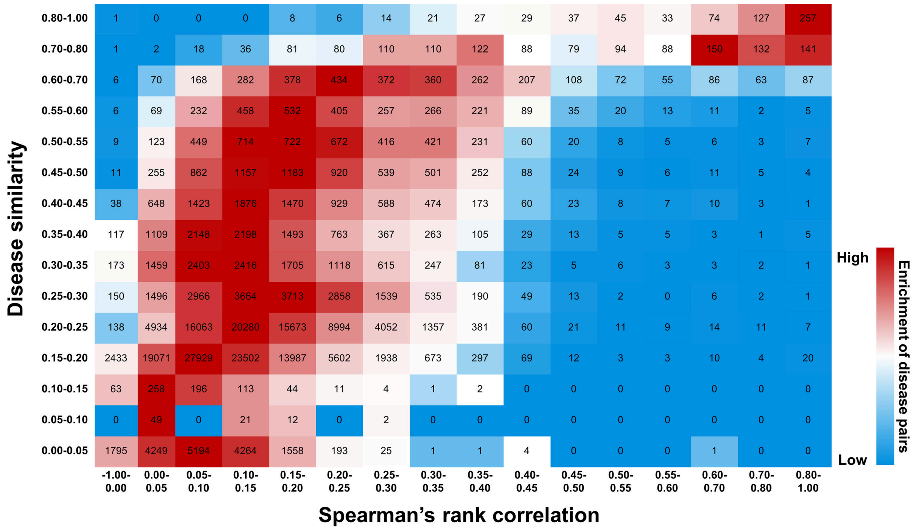


FIGURE 11 | Efficiency of druggability scoring system in characterizing gene-disease links. Diseases exhibiting similar gene profiles, calculated by Spearman's rank correlation, display similar symptoms measured by UMLS::similarity. The number of disease pairs is shown in the box. The color exhibits enrichment of the number in each row, with red representing the strong enrichment and blue representing the weak.

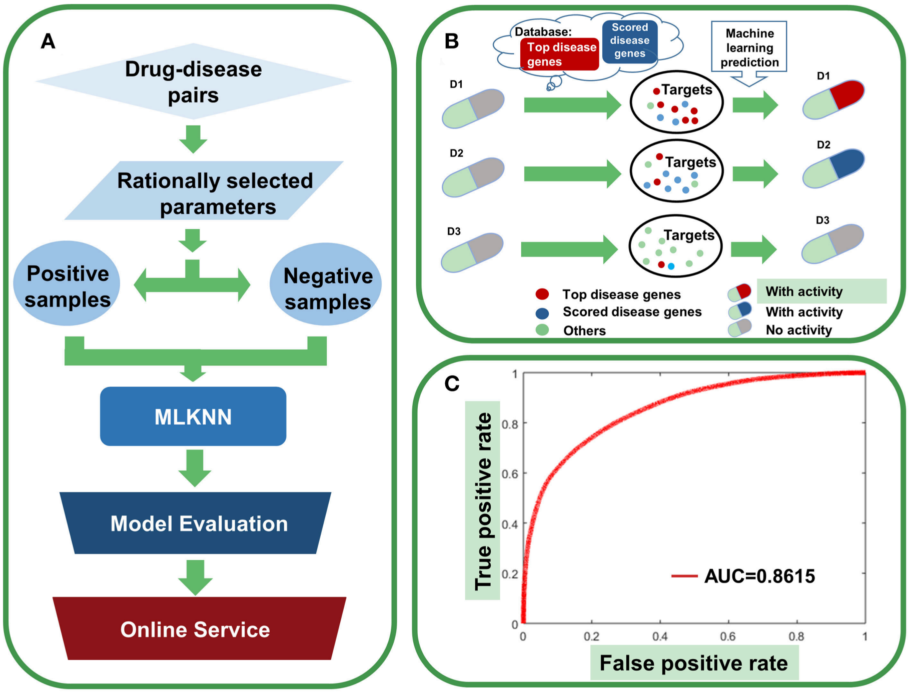


FIGURE 12 | Agent activity prediction with machine-learning models. **(A)** Workflow for the machine-learning model establishment. **(B)** Sketch view for the rationale of agent activity prediction. **(C)** The overall performance of the ensemble classifier.

genetic links between genes and diseases depends on progress in medical genetics.

A number of systems genetics methods have been developed for enriching and screening the driver genes underlying complex traits in the post-GWAS era. For example, Zhu et al. identified 126 genes related to human complex traits through the integration of summary-level GWAS results and eQTL data (Zhu et al., 2016). Based on the exome sequencing, array copy number and RNA sequencing (RNA-seq) data from 3,281 samples across 12 cancer types, Leiserson

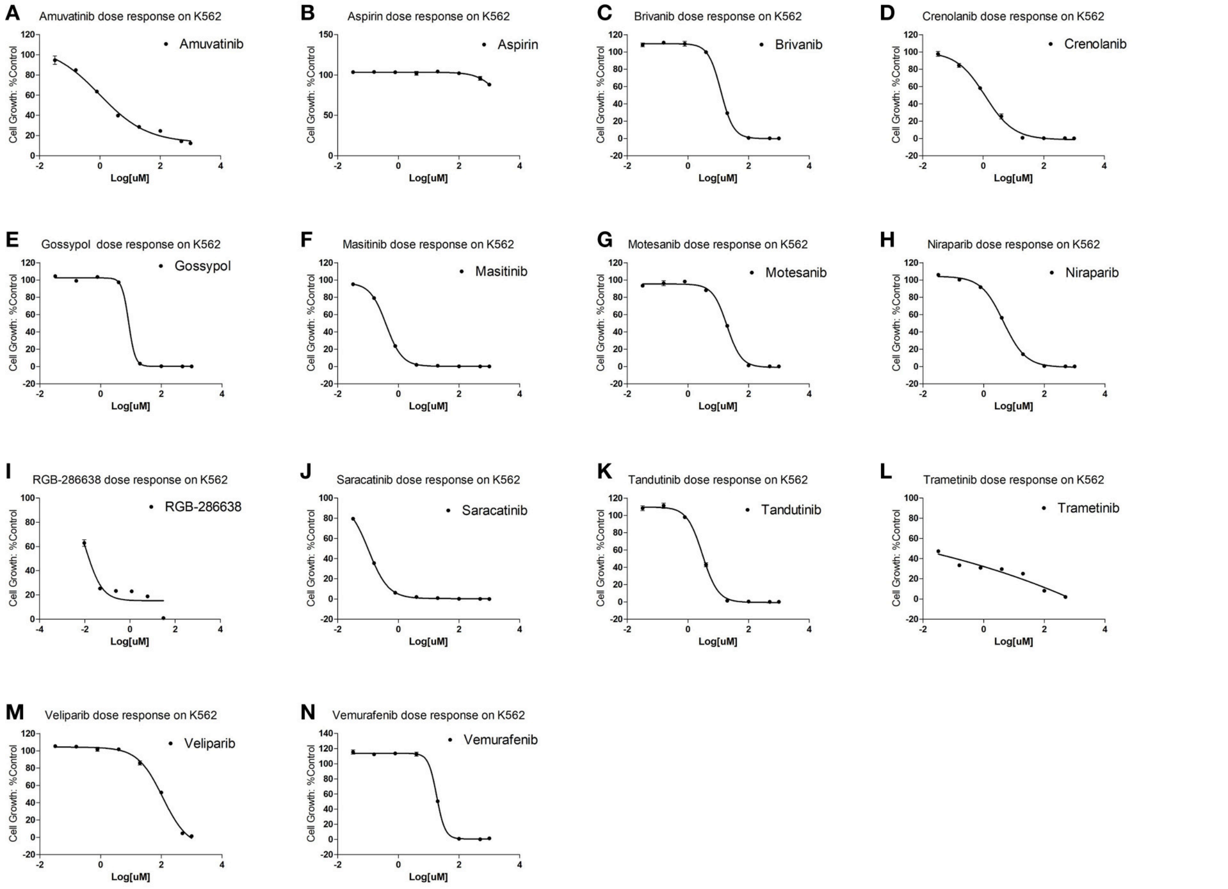


FIGURE 13 | Cytotoxicity of 14 predicted anti-leukemia agents. K562 cells were treated with **(A)** Amuvatinib, **(B)** Aspirin, **(C)** Brivanib, **(D)** Crenolanib, **(E)** Gossypol acetic acid, **(F)** Masitinib, **(G)** Motesanib, **(H)** Niraparib, **(I)** RGB-286638, **(J)** Saracatinib, **(K)** Tandutinib, **(L)** Trametinib, **(M)** Veliparib, **(N)** Vemurafenib. The results show that 10 agents (Amuvatinib, Brivanib, Crenolanib, Masitinib, Motesanib, Niraparib, Saracatinib, Tandutinib, Veliparib, Vemurafenib) (71.43%) can efficiently inhibit the growth of K562.

et al. performed a pan-cancer analysis of mutated networks utilizing a HotNet2 (HotNet diffusion-oriented sub-networks) algorithm, by which they identified 16 significantly mutated subnetworks containing 147 genes. Many of these genes have been validated to play a critical role in cancer pathogenesis (Leiserson et al., 2015). Gamazon et al. proposed a gene-based association method called PrediXcan that directly tests the molecular mechanisms through which genetic variation affects phenotype (Gamazon et al., 2015). Greene et al. introduced a Network-guided GWAS Analysis method called NetWAS, which integrated tissue-specific networks and nominally significant *P*-values in GWAS to identify biologically important disease-gene associations (Greene et al., 2015). Although these methods are helpful to identify reliable genes associated with a complex disease trait, the complex application procedures hinder their convenient use. In this study, we endorsed the possibility of using ohnolog genes as a source of "top disease genes." The high accessibility of ohnologs will facilitate the identification of disease driver genes and the genetics-based drug discovery.

The above discoveries inspired us to establish systems chemical genetic models for predicting drug activities. Because drug repurposing is a hot spot in the pharmaceutical industry, a number of theoretical methods, including cheminformatics-based, bioinformatics-based and systems biology-based methods, have been proposed to predict drug activities (Jin and Wong, 2014). However, most of these methods were derived from parameters trained using large datasets, suggesting that these methods may be sensitive to datasets and poor in generalization capabilities. The identification of the genetic determinants of drug activities facilitates the rational selection of parameters to establish machine-learning models for drug activity prediction. Because this model was built on the fundamental principle of drug activity determination, it is expected to be robust when generalized to different datasets and explainable to certain extent. Moreover, to maximize the convenience for researchers, a user-friendly online service (SCG-Drug) was provided for drug-activity prediction and data retrieval as well. These systems chemical genetics methods are of high value in prioritizing drug candidates, also highlighting the importance of modern genetics in facilitating the paradigm shift of pharmaceutical industry.

MATERIALS AND METHODS

Data Sources and Pre-processing

Agent Information

We collected agents and agent-target associations from three databases: DrugBank, TTD, and DGIdb (Law et al., 2014; Qin et al., 2014; Wagner et al., 2015). By integrating the 6,841 agents covering 3,692 targets from DrugBank, the 5,208 agents covering 569 targets from TTD, and the 10,941 agents covering 3,090 targets from DGIdb, we obtained 35,860 agent-target associations, comprising 16,021 agents and 4,613 target genes. The indication information for the agents were collected from DrugBank, TTD, and ClinicalTrials (Zarin et al., 2011; Law et al., 2014; Qin et al., 2014). Totally, we obtained 80, 90 agents with corresponding target genes and pharmacological activity records. Using the disease classes provided by Pharmaprojects (similarity threshold: 0.75, for more details see the ***Disease standardization*** section), we finally acquired 5,759 agents covering 667 types of diseases and 2,813 target genes.

Disease-Associated Genes

Eight databases were used to collect disease-related genes, including the Genetic Association Database (GAD, https://geneticassociationdb.nih.gov/) (Becker et al., 2004), Online Mendelian Inheritance in Man (OMIM, http://omim.org/) (Hamosh et al., 2005), Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/) (Landrum et al., 2014), Orphanet (http://www.orpha.net/consor/cgi-bin/index.php), DisGeNET (http://www.disgenet.org/web/DisGeNET/menu/rdf) (Piñero et al., 2015), INtegrated TaRget gEne PredItion (INTREPID) (Chen and Tian, 2016), GWASdb (http://jjwanglab.org/gwasdb) (Nelson et al., 2015) and The Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php) (Wang X. et al., 2012). A total of 19,233 disease-associated genes were collected for use in the present analysis. Genes that could not be mapped to an Entrez ID were excluded. The available URLs, version information, access dates, and number of records from the above eight databases are provided in **Table S8**.

Disease Standardization

We used the Unified Medical Language System (UMLS), which provides a comprehensive set of medical concepts, to standardize disease annotations of genes, and agents. UMLS is a medical terminology system that has been developed by the National Library of Medicine for more than 20 years and contains a large number of standardized medical concepts. The natural language processing program MetaMap was used to convert disease annotations to the corresponding disease concepts (Aronson, 2001). We selected Medical Subject Headings (MeSH) as the vocabulary, and limited the semantic type to "Pathologic Function," "Injury or Poisoning," and "Anatomical Abnormality" to obtain the disease-related concepts (Liu et al., 2014). We processed all gene-related phenotypes and agents' indications using the UMLS concept. As MeSH defines disease concepts using a hierarchical system, it classifies each disease to a narrow disease type; for example, "Alzheimer disease 15" is a subtype of "Alzheimer disease." The latter is simply a broader term for the former. In our work, all subtype disease concepts were converted to the appropriate broader term using a Perl module UMLS::Interface. Disease annotations that could not be mapped to any disease concept were excluded from subsequent analyses. Using the disease classes provided by Pharmaprojects (similarity threshold: 0.75) (Mcinnes et al., 2009), we obtained 914,190 gene-disease pairs (involving 703 types of diseases) and 74,902 agent-disease pairs (involving 667 types of diseases).

Sequence Similarity Analysis

The needle program of EMBOSS package (Version: 6.6.0.0) (Rice et al., 2000) was employed to perform sequence similarity analysis of agent-targeted proteins, because of its accurate production of Needleman-Wunsch global pairwise alignments.

Gene Ontology (GO) Terms Similarity Measurement

We used the GO-based Czekanowski–Dice distance to evaluate the GO terms similarity of the target pairs. The Czekanowski–Dice functional distance was calculated using a previously described method (Ovaska et al., 2008). The GO term information of the gene pairs was obtained from the Ensembl database (version 72).

"Top Genes" and Ohnolog Genes

The AlzGene database contains 650 genes for Alzheimer's disease (Bertram et al., 2007); the SzGene database contains 937 genes for schizophrenia (Allen et al., 2008); the PDGene database contains 571 genes for Parkinson's disease (Lill et al., 2012); and the MSGene database contains 675 genes for multiple sclerosis (Lill et al., 1994). From these databases, 44, 43, 31, and 43 genes strongly associated with Alzheimer's disease, schizophrenia, Parkinson's disease and multiple sclerosis, respectively, were identified. These genes were termed "top genes," meaning that relatively reliable associations have been established between these genes and certain diseases. In addition, the ohnologs served as an alternative source of "top disease genes," because ohnologs are significantly enriched with disease genes due to their strong dosage balance (Makino and Mclysaght, 2010; McLysaght et al., 2014). From Makino et al.'s work (Makino and Mclysaght, 2010), we extracted 9,057 ohnolog pairs covering 7,295 genes from the human genome.

Druggability Score of Disease Genes

Based on clinically active ratio of genes from eight disease databases (Clinvar, OMIM, HGMD, Orphanet, GWASdb, INTREPID, GAD, and DisGeNET), we proposed a parameter named druggability score for quantitatively measuring the druggability of disease genes. First, the genes derived from different databases were given different scores, with the highest-clinically active ratio database (i.e., Clinvar) being assigned with the highest score (eight points), the disease genes obtained from the second-ranked database of the clinically active ratio (i.e., OMIM) was given seven points, and so on, from HGMD was given six points, from Orphanet was given five points, from GWASdb was given four points, from INTREPID was given three points, from GAD was given two points, while the lowest clinically active ratio (i.e., DisGeNET) with the lowest score (one point) (**Table S3**). Then, if a disease gene is recorded in multiple databases, the scores of the corresponding multiple databases were summed up for this disease gene to define its druggability:

$$Druggability\ score = \sum_{j=1}^{k} score_{ij} \quad (1)$$

where $score_{ij}$ denotes the assigned score of a pathogenic gene i in the jth database (**Table S3**); $i = 1, 2, ..., m$; $j = 1, 2, ..., k$, where m is the number of disease genes, k is the number of databases ($k = 8$ in this study).

Statistical Analysis

Disease Similarity Measurement

First, the disease terms of genes and indication annotations of agents were converted to the standardized medical concepts of UMLS by a natural language processing tool MetaMap. Then, through using the disease classes provided by pharmaprojects (Similarity threshold: 0.75), the disease similarity was measured using UMLS::similarity. Lin, which is calculated using the information content and path of concepts, shows good performance for disease similarity measurement (Nelson et al., 2015). In this study, we used the Lin to evaluate the disease term similarity of all disease concepts. The Lin is calculated using the following equation:

$$Lin = \frac{IC(lcs)}{IC\left(concept1\right) + IC(concept2)} \quad (2)$$

where IC is the negative log of the probability of the concept, the probability is pre-calculated by the Perl module by summing the probability of the concept occurring in some text plus the probability of its descendants occurring in some text, and lcs is the least common subsuming concept of concept1 and concept2.

Tanimoto Coefficient Calculation

To assess the correlations between disease concepts and their corresponding causal genes or drugs, we characterized the distance between disease gene sets or drug sets using the Tanimoto coefficient. The Tanimoto coefficient (TC) is calculated using the following equation:

$$TC = \frac{N_{AB}}{N_A + N_B - N_{AB}} \quad (3)$$

where $\mathrm{N_A}$ is the number of disease A-related genes or drugs, $\mathrm{N_B}$ is the number of disease B-related genes or drugs, and $\mathrm{N_{AB}}$ is the number of common genes or drugs for disease A and disease B.

Permutation Test

To evaluate the quality of agent-target pairs, we did a 10000-permutation test on the three sets of agent-target pairs derived from DGIdb, TTD and DrugBank (Law et al., 2014; Qin et al., 2014; Wagner et al., 2015), respectively. The agents were randomly assigned with targets and the clinically active ratio of agents was calculated. This random shuffling procedure was repeated for 10,000 times.

Machine-Learning Modeling

Feature Generation

We rationally selected four parameters to build the model. The first parameter characterizes the overall druggability score of the pathogenic genes within drug targets. The second parameter is the average value of the first parameter and is normalized by 36 (namely 8∼). For example, if an agent targets two related disease genes derived from Clinvar and DisGeNET, respectively, the first parameter will be 9 (8 + 1), and the second parameter will be 0.125 (9/2 × 36). The third and fourth parameters are the absolute number and relative ratio of ohnologous disease genes within drug targets, respectively.

Positive Sample Generation

An agent-disease pair was regarded as a positive, if the drug hits one or more disease genes and is indicated for treating this disease. The positive samples were generated as 74,902 agent-disease pairs.

Negative Sample Generation

An agent-disease pair was regarded as a negative, if the drug targets one or more disease genes but is not annotated for controlling this disease. The negative samples were generated as 3,778,517 pairs. In the web server SCG-Drug (http://zhanglab.hzau.edu.cn/scgdrug), the model with all samples is provided.

MLKNN

Given the training set $\left\{\left(x_i, y_i\right)\right\}_{i=1}^{n}$, x_i is the ith instance (drug), and y_i is the corresponding disease vector. $y_i\left(l\right) = 1$. If the ith instance can treat the lth disease, otherwise $y_i\left(l\right) = 0, \quad l = 1, 2, \ldots, q$. The k nearest neighbors (in training set) of instance x_i are denoted by $N\left(x_i\right), \quad i = 1, 2, \ldots, n$. Thus, based on lth disease of these neighbors, a membership counting vector can be denoted as:

$$C_{x_i}\left(l\right) = \sum_{a \in N(x_i)} y_a\left(l\right), \quad l = 1, 2, \ldots, q \quad (4)$$

where $C_{x_i}\left(l\right)$ counts the number of neighbors of x_i treating the lth disease, and $0 \le C_{x_i}\left(l\right) \le k$.

For a test drug t, MLKNN identifies its k nearest neighbors in the training set and calculate $C_t\left(l\right)$. Let H_1^l be the event that a drug has lth disease and H_0^l be the event that a drug does not treat lth disease. Let E_j^l be the event that a drug just has j neighbors with lth disease in its k nearest neighbors. For the instance t, its label for lth disease $y_t\left(l\right)$ is determined by the following principle:

$$y_t\left(l\right) = \arg max_{b \in \{0,1\}} P\left(H_b^l | E_{C_t(l)}^l\right), \quad l = 1, 2, \ldots, q \quad (5)$$

Using the Bayesian rule, above Equation (5). can be rewritten as:

$$\begin{aligned} y_t\left(l\right) &= \arg max_{b \in \{0,1\}} \frac{P\left(H_b^l\right) P\left(E_{C_t(l)}^l | H_b^l\right)}{P\left(E_{C_t(l)}^l\right)} \\ &= \arg max_{b \in \{0,1\}} P\left(H_b^l\right) P\left(E_{C_t(l)}^l | H_b^l\right) \end{aligned} \quad (6)$$

In the prediction model, $P\left(H_b^l\right)$ and $P\left(E_{C_t(l)}^l|H_b^l\right)$ are calculated based on the training set. The prior probabilities are calculated.

$$P\left(H_1^l\right) = \frac{\left(s + \sum_{i=1}^{n} y_i\left(l\right)\right)}{\left(s \times 2 + n\right)} \text{and } P\left(H_0^l\right) = 1 - P\left(H_1^l\right) \quad (7)$$

Then, the posterior probabilities $P\left(E_{C_{x_i}(l)}^l|H_0^l\right)$, $P\left(E_{C_{x_i}(l)}^l|H_1^l\right)$ are calculated by following equations,

$$P\left(E_j^l|H_1^l\right) = \frac{\left(s + c\left[j\right]\right)}{\left(s \times \left(k+1\right) + \sum_{i=0}^{k} c_l\left[i\right]\right)} \quad (8)$$

$$P\left(E_j^l|H_0^l\right) = \frac{\left(s + c'\left[j\right]\right)}{\left(s \times \left(k+1\right) + \sum_{i=0}^{k} c_l'\left[i\right]\right)}$$

$$l = 1, 2, \ldots, q, j = 1, 2, \ldots, k \quad (9)$$

where s is the smooth factor. $c_l\left[i\right]$ is the number of instances which just has i neighbors with lth disease in their k nearest neighbors; $c_l^{'}\left[i\right]$ is the number of instances which just has i neighbors without lth disease in their k nearest neighbors (Zhang and Zhou, 2007).

Cross-Validation

We used 5-fold stratified cross-validation with 1,000 repeats to avoid arbitrariness.

Ensemble Learning Method

In this paper, an ensemble learning method was designed to combine various features and develop high-accuracy prediction models (Lee and Soo, 2013; Yang et al., 2014; Wen et al., 2015). Previous studies have shown that combining predictions from different methods could achieve better and more robust results than using one algorithm alone. In this study, an ensemble classifier was generated using the linear weighted sum of outputs from classifiers based on four features.

Given m features, we build m individual feature-based MLKNN models, and use them as base predictors. Since features may make different contributes, it is natural to adopt weighted scoring ensemble strategy, which assigns m base predictors with m weights $\{w_1, w_2, \ldots, w_m\}$. For a testing instance, the ith predictor will give scores for q diseases, denoted as $S_i = \left\{s_i^1, s_i^2, \ldots, s_i^q\right\}$, $i = 1, 2, \ldots, m$. The final prediction produced by the ensemble model is the linear weighted sum of outputs from base predictors.

$$\text{Ensemble Score} = \left[w_1, w_2, \ldots, w_m\right] \times \begin{bmatrix} S_1 \\ S_2 \\ \cdots \\ S_m \end{bmatrix} \quad (10)$$

$$= \left[w_1, w_2, \ldots, w_m\right] \times \begin{bmatrix} S_1^1 \cdots S_1^2 S_1^q \\ \vdots \ddots \vdots \\ S_m^1 S_m^2 \cdots S_m^q \end{bmatrix} \quad (11)$$

Tuning weights for base predictors are critical for the ensemble models. The weights are non-negative real values between 0 and 1, and the sum of weights equals 1. We adopt the internal 5-CV AUPR on training data is used as the fitness score (Lee and Soo, 2013; Yang et al., 2014; Wen et al., 2015).

Performance Evaluation

In the agent-activities prediction, the predicted scores for activities were usually merged for evaluation, and the metrics for ordinary binary classification were often adopted. The area under ROC curve (AUC) and the area under the precision-recall curve (AUPR) can be used to evaluate models regardless of any threshold. However, there are much more negative labels than positive labels in the agent-activities prediction, and machine-learning methods are likely to produce overestimated AUC scores. Since AUPR takes into account recall as well as precision, it is used as the most important metric.

We used the following evaluation metrics to evaluate the performance of machine-learning models: Precision, Accuracy (ACC), Recall, Specificity, Mathew's correlation coefficient (MCC) (12–16). These metrics can be calculated by the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$ACC = \frac{TP + TN}{TP + FN + TN + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$Specificity = \frac{TN}{TN + FP} \quad (15)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \quad (16)$$

Several metrics were designed for multi-label classification, i.e., Hamming loss, one-error, coverage, ranking loss and average precision. Hamming loss is the fraction of the wrong labels to the total number of labels. The one-error evaluates the fraction of examples whose top-ranked label is not in the relevant label set. The coverage evaluates how many steps are needed, on average, to move down the ranked label list so as to cover all the relevant labels of the example. The average precision evaluates whether the average fraction of relevant labels ranked higher than a particular label. Therefore, we adopt AUPR, average precision, one-error, coverage, ranking loss and hamming loss for the agent-activities prediction.

Cytotoxicity Assays

Cell Culture and Reagents

K562 cells were purchased from Shanghai Cell Bank, Chinese Academy of Sciences. Cells were cultured in RPMI-1640 (Procell, China) with 10% FBS (Biowest, France) and 1% penicillin/streptomycin (Procell, China) at 37°C, in 5% CO_2 humidified atmospheric air. All agents were purchased from TargetMol and dissolved in dimethyl sulfoxide (DMSO).

Cytotoxicity Assays

The effects of agents on K562 were determined using CellTiter-Glo® Luminescent Cell Viability Assay (Promega). Cells were seeded in 96-well plate at a density of 2 × 10^3 cells/well and treated with different agents for 72 h together. An equal volume of CellTiter-Glo reagents was added to the cells in 96-well plates and mixed for 2 min on an orbital shaker and incubated for a further 10 min at room temperature. The luminescence of each well was measured by FlexStation3(Molecular Devices). The IC50 values were calculated using Graphpad Prism software. All experiments were performed in triplicate.

Web Server Implementation

Systems Chemical Genetics-Drug (SCG-Drug, http://zhanglab.hzau.edu.cn/scgdrug) was built in Java, JavaScript, and Bootstrap with MySQL as the primary data store. The site is served with nginx on a server running CentOS 7.2. Two modules are used: the search module and the prediction module. The search module was implemented by an entry-name matching algorithm. By using this module, the server will return a list of partially matched terms and shows them in the dropdowns when users type only the starting characters of a gene, disease or drug in the search field. In the prediction module, there are two steps: data preprocessing and drug indication prediction. In the data preprocessing step, a Python script was used to produce the parameters matrix. In the drug indication prediction step, an R script was used to generate the result by calling the prediction model.

Code and Data Availability

The R and Python scripts used to process the data and conduct the analyses described herein are available upon request. All of the intermediate data are available from the authors by request.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://zhanglab.hzau.edu.cn/scgdrug.

AUTHOR'S NOTE

Finding novel drugs or new uses for old drugs is a costly process. Previous studies have shown that genetics, which is best dedicated to revealing gene-disease links, makes great contributions to the pharmaceutical industry. On the other hand, most diseases are caused by multiple pathogenic factors. In this paper, we proposed that aiming at multiple genes associated with certain diseases rather than a single pathogenic factor is more efficient in identifying potential drugs. In addition, our results demonstrated the therapeutic potential of agents can be enhanced with the consolidation of genetic links between targets and diseases. In other words, simultaneously increasing the quantity and quality of target-disease associations can significantly increase the activity/druggability of agents. According to the above theories, we have established a drug-activity predictor with multi-label classification model based on the genetic information of drug targets (online service is freely available at SCG-Drug, http://zhanglab.hzau.edu.cn/scgdrug), which is of high value in prioritizing drug candidates.

AUTHOR CONTRIBUTIONS

H-YZ: conceptualization. YQ, Z-HL, and L-DZ: data curation. YQ, Z-HL, and Q-YY: formal analysis. H-YZ: funding acquisition. QZ, Z-JC, and XQ: investigation. H-YZ and L-DZ: methodology. JL and Y-ML: software. B-ML: conceived and designed the experiments. Y-HX: performed the experiments. H-YZ and L-DZ: supervision. H-YZ and L-DZ: validation. YQ, Z-HL, Q-YY, L-DZ, and JL: visualization. YQ, Z-HL, Q-YY, L-DZ, and H-YZ: writing–original draft. YQ, Z-HL, Q-YY, L-DZ, and H-YZ: writing–review and editing.

ACKNOWLEDGMENTS

This research was partially supported by the Fundamental Research Funds for the Central Universities (Grant 2662017PY115). We are grateful to Lin Li for his helpful assistance in web server implementation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00474/full#supplementary-material

REFERENCES

Allen, N. C., Bagade, S., McQueen, M. B., Ioannidis, J. P., Kavvoura, F. K., Khoury, M. J., et al. (2008). Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat. Genet.* 40, 827–834. doi: 10.1038/ng.171

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* 2001, 17–21.

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36, 431–432. doi: 10.1038/ng0504-431

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., and Tanzi, R. E. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* 39, 17–23. doi: 10.1038/ng1934

Brinkman, R. R., Dubé, M. P., Rouleau, G. A., Orr, A. C., and Samuels, M. E. (2006). Human monogenic disorders-a source of novel drug targets. *Nat. Rev. Genet.* 7, 249–260. doi: 10.1038/nrg1828

Chen, J., and Tian, W. (2016). Explaining the disease phenotype of intergenic SNP through predicted long range regulation. *Nucleic Acids Res.* 44, 8641–8654. doi: 10.1093/nar/gkw519

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., and Himmelstein, D. S. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* 47, 569–576. doi: 10.1038/ng.3259

Guney, E., Menche, J., Vidal, M., and Barábasi, A. L. (2016). Network-based *in silico* drug efficacy screening. *Nat. Commun.* 7:10331. doi: 10.1038/ncomms10331

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

Hopkins, A. L. (2008). Network pharmacology: the next paradigm in, drug discovery. *Nat. Chem. Biol.* 4, 682–690. doi: 10.1038/nchembio.118

Jin, G., and Wong, S. T. C. (2014). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discov Today.* 19, 637–644. doi: 10.1016/j.drudis.2013.11.005

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

Landrum, M. J., Le, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 4, 980–985. doi: 10.1093/nar/gkt1113

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

Lee, P. F., and Soo, V. W. (2013). An ensemble rank learning approach for gene prioritization. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2013, 3507–3510. doi: 10.1109/EMBC.2013.6610298

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168

Lill, C. M., Roehr, J. T., McQueen, M. B., Bagade, S., Schjeide, B. M., Zipp, F., et al. (1994). *The MSGene Database*. Alzheimer Research Forum. Available online at http://www.msgene.org/ (accessed June, 2015).

Lill, C. M., Roehr, J. T., McQueen, M. B., Kavvoura, F. K., Bagade, S., Schjeide, B. M., et al. (2012). Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: the PDGene database. *PLoS Genet.* 8:e1002548. doi: 10.1371/journal.pgen.1002548

Liu, C. C., Tseng, Y. T., Li, W., Wu, C. Y., Mayzus, I., Rzhetsky, A., et al. (2014). DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res.* 42, 137–146. doi: 10.1093/nar/gku412

Makino, T., and Mclysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9270–9274. doi: 10.1073/pnas.0914697107

Mcinnes, B. T., Pedersen, T., and Pakhomov, S. V. (2009). UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. *AMIA Annu. Symp. Proc.* 14, 431–435.

McLysaght, A., Makino, T., Grayton, H., Tropeano, M., Mitchell, K. J., Vassos, E., et al. (2014). Ohnologs are overrepresented in pathogenic copy number mutations. *Proc. Natl. Acad. Sci. U.S.A.* 111, 361–364. doi: 10.1073/pnas.1309324111

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., et al. (2015). The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860. doi: 10.1038/ng.3314

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Ovaska, K., Laakso, M., and Hautaniemi, S. (2008). Fast gene ontology based clustering for microarray experiments. *BioData Min.* 1:11. doi: 10.1186/1756-0381-1-11

Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database.* 2015:bav028. doi: 10.1093/database/bav028

Plenge, R. M., Scolnick, E. M., and Altshuler, D. (2013). Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* 12, 581–594. doi: 10.1038/nrd4051

Qin, C., Zhang, C., Zhu, F., Xu, F., Chen, S. Y., Zhang, P., et al. (2014). Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res.* 42, D1118–D1123. doi: 10.1093/nar/gkt1129

Quan, Y., Wang, Z. Y., Chu, X. Y., and Zhang, H. Y. (2018). Evolutionary and genetic features of drug targets. *Med Res Rev.* 38, 1536–1549. doi: 10.1002/med.21487

Quan, Y., and Zhang, H. Y. (2016). A chemical-genetic criterion for identifying disease biomarkers. *Trends Mol. Med.* 22, 447–448. doi: 10.1016/j.molmed.2016.04.001

Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebbring, S. J., and Lin, S. M. (2015). Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol.* 33, 342–345. doi: 10.1038/nbt.3183

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., et al. (2012). Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* 30, 317–320. doi: 10.1038/nbt.2151

Sekine, M., and Makino, T. (2017). Inference of causative genes for Alzheimer's disease due to dosage imbalance. *Mol. Biol. Evol.* 34, 2396–2407. doi: 10.1093/molbev/msx183

Wagner, A. H., Coffman, A. C., Ainscough, B. J., Spies, N. C., Skidmore, Z. L., Campbell, K. M., et al. (2015). DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.* 44, D1036–D1044. doi: 10.1093/nar/gkv1165

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., and Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* 30, 159–164. doi: 10.1038/nbt.2106

Wang, Z. Y., Fu, L. Y., and Zhang, H. Y. (2012). Can medical genetics and evolutionary biology inspire drug target identification? *Trends Mol. Med.* 18, 69–71. doi: 10.1016/j.molmed.2011.11.004

Wang, Z. Y., and Zhang, H. Y. (2013). Rational drug repositioning by medical genetics. *Nat. Biotechnol.* 31, 1080–1082. doi: 10.1038/nbt.2758

Wen, Z., Feng, L., Longqiang, L., and Jingxia, Z. (2015). Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinformatics.* 16:365. doi: 10.1186/s12859-015-0774-y

Xie, T., Yang, Q. Y., Wang, X. T., McLysaght, A., and Zhang, H. Y. (2016). Spatial colocalization of human ohnolog pairs acts to maintain dosage-balance. *Mol. Biol. Evol.* 33, 2368–2375. doi: 10.1093/molbev/msw108

Xiong, M., Li, B., Zhu, Q., Wang, Y. X., and Zhang, H. Y. (2014). Identification of transcription factors for drug-associated gene modules and biomedical implications. *Bioinformatics* 30, 305–309. doi: 10.1093/bioinformatics/btt683

Yang, P., Yoo, P. D., Fernando, J., Zhou, B. B., Zhang, Z., and Zomaya, A. Y. (2014). Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Trans. Cybern.* 44, 445–455. doi: 10.1109/TCYB.2013.2257480

Yildirim, M. A., Goh, K. I., Cusick, M. E., Barabási, A., and Vidal, M. (2007). Drug-target network. *Nat. Biotechnol.* 25, 1119–1126. doi: 10.1038/nbt1338

Zarin, D. A., Tse, T., Williams, R. J., and Califf RM Ide, N. C. (2011). The clinicaltrials.gov results database-update and key issues. *N. Engl. J. Med.* 364, 852–860. doi: 10.1056/NEJMsa1012065

Zhang, M. L., and Zhou, Z. H. (2007). ML-KNN: a lazy learning approach to multi-label learning. *Pattern. Recogn.* 40, 2038–2048. doi: 10.1016/j.patcog.2006.12.019

Zhang, W., Niu, Y., Zou, H., Luo, L., Liu, Q., and Wu, W. (2015). Accurate prediction of immunogenic T-cell epitopes from epitope sequences using the genetic algorithm-based ensemble learning. *PLoS ONE* 10:e0128194. doi: 10.1371/journal.pone.0128194

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48, 481–487. doi: 10.1038/ng.3538

Conflict of Interest Statement: H-YZ has received research/grant support from Wuhan Bio-Links Technology Co., Ltd. Huazhong Agricultural University and the developers of the methods for drug discovery and drug repositioning may benefit financially pursuant to the University's Policy on Inventions, Patents and Technology Transfer, even if these methods are not used in the commercialized therapy.

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

ORIGINAL RESEARCH
published: 25 June 2019
doi: 10.3389/fgene.2019.00598

SCIA: A Novel Gene Set Analysis Applicable to Data With Different Characteristics

Yiqun Li[1†], *Ying Wu*[2†], *Xiaohan Zhang*[1], *Yunfan Bai*[1], *Luqman Muhammad Akthar*[1], *Xin Lu*[1], *Ming Shi*[1], *Jianxiang Zhao*[1], *Qinghua Jiang*[1*] and *Yu Li*[1*]

[1] *Department of Laboratory of Cancer Biology, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China,* [2] *Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China*

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Qinghua Cui,
Peking University, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Qinghua Jiang
qhjiang@hit.edu.cn
Yu Li
liyugene@hit.edu.cn

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 May 2019*
Accepted: *05 June 2019*
Published: *25 June 2019*

Citation:
Li Y, Wu Y, Zhang X, Bai Y, Akthar LM, Lu X, Shi M, Zhao J, Jiang Q and Li Y (2019) SCIA: A Novel Gene Set Analysis Applicable to Data With Different Characteristics. Front. Genet. 10:598. doi: 10.3389/fgene.2019.00598

Gene set analysis is commonly used in functional enrichment and molecular pathway analyses. Most of the present methods are based on the competitive testing methods which assume each gene is independent of the others. However, the false discovery rates of competitive methods are amplified when they are applied to datasets with high inter-gene correlations. The self-contained testing methods could solve this problem, but there are other restrictions on data characteristics. Therefore, a statistically rigorous testing method applicable to different datasets with various complex characteristics is needed to obtain unbiased and comparable results. We propose a self-contained and competitive incorporated analysis (SCIA) to alleviate the bias caused by the limited application scope of existing gene set analysis methods. This is accomplished through a novel permutation strategy using *a priori* biological networks to selectively permute gene labels with different probabilities. In simulation studies, SCIA was compared with four representative analysis methods (GSEA, CAMERA, ROAST, and NES), and produced the best performance in both false discovery rate and sensitivity under most conditions with different parameter settings. Further, the KEGG pathway analysis on two real datasets of lung cancer showed that the results found by SCIA in both of the two datasets are much more than that of GSEA and most of them could be supported by literature. Overall, SCIA promisingly offers researchers more reliable and comparable results with different datasets.

Keywords: GSA, competitive method, self-contained method, topology-based method, functional enrichment analysis

INTRODUCTION

In recent years, gene set analysis (GSA) has become the most common method in functional genomics studies, because evaluating a single *p*-value for a gene set is statistically more powerful than genewise tests. Typically, by choosing gene sets that represent biological pathways, GSA can help to bring insights into biological mechanisms, cellular functions, and disease states (Kanehisa et al., 2012). Various statistical procedures for gene set testing have been proposed and can be divided into three generations roughly in chronological order (Khatri et al., 2012; Zyla et al., 2017).

The first generation of GSA used over-representation analysis (ORA), where the first step is to define differentially expressed genes (DEGs) and non-DEGs in the input gene list by a certain threshold (Beissbarth and Speed, 2004). Then, the proportion of DEGs between a given functional gene set and the background gene set are tested by hypergeometric, binomial, or chi-square distribution. This comparison of the DEG proportions is the original theory of competitive testing. ORA has been reported with minor variations by many different authors (Khatri and Draghici, 2005). Even though the ORA method seems simple and effective, there are two serious drawbacks. First, the information about the strength of gene expression is lost by gene binarization. Second, the assumption of inter-gene independence needed by the testing methods is not satisfied in most cases.

The second generation of GSA, known as functional class sorting (FCS), was proposed to avoid these deficiencies. Instead of defining genes as DEGs and non-DEGs, different univariate gene-level statistics such as *t*-statistic (Al-Shahrour et al., 2005; Tian et al., 2005), *Q*-statistic (Goeman et al., 2004), signal-to-noise ratio (Subramanian et al., 2005), fold change score and *Z*-score (Kim and Volsky, 2005), or their trans-formations (Tian et al., 2005; Ackermann and Strimmer, 2009) are used to measure DEGs and overcome the first problem of ORA. Then, a gene-set-level statistic is aggregated by these gene-level statistics. Aggregation approaches can be sum, mean, median of the gene-level statistics (Jiang and Gentleman, 2007), or calculating statistics such as the Kolmogorov-Smirnov statistic (Mootha et al., 2003; Subramanian et al., 2005), Wilcoxon rank sum (Barry et al., 2005), or the max-mean statistic (Efron and Tibshirani, 2006). Because the distributions of gene-set-level statistics are usually unknown, permutation procedures are used to complete FCS tests. According to different null hypotheses and corresponding permutation objects, FCSs can be classified as competitive or self-contained methods.

Assuming that all the input genes are independent of each other, competitive methods usually permute gene labels but lose the inter-gene information, which causes the false discovery rate (FDR) to be uncontrolled when the inter-gene correlations are high. Self-contained methods test each gene set independently by permuting sample labels but lose all the information outside the given gene set, which causes the FDR to be uncontrolled when the percentage of DEGs in the background genes is high. Irrespective of the prerequisites for the permutation procedure, the ORA methods can be considered as generalized competitive methods, whereas the classical methods based on multiple linear regression (Mansmann and Meister, 2005; Kong et al., 2006), by definition, are special cases of self-contained methods.

To address the second problem of ORA, some competitive FCS methods that take account of the correlations among genes have been proposed. The method of Nam (2010) removed the bias caused by the inter-gene correlations, while the method of Wu and Smyth (2012) alleviated the problem by estimating the variance inflation factor. However, the information of inter-gene correlations is partially neglected in these procedures, which causes reduced sensitivity or uncontrolled FDR. Self-contained FCS methods seem to be more powerful than competitive ones and do not assume that all the genes are independent, but their null hypothesis is usually over restrictive (Goeman et al., 2004; Tian et al., 2005; Khatri et al., 2012). They assume that the gene set does not contain any genes with expression levels that are associated with different experimental conditions. Under this hypothesis, a few DEGs may cause a given pathway to be defined as a significant differential pathway (Khatri et al., 2012). Although the method of Wu et al. (2010) moderated this hypothesis using a Monte Carlo based testing method, the parameter describing the least proportion of DEGs in a pathway is given arbitrarily instead of calculated by the expression of genes outside the gene set. Even though competitive methods are overwhelmingly more commonly used than self-contained methods in the genomic literature (Gatti et al., 2010), information is still lost during the permutation procedures. Thus, the collision of applicable scopes between self-contained and competitive methods remains unsolved.

The third generation of GSA, known as the pathway topology (PT)-based approach, is based on the large amount of publicly available pathway knowledge. Mitrea et al. (2013) introduced dozens of PT-based methods with different principles and applicable conditions. Most of these methods consider topological information as a weight that measures the centrality of nodes but ignores the spatiotemporal specificity of topological information and changes in the topological structure between different experimental conditions (Fang et al., 2012; Gu et al., 2012; Dona et al., 2017). On this basis, the method of Yuan et al. (2016) proposed a novel statistic that combines node (gene expression) changes with edge (inter-gene correlation) changes. The utilization of biological information greatly improved the performance of PT-based methods, however, the testing methods of them are essentially the same as FCS methods in that they perform the same pipeline (Mitrea et al., 2013). Therefore, the above defects of FCS methods are not solved by PT-based methods.

Here, we propose a new GSA method with less information loss that can alleviate the bias of self-contained and competitive methods caused by their limited applicability. First, to capture all the information within a given gene set like other self-contained methods, a powerful multivariate statistic C is developed to test node changes and edge changes simultaneously. We chose Hotelling's T^2, a self-contained statistic with the ability to penalize gene collinearity (Ackermann and Strimmer, 2009), for node testing because of its suitability for overcoming the limitation of competitive methods, and linear regression to test the edge changes among genes. Because of the additivity of chi-square distributed variables, these two statistics are transformed to the chi-square scale and summed up to get the C statistic. Second, we developed a novel permutation procedure based on a condition-specific shortest-path network (CSSPN, proposed by Dezso et al., 2009). The genes in the CSSPN are selectively permuted instead of permuting the whole gene labels as usual. This procedure does not disrupt inter-gene correlations but uses

Abbreviations: GSA, gene set analysis; ORA, over-representation analysis; DEGs, differentially expressed genes; FCS, functional class sorting; FDR, false discovery rate; CSSPN, condition-specific shortest-path network; SCIA, self-contained and competitive incorporated analysis.

inter-pathway information from *a priori* biological networks, which creates a platform for the incorporation of self-contained, competitive, and PT-based methods. The whole pipeline is called self-contained and competitive incorporated analysis (SCIA), which has been implemented in an R package "SCIA" available on GitHub https://github.com/YiqunLiHIT/SCIA. Results from this study showed that the sensitivity and FDR of SCIA outperform four other commonly used GSA methods in most conditions in simulated datasets and the results are more stable with different real datasets of lung cancer.

STATISTICAL MODELS AND METHODS

Notations and Background Network

The main objective of SCIA is to detect gene sets that are differentially expressed under different experimental conditions. Here, we consider the gene set as pathway P for one experimental condition and P' for another. N_1 and N_2 are the sample size for P and P', respectively. For convenience, we assumed that P and P' are under linear models:

$$X_1 \xrightarrow{\beta_1} X_2 \xrightarrow{\beta_2} \ldots\ldots\ X_{n-1} \xrightarrow{\beta_{n-1}} X_n$$
$$X_1' \xrightarrow{\beta_1{}'} X_2{}' \xrightarrow{\beta_2{}'} \ldots\ldots X'{}_{n-1} \xrightarrow{\beta'{}_{n-1}} X_n'$$

with n nodes and $n-1$ edges, where β_i $(1 \leq i < n)$ represent the regression coefficient of X_i and X_{i+1}. Let $U = \left(\overline{X}_1 - \overline{X}_1{}', \overline{X}_2 - \overline{X}_2{}', \ldots\ldots, \overline{X}_n - \overline{X}_n{}'\right)$ denote the vector of difference in the means of two groups. S and S' are the covariance matrices of P and P', respectively. These notations are also used in the simulation studies.

We chose the background network of CSSPN as the Human Protein Reference Database (HPRD) network (Library et al., 2009), a centralized platform to visually depict and integrate information pertaining to do-main architecture, post-translational modifications, interaction networks, and disease associations for each protein in the human proteome. Other comprehensive networks, such as the integrated network of seven common used networks in Edge Set Enrichment Analysis (Han et al., 2015) can also be used as the background network of SCIA.

C Statistic

The C statistic is proposed to measure the difference of a given gene set in different experimental conditions. It consists of two parts, the node difference model and the edge difference model. The node difference model is based on Hotelling's T^2 method:

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} U^T S_c^{-1} U$$

where,

$$S_c = \frac{(N_1 - 1)\, S + (N_2 - 1)\, S'}{N_1 + N_2 - 2}$$

Under the self-contained null hypothesis H_0: $U = 0$, T^2 follows a chi-square distribution with degrees of freedom equal to n representing genes in the given pathway with a sufficient sample size. This allows Hotelling's T^2 statistic to be combined with other statistics that also follow a chi-square distribution, because chi-square distributions are additive on the freedoms. There are many transformations of Hotelling's T^2 statistic which show its different characteristics. It can be transformed as:

$$F = \frac{N_1 + N_2 - n - 1}{(N_1 + N_2 - 2)n} T^2$$

following an F distribution with the degree of freedom of n and $N_1 + N_2 - n - 1$ under a relatively small sample size. This allows Hotelling's T^2 statistic to be used alone when the sample size is insufficient. Typically, Hotelling's T^2 test is not only a node testing method but is related to the Pearson correlation coefficient. For convenience, assuming $n = 2$ and N_2 is big enough, the estimated value $\overline{X}_i{}'$ $(1 \leq i \leq 2)$ can be considered as constants μ_i $(1 \leq i \leq 2)$, then Hotelling's T^2 statistic can be transformed as:

$$T^2 = \frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{1 - \rho^2}$$

where t_1 *and* t_2 denote the t-statistics for the two component genes, and ρ represents the Pearson correlation coefficient between X_1 and X_2. If $t_1 = t_2$, Hotelling's T^2 statistic can be simplified to:

$$T^2 = \frac{2t_1^2}{1 + \rho}$$

This transformation of T^2 indicates that when X_1 and X_2 are positively correlated and have similar changes in different experimental conditions, there would be a penalty on the Pearson correlation coefficient, which can avoid the disadvantages of the competitive methods. When X_1 and X_2 are negatively correlated but both have positive changes in different experimental conditions, which indicates that the correlation of X_1 and X_2 has changed in different experimental conditions, the T^2 statistic is would be more sensitive.

Although Hotelling's T^2statistic only slightly considers the correlations between genes, a statistically rigorous edge testing statistic is still needed. Based on the linear regression method, a Z-score-like statistic is combined with Hotelling's T^2 statistic in the C statistic. $\hat{\beta}_i$ and $\hat{\beta}_i{}'$ can be estimated by the least square method. Then the Z-score-like B statistic can be written as:

$$B_i = \frac{\hat{\beta}_i - \hat{\beta}_i{}'}{\sqrt{var\left(\hat{\beta}_i\right) + var\left(\hat{\beta}_i{}'\right)}}$$

under the null hypothesis H_0: $\hat{\beta}_i = \hat{\beta}_i{}'$, B_i follows a standard normal distribution the same as the Z-score, and B_i^2 follows a chi-square distribution and can be combined with Hotelling's T^2 statistic. Thus, we obtained the C statistic as:

$$C = T^2 + \sum_{i=1}^{n-1} \mathrm{B}_i^2$$

which follows a chi-square distribution with the degrees of freedom equal to $n+(n-1)$, and can be used to test node changes and edge changes simultaneously. Notably, when the sample size is very small, T^2 and $B_i^{\mathbf{2}}$ will not obey the chi-square distribution, the parameter of SCIA about the correlation test should be set as "FALSE."

CSSPN-Based Permutation Procedure

To avoid the shortcoming of self-contained methods and utilize additional inter-pathway information from *a priori* biological networks, a CSSPN is built by SCIA. First, a set of DEGs should be selected as the terminal genes of CSSPN, and a set of initial genes can usually be selected in the same way. For each pair of genes (X_i , X_t), where X_i is in the initial gene set and X_t is in the terminal gene set, all the shortest pathways are searched under a background network, such as HPRD (see section Notations and Background Network). When the results are not unique, the pathway with the highest C score will be chosen for a sub-pathway permutation procedure. In this procedure, 1,000 nodes are selected randomly as the initial gene set for each X_t, which is the only terminal gene in this procedure. Assuming there are x shortest pathways, built by the randomly selected genes and X_t, that have higher C scores than the given gene pair (X_i, X_t), the permutation p-value of the sub-pathway (X_i, X_t) is x/1,000. The permutation p-value and C statistic p-value are both adjusted using the method of Benjamini and Hochberg (1995), and only if the two p-values are <0.05, the sub-pathway is defined as a statistically significant pathway. Then, all the significant sub-pathways among the initial gene set and the terminal gene set are used to build the CSSPN. All the genes in the CSSPN can be considered as DEGs with edges and can be used in classical functional enrichment analysis.

In SCIA, background genes are used selectively in the CSSPN-based permutation procedure. Essentially, the selection of background genes means the information from the *a priori* biological network is utilized, because all the genes neighboring DEGs in the background network are used at a higher probability to establish the CSSPN. Additionally, because the permutation procedure does not destroy any inter-gene or inter-pathway structures, almost no information is lost in SCIA.

RESULTS

Simulated Data and Scenarios

Simulated Data

The simulated data were generated under a linear model (Formula 1). Firstly, we generated the initial node X_1 of a given pathway P from the normal distribution $N\left(\mu_1,\ \sigma_1{}^2\right)$. And then, the neighbor node $X_2 = \beta_1 X_1 + \varepsilon_1$, $X_3 = \beta_2 X_2 + \varepsilon_2$ …… $X_n = \beta_{n-1} X_{n-1} + \varepsilon_n$ were generated in the same way. Where $\varepsilon_i \sim N\left(0, \tau_i{}^2\right)(1 < i \le n)$ was the residual error term. Similarly, we generated $X_1{}' \sim N\left(\mu_1{}',\ \sigma_1{}'^2\right)$, $X_i{}' = \beta_{i-1}{}' X_{i-1}{}' + \varepsilon_i{}'$ with $\varepsilon_i{}' \sim N\left(0,\ \tau_i{}'^2\right)$ $(1 < i \le n)$representing the pathway P'under another experimental condition. Under the H_0 hypothesis that there is no change in nodes and edges between different experimental conditions, we set the default simulating parameters as:$\mu_1 = \mu_1{}' = 1$, $\sigma_1{}^2 = \sigma_1{}'^2 = 1$,$\tau_i{}^2 = \tau_i{}'^2 = 1$, and $\beta_i = \beta_i{}' = 0.5$. In most of the following simulations without mentioned specially, the gene number n in a pathway was set as 5, the sample sizes N_1 and N_2 of different experimental conditions were both set as 100, and the simulations were repeated 1,000 times.

Scenarios

Four scenarios and 16 conditions were used to simulate different data structures and prove the extensive applicability of SCIA. The H_0 hypothesis condition was designed to evaluate the FDR and the H_1 hypothesis condition was designed to evaluate the sensitivity. The basic setting for the H_1 hypothesis is node or edge changes, with three additional conditions: sample size, inter-gene correlation, and percentages of DEGs in background genes that are outside the given pathway. In each scenario, only one additional condition is set as different values to highlight the robustness of SCIA. Thus, the four scenarios are:

(1) Node change, 0% background DEGs, different correlations, and fixed sample size.
(2) Node change, 10% background DEGs, different correlations, and fixed sample size.
(3) Node change, different percentages of background DEGs, fixed correlations, and fixed sample size.
(4) Edge change, 0% background DEGs, fixed correlations, and different sample sizes.

Scenarios 1 and 2 were designed to simulate datasets with different inter-gene correlations, scenario 3 was designed to simulate datasets with different percentages of DEGs in background genes, and scenario 4 was designed to simulate datasets with edge changes under different sample sizes. Details of the parameter settings under these scenarios are listed in **Supplementary Data Section 1**.

Evaluation of SCIA Performance With Simulated Data

To evaluate its performance, SCIA was compared with two powerful self-contained approaches, ROAST and NES, and two commonly used competitive approaches, CAMERA and GSEA (More details about these methods are stated in **Supplementary Data Section 2**). The application scope of these methods is quite different, so we compared SCIA with them under corresponding application conditions. As shown in **Table 1**, only competitive methods are suitable for scenario 3, and only self-contained methods are suitable for scenario 4.

SCIA Successfully Controls the FDR Under Different Inter-gene Correlations in Simulated Datasets

First, we compared SCIA with self-contained methods in scenario 1 under different inter-gene correlations in simulated datasets. The FDRs were well-controlled by all the three methods (**Table 2**), and **Figure 1** clearly shows the sensitivities of the three methods were quite similar, indicating the C statistic allowed SCIA to match the advantages of the self-contained methods. Noticeably, ROAST had high sensitivity under the

TABLE 1 | Application scope of the different methods evaluated in this study.

Conditions	SCIA	Self-contained		Competitive	
		NES	ROAST	CAMERA	GSEA
High intergene correlations	√	√	√	√	×
High prop. of background DEGs	√	×	×	√	√
Correlation changes testing	√	√	×	×	×

"√" indicates the method was designed for the condition; "×" indicates the method was not designed for the condition and may have problems in sensitivity or FDR.

TABLE 2 | FDR is well-controlled by SCIA similar to other self-contained methods under different inter-gene correlations in simulated datasets.

Correlations	SCIA	NES	ROAST
0.0	0.048	0.056	0.052
0.3	0.046	0.045	0.046
0.6	0.056	0.049	0.061
0.9	0.044	0.082	0.038

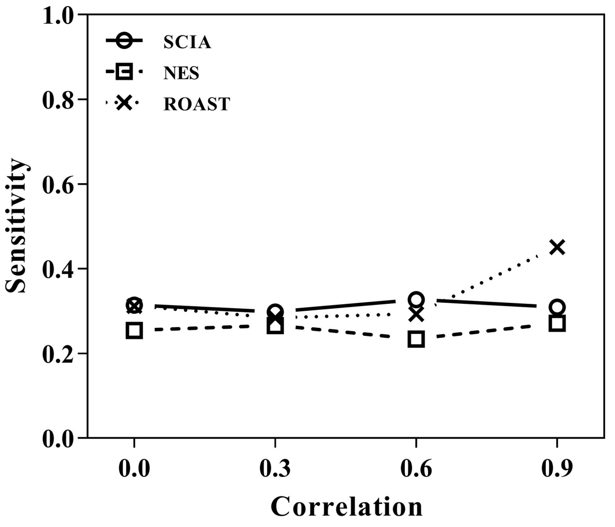


FIGURE 1 | SCIA produces sensitivities similar to those for the self-contained methods under different inter-gene correlations in simulated datasets.

high inter-gene correlation. However, high sensitivity with inter-gene correlations close to 1 is not useful for combination with competitive approaches because a small percentage of highly correlated DEGs may produce unreasonable significant results.

Second, we compared SCIA with competitive methods under scenario 2. **Table 3** clearly shows that the FDR of GSEA lost control, which is common for competitive methods due to the correlation between genes, whereas CAMERA adjusted the high FDR only under a moderate inter-gene correlation of all genes but failed to control the FDRs under high inter-gene correlations. SCIA was the most robust method with well-controlled FDRs and similar sensitivities as CAMERA with comparable FDRs. Because there were no randomly selected DEGs in the given pathway, the SCIA results in scenarios 1 and 2 are comparable, which indicated that the information of background genes outside the given gene set was well-utilized by SCIA. A notable question is that the intersection ratio of the results obtained from SCIA and GSEA is decreasing with the increasing of inter-gene correlation, because GSEA is more sensitive in finding significant pathways with less but consistent expression changes. This result indicated that SCIA and GSEA could find different types of differentially expressed gene sets.

SCIA Has Higher Sensitivity and Lower FDR Than Two Competitive Methods Under High Percentages of DEGs in Background Genes

When the percentages of DEGs in background genes are high, there are likely to be relatively high overlaps between a given gene set and background DEGs. Therefore, self-contained methods are invalid in scenario 3 and SCIA was compared with competitive methods. **Table 4** shows that SCIA had higher sensitivity than the other two methods and, interestingly, the FDR was negatively correlated with the percentage of DEGs in background genes. These results are reasonable and reflect the incorporation of different GSA methods in SCIA. Like other competitive methods, when the percentage of DEGs in background genes was high, SCIA assigned a competitive penalty of the significance to the given pathway, and when the percentage of DEGs in background genes was low, SCIA assumed only a few percentages of the DEGs would produce a significant result for the given pathway because there was no other explanation for these DEGs. Notably, in complex diseases such as cancer, DEGs usually account for more than 40% of the genes in a dataset, under which condition SCIA was the best method both in sensitivity and FDR.

SCIA Has Higher Sensitivity Than the Two Self-Contained Methods in Testing Changes of Inter-gene Correlations

Most competitive methods cannot simultaneously test node and edge changes; hence, we compared SCIA with self-contained methods under scenario 4 with the same H_0 hypothesis and FDRs (**Table 2**) as scenario 1. The influence of different sample sizes was measured at the same time. **Figure 2** shows that SCIA had the highest sensitivity and the slowest drop in sensitivity with decreasing sample sizes. However, when the sample size was 10 pairs, the sensitivity of SCIA dropped sharply because of the approximation of chi-square distribution (see method), which needs sample sizes of 15–30 pairs. Unsurprisingly, ROAST had the lowest sensitivity because it was not designed for this purpose. Besides, although the edge testing modules of SCIA and NES are quite similar, SCIA was more sensitive because edge changes are also considered by Hotelling's T^2 (see method), indicating SCIA does not simply superpose node testing and edge testing methods like NES.

TABLE 3 | SCIA has lower FDRs than the competitive methods under different inter-gene correlations in simulated datasets.

Pearson correlation coefficients	FDR			Sensitivity		
	SCIA	CAMERA	GSEA	SCIA	CAMERA	GSEA
0.0	0.016	0.048	0.042	0.286	0.183	0.126
0.3	0.018	0.065	0.112	0.257	0.287	0.304
0.6	0.029	0.104	0.216	0.256	0.442	0.529
0.9	0.033	0.381	0.424	0.304	0.821	0.297

TABLE 4 | SCIA has higher sensitivity than the competitive methods under different percentages of DEGs in background genes.

Proportion	FDR			Sensitivity		
	SCIA	CAMERA	GSEA	SCIA	CAMERA	GSEA
0.2	0.157	0.124	0.188	0.760	0.580	0.507
0.4	0.112	0.138	0.161	0.788	0.559	0.513
0.6	0.093	0.151	0.169	0.816	0.528	0.413

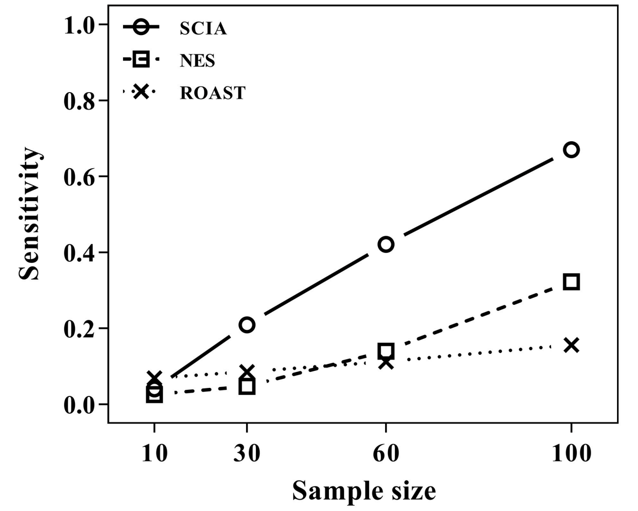


FIGURE 2 | SCIA performs better in edge testing than the self-contained methods under different inter-gene correlations and different sample sizes.

Evaluation of SCIA Performance With Real Datasets

We applied SCIA to recover differentially expressed genes and pathways involved in lung squamous cell carcinoma (LUSC), a common type of non-small-cell lung cancer using two datasets, one from the NCBI's GEO (Gene Expression Omnibus) and one from TCGA (The Cancer Genome Atlas) database. The GEO dataset (Series Accession: GSE103512, Brouwer-Visser et al., 2017) contains 23 LUSC sub-type cancer samples and 9 normal samples. The LUSC dataset from TCGA contains 502 LUSC samples and 51 normal samples.

The two LUSC datasets were used as input to compare the sensitivity and robustness of SCIA and GSEA. In the CSSPN-base permutation procedure of SCIA, all the genes were mapped to the HPRD network, then the top 2% of DEGs (about 200 in each dataset) were defined as the initial and terminal genes of CSSPN (see method). All the nodes in the CSSPN were used for classical functional enrichment analysis based on a hypergeometric test. Unlike the simulation studies, the adjustment of permutation *p*-values (see method) should be moderate here. This is because, under the H_0 hypothesis of simulation studies, there is no relation between the background network and the given gene set, whereas, in real organisms, hundreds of genes in the background network will differentially expressed in response to the DEGs in the given gene set. Due to the *C* statistic *p*-values of all the single pathways were already Benjamini and Hochberg (1995) adjusted, we did not adjust the permutation *p*-value in the following analysis, indicating there are approximate 500 genes in the HPRD background network that, on average, are affected by the terminal DEGs. This *p*-value threshold is a parameter of SCIA and can be set as different scores according to different data and requirements.

The results of the KEGG functional enrichment analysis are shown in **Supplementary Tables S1–S4**. SCIA found 131 and 64 pathways and GSEA found 46 and 40 pathways in the GSE103512 and TCGA LUSC datasets, respectively. Among them, 55 (42%) SCIA pathways were common between the two datasets, whereas only 5 (11%) of the GSEA pathways were common between the two datasets. These results illustrated that there was little comparability between the two results of GSEA, while, SCIA could demonstrate common results in different lung cancer datasets and the individual differences in the two researches, implying the two results of SCIA with different datasets were comparable. More than 33 of the 55 SCIA pathways found in both of the two datasets have been reported previously to have relationships with lung cancer (**Table 5**), including the non-small cell lung cancer. While, most of these pathways were not detected by GSEA. This result showed that SCIA could find many positive pathways that GSEA could not, and the high proportion of results with literature supporting indicated that the intersection of results of SCIA with different datasets could increase the reliability. Further, SCIA produces a CSSPN, which can be considered simply as a set of DEGs. SCIA detected 41 DEGs in the two datasets, and more than 27 (**Supplementary Table S5**) of these genes have been reported previously to be related with lung cancer.

TABLE 5 | SCIA found more literature supported KEGG pathways than GSEA in two non-small-cell lung cancer datasets.

KEGG pathway name	Adjusted *p*-value of SCIA	GSEA
Cell cycle	3.89E-45	Yes
Cellular senescence	3.99E-12	No
Epstein-Barr virus infection	2.31E-11	Yes
Viral carcinogenesis	5.59E-10	Yes
p53 signaling pathway	4.81E-09	Yes
FoxO signaling pathway	1.19E-08	No
Platinum drug resistance	2.16E-07	Yes
Hepatitis B	1.43E-06	No
Transcriptional misregulation in cancer	1.92E-06	No
Small cell lung cancer	5.74E-06	No
Human papillomavirus infection	1.39E-05	No
MicroRNAs in cancer	1.62E-05	No
Glioma	3.25E-05	No
Kaposi's sarcoma-associated herpesvirus infection	3.10E-05	Yes
Apoptosis	3.51E-05	No
Non-small cell lung cancer	5.11E-05	No
Hepatocellular carcinoma	9.52E-05	No
Hippo signaling pathway	0.0001275	No
TGF-beta signaling pathway	0.0004040	No
Adherens junction	0.0006536	No
PI3K-Akt signaling pathway	0.0006624	No
Proteoglycans in cancer	0.0058405	No
Wnt signaling pathway	0.0084030	No
AGE-RAGE signaling pathway in diabetic complications	0.0151588	No
HIF-1 signaling pathway	0.0302121	No
Hepatitis C	0.0339220	No
Basal cell carcinoma	0.0343406	No
Mitophagy—animal	0.0362401	No
ErbB signaling pathway	0.0418948	No
Insulin resistance	0.0418948	No
Apoptosis—multiple species	0.0427196	No
Measles	0.0427196	No
Amyotrophic lateral sclerosis (ALS)	0.0427196	No

"Yes" means the pathway is found by both SCIA and GSEA with adjusted p-value < 0.05. "No" means the pathway is found by SCIA but not by GSEA.

DISCUSSION

SCIA is the first GSA method that combines the advantages of self-contained, competitive, and PT-based methods. SCIA has three main advantages over the other methods as was shown by the simulation studies. First, SCIA is powerful and statistically rigorous under high inter-gene correlations, which are conditions under which most competitive methods lose control of FDR. Second, SCIA has higher sensitivity and minimum FDR compared to two competitive methods (GSEA, CAMERA) under a high proportion of DEGs in background genes, which are conditions that make most self-contained methods invalid. Moreover, SCIA uses an *a priori* biological network and performs better than ROAST and NES in testing edge (inter-gene correlation) changes. Overall, the FDR of SCIA was well-controlled and its sensitivity was higher than that of the other four methods tested (GSEA, CAMERA, ROAST, and NES) under most simulated conditions, highlighting the extensive applicability and unbiased results of SCIA.

The robustness of SCIA can be attributed to two aspects. First, its extensive applicability with reliable and unbiased results, as mentioned above, are the most important reasons. Second, through the CSSPN-based permutation strategy in SCIA, a reasonable hypothesis is innovatively combined with *a priori* biological information. Briefly, if DEGs can be mapped only in one gene set, a positive weight is added to them because there is no other explanation for the differential expressions of these genes. Therefore, for SCIA, comprehensiveness of the background networks is more important than its accuracy. However, when the *a priori* biological networks are more comprehensive, the hypothesis of SCIA becomes more reasonable and the results are more precise. This robustness gives SCIA the ability to calculate with different datasets and to integrate the results of SCIA with different datasets.

There are many potential applications for SCIA, including differential expression analysis (Dona et al., 2017), sub-pathway analysis (Martini et al., 2013), and micorRNA target gene prediction (Wang, 2008). First, all of the genes in the CSSPN can be considered as DEGs and used independently. In addition, CSSPN itself can be considered as a cascading effect pathway when the input data are from a knockout/over-expression experiment of a single gene. Second, if the function of differential pathways can be biologically confirmed, the sub-pathway of the given functional pathway can be built without the permutation procedure. Third, the choice of initial gene set is very flexible and can be tailored for different purposes. For instance, if the input data are derived from a microRNA knockout/over-expression experiment, the initial gene set can be select as the predicted target genes of the microRNAs, and the significant predicted targets will have more potential to be the targets of these microRNA in a specific experimental condition.

DATA AVAILABILITY

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=GSE103512; https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

AUTHOR CONTRIBUTIONS

YuL and QJ designed the experiments. YiL, YW, YB, XZ, and JZ performed the experiments and data analysis. LA, XL, and MS have contributed to the writing of this article

FUNDING

This work was supported by the National Nature Science Foundation of China [31571323, 61571152, 81703322] and the National Science and Technology Major Project [2016YFC1202302].

ACKNOWLEDGMENTS

We thank Margaret Biswas, Ph.D., from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00598/full#supplementary-material

REFERENCES

Ackermann, M., and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10:47. doi: 10.1186/1471-2105-10-47

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. (2005). Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21, 2988–2993. doi: 10.1093/bioinformatics/bti457

Barry, W. T., Nobel, A. B., and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21, 1943–1949. doi: 10.1093/bioinformatics/bti260

Beissbarth, T., and Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465. doi: 10.1093/bioinformatics/bth088

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Brouwer-Visser, J., Cheng, W. Y., Bauer-Mehren, A., Maisel, D., Lechner, K., Andersson, E., et al. (2017). Regulatory T-cell genes drive altered immune microenvironment in adult solid cancers and allow for immune contextual patient subtyping. *Cancer Epidemiol. Biomark. Prev.* 27, 103–112. doi: 10.1158/1055-9965.EPI-17-0461

Dezso, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C., et al. (2009). Identifying disease-specific genes based on their topologi-cal significance in protein networks. *BMC Syst. Biol.* 3:36. doi: 10.1186/1752-0509-3-36

Dona, M. S., Prendergast, L. A., Mathivanan, S., Keerthikumar, S., and Salim, A. (2017). Powerful differential expression analysis incorporating network topology for next-generation sequencing data. *Bioinformatics* 33, 1505–1513. doi: 10.1093/bioinformatics/btw833

Efron, B., and Tibshirani, R. (2006). On testing the significance of sets of genes. *Ann. Appl. Stat.* 1, 107–129. doi: 10.1214/07-AOAS101

Fang, Z., Tian, W., and Ji, H. (2012). A network-based gene-weighting approach for pathway analysis. *Cell Res.* 22, 565–580. doi: 10.1038/cr.2011.149

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I., and Wright, F. A. (2010). Heading down the wrong pathway: on the influence of correlation within gene sets. *BMC Genomics* 11:574. doi: 10.1186/1471-2164-11-574

Goeman, J. J., van de Geer, S. A., de Kort, F., and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20, 93–99. doi: 10.1093/bioinformatics/btg382

Gu, Z., Liu, J., Cao, K., Zhang, J., and Wang, J. (2012). Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst. Biol.* 6:56. doi: 10.1186/1752-0509-6-56

Han, J., Shi, X., Zhang, Y., Xu, Y., Jiang, Y., Zhang, C., et al. (2015). ESEA: Discovering the dysregulated pathways based on edge set enrichment analysis. *Sci. Rep.* 5:13044. doi: 10.1038/srep13044

Jiang, Z., and Gentleman, R. (2007). Extensions to gene set enrichment. *Bioinformatics* 23, 306–313. doi: 10.1093/bioinformatics/btl599

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Khatri, P., and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587–3595. doi: 10.1093/bioinformatics/bti565

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi: 10.1371/journal.pcbi.1002375

Kim, S. Y., and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6:144. doi: 10.1186/1471-2105-6-144

Kong, S. W., Pu, W. T., and Park, P. J. (2006). A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics* 22, 2373–2380. doi: 10.1093/bioinformatics/btl401

Library, W. P., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database. *Nucleic Acids Res.* 37, D767–D772. doi: 10.1093/nar/gkn892

Mansmann, U., and Meister, R. (2005). Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.* 44, 449–453. doi: 10.1055/s-0038-1633992

Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2013). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res.* 41:e19. doi: 10.1093/nar/gks866

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., et al. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.* 4:278. doi: 10.3389/fphys.2013.00278

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273. doi: 10.1038/ng1180

Nam, D. (2010). De-correlating expression in gene-set analysis. *Bioinformatics* 26, i511–i516. doi: 10.1093/bioinformatics/btq380

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. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13544–13549. doi: 10.1073/pnas.0506577102

Wang, X. (2008). miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *RNA* 14, 1012–1017. doi: 10.1261/rna.965408

Wu, D., Lim, E., Vaillant, F., Asselin-Labat, M. L., Visvader, J. E., and Smyth, G. K. (2010). ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* 26, 2176–2182. doi: 10.1093/bioinformatics/btq401

Wu, D., and Smyth, G. K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* 40:e133. doi: 10.1093/nar/gks461

Yuan, Z., Ji, J., Zhang, T., Liu, Y., Zhang, X., Chen, W., et al. (2016). A novel chi-square statistic for detecting group differences between pathways in systems epidemiology. *Stat. Med.* 35, 5512–5524. doi: 10.1002/sim.7094

Zyla, J., Marczyk, M., and Polanska, J. (2017). Reproducibility of finding enriched gene sets in biological data analysis. *Int. Conf. Pract. Appl. Comput. Biol. Bioinformatics* 146–154. doi: 10.1007/978-3-319-60816-7_18

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

ORIGINAL RESEARCH
published: 10 September 2019
doi: 10.3389/fgene.2019.00793

iDNA6mA-Rice: A Computational Tool for Detecting N6-Methyladenine Sites in Rice

Hao Lv[1], Fu-Ying Dao[1], Zheng-Xing Guan[1], Dan Zhang[1], Jiu-Xin Tan[1], Yong Zhang[1]*, Wei Chen[2]* and Hao Lin[1]*

[1] Key Laboratory for Neuro-Information of Ministry of Education, School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [2] Innovative Institute of Chinese Medicine and Pharmacy, Chengdu University of Traditional Chinese Medicine, Chengdu, China

DNA N6-methyladenine (6mA) is a dominant DNA modification form and involved in many biological functions. The accurate genome-wide identification of 6mA sites may increase understanding of its biological functions. Experimental methods for 6mA detection in eukaryotes genome are laborious and expensive. Therefore, it is necessary to develop computational methods to identify 6mA sites on a genomic scale, especially for plant genomes. Based on this consideration, the study aims to develop a machine learning-based method of predicting 6mA sites in the rice genome. We initially used mono-nucleotide binary encoding to formulate positive and negative samples. Subsequently, the machine learning algorithm named Random Forest was utilized to perform the classification for identifying 6mA sites. Our proposed method could produce an area under the receiver operating characteristic curve of 0.964 with an overall accuracy of 0.917, as indicated by the fivefold cross-validation test. Furthermore, an independent dataset was established to assess the generalization ability of our method. Finally, an area under the receiver operating characteristic curve of 0.981 was obtained, suggesting that the proposed method had good performance of predicting 6mA sites in the rice genome. For the convenience of retrieving 6mA sites, on the basis of the computational method, we built a freely accessible web server named iDNA6mA-Rice at http://lin-group.cn/server/iDNA6mA-Rice.

Keywords: N6-methyladenine, mono-nucleotide binary encoding, random forest, cross-validation, web-server

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University,
China

Reviewed by:
Jianzhao Gao,
Nankai University, China
Xiangxiang Zeng,
Xiamen University, China

***Correspondence:**
Yong Zhang
zhangyong916@uestc.edu.cn
Wei Chen
chenweiimu@gmail.com
Hao Lin
hlin@uestc.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *13 June 2019*
Accepted: *26 July 2019*
Published: *10 September 2019*

Citation:
Lv H, Dao F-Y, Guan Z-X, Zhang D, Tan J-X, Zhang Y, Chen W and Lin H (2019) iDNA6mA-Rice: A Computational Tool for Detecting N6-Methyladenine Sites in Rice. Front. Genet. 10:793. doi: 10.3389/fgene.2019.00793

INTRODUCTION

Methylated bases, such as N4-methylcytosine (4mC), N6-methyladenine (6mA), and 5-methylcytosine (5mC), exist in genomic DNA of diverse species (Cheng, 1995; Ratel et al., 2006). All these DNA methylation modifications play important roles in controlling many biological functions (Tang et al., 2018b). As an epigenetic mechanism, DNA methylation refers to a process that methyl groups are transferred to DNA molecules and is essential in the normal development of organisms (Bergman and Cedar, 2013; Smith and Meissner, 2013; von Meyenn et al., 2016). Through DNA methylation, the activity of a DNA segment can be changed without changing its sequence. For example, gene transcription can be repressed when DNA methylation occurs at its promoter (Bird, 1992).

As shown in **Figure 1**, after a methyl group is transferred to the sixth position of adenine ring, under the catalysis action of methyltransferases, 6mA is formed. 6mA is a noncanonical DNA

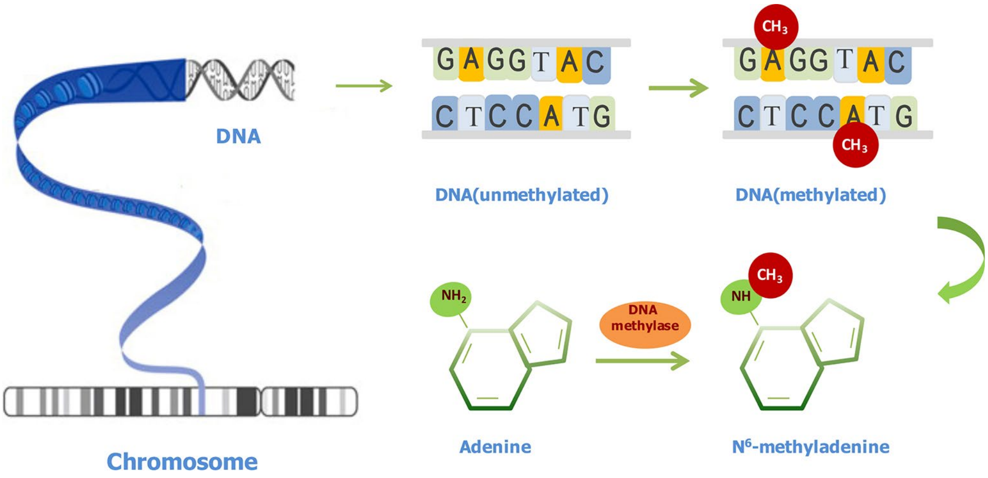


FIGURE 1 | Illustration of N6-methyladenine (6mA) modifications in DNA. The conversion of adenine to 6mA is mediated by methyl-transferases.

modification form in different eukaryotes at low levels (Fu et al., 2015; Greer et al., 2015; Zhang et al., 2015; Koziol et al., 2016; Liu et al., 2016; Mondo et al., 2017; Wang et al., 2017). 6mA in prokaryotes and eukaryotes shows similar characteristics (Heyn and Esteller, 2015). It has diverse functions, including guiding the discrimination of an original DNA strand from a newly synthesized DNA strand (Wion and Casadesus, 2006), regulating gene transcription (Cheng et al., 2016), repressing transposable elements, and reducing the stability of base pairings (Fang et al., 2012). Surprisingly, the methylation protection is an inheritable state, although it may be changed by environmental factors (Wion and Casadesus, 2006). Therefore, it is worth underscoring the importance of 6mA throughout generations.

Recent studies revealed the genome-wide distributions of 6mA in *Tetrahymena* (Wang et al., 2017), *Chlamydomonas reinhardtii* (Fu et al., 2015), *Drosophila melanogaster* (Zhang et al., 2015), *Caenorhabditis elegans* (Greer et al., 2015), vertebrates (e.g. frog and fish) (Koziol et al., 2016; Liu et al., 2016), mammals (e.g., human and *Mus. musculus*) (Wu et al., 2016; Yao et al., 2017; Xiao et al., 2018; Zou et al., 2018a), fungi (Mondo et al., 2017), and vascular plants (e.g. rice) (Zhou et al., 2018). Although these studies testified the presence of 6mA in eukaryotic genomes based on experimental means and indeed achieved encouraging results, the implication of 6mA in epigenetics is still obscure (Ratel et al., 2006). In addition, in eukaryotes, the level of 6mA was so low that it could only be detected by advanced techniques. In rice, with two antibodies, based on SMRT and IP-seq, Zhou et al. (2018) found that AGG-rich sequences were the most significantly enriched for 6mA. Thus, the computational prediction of 6mA sites may be a good choice to reduce experimental costs and guide the experimental study on plant 6mA.

In fact, several computational methods have been applied in the identification of DNA methylation sites. Based on the data of experimentally confirmed 4mC sites, Chen et al. (2017) firstly developed a predictor called iDNA4mC to identify 4mC sites, in which DNA samples were formulated with nucleotide frequency and nucleotide chemical property. Then, based on the dataset (Chen et al., 2017), He et al. (2018a) established another tool named 4mCPred, and Wei et al. (2018b) built a new predictor (4mcPred-SVM) to predict 4mC sites. Recently, a free tool called iDNA6mA-PseKNC was constructed for the computational prediction of 6mA sites (Feng et al., 2019). The tool could be used to identify 6mA sites in *Mus. musculus* genome. However, the tool could not provide valuable data contained in plant genomes due to the difference between mammal and plant genomes. Thus, it is necessary to develop a 6mA site predictor for plant genomes. Recently, a tool named i6mA-Pred was constructed to identify 6mA site in rice (Chen et al., 2019). The tool could realize the area under the receiver operating characteristic curve (auROC) of 0.886 in jackknife cross-validation. However, the database used was not large enough, and the accuracy should be further improved.

In view of the aforementioned descriptions, this study aims to develop a new method and establish an efficient tool to identify 6mA sites in the rice genome. A flowchart is shown in **Figure 2**. We firstly collected the existing data in the rice genome, including experimentally confirmed non-6mA sequences and 6mA sequences and built a benchmark dataset based on the report by Zhou et al. (2018). Subsequently, three kinds of sequence encoding features were proposed to formulate samples as the input of the Random Forest algorithm (RF) to discriminate 6mA sequences from non-6mA sequences. Then, several experiments were performed to investigate the prediction capability of the proposed method. Finally, on the basis of the method, we established a predictor called iDNA6mA-Rice.

MATERIALS AND METHODS

Benchmark Dataset

A benchmark dataset is important in building a reliable prediction model. By combining immunoprecipitation with single-molecular real-time sequencing approach, 6mA sites

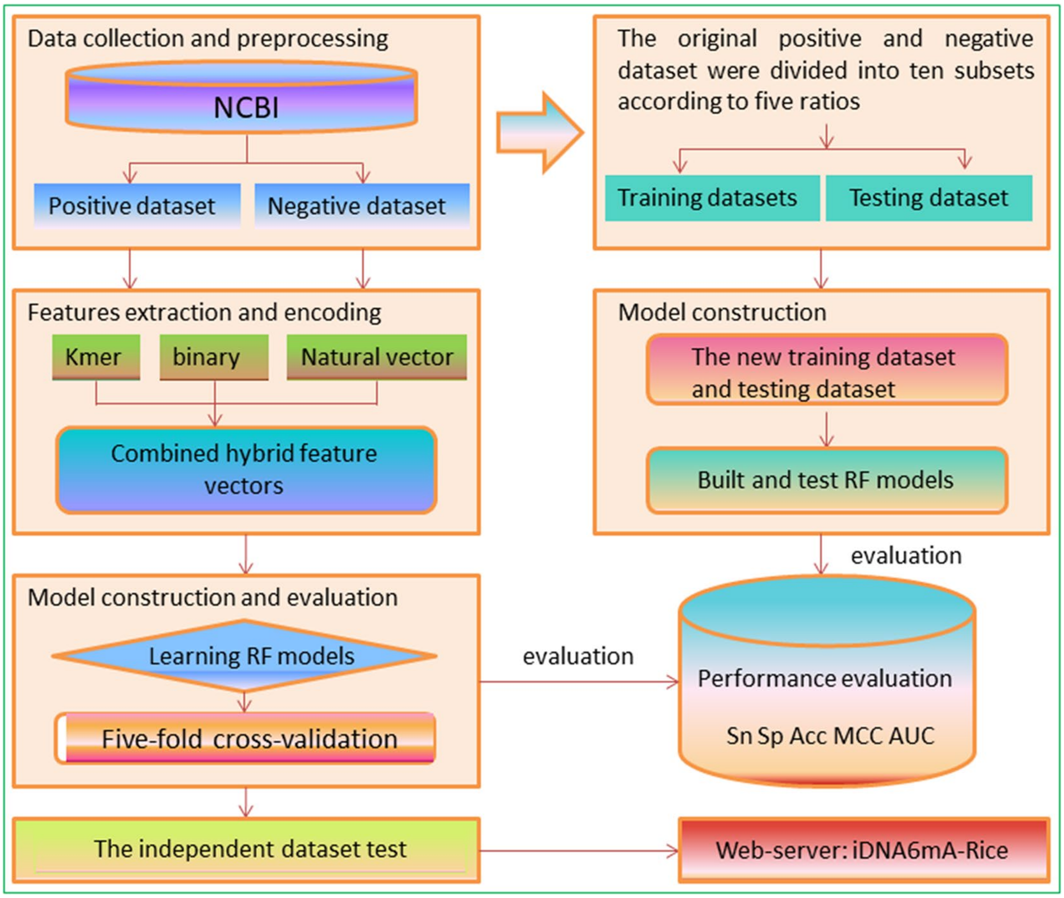

Data collection and preprocessing
NCBI
Positive dataset
Negative dataset
The original positive and negative dataset were divided into ten subsets according to five ratios
Training datasets
Testing dataset
Features extraction and encoding
Kmer
binary
Natural vector
Combined hybrid feature vectors
Model construction
The new training dataset and testing dataset
Built and test RF models
evaluation
Model construction and evaluation
Learning RF models
Five-fold cross-validation
evaluation
Performance evaluation
Sn Sp Acc MCC AUC
The independent dataset test
Web-server: iDNA6mA-Rice

FIGURE 2 | A flowchart used in this study.

in the rice genome had been detected (Zhou et al., 2018) and deposited in Gene Expression Omnibus (GEO) database, which was created and is maintained by the National Center for Biotechnology Information (NCBI) (Long et al., 2019). Therefore, a total of 265,290 6mA sites containing sequences were obtained from GEO. All of these sequences in GEO are 41 nt long with the 6mA site at the center. To reduce homologous bias and avoid redundancy (Dao et al., 2018; Su et al., 2018; Tang et al., 2018a; Zou et al., 2018b; Feng et al., 2019), sequences with the similarity above 80% were excluded by using the CD-HIT program (Li and Godzik, 2006). Finally, we obtained 154,000 6mA sites-contained sequences as positive samples.

Negative samples were collected from NCBI (https://www.ncbi.nlm.nih.gov/genome/10) and according to the following three rules. Firstly, the 41-nt long sequences with adenine at the center were selected. Secondly, experimental results proved that the centered adenine was not methylated. Thirdly, Zhou et al. (2018) believed that 6mA most frequently occurred at GAGG, AGG, and AG motifs, so we statistically analyzed the ratios of GAGG, AGG, and AG motifs in positive samples and reported the result in **Table 1**. Based on the result in **Table 1**, we selected the negative samples with the same ratio of motifs so that the negative data were more objective. In this way, a large number of negative samples were obtained. In machine learning processes, imbalanced datasets lead to unreliable results. To balance positive and negative samples, 154,000 non-modified segments were randomly picked out as negative samples in model training. Finally, the benchmark dataset contained 154,000 positive samples and 154,000 negative samples. The benchmark dataset **S** is formulated as:

TABLE 1 | Details of the three motifs in positive samples.

Motifs	Numbers	Proportions (%)
GAGG	26,300	17.08
AGG	24,264	15.76
AG	22,206	14.42

$$\mathbf{S} = \mathbf{S}^{+} \cup \mathbf{S}^{-} \quad (1)$$

where the $\mathbf{S}^{+}$ contains 154,000 positive samples; the $\mathbf{S}^{-}$ contains 154,000 negative samples; $\cup$ is the symbol of "union" in the set theory. The benchmark dataset is available at http://lin-group.cn/server/iDNA6mA-Rice.

Feature Descriptions

Feature extraction is a key step in establishing an excellent predictor (Song et al., 2012; Zuo et al., 2017; Stephenson et al., 2018; Manavalan et al., 2018a; Wei et al., 2018a; Manavalan et al., 2018b; Song et al., 2018b; Song et al., 2018c). The following three feature extraction techniques were adopted to formulate 6mA samples.

K-tuple Nucleotide Frequency Component

As a special form of PseKNC (Guo et al., 2014; Lin et al., 2014), the K-tuple nucleotide frequency component has been widely used in a variety of bioinformatics problems (Lin and Li, 2011; Yang et al., 2018b).

A DNA sequence **D** can be expressed as:

$$\mathbf{D} = R_1 R_2 R_3 R_4 \cdots R_i \cdots R_{L-1} R_L, \tag{2}$$

where R_i represents the nucleotide [Adenine (A), Thymine (T), Cytosine (C), and Guanine (G)] at the *i*th position; L is the length of sequence **D** and equals to 41 in this study. The strategy of k-tuple composition is to convert each sample into a 4^k dimension vector expressed as:

$$\mathbf{D} = \left[f_1^{k-tuple} f_2^{k-tuple} \cdots f_i^{k-tuple} \cdots f_{4^k}^{k-tuple} \right]^T \tag{3}$$

where T represents the transposition of the vector and $f_i^{k-tuple}$ represents the frequency of the *i*th *k*-tuple composition in the DNA sequence sample. The feature has been applied in DNA element identification (Wei et al., 2018b). Here, we set $k = 2, 3, 4$.

Mono-Nucleotide Binary Encoding

The second feature technique is to transfer nucleotide into a binary code formulated as:

$$n = \begin{cases} (1,0,0,0), & when\ n = A \\ (0,1,0,0), & when\ n = C \\ (0,0,1,0), & when\ n = G \\ (0,0,0,1), & when\ n = T \end{cases} \tag{4}$$

Thus, an arbitrary DNA sequence with L nucleotides can be described as a vector of $4 \times L$ features (Song et al., 2018a; Wei et al., 2018b).

Natural Vector

In the natural vector method proposed by Deng et al. (2011), sequences are represented as points in high-dimensional space based on statistical characteristics (Liu et al., 2018). With the sequence data, such as occurrence frequencies, the central moments, and average positions of nucleotides, the natural vector method is used to describe the distributions and numbers of nucleotides, cluster sequences, and predict their various attributes.

Based on Eq. (3), each nucleotide R can be defined as follows:

$$W_k(\cdot): \{\text{A,C,G,T}\}, \rightarrow \{0,1\}, \tag{5}$$

where $W_R\ (R_i) = 1$ if $D_i = R$ and $W_R\ (D_i) = 0$, otherwise

$$n_R = \sum\nolimits_{i=1}^{n} W_R(D_i), \tag{6}$$

where n_R represents the number of nucleotide R in the DNA sequence D:

$$S_{[R][i]} = i \cdot W_R(D_i), \tag{7}$$

where $S_{[R][i]}$ represents the distance from the first nucleotide to the *i*th nucleotide R.

$$T_R = \sum\nolimits_{i=1}^{n_R} S_{[R][i]}, \tag{8}$$

where T_R represents the total distance of each set of the four nucleotides.

$$\mu_R = T_R / n_R, \tag{9}$$

where μ_R represents the mean position of the nucleotide R.

Finally, the second-order normalized central moments can be defined as:

$$D_2^R = \sum\nolimits_{i=1}^{n_R} \frac{(S_{[R][i]} - \mu_R)^2}{n n_R} \tag{10}$$

Then, the natural vector of sequence D is expressed as (Tian et al., 2018):

$$\left(n_A, \mu_A, D_2^A, n_c, \mu c, D_2^C, n_G, \mu_G, D_2^G, n_T, \mu_T, D_2^T \right). \tag{11}$$

Random Forest Algorithm

The RF algorithm has been extensively applied in computational biology (Zhao et al., 2014; Zhang et al., 2016; Lv et al., 2019), since it is a flexible and practical machine learning method and can deal with many input variables without variable deletion and provide an internal unbiased estimate of the generalization error. According to the principle of RF, many trees are randomly generated with the recursive partitioning approach, and then, the results are aggregated according to voting rules. In this study, the number of trees is set to 100 with the seed of 1. The details of RF had been described by Breiman (2001).

Performance Evaluation

Cross-validation test is a statistical analysis method for assessing a classifier. For the purpose of saving computation time, the fivefold cross-validation test was performed to assess the method proposed in this study. We used four metrics [Matthew's correlation coefficient (*MCC*), sensitivity (*Sn*), overall accuracy (*Acc*), and specificity (*Sp*)] to measure the predictive capability of our model (Zuo et al., 2014; Zou et al., 2016; Manavalan and Lee, 2017; Manavalan et al., 2017; Cao et al., 2017a; Cao et al., 2017b; Cheng et al., 2018a; Yang et al., 2018a; Zhu et al., 2019).

$$\begin{cases} Sn = 1 - \dfrac{N_-^+}{N^+} & 0 \le Sn \le 1 \\ Sp = 1 - \dfrac{N_+^-}{N^-} & 0 \le Sp \le 1 \\ Acc = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le Acc \le 1 \\ MCC = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+}\right)\left(1 + \dfrac{N_-^+ - N_+^-}{N^-}\right)}} & 0 \le MCC \le 1 \end{cases}, \tag{12}$$

where N^+ and N^- are, respectively, the numbers of 6mA sites and non-6mA sites in benchmark dataset; N_-^+ indicates the number of the 6mA sites recognized as non-6mA sites; and N_+^- indicates the number of the wrongly predicted non-6mA sites. *Sn* and *Sp* represent the ability of a model to correctly identify 6mA sites and non-6mA sites, respectively. The value of *Acc* indicates the overall accuracy of our model distinguishing 6mA sites from non-6mA sites. *MCC* indicates the performance of our model based on real and predicted values. When $N_-^+ = N_+^- = 0$, meaning that none of the 6mA sites in the dataset S^+ and none of the non-6mA sites in the dataset S^- was mispredicted, we have *MCC* = 1; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, we have *MCC* = 0, meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$ we have *MCC* = -1, meaning total disagreement between prediction and observation.

In addition to the analysis based on the previously discussed indicators, the ROC curves (Metz, 1989; Chen et al., 2016; Dao et al., 2018; Feng et al., 2018; Lai et al., 2019; Tan et al., 2019) were plotted, and then, the area under the receiver operating characteristic curve (AUC) was calculated to objectively evaluate our proposed model.

RESULTS AND DISCUSSION

Sequence Analysis

To investigate the nucleotide distribution around the 21st site (6mA or non 6mA) in positive and negative samples, the pLogo (O'Shea et al., 2013) was plotted to analyze the statistical difference of nucleotide occurrence between two kinds of samples. The 6mA samples were dramatically different from non-6mA samples in terms of nucleotide compositions (**Figure 3**). The nucleotide composition bias regions existed in the ranges from -8 to +10 sites and from +15 to +18 downstream of the 6mA site. Unlike the distribution in the non-6mA samples, a consensus motif of AAAA was observed in the upstream of the 6mA site. These results suggested that it was feasible to construct a machine learning model for identifying 6mA sites with extracted sequence features.

Performance Evaluation on Different Features

The prediction performances of three features [K-tuple nucleotide frequency component (KNFC), mono-nucleotide binary encoding (MNBE), and natural vector (NV)] and their combinations were firstly explored with RF. Accordingly, we built four computational models and evaluated them through the fivefold cross-validation test. The prediction results are provided in **Figure 4** and **Table 2**. It was found that MNBE could produce the best prediction performance among all features, indicating that it was the best descriptor for 6mA samples.

KNFC is a commonly used feature extractor technique and has been successfully applied in DNA regulatory element prediction. However, the results in **Table 2** showed that the accuracy of KNFC was only 68.3%, which was far from satisfactory. For the 41-nt long 6mA samples, KNFC is a high-dimension vector (16 + 64 + 256), which is so large that many elements in feature vector are zero. Although

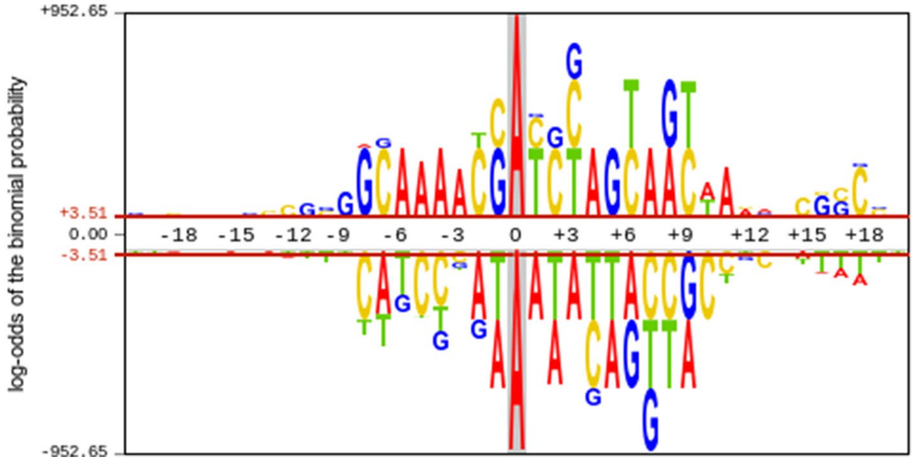


FIGURE 3 | Nucleotide distribution preferences around 6mA and non-6mA sites. The upper half of the x-axis indicates the nucleotide distribution in 6mA site containing sequence, whereas the lower half of the x-axis indicates the nucleotide distribution in non-6mA site containing sequences.

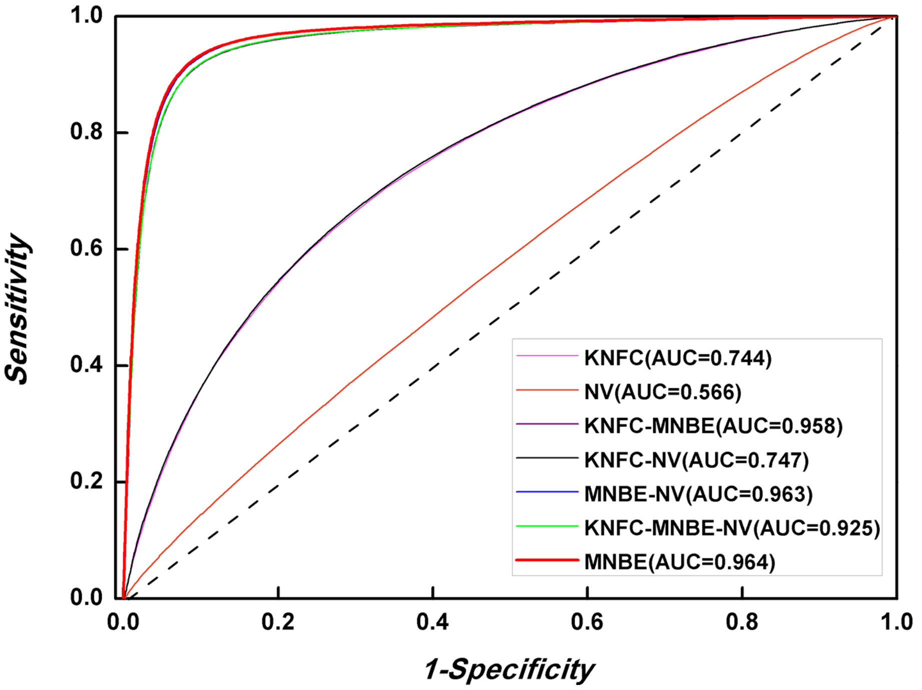


FIGURE 4 | Performance evaluation based on three features and their combinations.

TABLE 2 | Predictive performances of KNFC, MNBE, and NV.

Methods	*Sn* (%)	*Sp*(%)	*Acc*(%)	MCC	AUC
KNFC (k = 2, 3, 4)	70.3	66.3	68.3	0.366	0.744
MNBE	93.0	90.5	91.7	0.835	0.964
NV	58.1	50.6	54.3	0.087	0.566
KNFC-MNBE	91.8	90.1	90.9	0.819	0.958
KNFC-NV	70.4	66.5	68.4	0.369	0.747
MNBE-NV	92.8	90.3	91.6	0.832	0.963
KNFC-MNBE-NV	91.7	90.3	91.0	0.820	0.925

high-dimension features contain more information, more noise and redundant information are also included, thus decreasing the discrimination capability. Therefore, KNFC is not suitable for 6mA identification. In fact, the NV is the worst descriptor among all features in this study, since it can only obtain the overall accuracy of 54.3%, which almost equals the accuracy of random guess. The reason for the poor performance of NV in 6mA prediction is that NV contains too few features to capture enough sequence information of 6mA and non-6mA samples.

For the combinations of different features, if MNBE was included, the prediction performances are always good. However, they are still not higher than those obtained with MNBE alone. Thus, subsequent studies were based on MNBE.

Performance Evaluation of Different Algorithms

It is natural to ask whether other classification is better than RF in 6mA identification. Thus, we investigated the discriminant capabilities of three algorithms, namely, Naïve Bayes, Bayes Net, and Logistic Regression, with the benchmark dataset through fivefold cross-validation. All algorithms were implemented in WEKA (Frank et al., 2004). The ROC curves were plotted (**Figure 5**). It is obvious that RF is the best one for 6mA prediction among four algorithms. Thus, the final model was built with RF.

Performance Evaluation Based on Different Data Ratios

In order to further assess the proposed method, the benchmark dataset was randomly divided into two parts according to five ratios (5:5, 6:4, 7:3, 8:2, and 9:1): training dataset and testing dataset. The former part was used to train the model, whereas the other part was used to test corresponding model. In this way, the training dataset and testing dataset are independent of each other. The predictive results are listed in **Table 3**. For each ratio between training and testing datasets, the model could always

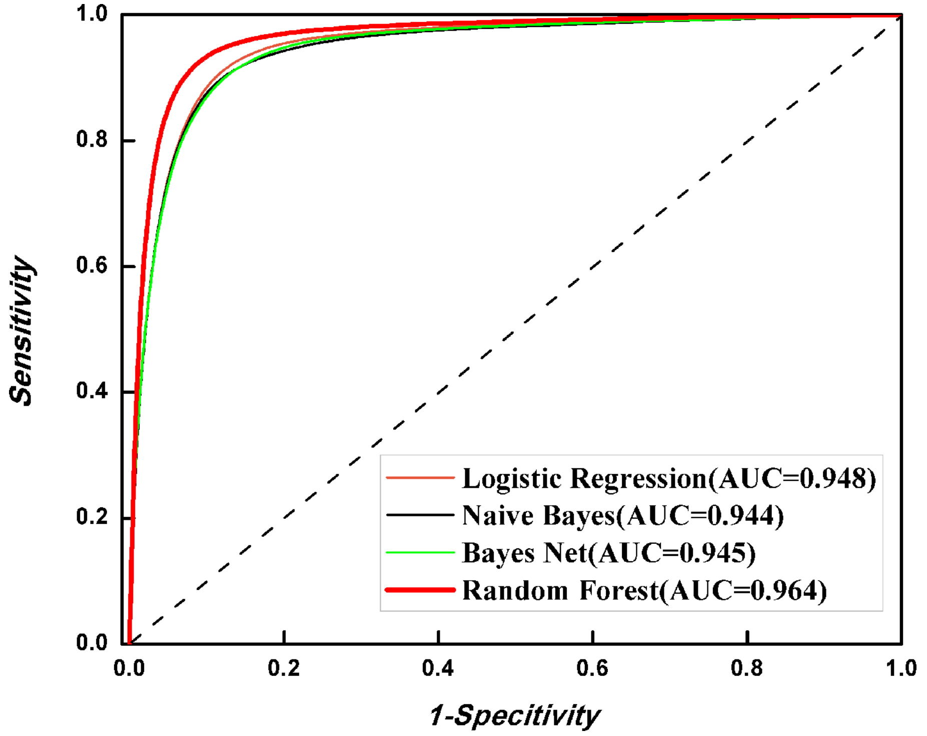


FIGURE 5 | Performance evaluation of different algorithms.

TABLE 3 | Predictive performances of five ratios on the testing and training datasets.

Ratios	**5:5**		**6:4**		**7:3**		**8:2**		**9:1**	
	testing	**training**	**testing**	**training**	**testing**	**training**	**testing**	**training**	**testing**	**training**
Sn (%)	91.4	91.8	92.0	91.9	92.2	92.4	92.4	92.5	92.7	92.7
Sp (%)	70.9	90.5	87.7	90.0	90.6	90.0	91.7	90.1	92.1	90.4
Acc (%)	81.1	91.1	89.9	90.9	91.4	91.2	92.1	91.3	92.2	91.8
MCC	0.636	0.822	0.798	0.819	0.828	0.824	0.841	0.827	0.853	0.835
AUC	**0.904**	**0.969**	**0.953**	**0.963**	**0.963**	**0.963**	**0.967**	**0.963**	**0.969**	**0.964**

produce the AUC of >0.90, suggesting that our method was robust and reliable.

Performance Evaluation With an Independent Dataset

We designed the third experiment to investigate the performance of our proposed predictor. In the experiment, an independent test set was collected from NCBI Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE103145 (Zhou et al., 2018). All the sequences were 41 nt long with the 6mA site at the center. After removing redundant information with CD-HIT program according to the cutoff of 60%, a total of 880 positive samples were obtained (Chen et al., 2019). The negative samples were also obtained from the rice genome. In the report by Zhou et al., 6mA most frequently occurs at GAGG motifs and seldom occurs in coding sequences (CDSs). Thus, negative samples were extracted from CDSs with GAGG motifs in the rice genome. In total, 880 negative samples with the sequence identity less than 60% were obtained. All negative samples were also 41 nt long with non-methylated adenosine at the center. The data were utilized as the benchmark dataset in i6mA-Pred (Chen et al., 2019). The details for the benchmark dataset are available at http://lin-group.cn/server/iDNA6mA-Rice.

We utilized these data to examine our proposed model (**Table 4**). In total, 95.8% 6mA sites and 93.3% non-6mA sites were correctly identified, suggesting that the method was a powerful tool for identifying 6mA sites in rice genome.

Comparison With Published Methods

Till now, i6mA-Pred (Chen et al., 2019) is the only computational-based predictor for 6mA site prediction in the

TABLE 4 | Comparison of different methods for predicting 6mA sites in independent dataset.

Method	*Sn* (%)	*Sp* (%)	*Acc* (%)	MCC	auROC
Our method	95.8	93.3	94.6	0.891	0.981
iDNA6mA-PseKNC	76.6	94.3	85.5	0.721	–

TABLE 5 | Comparison of different methods for predicting 6mA sites in the rice genome with jackknife test.

Methods	*Sn* (%)	*Sp* (%)	*Acc* (%)	MCC	auROC
This study	83.86	83.41	83.63	0.67	0.910
i6mA-Pred	82.95	83.30	83.13	0.66	0.886

rice genome. To provide an objective and strict comparison, we investigated the performance of our method with the same data through jackknife cross-validation. The method could produce the auROC of 0.910 (**Table 5**), which was higher than that of i6mA-Pred. This comparison demonstrated that our method was powerful.

Subsequently, iDNA6mA-PseKNC (Feng et al., 2019) is a tool to identify 6mA sites in *Mus. musculus* genome, and it can identify 6mA sites in many other species with high success rates. Thus, it is necessary to compare our proposed method with it. We investigated the performance of our predictor and iDNA6mA-PseKNC based on the independent dataset used in this work. All compared results were recorded in **Table 4**. It is obvious that the model proposed in this paper is superior to iDNA6mA-PseKNC for identifying 6mA sites.

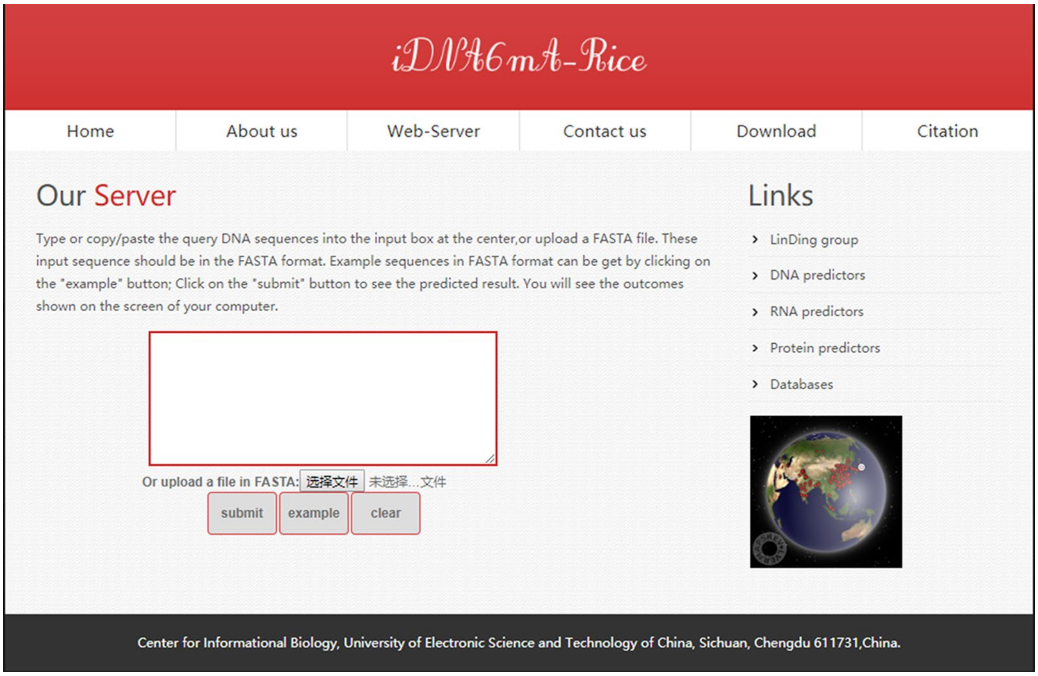
Web Server

Databases and web servers (Wang et al., 2014; Liang et al., 2017; Yi et al., 2017; Zhang et al., 2017; Cui et al., 2018; Dao et al., 2018; Cheng et al., 2018b; He et al., 2018b; Hu et al., 2019; Cheng et al., 2019a; Cheng et al., 2019b) can provide scholars with more convenient services. Thus, the basis of the novel method, we built a web server named iRNA6mA-Rice to identify 6mA sites in the rice genome. The web server can be freely accessible at http://lin-group.cn/server/iDNA6mA-Rice.

Users can open the homepage shown in **Figure 6** to see a short introduction about iDNA6mA-Rice. One may firstly click the “Web-server” button, then type or copy/paste DNA sequences in the input box, or upload the FASTA format file. Note that the length of each sequence should be greater than 41 nt. Subsequently, after clicking the “submit” button, the predicted results will appear on a new page. As described previously, the tool is simple and can provide a convenient way for users to identify putative 6mA sites in DNA of their interest. Moreover, in order to facilitate the processing of large-scale data, the stand-alone package can be downloaded at http://lin-group.cn/server/iDNA6mA-Rice/download.html.

CONCLUSIONS

This paper developed a computational method for the identification of 6mA sites in the rice genome. We designed several kinds of experiments to examine the performance of the proposed method, for example, the performance evaluation on different features, performance evaluation on different algorithms, performance evaluation based on different data ratios, performance evaluation with an independent dataset, and



Our Server

Type or copy/paste the query DNA sequences into the input box at the center,or upload a FASTA file. These input sequence should be in the FASTA format. Example sequences in FASTA format can be get by clicking on the "example" button; Click on the "submit" button to see the predicted result. You will see the outcomes shown on the screen of your computer.

Or upload a file in FASTA: 选择文件 未选择...文件

Links

- LinDing group
- DNA predictors
- RNA predictors
- Protein predictors
- Databases

Center for Informational Biology, University of Electronic Science and Technology of China, Sichuan, Chengdu 611731,China.

FIGURE 6 | A semi-screenshot for the web server page of the iDNA6mA-Rice web server at http://lin-group.cn/server/iDNA6mA-Rice.

comparison with published methods. All results demonstrated that our proposed method could accurately recognize 6mA sites in the rice genome. For the convenience of most wet-experimental scholars, we established a free web server to predict 6mA sites. We anticipate that the web server can promote the efficient discovery of novel potential 6mA sites in the rice genome and facilitate the exploration of their functional mechanisms in gene regulation.

DATA AVAILABILITY

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

AUTHOR CONTRIBUTIONS

WC, YZ, and HLin conceived the study. HLv and F-YD implemented the study and drafted the manuscript. HLv, Z-XG, and DZ wrote the custom scripts and performed analysis. HLv, WC, and YZ interpreted the data. All authors read and approved the manuscript.

FUNDING

This work has been supported by the National Nature Scientific Foundation of China (grant nos. 61772119 and 31771471) and the Science Strength Promotion Programme of UESTC.

REFERENCES

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* 20, 274–281. doi: 10.1038/nsmb.2518

Bird, A. (1992). The essentials of DNA methylation. *Cell* 70, 5–8. doi: 10.1016/0092-8674(92)90526-I

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017a). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22. doi: 10.3390/molecules22101732

Cao, R. Z., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., and Cheng, J. L. (2017b). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics* 33. doi: 10.1093/bioinformatics/btw694

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*. doi: 10.1093/bioinformatics/btz015

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Chen, X. X., Tang, H., Li, W. C., Wu, H., Chen, W., Ding, H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2016, 1654623. doi: 10.1155/2016/1654623

Cheng, J. H., Yang, H., Liu, M. L., Su, W., Feng, P. M., Ding, H., et al. (2018a). Prediction of bacteriophage proteins located in the host cell using hybrid features. *Chemometr. Intell. Lab. Syst.* 180, 64–69. doi: 10.1016/j.chemolab.2018.07.006

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018b). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019a). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019b). 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

Cheng, X. (1995). Structure and function of DNA methyltransferases. *Annu. Rev. Biophys. Biomol. Struct.* 24, 293–318. doi: 10.1146/annurev.bb.24.060195.001453

Cui, T., Zhang, L., Huang, Y., Yi, Y., Tan, P., Zhao, Y., et al. (2018). MNDR v2.0: an updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* 46, D371–D374. doi: 10.1093/nar/gkx1025

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2018). Identify origin of replication in Saccharomyces cerevisiae using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Deng, M., Yu, C., Liang, Q., He, R. L., and Yau, S. S. (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* 6, e17293. doi: 10.1371/journal.pone.0017293

Fang, G., Munera, D., Friedman, D. I., Mandlik, A., Chao, M. C., Banerjee, O., et al. (2012). Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. *Nat. Biotechnol.* 30, 1232–1239. doi: 10.1038/nbt.2432

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2018). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K. C. (2019). iDNA6mA-PseKNC: Identifying DNA N(6)-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. *Genomics* 111, 96–102. doi: 10.1016/j.ygeno.2018.01.005

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Fu, Y., Luo, G. Z., Chen, K., Deng, X., Yu, M., Han, D., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* 161, 879–892. doi: 10.1016/j.cell.2015.04.010

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizabal-Corrales, D., et al. (2015). DNA Methylation on N6-adenine in C. elegans. *Cell* 161, 868–878. doi: 10.1016/j.cell.2015.04.005

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., Lin, H., Chen, W., et al. (2014). iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition. *Bioinformatics* 30, 1522–1529. doi: 10.1093/bioinformatics/btu083

He, W., Jia, C., and Zou, Q. (2018a). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 35, 593–601. doi: 10.1093/bioinformatics/bty668

He, W., Jia, C., Duan, Y., and Zou, Q. (2018b). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12, 44. doi: 10.1186/s12918-018-0570-1

Heyn, H., and Esteller, M. (2015). An adenine code for DNA: a second life for N6-Methyladenine. *Cell* 161, 710–713. doi: 10.1016/j.cell.2015.04.021

Hu, B., Zheng, L., Long, C., Song, M., Li, T., Yang, L., et al. (2019). EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. *Open Biol.* 9, 190054. doi: 10.1098/rsob.190054

Koziol, M. J., Bradshaw, C. R., Allen, G. E., Costa, A. S. H., Frezza, C., and Gurdon, J. B. (2016). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. *Nat. Struct. Mol. Biol.* 23, 24–30. doi: 10.1038/nsmb.3145

Lai, H. Y., Zhang, Z. Y., Su, Z. D., Su, W., Ding, H., Chen, W., et al. (2019). iProEP: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids* 17, 337–346. doi: 10.1016/j.omtn.2019.05.028

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liang, Z. Y., Lai, H. Y., Yang, H., Zhang, C. J., Yang, H., Wei, H. H., et al. (2017). Pro54DB: a database for experimentally verified sigma-54 promoters. *Bioinformatics* 33, 467–469. doi: 10.1093/bioinformatics/btw630

Lin, H., and Li, Q. Z. (2011). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory Biosci.* 130, 91–100. doi: 10.1007/s12064-010-0114-8

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

Liu, D., Li, G., and Zuo, Y. (2018). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes. *Brief. Bioinform.* doi: 10.1093/bib/bby053

Liu, J., Zhu, Y., Luo, G. Z., Wang, X., Yue, Y., Wang, X., et al. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. *Nat. Commun.* 7, 13052. doi: 10.1038/ncomms13052

Long, C. S., Li, W., Liang, P. F., Liu, S., and Zuo, Y. C. (2019). Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. *IEEE Access* 7, 7794–7802. doi: 10.1109/ACCESS.2018.2889809

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief. Bioinform.* doi: 10.1093/bib/bbz048

Manavalan, B., and Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics* 33, 2496–2503. doi: 10.1093/bioinformatics/btx222

Manavalan, B., Shin, T. H., and Lee, G. (2018a). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9, 476. doi: 10.3389/fmicb.2018.00476

Manavalan, B., Shin, T. H., and Lee, G. (2018b). DHSpred: support-vector-machine-based human DNase I hypersensitive sites prediction using the optimal features selected by random forest. *Oncotarget* 9, 1944–1956. doi: 10.18632/oncotarget.23099

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Invest. Radiol.* 24, 234–245. doi: 10.1097/00004424-198903000-00012

Mondo, S. J., Dannebaum, R. O., Kuo, R. C., Louie, K. B., Bewick, A. J., LaButti, K., et al. (2017). Widespread adenine N6-methylation of active genes in fungi. *Nat. Genet.* 49, 964–968. doi: 10.1038/ng.3859

O'Shea, J. P., Chou, M. F., Quader, S. A., Ryan, J. K., Church, G. M., and Schwartz, D. (2013). pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* 10, 1211–1212. doi: 10.1038/nmeth.2646

Ratel, D., Ravanat, J. L., Berger, F., and Wion, D. (2006). N6-methyladenine: the other methylated base of DNA. *Bioessays* 28, 309–315. doi: 10.1002/bies.20342

Smith, Z. D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* 14, 204–220. doi: 10.1038/nrg3354

Song, J., Zhai, J., Bian, E., Song, Y., Yu, J., and Ma, C. (2018a). Transcriptome-wide annotation of m5c RNA modifications using machine learning. *Front. Plant Sci.* 9, 519. doi: 10.3389/fpls.2018.00519

Song, J., Tan, H., Perry, A. J., Akutsu, T., Webb, G. I., Whisstock, J. C., et al. (2012). PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. *PLoS One* 7, e50300. doi: 10.1371/journal.pone.0050300

Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N. D., Webb, G. I., et al. (2018b). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. *Brief. Bioinform.* 20, 638–658. doi: 10.1093/bib/bby028

Song, J., Li, F., Leier, A., Marquez-Lago, T. T., Akutsu, T., Haffari, G., et al. (2018c). PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics* 34, 684–687. doi: 10.1093/bioinformatics/btx670

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2018). Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457

Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 34, 4196–4204. doi: 10.1093/bioinformatics/bty508

Tan, J. X., Li, S. H., Zhang, Z. M., Chen, C. X., Chen, W., Tang, H., et al. (2019). Identification of hormone binding proteins based on machine learning methods. *Math. Biosci. Eng.* 16, 2466–2480. doi: 10.3934/mbe.2019123

Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., et al. (2018a). HBPred: a tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* 14, 957–964. doi: 10.7150/ijbs.24174

Tang, W., Wan, S., Yang, Z., Teschendorff, A. E., and Zou, Q. (2018b). Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics* 34, 398–406. doi: 10.1093/bioinformatics/btx622

Tian, K., Zhao, X., and Yau, S. S. (2018). Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. *J. Theor. Biol.* 456, 34–40. doi: 10.1016/j.jtbi.2018.07.035

von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N. Q., Salehzadeh-Yazdi, A., Santos, F., et al. (2016). Impairment of DNA methylation maintenance is the main cause of global demethylation in naive embryonic stem cells. *Mol. Cell* 62, 848–861. doi: 10.1016/j.molcel.2016.04.025

Wang, M. J., Zhao, X. M., Tan, H., Akutsu, T., Whisstock, J. C., and Song, J. N. (2014). Cascleave 2.0, a new approach for predicting caspase and granzyme cleavage targets. *Bioinformatics* 30, 71–80. doi: 10.1093/bioinformatics/btt603

Wang, Y., Chen, X., Sheng, Y., Liu, Y., and Gao, S. (2017). N6-adenine DNA methylation is associated with the linker DNA of H2A.Z-containing well-positioned nucleosomes in Pol II-transcribed genes in Tetrahymena. *Nucleic Acids Res.* 45, 11594–11606. doi: 10.1093/nar/gkx883

Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018a). ACPred-FL: a sequence-based predictor based on effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016. doi: 10.1093/bioinformatics/bty451

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018b). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi: 10.1093/bioinformatics/bty824

Wion, D., and Casadesus, J. (2006). N6-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* 4, 183–192. doi: 10.1038/nrmicro1350

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., et al. (2016). DNA methylation on N(6)-adenine in mammalian embryonic stem cells. *Nature* 532, 329–333. doi: 10.1038/nature17640

Xiao, C. L., Zhu, S., He, M., Chen, Zhang, Q., Chen, Y., Yu, G., et al. (2018). N(6)-Methyladenine DNA modification in the human genome. *Mol. Cell* 71, 306–318 e7. doi: 10.1016/j.molcel.2018.06.015

Yang, H., Lv, H., Ding, H., Chen, W., and Lin, H. (2018a). iRNA-2OM: a sequence-based predictor for identifying 2'-O-methylation sites in homo sapiens. *J. Comput. Biol.* 25, 1266–1277. doi: 10.1089/cmb.2018.0004

Yang, H., Qiu, W. R., Liu, G. Q., Guo, F. B., Chen, W., Chou, K. C., et al. (2018b). iRSpot-Pse6NC: identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. *Int. J. Biol. Sci.* 14, 883–891. doi: 10.7150/ijbs.24616

Yao, B., Cheng, Y., Wang, Z., Li, Y., Chen, L., Huang, L., et al. (2017). DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. *Nat. Commun.* 8, 1122. doi: 10.1038/s41467-017-01195-y

Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., et al. (2017). RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Res.* 45, D115–D118. doi: 10.1093/nar/gkw1052

Zhang, C. J., Tang, H., Li, W. C., Lin, H., Chen, W., and Chou, K. C. (2016). iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 7, 69783–69793. doi: 10.18632/oncotarget.11975

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., et al. (2015). N6-methyladenine DNA modification in Drosophila. *Cell* 161, 893–906. doi: 10.1016/j.cell.2015.04.018

Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., et al. (2017). RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.* 45, D135–D138. doi: 10.1093/nar/gkw728

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2014). Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteomics* 11, 289–299. doi: 10.2174/157016461104150121115154

Zhou, C., Wang, C., Liu, H., Zhou, Q., Liu, Q., Guo, Y., et al. (2018). Identification and analysis of adenine N(6)-methylation sites in the rice genome. *Nat. Plants* 4, 554–563. doi: 10.1038/s41477-018-0214-x

Zhu, X. J., Feng, C. Q., Lai, H. Y., Chen, W., and Lin, H. (2019). Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl.-Based Syst.* 163, 787–793. doi: 10.1016/j.knosys.2018.10.007

Zou, Q., Xing, P., Wei, L., and Liu, B. (2018a). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). 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., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2018b). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform*. doi: 10.1093/bib/bby090

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

Zuo, Y. C., Peng, Y., Liu, L., Chen, W., Yang, L., and Fan, G. L. (2014). Predicting peroxidase subcellular location by hybridizing different descriptors of Chou' pseudo amino acid patterns. *Anal. Biochem*. 458, 14–19. doi: 10.1016/j.ab.2014.04.032

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

ORIGINAL RESEARCH
published: 12 September 2019
doi: 10.3389/fgene.2019.00832

Predicting circRNA-Disease Associations Based on circRNA Expression Similarity and Functional Similarity

Yongtian Wang, Chenxi Nie, Tianyi Zang and Yadong Wang**

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Circular RNAs (circRNAs) are a novel class of endogenous noncoding RNAs that have well-conserved sequences. Emerging evidence has shown that circRNAs can be novel biomarkers or therapeutic targets for many diseases and play an important role in the development of various pathological conditions. Therefore, identifying potential disease-related circRNAs is helpful in improving the efficiency of finding therapeutic targets for diseases. Here, we propose a computational model (PreCDA) to predict potential circRNA–disease associations. First, we calculated the circRNA expression similarity based on circRNA expression profiles. The circRNA functional similarity is calculated based on cosine similarity, and the disease similarity is used as the dimension of each circRNA vector. The associations between circRNAs and diseases are defined based on the circRNA functional similarity and expression similarity. We constructed a disease-related circRNA association network and used a graph-based recommendation algorithm (PersonalRank) to sort candidate disease-related circRNAs. As a result, PreCDA has an average area under the receiver operating characteristic curve value of 78.15% in predicting candidate disease-related circRNAs. In addition, we discuss the factors that affect the performance of this method and find some unknown circRNAs related to diseases, with several common diseases used as case studies. These results show that PreCDA has good performance in predicting potential circRNA–disease associations and is helpful for the diagnosis and treatment of human diseases.

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Leyi Wei,
The University of Tokyo, Japan
Hui Ding,
University of Electronic Science and Technology of China, China

****Correspondence:***
Tianyi Zang
tianyi.zang@hit.edu.cn
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 June 2019*
Accepted: *13 August 2019*
Published: *12 September 2019*

Citation:
Wang Y, Nie C, Zang T and Wang Y (2019) Predicting circRNA-Disease Associations Based on circRNA Expression Similarity and Functional Similarity. Front. Genet. 10:832. doi: 10.3389/fgene.2019.00832

Keywords: circRNA, disease, circRNA expression similarity, circRNA functional similarity, PersonalRank

INTRODUCTION

Circular RNAs (circRNAs) are a type of RNA molecule that forms a covalently closed continuous loop from exon circularization (Motieghader et al., 2017; Xu, 2017). In recent years, advances in high-throughput sequencing technology have greatly facilitated the study of circRNAs (Jeck and Sharpless, 2014). When compared to other ncRNAs (Danan et al., 2012), circRNAs are highly stable. Circular RNAs have evolutionarily conserved sequence features across species, tissues, and developmental stages (Jens, 2013; Conn et al., 2015; Rybak-Wolf et al., 2015). Therefore, circRNAs have become hotspots in transcriptomics research.

Recent studies have shown that alterations in the expression of circRNAs play important roles in human disease and other biological processes (Xu, 2017; Zhao and Shen, 2017; Xia et al., 2018). For example, the best-known circRNA, CDR1as, as the inhibitor of miR-7, is a critical ncRNA known to be involved in cancer, neurodegenerative diseases, diabetes, and atherosclerosis (Li et al., 2015; Xu et al., 2018).

Researchers found that the circRNA ciRS-7 may be a promising target for neurodegenerative disorder (Lukiw, 2013) and myocardial infarction (Lin et al., 2018). The circRNA CircCCDC66 has been demonstrated to regulate colon cancer growth and metastasis as a miRNA sponge (Hsiao et al., 2017). The circRNA hsa_circ_0001895 is involved in the expression of cancer-related proteins in gastric cancer (Shao et al., 2017). The circRNA CircHIPK3 plays an important role in cell growth by sponging multiple miRNAs (Zheng et al., 2016). Moreover, circRNAs can be found in exosomes, cell-free saliva, and plasma (Li Y et al., 2015). Circular RNAs are emerging as novel biomarkers or therapeutic targets for many diseases due to their conservation, cell type–specific expression, and tissue-specific expression, and they play roles in the development of various pathological conditions (Meng et al., 2017; Vo et al., 2018).

Although a large number of circRNAs have been discovered, the mechanisms of circRNAs in many diseases remain unclear (Xu et al., 2018). To enable research on circRNAs and diseases, several databases have been constructed, such as circRNADisease (Zhao et al., 2018), CircR2Disease (Fan et al., 2018), and Circ2Disease (Yao et al., 2018). They provide important data support for circRNA–disease association analyses. Some methods have been proposed to provide the most promising disease-related biomarkers, including those involving lncRNAs (Chen et al., 2015; Gu et al., 2017; Cheng et al., 2018a; Cheng et al., 2019), miRNAs (Peng et al., 2019b; Shao et al., 2018), genes (Cheng et al., 2016; Hu et al., 2019; Peng et al., 2019a), and drugs (Jiang et al., 2017; Zhang et al., 2017), for further experimental validation. These methods can decrease the time and cost of biological experiments. However, very few methods have been developed to predict potential circRNA–disease associations (Lei et al., 2018), and both disease functional similarity and semantic similarity were not considered in these methods. Improved knowledge has suggested that exploring both the semantic and functional associations of diseases, which are two types of significant associations, is beneficial in measuring disease similarity (Cheng et al., 2014; Peng et al., 2018).

In this study, we proposed a computational model (PreCDA) for potential disease-related circRNA identification. In view of the limited number of circRNA–disease associations, we introduced disease similarity to solve possible sparse problems and built a disease-related circRNA similarity network. However, relying entirely on circRNA-related diseases greatly limits the utility of the method because many circRNAs still have very few or no associated diseases. To overcome this limitation, we calculated the circRNA expression similarity based on the existing data resources. Subsequently, we built a new disease-associated circRNA network by fusing circRNA functional associations and expression similarities. To assess the practicability and accuracy of this method, we designed a validation process with different datasets of circRNA–disease associations, as good computational models must perform well on different data sources. Finally, PreCDA proved successful in predicting potential disease-related circRNAs.

MATERIALS AND METHODS

Workflow

A flowchart of the PreCDA workflow is shown in **Figure 1**. We preprocessed circRNA and disease data because of the lack of uniform identification of circRNAs and diseases. We extracted the synonym vocabulary from the two circRNA databases, including circRNADisease (Zhao et al., 2018) and circBase (Glažar et al., 2014). Then, we unified different representations of the same circRNA in different databases. Additionally, the identification of the Human Disease Ontology (DO) (Kibbe et al., 2015) was used as the unified marker of diseases in the computational model. We measured the similarity between circRNAs in two ways, including the circRNA expression similarity and functional similarity. We extracted circRNA expression profiles from circBase (Glažar et al., 2014) and CIRCpedia (Dong et al., 2018). The circRNA expression similarity was calculated based on the Spearman correlation coefficient. The disease similarity was used as the dimension of each circRNA vector, and the circRNA functional similarity was calculated based on cosine similarity. A disease-related circRNA association network was built based on the circRNA expression similarity and functional similarity. Finally, we identified potential candidate disease-related circRNAs based on the PersonalRank algorithm (PR) (Haveliwala, 2002).

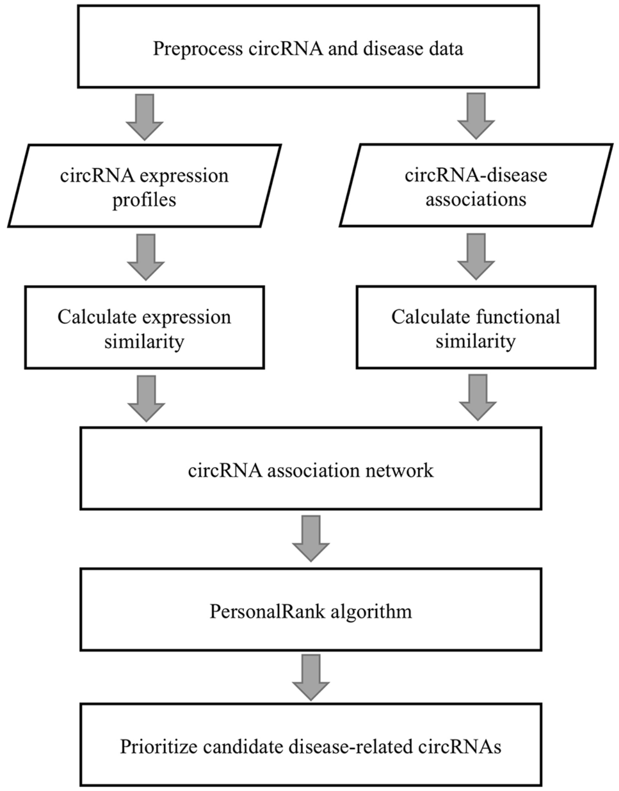

Preprocess circRNA and disease data
circRNA expression profiles
circRNA-disease associations
Calculate expression similarity
Calculate functional similarity
circRNA association network
PersonalRank algorithm
Prioritize candidate disease-related circRNAs

FIGURE 1 | Flowchart of the PreCDA workflow.

Data Preprocessing

circRNA Data

In this study, we used three circRNA databases for experiments and validations. The circRNADisease database is a manually curated database of experimentally supported circRNA and disease associations, which collected 330 circRNAs and 48 diseases in 354 associations. Each entry in the circRNADisease database includes detailed information on a circRNA–disease association, including the circRNA and disease name, the circRNA expression pattern, literature references, and other annotation information. CircR2Disease is a database for experimentally supported circRNA–disease associations and provides a platform for investigating the mechanism of disease-related circRNAs. The present version of CircR2Disease collected 661 circRNAs and 100 diseases. Circ2Disease is a database that curates experimentally supported human circRNAs and provides comprehensive associations between circRNAs and human diseases. It contains 273 manually curated associations between 237 circRNAs and 54 human diseases from 120 studies. However, currently, the naming of circRNAs has not yet been unified (Xu et al., 2018), which leads to the underutilization of information from different public circRNA databases. Therefore, we designed and collected mappings among different circRNA names provided by different circRNA databases, including circRNADisease and circBase. circRNADisease contains circRNA synonyms, and circBase is a database that merged and unified datasets of circRNAs. We mapped circRNAs from the three circRNA databases to circBase referring to circRNA synonyms. Then, we used circRNA IDs from circBase as the unified IDs of circRNAs in this work.

Disease Data

Human Disease Ontology represents common and rare human disease concepts captured across biomedical resources. Each node in DO represents one disease term and is organized in a directed acyclic graph with the relationship of “is_a”. MEDIC (Davis et al., 2012) integrates OMIM (Online Mendelian Inheritance in Man) terms (Amberger et al., 2015), synonyms and identifiers with MeSH terms (Lipscomb, 2000), synonyms, definitions, identifiers, and hierarchical relationships.

We extracted disease terms and synonyms from MEDIC to annotate DO by the same external references in DO and MEDIC, as shown in **Figure 2**. If a disease term was recorded in both DO and MEDIC, the term and its synonyms in MEDIC were used to annotate DO. With this approach, a given disease name can be matched to DO to a great extent by string matching, considering that the naming rules for diseases in different disease-related circRNA databases are different. The diseases described by different names are considered to be the same disease that has a unique id in DO if these disease names can match the disease term or its extended synonyms in DO.

circRNA Expression Similarity

Considering that comprehensive circRNA expression data are still unavailable, we extracted circRNA expression profiles from circBase and CIRCpedia, including the expression profiles of 92488 circRNAs in 78 human cell types or tissues. We used the Spearman correlation coefficient between the expression profiles of each circRNA as the circRNA expression similarity, as shown in Formula 1.

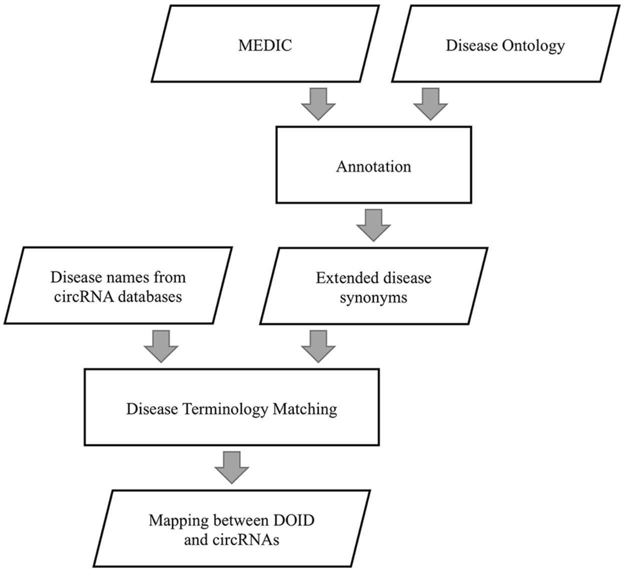

MEDIC
Disease Ontology
Annotation
Disease names from circRNA databases
Extended disease synonyms
Disease Terminology Matching
Mapping between DOID and circRNAs

FIGURE 2 | Flowchart of establishing mappings between circRNAs and disease ontology terms.

$$\rho = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

where d_i is the difference between the two ranks of the expression scores in the *i*th human cell type or tissue, and *n* is the number of the human cell types or tissues from circBase or CIRCpedia. Matrix *CB* and Matrix *CP* are, respectively, denoted as the circRNA expression similarity matrix of circBase and CIRCpedia, where *CB(i,j)* and *CP(i,j)* are the expression similarities between circRNA *c(i)* and *c(j)*. Then, to obtain reliable performance for circRNA expression data, we defined the expression similarity between circRNA *c(i)* and *c(j)* as shown in Formula 2 if circRNA *c(i)* and *c(j)* are included in both circBase and CIRCpedia.

$$ExSim\left(i,j\right) = \begin{cases} Max\left(CB\left(i,j\right), CP\left(i,j\right)\right) & \mathrm{Max}\left(CB\left(i,j\right), CP\left(i,j\right)\right) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

To reduce the impact of data noise, we set a threshold τ to filter out those weak similarities between circRNAs. The threshold τ is set to 0.7 based on our experiments.

circRNA Functional Similarity

We extracted circRNA–disease associations from these above circRNA databases and defined a relational matrix of circRNAs and diseases. For each circRNA, all diseases in the matrix can be used to make a vector in a multidimensional space. Because of the limited number of available disease–circRNA pairs, there is a data sparsity problem in the matrix. Therefore, we calculated the circRNA-related disease similarity and filled this matrix with predicted association scores based on disease–circRNA associations and the disease similarity. Here, we use FNSemSim (Wang et al., 2017) to calculate disease similarity. This method, which combines disease functional similarity and semantic similarity, has good performance for calculating similarities between diseases. The workflow of calculating circRNA functional similarity is shown in **Figure 3**.

To calculate the association between one circRNA and any disease, the similarities between this disease and all diseases that are directly related to this circRNA are calculated by FNSemSim. C is defined as the set of disease-related circRNAs, and D represents the set of circRNA-related diseases. DisSet(c) is defined as the set of diseases directly related to circRNA c. The association score between disease *dis* and circRNA c is defined as follows:

$$Score(\text{dis},c)=\begin{cases} Max\left(FNSemSim\left(\text{dis},\text{dis}_i\right)\right) & \text{dis}_i\in \text{DisSet}(c),\quad \text{dis}\notin \text{DisSet}(c) \\ 1 & \text{dis}\in \text{DisSet}(c) \end{cases} \tag{3}$$

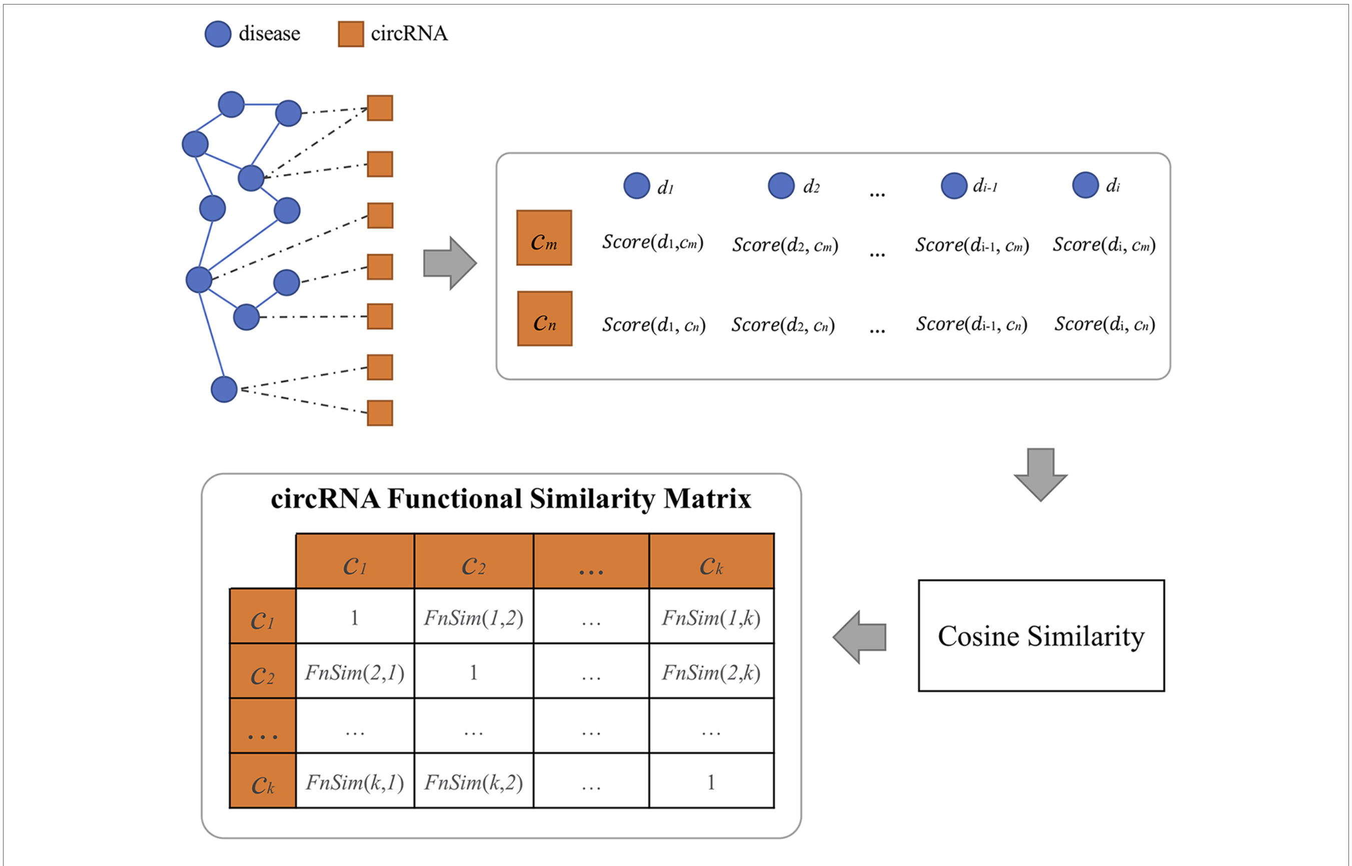
where DisSet(c) ⊆D, $1\le i\le |\text{DisSet}(c)|$; |DisSet(c)| is denoted as the number of diseases in DisSet(c). If this disease belongs to DisSet(c), the score is 1; otherwise, the score is defined as the maximum of similarities between this disease and all the diseases related to circRNA c. Therefore, circRNA c can be depicted by a vector that is composed of circRNA-related diseases in a multidimensional space. We can calculate the functional similarity between any two circRNAs based on cosine similarity. The functional similarity between circRNA *c(m)* and *c(n)* is defined as follows:

$$FnSim(m,n)=\frac{\sum_{i=1}^{|D|}\text{Score}\left(\text{dis}_i,c(m)\right)\times \text{Score}\left(\text{dis}_i,c(n)\right)}{\sqrt{\sum_{i=1}^{|D|}\text{Score}\left(\text{dis}_i,c(m)\right)^2}\sqrt{\sum_{i=1}^{|D|}\text{Score}\left(\text{dis}_i,c(n)\right)^2}} \tag{4}$$

where |D| represents the size of the circRNA-related disease set D, and dis_i is the ith disease in the circRNA-related disease set D.

Prediction of Candidate Disease-Related circRNAs

We take circRNA functional similarity and expression similarity as weights to construct a circRNA association network. In this network, the weight between circRNA *c(i)* and *c(j)* is defined as shown in Formula 5. If *ExSim(i,j)* is greater than 0, the weight between circRNA *c(i)* and *c(j)* is the average value of their



	d_1	d_2	...	d_{i-1}	d_i
C_m	Score(d_1,c_m)	Score(d_2, c_m)	...	Score(d_{i-1}, c_m)	Score(d_i, c_m)
C_n	Score(d_1, c_n)	Score(d_2, c_n)	...	Score(d_{i-1}, c_n)	Score(d_i, c_n)

circRNA Functional Similarity Matrix

	C_1	C_2	…	C_k
C_1	1	*FnSim(1,2)*	…	*FnSim(1,k)*
C_2	*FnSim(2,1)*	1	…	*FnSim(2,k)*
…	…	…	…	…
C_k	*FnSim(k,1)*	*FnSim(k,2)*	…	1

FIGURE 3 | Flowchart of calculating circRNA functional similarity.

functional similarity and expression similarity; otherwise, the weight is defined as the functional similarity between them.

$$CircWeight(i,j)=\begin{cases}\left(\mathrm{FnSim}(i,j)+\mathrm{ExSim}(i,j)\right)/2 & \text{if } \mathrm{ExSim}(i,j)>0\\ FnSim(i,j) & \text{otherwise}\end{cases} \quad (5)$$

To predict candidate disease-related circRNAs, the associations between diseases and circRNAs are also considered in this network. The weight between circRNA *c* and disease *dis* is defined as shown in Formula 6. If the disease is directly related to circRNA *c*, the weight between them is 1; otherwise, the weight is 0.

$$CircDisWeight(i,j)=\begin{cases}1 & \text{if dis} \in \mathrm{DisSet}(c)\\ 0 & \text{otherwise}\end{cases} \quad (6)$$

In this network composed of circRNAs and diseases, we identify novel candidate disease-related circRNAs based on the PR. PersonalRank algorithm, as a recommendation algorithm based on random walking, can reveal more information between a target node and all the others in a specific network. PersonalRank algorithm is defined as follows:

$$PR(i)=(1-d)r_i+d\sum_{j\in \mathrm{in}(i)}\frac{\mathrm{PR}(j)}{|\mathrm{out}(j)|} \quad (7)$$

where PR(i) represents the possibility value that node i is accessed; d is the transfer probability; out(j) represents the out-degree of node j; in(i) is the in-degree of node i; and r_i is defined as follows:

$$r_i=\begin{cases}1 & \text{if } i=t\\ 0 & \text{if } i\neq t\end{cases} \quad (8)$$

where t represents the target node. According to previous studies (Kang et al., 2014; Cheng et al., 2018b), d is set to 0.85. The target node t in the network randomly moves to adjacent nodes with the probabilities of the edges between these nodes. After enough iterations, the probabilities from the target node to all the other nodes will become stable. Eventually, the algorithm outputs the relevance degrees between all the nodes and this target node.

RESULTS

circRNAs and Diseases

We calculated similarities between 323 circRNAs from circBase and CIRCpedia based on circRNA expression profiles. Then, we obtained 11,281 circRNA pairs based on the preset threshold. Additionally, we found 507 relationships between 58 diseases and 445 circRNAs by mapping DO terms to the diseases in CircR2Disease. We matched 26 diseases based on DO terms and extracted 293 relationships between 277 circRNAs and these diseases from circRNADisease. In Circ2Disease, 218 relationships between 37 diseases and 199 circRNAs were found. Based on DO terms and the unification of circRNA naming, we analyzed the three circRNA databases and found the same circRNAs and diseases among these databases, as shown in **Figure 4**. This provided the test data for the performance evaluation of PreCDA.

We separately calculated the similarities between 445 circRNAs from CircR2Disease, 277 circRNAs from circRNADisease and 199 circRNAs from Circ2Disease. Three circRNA association networks were built that in turn contained 96,580 associations

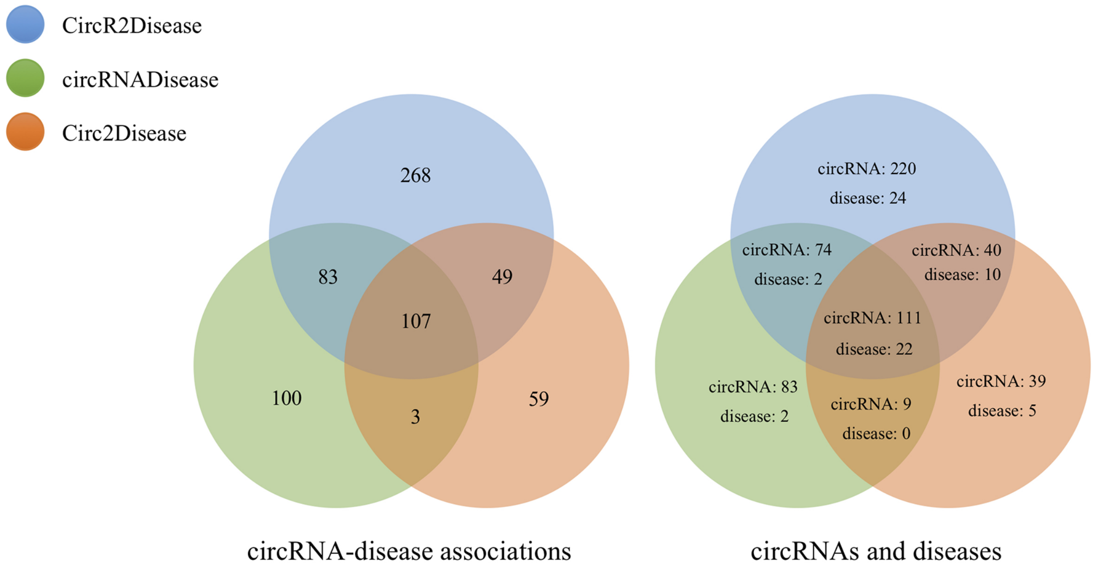


FIGURE 4 | Data distribution in the three databases.

TABLE 1 | Information on the three circRNA association networks.

Database	circRNA association network		
	circRNA	Disease	Association
CircR2Disease	440	56	96,580
circRNADisease	277	26	38,226
Circ2Disease	195	36	18,915

between 440 circRNAs associated with 56 diseases; 38,226 associations between 277 circRNAs associated with 26 diseases; and 18,915 associations between 195 circRNAs associated with 36 diseases. The detailed statistics of the circRNAs and diseases are shown in **Table 1**.

Performance

We designed a test scheme to assess the performance of PreCDA. First, we selected two circRNA–disease databases, one to build the circRNA association network and the other to provide test data. Then, we extracted the same diseases from the circRNA association network and the reference database. For a given disease, if any circRNA related to this disease in the reference database exists in the network, but the association between the circRNA and the disease does not, the circRNA can be used as a test case for the disease to assess the performance of this circRNA association network. The test scheme is shown in **Figure 5**.

In this article, we used three circRNA–disease databases, including CircR2Disease, circRNADisease, and Circ2Disease. For example, both circRNA hsa_circ_0000284 and liver cancer (DOID: 3571) were recorded in Circ2Disease and CircR2Disease. The circRNA hsa_circ_0000284 was related to liver cancer (DOID: 3571) in Circ2Disease but not in CircR2Disease. Therefore, we built a circRNA association network based on CircR2Disease and calculated the relevance degrees between liver cancer and all circRNAs unrelated to the disease. We calculated the area under the receiver operating characteristic curve (AUC) according to the ranking of the circRNA hsa_circ_0000284 among these circRNAs to measure the prediction results. To validate the reliability of the computational model, we conducted nine validation experiments based on this scheme involving 18 diseases. We built three circRNA association networks based on the three different circRNA–disease databases. The three data sources were also used as the reference data. Additionally, we merged the known circRNA–disease associations in the three databases as an additional control data source.

PreCDA had an average AUC value of 78.15% in predicting candidate disease-related circRNAs. Furthermore, it had an outstanding performance on some diseases. For example, diabetes mellitus (DOID: 9351) in the network from Circ2Disease had an AUC of 98.48% based on the control data from circRNADisease and an AUC of 93.04% based on the control data from CircR2Disease. Based on the control data from Circ2Disease, the AUC of osteoarthritis (DOID: 8398) was 97.44% in the network from CircR2Disease and 98% in the network from circRNADisease. In the network from Circ2Disease, the AUC of stomach cancer (DOID: 10534) was 56.41% based on the control data from circRNADisease; it had an AUC of 73.88% in CircR2Disease. This shows that the networks from the different data sources have different results for a disease based on the same control database. However, the AUCs in the other validation experiments achieved more than 65%. Even so, the performance of PreCDA is excellent in predicting candidate disease-related circRNAs. The performance of PreCDA based on the different databases and the different control data sources is shown in **Figure 6**.

Case Study

To further evaluate the performance of PreCDA in predicting potential disease-related circRNAs, we conducted some case studies, including prostate cancer (DOID: 10283), liver cancer (DOID: 3571), breast carcinoma (DOID: 3459), Alzheimer disease, and pancreatic cancer (DOID: 1793). We integrated the known associations between circRNAs and diseases in the three databases and prioritized candidate disease-related circRNAs based on PreCDA.

In the ranking of candidate circRNAs related to liver cancer (DOID: 3571), hsa_circ_0001727 (Qiu et al., 2018) ranked 4th, hsa_circ_0001946 (Yu et al., 2016) ranked 7th, and hsa_circ_0001141 (Guo et al., 2017) ranked 19th. They ranked in the

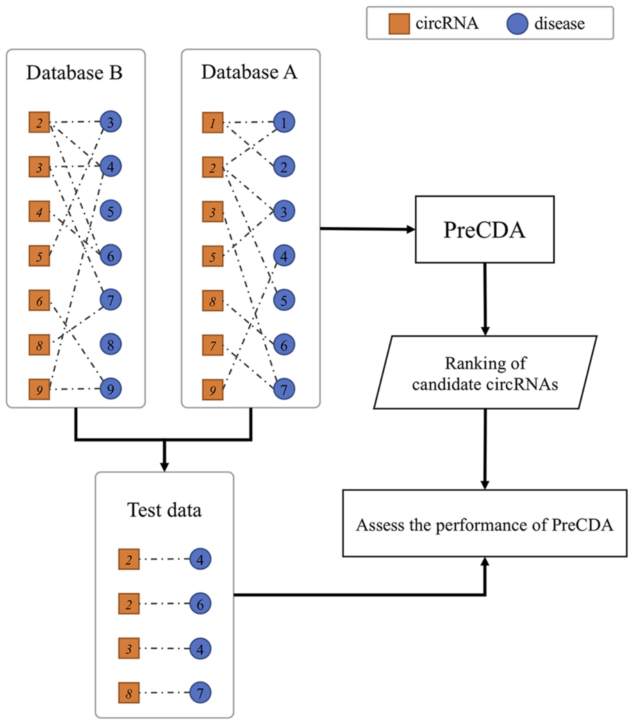


FIGURE 5 | The validation scheme of the computational model. For comparison with database B, test data are extracted from database A according to the test scheme. PreCDA outputs the ranks of candidate circRNAs with the circRNA–disease associations from database A as the input. The performance of PreCDA is assessed based on the test data.

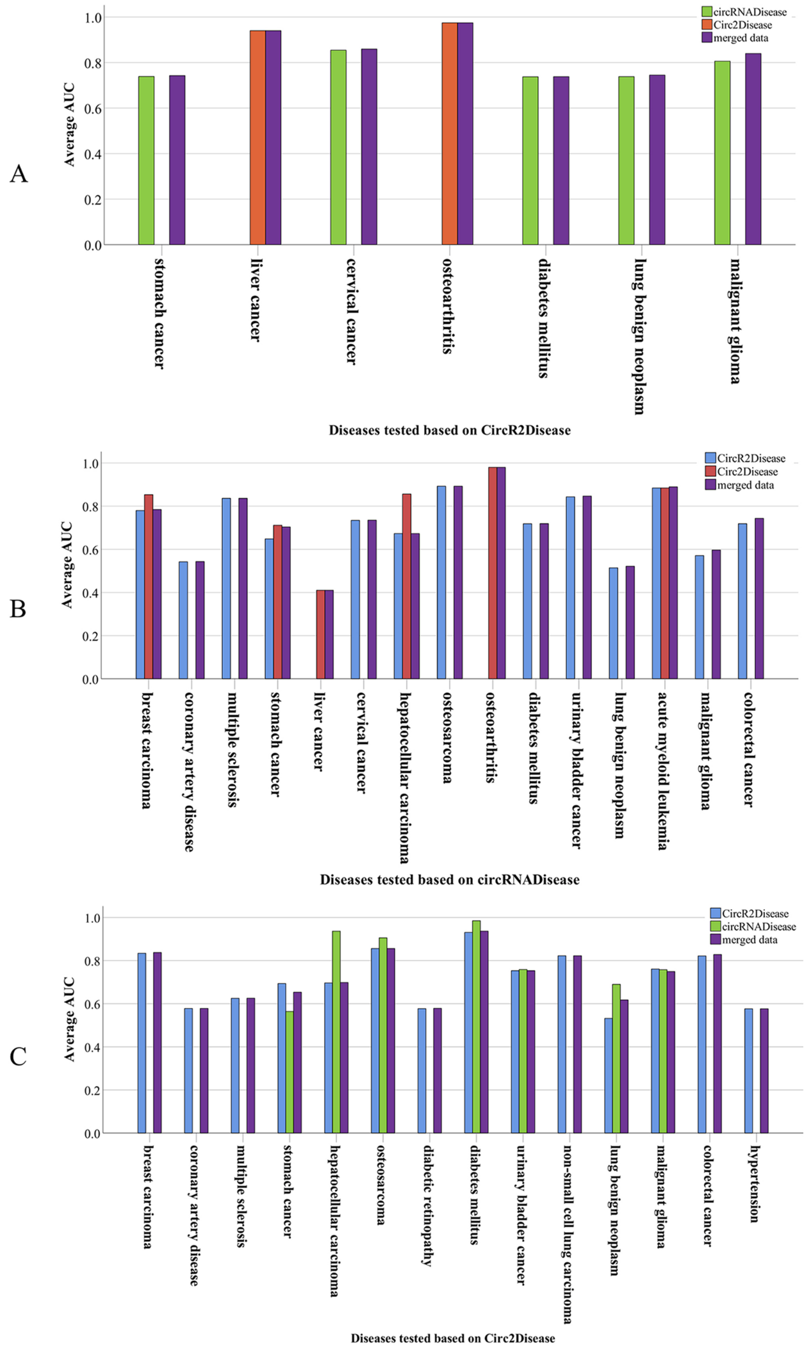


FIGURE 6 | The performance in predicting circRNA-associated diseases. **(A)** Seven diseases were tested based on CircR2Disease with reference to circRNADisease, Circ2Disease, and all circRNA–disease associations from the three data sources. **(B)** Fifteen diseases were tested based on circRNADisease with reference to CircR2Disease, Circ2Disease, and all circRNA–disease associations from the three data sources. **(C)** Fourteen diseases were tested based on Circ2Disease with reference to CircR2Disease, circRNADisease, and all circRNA–disease associations from the three data sources.

top 3% and were associated with liver cancer. For prostate cancer (DOID: 10283), hsa_circ_0001946 (Zhang et al., 2018) and hsa_circ_0001649 (Yi et al., 2016) ranked 3rd and 5th in the ranking, respectively. They were documented to be related to prostate cancer. For pancreatic cancer (DOID: 1793), CircRNA_100782 (Chen et al., 2017), which ranked 1st in the ranking, was validated to regulate pancreatic carcinoma proliferation through the IL6-STAT3 pathway. We found that some candidate circRNAs related to these diseases were included by Circ2Traits (Ghosal et al., 2013), which is a comprehensive database for circRNAs potentially associated with disease and traits. For example, hsa_circ_0000118, which ranked 1st in the ranking of candidate

TABLE 2 | The prediction results of predicting candidate circRNAs for five diseases.

Disease name	DOID	circRNA	Rank	Evidence
Prostate cancer	10283	hsa_circ_0000118	1	Circ2Traits
		hsa_circ_0001946	3	Zhang et al., 2018
		hsa_circ_0001649	5	Yi et al., 2016
		hsa_circ_0001070	7	Circ2Traits
		hsa_circ_0001512	16	Circ2Traits
		hsa_circ_0000437	18	Circ2Traits
		hsa_circ_0001727	45	Circ2Traits
		hsa_circ_0000130	52	Circ2Traits
Breast carcinoma	3459	hsa_circ_0001070	7	Circ2Traits
		hsa_circ_0001727	19	Circ2Traits
		hsa_circ_0001333	35	Circ2Traits
		hsa_circ_0000190	54	Circ2Traits
Liver cancer	3571	hsa_circ_0001727	4	Qiu et al., 2018
		hsa_circ_0001946	7	Yu et al., 2016
		hsa_circ_0001141	19	Guo et al., 2017
Pancreatic cancer	1793	hsa_circ_0000284	1	Chen et al., 2017
		hsa_circ_0002702	5	Circ2Traits
		hsa_circ_0001667	29	Circ2Traits
Alzheimer disease	10652	hsa_circ_0000284	8	Circ2Traits
		hsa_circ_0001141	28	Circ2Traits
		hsa_circ_0000096	32	Circ2Traits

circRNAs associated with prostate cancer, was documented to be potentially related to this disease in Circ2Traits. The prediction results of the case studies are presented in **Table 2**.

DISCUSSION

Although functional associations between circRNAs are measured based on circRNA expression profiles, there are many weak connections among them. To reduce the impact of data noise, we set a threshold to filter out those weak connections between circRNAs. Based on the above validation strategy and different thresholds, we conducted nine groups of experiments in which these three databases were used as a reference to each other and to test the performance of PreCDA. As shown in **Figure 7**, the average AUC of PreCDA varied with the change in the threshold, and the computational model worked best when the threshold was set to 0.7.

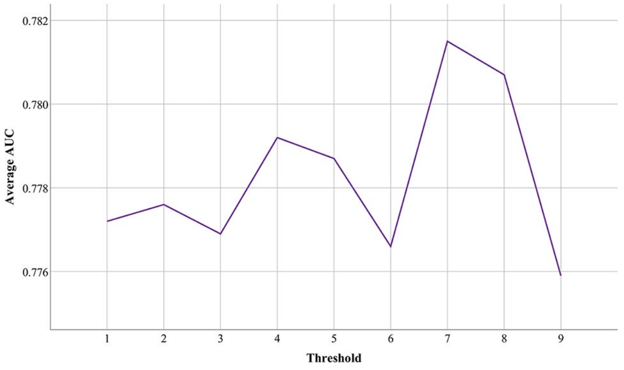


FIGURE 7 | The impact of different thresholds on the performance of PreCDA.

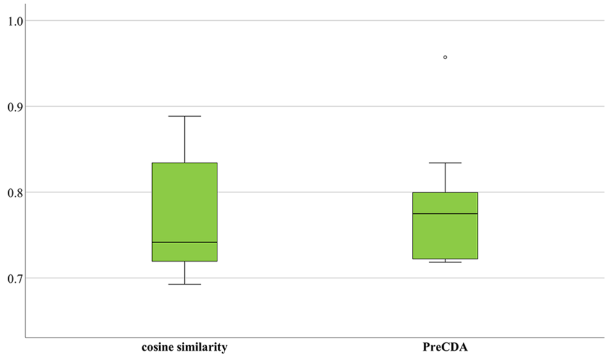


FIGURE 8 | The performance of different computational models.

We calculated circRNA similarities by only cosine similarity and built a circRNA association network. Additionally, we merged the known circRNA–disease associations in these three databases

TABLE 3 | Performance differences of predicting circRNA–disease pairs based on different data sources.

References database	Disease	DOID	AUC		circRNA
CircR2Disease			circRNADisease	Circ2Disease	
	Colorectal cancer	9256	71.86%	82.17%	hsa_circ_0001649
					hsa_circ_0000284
					hsa_circ_0014717
					hsa_circ_0001141
	Malignant glioma	3070	57.1%	76.1%	hsa_circ_0000284
					hsa_circ_0001649
					hsa_circ_0001445
	Lung benign neoplasm	3683	51.4%	53.18%	hsa_circ_0001821
					circUBAP2
	Diabetes mellitus	9351	71.85%	93.04%	hsa_circ_0000284
	Coronary artery disease	3393	54.21%	57.78%	hsa_circ_0000615
			CircR2Disease	Circ2Disease	
circRNADisease	Diabetes mellitus	9351	73.73%	98.48%	hsa_circ_0054633
	Malignant glioma	3070	80.6%	75.77%	hsa_circ_0001946
					hsa_circ_0004214
			CircR2Disease	circRNADisease	
Circ2Disease	Osteoarthritis	8398	97.44%	98%	hsa_circ_0000026

as an additional control data source. Based on the validation strategy mentioned above, we used these three databases to test the performance of the network. As shown in **Figure 8**, the average AUC was 77.22%, the minimum AUC was 69.26%, and the maximum AUC was 88.85%. In comparison, PreCDA has a more stable performance, with an average AUC of 78.15%. Its minimum and maximum AUCs are 71.83% and 95.72%, respectively.

We found that the performance of predicting potential disease–circRNA pairs in the disease-related circRNA association network was impacted by different data sources. The result of predicting the associations between the same diseases and circRNAs was different based on the different data sources that were used to build networks. For example, referring to CircR2Disease, some of the data to be tested in the networks built based on circRNADisease and Circ2Disease were the same. However, the AUC values of predicting the associations between them were different. As shown in **Table 3**, we predicted the associations between colorectal cancer (DOID: 9256) and four circRNAs, including hsa_circ_0001649, hsa_circ_0000284, hsa_circ_0014717, and hsa_circ_0001141. The AUC value for the network of circRNADisease was 71.86%. The performance of identifying the associations between colorectal cancer and these four circRNAs based on Circ2Disease was improved, and its AUC achieved 82.17%.

CONCLUSIONS

Circular RNA plays an important role in the development of various pathological conditions. Research on circRNA is invaluable in explaining the underlying pathogenesis. Therefore, we proposed a computational model to identify candidate disease-related circRNAs. First, we calculated the circRNA expression similarity with the circRNA expression profiles. Then, the disease similarity was used as dimensions of circRNA vectors, and the circRNA functional similarity was calculated based on cosine similarity. We defined the associations between circRNAs and diseases based on the circRNA expression similarity and functional similarity. A disease-related circRNA association network was built, and potential candidate disease-related circRNAs were ranked by the PR.

We evaluated the performance of PreCDA with the help of data differences among these three databases, including CircR2 Disease, circRNADisease, and Circ2Disease. The results showed that the average AUC of PreCDA was 78.15%, and it had good performance in predicting potential disease-related circRNA signatures. We discussed the selection of the threshold and the impact of different data sources on the performance of PreCDA. Then, we used several common diseases as case studies and found some unknown circRNAs that could be related to these diseases based on PreCDA. The findings of this study could be further applied in analyzing diseases in a system biology perspective (Cheng and Hu, 2018) and helpful for researchers to improve disease diagnostics and treatments.

DATA AVAILABILITY

PreCDA is implemented using a combination of Java and scala, and it is freely available from the website at https://github.com/wythit/PreCDA.

AUTHOR CONTRIBUTIONS

YoW and CN did data collection and preprocessing. And with the guidance of TZ and YaW, YoW finished the algorithm design and validation. YoW was the major contributor in writing the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

Publication costs were funded by the National Key Research and Development Program of China (grant no. 2016YFC0901605, 2016YFC1201702-01) and the National High-tech R&D Program of China (grant no. 2012AA02A604, 2015AA020108).

ACKNOWLEDGMENTS

TZ and YaW are he corresponding authors. We thank them for their guidance. We also thank Ling Wang and Zhenxing Wang for their valuable suggestions on our work. We are grateful to Rongjie Wang and Yanshuo Chu for their helpful assistance in writing.

REFERENCES

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM (R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43(D1), D789-D798. doi: 10.1093/nar/gku1205

Chen, G. W., Shi, Y. T., Zhang, Y., and Sun, J. Y. (2017). CircRNA_100782 regulates pancreatic carcinoma proliferation through the IL6-STAT3 pathway. *Onco. targets Ther.* 10, 5783–5794. doi: 10.2147/ott.s150678

Chen, X., Yan, C. G. C., Luo, C., Ji, W., Zhang, Y. D., and Dai, Q. H. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 12. doi: 10.1038/srep11338

Cheng, L., and Hu, Y. (2018). Human disease system biology. *Curr. Gene Ther.* 18, 255–256. doi: 10.2174/1566523218666181010101114

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. H. (2018a). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34(11), 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J. J., Zhou, M., et al. (2018b). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *Bmc Genomics* 19, 10. doi: 10.1186/s12864-017-4338-6

Cheng, L., Li, J., Ju, P., Peng, J. J., and Wang, Y. D. (2014). SemFunSim: A New Method for Measuring Disease Similarity by Integrating Semantic and Gene Functional Association. *PLoS One* 9(6), 11. doi: 10.1371/journal.pone.0099415

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Wang, P. P., Tian, R., Wang, S., Guo, Q. H., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47(D1), D140-D144. doi: 10.1093/nar/gky1051

Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., et al. (2015). The RNA Binding Protein Quaking Regulates Formation of circRNAs. *Cell* 160(6), 1125–1134. doi: 10.1016/j.cell.2015.02.014

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* 40(7), 3131–3142. doi: 10.1093/nar/gkr1009

Davis, A. P., Wiegers, T. C., Rosenstein, M. C., and Mattingly, C. J. (2012). MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database* (*oxford*), 9. doi: 10.1093/database/bar065

Dong, R., Ma, X. K., Li, G. W., and Yang, L. (2018). CIRCpedia v2: An Updated Database for Comprehensive Circular RNA Annotation and Expression Comparison. *Genomics Proteomics Bioinf.* 16(4), 226–233. doi: 10.1016/j.gpb.2018.08.001

Fan, C. Y., Lei, X. J., Fang, Z. Q., Jiang, Q. H., and Wu, F. X. (2018). CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database* (*oxford*), 6. doi: 10.1093/database/bay044

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

Glažar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *Rna* 20(11), 1666–1670. doi: 10.1261/rna.043687.113

Gu, C. L., Liao, B., Li, X. Y., Cai, L. J., Li, Z. J., Li, K. Q., et al. (2017). Global network random walk for predicting potential human lncRNA-disease associations. *Sci. Rep.* 7, 11. doi: 10.1038/s41598-017-12763-z

Guo, W. Z., Zhang, J. K., Zhang, D. Y., Cao, S. L., Li, G. Q., Zhang, S. J., et al. (2017). Polymorphisms and expression pattern of circular RNA circ-ITCH contributes to the carcinogenesis of hepatocellular carcinoma. *Oncotarget* 8(29), 48169–48177. doi: 10.18632/oncotarget.18327

Haveliwala, T. H. (2002). "Topic-sensitive PageRank", in: *Proceedings of the 11th international conference on World Wide Web.* (Honolulu, Hawaii, USA: ACM). doi: 10.1145/511446.511513

Hsiao, K. Y., Lin, Y. C., Gupta, S. K., Chang, N., Yen, L., Sun, H. S., et al. (2017). Noncoding Effects of Circular RNA CCDC66 Promote Colon Cancer Growth and Metastasis. *Cancer Res.* 77(9), 2339–2350. doi: 10.1158/0008-5472.can-16-1883

Hu, Y., Zhao, T. Y., Zang, T. Y., Zhang, Y., and Cheng, L. (2019). Identification of Alzheimer's Disease-Related Genes Based on Data Integration Method. *Front. Genet.* 9, 7. doi: 10.3389/fgene.2018.00703

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

Jens, M. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928

Jiang, J. J., Wang, N., Chen, P., Zhang, J., and Wang, B. (2017). DrugECs: An Ensemble System with Feature Subspaces for Accurate Drug-Target Interaction Prediction. *Biomed Res. Int.* 10. doi: 10.1155/2017/6340316

Kang, Z. Z., Pei, Y. J., and Wu, H. (2014). *RWR-based Resources Recommendation on Weighted and Clustered Folksonomy Graph*. New York: Ieee. doi: 10.1109/icebe.2014.30

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(D1), D1071-D1078. doi: 10.1093/nar/gku1011

Lei, X. J., Fang, Z. Q., Chen, L. N., and Wu, F. X. (2018). PWCDA: Path Weighted Method for Predicting circRNA-Disease Associations. *Int. J. Of Mol. Sci.* 19(11), 13. doi: 10.3390/ijms19113410

Li, P., Qing, Y. X., and Cheng, L. G. (2015). The emerging landscape of circular RNA ciRS-7 in cancer (Review). *Oncol. Rep.* 33, 2669–2674. doi: 10.3892/or.2015.3904

Li, Y., Zheng, Q. P., Bao, C. Y., Li, S. Y., Guo, W. J., Zhao, J., et al. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Research* 25(8), 981–984. doi: 10.1038/cr.2015.82

Lin, F., Zhao, G. A., Chen, Z. G., Wang, X. H., Lu, F. H., Zhang, Y. C., et al. (2018). Network correlation of circRNA-miRNA and the possible regulatory mechanism in acute myocardial infarction. *Zhonghua yi xue za zhi* 98(11), 851–854. doi: 10.3760/cma.j.issn.0376-2491.2018.11.012

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bull. Med. Lib. Assoc.* 88(3), 265–266.

Lukiw, W. (2013). Circular RNA (circRNA) in Alzheimer's disease (AD). *Front. Genet.* 4, 307. doi: 10.3389/fgene.2013.00307

Meng, S. J., Zhou, H. C., Feng, Z. Y., Xu, Z. H., Tang, Y., Li, P. Y., et al. (2017). CircRNA: functions and properties of a novel potential biomarker for cancer. *Molecular Cancer* 16, 8. doi: 10.1186/s12943-017-0663-2

Motieghader, H., Kouhsar, M., Najafi, A., Sadeghi, B., and Masoudi-Nejad, A. (2017). mRNA-miRNA bipartite network reconstruction to predict prognostic module biomarkers in colorectal cancer stage differentiation. *Mol. Biosyst.* 13(10), 2168-2180. doi: 10.1039/c7mb00400a

Peng, J. J., Guan, J. J., and Shang, X. Q. (2019a). Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder. *Front. Genet.* 10, 6. doi: 10.3389/fgene.2019.00226

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019b). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* (in press). doi: 10.1093/bioinformatics/btz254

Peng, J. J., Zhang, X. S., Hui, W. W., Lu, J. Y., Li, Q. Q., Liu, S. H., et al. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *Bmc Syst. Biol.* 12, 8. doi: 10.1186/s12918-018-0539-0

Qiu, L. P., Wu, Y. H., Yu, X. F., Tang, Q., Chen, L., and Chen, K. P. (2018). The Emerging Role of Circular RNAs in Hepatocellular Carcinoma. *J. Of Cancer* 9(9), 1548–1559. doi: 10.7150/jca.24566

Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., et al. (2015). Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Mol. Cell* 58(5), 870–885. doi: 10.1016/j.molcel.2015.03.027

Shao, B., Liu, B., and Yan, C. (2018). SACMDA: miRNA-disease association prediction with short acyclic connections in heterogeneous graph. *Neuroinformatics* 16, 373–382. doi: 10.1007/s12021-018-9373-1

Shao, Y. F., Chen, L. B., Lu, R. D., Zhang, X. J., Xiao, B. X., Ye, G. L., et al. (2017). Decreased expression of hsa_circ_0001895 in human gastric cancer and its clinical significances. *Tumor Biol.* 39(4), 6. doi: 10.1177/1010428317699125

Vo, J. N., Zhang, Y. J., Shukla, S., Xiao, L. B., Robinson, D., Wu, Y. M., et al. (2018). The landscape of circular RNA in cancer. *Cancer Res.* 78(13), 2. doi: 10.1158/1538-7445.am2018-3288

Wang, Y. T., Juan, L. R., Chu, Y. S., Wang, R. J., Zang, T. Y., and Wang, Y. D. (2017). "FNSemSim: an improved disease similarity method based on network fusion", in: *2017 Ieee International Conference on Bioinformatics And Biomedicine*. (Kansas City, MO, USA: Ieee). doi: 10.1109/BIBM.2017.8217726

Xia, S. Y., Feng, J., Chen, K., Ma, Y. B., Gong, J., Cai, F. F., et al. (2018). CSCD: a database for cancer-specific circular RNAs. *Nucleic Acids Res.* 46(D1), D925-D929. doi: 10.1093/nar/gkx863

Xu, S., Zhou, L. Y., Ponnusamy, M., Zhang, L. X., Dong, Y. H., Zhang, Y. H., et al. (2018). A comprehensive review of circRNA: from purification and identification to disease marker potential. *Peerj* 6, 28. doi: 10.7717/peerj.5503

Xu, Y. (2017). An overview of the main circRNA databases. *Non-coding RNA Investigation* 1(4). doi: 10.21037/ncri.2017.11.05

Yao, D. X., Zhang, L., Zheng, M. Y., Sun, X. W., Lu, Y., and Liu, P. Y. (2018). Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci. Rep.* 8, 6. doi: 10.1038/s41598-018-29360-3

Yi, Q., Gharbi, N., Xing, Y., Olsen, J. R., Blicher, P., Dalhus, B., et al. (2016). Axitinib blocks Wnt/beta-catenin signaling and directs asymmetric cell division in cancer. *Proc. Natl. Acad. Sci. U. S. A* 113(33), 9339–9344. doi: 10.1073/pnas.1604520113

Yu, L., Gong, X. J., Sun, L., Zhou, Q. Y., Lu, B. L., and Zhu, L. Y. (2016). The Circular RNA Cdr1as Act as an Oncogene in Hepatocellular Carcinoma through Targeting miR-7 Expression. *PLoS One* 11(7), 10. doi: 10.1371/journal.pone.0158347

Zhang, C. L., Xiong, J., Yang, Q., Wang, Y., Shi, H. Q., Tian, Q. Q., et al. (2018). Profiling and bioinformatics analyses of differential circular RNA expression in prostate cancer cells. *Future Sci. Oa* 4(9), 21. doi: 10.4155/fsoa-2018-0046

Zhang, W., Chen, Y., and Li, D. (2017). Drug-target interaction prediction through label propagation with linear neighborhood information. *Molecules* 22, 2056. doi: 10.3390/molecules22122056

Zhao, Z., Wang, K. Y., Wu, F., Wang, W., Zhang, K. N., Hu, H.M., et al. (2018). circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis.* 9, 2. doi: 10.1038/s41419-018-0503-3

Zhao, Z. J., and Shen, J. (2017). Circular RNA participates in the carcinogenesis and the malignant behavior of cancer. *Rna Biol.* 14(5), 514-521. doi: 10.1080/15476286.2015.1122162

Zheng, Q. P., Bao, C. Y., Guo, W. J., Li, S. Y., Chen, J., Chen, B., et al. (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Comm.* 7, 13. doi: 10.1038/ncomms11215

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

TECHNOLOGY AND CODE
published: 18 September 2019
doi: 10.3389/fgene.2019.00842

iRO-PsekGCC: Identify DNA Replication Origins Based on Pseudo k-Tuple GC Composition

Bin Liu[1,2*†], *Shengyu Chen*[3†], *Ke Yan*[4] *and Fan Weng*[4]

[1] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China,* [2] *Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing, China,* [3] *School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN, United States,* [4] *School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China*

Summary: Identification of replication origins is playing a key role in understanding the mechanism of DNA replication. This task is of great significance in DNA sequence analysis. Because of its importance, some computational approaches have been introduced. Among these predictors, the iRO-3wPseKNC predictor is the first discriminative method that is able to correctly identify the entire replication origins. For further improving its predictive performance, we proposed the Pseudo k-tuple GC Composition (PsekGCC) approach to capture the "GC asymmetry bias" of yeast species by considering both the GC skew and the sequence order effects of *k*-tuple GC Composition (*k*-GCC) in this study. Based on PseKGCC, we proposed a new predictor called iRO-PsekGCC to identify the DNA replication origins. Rigorous jackknife test on two yeast species benchmark datasets (*Saccharomyces cerevisiae*, *Pichia pastoris*) indicated that iRO-PsekGCC outperformed iRO-3wPseKNC. It can be anticipated that iRO-PsekGCC will be a useful tool for DNA replication origin identification.

Availability and implementation: The web-server for the iRO-PsekGCC predictor was established, and it can be accessed at http://bliulab.net/iRO-PsekGCC/.

Keywords: replication origin identification, pseudo *k*-tuple GC composition, random forest, web-server, DNA sequence analysis

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Xun Lan,
Tsinghua University, China
Qiwen Dong,
East China Normal University, China

****Correspondence:***
Bin Liu
bliu@bliulab.net

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 July 2019*
Accepted: *13 August 2019*
Published: *18 September 2019*

Citation:
Liu B, Chen S, Yan K and Weng F (2019) iRO-PsekGCC: Identify DNA Replication Origins Based on Pseudo k-Tuple GC Composition. Front. Genet. 10:842. doi: 10.3389/fgene.2019.00842

INTRODUCTION

In the process of the cell cycle, DNA replication is one of the most important steps (Shirahige et al., 1998). Since the DNA replication is initiated from a specific region, which is called replication origin, identifying the DNA replication origin is especially important for studying drug developments, cell life activities, genetic engineering, etc. (Méchali, 2010). Experimental methods detect the replication origins by using Chromatin immunoprecipitation (Chip) with high cost (Lubelsky et al., 2012). Therefore, researchers are seeking computational methods to efficiently predict the replication origins only based on the sequence information. Compared with non-replication origins, replication origins show uneven distribution of G (guanine) and C (cytosine) (Lobry, 1996), and the concept of "GC Skew" (Grigoriev, 1998) was proposed. Later, some computational methods incorporated these characteristics into the predictors based on the replication origins (Zhang and Zhang, 1991; Zhang and Zhang, 1994; Grigoriev, 1998; Roten et al., 2002; Thomas et al., 2007; Gao and Zhang, 2008; Luo et al., 2014; Bu et al., 2018). In order to further improve the predictive performance, the discriminative methods were proposed by using both the information of the positive and negative samples (Chen et al., 2012; Li et al., 2015; Zhang et al., 2016), and all of these methods mentioned above achieved the-state-of-the-art performance. A recent method iRO-3wPseKNC incorporated the "GC asymmetry bias" (Lobry, 1996; Grigoriev, 1998; Lubelsky et al., 2012; Li et al., 2014) into the prediction by representing the entire replication origins based on three-window-based PseKNC (3wPseKNC) (Liu et al., 2018b). Feature extraction methods are the keys for the performance improvement. In this regard, many features have been proposed, which can be easily generated by some software tools.

These existing computational methods have significantly enhanced the development of this hot area, but they all suffer from certain disadvantages or limitations, for example, as discussed above the GC Skew is an important feature of replication origins, but all the existing discriminative methods failed to directly use GC Skew to construct the predictors. Furthermore, the existing feature extraction methods cannot reflect the uneven distribution of G and C. To solve these problems, we followed the framework of iRO-3wPseKNC (Liu et al., 2018b), and proposed an improved predictor called iRO-PsekGCC for replication origin identification. iRO-PsekGCC cannot only capture the CG asymmetry bias by using *k*-tuple GC composition (or *k*-GCC), but can also incorporate the GC Skew into the concept of PseKNC (Chen et al., 2014a; Chen et al., 2014b).

MANUSCRIPT FORMATTING

Benchmark Datasets

In order to evaluate the performance of the proposed method, two recently established benchmark datasets of the *Saccharomyces cerevisiae* and *Pichia pastoris* (Liu et al., 2018b) were employed in this study, because they showed clear CG asymmetry distributions, which can be represented as:

$$\mathbb{S}_\tau = \mathbb{S}_\tau^+ \cup \mathbb{S}_\tau^-,\ \tau = \begin{cases} 1 \text{ for } \textit{Saccharomyces cerevisiae} \\ 2 \text{ for } \textit{Pichia pastoris} \end{cases} \tag{1}$$

where the symbol ∪ represents the union, and $\mathbb{S}_-^+$ represents the positive dataset containing 340 replication origins, and $\mathbb{S}_1^-$ represents the negative dataset containing 342 non-replication origins; 305 replication origins are in positive dataset $\mathbb{S}_2^+$, and 302 non-replication origins are in the negative dataset $\mathbb{S}_2^-$. For both of the two benchmark datasets, the redundant samples have been removed by using CD-HIT software tool (Li and Godzik, 2006) with the most stringent cut-off threshold (80%).

Pseudo *k*-Tuple GC Composition (PsekGCC)

One of the key steps for constructing machine-learning predictors for analyzing biological sequences is feature extraction. Following the framework of three-window-based PseKNC (3wPseKNC) (Liu et al., 2018b), we proposed a feature extraction method called "Pseudo k-tuple GC composition (PseKGCC)" to directly incorporate the CG asymmetry bias (Lobry, 1996; Grigoriev, 1998; Lubelsky et al., 2012; Li et al., 2014) and GC skew (Grigoriev, 1998) into the predictor. In the following sections, we will introduce how to represent DNA samples by using PseKGCC.

A DNA sequence D can be formulated as follow:

$$\mathbf{D} = \mathrm{N}_1\mathrm{N}_2\mathrm{N}_3 \cdots \mathrm{N}_i \cdots \mathrm{N}_L \quad (i = 1, 2, 3 \cdots, L) \tag{2}$$

where L denotes the length of **D**, and

$$\begin{aligned} N_i \in \{&\mathrm{A(adenine), C(cytosine), G(guanine),} \\ &\mathrm{T(thymine)}\}, \quad (i = 1, 2, 3, \cdots, L) \end{aligned} \tag{3}$$

which represents the *i*-th nucleobase in the sequence, and fi ∈ denotes the "member of"" in the set theory. Following the study (Liu et al., 2018b), **D** is divided into three windows by two parameters ε and δ, including front window, middle window, and rear window respectively. ε and 1 − δ denote the percentage of total nucleobases of **D** in the front window and rear window, respectively. The front window, middle window and rear window can be represented as **D**[1,η], D[η + 1,ξ], and **D**[ξ + 1, L], respectively, where η and ξ are defined as (Liu et al., 2018b),

$$\begin{cases} \eta = \mathrm{Int}^{\mathrm{C}}[L \times \varepsilon] \\ \xi = \mathrm{Int}^{\mathrm{C}}[L \times \delta] \end{cases}, \quad (0 < \varepsilon < \delta < 1.0) \tag{4}$$

where the symbol $\mathrm{Int}^{\mathrm{C}}$ represents the ceiling operator, which means to return the smallest integer value greater than or equal to the float number .

According to (Liu et al., 2018b), if D is formulated by the *k*-tuple nucleotide (or *k*-mer) (Liu et al., 2019b; Liu, 2017) based on the three windows strategy, it can be represented as follow:

$$\begin{aligned} \mathbf{D} = \Big[& f_1^{(1)} \cdots f_v^{(1)} \cdots f_{4^k}^{(1)} f_{4^k+1}^{(2)} \cdots f_{4^k+v}^{(2)} \cdots f_{4^k+v}^{(2)} \cdots \\ & f_{2\times 4^k}^{(2)} f_{2\times 4^k+1}^{(3)} \cdots f_{2\times 4^k+v}^{(3)} \cdots f_{3\times 4^k}^{(3)} \Big]^{\mathbf{T}} \end{aligned} \tag{5}$$

where in vector operations, symbol 'T' denotes the transformation symbol, and in the sample D, the normalized frequency values of the corresponding *k*-tuple nucleotides appearing in the front window, middle window and rear window are represented as $f^{(1)}$, $f^{(2)}$, $f^{(3)}$, respectively. The feature vector's dimension is 3×4^k.

This strategy was proposed to capture the patterns of "GC asymmetry bias" in yeast species genomes, and it is able to improve the predictive performance for identifying replication origins among multiple yeast species genomes. However, this approach has the following disadvantages: 1) the three windows strategy can only capture the local GC asymmetry bias of replication origins, but it cannot incorporate the GC asymmetry bias in a global fashion; 2) for large *k* values of *k*-tuple nucleotide, the dimension of the resulting feature vectors is high, which will cause high dimension disaster.

In order to overcome these disadvantages, we proposed a new composition of DNA sequence called "*k*-tuple GC composition (or *k*-GCC)" to capture the GC preference in the replication origins and their global interactions. *k*-GCC treats A (adenine) and T (thymine) as one nucleotide type represented as *. Therefore, the alphabet of *k*-GCC is

$$\mathrm{N}_i \in \{\mathrm{G(guanine)}, \mathrm{C(cytosine)}, *\}, \quad (i = 1, 2, 3, \cdots, L) \tag{6}$$

Therefore, by replacing the *k*-tuple by k-GCC, a DNA sequence D can be represented as:

$$\mathbf{D} = \Big[f_1^{(1)} \cdots f_v^{(1)} \cdots f_{3^k}^{(1)} f_{3^k+1}^{(2)} \cdots f_{3^k+v}^{(2)} \cdots f_{2\times3^k}^{(2)} f_{2\times3^k+1}^{(3)} \cdots \\ f_{2\times3^k+1}^{(3)} \cdots f_{2\times3^k+v}^{(3)} \cdots f_{3\times3^k}^{(3)} \Big]^{\mathbf{T}} \tag{7}$$

Compared with Equation 5, the *k*-GCC can efficiently reduce the dimension of the feature vector from 3 × 4^k to 3 × 3^k by focusing on the GC composition.

The proposed Pse-KGCC incorporates both the *k*-GCC and GC skew into the framework of PseKNC (Chen et al., 2014a), which can be represented as:

$$\mathbf{D} = \begin{matrix} [\phi_1 \cdots \phi_{3^k} \cdots \phi_{3^k+\lambda} \ \phi_{3^k+\lambda+1} \cdots \phi_{(3^k+\lambda)+3^k} \cdots \phi_{2\times(3^k+\lambda)} \ \phi_{2\times(3^k+\lambda)+1} \\ \cdots \phi_{2\times(3^k+\lambda)+3^k} \cdots \phi_{3\times(3^k+\lambda)}]^{\mathbf{T}} \end{matrix} \tag{8}$$

where

$$\phi_u = \begin{cases} \dfrac{f_u^{(1)}}{\sum_{i=1}^{3^k} f_i^{(1)} + w\sum_{j=1}^{\lambda} \theta_j^{(1)}} & 1 \le u \le 3^k \\ \dfrac{w\theta_{u-3^k}^{(1)}}{\sum_{i=1}^{3^k} f_i^{(1)} + w\sum_{j=1}^{\lambda} \theta_j^{(1)}} & 3^k + 1 \le u \le 3^k + \lambda \\ \dfrac{f_u^{(2)}}{\sum_{i=3^k+1}^{2\times3^k} f_i^{(2)} + w\sum_{j=1}^{\lambda} \theta_j^{(2)}} & 3^k + \lambda + 1 \le u \le 2\times3^k + \lambda \\ \dfrac{w\theta_{u-(3^k+\lambda)-3^k}^{(2)}}{\sum_{i=3^k+1}^{2\times3^k} f_i^{(2)} + w\sum_{j=1}^{\lambda} \theta_j^{(2)}} & 2\times3^k + \lambda + 1 \le u \le 2\times3^k + 2\lambda \\ \dfrac{f_u^{(3)}}{\sum_{i=2\times3^k+1}^{3\times3^k} f_i^{(3)} + w\sum_{j=1}^{\lambda} \theta_j^{(3)}} & 2\times3^k + 2\lambda + 1 \le u \le 3\times3^k + 2\lambda \\ \dfrac{w\theta_{u-2\times(3^k+\lambda)-3^k}^{(3)}}{\sum_{i=2\times3^k+1}^{3\times3^k} f_i^{(3)} + w\sum_{j=1}^{\lambda} \theta_j^{(3)}} & 3\times3^k + 2\lambda + 1 \le u \le 3\times3^k + 3\lambda \end{cases} \tag{9}$$

where λ denotes the highest tier correlation of the *k*-GCC nucleotides in each local window of **D**, whose the value is an integer. *w* is a float number that represents the weight factor, and the value of *w* is between 0 and 1. In the front window, the middle window and the rear window, the correlation factor of the *j*-th tier is represented as $\theta_j^{(1)}$, $\theta_j^{(2)}$, and $\theta_j^{(3)}$, respectively. The GC skew value of the *k*-GCC nucleotides separated by *j* nucleotides is used to represent the correlation factor of the *j*-th tier in each local window. (**Figure 1**). $\theta_j^{(1)}$, $\theta_j^{(2)}$, and $\theta_j^{(3)}$ can be calculated by

$$128\begin{cases} \theta_j^{(1)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{\eta-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}[\frac{\eta-k}{j}]} \Theta(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}) \\ \theta_j^{(2)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{\xi-\eta-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}[\frac{\xi-\eta-k}{j}]} \Theta(\mathrm{N}_{\eta+i\times j+1}\mathrm{N}_{\eta+i\times j+2}\cdots\mathrm{N}_{\eta+i\times j+k}) & \begin{matrix} j = 1, 2, \cdots, \lambda; \\ \lambda \le min(\eta, \xi-\eta, L-\xi) \end{matrix} \\ \theta_j^{(3)} = \dfrac{1}{\mathrm{Int}^{\mathrm{C}}[\frac{L-\xi-k}{j}]+1} \sum_{i=0}^{\mathrm{Int}^{\mathrm{C}}\left[\frac{L-\xi-k}{j}\right]} \Theta\left(\mathrm{N}_{\xi+i\times j+1}\mathrm{N}_{\xi+i\times j+2}\cdots\mathrm{N}_{\xi+i\times j+k}\right) \end{cases} \tag{10}$$

where $\mathrm{Int}^{\mathrm{C}}[\frac{\eta-k}{j}]+1$ denotes the number of the *k*-GCC in the corresponding local window, and Θ($\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2} \cdots \mathrm{N}_i \times j + k$) is the GC Skew (Lobry, 1996; Grigoriev, 1998; Li et al., 2014) of the *i*-th *k*-GCC in the local window, which can be calculated by

$$\Theta\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) = \frac{f_G\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) - f_C\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right)}{f_G\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right) + f_C\left(\mathrm{N}_{i\times j+1}\mathrm{N}_{i\times j+2}\cdots\mathrm{N}_{i\times j+k}\right)} \tag{11}$$

where f_{G}($\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ N_i × j + k) denotes the frequency of G in the subsequence $\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ $\mathrm{N}_{i \times j + k}$$f_{\mathrm{C}}$($\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + k}$ ⋯ $\mathrm{N}_{i \times j + k}$) denotes the frequency of C in the subsequence $\mathrm{N}_{i \times j + 1}$ $\mathrm{N}_{i \times j + 2}$ ⋯ $\mathrm{N}_{i \times j + k}$, reflecting the CG asymmetry bias directly. Please note that for the terminal subsequence, if its length is less than *k*, then the GC skew will be calculated by all the available nucleotide residues.

Random Forest

Being widely used in bioinformatics (Zhao et al., 2014; Su et al., 2019), Random Forest (RF) (Ho, 1995; Barandiaran, 1998) is a machine learning classifier. Its training process can prevent overfitting (Hastie et al., 2008). The Random Forest model was implemented by calling the command line RandomForestClassifier ("max_features='sqrt', min_samples_leaf=1, min_samples_split=2, criterion = 'gini', $\mathcal{F}$ = optimize-d value ") with the help of the Scikit-learn package (Pedregosa et al., 2011), where the values of $\mathcal{F}$ represents the number of the trees in the forest, and it was set as 600 for both the two benchmark datasets (cf. Equation 1).

Ensemble Learning

Previous studies (Zou et al., 2015; Liu et al., 2016a; Chen et al., 2016b; Chen et al., 2017a; Chen et al., 2017b; Liu et al., 2018a) have demonstrated that fusing a series of individual predictors

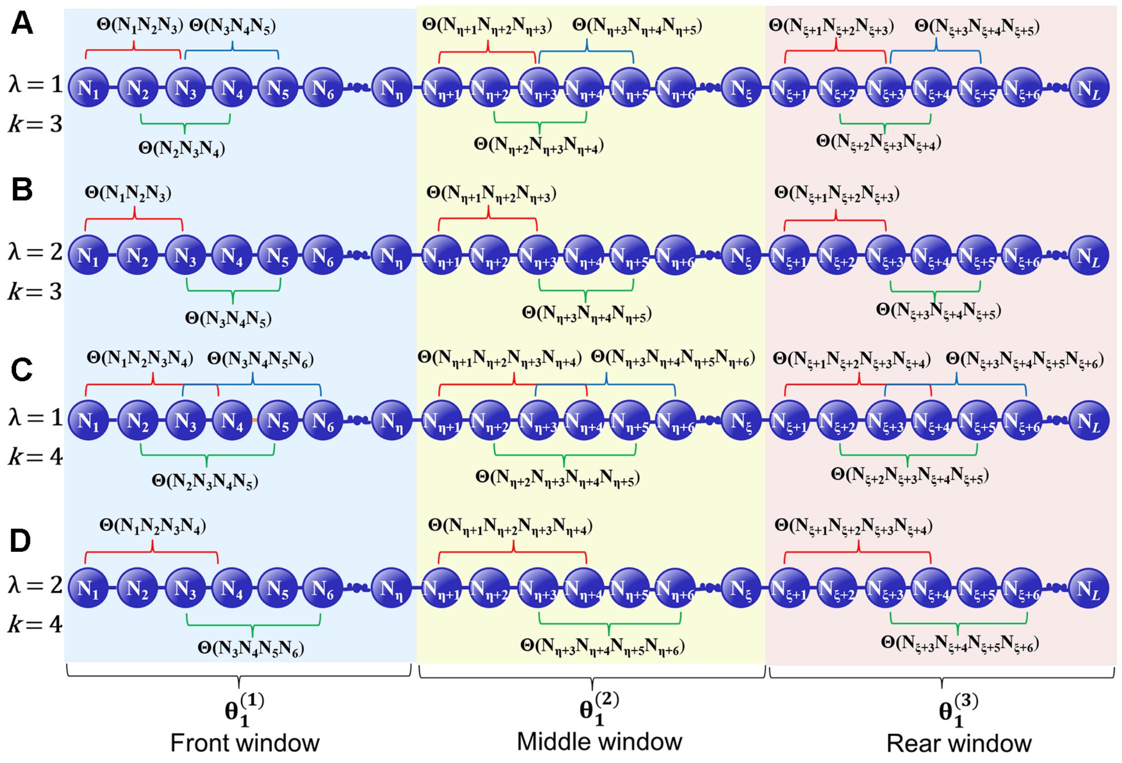

A
$\Theta(N_1N_2N_3)$ $\Theta(N_3N_4N_5)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3})$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5})$
$\lambda = 1$
$k = 3$
N_1 N_2 N_3 N_4 N_5 N_6 N_η $N_{\eta+1}$ $N_{\eta+2}$ $N_{\eta+3}$ $N_{\eta+4}$ $N_{\eta+5}$ $N_{\eta+6}$ N_ξ $N_{\xi+1}$ $N_{\xi+2}$ $N_{\xi+3}$ $N_{\xi+4}$ $N_{\xi+5}$ $N_{\xi+6}$ N_L
$\Theta(N_2N_3N_4)$ $\Theta(N_{\eta+2}N_{\eta+3}N_{\eta+4})$ $\Theta(N_{\xi+2}N_{\xi+3}N_{\xi+4})$
B
$\Theta(N_1N_2N_3)$ $\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3})$ $\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3})$
$\lambda = 2$
$k = 3$
N_1 N_2 N_3 N_4 N_5 N_6 N_η $N_{\eta+1}$ $N_{\eta+2}$ $N_{\eta+3}$ $N_{\eta+4}$ $N_{\eta+5}$ $N_{\eta+6}$ N_ξ $N_{\xi+1}$ $N_{\xi+2}$ $N_{\xi+3}$ $N_{\xi+4}$ $N_{\xi+5}$ $N_{\xi+6}$ N_L
$\Theta(N_3N_4N_5)$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5})$
C
$\Theta(N_1N_2N_3N_4)$ $\Theta(N_3N_4N_5N_6)$
$\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3}N_{\eta+4})$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5}N_{\eta+6})$
$\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3}N_{\xi+4})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5}N_{\xi+6})$
$\lambda = 1$
$k = 4$
N_1 N_2 N_3 N_4 N_5 N_6 N_η $N_{\eta+1}$ $N_{\eta+2}$ $N_{\eta+3}$ $N_{\eta+4}$ $N_{\eta+5}$ $N_{\eta+6}$ N_ξ $N_{\xi+1}$ $N_{\xi+2}$ $N_{\xi+3}$ $N_{\xi+4}$ $N_{\xi+5}$ $N_{\xi+6}$ N_L
$\Theta(N_2N_3N_4N_5)$ $\Theta(N_{\eta+2}N_{\eta+3}N_{\eta+4}N_{\eta+5})$ $\Theta(N_{\xi+2}N_{\xi+3}N_{\xi+4}N_{\xi+5})$
D
$\Theta(N_1N_2N_3N_4)$ $\Theta(N_{\eta+1}N_{\eta+2}N_{\eta+3}N_{\eta+4})$ $\Theta(N_{\xi+1}N_{\xi+2}N_{\xi+3}N_{\xi+4})$
$\lambda = 2$
$k = 4$
N_1 N_2 N_3 N_4 N_5 N_6 N_η $N_{\eta+1}$ $N_{\eta+2}$ $N_{\eta+3}$ $N_{\eta+4}$ $N_{\eta+5}$ $N_{\eta+6}$ N_ξ $N_{\xi+1}$ $N_{\xi+2}$ $N_{\xi+3}$ $N_{\xi+4}$ $N_{\xi+5}$ $N_{\xi+6}$ N_L
$\Theta(N_3N_4N_5N_6)$ $\Theta(N_{\eta+3}N_{\eta+4}N_{\eta+5}N_{\eta+6})$ $\Theta(N_{\xi+3}N_{\xi+4}N_{\xi+5}N_{\xi+6})$
$\theta_1^{(1)}$
Front window
$\theta_1^{(2)}$
Middle window
$\theta_1^{(3)}$
Rear window

FIGURE 1 | A schematic diagram to illustrate how to calculate the GC Skew in the front, middle, and rear windows along a DNA sequence. **(A)** The coupling between all the contiguous *k*-GCC (*k* = 3); **(B)** The coupling between the second most contiguous *k*-GCC (*k* = 3); **(C)** The coupling between all the contiguous *k*-GCC (*k* = 4); **(D)** The coupling between the second most contiguous *k*-GCC (*k* = 4).

by a voting strategy can improve the predictive performance. In this regard, in this study an ensemble predictors was constructed by fusing 10 top performing individual predictors constructed by different parameter combinations of PseKGCC (see **Supplementary Information S1**), which can be represented as (Liu et al., 2016a):

$$\mathbb{RF}^{\mathrm{E}} = \mathrm{RF}(1) \forall \mathrm{RF}(2) \forall \cdots \forall \mathrm{RF}(i) \forall \cdots \forall \mathrm{RF}(10) = \forall_{i=1}^{10} \mathrm{RF}(i) \tag{12}$$

where $\mathbb{RF}^{\mathrm{E}}$ represents the ensemble classifier, ∀ represents the fusing operator, and RF(i) represents the basic Random Forest predictor.

The ensemble predictor is constructed based on the fusion score ß of the probabilities predicted by the 10 basic predictors, which can be calculated by

$$ß = \sum_{i=1}^{10} q_i P_i \tag{13}$$

where q_i is the weight of the i-th basic RF predictor, which was optimized by the genetic algorithm (Mitchell, 1998), and their values were listed in **Supplementary Information S1**. If the value of ß is higher than 0.5, it is a replication origin, otherwise, it is a non-replication origin. The flowchart of the iRO-PseKGCC is illuminated in **Figure 2**.

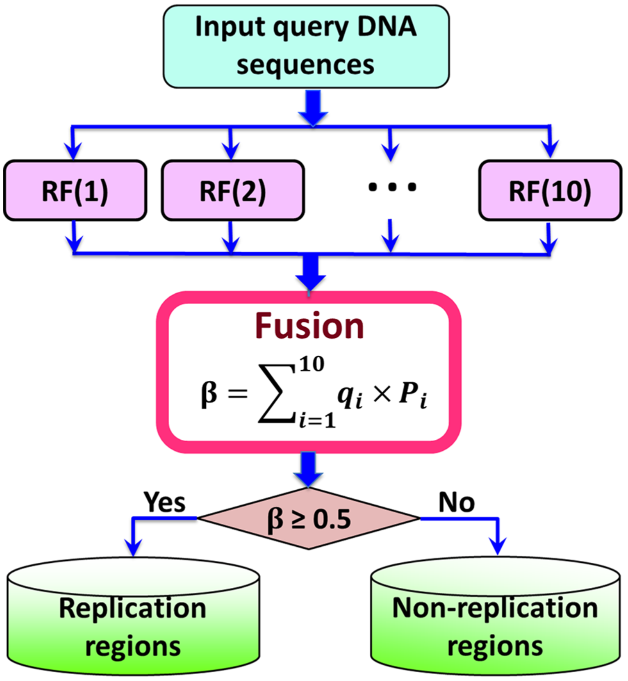


FIGURE 2 | A flowchart illustration to show how the iRO-PseKGCC predictor works.

Cross Validation

Three widely used cross-validation strategies include: i) independent test, ii) K-fold cross validation, and iii) jackknife test. Among these methods, only the jackknife test can achieve the unique results for the same benchmark dataset. Therefore, in this study, the jackknife test was employed to give the final predictive results. However, considering its high computational cost, during the parameter optimization process, the 5-fold cross-validation was used to reduce the computational cost (see *Optimize Parameters* section).

Evaluation Method of Performance

To evaluate the quality of the classifier for prediction of the replication origins, the four metrics are used (Feng et al., 2013; Chen et al., 2016c; Chen et al., 2019): i) the sensitivity, Sn, ii) the specificity, Sp, iii) the overall accuracy of the predictive results, Acc, iv) the Mathew's correlation coefficient, MCC, and v) Arear under ROC Curve, AUC (Chen et al., 2016a), defined as:

$$\begin{cases} \text{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \le \text{Sn} \le 1 \\ \text{Sp} = 1 - \frac{N_+^-}{N^-} & 0 \le \text{Sp} \le 1 \\ \text{Acc} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \text{ACC} \le 1 \\ \text{MCC} = \dfrac{1 - \left(\frac{N_-^+ + N_+^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- + N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ + N_+^-}{N^-}\right)}} & -1 \le \text{MCC} \le 1 \\ AUC & \text{Arear under ROC Curve} \end{cases} \tag{14}$$

where N^+ denotes the number of all the positive samples (replication origins), N^- denotes the number of all the negative samples (non-replication origins), N_-^+ denotes the number of the positive samples (replication origins) incorrectly predicted as the negative samples (non-replication origins), N_+^- denotes the number of the negative samples (non-replication origins) incorrectly predicted as the positive samples (replication origins). More information of these performance measures can refer to Liu et al. (2016b).

RESULTS AND DISCUSSION

Optimize Parameters

There are five parameters in PseKGCC according to Equations 4–9. These parameters were optimized by the following equations:

$$\begin{cases} 0.15 \le \varepsilon \le 0.5, & \text{with step} \triangle\varepsilon = 0.05 \\ 0.5 < \delta \le 0.85, & \text{with step} \triangle\delta = 0.05 \\ 3 \le k \le 7, & \text{with step} \triangle k = 1 \\ 1 \le \lambda \le 10, & \text{with step} \triangle\lambda = 3 \\ 0.1 \le w \le 1, & \text{with step} \triangle w = 0.1 \end{cases} \tag{15}$$

The fivefold cross-validation was employed to search the optimal parameters by gridding method so as to reduce the time consumption, and the predictive results of the top 10 performing predictors, and their optimized parameters were listed in **Supplementary Information S1**.

Comparison With Other Methods

To the best knowledge of ours, iRO-3wPseKNC (Liu et al., 2018b) is the only existing predictor that is able to predict the entire replication origins. All the other predictors can only predict the fragments of replication origins. Therefore, the performance of the proposed iRO-PseKGCC was compared with iRO-3wPseKNC on the two benchmark datasets, and the results were listed in **Table 1**, from which we can see that iRO-PseKGCC obviously outperformed iRO-3wPseKNC in terms of the five performance measures (cf. Equation 14), indicating that the proposed PseKGCC feature is able to capture the GC asymmetry bias, and incorporate the GC skew into the predictor. Therefore, iRO-PseKGCC is an efficient approach for improving the predictive performance.

Feature Analysis

Random forest is a combination classifier model composed of decision tree classifiers. During the process of constructing each tree by the "Bootstrap" method (Efron, 1992), samples not extracted for training the corresponding tree can be used to make "Out Of Bag" (OOB) error estimate (Breiman, 1996) to evaluate the generalization performance of a predictor. Based on the OOB error, the Mean Decrease Accuracy (MDA) (Jiang et al., 2007) can

TABLE 1 | The results of the iRO-PseKGCC Predictor and comparison with iRO-PseKGCC on the two benchmark datasets (cf. Equation 1) obtained by using jackknife test.

Species	Method	Acc(%)	MCC	Sn(%)	Sp(%)	AUC
Saccharomyces cerevisiae $\mathbb{S}_1$	iRO-PseKGCC[a]	76.46	0.5298	73.90	78.13	0.8129
	iRO-3wPseKNC[b]	72.95	0.4594	70.67	75.22	0.8084
Pichia pastoris $\mathbb{S}_2$	iRO-PseKGCC[a]	74.22	0.4844	74.51	73.93	0.8002
	iRO-3wPseKNC[c]	71.10	0.4222	69.93	72.28	0.7962

[a] *The parameters are listed in* ***Supplementary Information S1****.*

[b] *The predictor reported in (Liu et al., 2018b) with parameter* $\varepsilon = 0.25$, $\delta = 0.85$, $k = 5$, $\lambda = 6$, $w = 0.3$, *and* $\mathcal{F} = 700$.

[c] *The predictor reported in (Liu et al., 2018b) with parameter* $\varepsilon = 0.15$, $\delta = 0.55$, $k = 4$, $\lambda = 9$, $w = 0.3$, *and* $\mathcal{F} = 800$.

be used to estimate the importance of the features. The details of the process are (Jiang et al., 2007): 1) When training a Random Forest model, using the OOB samples to test the accuracy of each tree in the model; 2) Randomly disturb the value of the feature variable v in the OOB samples, and retest the accuracy of each tree; 3) Calculate the mean value of the decreasing accuracy between the two tests in all decision trees in the Random Forest model. The MDA value can reflect the importance of the corresponding feature.

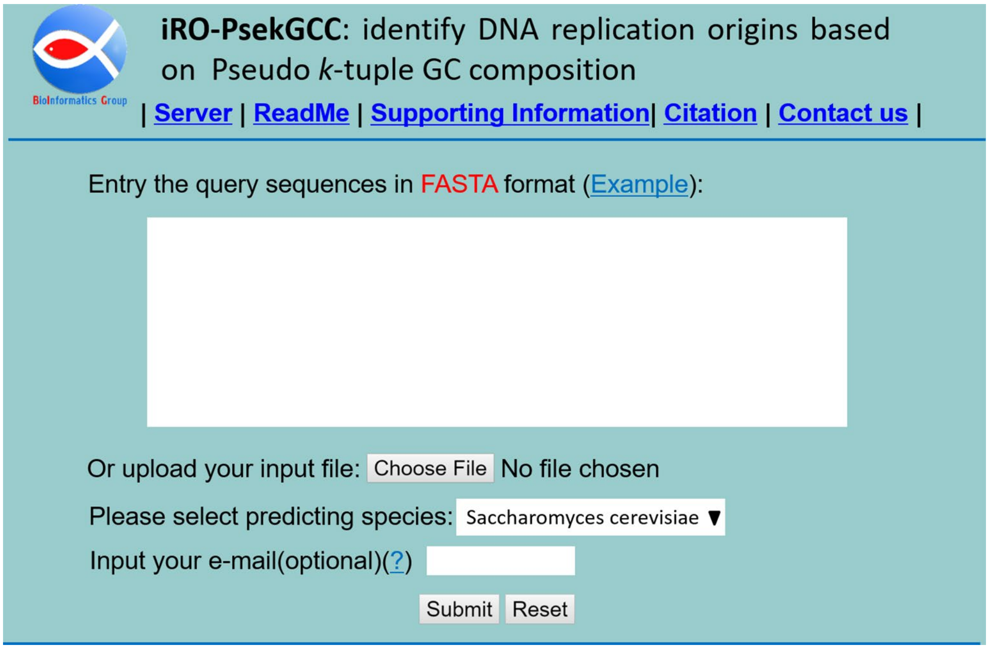
As shown in previous studies (Liu and Zhu, 2019; Liu et al. 2019a), feature analysis is critical for exploring the characteristics of the predictors. To explore the reason why the proposed predictor iRO-PseKGCC works so well, we analyzed the features of the two top performing iRO-PseKGCC predictors (see **Supplementary Information S1**) on the two benchmark datasets (cf. Equation 1) by MDA approach, and the results are listed in the **Table 2**, from which we can see that: 1) for both the two RF-based predictors, their most important features are the "***" and "*****," indicating the importance of the *k*-GCC; 2) The global sequence order effects measured by different λ values and GC skew values contribute to the performance improvement; 3) Features in certain local window show more discriminative powers than those in other windows, for examples, for *Pichia pastoris*, all the top 10 most important features are in the middle window, which is consistent with the previous observations that the nucleobase composition distribution is uneven along the replication origins (Lobry, 1996; Grigoriev, 1998; Frank and Lobry, 1999; Tillier and Collins, 2000; Liu et al., 2018b).

Web Server and User Guide

Web-servers are important for the researchers to implement the corresponding computational predictors. In this regard, for the user's convenience, we established a web-server named

TABLE 2 | The top 10 most important features of the top two performing RF-based predictors on the two benchmark datasets (cf. Equation 1).

Rank	Saccharomyces cerevisiae			Pichia pastoris		
	Feature	Window	MDA (%)	Feature	Window Index	MDA (%)
1	***	Rear window	20.49	*****	Middle window	15.89
2	***	Middle window	19.62	****G	Middle window	5.69
3	*GG	Rear window	9.04	G****	Middle window	5.38
4	GG*	Rear window	8.35	*C***	Middle window	5.23
5	*GG	Middle window	8.26	*G***	Middle window	5.14
6	λ = 1	Rear window	7.67	*CGCG	Middle window	3.99
7	GG*	Middle window	7.45	****C	Middle window	3.94
8	CC*	Middle window	7.31	***G*	Middle window	3.77
9	G*G	Rear window	6.64	*C*GG	Middle window	3.47
10	λ = 2	Rear window	6.12	C**G*	Middle window	3.40



iRO-PsekGCC: identify DNA replication origins based on Pseudo *k*-tuple GC composition

| Server | ReadMe | Supporting Information| Citation | Contact us |

Entry the query sequences in FASTA format (Example):

Or upload your input file: Choose File No file chosen

Please select predicting species: Saccharomyces cerevisiae ▼

Input your e-mail(optional)(?)

FIGURE 3 | A semi-screen shot to show the homepage of the web-server iRO-PseKGCC, which can be accessed at http://bliulab.net/iRO-PsekGCC/.

"iRO-PseKGCC." For users' convenience, a detailed user guide explaining how to use the web-server is given.

Step 1. Click on the web sites address http://bliulab.net/iRO-PsekGCC/ to open the web-server, then the main pages on the website as shown in **Figure 3** will appear in front of you. To see a brief introduction about the server, please click on the "Read Me" button.

Step 2. Choose the one specie from *Saccharomyces cerevisiae* or *Pichia pastoris*.

Step 3. The input sequences should be in the FASTA format. The sequence data can be uploaded *via* the "Browse" button or copy and paste or type into the input box directly.

Step 4. To see the predicted results, please click on the "Submit" button. For example, if the four query DNA sequences in the Example window are used as the queried data, the predictive results are the 1st and 2nd query sequences are replication origins, and the 3rd and 4th are non-replication origins.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. These data can be found here: https://academic.oup.com/bioinformatics/article-abstract/34/18/3086/4978052?redirectedFrom=fulltext

AUTHOR CONTRIBUTIONS

BL provided the main idea of the manuscript and wrote the manuscript. SC did the experiments and revised the manuscript. KY revised the manuscript and did the typesetting. FW did the experiments.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 61672184, 61732012, 61822306), Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (161063), Shenzhen Overseas High Level Talents Innovation Foundation (Grant No. KQJSCX20170327161949608), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008), and Scientific Research Foundation in Shenzhen (JCYJ20180306172207178).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00842/full#supplementary-material

REFERENCES

Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20 (8), 832–844. doi: 10.1109/34.709601

Breiman, L. (1996). "Out-of-bag estimation". Citeseer).

Bu, H. D., Hao, J. Q., Guan, J. H., and Zhou, S. G. (2018). Predicting enhancers from multiple cell lines and tissues across different developmental stages based on svm method. *Curr. Bioinform.* 13 (6), 655–660. doi: 10.2174/1574893613666180726163429

Chen, H., Peng, S., Dai, L., Zou, Q., Yi, B., Yang, X., et al. (2017a). Oral microbial community assembly under the influence of periodontitis. *PloS One* 12 (8), e0182259. doi: 10.1371/journal.pone.0182259

Chen, J., Guo, M., Li, S., and Liu, B. (2017b). Protdec-ltr2. 0: an improved method for protein remote homology detection by combining pseudo protein and supervised learning to rank. *Bioinformatics* 33 (21), 3473–3476. doi: 10.1093/bioinformatics/btx429

Chen, J., Guo, M., Wang, X., and Liu, B. (2016a). A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinform.* 19 (2), 231–244. doi: 10.1093/bib/bbw108

Chen, J., Long, R., Wang, X.-l., Liu, B., and Chou, K.-C. (2016b). dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Sci. Rep.* 6, 32333. doi: 10.1038/srep32333

Chen, J., Wang, X., and Liu, B. (2016c). IMiRNA-SSF: improving the identification of MicroRNA precursors by combining negative sets with different distributions. *Sci. Rep.* 6, 19062. doi: 10.1038/srep19062

Chen, W., Feng, P., and Lin, H. (2012). Prediction of replication origins by calculating DNA structural properties. *Febs Letters* 586 (6), 934–938. doi: 10.1016/j.febslet.2012.02.034

Chen, W., Lei, T.-Y., Jin, D.-C., Lin, H., and Chou, K.-C. (2014a). PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.* 456, 53–60. doi: 10.1016/j.ab.2014.04.001

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: Identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics*. doi: 10.1093/bioinformatics/btz015

Chen, W., Zhang, X., Brooker, J., Lin, H., Zhang, L., and Chou, K.-C. (2014b). PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics* 31 (1), 119–120. doi: 10.1093/bioinformatics/btu602

Efron, B. (1992). "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics* (New York: Springer), 569–593. doi: 10.1007/978-1-4612-4380-9_41

Feng, P.-M., Chen, W., Lin, H., and Chou, K.-C. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Anal. Biochem.* 442 (1), 118–125. doi: 10.1016/j.ab.2013.05.024

Frank, A., and Lobry, J. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238 (1), 65–77. doi: 10.1016/S0378-1119(99)00297-8

Gao, F., and Zhang, C.-T. (2008). Ori-Finder: a web-based system for finding oriC s in unannotated bacterial genomes. *BMC Bioinform.* 9 (1), 79. doi: 10.1186/1471-2105-9-79

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.* 26 (10), 2286–2290. doi: 10.1093/nar/26.10.2286

Hastie, T., Tibshirani, R., and Friedman, J., (2008). *The elements of statistical learning (2nd ed.)*. New York: Springer series in statistics New York.

Ho, T. K. (1995). "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition* (Washington: IEEE), 278–282.

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., and Lu, Z. (2007). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.* 35 (suppl_2), W339–W344. doi: 10.1093/nar/gkm368

Li, W.-C., Deng, E.-Z., Ding, H., Chen, W., and Lin, H. (2015). iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemom. Intell. Lab. Syst.* 141, 100–106. doi: 10.1016/j.chemolab.2014.12.011

Li, W.-C., Zhong, Z.-J., Zhu, P.-P., Deng, E.-Z., Ding, H., Chen, W., et al. (2014). Sequence analysis of origins of replication in the Saccharomyces cerevisiae genomes. *Front. Microbiol.* 5, 574. doi: 10.3389/fmicb.2014.00574

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22 (13), 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, B. (2017). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform*. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019b). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* doi: 10.1093/nar/gkz740

Liu, B., Li, C., and Yan, K. (2019a). DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.* doi: 10.1093/bib/bbz098

Liu, B., Li, K., Huang, D.-S., and Chou, K.-C. (2018a). iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 34 (22), 3835–3842. doi: 10.1093/bioinformatics/bty458

Liu, B., Long, R., and Chou, K.-C. (2016a). iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics* 32 (16), 2411–2418. doi: 10.1093/bioinformatics/btw186

Liu, B., Wang, S., Long, R., and Chou, K.-C. (2016b). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33 (1), 35–41. doi: 10.1093/bioinformatics/btw539

Liu, B., Weng, F., Huang, D.-S., and Chou, K.-C. (2018b). iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC. *Bioinformatics* 34 (18), 3086–3093. doi: 10.1093/bioinformatics/bty312

Liu, B., and Zhu, Y. (2019). ProtDec-LTR3.0: protein remote homology detection by incorporating profile-based features into Learning to Rank. *IEEE Access* 7, 102499–102507. doi: 10.1109/ACCESS.2019.292963

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.* 13 (5), 660–665. doi: 10.1093/oxfordjournals.molbev.a025626

Lubelsky, Y., MacAlpine, H. K., and MacAlpine, D. M. (2012). Genome-wide localization of replication factors. *Methods* 57 (2), 187–195. doi: 10.1016/j.ymeth.2012.03.022

Luo, H., Zhang, C.-T., and Gao, F. (2014). Ori-Finder 2, an integrated tool to predict replication origins in the archaeal genomes. *Front. Microbiol.* 5, 482. doi: 10.3389/fmicb.2014.00482

Méchali, M. (2010). Eukaryotic DNA replication origins: many choices for appropriate answers. *Nat. Rev. Mol. Cell Biol.* 11 (10), 728. doi: 10.1038/nrm2976

Mitchell, M. (1998). *An introduction to genetic algorithms*. Boston: MIT press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12 (Oct), 2825–2830. doi: 10.1524/auto.2011.0951

Roten, C.-A. H., Gamba, P., Barblan, J.-L., and Karamata, D. (2002). Comparative Genometrics (CG): a database dedicated to biometric comparisons of whole genomes. *Nucleic Acids Res.* 30 (1), 142–144. doi: 10.1093/nar/30.1.142

Shirahige, K., Hori, Y., Shiraishi, K., Yamashita, M., Takahashi, K., Obuse, C., et al. (1998). Regulation of DNA-replication origins during cell-cycle progression. *Nature* 395 (6702), 618. doi: 10.1038/27007

Su, R., Liu, X., Wei, L., and Zou, Q. (2019). Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods (San Diego, Calif.)* 166 (2019), 91–102. doi: 10.1016/j.ymeth.2019.02.009

Thomas, J. M., Horspool, D., Brown, G., Tcherepanov, V., and Upton, C. (2007). GraphDNA: a Java program for graphical display of DNA composition analyses. *BMC Bioinform.* 8 (1), 21. doi: 10.1186/1471-2105-8-21

Tillier, E. R., and Collins, R. A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.* 50 (3), 249–257. doi: 10.1007/s002399910029

Zhang, C.-J., Tang, H., Li, W.-C., Lin, H., Chen, W., and Chou, K.-C. (2016). iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 7 (43), 69783–69793. doi: 10.18632/oncotarget.11975

Zhang, C.-T., and Zhang, R. (1991). Analysis of distribution of bases in the coding sequences by a digrammatic technique. *Nucleic Acids Res.* 19 (22), 6313–6317. doi: 10.1093/nar/19.22.6313

Zhang, R., and Zhang, C.-T. (1994). Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.* 11 (4), 767–782. doi: 10.1080/07391102.1994.10508031

Zhao, X., Zou, Q., Liu, B., and Liu, X. (2014). Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteomics* 11 (4), 289–299. doi: 10.2174/157016461104150121115154

Zou, Q., Guo, J., Ju, Y., Wu, M., Zeng, X., and Hong, Z. (2015). Improving tRNAscan-SE annotation results *via* ensemble classifiers. *Mol. Inf.* 34 (11–12), 761–770. doi: 10.1002/minf.201500031

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

ORIGINAL RESEARCH
published: 20 September 2019
doi: 10.3389/fgene.2019.00855

Variance-Preserving Estimation of Intensity Values Obtained From Omics Experiments

Adèle H. Ribeiro[1]*, *Julia Maria Pavan Soler*[2] *and Roberto Hirata Jr.*[1]

[1] *Department of Computer Science, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil,*
[2] *Department of Statistics, Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil*

Faced with the lack of reliability and reproducibility in omics studies, more careful and robust methods are needed to overcome the existing challenges in the multi-omics analysis. In conventional omics data analysis, signal intensity values (denoted by *M* and values) are estimated neglecting pixel-level uncertainties, which may reflect noise and systematic artifacts. For example, intensity values from two-color microarray data are estimated by taking the mean or median of the pixel intensities within the spot and then subjected to a within-slide normalization by LOWESS. Thus, focusing on estimation and normalization of gene expression profiles, we propose a spot quantification method that takes into account pixel-level variability. Also, to preserve relevant variation that may be removed in LOWESS normalization with poorly chosen parameters, we propose a parameter selection method that is parsimonious and considers intrinsic characteristics of microarray data, such as heteroskedasticity. The usefulness of the proposed methods is illustrated by an application to real intestinal metaplasia data. Compared with the conventional approaches, the analysis is more robust and conservative, identifying fewer but more reliable differentially expressed genes. Also, the variability preservation allowed the identification of new differentially expressed genes. Using the proposed approach, we have identified differentially expressed genes involved in pathways in cancer and confirmed some molecular markers already reported in the literature.

Keywords: delta method, pixel-level uncertainty, spot quantification, optimal LOWESS normalization, two-color microarray, variability preservation, parameter selection

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Tian Qing Zheng,
Chinese Academy of Agricultural Sciences, China
Tianyi Zhao,
Harbin Institute of Technology, China

****Correspondence:***
Adèle H. Ribeiro
adele.ribeiro@usp.br

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 April 2019*
Accepted: *16 August 2019*
Published: *20 September 2019*

Citation:
Ribeiro AH, Soler JMP and Hirata R Jr (2019) Variance-Preserving Estimation of Intensity Values Obtained From Omics Experiments. Front. Genet. 10:855. doi: 10.3389/fgene.2019.00855

INTRODUCTION

The growing number of omics datasets (e.g., genomics, transcriptomics, proteomics, metabolomics) and the recent advances in multi-omics integration approaches have contributed to the better understanding of biological mechanisms and also the emergence of the personalized medicine. However, the lack of reliability and reproducibility in omics studies stands as one of the biggest obstacles in bridging the gap between research and practice of personalized medicine (Alyass et al., 2015; Karczewski and Snyder, 2018). Considering that inflated variability and non-robust estimation may lead to inaccurate and misleading results, this paper proposes improvements to the conventional estimation and normalization of the intensity values obtained from omics experiments. Specifically, the proposal is to estimate the intensity values by a method that accounts for the variability due to pixel-level uncertainties and to normalize these values by using LOWESS with suitably selected parameter values, preserving variation that may be relevant to subsequent analyses.

Image processing and fluorescence analysis are the preferred approaches for data quantification in microarray technologies. Although microarrays have been predominantly used since the end of the nineties to measure gene expression levels, they remain widely used to detect other omics data types, including microRNA expression, DNA methylation, single-nucleotide polymorphisms (SNPs), and copy number variants (CNVs) (Goodwin et al., 2016). After hybridization and cleaning of the target molecules, the array is scanned by activation with lasers at different wavelengths (one for each of the fluorophores used), and each laser channel generates an image. The pixel intensities within each spot in these microarray images are summarized to represent the hybridization signal. Depending on the platform (e.g., gene expression array, DNA methylation array, SNP array, and comparative genomic hybridization [CGH] array), the interpretation of this signal is different (e.g., gene expression levels, methylation levels, allele frequencies, and copy number alterations).

The continuance of the microarray technology can be mainly explained by the availability of many datasets in public repositories, such as the Gene Expression Omnibus (GEO) (Edgar et al., 2002; Barrett et al., 2012) and ArrayExpress (Kolesnikov et al., 2015), by the existence of well-established strategies for data analysis and experimental design, and by the low cost compared with the next-generation sequencing technologies. However, given that microarray analysis is still facing reliability and reproducibility problems, more robust and rigorous methods are needed to account for the high variability and biases introduced in all steps of a microarray experiment.

Several preprocessing and normalization procedures have been proposed to remove biases due to the inhomogeneity of the background and the different fluorescence properties of the dyes. However, biases introduced in the image analysis step, which includes spot segmentation and signal extraction, have not received the same attention, and those may partially explain the existing reliability and reproducibility problems in omics studies. Particularly, several factors, including image resolution, scanner settings, effectiveness of the segmentation algorithm, and unexpected behaviors during hybridization, may lead to errors in spot localization and classification of the pixels (as foreground or background, depending on whether it is situated within or around the spot). Thus, spot intensities are usually noisy and that high pixel–level variability leads to uncertainty in microarray quantification and correlates with variability between replicate spots on duplicate slides (Brown et al., 2001).

Given that even state-of-art image processing tools are susceptible to errors that significantly influence the variability of the data derived from microarray images (Ahmed et al., 2004), new segmentation and intensity extraction algorithms are still being developed in order to improve precision in spot quantification (Li et al., 2017; Karthik and Manjunath, 2018; Shao et al., 2019). Usually, these tools combine sophisticated algorithms and pixel-level analyses in order to obtain an accurate estimate of the signal intensity in each spot. However, to allow subsequent analyses to take into account possible errors and uncertainties arising from the image processing, the method output usually includes not only statistical measures of location (e.g., mean and median) of the foreground and background intensities of each channel of each spot but also measures of dispersion, including standard deviation and covariance between both channels.

Despite the common use of pixel-level variability measures as data quality criteria for filtering purpose, the conventional microarray analysis is solely based on statistical measures of location of the spot intensities (Yang et al., 2002; Sun et al., 2011; Brady and Vermeesch, 2012). To improve robustness and reliability in microarray analysis, pixel-level uncertainties should be accounted for in the intensity log-ratio estimation and propagated to the next steps of the analysis.

Pixel-level uncertainties have been taken into account by many spot quantification algorithms in the literature, but requiring all pixel values to be available. Some of them are interested in improving the log-ratio estimator. Particularly, the method proposed by (Dodd et al., 2004) is a log-ratio estimator that corrects for signal saturation by regressing all pixel intensities at both test and control channels using a censored regression model. The META algorithm (Chan and Chang, 2009) estimates the intensity log-ratio by grouping the pixels according to their distance to the center of the spot and then weighting the log-ratio of each group in inverse proportion to its sample variance. A method that only uses pixel-level mean and variance summary statistics is the hierarchical maximum-likelihood estimator (Bakewell and Wit, 2005). However, it is not exactly based on the standard log-ratio representation of the spot intensity. It models the gene expression signal at control and treatment channels separately, incorporating the sample within-spot deviation and then performs the estimation using maximum likelihood. To the best of our knowledge, there is no intensity log-ratio estimator to be used after the image analysis phase (i.e., based solely on the pixel-level summary statistics) that takes into account pixel-level uncertainties.

The first contribution of this paper is a more robust estimator for the intensity log-ratio (M) and average log intensity (A) of a microarray spot that accounts for pixel-level variance and covariance between channels. For a spot t, these values are denoted by M_t and A_t, respectively (Dudoit et al., 2002). We derive these estimators by using the multivariate delta method (Casella and Berger, 1990). Specifically, we approximate the expected values of M_t and A_t by using their second-order Taylor's expansions, and the variance of M_t and A_t by using their first-order Taylor's expansions. These expansions depend on the pixel-level variance and covariance between channels of the spot, whose sample estimates are readily accessible through standard output files of microarray image analysis tools.

After spot intensity estimation, it is necessary to perform a within-slide normalization to remove array-specific effects, intensity-dependent dye biases, and other systematic trends of the microarray data. The within-slide normalization based on the robust locally weighted regression (LOWESS) (Cleveland, 1979) is one of the most used techniques. The choice of the LOWESS parameters, particularly the smoothing parameter (also known as neighborhood size or bandwidth), dramatically affects the intensity and quality of the microarray data calibration. Although the smoothing parameter is still commonly set arbitrarily (around 0.2 and 0.4) (Dudoit et al., 2002; Smyth and Speed, 2003; Drăghici, 2012), some data-driven methods have been proposed to select its optimal value (Berger et al., 2004; Futschik and Crompton, 2004a; Lee et al., 2008). All these methods are similar in that they choose the smoothing parameter by minimizing a measure of error of the LOWESS fit. Berger et al. (2004) use the mean-squared difference between the LOWESS estimates and the corresponding normalization reference levels as cost function. These normalization levels are the true spot-specific calibration errors, which are usually unknown. Thus, Berger et al. suggest to estimate them from control transcripts and replicate slides. However, they are not always available for all genes in a typical microarray experiment, making it hard to reliably use the method. Futschik and Crompton's selection method, named OLIN (Futschik and Crompton, 2004a; Futschik and Crompton, 2004b), has the advantage of not relying on a reference level. Its optimization procedures use the generalized cross-validation (GCV) criterion, an estimator of the prediction mean square error (PMSE), as cost function. Lee et al. (2008) proposes to select the smoothing parameter by minimizing the bootstrap estimate of the mean integrated square error (MISE) and show that their results are comparable to OLIN.

Although all these methods have shown superiority over LOWESS normalization with a fixed arbitrarily chosen smoothing parameter, they lack in taking into account any heteroskedasticity in the data. In addition, they usually suffer from a poor bias–variance trade-off, tending to choose small smoothing values, which yield unnecessarily complicated (with high variance) LOWESS fits.

The second contribution of this paper is a data-driven method for selecting the smoothing parameter of the LOWESS normalization process. Inspired by the previous proposed methods, we choose the optimal smoothing value by minimizing a mean squared error criterion. However, our selection method also takes into account heteroskedasticity of the microarray data and offers a better bias–variance trade-off by selecting from among the low-MSE fits the one that is the most parsimonious. The parameter selection is obtained by solving a discrete optimization problem and is based on conventionally accepted ideas for analysis of M-plots—a graphical tool showing the curve of the MSE against the effective degrees of freedom of the estimate (Cleveland et al., 1988).

Given that the primary application of DNA microarrays has been to measure gene expression levels, we focus in this paper on variation-preserving estimation and normalization methods for gene expression levels from two-channel (or two-color) microarrays. However, it is straightforward to adapt the same ideas to improve analysis of other types of microarray data, even from single-channel technologies.

The proposed methods were evaluated by a differential gene expression analysis from real intestinal metaplasia and normal microarray samples. The proposed estimators for the M_t and A_t values were compared with the conventional estimators that neglect the pixel-level variability. In addition, we compared the proposed method for selecting the LOWESS smoothing parameter with OLIN, as it is conceptually similar to the other existing methods and can be applied even to microarray experiments with few or no replicates. Results show that a more robust and conservative analysis is performed when the LOWESS smoothing parameter is selected by our method, potentially reducing the number of false-positive differential expressions. Besides, both the pixel-level variabilities incorporated by the proposed estimators for the M_t and A_t values and the variability preserved by our more parsimonious normalization method contributed to the identification of new differentially expressed genes. Thus, the proposed methods may also reduce the false-negative rate.

MATERIALS AND METHODS

Two procedures that critically affect the adequacy of microarray data analysis are the spot quantification, which extracts summarized quantitative measures of the pixel intensities within each spot of the microarray slide, and the within-slide normalization, which removes dye-specific biases and other systematic noises simultaneously from all logged spot intensities (M_t and A_t values).

In the section Intestinal Metaplasia Database, we describe a gene expression dataset used to illustrate the application of our proposed methods. In the section Improved Estimators for the M_t and A_t values, we show our improved estimation method for the M_t and A_t values that incorporates pixel-level variability. In the section Estimators for the Variances of the M_t and A_t Values, we discuss some criteria that can be used for proper setting of the parameters of the LOWESS within-slide normalization and we propose an algorithm for selecting the optimal value for the smoothing parameter (denoted by f).

Intestinal Metaplasia Database

Due to a chronic inflammatory process, the normal squamous mucosa of the stomach may be replaced by columnar intestinal-type epithelium, characterizing a disease called intestinal metaplasia of the stomach. Since adenocarcinoma of the stomach and inflamed intestinal mucosa are strongly associated (Coussens and Werb, 2002), intestinal metaplasia may be a significant risk factor for gastric cancer.

We analyzed data from a two-color microarray experiment with tissues samples from 90 different subjects, being 35 from tissues representing type II intestinal metaplasia and 55 from tissues representing the normal condition, obtained from the Tumor Bank at A.C. Camargo Cancer Center/Antonio Prudente Foundation.

It was used the standard reference design (Churchill, 2002), in which each sample is hybridized against a pool of normal tissues using the same orientation of dye labeling. Gene expression levels were measured on Agilent Whole Human Genome Microarrays 4x44K G4112F (design ID 014850), each slide containing 41,093 unique probes. The scanned images of the microarray slides were processed by *Agilent Feature Extraction* software, version 9.5, where statistics (mean, standard deviation, and covariance) of the foreground and local background pixels were computed for each spot, in both test and reference channels. Each microarray spot contains about 60 foreground pixels.

This study was carried out in accordance with the recommendations of the international guidelines for investigations involving human beings with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Institutional Committee of the A.C. Camargo Cancer Center (process number 1023/07).

Improved Estimators for the M_t and A_t Values

Usually, in microarray analysis, the test channel is denoted by (red), and the reference channel is denoted by G (green), following this usual notation, denoted by R_{tj} and by G_{tj}, the intensity value of the jth pixel within the th spot, respectively, in the test and reference channel. The relative expression of pixel j within spot is denoted by M_{tj} and defined as follows:

$$M_{tj} \doteq \log_2\left(\frac{R_{tj}}{G_{tj}}\right) = \log_2(R_{tj}) - \log_2(G_{tj}). \tag{1}$$

The average expression of pixel within spot is denoted by A_{tj} and defined as follows:

$$A_{tj} \doteq \frac{1}{2}\left(R_{tj}G_{tj}\right) = \frac{\log_2(R_{tj}) + \log_2(G_{tj})}{2}. \tag{2}$$

Usually, image analysis software does not provide all pixel intensity values within each spot. Nonetheless, it provides several descriptive statistics of the foreground and background pixel intensities, including sample estimates for the mean, median, variance, and covariance between the two channels.

To incorporate the pixel-level variability in the analysis, we derived an approximation of the expected values of M_{tj} and A_{tj} by using the *multivariate delta method* (Casella and Berger, 1990). Assuming that the functions (1) and (2) are twice differentiable on an open interval which contains the point $\left(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj})\right)$, we computed their second-order Taylor's expansions, around the point $\left(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj})\right)$, and then derived their expected values. The derivation is presented in Appendix 4.

It is reasonable to assume that the variables R_{tj}, G_{tj}, M_{tj} and A_{tj} have a distribution with well-defined mean and variance. Particularly, Hoyle et al. (Hoyle et al., 2002) empirically showed that the distribution of the pixels within a spot is heavy-tailed (a non-Gaussian distribution) and well-approximated by a log-normal distribution. Consequently, M_{tj} and A_{tj} follow a distribution which is well-approximated by a Gaussian distribution and all the variables have at least the first and second moments finite.

Let $\bar{R}_{tc}$ and $\bar{G}_{tc}$ be non-zero estimates of, respectively, $\mathbb{E}(R_{tj})$ and $\mathbb{E}(G_{tj})$, which represent average foreground signals after correction for removing the background influence. The subscript indicates dependence on the background correction. Also, let $\hat{\sigma}^2(R_t)$ and $\hat{\sigma}^2(G_t)$ be estimates of, respectively, Var (R_{tj}) and Var (G_{tj}), which are assumed to be independent of the background correction. Note that mean and variance estimates are calculated across observed foreground pixel intensities within the spot at the respective channel.

We can derive improved estimators for $\mathbb{E}(M_{tj})$ and $\mathbb{E}(A_{tj})$ as follows:

$$\tilde{M}_t \doteq \mathbb{E}(M_{tj}) \approx \log_2(\bar{R}_{tc}) - \log_2(\bar{G}_{tc}) + \frac{1}{2ln\,(2)}\left(-\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2} + \frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}\right), \tag{3}$$

$$\tilde{A}_t \doteq \mathbb{E}(A_{tj}) \approx \frac{1}{2}\left(\log_2(\bar{R}_{tc}) + \log_2(\bar{G}_{tc})\right) - \frac{1}{4ln\,(2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2} + \frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}\right). \tag{4}$$

Note that the conventional estimators for the M_{tj} and A_{tj} values, given by

$$\hat{M}_t \doteq \log_2(\bar{R}_{tc}) - \log_2(\bar{G}_{tc}), \tag{5}$$

$$\hat{A}_t \doteq \frac{\log_2(\bar{R}_{tc}) + \log_2(\bar{G}_{tc})}{2}, \tag{6}$$

are approximations of, respectively, $\mathbb{E}(M_{tj})$ and $\mathbb{E}(A_{tj})$ derived from only the zeroth-order Taylor's expansion of the functions that define M_{tj} and A_{tj}. Thus, the conventional estimators ignore the known measures of pixel-variability, which represent uncertainties in the gene expression measurements.

Figure 1 illustrates the differences between the estimators for the $\mathbb{E}(M_{tj})$ and $\mathbb{E}(A_{tj})$ for a randomly chosen microarray slide of the database described in the section *Intestinal Metaplasia Database*. Since these estimators may suffer from numerical instability if the corrected foreground signals, $\bar{R}_{tc}$ and $\bar{G}_{tc}$, are very close to zero, we removed the background influence by applying the *normexp* method (Ritchie et al., 2007) with offset equals to 50. The top 20 spots with the highest pixel-level variability are highlighted in red plus symbols. Several of these spots have low average intensity (small estimates for $\mathbb{E}(A_{tj})$) and a small difference between the intensities of the two channels (estimates for $\mathbb{E}(M_{tj})$ close to zero), but they are not the majority. The differences between the proposed estimators, defined in Eq. (3) and (4), and the conventional estimators, defined in Eq. (5) and (6), are shown in **Figures 1C**, **D**. These differences are due to the distinct parts between their respective formulas. When computing the $\tilde{M}_j$ estimates, the ratio of the pixel-level variability to the squared expected value in the test channel appears in Eq. (3) with an opposite sign to the same term in the reference channel. Thus, positive and negative differences between the estimates for $\mathbb{E}(M_{tj})$ may occur if such terms do not cancel each other out. **Figure 1C** shows the *ilde* $\tilde{M}_t$ estimates were smaller than the $\hat{M}_t$ estimates for the genes with highest pixel-level variance, indicating a larger variance in their test channels. **Figure 1D** shows some $\tilde{A}_t$ estimates were smaller than

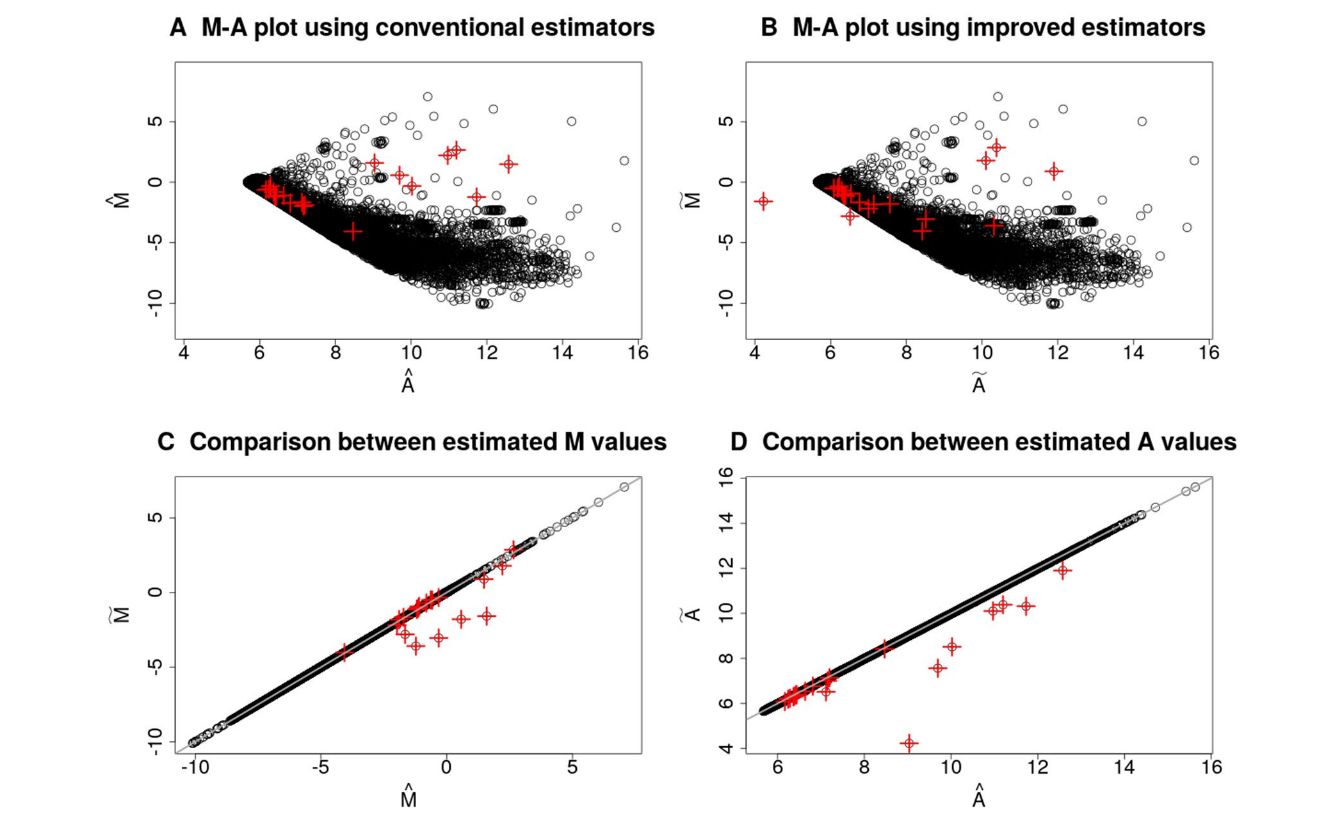


FIGURE 1 | Comparison between conventional and proposed estimation methods for the M_t and A_t values for the microarray slide with ID 251485069395_1.4. The M-A plots in **(A)** and **(B)** were obtained by using, respectively, the conventional and improved estimators for the M_t and A_t values. Plots **(C)** and **(D)** show the conventional against the improved estimates for, respectively, the M_t and A_t values. Top 20 genes with the highest pixel-level variance are highlighted in red plus symbols. The test channel contains RNA samples of normal gastric mucosa, and the control channel contains samples from a common reference. The background influence was removed from the foreground signals by the *normexp* method with offset.

the $\hat{A}_t$ estimates. The reduction is explained by the fact that the additional terms in Eq. (4) are negative for any positive pixel-level variability in any channel.

Estimators for the Variances of the M_t and A_t Values

Since we have also available the sample covariance between R_{tj} and G_{tj}, denoted by $\hat{\sigma}(R_t, G_t)$, we applied the multivariate delta method for deriving estimators for the variances of the M_{tj} and A_{tj}. We calculated the variance of the first order Taylor's expansion of the functions (1) and (2) that define, respectively, M_{tj} and A_{tj}, as shown in Appendix 5. The variance estimators for M_{tj} and A_{tj}, for pixels j within spot t are:

$$\hat{\sigma}^2(M_t) \doteq \frac{1}{\ln^2(2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}-2\frac{\hat{\sigma}(R_t,G_t)}{\bar{R}_{tc}\bar{G}_{tc}}\right), \qquad (7)$$

$$\hat{\sigma}^2(A_t) \doteq \frac{1}{4\ln^2(2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}-2\frac{\hat{\sigma}(R_t,G_t)}{\bar{R}_{tc}\bar{G}_{tc}}\right). \qquad (8)$$

The variances of M_{tj} and represent pixel-level uncertainties of the th spot. They can be used, for instance, for assessing the quality of the th spot or for constructing confidence intervals for the parameters $\mathbb{E}(M_{tj})$ and $\mathbb{E}(A_{tj})$.

Optimal Selection of the LOWESS Parameters

To simplify the notation, we will denote the estimates for $\mathbb{E}(M_{tj})$ and $\mathbb{E}(A_{tj})$, independently of the estimation method used, by, respectively, M_t and A_t values.

It is necessary to remove from these M_{tj} intensity values the dependent dye-specific biases and other systematic errors by using some within-slide normalization method.

In the LOWESS within-slide normalization method, one estimates for each microarray slide a smoothing function $\hat{\mu}$ that maps each A_t observed value to a smoothed M_t value, $\hat{\mu}(A_t)$. Since $\hat{\mu}(A_t)$ is considered an estimate of a dye-dependent bias, it must be subtracted from the corresponding observed M_t value to obtain a residual value representing, presumably, the biologically relevant gene expression level.

An appropriate LOWESS estimation depends on the choice of its parameters. According to loader (Loader, 1999), the weight function and the number of iterations of the robustness algorithm are not critical parameters. Cleveland (Cleveland, 1979) comments that good choices for these parameters are, respectively, the tricube function and three iterations. However, the degree of the local polynomials and the smoothing parameter *f*, which, in the nearest neighbor method, is a number between and indicating the proportion of data used in each local fit, affects the bias and the variance of the fit.

Specifically, the higher the degree of the local polynomial (related to the complexity of the model), the lower the bias of the fit (probably, fitting the data very well). However, the additional parameters of this more complex model increase the variance of the fitted values, yielding a poor generalization ability (i.e., the model will have a large error). Thus, to avoid unstable LOWESS estimates, several references as (Loader, 1999; Yang et al., 2001; Dudoit et al., 2002; Smyth and Speed, 2003) recommend using local polynomials of degree one, mainly in the presence of sparsity, as is the case of microarray data.

The effects of the smoothing parameter *f* on the bias and variance of the fit are opposite to those of the degree of the local polynomials. Since the *f* parameter indicates the number of observations that will be used in the local polynomial estimation, when *f* value is large, a simple polynomial may not fit well to all observations in the neighborhood, distorting or ignoring essential features. In other words, the estimation of the smoothing function can be significantly biased. On the other hand, when a low *f* value is chosen, the number of observations may be insufficient to capture the general behavior of the data, resulting in a very noisy (large variance) fitness function.

In the next section, we propose a method for selecting a value for the *f* parameter, focusing on microarray data normalization. Our method takes into account the intrinsic characteristics of the bias and variance of the fit as well as of gene expression data.

Lowess Smoothing Parameter Selection

For microarray data normalization, the ideal LOWESS fitted curve captures only trends and effects from systematic errors, retaining all biological variation. However, it critically depends on the choice of the *f* parameter value.

Figure 2 illustrates the MA plot of the microarray slide shown in **Figure 1B**, with different LOWESS fits yielded by *f* values varying from 0.05 to 0.9. The improved estimation method was used to obtain the M_t and A_t values, that is, the $\tilde{M}_t$ and $\tilde{A}_t$ estimates.

The quality of a LOWESS estimator can be assessed by the MSE, which measures how close the estimator $\hat{\mu}$ is of the true mean function μ :

$$MSE(\hat{\mu}) = \mathbb{E}[(\mu - \hat{\mu})^2].$$

Since the real curve μ is unknown, we need a criterion to evaluate the MSE. Under the assumption of heteroskedasticity, Cleveland and Devlin (Cleveland and Devlin, 1988) propose the Mallows' Cp criterion for local fitting that can be used as as MSE estimator. In the presence of heteroskedasticity, as usual for microarray data, the heteroskedasticity-robust Cp (HRCp) criterion, proposed by Liu and Okui (Liu and Okui, 2013), may be a more appropriate MSE estimator. We detail this MSE estimator next.

Considering $\{(A_t, M_t)\}_{t=1}^{T}$ within-slide data points, the evaluation of the LOWESS smoothing function on any point is given by a linear combination of the observed points, whose weights $\{(l_t(A)\}_{t=1}^{T}$ are assigned according to the distance of A to the A_t observed points:

$$\hat{\mu}(A) = \sum_{t=1}^{T} l_t(A) M_t.$$

Consider the $T \times T$ matrix $\boldsymbol{L}$ which maps the observed to the fitted values:

$$\begin{pmatrix} \hat{\mu}(A_1) \\ \vdots \\ \hat{\mu}(A_T) \end{pmatrix} = \boldsymbol{L}M = \begin{pmatrix} l_1(A_1) \ldots l_T(A_1) \\ \vdots \\ l_1(A_T) \ldots l_T(A_T) \end{pmatrix} \begin{pmatrix} M_1 \\ \vdots \\ M_T \end{pmatrix}.$$

Two commons definitions of the effective degrees of freedom of $\hat{\mu}$ are: (1) $v_1 \doteq \text{tr}\ (\boldsymbol{L})$ and (2) $v_2 \doteq \text{tr}\ (\boldsymbol{L'L})$, where tr stands for the trace operator.

Supposing that the variance of M_t, across T spots of a microarray slide, is constant and equals to σ^2, the Mallows' Cp for local fitting is defined as:

$$Cp(\hat{\mu}) = \frac{1}{\sigma^2} \sum_{t=1}^{T} (M_t - \hat{\mu}(A_t))^2 - T + 2v_1.$$

Cleveland et al. (1988) shows that σ^2 can be estimated as follows:

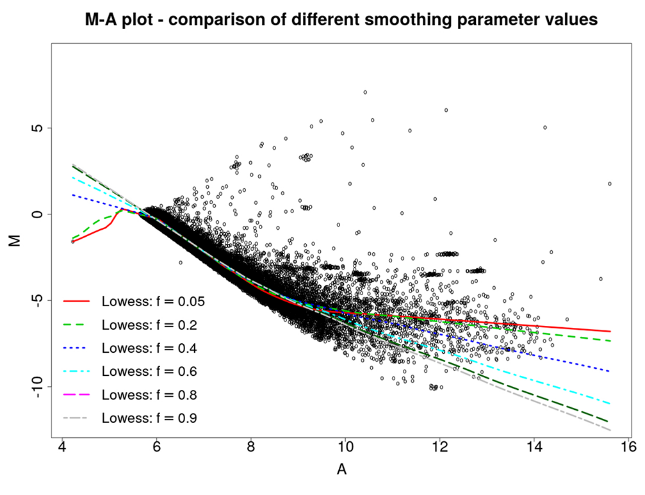


FIGURE 2 | MA plot for the slide 251485069395_1.4, with M_t and A_t values estimated by the proposed method and LOWESS fits yielded by *f* values ranging from 0.05 to 0.9.

$$\hat{\sigma}^2 \doteq \frac{\Sigma_{t=1}^{T}[M_t - \hat{\mu}(A_t)]^2}{n+\nu_2-2\nu_1}.$$

When heteroskedasticity is present, Mallows' Cp criterion is not an appropriate MSE estimator. Considering the $T \times T$ diagonal matrix Σ, whose th diagonal element is given by a non-homogeneous variance σ_t^2 of M_t, a robust MSE estimation can be achieved by using the HRCp criterion, defined as:

$$HRCp(\hat{\mu}) = \sum_{t=1}^{T}(M_t - \hat{\mu}(A_t))^2 + 2\,\mathrm{tr}\,(\Sigma \boldsymbol{L}).$$

According to Loader (1999), σ_t^2 can be estimated locally by calculating the error variance (the residual sum of squares divided by the corresponding degrees of freedom) of a nearly unbiased LOWESS fit, which can be yielded using a very small value for the smoothing parameter (e.g., f = 0.1. Since the local variance estimates can be very noisy, it may be appropriate to smooth them using a gamma kernel.

Several authors suggest to choose the *f* value which minimizes a measure of error of the LOWESS fit, such as the MSE criterion (Berger et al., 2004; Futschik and Crompton, 2004a; Lee et al., 2008). However, other authors (Mallows, 1973; Cleveland and Devlin, 1988; Loader, 1999) argue that a selection based only on minimizing the MSE criterion is a poor procedure since it ignores the intrinsic information of the bias and variance of the fit. Therefore, following their suggestion, we propose a method based on a graphical tool called M-plot. It is a graph of the MSE estimate as a function of the effective degrees of freedom of the fit.

M-plots illustrating the *f* parameter selection method for a typical microarray slide (ID 251485069395_1.4) are shown in **Figure 3**. Dots show MSE estimates (by HRCp criterion) and respective degrees of freedom (by ν_2 definition) of LOWESS fits (on the $\hat{M}_t$ and $\hat{A}_t$ estimates, in the first M-plot, and on the $\tilde{M}_t$ and $\tilde{A}_t$, in the second M-plot) obtained with *f* parameter varying from to 0.2 We fixed the other LOWESS parameters (local polynomials of degree one, tricube weight function, and three iterations) so that the M-plot curve shows only the effect of the *f* parameter on the bias–variance compromise. Large *f* values tend to yield simple fits (with fewer degrees of freedom), which have a small variance, but a large bias. On the other hand, minimal *f* values tend to yield complex fits (with many degrees of freedom), which have a small bias, but a large variance.

For the microarray slide in **Figure 3**, a selection method based only on the minimization of the MSE curve would choose the smallest evaluated *f* value (0.2). However, any *f* value within the flattening region near to the minimum (the region with light-colored dots) is a good choice, in the sense that it yields a low-MSE fit (Cleveland and Devlin, 1988; Loader, 1999). Depending on the type of application, we can choose between one value which yields a low-bias fit (with more degrees of freedom) or a low-variance fit (with fewer degrees of freedom). Since we want to estimate a natural phenomenon behavior, we propose to select from the flattening region the *f* value which yields the simplest LOWESS fit (the one with fewest effective degrees of freedom). The biggest dot in each M-plot indicates the selected *f* value. The detection of the flattening region is made by searching points for which the derivative of the MSE curve is small. We check for each sequence of three points near the minimum whether the difference between the MSE values

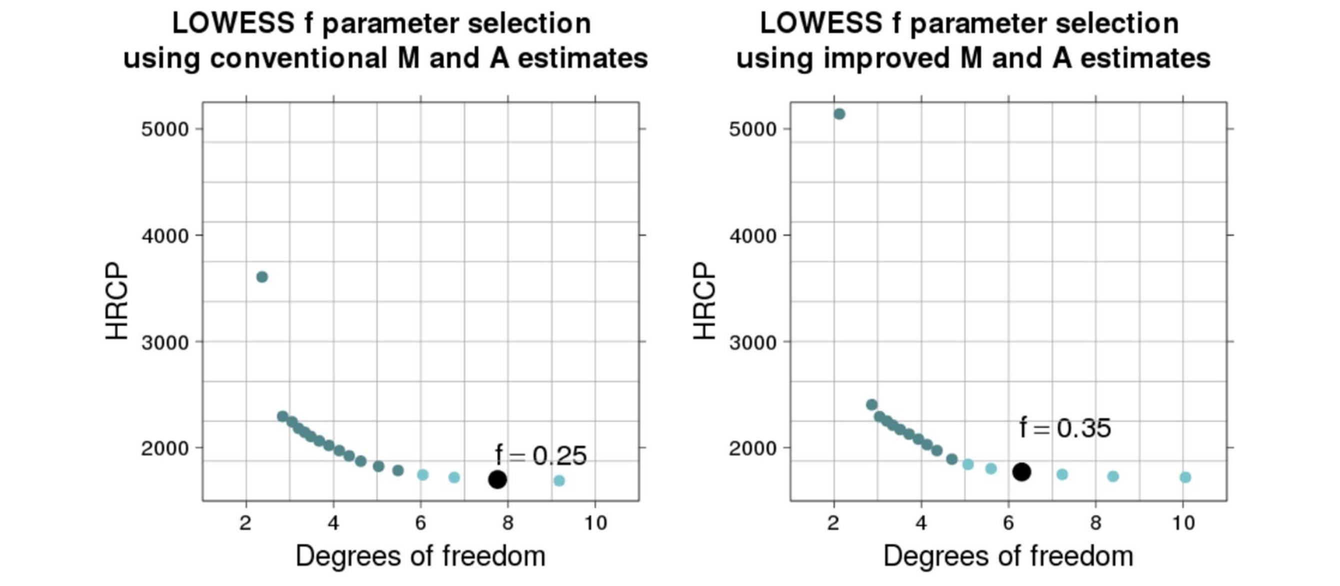


FIGURE 3 | Selection of the LOWESS *f* parameter by using HRCp criterion. The M-plots illustrate the selection process for a particular microarray slide (ID: 251485069395_1.4). The flattening region is represented by the light-colored dots and the selected *f* value by the biggest dot. The LOWESS fits were yielded using values ranging from 1 to 0.2 (from lowest to highest degree of freedom).

is small. If so, these points are considered as belonging to the flattening region.

The f parameter selection method can be summarized in the following discrete and constrained optimization problem. Consider a sequence of l different values for f, $\{f_1, f_2, \ldots , f_l\}$, and denoted by $\hat{\mu}_{f_k}$, the LOWESS fit yielded by using the value f_k for the f parameter. Also, let:

$$\begin{aligned}
&\mathcal{F} = \{\hat{\mu}_{f_k}; f_k \in \{f_1, f_2, \ldots, f_l\},\ f_{k+1} < f_k,\ \mathit{for}\ k = 1, \ldots, l-1\};\\
&f_{min} = \underset{f_k}{\arg\min}\ \mathit{HRCp}\ (\hat{\mu}_{f_k}), \mathit{such\ that}\ \hat{\mu}_{f_k} \in \mathcal{F};\\
&f_{max} = \underset{f_k}{\arg\max}\ \mathit{HRCp}\ (\hat{\mu}_{f_k}), \mathit{such\ that}\ \hat{\mu}_{f_k} \in \mathcal{F}; \mathit{and}\\
&\Delta_{MSE} = 0.05(\mathit{HRCp}\ (\hat{\mu}_{f_{\max}}) - \mathit{HRCp}\ (\hat{\mu}_{f_{\min}})).
\end{aligned}$$

Since v_2 function provides the effective degrees of freedom of a given fit, the selected f value is the solution f^*, if it exists, of the following problem:

$$\begin{aligned}
&f^\star \doteq \underset{f_k}{\arg\min}\ v_2(\hat{\mu}_{f_k})\\
&\text{subject to}:\\
&\hat{\mu}_{f_k} \in \mathcal{F};\\
&\text{HRCp}\,(\hat{\mu}_{f_k}) \ \le\ \text{HRCp}\,(\hat{\mu}_{f_{min}}) \ +\ \Delta_{MSE}, \text{for } k = 1, 2;\\
&\text{HRCp}\,(\hat{\mu}_{f_{k-2}}) \ \le\ \text{HRCp}\,(\hat{\mu}_{f_{min}}) \ +\ \Delta_{MSE}, \text{for } k = 3, \ldots, l;\\
&|\,\text{HRCp}\,(\hat{\mu}_{f_k}) \ -\ \text{HRCp}\,(\hat{\mu}_{f_{k-1}})| \ <\ \Delta_{MSE}, \text{for } k = 2, \ldots, l; \text{and}\\
&|\,\text{HRCp}\,(\hat{\mu}_{f_k}) \ -\ \text{HRCp}\,(\hat{\mu}_{f_{k-2}})| \ <\ \Delta_{MSE}, \text{for } k = 3, \ldots, l.
\end{aligned}$$

If the minimum of the M-plot curve is far away of the point corresponding to the second lowest MSE estimate, the previous problem has no solution. In that case, the f value that yields the fit with lowest MSE estimate is selected. Specifically, the f parameter value is selected by solving the following problem:

$$f^\star \doteq \underset{f_k}{\arg\min}\ \text{HRCp}\,(\hat{\mu}_{f_k}), \text{such that } \hat{\mu}_{f_k} \in \mathcal{F}.$$

APPLICATION ON INTESTINAL METAPLASIA DATA

To investigate the effects of the proposed methods, we preprocessed the data described in the section *Intestinal Metaplasia Database* by using all discussed methods and compared the identified differentially expressed genes.

First, we applied the *normexp* method with offset value of for removing the background influence. Then, we compute the M_t and A_t values both by the conventional estimation methods, defined in Eq. (5) and (6), and by the proposed estimation methods, defined in Eq. (3) and (4). The LOWESS within-slide normalization was carried out as discussed in the section *Optimal Selection of the LOWESS Parameters*. For comparison purpose, the f smoothing parameter was selected both by the OLIN method (considered by us as a conventional approach) and by the proposed method, discussed in the section *LOWESS Smoothing Parameter Selection*. Since data from all microarray slides present overdispersion, we used the HRCp criterion as cost function of our selection method.

Therefore, the following four preprocessing procedures were applied separately to the original data:

1. Conventional estimation for M_t and A_t and LOWESS within-slide normalization using f parameter selected by OLIN;
2. Improved estimation of M_t and A_t and LOWESS within-slide normalization using f parameter selected by OLIN;
3. Conventional estimation of M_t and A_t and LOWESS within-slide normalization using f parameter selected by the proposed method;
4. Improved estimation of M_t and A_t and LOWESS within-slide normalization using parameter selected by the proposed method.

Figure 4 shows the distribution of the optimal values for the LOWESS f parameter, according to the proposed selection method with HRCp criterion, for the entire database, separated by normal and intestinal metaplasia conditions (both, hybridized against a pool of normal tissues). In the first plot, the LOWESS curve was fitted on the $\hat{M}_t$ and $\hat{A}_t$ estimates and, in the second plot, on the $\tilde{M}_t$ and $\tilde{A}_t$ estimates. The average of the selected f values was close to 0.5.

As expected from a method that neither takes into account heteroskedasticity of the data nor attempts to make a good balance between bias and variance, the OLIN method selected the smallest evaluated value (0.2) for most of the slides. Same results were obtained when the M_t and A_t values were estimated by the conventional and by the proposed estimator. Such behavior has been reported in the literature, implying that the optimal f values according OLIN are usually close to the default one (Chiogna et al., 2009).

After preprocessing the data, a two-sample t-test assuming unequal variance was performed for each spotted gene to determine whether its expression is statistically different between gastric tissues in normal and intestinal metaplasia groups. However, since we are interested in directly assessing the impact of each proposed method on the t-statistics and p-values rather than making inference about differential expression, the comparative study was performed before applying a multiple testing correction.

Comparison of the Results

Results of a pairwise comparison among the p-values and t-statistics obtained by the four preprocessing methods are shown in **Figure 5**. In the left-column plots, we compare the p-values and, in the right-column plots, we show the changes in the difference between the group means (the absolute value of the t-statistic numerator) and in the within-group variability (the denominator of the t-statistic). Only genes with p-value less than 5% were considered.

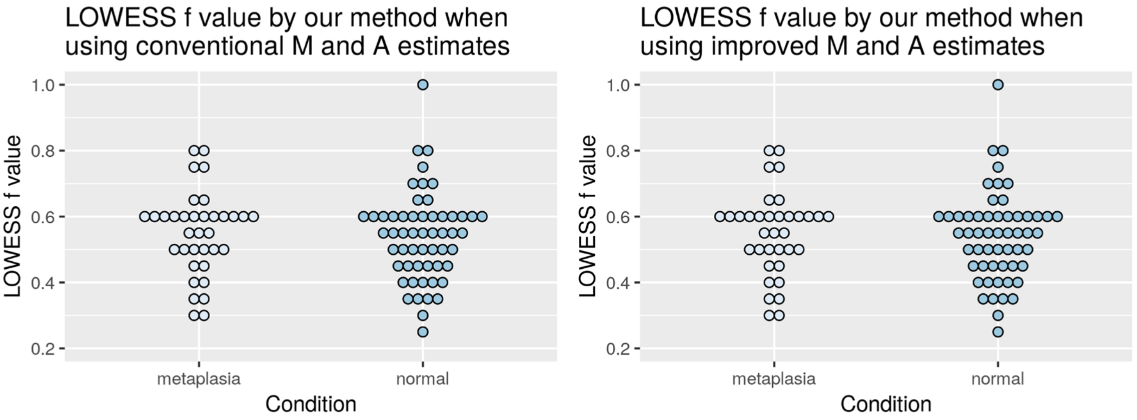


FIGURE 4 | Distribution of the selected *f* values by normal and metaplasia intestinal conditions when the M_t and A_t values are estimated by using the conventional (left) and the proposed (right) method.

The left-column plots show that most of the points are distributed around the 45-degree line. Thus, the p-values and, consequently, the total number of differentially expressed genes, even at a lower significance level, were similar among the four methods.

The first- and second-row plots show how p-values and t-statistics were affected by estimating the M_t and A_t values with the proposed method, which takes into account the pixel-level uncertainties. The genes represented by blue plus signs were identified as differentially expressed only when using the proposed estimator for the M_t and A_t values.

The genes represented by green crosses were identified as differentially expressed only when using the conventional estimator for the M_t and A_t values.

When the LOWESS *f* parameter is selected by OLIN (first-row plots), it is clear that the within-group variability decreases when using the proposed estimators for the M_t and A_t values. When the LOWESS parameter is selected by our method (second-row plots), there is still a reduction in the within-group variability. However, this impact is less clear because of the variability introduced when the LOWESS *f* parameter is selected by our method.

The third- and fourth-row plots compare p-values and t-statistics obtained by OLIN and the proposed approach for selecting the LOWESS *f* parameter. The genes represented by blue plus signs were identified as differentially expressed only when *f* was selected by the proposed method. The genes represented by green crosses were identified as differentially expressed only when selecting *f* by OLIN. It is clear that, for most genes, both within-group variabilities increased, implying that the normalization procedure was more conservative, and thus, more potentially relevant information is retained. In addition, for many genes, the increase in the within-group variability was counterbalanced by an increase in the distance between the groups. Such effect is even most pronounced when the proposed estimator for the M_t and A_t values are used. Thus, their respective p-values reduced enough to consider them as differentially expressed genes.

The diagrams in **Figure 6** show a comparison of the methods with respect to the total number of genes with p-value less than 5%. On the left, the p-values were not corrected for multiple tests, while on the right, the p-values were adjusted by the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

Note that the four methodologies are quite different in terms of which genes were identified as differentially expressed. As a consequence of the more conservative (milder) noise reduction performed in the LOWESS within-slide normalization procedure with *f* parameter selected by our method, fewer genes are identified as differentially expressed. However, regardless of the normalization method, more genes could be identified as differentially expressed when the M_t and A_t values were estimated by the proposed estimation method that incorporates pixel-level variability. Given that both proposed methods make the analysis more robust by incorporating and preserving information neglected by the conventional methods, we can argue that they are contributing to the reduction of both false-positive and false-negative rates.

Validation Analysis

To check the consistency of our analysis, we compared our results with those reported in the literature. Out of the genes which are associated with intestinal metaplasia according to the Gene Expression Omnibus platform (Edgar et al., 2002) of the NCBI (National Center for Biotechnology Information) website, 80 spotted genes (corresponding to 63 unique genes) have p-value (before FDR correction) less than 5%, and 35 spotted genes (corresponding to 29 unique genes) have p-value (after FDR correction) less than 5%. These findings are summarized respectively in **Tables 1**, **2**. In addition, **Figure 7** compares the total number of validated genes identified by each method with p-value less than 5% (before FDR correction).

Greater differences in inference were observed among the genes whose p-value is close to the significance level. These

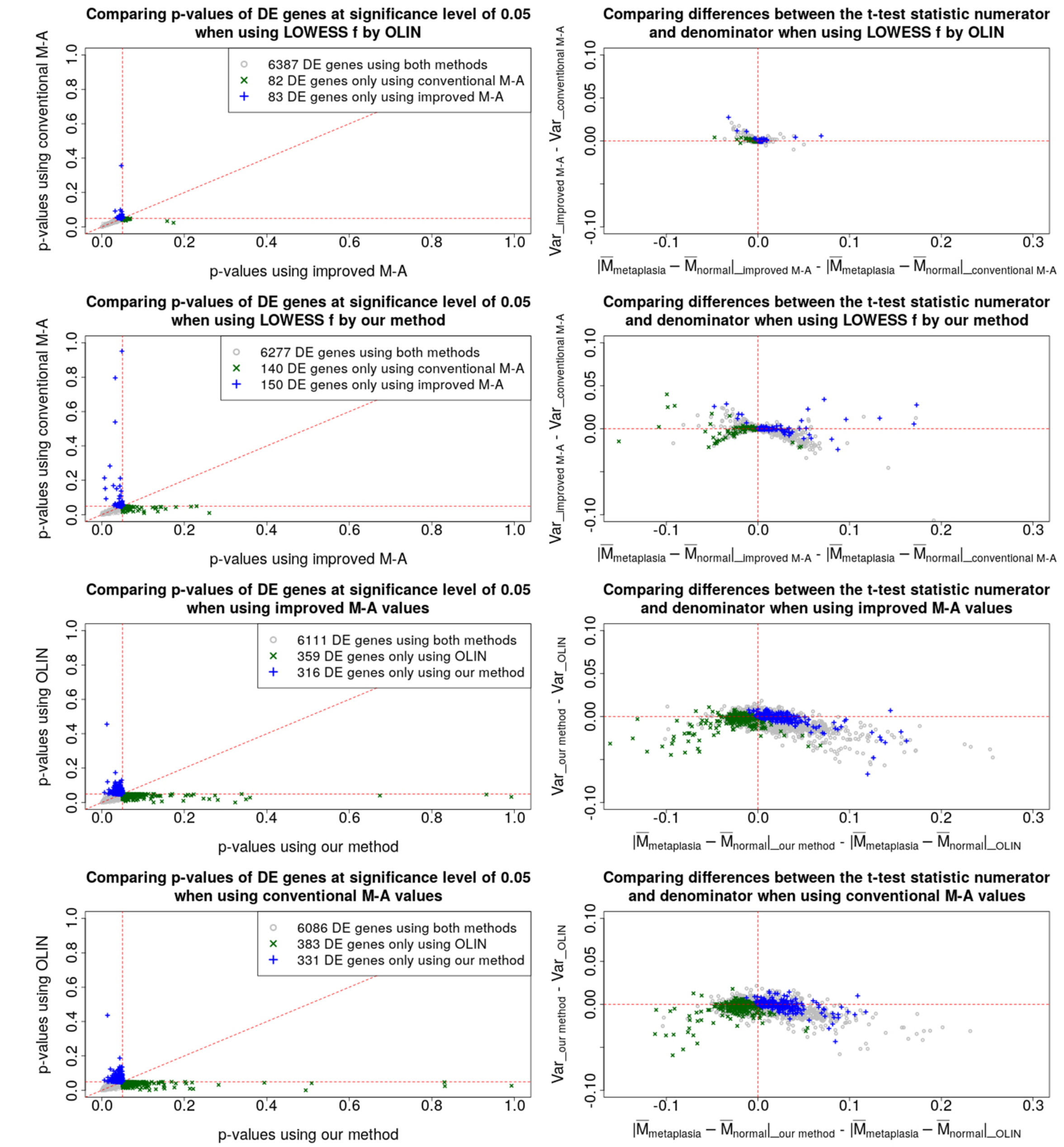


FIGURE 5 | Pairwise comparison between the proposed and the conventional methods. Left-column plots compare the FDR-corrected p-values, and the right-column plots compare the difference between the absolute values of the numerators with the difference between the denominators of the t-test statistic.

genes have a more subtle differential expression, which can be easily damaged by measurement errors and poor estimation and normalization methods. Thus, the more accurate and careful analysis provided by the proposed methods is especially important for making decisions on the differential expression of these more sensitive genes.

Two replicates of the HSPB1 gene could not be identified as differentially expressed when using both the conventional estimators for the M_t and A_t values and our selection method for the LOWESS *f* parameter. Thus, the estimation of the M_t and A_t values by the proposed estimators was crucial in determining the differential expression of the HSPB1 gene.

The genes PTEN, CTNNB1, MLH1, CXCR4, and CXCR1 could only be identified as differentially expressed when the LOWESS parameter was selected by our proposed method. Particularly, the gene CXCR4 only was determined as differentially expressed when the improved estimators for the M_t and A_t values were also used. In contrast, the gene KRT14 was no longer identified as differentially expressed when the LOWESS *f* parameter was selected by our proposed method.

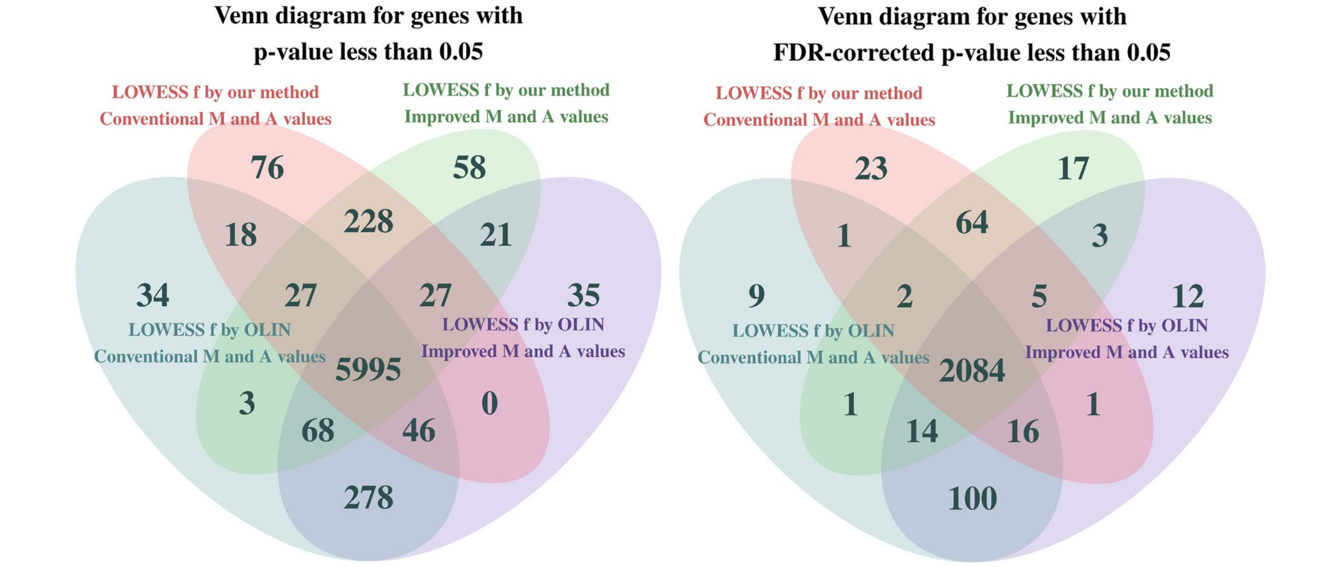


FIGURE 6 | Venn diagram illustrating the total number of differentially expressed genes identified in each variant of the database at a significance level of 5%. On the left, p-values were not corrected for multiple tests, while on the right, p-values were adjusted by the false discovery rate (FDR) correction.

In the following, we briefly describe the association of those genes with intestinal metaplasia of the stomach according to the literature data:

- HSPB1 (heat-shock protein beta-1, also known as HSP27—heat-shock protein 27): It has a protective role against stress-induced cell damage, and its expression has been considered critical for mucosal protection in the stomach (Ebert et al., 2005). Also, it has been reported as down-regulated in esophageal adenocarcinoma (Lv et al., 2019).
- PTEN (phosphatase and tensin homolog): It has been identified as overexpressed in intestinal metaplasia and is a known marker for tumorigenesis and progression of gastric carcinoma (Yang et al., 2003).
- CTNNB1 (beta-catenin 1): It is a canonical oncogene that has been identified as overexpressed in intestinal metaplasia and gastric adenocarcinomas (Werner et al., 2001; Huang et al., 2018).
- MLH1 (mutL homolog 1): Its expression has been reported as absent or downregulated in intestinal metaplasia, dysplasia, and gastric cancers (Takeda et al., 2012; Hu et al., 2018).
- CXCR4 (chemokine receptor type 4): Its expression has been associated with the staging of gastric cancer, being reduced in the majority of gastrointestinal tumors and significantly higher in patients with advanced stages of gastric cancer (Shibuta et al., 1997; Hannelien et al., 2012; Nikzaban et al., 2014).
- CXCR1 (C-X-C motif chemokine receptor 1): It has been reported to be strongly expressed in gastric carcinoma (Eck et al., 2003; Hannelien et al., 2012).
- KRT14 (keratin 14): It is a squamous cell marker that is down-regulated by CDX2 transfection (Liu et al., 2007). In addition, although it has been determined as significantly overexpressed in intestinal metaplasia by our analysis when the parameter was selected by OLIN, it has been reported as down-regulated in esophageal adenocarcinoma when compared to normal esophagus (Lv et al., 2019).

Genes Involved in Cancer

By performing a gene enrichment analysis, we identified, at a significance level of 5% (after FDR correction), 31 differentially expressed genes that are involved in cancer. Their respective p-values and fold changes are shown in **Table 3**. We remark that their association with intestinal metaplasia has not been clearly demonstrated yet. Thus, further investigation has to be done to confirm such conclusions.

Particularly, two replicates of the CCND1 gene and the LAMB2 gene were identified as differentially expressed only by the conventional approaches, suggesting that they may be false positives. Next, we briefly describe their association with cancer:

- CCND1 (cyclin D1): In contrast to its underexpression identified by the conventional analyses, it has been frequently reported as overexpressed in intestinal metaplasia, human neoplasias, and several tumors (Hosokawa and Arnold, 1998; Franchi et al., 2015).
- LAMB2 (laminin subunit beta 2): Although its expression has been associated with some carcinomas, ts expression is tightly regulated in normal human tissues and in disease (Wewer et al., 1994; Ljubimova et al., 2006).

DISCUSSIONS

Faced with the growing trend of multi-omics data integration in the midst of a replication crisis, improved microarray data analyses are crucial to identifying more reliable results (Ritchie et al., 2015a).

TABLE 1 | Genes reported in the literature as associated with intestinal metaplasia of the stomach that were identified as differentially expressed in our analysis at a significance level of 5% (after FDR correction).

Gene	**Improved estimation for the and values**						**Conventional estimation for the M_t and A_t values**					
	***f* by our method**			***f* by OLIN**			***f* by our method**			***f* by OLIN**		
	p	**adj. p**	**FC**	**p**	**adj. p**	**FC**	**p**	**adj. p**	**FC**	**p**	**adj. p**	**FC**
CLND3	2.70×10^{-12}	4.28×10^{-8}	2.86	1.84×10^{-12}	2.32×10^{-8}	2.74	$2.77{\times}10^{-12}$	4.01×10^{-8}	2.86	1.87×10^{-12}	2.33×10^{-8}	2.74
CLND3	2.23×10^{-5}	1.35×10^{-3}	0.59	1.63×10^{-5}	1.07×10^{-3}	0.60	2.23×10^{-5}	1.35×10^{-3}	0.59	1.55×10^{-5}	1.04×10^{-3}	0.60
MUC2	3.51×10^{-11}	1.32×10^{-7}	1.73	3.14×10^{-11}	1.06×10^{-7}	1.71	3.21×10^{-11}	1.21×10^{-7}	1.73	3.06×10^{-11}	1.04×10^{-7}	1.71
MUC2	1.90×10^{-4}	$6.56{\times}10^{-3}$	0.24	2.14×10^{-4}	7.19×10^{-3}	0.24	1.96×10^{-4}	6.69×10^{-3}	0.24	2.35×10^{-4}	7.74×10^{-3}	0.23
CDX1	4.22×10^{-10}	6.05×10^{-7}	2.15	4.53×10^{-10}	6.74×10^{-7}	2.13	4.03×10^{-7}	5.94×10^{-7}	2.16	4.40×10^{-10}	6.98×10^{-7}	2.14
ANPEP	4.28×10^{-10}	6.05×10^{-7}	3.14	5.31×10^{-10}	7.19×10^{-7}	3.08	4.37×10^{-10}	6.17×10^{-7}	3.13	5.19×10^{-10}	7.03×10^{-7}	3.07
CLCA1	2.55×10^{-9}	1.69×10^{-6}	3.75	7.18×10^{-10}	8.49×10^{-7}	3.85	2.71×10^{-9}	1.70×10^{-6}	3.75	7.15×10^{-10}	8.93×10^{-7}	3.85
DMBT1	2.79×10^{-9}	1.75×10^{-6}	3.39	4.22×10^{-9}	2.43×10^{-6}	3.26	2.77×10^{-9}	1.71×10^{-6}	3.39	3.98×10^{-9}	2.33×10^{-6}	3.26
GUCY2C	3.07×10^{-9}	1.86×10^{-6}	2.31	9.58×10^{-9}	4.07×10^{-6}	2.20	3.10×10^{-9}	1.84×10^{-6}	2.31	9.70×10^{-9}	4.06×10^{-6}	2.19
CLDN7	3.78×10^{-9}	2.17×10^{-6}	2.37	2.21×10^{-9}	1.56×10^{-6}	2.23	1.24×10^{-9}	1.13×10^{-6}	2.27	2.30×10^{-9}	1.59×10^{-6}	2.22
CDH17	4.21×10^{-9}	2.27×10^{-6}	2.69	4.83×10^{-9}	2.64×10^{-6}	2.65	4.16×10^{-9}	2.24×10^{-6}	2.69	4.73×10^{-9}	2.59×10^{-6}	2.65
CDX2	5.67×10^{-9}	2.80×10^{-6}	1.01	7.29×10^{-9}	3.40×10^{-6}	1.00	6.00×10^{-9}	2.82×10^{-6}	1.01	7.67×10^{-9}	3.51×10^{-6}	1.00
DEFA5	1.17×10^{-7}	2.48×10^{-5}	3.33	1.17×10^{-7}	2.45×10^{-5}	3.29	1.18×10^{-7}	2.46×10^{-5}	3.32	1.17×10^{-7}	2.43×10^{-5}	3.28
VDR	2.82×10^{-7}	4.94×10^{-5}	1.15	1.61×10^{-7}	3.23×10^{-5}	1.12	2.60×10^{-7}	4.64×10^{-5}	1.15	1.57×10^{-7}	3.17×10^{-5}	1.12
ISX	5.26×10^{-7}	8.04×10^{-5}	1.33	5.57×10^{-7}	8.25×10^{-5}	1.32	5.37×10^{-7}	8.06×10^{-5}	1.33	5.83×10^{-7}	8.03×10^{-5}	1.31
CLDN4	1.15×10^{-6}	1.43×10^{-4}	1.20	1.33×10^{-6}	1.62×10^{-4}	1.19	1.15×10^{-6}	1.40×10^{-4}	1.19	1.33×10^{-6}	1.60×10^{-4}	1.18
ACSL5	2.44×10^{-6}	2.49×10^{-4}	1.45	2.29×10^{-6}	2.42×10^{-4}	1.45	2.17×10^{-6}	2.26×10^{-4}	1.46	2.16×10^{-6}	2.30×10^{-4}	1.45
REG4	3.24×10^{-6}	3.06×10^{-4}	2.50	3.53×10^{-6}	3.35×10^{-4}	2.45	3.21×10^{-6}	3.02×10^{-4}	2.50	3.49×10^{-6}	3.31×10^{-4}	2.45
REG4	3.62×10^{-4}	1.08×10^{-2}	1.28	1.41×10^{-3}	2.84×10^{-2}	1.11	3.57×10^{-4}	1.06×10^{-2}	1.28	1.35×10^{-3}	2.76×10^{-2}	1.11
RUNX1	1.11×10^{-5}	7.87×10^{-4}	–0.56	6.91×10^{-6}	5.50×10^{-4}	–0.57	1.01×10^{-5}	7.16×10^{-4}	–0.55	7.59×10^{-6}	5.93×10^{-4}	–0.57
FOXA2	1.12×10^{-5}	7.90×10^{-4}	–1.13	9.18×10^{-6}	6.75×10^{-4}	–1.14	1.10×10^{-5}	7.72×10^{-4}	–1.13	9.51×10^{-6}	6.96×10^{-4}	–1.13
FOXA2	1.67×10^{-4}	5.93×10^{-3}	–0.86	2.12×10^{-4}	7.16×10^{-3}	–0.85	1.73×10^{-4}	6.10×10^{-3}	–0.86	2.22×10^{-4}	7.42×10^{-3}	–0.85
FOXA2	7.25×10^{-3}	**8.39×10^{-2}**	–0.61	8.20×10^{-3}	**9.01×10^{-2}**	–0.60	7.71×10^{-3}	**8.67×10^{-2}**	–0.61	8.13×10^{-3}	**9.03×10^{-2}**	–0.60
SOX2	1.62×10^{-5}	1.05×10^{-3}	–0.87	1.44×10^{-5}	9.73×10^{-4}	–0.87	1.56×10^{-5}	1.01×10^{-3}	–0.87	1.38×10^{-5}	9.41×10^{-4}	–0.87
SOX2	1.48×10^{-4}	5.50×10^{-3}	–0.77	3.23×10^{-4}	9.87×10^{-3}	–0.74	1.55×10^{-4}	5.62×10^{-3}	–0.76	3.26×10^{-4}	1.00×10^{-2}	–0.73
SERPINB5	2.42×10^{-5}	1.44×10^{-3}	1.04	2.55×10^{-5}	1.51×10^{-3}	1.03	2.46×10^{-5}	1.45×10^{-3}	1.05	2.67×10^{-5}	1.58×10^{-2}	1.03
SERPINB5	1.15×10^{-4}	4.59×10^{-3}	0.64	1.18×10^{-4}	4.65×10^{-3}	0.64	1.13×10^{-4}	4.49×10^{-3}	0.64	1.14×10^{-4}	4.52×10^{-3}	0.64
SERPINB5	1.73×10^{-2}	**1.42×10^{-1}**	0.11	1.18×10^{-2}	**1.14×10^{-1}**	0.12	1.66×10^{-2}	**1.39×10^{-1}**	0.11	1.22×10^{-2}	**1.16×10^{-1}**	0.11
FAS	6.35×10^{-5}	2.95×10^{-3}	0.41	6.54×10^{-5}	3.02×10^{-3}	0.41	6.46×10^{-5}	2.97×10^{-3}	0.41	6.93×10^{-5}	3.12×10^{-3}	0.41
CDHI	2.13×10^{-4}	7.14×10^{-3}	0.62	1.97×10^{-4}	6.74×10^{-3}	0.60	1.88×10^{-4}	6.50×10^{-3}	0.62	2.05×10^{-4}	6.94×10^{-3}	0.60
EMPI	6.05×10^{-4}	1.57×10^{-2}	0.94	6.45×10^{-4}	1.65×10^{-2}	0.90	5.77×10^{-4}	1.51×10^{-2}	0.94	6.61×10^{-4}	1.68×10^{-2}	0.90
EMPI	7.22×10^{-3}	**8.36×10^{-2}**	0.37	5.94×10^{-3}	**7.37×10^{-2}**	038	7.02×10^{-3}	**8.19×10^{-2}**	0.37	5.95×10^{-3}	**7.39×10^{-2}**	0.37
FGFR2	7.50	1.86×10^{-2}	–0.57	9.15×10^{-4}	2.12×10^{-2}	–0.58	7.44×10^{-4}	1.83×10^{-2}	–0.57	9.13×10^{-4}	2.11×10^{-2}	–0.57
FGFR2	8.37×10^{-3}	**9.20×10^{-2}**	–0.12	7.95×10^{-3}	**8.85×10^{-2}**	–0.12	9.07×10^{-3}	**9.65×10^{-2}**	–0.12	8.47×10^{-3}	**9.23×10^{-2}**	–0.12
PGC	9.29×10^{-4}	2.15×10^{-2}	–1.71	1.47×10^{-3}	2.92×10^{-2}	–1.45	$7.65 \times \times 10^{-4}$	1.87×10^{-2}	–1.64	1.46×10^{-3}	2.91×10^{-2}	–1.45
LRIG1	9.74×10^{-4}	2.22×10^{-2}	–0.67	4.82×10^{-4}	1.34×10^{-2}	–0.67	8.72×10^{-4}	2.04×10^{-2}	–0.66	5.07×10^{-4}	1.39×10^{-2}	–0.67
KRT20	1.05×10^{-3}	2.32×10^{-2}	1.49	1.18×10^{-3}	2.52×10^{-2}	1.46	1.02×10^{-3}	2.26×10^{-2}	1.49	1.17×10^{-3}	2.50×10^{-2}	1.46

Each column shows the p-value (p), the FDR-corrected p-value (adj. p), and the fold change (FC) obtained in a variant of the database. P-values greater than 5% are shown in bold type.

TABLE 2 | Other genes reported in the literature as associated with intestinal metaplasia of the stomach that were identified as differentially expressed in our analysis at a significance level of 5% (without FDR correction).

Gene	Improved estimation for the M_t and A_t values						Conventional estimation for the M_t and A_t values					
	***f* by our method**			***f* by OLIN**			***f* by our method**			***f* by OLIN**		
	p	**adj. p**	**FC**	**p**	**adj. p**	**FC**	**p**	**adj. p**	**FC**	**p**	**adj. p**	**FC**
VEGFA	3.76×10^{-3}	**5.52×10^{-2}**	−0.76	4.16×10^{-3}	**5.84×10^{-2}**	−0.75	3.54×10^{-3}	**5.28×10^{-2}**	−0.76	4.21×10^{-3}	**$\times 10^{-2}$**	−0.75
VEGFA	4.03×10^{-2}	**2.35×10^{-1}**	−0.25	3.93×10^{-2}	**2.29×10^{-1}**	−0.25	4.65×10^{-2}	**2.54×10^{-1}**	−0.25	4.52×10^{-2}	**$\times 10^{-1}$**	−0.25
PPP1R1B	3.96×10^{-3}	**5.70×10^{-2}**	0.76	4.07×10^{-3}	**5.76×10^{-2}**	0.75	3.89×10^{-3}	**5.60×10^{-2}**	0.76	4.03×10^{-3}	**$\times 10^{-2}$**	0.75
MUC5AC	4.07×10^{-3}	**5.79×10^{-2}**	−1.08	3.54×10^{-3}	**5.24×10^{-2}**	−1.08	4.18×10^{-3}	**5.87×10^{-2}**	−1.07	3.58×10^{-3}	**$\times 10^{-2}$**	−1.08
MUC5AC	4.60×10^{-3}	**6.30×10^{-2}**	−0.83	4.51×10^{-3}	**6.15×10^{-2}**	−0.82	4.78×10^{-3}	**6.40×10^{-2}**	−0.82	4.50×10^{-3}	**$\times 10^{-2}$**	−0.82
CLDN18	4.78×10^{-3}	**6.46×10^{-2}**	−1.05	5.12×10^{-3}	**6.69×10^{-2}**	−1.03	4.83×10^{-3}	**6.44×10^{-2}**	−1.04	5.03×10^{-3}	**$\times 10^{-2}$**	−1.03
ASCC1	6.62×10^{-3}	**7.90×10^{-2}**	0.18	1.42×10^{-2}	**1.27×10^{-1}**	0.17	6.57×10^{-3}	**7.85×10^{-2}**	0.18	1.43×10^{-2}	**$\times 10^{-1}$**	0.17
FOXA3	6.85×10^{-3}	**8.09×10^{-2}**	−0.57	4.87×10^{-3}	**6.47×10^{-2}**	−0.57	6.98×10^{-3}	**8.15×10^{-2}**	−0.56	5.01×10^{-3}	**$\times 10^{-2}$**	−0.57
FOXA3	1.96×10^{-2}	**1.54×10^{-1}**	−0.53	2.05×10^{-2}	**1.58×10^{-1}**	−0.52	1.98×10^{-2}	**1.54×10^{-1}**	−0.53	1.98×10^{-2}	**$\times 10^{-1}$**	−0.52
GAST	8.99×10^{-3}	**9.60×10^{-2}**	−1.48	1.24×10^{-2}	**1.17×10^{-1}**	−1.31	9.15×10^{-3}	**9.69×10^{-2}**	−1.48	1.21×10^{-2}	**$\times 10^{-1}$**	−1.32
PIK3CA	1.02×10^{-2}	**1.04×10^{-1}**	−0.16	7.28×10^{-3}	**8.42×10^{-2}**	−0.17	9.62×10^{-3}	**9.97×10^{-2}**	−0.16	6.62×10^{-3}	**$\times 10^{-2}$**	−0.17
BHLHA15	1.04×10^{-2}	**1.05×10^{-1}**	−0.63	9.50×10^{-3}	**9.93×10^{-2}**	−0.63	1.11×10^{-2}	**1.09×10^{-1}**	−0.62	9.79×10^{-3}	**$\times 10^{-1}$**	−0.63
SLPI	1.07×10^{-2}	**1.06×10^{-1}**	−0.71	7.96×10^{-3}	**8.86×10^{-2}**	−0.70	1.41×10^{-2}	**1.26×10^{-1}**	−0.70	7.91×10^{-3}	**$\times 10^{-2}$**	−0.70
SLPI	1.80×10^{-2}	**1.46×10^{-1}**	−0.64	1.13×10^{-2}	**1.10×10^{-1}**	−0.66	1.74×10^{-2}	**1.43×10^{-1}**	−0.64	1.18×10^{-2}	**$\times 10^{-1}$**	−0.65
KLF5	1.22×10^{-2}	**1.15×10^{-1}**	0.54	1.60×10^{-2}	**1.36×10^{-1}**	0.49	1.24×10^{-2}	**1.16×10^{-1}**	0.54	1.55×10^{-2}	**$\times 10^{-1}$**	0.49
CXCR2	1.26×10^{-2}	**1.18×10^{-1}**	0.23	1.30×10^{-2}	**1.20×10^{-1}**	0.23	1.25×10^{-2}	**1.17×10^{-1}**	0.23	1.34×10^{-2}	**$\times 10^{-1}$**	0.23
MGMT	1.28×10^{-2}	**1.19×10^{-1}**	−0.30	1.09×10^{-2}	**1.08×10^{-1}**	−0.31	1.30×10^{-2}	**1.20×10^{-1}**	−0.30	1.09×10^{-2}	**$\times 10^{-1}$**	−0.31
MOS	1.32×10^{-2}	**1.21×10^{-1}**	0.14	5.84×10^{-3}	**7.29×10^{-2}**	0.16	1.24×10^{-2}	**1.16×10^{-1}**	0.14	6.22×10^{-3}	**$\times 10^{-2}$**	0.16
IL10	1.35×10^{-2}	**1.23×10^{-1}**	0.05	1.74×10^{-2}	**1.43×10^{-1}**	0.05	1.26×10^{-2}	**1.17×10^{-1}**	0.05	1.73×10^{-2}	**$\times 10^{-1}$**	0.05
GHRL	1.39×10^{-2}	**1.26×10^{-1}**	1.08	1.24×10^{-2}	**1.17×10^{-1}**	1.06	1.34×10^{-2}	**1.22×10^{-1}**	1.08	1.23×10^{-2}	**$\times 10^{-1}$**	1.06
KRT7	1.56×10^{-2}	**1.35×10^{-1}**	0.40	1.81×10^{-2}	**1.47×10^{-1}**	0.39	1.58×10^{-2}	**1.35×10^{-1}**	0.40	1.81×10^{-2}	**$\times 10^{-1}$**	0.39
CDKN1A	1.70×10^{-2}	**1.41×10^{-1}**	0.25	1.91×10^{-2}	**1.51×10^{-1}**	0.24	1.70×10^{-2}	**1.40×10^{-1}**	0.24	1.94×10^{-2}	**$\times 10^{-1}$**	0.24
CDKN1A	3.48×10^{-2}	**2.17×10^{-1}**	0.42	4.19×10^{-2}	**2.37×10^{-1}**	0.39	3.34×10^{-2}	**2.11×10^{-1}**	0.42	4.17×10^{-2}	**$\times 10^{-1}$**	0.39
PDPK1	2.65×10^{-2}	**1.85×10^{-1}**	0.17	4.31×10^{-2}	**2.41×10^{-1}**	0.15	2.61×10^{-2}	**1.82×10^{-1}**	0.17	4.25×10^{-2}	**$\times 10^{-1}$**	0.15
PDX1	2.72×10^{-2}	**1.87×10^{-1}**	0.06	2.29×10^{-2}	**1.69×10^{-1}**	0.06	2.28×10^{-2}	**1.68×10^{-1}**	0.06	2.07×10^{-2}	**$\times 10^{-1}$**	0.06
HSPB1	3.22×10^{-2}	**2.07×10^{-1}**	−0.58	4.43×10^{-2}	**2.45×10^{-1}**	−0.55	4.65×10^{-2}	**2.53×10^{-1}**	−0.53	4.43×10^{-2}	**$\times 10^{-1}$**	−0.55
HSPB1	3.66×10^{-2}	**2.23×10^{-1}**	−0.56	3.55×10^{-2}	**2.17×10^{-1}**	−0.55	**5.03×10^{-2}**	**2.65×10^{-1}**	−0.52	3.63×10^{-2}	**$\times 10^{-1}$**	−0.55
HSPB1	3.66×10^{-2}	**2.23×10^{-1}**	−0.51	4.63×10^{-2}	**2.52×10^{-1}**	−0.48	**5.63×10^{-2}**	**2.80×10^{-1}**	−0.46	4.73×10^{-2}	**$\times 10^{-1}$**	−0.48
THBSI	3.27×10^{-2}	**2.08×10^{-1}**	−0.10	3.86×10^{-2}	**2.27×10^{-1}**	−0.10	3.36×10^{-2}	**2.11×10^{-1}**	−0.10	3.94×10^{-2}	**$\times 10^{-1}$**	−0.10
PTEN	3.30×10^{-2}	**2.09×10^{-1}**	0.16	**6.99×10^{-2}**	**3.12×10^{-1}**	0.14	3.17×10^{-2}	**2.04×10^{-1}**	0.16	**6.90×10^{-2}**	**$\times 10^{-1}$**	0.14
LGR5	3.63×10^{-2}	**2.22×10^{-1}**	−0.07	3.64×10^{-2}	**2.20×10^{-1}**	−0.07	4.22×10^{-2}	**2.41×10^{-1}**	−0.07	3.88×10^{-2}	**$\times 10^{-1}$**	−0.07
SHH	3.96×10^{-2}	**2.32×10^{-1}**	−0.07	2.68×10^{-2}	**1.85×10^{-1}**	−0.08	4.82×10^{-2}	**2.59×10^{-1}**	−0.07	3.10×10^{-2}	**$\times 10^{-1}$**	−0.08
TJP1	3.98×10^{-2}	**2.33×10^{-1}**	0.31	4.33×10^{-2}	**2.41×10^{-1}**	0.30	4.14×10^{-2}	**2.39×10^{-1}**	0.31	4.56×10^{-2}	**$\times 10^{-1}$**	0.29
PTGS2	4.02×10^{-2}	**2.35×10^{-1}**	0.21	3.90×10^{-2}	**2.28×10^{-1}**	0.20	4.00×10^{-2}	**2.34×10^{-1}**	0.21	3.73×10^{-2}	**$\times 10^{-1}$**	0.21
SOX9	4.48×10^{-2}	**2.48×10^{-1}**	−0.29	4.02×10^{-2}	**2.32×10^{-1}**	−0.30	4.45×10^{-2}	**2.48×10^{-1}**	−0.29	4.04×10^{-2}	**$\times 10^{-1}$**	−0.30
CTNNB1	4.53×10^{-2}	**2.50×10^{-1}**	0.33	**5.05×10^{-2}**	**2.63×10^{-1}**	0.33	4.83×10^{-2}	**2.59×10^{-1}**	0.33	**5.31×10^{-2}**	**$\times 10^{-1}$**	0.32
MLH1	4.55×10^{-2}	**2.51×10^{-1}**	−0.23	**6.82×10^{-2}**	**3.08×10^{-1}**	−0.22	4.97×10^{-2}	**2.63×10^{-1}**	−0.22	**6.80×10^{-2}**	**$\times 10^{-1}$**	−0.22
CDKN1B	4.56×10^{-2}	**2.51×10^{-1}**	−0.22	4.90×10^{-2}	**2.59×10^{-1}**	−0.22	4.41×10^{-2}	**2.46×10^{-1}**	−0.23	4.69×10^{-2}	**$\times 10^{-1}$**	−0.22
CXCR4	4.83×10^{-2}	**2.58×10^{-1}**	−0.43	**5.72×10^{-2}**	**2.81×10^{-1}**	−0.42	**5.00×10^{-2}**	**2.64×10^{-1}**	−0.43	**5.77×10^{-2}**	**$\times 10^{-1}$**	−0.42
CXCR1	4.98×10^{-2}	**2.63×10^{-1}**	0.19	**5.38×10^{-2}**	**2.72×10^{-1}**	0.18	4.64×10^{-2}	**2.53×10^{-1}**	0.19	**5.18×10^{-2}**	**$\times 10^{-1}$**	0.18
KRT14	**5.11×10^{-2}**	**2.67×10^{-1}**	0.19	3.65×10^{-2}	**2.20×10^{-1}**	0.19	**5.15×10^{-2}**	**2.68×10^{-1}**	0.19	3.95×10^{-2}	**$\times 10^{-1}$**	0.19

Each column shows the p-value (p), the FDR-corrected p-value (adj. p), and the fold change (FC) obtained in a variant of the database. P-values greater than 5% are shown in bold type.

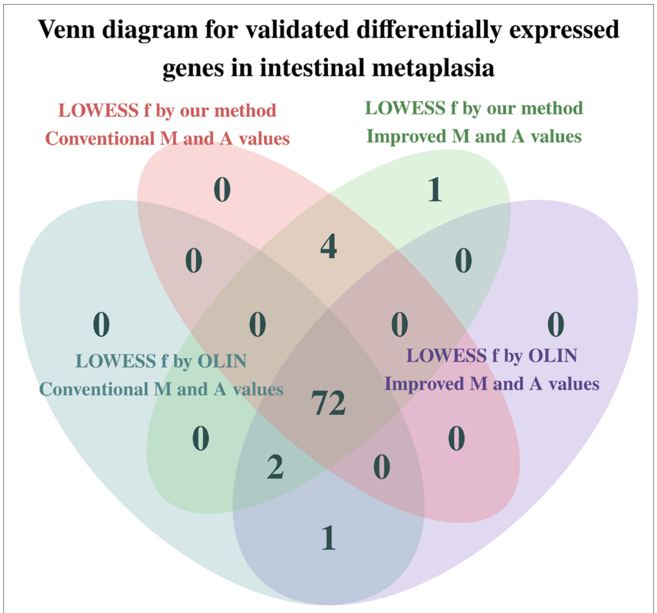


FIGURE 7 | Venn diagram for the total number of genes already identified as differentially expressed in intestinal metaplasia according to the literature. Inferences were made at a significance level of 5%.

Given that several pixel-level summary statistics are readily available in microarray databases, but are usually discarded in conventional approaches, we propose an improved estimation method for the M_t and A_t values, which takes into account the pixel-level variability. Specifically, we applied the multivariate delta method to derive estimators for the expected values of M_t and A_t, considering their Taylor's expansion up to the second-order terms. The conventional estimators, nonetheless, approximate the expected values considering only the zeroth-order term. Since the functions that define M_t and A_t are analytic (they are combinations of logarithmic function through addition or subtraction), the higher the number of terms of the Taylor expansion, the better the approximation of the function. Thus, we expect that the proposed estimators provide a better quantification of the hybridization signal. Also, by using these improved estimators, pixel-level dispersion can play an essential role in the analysis, increasing reliability.

To minimize the propagation of errors, the M_t and A_t values have to be properly normalized. Thus, we also propose a method for selecting the LOWESS smoothing parameter f that provides an optimal bias–variance compromise, considering some specific characteristics of microarray experiments, such as heteroskedasticity. This optimal normalization method leads to a more parsimonious correction of the systematic biases and, consequently, to greater preservation of the biological variation of interest.

By using the proposed methods, more variability information is considered and retained, improving inferences and preventing false conclusions. Thus, we expect to perform a more conservative analysis, where possibly fewer but more reliable differentially expressed genes are identified. In other words, we expect a reduction in both the false-positive and false-negative error rates.

Besides the theoretical support, relevant empirical observations could be drawn by a comparative study between the methods using real intestinal metaplasia microarray data. The results shows that inferences on differential gene expression were moderately affected by the incorporation of the pixel-level variability in the estimation of the M_t and A_t values and significantly affected by the LOWESS within-slide normalization using a smoothing parameter selected by the method. Both proposed methods tend to increase the within-group variability (the denominator of the t-statistic). However, for many genes, such increase occurred along with an increase in the difference between the group means (the absolute value of the t-statistic numerator), significantly reducing their respective p-values. Thus, many genes were identified as differentially expressed only when the proposed methods were used and some of them have been validated by other studies.

It is important to remark that most of the genes reported in the literature as differentially expressed in intestinal metaplasia were validated with a very strong association with the disease. Thus, these genes are probably more robust to difference approaches for estimating and normalizing the gene expression levels. On the other hand, genes sensitive to methods that address essential uncertainties in measurements are precisely those plagued with major reproducibility issues. Measurement error is one of the most damaging sources of error and has been neglected in many published analyses, thereby increasing uncertainty in parameter estimates and even inflating the estimates of effect sizes (Loken and Gelman, 2017). Thus, particularly for those sensitive genes, a more robust analysis is needed so that false conclusions are not made.

In this paper, we focused on gene expression from two-color microarray data, but it is possible to use the same ideas to improve estimation and normalization of any fluorescent signal quantified by microarray image analysis. Also, the proposed methods could be adapted for oligonucleotide (one-color) microarray data. Particularly, the cyclic LOWESS normalization method (Bolstad et al., 2003) could be extended by just considering that the M_t and A_t values are defined by comparing pairs of arrays instead of pairs of channels and that the LOWESS normalization is applied to all distinct combination of two arrays. Although not so straightforward, it is also possible to adapt our methods to handle next-generation sequencing (NGS) data. Recently, Law et al. (Law et al., 2014) showed that RNA-Seq counts after log transformation and normalization by sequencing depth (log-counts per million, or log-cpm) can be properly analyzed by methods based on the normal distribution if a precision weight for each observation is taken into account. It was used to adapt all methods in the `limma` package (initially developed for microarrays) to also handle RNA-Seq and other sequence count data (Ritchie et al., 2015b). Therefore, considering the current need for accounting and propagating measurement uncertainties through analyses of NGS data (O'Rawe et al., 2015), a possible future work is to adapt our ideas to improve transcriptome profiling from RNA-Seq data. Specifically, one could investigate whether it is possible to use the delta method for incorporating a measure of uncertainty

TABLE 3 | Genes belonging to the "pathways in cancer" category identified as differentially expressed between normal and intestinal metaplasia groups at a significance level of 5% (after FDR correction).

Gene	Improved estimation for the M_t and A_t values						Conventional estimation for the M_t and A_t values					
	f by our method			*f* by OLIN			*f* by our method			*f* by OLIN		
	p	adj. p	FC	p	adj. p	FC	p	adj. p	FC	p	adj. p	FC
PLD1	4.08×10^{-7}	6.60×10^{-5}	1.03	3.54×10^{-7}	5.86×10^{-5}	0.99	4.31×10^{-7}	6.73×10^{-5}	1.03	3.60×10^{-7}	5.89×10^{-5}	0.99
PLD1	2.50×10^{-6}	2.53×10^{-4}	0.43	3.49×10^{-6}	3.32×10^{-4}	0.42	2.36×10^{-6}	2.41×10^{-4}	0.43	3.34×10^{-6}	3.24×10^{-4}	0.42
PLD1	9.73×10^{-5}	4.06×10^{-3}	0.49	9.90×10^{-5}	4.14×10^{-3}	0.49	1.07×10^{-4}	4.35×10^{-3}	0.48	1.08×10^{-4}	4.37×10^{-3}	0.48
MITF	2.68×10^{-6}	2.68×10^{-4}	–0.69	6.38×10^{-6}	5.19×10^{-4}	–0.69	2.70×10^{-6}	2.67×10^{-4}	–0.68	6.29×10^{-6}	5.19×10^{-4}	–0.69
MAX	6.06×10^{-6}	4.93×10^{-4}	0.43	7.72×10^{-6}	6.00×10^{-4}	0.43	5.26×10^{-6}	4.37×10^{-4}	0.43	7.13×10^{-6}	5.67×10^{-4}	0.43
MAX	1.61×10^{-3}	3.10×10^{-2}	0.35	1.35×10^{-3}	2.75×10^{-2}	0.35	1.36×10^{-3}	2.77×10^{-2}	0.35	1.31×10^{-3}	2.68×10^{-2}	0.35
NOS2	7.08×10^{-6}	5.52×10^{-4}	1.37	7.61×10^{-6}	5.93×10^{-4}	1.34	6.59×10^{-6}	5.19×10^{-4}	1.37	7.28×10^{-6}	5.76×10^{-4}	1.34
CDKN2B	8.14×10^{-6}	6.14×10^{-4}	0.98	8.41×10^{-6}	6.38×10^{-4}	0.97	7.79×10^{-6}	5.94×10^{-4}	0.98	8.20×10^{-6}	6.25×10^{-4}	0.97
CDKN2B	4.00×10^{-4}	1.16×10^{-2}	0.24	5.72×10^{-4}	1.51×10^{-2}	0.23	3.33×10^{-4}	1.01×10^{-2}	0.24	4.84×10^{-4}	1.34×10^{-2}	0.24
VEGFB	1.23×10^{-5}	8.41×10^{-4}	–0.95	7.23×10^{-6}	5.68×10^{-4}	–0.89	4.36×10^{-6}	3.78×10^{-4}	–0.94	6.65×10^{-6}	5.35×10^{-4}	–0.89
VEGFB	1.09×10^{-4}	4.40×10^{-3}	–0.55	1.05×10^{-4}	4.32×10^{-3}	–0.55	1.09×10^{-4}	4.38×10^{-3}	–0.54	1.04×10^{-4}	4.26×10^{-3}	–0.55
ITGA6	2.80×10^{-5}	1.60×10^{-3}	0.63	3.92×10^{-5}	2.06×10^{-3}	0.59	2.43×10^{-5}	1.44×10^{-3}	0.64	3.63×10^{-5}	1.96×10^{-3}	0.59
RXRA	3.03×10^{-5}	1.71×10^{-3}	0.25	4.33×10^{-5}	2.23×10^{-3}	0.26	3.05×10^{-5}	1.72×10^{-3}	0.25	4.76×10^{-5}	2.39×10^{-3}	0.25
PIAS3	4.53×10^{-5}	2.29×10^{-3}	–0.55	2.93×10^{-5}	1.68×10^{-3}	–0.57	4.81×10^{-5}	2.38×10^{-3}	–0.55	2.85×10^{-5}	1.65×10^{-3}	–0.57
ITGA2	5.24×10^{-5}	2.53×10^{-3}	0.48	7.52×10^{-5}	3.33×10^{-3}	0.47	5.88×10^{-5}	2.76×10^{-3}	0.48	7.43×10^{-5}	3.30×10^{-3}	0.47
FZD8	6.00×10^{-5}	2.83×10^{-3}	–0.60	5.09×10^{-5}	2.51×10^{-3}	–0.60	6.05×10^{-5}	2.81×10^{-3}	–0.60	4.83×10^{-5}	2.42×10^{-3}	–0.61
FOXO1	1.54×10^{-4}	5.65×10^{-3}	–0.53	1.03×10^{-4}	4.25×10^{-3}	–0.53	1.39×10^{-4}	5.24×10^{-3}	–0.53	1.00×10^{-4}	4.16×10^{-3}	–0.54
FOXO1	2.70×10^{-3}	4.46×10^{-2}	–0.20	2.66×10^{-3}	4.33×10^{-2}	–0.20	2.80×10^{-3}	4.51×10^{-2}	–0.20	2.42×10^{-3}	4.06×10^{-2}	–0.21
EGLN1	1.85×10^{-4}	6.42×10^{-3}	0.50	4.00×10^{-4}	1.16×10^{-2}	0.46	1.73×10^{-4}	6.10×10^{-3}	0.50	3.96×10^{-4}	1.16×10^{-2}	0.46
TGFBR2	2.88×10^{-4}	9.06×10^{-3}	–0.36	8.86×10^{-5}	3.78×10^{-3}	–0.37	2.68×10^{-4}	8.46×10^{-3}	–0.36	8.71×10^{-5}	3.73×10^{-3}	–0.37
WNT3	4.16×10^{-4}	1.19×10^{-2}	0.51	4.13×10^{-4}	1.19×10^{-2}	0.51	4.00×10^{-4}	1.15×10^{-2}	0.51	4.22×10^{-4}	1.21×10^{-2}	0.50
CKS1B	7.02×10^{-4}	1.76×10^{-2}	–0.29	1.91×10^{-3}	3.46×10^{-2}	–0.27	1.04×10^{-3}	2.29×10^{-2}	–0.27	2.01×10^{-3}	3.56×10^{-2}	–0.27
AXIN2	7.63×10^{-4}	1.88×10^{-2}	–0.53	8.64×10^{-4}	2.02×10^{-2}	–0.53	7.62×10^{-4}	1.86×10^{-2}	–0.53	8.52×10^{-4}	2.01×10^{-2}	–0.53
CCND1	9.74×10^{-4}	2.22×10^{-2}	–0.55	7.00×10^{-4}	1.75×10^{-2}	–0.55	9.79×10^{-4}	2.21×10^{-2}	–0.55	6.73×10^{-4}	1.70×10^{-2}	–0.56
CCND1	3.34×10^{-3}	**5.12×10^{-2}**	–0.76	2.81×10^{-3}	4.51×10^{-2}	–0.77	3.45×10^{-3}	**5.19×10^{-2}**	–0.76	2.88×10^{-3}	4.58×10^{-2}	–0.77
CCND1	3.49×10^{-3}	**5.23×10^{-2}**	–0.26	4.11×10^{-3}	**5.80×10^{-2}**	–0.26	3.19×10^{-3}	4.95×10^{-2}	–0.27	3.75×10^{-3}	**5.45×10^{-2}**	–0.26
ITGAV	1.03×10^{-3}	2.30×10^{-2}	–0.36	1.06×10^{-3}	2.34×10^{-2}	–0.35	9.39×10^{-4}	2.15×10^{-2}	–0.36	1.04×10^{-3}	2.29×10^{-2}	–0.35
CEBPA	1.50×10^{-3}	2.96×10^{-2}	0.63	1.79×10^{-3}	3.32×10^{-2}	0.60	1.36×10^{-3}	2.77×10^{-2}	0.63	1.76×10^{-3}	3.27×10^{-2}	0.60
JUN	1.60×10^{-3}	3.09×10^{-2}	–0.58	1.57×10^{-3}	3.04×10^{-2}	–0.54	1.94×10^{-3}	3.48×10^{-2}	–0.56	1.56×10^{-3}	3.03×10^{-2}	–0.54
WNT11	2.98×10^{-3}	4.76×10^{-2}	0.28	2.96×10^{-3}	4.65×10^{-2}	0.28	3.06×10^{-3}	4.81×10^{-2}	0.28	2.97×10^{-3}	4.67×10^{-2}	0.28
LAMB2	5.18×10^{-3}	**6.76×10^{-2}**	–0.52	2.58×10^{-3}	4.25×10^{-2}	–0.49	4.42×10^{-3}	**6.10×10^{-2}**	–0.49	2.61×10^{-3}	4.28×10^{-2}	–0.49

Each column shows the p-value (p), the FDR-corrected p-value (adj. p), and the fold change (FC) obtained in a variant of the database. P-values greater than 5% are shown in bold type.

to each base call, usually provided by base-calling algorithms, into the log-cpm estimator, leading to a more accurate gene expression quantification from RNA-Seq data.

DATA AVAILABILITY

The omicsMA R package contains the source code of the proposed methods and part of the metaplasia dataset analyzed in this study. It was implemented using R, version 3.5.1, and depends on the locfit (Loader, 2013), maigesPack (Esteves et al., 2016), and limma (Ritchie et al., 2015b) R packages. The omicsMA R package is available at https://github.com/adele/omicsMA, and the latest release is available at https://github.com/adele/omicsMA/releases/latest.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the international guidelines for investigations involving human beings with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Institutional Committee of the A.C. Camargo Cancer Center (process number 1023/07).

AUTHOR CONTRIBUTIONS

AR and RH conceived of the presented ideas. AR derived the models, implemented the methods, and analyzed the data. AR wrote the manuscript with support from RH and JS. All authors discussed the results and contributed to the final manuscript. RH and JS supervised the project.

FUNDING

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; National Council of Technological and Scientific Development (CNPq); and NAP eScience - PRP - USP. It was also supported by the Foundation for Research Support of the State of São Paulo (FAPESP) [grants 06/03227-2, 2011/50761-2 and 2015/01587-0].

ACKNOWLEDGMENTS

We are greatly thankful to Luiz F. L. Reis, Director of the Research and Education, Hospital Sírio-Libanês, São Paulo, for making the intestinal metaplasia database available to us, and to Professors Eduardo Jordão Neves and Luís Gustavo Estevesfor helping us with data preprocessing and analysis.

REFERENCES

Ahmed, A. A., Vias, M., Iyer, N. G., Caldas, C., and Brenton, J. D. (2004). Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res.* 32 (5), e50–e50. doi: 10.1093/nar/gnh047

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genom.* 8 (1), 33. doi: 10.1186/s12920-015-0108-y

Bakewell, D. J., and Wit, E. (2005). Weighted analysis of microarray gene expression using maximum-likelihood. *Bioinformatics*, 21 (6), 723–729. doi: 10.1093/bioinformatics/bti051

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41 (D1), D991–D995. doi: 10.1093/nar/gks1193

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57 (1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berger, J. A., Hautaniemi, S., Järvinen, A. K., Edgren, H., Mitra, S. K., and Astola, J. (2004). Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinform.* 5 (1), 194. doi: 10.1186/1471-2105-5-194

Bolstad, B. M., Irizarry, R. A., Åstrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19 (2), 185–193. doi: 10.1093/bioinformatics/19.2.185

Brady, P. D., and Vermeesch, J. R. (2012). Genomic microarrays: a technology overview. *Pren. Diagn.* 32 (4), 336–343. doi: 10.1002/pd.2933

Brown, C. S., Goodwin, P. C., and Sorger, P. K. (2001). Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. U. S. A.* 98 (16), 8944–8949. doi: 10.1073/pnas.161242998

Casella, G., and Berger, R. L. (1990). *Statistical Inference* Vol. 70. CA: Duxbury Press Belmont, 240–245.

Chan, S. H., and Chang, W. C. (2009). A robust ratio estimator of gene expression *via* inverse-variance weighting for cDNA microarray images. *Comput. Stat. Data Anal.* 53 (5) 1577–1589. doi: 10.1016/j.csda.2008.06.003

Chiogna, M., Massa, M. S., Risso, D., and Romualdi, C. (2009). A comparison on effects of normalisations in the detection of differentially expressed genes. *BMC Bioinform.* 10 (1), 61. doi: 10.1186/1471-2105-10-61

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* 32, 490–495. doi: 10.1038/ng1031

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 74 (368), 829–836. doi: 10.1080/01621459.1979.10481038

Cleveland, W., and Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* 83 (403), 596–610. doi: 10.1080/01621459.1988.10478639

Cleveland, W. S., Devlin, S. J., and Grosse, E. (1988). Regression by local fitting: methods, properties, and computational algorithms. *J. Econom.* 37 (1), 87–114. doi: 10.1016/0304-4076(88)90077-2

Coussens, L. M., and Werb, Z. (2002). Inflammation and cancer. *Nature*, 420 (6917), 860–867. doi: 10.1038/nature01322

Dodd, L. E., Korn, E. L., McShane, L. M., Chandramouli, G., and Chuang, E. Y. (2004). Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics*, 20 (16), 2685–2693. doi: 10.1093/bioinformatics/bth309

Drăghici, S. (2012). *Statistics and data analysis for microarrays using R and bioconductor* Vol. 4. (CRC Press).

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.* 12 (1), 111–140.

Ebert, M. P., Schäfer, C., Chen, J., Hoffmann, J., Gu, P., Kubisch, C., et al. (2005). Protective role of heat shock protein 27 in gastric mucosal injury. *J. Pathol. J. Pathol. Soc. G. B. Irel.* 207 (2), 177–184. doi: 10.1002/path.1815

Eck, M., Schmausser, B., Scheller, K., Brändlein, S., and Müller-Hermelink, H. (2003). Pleiotropic effects of cxc chemokines in gastric carcinoma: differences in cxcl8 and cxcl1 expression between diffuse and intestinal types of gastric carcinoma. *Clin. Exp. Immunol.* 134 (3), 508–515. doi: 10.1111/j.1365-2249.2003.02305.x

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1) 207–210. doi: 10.1093/nar/30.1.207

Esteves, G., Hirata Jr, R., Neves, E., Cristo, E., Simoes, A., and Fahham, L. (2016). *maigesPack: Functions to handle cDNA microarray data, including several methods of data analysis*. R package version 1.36.0.

Franchi, A., Palomba, A., Miligi, L., Ranucci, V., Degli Innocenti, D. R., Simoni, A., et al. (2015). Intestinal metaplasia of the sinonasal mucosa adjacent to intestinal-type adenocarcinoma. A morphologic, immunohistochemical, and molecular study. *Virchows Arch.* 466 (2), 161–168. doi: 10.1007/s00428-014-1696-1

Futschik, M., and Crompton, T. (2004a). Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biol.* 5 (8), R60. doi: 10.1186/gb-2004-5-8-r60

Futschik, M. E., and Crompton, T. (2004b). OLIN: optimized normalization, visualization and quality testing of two-channel microarray data. *Bioinformatics*, 21 (8), 1724–1726. doi: 10.1093/bioinformatics/bti199

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17 (6), 333–351. doi: 10.1038/nrg.2016.49

Hannelien, V., Karel, G., Sofie, S., et al. (2012). The role of cxc chemokines in the transition of chronic inflammation to esophageal and gastric cancer. *Biochim. Biophys. Acta (BBA)-Rev. Cancer*, 1825 (1), 117–129. doi: 10.1016/j.bbcan.2011.10.008

Hosokawa, Y., and Arnold, A. (1998). Mechanism of cyclin d1 (ccnd1, prad1) overexpression in human cancer cells: analysis of allele-specific expression. *Genes, Chromosome Cancer*, 22 (1), 66–71. doi: 10.1002/(SICI)1098-2264(199805)22:1<66::AID-GCC9>3.0.CO;2-5

Hoyle, D. C., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18 (4), 576–584. doi: 10.1093/bioinformatics/18.4.576

Hu, G., Qin, L., Zhang, X., Ye, G., and Huang, T. (2018). Epigenetic silencing of the mlh1 promoter in relation to the development of gastric cancer and its use as a biomarker for patients with microsatellite instability: a systematic analysis. *Cell. Physiol. Biochem.* 45 (1), 148–162. doi: 10.1159/000486354

Huang, K. K., Ramnarayanan, K., Zhu, F., Srivastava, S., Xu, C., Tan, A. L. K., et al. (2018). Genomic and epigenomic profiling of high-risk intestinal metaplasia reveals molecular determinants of progression to gastric cancer. *Cancer Cell*, 33(1), 137–150. doi: 10.1016/j.ccell.2017.11.018

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19 (5), 299–310. doi: 10.1038/nrg.2018.4

Karthik, S., and Manjunath, S. (2018). An enhanced approach for spot segmentation of microarray images. *Procedia Comput. Sci.* 132, 226–235. doi: 10.1016/j.procs.2018.05.192

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al. (2015). Array express update–simplifying data submissions. *Nucleic Acids Res.* 43, D1113–6. doi: 10.1093/nar/gku1057

Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-Seq read counts. *Genome Biol.*15 (2), R29. doi: 10.1186/gb-2014-15-2-r29

Lee, J. W., Jhun, M., Kim, J. Y., and Lee, J. (2008). An optimal choice of window width for LOWESS normalization of microarray data. *OR Spectr.* 30 (2), 235–248. doi: 10.1007/s00291-007-0092-5

Li, Y., Păun, A., and Păun, M. (2017). Improvements on contours based segmentation for DNA microarray image processing. *Theor. Comput. Sci.* 701, 174–189. doi: 10.1016/j.tcs.2017.04.013

Liu, Q., and Okui, R. (2013). Heteroscedasticity-robust cp model averaging. *Econom. J.* 16 (3), 463–472. doi: 10.1111/ectj.12009

Liu, T., Zhang, X., So, C. K., Wang, S., Wang, P., Yan, L., et al. (2007). Regulation of cdx2 expression by promoter methylation, and effects of cdx2 transfection on morphology and gene expression of human esophageal epithelial cells. *Carcinogenesis*, 28 (2), 488–496. doi: 10.1093/carcin/bgl176

Ljubimova, J. Y., Fujita, M., Khazenzon, N. M., Ljubimov, A. V., and Black, K. L. (2006). Changes in laminin isoforms associated with brain tumor invasion and angiogenesis. *Front. Biosci. J. Virtual Libr.* 11, 81–88. doi: 10.2741/1781

Loader, C. (1999). *Local regression and likelihood* Vol. 47. New York: Springer-Verlag.

Loader, C. (2013). *locfit: Local Regression, Likelihood and Density Estimation*. R package version 1.5-9.1.

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Sci.* 355 (6325), 584–585. doi: 10.1126/science.aal3618

Lv, J., Guo, L., Wang, J. H., Yan, Y. Z., Zhang, J., Wang, Y. Y., et al. (2019). Biomarker identification and trans-regulatory network analyses in esophageal adenocarcinoma and Barrett's esophagus. *World J. Gastroenterol.* 25 (2), 233–244. doi: 10.3748/wjg.v25.i2.233

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15 (4), 661–675. doi: 10.1080/00401706.1973.10489103

Nikzaban, M., Hakhamaneshi, M. S., Fakhari, S., Sheikhesmaili, F., Roshani, D., Ahsan, B., et al. (2014). The chemokine receptor cxcr4 is associated with the staging of gastric cancer. *Adv. Biomed. Res.* 3 (1), 16.

O'Rawe, J. A., Ferson, S., and Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends Genet.* 31 (2), 61–66. doi: 10.1016/j.tig.2014.12.002

Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., et al. (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, 23 (20), 2700–2707. doi: 10.1093/bioinformatics/btm412

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015a). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16 (2), 85–97. doi: 10.1038/nrg3868

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015b). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43(7), e47. doi: 10.1093/nar/gkv007

Shao, G., Li, D., Zhang, J., Yang, J., and Shangguan, Y. (2019). Automatic microarray image segmentation with clustering-based algorithms. *PloS One*, 14 (1), e0210075. doi: 10.1371/journal.pone.0210075

Shibuta, K., Begum, N. A., Mori, M., Shimoda, K., Akiyoshi, T., and Barnard, G. F. (1997). Reduced expression of the cxc chemokine hirh/sdf-1 mRNA in hepatoma and digestive tract cancer. *Int. J. Cancer*, 73 (5), 656–662. doi: 10.1002/(SICI)1097-0215(19971127)73:5<656::AID-IJC8>3.0.CO;2-W

Smyth, G. K., and Speed, T. (2003). Normalization of cDNA microarray data. *Methods*, 31(4), 265–273. doi: 10.1016/S1046-2023(03)00155-5

Sun, S., Huang, Y. W., Yan, P. S., Huang, T. H., and Lin, S. (2011). Preprocessing differential methylation hybridization microarray data. *BioData Mining* 4 (1), 13. doi: 10.1186/1756-0381-4-13

Takeda, Y., Yashima, K., Hayashi, A., Sasaki, S., Kawaguchi, K., Harada, K., et al. (2012). Expression of aid, p53, and mlh1 proteins in endoscopically resected differentiated-type early gastric cancer. *World J. Gastrointest. Oncol.* 4 (6), 131–137. doi: 10.4251/wjgo.v4.i6.131

Werner, M., Becker, K. F., Keller, G., and Höfler, H. (2001). Gastric adenocarcinoma: pathomorphology and molecular pathology. *J. Cancer Res. Clin. Oncol.* 127 (4), 207–216. doi: 10.1007/s004320000195

Wewer, U. M., Gerecke, D. R., Durkin, M. E., Kurtz, K. S., Mattei, M. G., Champliaud, M. F., et al. (1994). Human 2 chain of laminin (formerly s chain): cDNA cloning, chromosomal localization, and expression in carcinomas. *Genomics*, 24 (2), 243–252. doi: 10.1006/geno.1994.1612

Yang, Y., Dudoit, S., Luuc, P., and Speed, T. (2001). Normalization for cDNA microarray data. *Microarrays: Opt. Technol. Inform.* 4266, 141–152. doi: 10.1117/12.427982

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *J. Comput. Graph. Stat.* 11 (1), 108–136. doi: 10.1198/106186002317375640

Yang, L., Kuang, L. G., Zheng, H. C., Li, J. Y., Wu, D. Y., Zhang, S. M., et al. (2003). PTEN encoding product: a marker for tumorigenesis and progression of gastric carcinoma. *World J. Gastroenterol.* 9 (1), 35–39. doi: 10.3748/wjg.v9.i1.35

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.

APPENDIX

Estimation of *E*(*M*$_{tj}$) and *E*(*A*$_{tj}$) by the Delta Method

Let $f\,(R_{tj},\,G_{tj})$ be a twice differentiable function of two random variables, R_{tj} and G_{tj}.The second-order Taylor's expansion of at $(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))$ is:

$$\begin{aligned}
f(R_{tj},G_{tj}) \approx f(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))+\frac{\partial f}{\partial R_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))(R_{tj}-\mathbb{E}(R_{tj}))+\\
\frac{\partial f}{\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))(G_{tj}-\mathbb{E}(G_{tj}))+\\
\frac{1}{2}\left(\frac{\partial^2 f}{\partial R_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))(R_{tj}-\mathbb{E}(R_{tj}))^2+2\frac{\partial^2 f}{\partial R_{tj}\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\right.\\
\left.[(R_{tj}-\mathbb{E}(R_{tj}))(G_{tj}-\mathbb{E}(G_{tj}))]+\frac{\partial^2 f}{\partial G_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))(G_{tj}-\mathbb{E}(G_{tj}))^2\right).
\end{aligned}$$

An approximation of $(\mathbb{E}(f(R_{tj},G_{tj}))$ can be determined by the expected value of the second-order Taylor's expansion of *f*:

$$\begin{aligned}
\mathbb{E}(f(R_{tj},G_{tj})) \approx \mathbb{E}[f(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj})))]+\frac{\partial f}{\partial R_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathbb{E}(R_{tj}-\mathbb{E}(R_{tj}))+\\
\frac{\partial f}{\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathbb{E}(G_{tj}-\mathbb{E}(G_{tj}))+\\
\frac{1}{2}\left(\frac{\partial^2 f}{\partial R_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathbb{E}[(R_{tj}-\mathbb{E}(R_{tj}))^2]+\right.\\
2\frac{\partial^2 f}{\partial R_{tj}\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathbb{E}[(R_{tj}-\mathbb{E}(R_{tj}))(G_{tj}-\mathbb{E}(G_{tj}))]+\\
\left.\frac{\partial^2 f}{\partial G_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathbb{E}[(\mathrm{G}_{tj}-\mathbb{E}(G_{tj}))^2])\right).
\end{aligned}$$

Considering that

$$\begin{aligned}
&\mathrm{Var}\ (R_{tj})=\mathbb{E}[(R_{tj}-\mathbb{E}(R_{tj}))^2],\\
&\mathrm{Var}\ (G_{tj})=\mathbb{E}[(G_{tj}-\mathbb{E}(G_{tj}))^2],\ \text{and}\\
&\mathrm{Cov}\ (R_{tj},G_{tj})=\mathbb{E}[(R_{tj}-\mathbb{E}(R_{tj}))(G_{tj}-\mathbb{E}(G_{tj}))],
\end{aligned}$$

the following simplified expression for the expected value of $f\,(R_{tj},\,G_{tj})$ is obtained:

$$\begin{aligned}
\mathbb{E}(f(R_{tj},G_{tj}))\approx f(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))+\frac{1}{2}\left(\frac{\partial^2 f}{\partial R_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathrm{Var}\ (R_{tj})+\right.\\
2\frac{\partial^2 f}{\partial R_{tj}\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathrm{C}\mathrm{ov}\ (R_{tj},G_{tj})+\\
\left.\frac{\partial^2 f}{\partial G_{tj}^2}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\mathrm{Var}\ (G_{tj})\right).
\end{aligned}$$

Since

$$M_{tj}=f(R_{tj},G_{tj})\doteq\log_2(R_{tj})-\log_2(G_{tj}),$$

the first and second derivatives of the function that defines M_{tj} are:

$$\begin{aligned}
&\frac{\partial f}{\partial R_{tj}}=\frac{1}{R_{tj}ln\ (2)}; \qquad \frac{\partial f}{\partial G_{tj}}=-\frac{1}{G_{tj}ln\ (2)};\\
&\frac{\partial^2 f}{\partial R_{tj}^2}=-\frac{1}{R_{tj}^2 ln\ (2)}; \qquad \frac{\partial^2 f}{\partial G_{tj}^2}=\frac{1}{G_{tj}^2 ln\ (2)}; \qquad \frac{\partial^2 f}{\partial R_{tj}\partial G_{tj}}=0.
\end{aligned}$$

Assuming that $\mathbb{E}(R_{tj})$ and $\mathbb{E}(G_{tj})$ are non-zero, an approximation of $\mathbb{E}(M_{tj})=\mathbb{E}(\log_2(R_{tj})-\log_2(G_{tj}))$ can be obtained by using its second-order Taylor's expansion:

$$\begin{aligned}
\mathbb{E}(M_{tj})&=\mathbb{E}(\log_2(R_{tj})-\log_2(G_{tj}))\\
&\approx\log_2(\mathbb{E}(R_{tj}))-\log_2(\mathbb{E}(G_{tj}))+\frac{1}{2}\left(-\frac{Var\ (R_{tj})}{ln\ (2)\mathbb{E}^2(R_{tj})}+\frac{Var\ (G_{tj})}{ln\ (2)\mathbb{E}^2(G_{tj})}\right)\\
&=\log_2(\mathbb{E}(R_{tj}))-\log_2(\mathbb{E}(G_{tj}))+\frac{1}{2ln\ (2)}\left(-\frac{Var\ (R_{tj})}{\mathbb{E}^2(R_{tj})}+\frac{Var\ (G_{tj})}{\mathbb{E}^2(G_{tj})}\right).
\end{aligned}$$

Let the non-zero background-corrected signals be estimators for the expected values of the foreground signals, i.e.,

$$\hat{\mathbb{E}}(R_{tj})=\bar{R}_{tc},\ \text{with}\ \bar{R}_{tc}\neq 0,$$

$$\hat{\mathbb{E}}(G_{tj})=\bar{G}_{tc},\ \text{with}\ \bar{G}_{tc}\neq 0.$$

Denote the sample variance estimators, obtained across the pixel intensities within each spot, as $\hat{\sigma}^2(R_t)$ (for the test channel) and $\hat{\sigma}^2(G_t)$ (for the control channel). Also, assume that these estimators do not depend on thebackground correction. We can derive an estimator for $\mathbb{E}(M_{tj})$ as follows:

$$\tilde{M}_t\doteq\log_2(\bar{R}_{tc})-\log_2(\bar{G}_{tc})+\frac{1}{2ln\ (2)}\left(-\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}\right).$$

Since

$$A_{tj}=f(R_{tj},G_{tj})\doteq\frac{\log_2(R_{tj})+\log_2(G_{tj})}{2},$$

we can estimate $\mathbb{E}(A_{tj})$ in a similar way to $\mathbb{E}(M_{tj})$. The first and second derivatives of A_{tj} are:

$$\frac{\partial f}{\partial R_{tj}}=\frac{1}{2ln\ (2)R_{tj}};\qquad \frac{\partial f}{\partial G_{tj}}=\frac{1}{2ln\ (2)G_{tj}};$$
$$\frac{\partial^2 f}{\partial R_{tj}^2}=-\frac{1}{2ln\ (2)R_{tj}^2};\qquad \frac{\partial^2 f}{\partial G_{tj}^2}=-\frac{1}{2ln\ (2)G_{tj}^2};\qquad \frac{\partial^2 f}{\partial R_{tj}\partial G_{tj}}=0,$$

An approximation of $\mathbb{E}(A_{tj})$ is obtained by using its second-order Taylor's expansion:

$$\mathbb{E}(A_{tj})\approx\frac{1}{2}(\log_2(\mathbb{E}(R_{tj}))+\log_2(\mathbb{E}(G_{tj})))+\frac{1}{2}\left(-\frac{Var\ (R_{tj})}{2ln\ (2)\mathbb{E}^2(R_{tj})}-\frac{Var\ (G_{tj})}{2ln\ (2)\mathbb{E}^2\left(G_{tj}\right)}\right)$$
$$=\frac{1}{2}\left(\log_2(\mathbb{E}(R_{tj}))+\log_2(\mathbb{E}(G_{tj}))\right)-\frac{1}{4ln\ (2)}\left(\frac{Var\ (R_{tj})}{\mathbb{E}^2(R_{tj})}+\frac{Var\ (G_{tj})}{\mathbb{E}^2(G_{tj})}\right).$$

Considering the sample estimators of the expected values and variances of R_{tj} and G_{tj}, we can derive the following estimator for $\mathbb{E}(A_{tj})$:

$$\tilde{A}_t\doteq\frac{1}{2}\left(\log_2(\bar{R}_{tc})+\log_2(\bar{G}_{tc})\right)-\frac{1}{4ln\ (2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}\right).$$

A.2.Estimation of *Var* (M_{tj}) and *Var* (A_{tj}) by the Delta Method

We can derive an estimator for Var (f (R$_{tj}$, G$_{tj}$)) by computing the variance of the first-order Taylor's expansion of f (R$_{tj}$, G$_{tj}$) at $(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))$:

$$\text{Var}\ (f(R_{tj},G_{tj}))\approx\left(\frac{\partial f}{\partial R_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\right)^2\text{Var}\ (R_{tj})+$$
$$\left(\frac{\partial f}{\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\right)^2\text{Var}\ (G_{tj})+$$
$$2\left(\frac{\partial f}{\partial R_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\right)\left(\frac{\partial f}{\partial G_{tj}}(\mathbb{E}(R_{tj}),\mathbb{E}(G_{tj}))\right)C\text{ov}\ (R_{tj},G_{tj}).$$

The second-order term was not considered because $\text{Var}\ (R_{tj}^2)$ and $\text{Var}\ (G_{tj}^2)$ cannot be usually estimated.

Since $M_{tj}=f(R_{tj},G_{tj})\doteq\log_2(R_{tj})-\log_2(G_{tj})$, with the first and second derivative showed in Appendix 5, we can obtain an approximation of Var (M_{tj}) as follows:

$$\text{Var}\ (M_{tj})\approx\left(\frac{1}{ln\ (2)\mathbb{E}(R_{tj})}\right)^2\text{Var}\ (R_{tj})+\left(-\frac{1}{ln\ (2)\mathbb{E}(G_{tj})}\right)^2\text{Var}\ (G_{tj})+$$
$$2\left(\frac{1}{ln\ (2)\mathbb{E}(R_{tj})}\right)\left(-\frac{1}{ln\ (2)\mathbb{E}(G_{tj})}\right)C\text{ov}\ (R_{tj},G_{tj})$$
$$=\frac{1}{\ln^2(2)}\left(\frac{Var\ (R_{tj})}{\mathbb{E}^2(R_{tj})}+\frac{Var\ (G_{tj})}{\mathbb{E}^2(G_{tj})}-2\frac{Cov\ (R_{tj},G_{tj})}{\mathbb{E}(R_{tj})\mathbb{E}(G_{tj})}\right).$$

Consider the sample estimators of the expected values of R_{tj} and G_{tj}, denoted by, respectively, $\bar{R}_{tc}$ and $\bar{G}_{tc}$, and assume that they are non-zero. Also, consider their variance and covariance sample estimators, denoted by, respectively, $\hat{\sigma}^2(R_t)$, $\hat{\sigma}^2(G_t)$, and $\hat{\sigma}(R_t,G_t)$, and assume that they are independent of the background correction. We can derive the following estimator for Var (M_{tj}) :

$$\hat{\sigma}^2(M_t)\doteq\frac{1}{\ln^2(2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}-2\frac{\hat{\sigma}(R_t,G_t)}{\bar{R}_{tc}\bar{G}_{tc}}\right).$$

Considering that A_{tj} is defined by the function

$$f(R_{tj},G_{tj})\doteq\frac{\log_2(R_{tj})+\log_2(G_{tj})}{2},$$

we can estimate Var (A_{tj}) in a similar way to Var (M_{tj}).

By using the first and second derivatives of A$_{tj}$, which are showed in Appendix (Barrett et al., 2012), we obtain the following approximation of Var (A_{tj}):

$$\text{Var}\ (A_{tj})\approx\left(\frac{1}{2\ ln\ (2)\mathbb{E}(R_{tj})}\right)^2\text{Var}\ (R_{tj})+\left(-\frac{1}{2\ ln\ (2)\mathbb{E}(G_{tj})}\right)^2\text{Var}\ (G_{tj})+$$
$$2\left(\frac{1}{2\ ln\ (2)\mathbb{E}(R_{tj})}\right)\left(-\frac{1}{2\ ln\ (2)\mathbb{E}(G_{tj})}\right)C\text{ov}\ (R_{tj},Gt_{tj})$$
$$=\frac{1}{4\ ln^2(2)}\left(\frac{Var\ (R_{tj})}{\mathbb{E}^2(R_{tj})}+\frac{Var\ (G_{tj})}{\mathbb{E}^2(G_{tj})}+2\frac{C\text{ov}\ (R_{tj},G_{tj})}{\mathbb{E}(R_{tj})\mathbb{E}(G_{tj})}\right).$$

Rewriting the above expression using the sample estimators for the expected value, variance and covariance of R_{tj} and G_{tj}, we derive the following estimator for Var (A_{tj}) :

$$\hat{\sigma}^2(A_t)\doteq\frac{1}{4\ln^2(2)}\left(\frac{\hat{\sigma}^2(R_t)}{\bar{R}_{tc}^2}+\frac{\hat{\sigma}^2(G_t)}{\bar{G}_{tc}^2}+2\frac{\hat{\sigma}\left(R_t,G_t\right)}{\bar{R}_{tc}\bar{G}_{tc}}\right).$$

ORIGINAL RESEARCH
published: 25 September 2019
doi: 10.3389/fgene.2019.00828

The Construction and Comprehensive Analysis of ceRNA Networks and Tumor-Infiltrating Immune Cells in Bone Metastatic Melanoma

Runzhi Huang[1,2,3†], Zhiwei Zeng[1†], Guangyu Li[1†], Dianwen Song[4], Penghui Yan[1], Huabin Yin[4], Peng Hu[1], Xiaolong Zhu[1], Ruizhi Chang[1], Xu Zhang[1], Jie Zhang[5], Tong Meng[2,3,4]* and Zongqiang Huang[1]**

[1] Department of Orthopaedics, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2] Division of Spine, Department of Orthopedics, Tongji Hospital affiliated to Tongji University School of Medicine, Shanghai, China, [3] Tongji University School of Medicine, Tongji University, Shanghai, China, [4] Department of Orthopedics, Shanghai General Hospital, School of Medicine, Shanghai Jiaotong University, Shanghai, China, [5] Shanghai East Hospital, Key Laboratory of Arrhythmias, Ministry of Education, Tongji University School of Medicine, Shanghai, China

Background/Aims: As a malignant and melanocytic tumor, cutaneous melanoma is the devastating skin tumor with high rates of recurrence and metastasis. Bone is the common metastatic location, and bone metastasis may result in pathologic fracture, neurologic damage, and severe bone pain. Although metastatic melanoma was reported to get benefits from immunotherapy, molecular mechanisms and immune microenviroment underlying the melanoma bone metastasis and prognostic factors are still unknown.

Methods: Gene expression profiling of 112 samples, including 104 primary melanomas and 8 bone metastatic melanomas from The Cancer Genome Atlas database, was assayed to construct a ceRNA network associated with bone metastases. Besides, we detected the fraction of 22 immune cell types in melanoma *via* the algorithm of "cell type identification by estimating relative subsets of RNA transcripts (CIBERSORT)." Based on the significant ceRNAs or immune cells, we constructed nomograms to predict the prognosis of patients with melanoma. Ultimately, correlation analysis was implemented to discover the relationship between the significant ceRNA and immune cells to reveal the potential signaling pathways.

Results: We constructed a ceRNA network based on the interaction among 8 pairs of long noncoding RNA–microRNA and 15 pairs of microRNA–mRNA. CIBERSORT and ceRNA integration analysis discovered that AL118506.1 has both significant prognostic value ($P = 0.002$) and high correlation with T follicular helper cells ($P = 0.033$). Meanwhile, T cells CD8 and macrophages M2 were negatively correlated ($P < 0.001$). Moreover, we constructed two satisfactory nomograms (area under curve of 3-year survival: 0.899; 5-year survival: 0.885; and concordance index: 0.780) with significant ceRNAs or immune cells, to predict the prognosis of patients.

OPEN ACCESS

Edited by:
*Liang Cheng,
Harbin Medical University,
China*

Reviewed by:
*Chuan-xing Li,
Karolinska Institute (KI), Sweden
Dapeng Hao,
Baylor College of Medicine,
United States*

***Correspondence:**
*Jie Zhang
jiezhang@tongji.edu.cn
Tong Meng
mengtong@medmail.com.cn
Zongqiang Huang
gzhuangzq@163.com*

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *26 June 2019*
Accepted: *12 August 2019*
Published: *25 September 2019*

Citation:
Huang R, Zeng Z, Li G, Song D, Yan P, Yin H, Hu P, Zhu X, Chang R, Zhang X, Zhang J, Meng T and Huang Z (2019) The Construction and Comprehensive Analysis of ceRNA Networks and Tumor-Infiltrating Immune Cells in Bone Metastatic Melanoma. Front. Genet. 10:828. doi: 10.3389/fgene.2019.00828

Conclusions: In this study, we suggest that bone metastasis in melanoma might be related to AL118506.1 and its role in regulating thrombospondin 2 and T follicular helper cells. Two nomograms were constructed to predict the prognosis of patients with melanoma and demonstrated their value in improving the personalized management.

Keywords: melanoma, bone metastasis, competing endogenous RNA network, immune cell, nomogram

INTRODUCTION

Cutaneous melanoma is a malignant, melanocytic tumor and considered as the most harmful skin cancer (Cymerman et al., 2016; Lombard et al., 2019). All over the world, it accounts for about 232,100 (1.7%) cases of all newly diagnosed primary malignant cancers (excluding nonmelanoma), and meanwhile approximately 55,500 (0.7%) deaths are derived from cutaneous melanoma each year (Schadendorf et al., 2018). Nowadays, its incidence rate is still escalating dramatically (Schadendorf et al., 2019).

Extensive local resection with clean margins, depending on Breslow thickness of the tumor tissue, is recommended as the primary treatment for localized disease [The Cancer Genome Atlas (TCGA), 2015]. However, distant metastases often occur even after complete tumor resection due to the aggressive nature. Bone is the common metastatic location, and bone metastasis often results in pathologic fracture, neurologic damage, and severe bone pain, which decreases the quality of life (Braeuer et al., 2014; Bier et al., 2016). Regarding some patients with metastasis, systemic therapies such as targeted therapy and immunotherapy have achieved promising survival outcome; however, prognosis remains poor in most patients with metastasis (Bostel et al., 2016). Hence, it is in a desperate need to explore the molecular mechanism and probe for the prognostic factors for cutaneous melanoma patients with bone metastasis. The relationship among microRNA (miRNA), long noncoding RNA (lncRNA), and mRNA, known as ceRNA networks, had been explored in many diseases. However, ceRNA network mechanism underlying melanoma and bone metastasis still remains unknown.

In this study, we constructed a ceRNA network based on the gene expression profiling retrieved from the TCGA database to identify the ceRNAs associated with melanoma and bone metastasis. Besides, we perform "The Cell Type Identification by Estimating Relative Subsets of RNA Transcripts algorithm (CIBERSORT)" algorithm to detect the immune cells and their proportions in tumor tissues of melanoma. Additionally, nomograms were developed to predict the prognosis of melanoma with bone metastasis based on significant immune cells and ceRNA. The relationship between bone metastasis–related immune cells and ceRNA networks was evaluated to identify the underlying signaling pathways.

MATERIALS AND METHODS

Data Collection and Differential Gene Expression Analysis

The Ethics Committee of the First Affiliated Hospital of Zhengzhou University approved this study (no. 2019-KY-107). We downloaded the RNA profiles of the primary melanomas and bone metastasis samples from the TCGA (https://tcga-data.nci.nih.gov/tcga/) database. HTseq-count and fragments per kilobase of exon per million reads mapped profiles of 112 samples, including 104 primary melanomas and 8 melanomas with bone metastasis, were assembled. Meanwhile, demographic and survival information of each patient was collected. The edgeR method was used to find differentially expressed mRNAs, lncRNAs, and miRNAs after removing nonmelanoma-specific expression genes (no expression in both the experimental group and control group). Only when the false discovery rate (FDR) $P < 0.05$ and the log (fold change) > 1.0 or <−1.0 could be regarded as differentially expressed gene of downregulation and upregulation, respectively.

The Construction of the ceRNA Network

Prior to the initial statistical analysis, the miRNA–mRNA and lncRNA–miRNA interaction data were retrieved from miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) (Chou et al., 2018) and Incbase v.2 Experimental Module (http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex-experimental) (Paraskevopoulou et al., 2016), respectively. Afterward, miRNAs, which illustrate significant outcomes in the aspect of regulating both IncRNAs and mRNAs in hypergeometric testing and correlation analysis, were collected for establishing the ceRNA network by Cytoscape v.3.5.1 (Shannon et al., 2003).

Survival Analysis and Nomograms of Key Members in the ceRNA Network

Kaplan–Meier (K-M) survival analysis was performed to show the relationship between the expression level of biomarkers with the prognostic value illustrated in the ceRNA network and survival outcomes in patients with melanoma. Afterward, the significant biomarkers were incorporated into the reduced Cox proportional hazards model by screening the significant variables in the initial Cox models to illustrate the variables with prognostic values. Besides, Lasso regression (least absolute shrinkage and selection operator regression), which is a kind of linear regression using shrinkage where data values are shrunk to a specific point, was implemented to confirm

Abbreviations: AUC, Area under curve; ceRNA, competitive endogenous RNA; lncRNA, long noncoding RNA; miRNA, microRNA; CIBERSORT, Cell type identification by estimating relative subsets of RNA transcripts; TCGA, The Cancer Genome Atlas; FDR, false discovery rate; SD, standard deviation; ROC, Receiver operating characteristic curves; THBS, Thrombospondin, Tfh, T follicular helper cells; IL-21, interleukin 21.

the fitness of the established multifactor models. Ultimately, a nomogram based on the multivariable models was developed to predict the prognosis of patients with melanoma. In accordance with the expression level of biomarkers with prognostic values, we can acquire the points of each biomarker and add up to obtain the total points, which can display the 3- and 5-year overall survival probability. Meanwhile, calibration curves and receiver operating characteristic (ROC) curves were performed to evaluate the discrimination and precision of the nomogram.

CIBERSORT Estimation

CIBERSORT is an analytical tool constructed by Newman et al. (2015) to identify the richness and proportion of the diversified cell types in a mixed cell population using gene expression data. Every cell type and their quantity in each sample can be conveniently acquired *via* CIBERSORT estimation. In this study, we use CIBERSORT algorithm to further probe for the cytological causes of molecular mechanisms of the pivotal biomarkers in the ceRNA network. The proportions of 22 immune cell types in the primary melanoma and melanoma with bone metastasis were estimated by CIBERSORT. Only when the CIBERSORT output of P < 0.05 could put the samples into further analysis. The Wilcoxon rank-sum test was performed to look for the significant immune cells in the aspect of the fraction between the primary melanoma and melanoma with bone metastasis. Then, K-M survival analysis was used to demonstrate the relationship between the overall survival of melanoma patients and proportion of specific immune cells. After being well filtered by Lasso regression, specific immune cells were incorporated into the Cox proportional hazards model. Then, nomogram was constructed to predict the prognosis for melanoma. Concordance index of Cox model was applied to access the discrimination and accuracy of the nomogram. Ultimately, Pearson correlation analysis was carried out to show the relationship between immune cells and biomarkers.

Online Database Validation

To minimize bias caused by the imbalanced sample size and get more complete annotation of key biomarkers, multiple online databases including the CellMarker (Zhang et al., 2019), LncRNA2Target (Cheng et al., 2019), Ontogene (Cheng et al., 2016), String (Szklarczyk et al., 2019), DincRNA (Cheng et al., 2018), SurvExpress (Aguirre-Gamboa et al., 2013), Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019), Genotype–Tissue Expression (GTEx) (Consortium, 2015), Oncomine (Elfilali et al., 2006), and Gene Expression Omnibus (GEO) (ID: GSE19234 (Bogunovic et al., 2009), GSE22153 (Jonsson et al., 2010) were used to detect gene expression levels of key biomarkers at the tissue and cellular levels.

Statistical Analysis

Only two-sided $P < 0.05$ was defined as statistical significance. All the statistical analyses were performed with R version 3.5.1 software (Institute for Statistics and Mathematics, Vienna, Austria; www.r-project.org) (package: GDCRNATools (Li et al., 2018), edgeR, ggplot2, rms, glmnet, preprocessCore, survminer, timeROC).

RESULTS

Identification of Significantly Differentially Expressed Genes

Figure 1 illustrates the analysis process of this study. The baseline features of all the patients retrieved from the TCGA database were described in **Table S1**. We defined the log (fold change) >1.0 or < −1.0 and FDR <0.05 as the critical point and found out that there were 701 differentially (550 down- and 151 up-) expressed

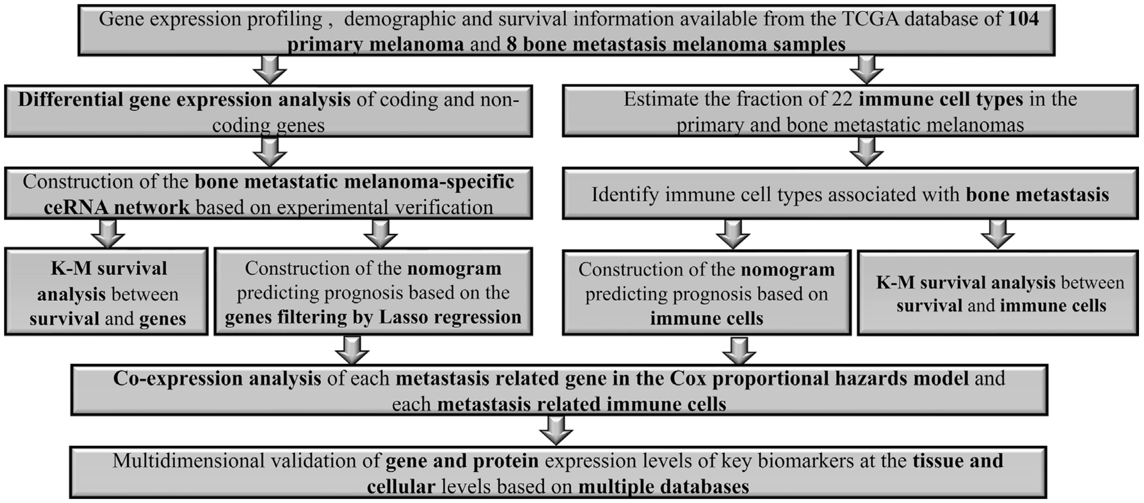

Gene expression profiling , demographic and survival information available from the TCGA database of 104 primary melanoma and 8 bone metastasis melanoma samples
Differential gene expression analysis of coding and non-coding genes
Estimate the fraction of 22 immune cell types in the primary and bone metastatic melanomas
Construction of the bone metastatic melanoma-specific ceRNA network based on experimental verification
Identify immune cell types associated with bone metastasis
K-M survival analysis between survival and genes
Construction of the nomogram predicting prognosis based on the genes filtering by Lasso regression
Construction of the nomogram predicting prognosis based on immune cells
K-M survival analysis between survival and immune cells
Co-expression analysis of each metastasis related gene in the Cox proportional hazards model and each metastasis related immune cells
Multidimensional validation of gene and protein expression levels of key biomarkers at the tissue and cellular levels based on multiple databases

FIGURE 1 | The flowchart of the analysis process.

protein-coding genes, along with 14 differentially (5 down- and 9 up-) expressed lncRNAs and 72 differentially (45 down- and 27 up-) expressed miRNAs between the bone metastatic melanoma and the primary melanoma from the TCGA database (**Figures 2A–F**).

ceRNA Network Establishment and Survival Analysis

A ceRNA network was established based on the interaction among 8 pairs of lncRNA–miRNA and 15 pairs of miRNA–mRNA (**Figure 3A**) (**Table 1**). Kaplan–Meier survival analysis was implemented to explore the relationship between the prognosis and biomarkers involved in ceRNA network related to the bone metastasis in melanoma. The results revealed that thrombospondin 2 (THBS2) (*P* = 0.040) and AL118506.1 (*P* = 0.002) displayed significance (**Figures 3B**, **C**). According to enrichment analysis, the significant genes associated with bone metastasis in melanoma were mostly functioned in extracellular matrix organization (**Figure S1**).

Construction of the Prediction Model Based on the ceRNA Network

The outcomes of Lasso regression illustrated that four genes, hsa-miR-137, hsa-miR-425-5p, VCAN, and AL118506.1, were critical to modeling and were then incorporated into the Cox regression, after which the nomogram, aimed to predict the prognosis, was constructed according to the Lasso regression. The areas under curve (AUC) of the 3- and 5-year survival were

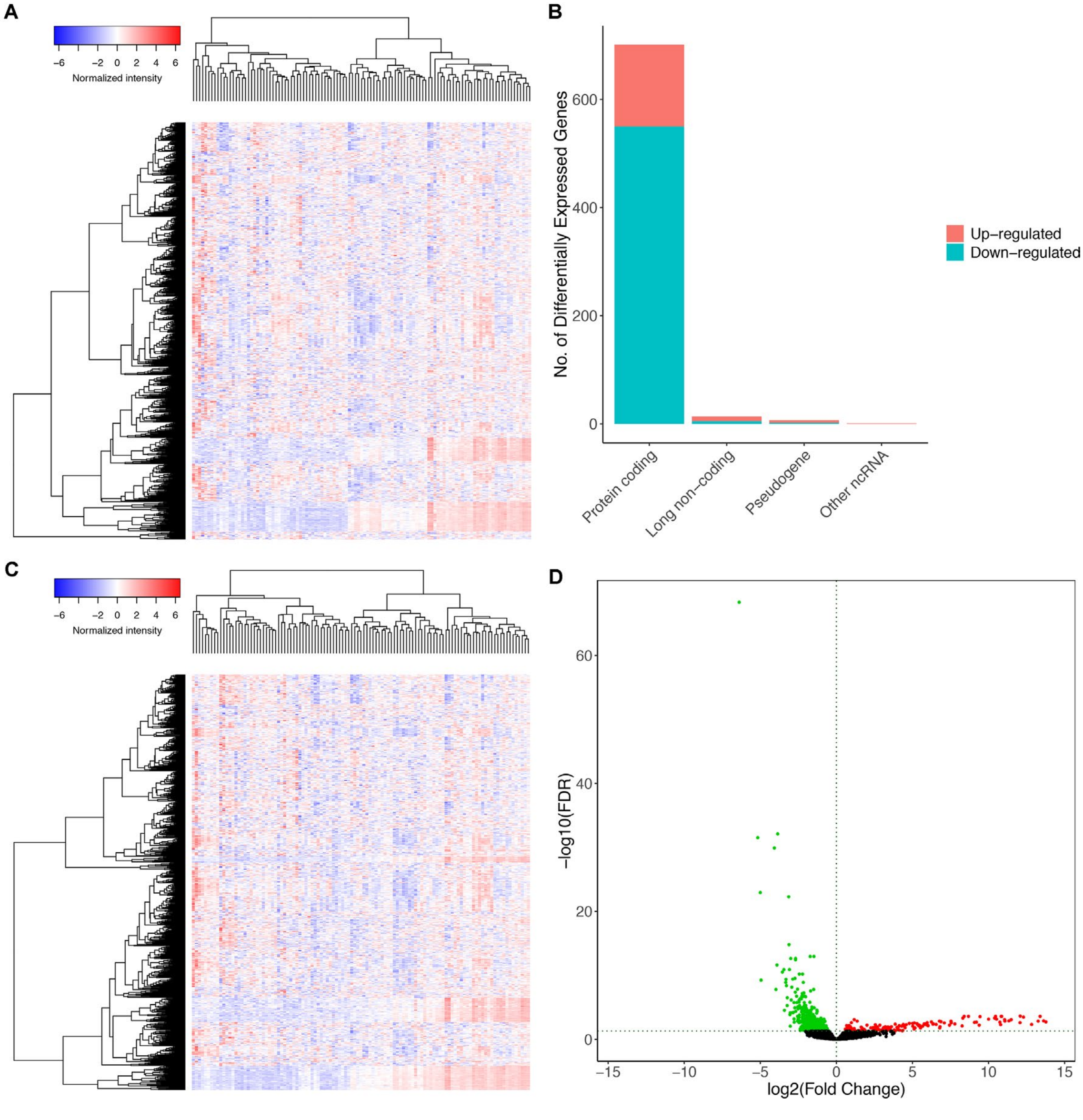


FIGURE 2 | Continued

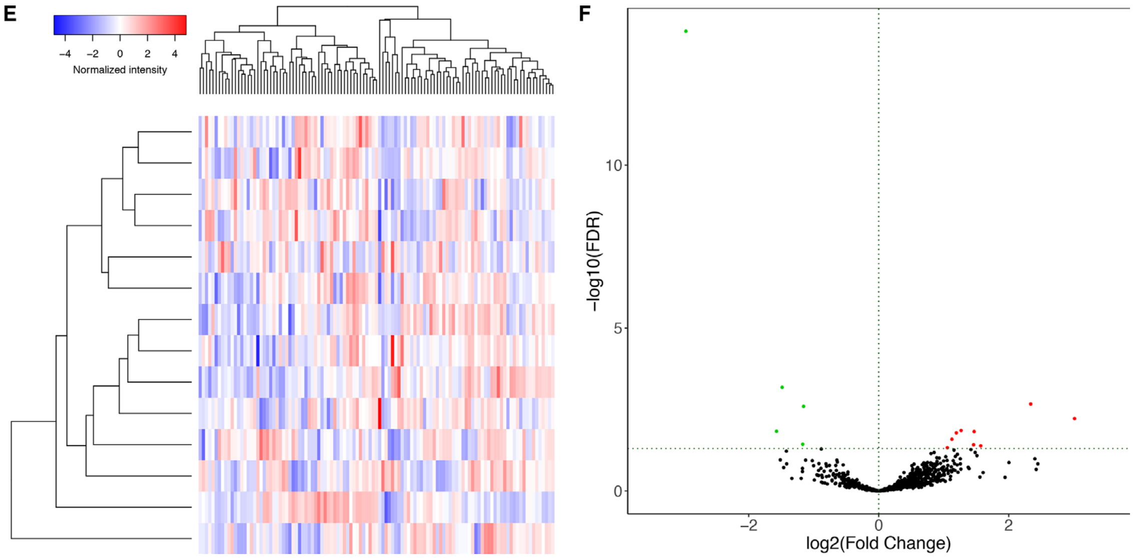


FIGURE 2 | The heat maps of differentially expressed **(A)** RNAs, **(C)** miRNAs, **(E)** lncRNAs between the bone-metastatic melanoma and the primary melanoma. **(B)** Bar plot showing differentially expressed protein-coding genes, long noncoding genes, pseudogenes, and other RNAs. Red and blue represent up-regulated and down-regulated RNAs, respectively. It shows that 550 of 701 differentially expressed protein-coding genes are down-regulated and 151 are up-regulated. Besides, among 14 differentially expressed lncRNAs, 5 lncRNAs are down-regulated, and 9 are up-regulated. Volcano plots of differentially expressed mRNAs **(D)** and lncRNAs **(F)**. We defined the log (fold change) >1.0 or <—1.0 and FDR <0.05 as the critical point. Thus, the red and blue dots in the plots represent high and low expression RNAs with statistical significance, respectively. Meanwhile, black dots represent mRNAs and lncRNAs without statistical significance between the primary and the bone-metastatic melanoma.

0.899 and 0.855, respectively, which reflects the satisfactory accuracy. Additionally, the discrimination of the nomogram was suggested by the calibration curves (**Figures 4A–F**).

Immune Cells Related to the Melanoma

The composition of the immune cells in the melanoma evaluated by CIBERSORT algorithm was illustrated in the histogram and the heat map (**Figures 5A**, **B**). The results of the Wilcoxon rank-sum test revealed that the proportion of the T follicular helper (Tfh) cells in the melanoma with bone metastasis was relatively less than that in the primary melanoma (P = 0.021), and macrophages M2 was relatively greater in the melanoma with bone metastasis (P = 0.036) (**Figure 5C**).

Construction of the Prediction Model Based on the Immune Cells

Similarly, 16 of 22 immune cells, which showed significant prognostic values in the initial Cox regression model, were integrated into the final multivariable model with satisfactory predictive power (concordance index 0.780) and were utilized to construct the nomogram (**Figures 6A**, **B**). The concordance curve and concordance index showed a good concordance of the model (**Figure 6C**). Based on the result of the Kolmogorov–Smirnov test, the fraction of regulatory T cells (Tregs) in stages T1, T2, T3, and T4 showed significant difference between patients with or without bone metastasis (**Figure S2**).

Comprehensive Analysis of Genes and Immune Cells

Correlation analysis (Pearson analysis) was applied to demonstrate the coexpression patterns among diversified immune cells (**Figure 7A**). Likewise, correlation relationship (Pearson analysis) between immune cells and biomarkers was further analyzed and illustrated (**Figure 7B**). As shown, hsa-miR-425-5p and Tfh cells (P = 0.019, R = 0.260) (**Figure 7C**), AL118506.1 and Tfh cells (P = 0.033, R = −0.240) (**Figure 7D**), and Tfh cells and hsa-miR-425-5p (**Figure S3**) represented good correlation. Eventually, bone metastasis–specific immune cells and ceRNAs significantly associated with prognosis were integrated into one multivariable model and one nomogram (**Figure S4**), which could decently predict the prognosis of SKCM (AUC of 3-year survival: 1.000; AUC of 5-year survival: 1.000). However, the model diagnostic information suggested that the prediction model had bias due to the small sample size.

Metastasis-Specific ceRNAs and Immune Cells' Surface Markers Coding Genes Showing Significant Results in Multidimensional Validation

In order to explore the expressions of metastasis-specific ceRNAs and immune cells' surface markers coding genes in different datasets, a dimensional validation applying multiple online databases was performed.

At the cellular level, BCL6 transcription repressor (BCL6), membrane metalloendopeptidase (MME), C-X-C motif

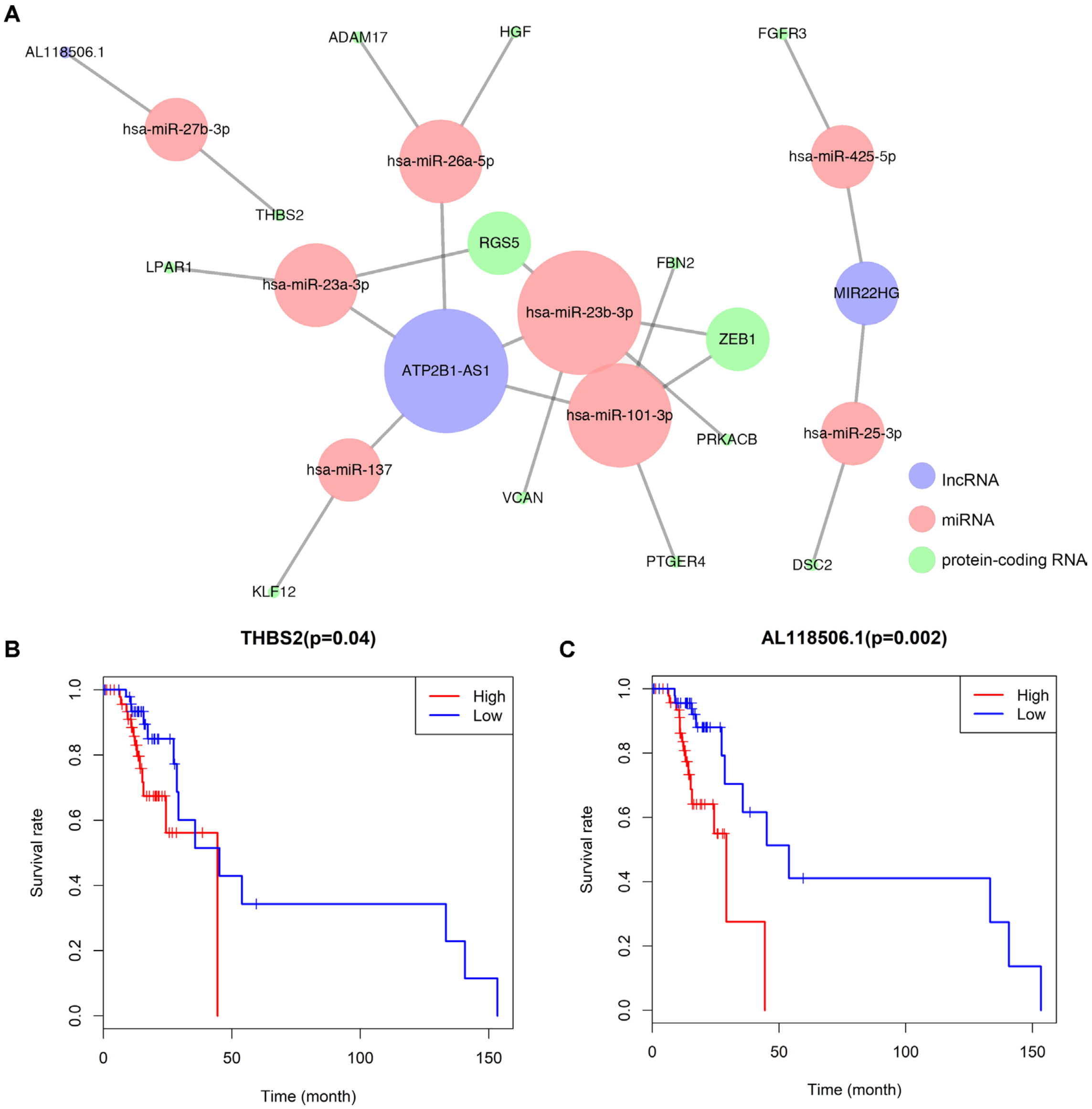


FIGURE 3 | (A) Overview of the lncRNA–miRNA–mRNA ceRNA network of melanoma with 8pairs of lncRNA–miRNA and 15 pairs of miRNA–mRNA. Red balls represent miRNAs, blue balls represent lncRNAs, and green balls represent protein-coding RNAs. Kaplan–Meier survival curves based on the expression of biomarkers involved in ceRNA network related to the bone metastasis in melanoma shows that **(B)** THBS2 (*P* = 0.040) and **(C)** AL118506.1 (*P* = 0.002) had significantly prognostic values.

chemokine ligand 13 (CXCL13), inducible T-cell costimulator (ICOS), and programmed cell death 1 (PDCD1) had been reported as the surface markers of Tfh cell in the CellMarker (**Figure S5**). AL118506.1 is a type of lncRNA (Ensemble ID: ENSG00000268858). According to DincRNA, Ontogene, and LncRNA2Target database, AL118506.1 is the antisense to Abhydrolase domain containing 16B (ABHD16B, also known as C20orf135), and it can down-regulate the expression level of hsa-miR-27b-3p. However, the function of AL118506.1 remains largely unknown. Thus, AL118506.1, ABHD16B, THBS2, BCL6, MME, CXCL13, ICOS, and PDCD1 were incorporated into further multidimensional validation.

First, **Figure S6** illustrates the protein–protein interaction network of these genes, indicating that there are many interactions between THBS2 protein and T infertile helper cell's surface markers. Besides, in the CCLE and GTEx, we found that THBS2 was expressed in various SKCM cell lines, and Tfh cell's surface marker coding gene expressions were low, while in normal skin tissue THBS2 and AL118506.1 were expressed, and surface marker coding gene expressions were also low (**Figures S7A**, **S7C**). Meanwhile, significant coexpression relationships between THBS2 and Tfh cell's surface marker coding genes had been observed in tissue levels, but not in cancer cell lines (**Figures S7B**, **S7D**). Besides, in meta-analysis of Oncomine,

TABLE 1 | Hypergeometric testing and correlation analysis results of ceRNAs network.

LncRNA	Protein-coding RNA	MiRNAs	Correlation *P*	Hypergeometric test *P*
AL118506.1	THBS2	hsa-miR-27b-3p	0.006581855	0.00747894
MIR22HG	FGFR3	hsa-miR-425-5p	0.022787186	0.006234399
MIR22HG	DSC2	hsa-miR-25-3p	0.000396455	0.001248439
ATP2B1-AS1	RGS5	hsa-miR-23a-3p,hsa-miR-23b-3p	2.58E–06	0.001872829
ATP2B1-AS1	FBN2	hsa-miR-101-3p	0.000158704	0.006866417
ATP2B1-AS1	KLF12	hsa-miR-137	0.009365181	0.020470827
ATP2B1-AS1	VCAN	hsa-miR-23b-3p	0.001403378	0.006866417
ATP2B1-AS1	LPAR1	hsa-miR-23a-3p	5.78E–09	0.020470827
ATP2B1-AS1	ZEB1	hsa-miR-101-3p,hsa-miR-23b-3p	2.43E–05	0.014703227
ATP2B1-AS1	HGF	hsa-miR-26a-5p	0.000176262	0.033905608
ATP2B1-AS1	PTGER4	hsa-miR-101-3p	0.016481894	0.006866417
ATP2B1-AS1	PRKACB	hsa-miR-23b-3p	9.06E–07	0.006866417
ATP2B1-AS1	ADAM17	hsa-miR-26a-5p	0.000901194	0.033905608

ceRNAs, competing endogenous RNAs; LncRNA, long noncoding RNA; MiRNA, microRNA.

THBS2 (Median rank 1,088, $P < 0.001$) (**Figures S8A, B**), ICOS (Median rank 1,008, COPA = 1.854) (**Figures S8C, D**), CXCL13 (Median rank 536.5, COPA = 30.145) (**Figures S8E, F**), BCL6 (Median rank 434.5, COPA = 2.016) (**Figures S8G, H**), MME (median rank 221.0, COPA = 8.940) (**Figures S8I, J**), and PDCD1 (median rank 7,680, $P = 0.350$) (**Figures S8C, D**) all showed significant results in multiple melanoma–related studies except PDCD1. Additionally, the reanalysis results of GSE19234 (**Figure S9**) and GSE22153 (**Figure S10**) in SurvExpress suggested that these genes have significant predictive value for metastasis (censoring event: metastasis, hazard ratio = 5.19 [95% confidence interval {CI}, 1.92–14.05], $P = 0.001$, **Figure S9C**) (censoring event: subcutaneous metastasis, hazard ratio = 4.01 [95% CI, 1.93–8.34], $P < 0.001$, **Figures S10C, D**) and prognosis (censoring event: overall death, hazard ratio = 3.15 [95% CI, 1.71–5.80], $P < 0.001$, **Figure S10B**).

DISCUSSION

Malignant melanoma is regarded as one of the most devastating and metastatic diseases with a drastic increasing incidence rate around the world (Bostel et al., 2016). Tumor metastasis is the advanced stage of disease and its complications often decrease the quality of life, especially for the bone metastasis. Although the mechanisms of tumorigenesis and metastasis are still unclear for melanoma, molecular and cellular features often changed during the process and are often viewed as important predictors (Braeuer et al., 2014; Rodina et al., 2016). Thus, the differentially expressed genes and tumor-infiltrating immune cells in the primary melanoma and bone metastasis attract our interest, which is seldom focused by previous studies.

In the current study, we first figured out the differently expressed and statistically significant ceRNA and tumor-infiltrating immune cells between the primary and metastatic melanoma. Afterward, two nomograms are constructed based on them to predict the outcomes of patients with melanoma. The high AUC value and concordance index in two nomograms might contribute to make an evaluation for bone metastasis and survival outcomes. At last, according to the results of K-M survival analysis and correlation analysis, we inferred that the ceRNA regulatory mechanism of AL18506.1 (lncRNA), THBS2 (mRNA), hsa-miR-27b-3p (miRNA), and Tfh cell might play a crucial role in bone metastasis of melanoma.

Recently, a myriad of studies had uncovered that no more than 2% of the whole genome encode protein-coding genes, which suggests that most of the human transcriptomes are represented by noncoding RNAs (Volders et al., 2013). mRNAs, miRNAs, and lncRNAs are connected through the competitive endogenous RNA networks in an intricate crosstalk (Tay et al., 2014). The interaction among miRNA, lncRNA, and mRNA, operating as ceRNA networks, had been drastically explored in many diseases, including lung cancer, gastric cancer, and gallbladder cancer, among others (Kumar et al., 2014; Chen et al., 2018; Chen et al., 2019). However, ceRNA network mechanism underlying melanoma and bone metastasis remains largely unknown. In our study, we identified that AL118506.1 (lncRNA) could down-regulate and up-regulate the level of hsa-miR-27b-3p and THBS2, respectively, to promote bone metastasis in patients with melanoma *via* ceRNA network. The role of hsa-miR-27b-3p was shown to be essential in malignant transformation, which is in conformity with our present study (Liu et al., 2015).

Thrombospondins (THBSs) had been verified to play important roles in various processes, including angiogenesis, cellular adhesion, extracellular matrix interaction, tumor formation, and metastasis (Roberts, 2008; Liu et al., 2018). Thrombospondin 2, one of members in THBSs, is revealed to regulate the antiangiogenic activity and prevent the development of focal adhesion in endothelial cells (Agostini et al., 2012). Moreover, the overexpression of THBS2 had been demonstrated to be positively correlated with node metastasis and over survival in many types of cancer, including colorectal adenocarcinoma, myxoid liposarcoma, prostate cancer, and gastric cancer (Kim et al., 2010; Slavin et al., 2014; Chang et al., 2016; Lin et al., 2016; Nezu et al., 2016; Zhuo et al., 2016; Qian et al., 2017; Wei et al., 2017). The role of THBS2 was also investigated in melanoma in a previous study, and metastatic uveal melanoma had a higher expression level of THBS2, which is consistent with our analysis (Liu and Ma, 2018).

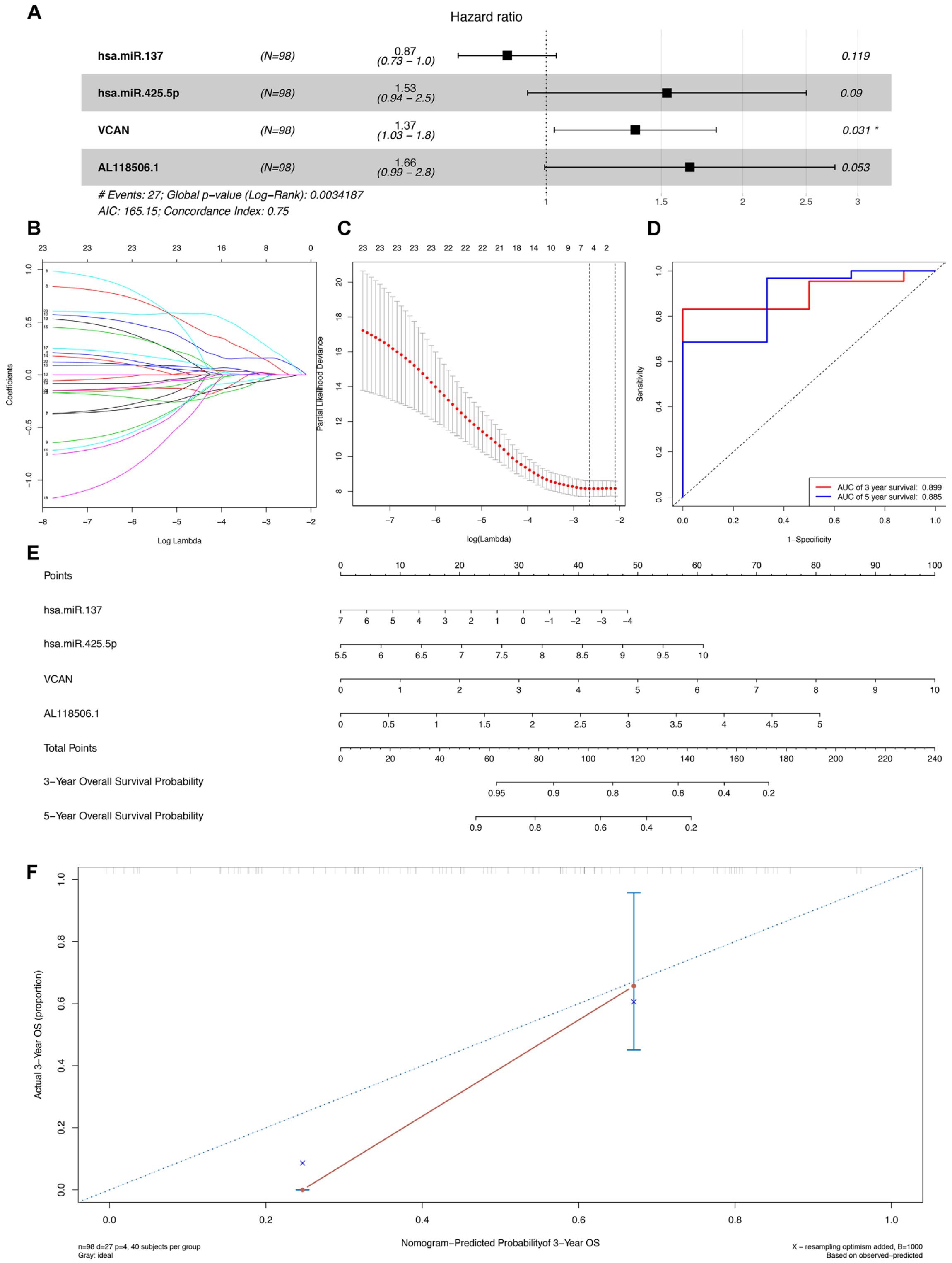


FIGURE 4 | (A) The Cox proportional hazards model based on RNAs selected by **(B) (C)** Lasso regression. hsa-miR-137, hsa-miR-425-5p, VCAN, and AL118506.1 are incorporated into the Cox proportional hazards model. **(E)** Nomogram for predicting patients' outcome based on RNAs (hsa-miR-137, hsa-miR-425-5p, VCAN, and AL118506.1) in Panel **(A)**. **(D)** ROC curves and **(F)** calibration curves for assessing the discrimination and accuracy of the nomogram. Besides, AUCs of the 3- and 5-year survival were 0.899 and 0.855, respectively. AUC, area under curve; ROC, receiver operating characteristic.

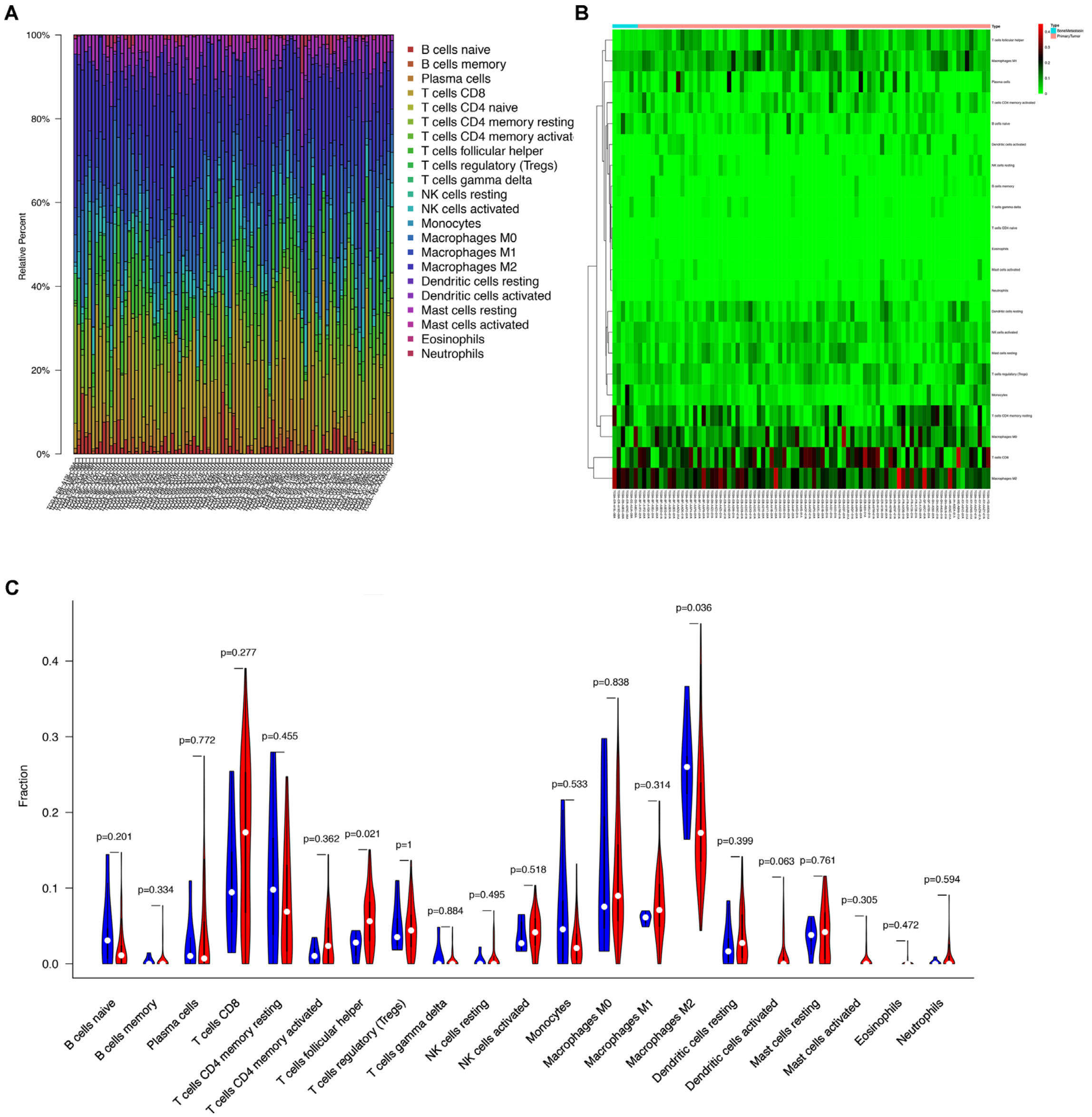


FIGURE 5 | (A) Bar plot showing cell types and relative percent in melanoma tissues. Different colors represent different cell types, which are listed in the right as *y* axis, while *x* axis represents different samples. **(B)** Heat map of tumor-infiltrating cells in tumor tissues in patients with the primary melanoma and the bone metastatic disease. Annotations on top show clustering of samples. While the blue represents the melanoma with bone metastasis, the red symbolizes the primary melanoma. **(C)** Violin plot for comparing cells' proportion between the primary and bone-metastatic disease. It illustrates that the proportion of the T follicular helper (Tfh) cells in the melanoma with bone metastasis was relatively less than that in the primary melanoma (*P* = 0.021), and macrophages M2 was relatively greater in the melanoma with bone metastasis (*P* = 0.036).

We also found out the different proportions of numerous immune cells in the primary melanoma and bone metastatic melanoma tissues. T follicular helper cells and macrophages M2 were demonstrated to be related to bone metastasis. The nomogram, composed of 16 kinds of immune cells, was constructed to predict the overall survival, which showed the great clinical utility with the concordance index of 0.78.

Generally, the CD8+ cytotoxic T cell is considered to be the main element of active antitumor immunity, whose full function greatly relied on adequate help from CD4+ T cells (Gillgrass et al., 2014). Naive CD4+ T cells could differentiate into different T helper (T_H) cells, including T_H1, T_H2, T_H17, Tregs, and Tfh cells (Zhu et al., 2010). The Tfh cell is one subtype of CD4+ T cells, which is defined by its surface phenotypes with the highest expression level of CXCR5(Vinuesa et al., 2016). It had been demonstrated that Tfh plays an important part in the construction of humoral immunity through regulating the formation and cellular reactions that happen in the germinal center (Qi, 2016). The dysregulated Tfh cells were found to be associated with several autoimmune or (and) immune-deficient diseases, including systemic lupus erythematosus, HIV, and lymphoma (Tangye et al., 2013). A few previous studies had revealed that there are ordered lymph node–like structures mainly

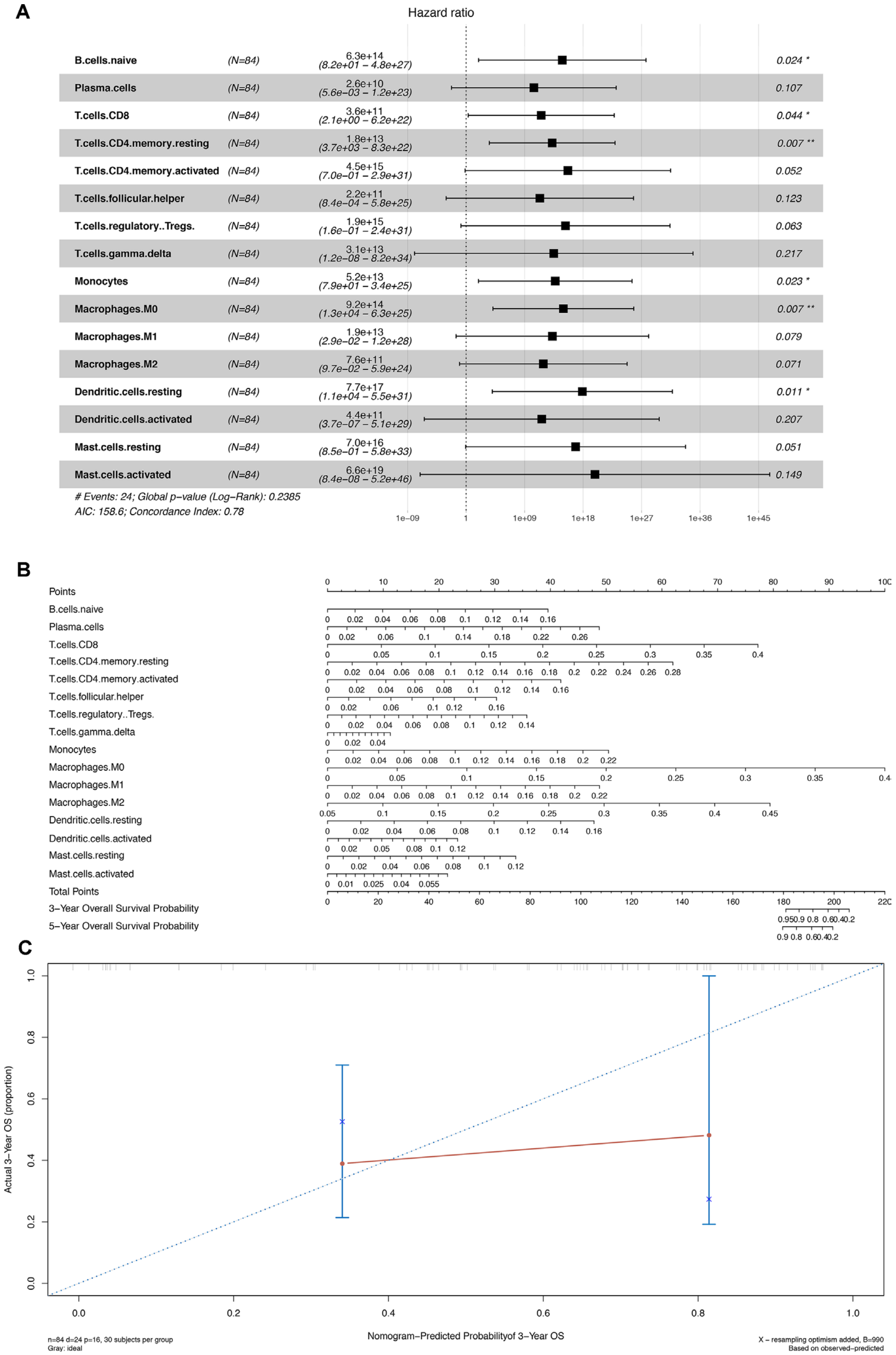


FIGURE 6 | (A) Cox proportional hazards model integrated by 16 different types of immune cells. **(B)** Nomogram for predicting patients' outcome based on 16 cells in Panel **(A)**. **(C)** Calibration curves for evaluating the accuracy of the nomogram. *$P < 0.05$; **$P < 0.001$.

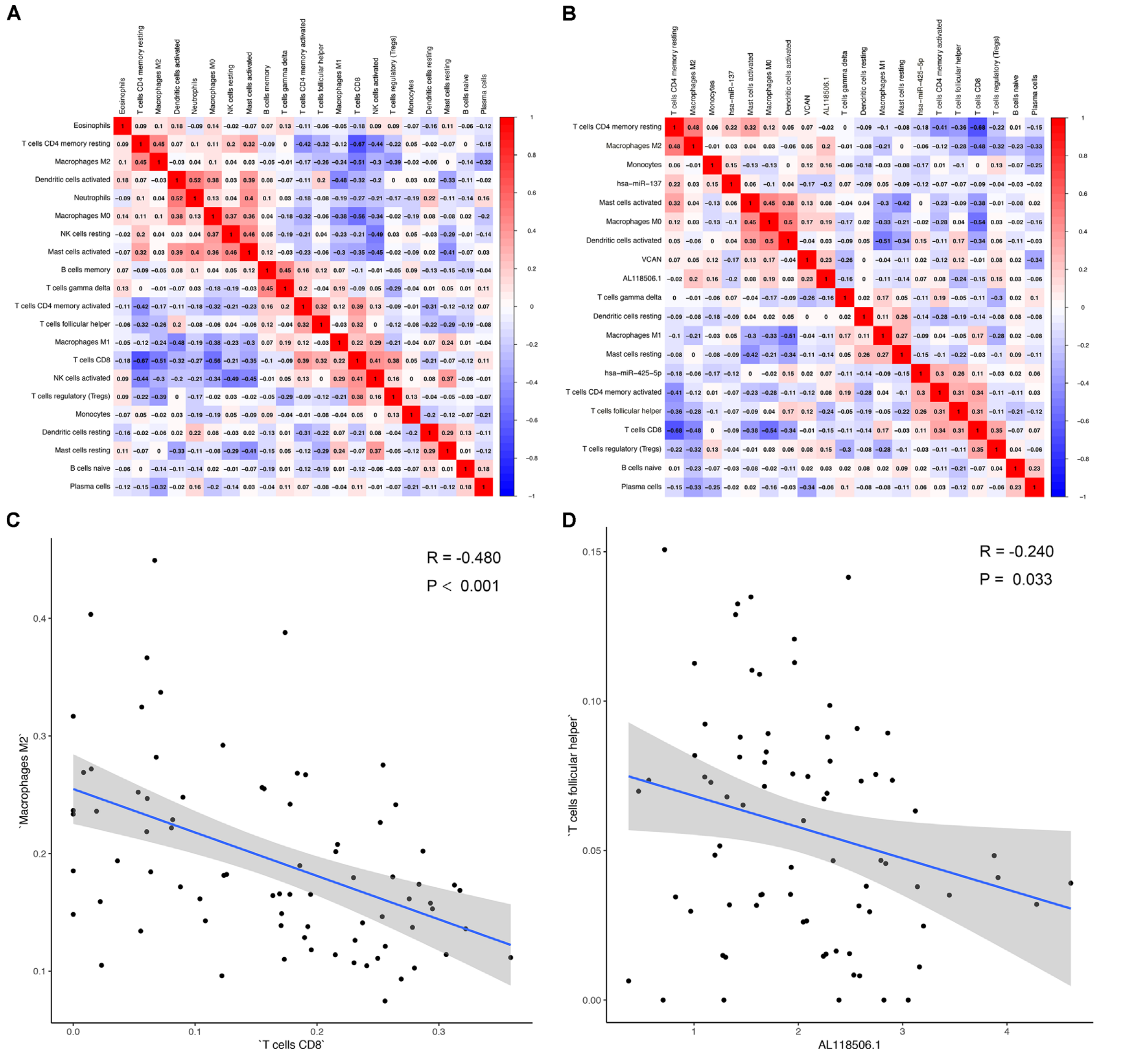


FIGURE 7 | (A) Correlation analysis (Pearson analysis) of different tumor-infiltrating cells and **(B)** the relationships between different tumor-infiltrating cells and differentially expressed genes in tumor tissues of melanoma. Scatterplots further illustrate the exact relationship between T cells CD8 and macrophages M2 (*P* < 0.001, *R* = −0.480) **(C)**, AL118506.1, and T follicular helper cells (*P* = 0.033, *R* = −0.240) **(D)**. Besides, gray-shaded areas in two graphs represent the standard errors of the blue regression lines. R, correlation coefficient.

formed by Tfh cells in extensively infiltrated tumors, including breast cancer, lung cancer, and colorectal cancer, with obviously detectable Tfh cells, which function in antitumor immunity with positive clinical outcome (Dieu-Nosjean et al., 2008; deLeeuw et al., 2012). Other human-related studies also identified that Tfh cells had great capacity in directly assisting B cells through releasing interleukin 21 (IL-21), and IL-21 could further help human antigen-specific cytotoxic T cells to generate and proliferate, which also suggests that Tfh cells had a direct antitumorigenic function (Chen et al., 2016). Thus, patients with fewer Tfh cells had a decreased immune response in fighting against tumor, while immunosuppression was positively correlated with tumor metastasis (Bidwell et al., 2012). In our study, our data indicate that Tfh cells had a lower expression level in patients with bone metastatic disease.

Similarly, the importance of CD4+ cells of high concentration in hindering melanoma metastasis and recurrence has also been reported (He et al., 2017). Antibody of anti–programmed death 1, situated on the surface of CD4+ cells, had been verified to prove the clinical outcomes of patients with melanoma (Yamaguchi et al., 2018). Additionally, the expression levels of tumor-infiltrating cells of CD8 and macrophages M2 are, to some extent, related to clinical outcomes. The extensively studied immune infiltrate in different cancer had established that macrophages M2 could suppress antitumor immunity and promote tumor progression (Gillgrass et al., 2014; Guerriero et al., 2017). The data presented in this study also showed that macrophages M2 expression is higher in samples of patients with bone metastasis. Furthermore, the correlation analysis led us to know that the level of macrophages

M2 was inversely correlated with that of CD8 T cells, and patients with more CD8 cells in tumor tissues had worse outcome, which was highly consistent with a previous study (Gillgrass et al., 2014).

The correlation analysis revealed that Tfh cells were associated with AL118506.1 ($R = -0.240$, $P = 0.033$). Based on the results of correlation analysis and hypergeometric testing of ceRNA network, AL118506.1 (lncRNA), THBS2 (protein-coding RNA), and hsa-miR-27b-3p (miRNA) were considerably correlated ($P = 0.007$). Therefore, we inferred that the interaction among hsa-miR-27b-3p, AL118506.1, THBS2, and Tfh cells was highly relevant with bone metastasis in patients with melanoma.

Nevertheless, there are several unavoidable limitations to our study that should be taken into consideration. First, the quantity of related data available from the public datasets is still limited. The idea of acquiring the same number of cases in the aspects of different genders, age groups, and races, among others, to decrease the potential error and bias is far too difficult to be achieved under the current circumstances, which leads to the lack of comprehensiveness of this study. Second, we have not taken into account the heterogeneity of the immune microenvironment associated with the location of immune infiltration. Third, all data series retrieved for the construction of nomograms aimed to predict outcomes were from the west. Therefore, if patients are from other countries, samples are tested by other platforms, but GPL96 or GPL570. Last but not least, the small sample size of bone metastasis melanoma may reduce the confidence and transformation of the predictive models into other cohorts. And to minimize bias, additional validation based on multiple databases was applied to detect gene expression levels of key biomarkers at the tissue and cellular levels, showing the key biomarkers were significantly associated with metastasis and prognosis of SKCM (**Figure S5–S10**).

CONCLUSIONS

According to ceRNA networks and tumor-infiltrating immune cells, two nomograms were built, respectively, in our study to predict survival and metastasis of melanoma patients and had great utility, which was verified by high concordance index and AUC values. Based on the comprehensive clinical information from the prediction nomograms, individual management of melanoma patients could be greatly improved. Furthermore, with sufficient evidence shown in this study, we speculate that melanoma bone metastasis may depend on the interaction among hsa-miR-27b-3p, AL118506.1, THBS2, and Tfh cells.

DATA AVAILABILITY

All datasets for this study are included in the TCGA-SKCM program.

ETHICS STATEMENT

The Ethics Committee of the First Affiliated Hospital of Zhengzhou University approved this study (no. 2019-KY-107).

AUTHOR CONTRIBUTIONS

Conception/design: RH, ZZ, GL, DS, PY, HY, PH, XiZ, RC, XuZ, TM, JZ, and ZH. Provision of study material: RH, ZZ, GL, TMeng, JZ, and ZH. Collection and/or assembly of data: RH, ZZ, GL, DS, PY, HY, PH, XiZ, RC, and XuZ. Data analysis and interpretation: RH, ZZ, GL, DS, PY, HY, PH, XiZ, RC, and XuZ. Manuscript writing: RH, ZZ, GL, TM, JZ, and ZH. Final approval of manuscript: RH, ZZ, GL, DS, PY, HY, PH, XiZ, RC, XuZ, TM, JZ, and ZH.

FUNDING

This study was supported in part by the National Natural Science Foundation of China (grant no. 81702659; 81772856; 81501203). Youth Fund of Shanghai Municipal Health Planning Commission (No.2017YQ054); Henan Medical Science and Technology Research Project (grant no. 201602031).

ACKNOWLEDGMENTS

We thank the TCGA team of the National Cancer Institute for using their data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00828/full#supplementary-material

TABLE S1| Baseline information of 112 patients diagnosed with Primary melanoma.

FIGURE S1 | The result of enrichment analysis showed that genes in melanoma tissues were significantly associated with extracellular matrix organization.

FIGURE S2 | The results of the Wilcoxon rank-sum test in T regulatory cells (Tregs) of different T stages.

FIGURE S3 | The correlation analysis revealed that T cells follicular helper was positively correlated with hsa-miR-425-5p ($P = 0.019$; $R = 0.260$).

FIGURE S4 | The results of Cox proportional hazards model and the nomogram integrating both biomarkers and immune cell portions significantly associated with prognosis. Bone metastasis–specific immune cells and ceRNAs significantly associated with prognosis were integrated into one multi-variable model and one nomogram **(A, E)**, which could decently predict the prognosis of SKCM (AUC of 3-year survival: 1.000; AUC of 5-year survival: 1.000) **(D)**. However, the model diagnostic information suggested that the prediction model had bias due to the small sample size **(A, B, C, F)**.

FIGURE S5 | Use CellMarker to explore the surface markers of T follicular helper cells. At the cellular level, BCL6 transcription repressor (BCL6), membrane metalloendopeptidase (MME), C-X-C motif chemokine ligand 13 (CXCL13), inducible T-cell costimulator (ICOS) and Programmed cell death 1 (PDCD1) had been reported as the surface markers of T follicular helper cell in the CellMarker.

FIGURE S6 | Protein–protein interaction network of ABHD16B, THBS2, BCL6, MME, CXCL13, ICOS, PDCD1, indicating that there are many interactions between THBS2 protein and T infertile helper cell's surface markers.

FIGURE S7 | The expression levels and co-expression analysis of AL118506.1, ABHD16B, THBS2, BCL6, MME, CXCL13, ICOS, PDCD1 in various SKCM cell lines and normal skin tissue in Cancer Cell Line Encyclopedia (CCLE) **(A, B)** and The Genotype–Tissue Expression (GTEx) database **(C, D)**.

FIGURE S8 | Validation of THBS2 **(A, B)**, ICOS **(C, D)**, CXCL13 **(E, F)**, BCL6 **(G, H)**, MME **(I, J)**, and PDCD1 **(K, L)** on a transcriptional level in multiple cancer types and multiple studies using the Oncomine database.

FIGURE S9 | The results of reanalysis of GSE19234 in SurvExpress. The reanalysis results of GSE19234 in SurvExpress suggested that these genes have significant predictive value for metastasis (Censoring event: metastasis, Hazard Ratio = 5.19 (95% CI, 1.92–14.05), *P* = 0.001)

FIGURE S10 | The results of reanalysis of GSE22153 in SurvExpress. The reanalysis results of GSE22153 in SurvExpress suggested that these genes have significant predictive value for metastasis (Censoring event: subcutaneous metastasis, Hazard Ratio = 4.01 (95% CI, 1.93–8.34), *P* < 0.001) and prognosis (Censoring event: overall death, Hazard Ratio = 3.15 (95% CI, 1.71–5.80), *P* < 0.001).

REFERENCES

Agostini, J., Benoist, S., Seman, M., Julie, C., Imbeaud, S., Letourneur, F., et al. (2012). Identification of molecular pathways involved in oxaliplatin-associated sinusoidal dilatation. *J. Hepatol.* 56 (4), 869–876. doi: 10.1016/j.jhep.2011.10.023

Aguirre-Gamboa, R., Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Chacolla-Huaringa, R., Rodriguez-Barrientos, A., et al. (2013). SurvExpress: an online biomarker validation tool and database for cancer gene expression data using survival analysis. *PLoS One* 8, e74250. doi: 10.1371/journal.pone.0074250

Bidwell, B. N., Slaney, C. Y., Withana, N. P., Forster, S., Cao, Y., Loi, S., et al. (2012). Silencing of Irf7 pathways in breast cancer cells promotes bone metastasis through immune escape. *Nat. Med.* 18 (8), 1224–1231. doi: 10.1038/nm.2830

Bier, G., Hoffmann, V., Kloth, C., Othman, A.E., Eigentler, T., Garbe, C., et al. (2016). CT imaging of bone and bone marrow infiltration in malignant melanoma—Challenges and limitations for clinical staging in comparison to 18FDG-PET/CT. *Eur. J. Radiol*. 85, 732–738. doi: 10.1016/j.ejrad.2016.01.012

Bogunovic, D., O'neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y. L., Adams, S., et al. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc. Natl. Acad. Sci. U. S. A.* 106, 20429–20434. doi: 10.1073/pnas.0905139106

Bostel, T., Forster, R., Schlampp, I., Wolf, R., Serras, A.F., Mayer, A., et al. (2016). Stability, prognostic factors and survival of spinal bone metastases in malignant melanoma patients after palliative radiotherapy. *Tumori* 102, 156–161. doi: 10.5301/tj.5000382

Braeuer, R.R., Watson, I.R., Wu, C.J., Mobley, A.K., Kamiya, T., Shoshan, E., et al. (2014). Why is melanoma so metastatic? *Pigment Cell Melanoma* Res. 27, 19–36. doi: 10.1111/pcmr.12172

Bremnes, R.M., Busund, L.T., Kilvaer, T.L., Andersen, S., Richardsen, E., et al. (2016). The Role of Tumor-Infiltrating Lymphocytes in Development, Progression, and Prognosis of Non-Small Cell Lung Cancer. *J. Thorac. Oncol.* 11, 789–800. doi: 10.1016/j.jtho.2016.01.015

Cancer Genome Atlas, N. (2015). Genomic classification of cutaneous melanoma. *Cell* 161, 1681–1696. doi: 10.1016/j.cell.2015.05.044

Chang, I.W., Li, C.F., Lin, V.C., He, H.L., Liang, P.I., Wu, W.J., et al. (2016). Prognostic impact of Thrombospodin-2 (THBS2) overexpression on patients with urothelial carcinomas of upper urinary tracts and bladders. *J. Cancer* 7, 1541–1549. doi: 10.7150/jca.15696

Chen, J., Yu, Y., Li, H., Hu, Q., Chen, X., He, Y., et al. (2019). Long non-coding RNA PVT1 promotes tumor progression by regulating the miR-143/HK2 axis in gallbladder cancer. *Mol. Cancer* 18 (1), 33. doi: 10.1186/s12943-019-0947-9

Chen, M. M., Xiao, X., Lao, X. M., Wei, Y., Liu, R. X., Zeng, Q. H., et al. (2016). Polarization of tissue-resident TFH-like cells in human hepatoma bridges innate monocyte inflammation and M2b macrophage polarization. *Cancer Discov.* 6 (10), 1182–1195. doi: 10.1158/2159-8290.CD-16-0329

Chen, X., Chen, Z., Yu, S., Nie, F., Yan, S., Ma, P., et al. (2018). Long Noncoding RNA LINC01234 Functions as a competing endogenous rna to regulate CBFB expression by sponging miR-204-5p in gastric cancer. *Clin. Cancer Res.* 24 (8), 2002–2014. doi: 10.1158/1078-0432.CCR-17-2376

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). 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

Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res*. 46, D296–d302. doi: 10.1093/nar/gkx1067

Consortium, G. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Cymerman, R.M., Shao, Y., Wang, K., Zhang, Y., Murzaku, E.C., Penn, L.A., et al. (2016). De novo vs nevus-associated melanomas: differences in associations with prognostic indicators and survival. *J. Natl. Cancer Inst*. 10, 108. doi: 10.1093/jnci/djw121

deLeeuw, R. J., Kost, S. E., Kakal, J. A., and Nelson, B. H. (2012). The prognostic value of FoxP3+ tumor-infiltrating lymphocytes in cancer: a critical review of the literature. *Clin. Cancer Res.* 18, 3022–3029. doi: 10.1158/1078-0432.CCR-11-3216

Dieu-Nosjean, M.C., Antoine, M., Danel, C., Heudes, D., Wislez, M., Poulot, V., et al. (2008). Long-term survival for patients with non-small-cell lung cancer with intratumoral lymphoid structures. *J. Clin. Oncol*. 26, 4410–4417. doi: 10.1200/jco.2007.15.0284

Elfilali, A., Lair, S., Verbeke, C., La Rosa, P., Radvanyi, F., and Barillot, E. (2006). ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis. *Nucleic Acids Res.* 34, D613–D616. doi: 10.1093/nar/gkj022

Fridman, W. H., Zitvogel, L., Sautes-Fridman, C., and Kroemer, G. (2017). The immune contexture in cancer prognosis and treatment. *Nat. Rev. Clin. Oncol.* 14, 717–734. doi: 10.1038/nrclinonc.2017.101

Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., Mcdonald, E.R., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508. doi: 10.1038/s41586-019-1186-3

Gillgrass, A., Gill, N., Babian, A., and Ashkar, A. A. (2014). The absence or overexpression of IL-15 drastically alters breast cancer metastasis via effects on NK cells, CD4 T cells, and macrophages. *J. Immunol.* 193 (12), 6184–6191. doi: 10.4049/jimmunol.1303175

Guerriero, J. L., Sotayo, A., Ponichtera, H. E., Castrillon, J. A., Pourzia, A. L., Schad, S., et al. (2017). Class IIa HDAC inhibition reduces breast tumours and metastases through anti-tumour macrophages. *Nature* 543 (7645), 428–432. doi: 10.1038/nature21409

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

He, K., Liu, P., and Xu, L. X. (2017). The cryo-thermal therapy eradicated melanoma in mice by eliciting CD4(+) T-cell-mediated antitumor memory immune response. *Cell Death Dis.* 8 (3), e2703. doi: 10.1038/cddis.2017.125

Jonsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringner, M., et al. (2010). Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin. Cancer Res.* 16, 3356–3367. doi: 10.1158/1078-0432.CCR-09-2509

Kim, H., Watkinson, J., Varadan, V., and Anastassio, D. (2010). Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics* 3, 51. doi: 10.1186/1755-8794-3-51

Kumar, M. S., Armenteros-Monterroso, E., East, P., Chakravorty, P., Matthews, N., Winslow, M. M., et al. (2014). HMGA2 functions as a competing endogenous RNA to promote lung cancer progression. *Nature* 505 (7482), 212–217. doi: 10.1038/nature12785

Li, R., Qu, H., Wang, S., Wei, J., Zhang, L., Ma, R., et al. (2018). GDCRNATools: an R/Bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics* 34, 2515–2517. doi: 10.1093/bioinformatics/bty124

Lin, X., Hu, D., Chen, G., Shi, Y., Zhang, H., Wang, X., et al. (2016). Associations of THBS2 and THBS4 polymorphisms to gastric cancer in a Southeast Chinese population. *Cancer Genet*. 209, 215–222. doi: 10.1016/j.cancergen.2016.04.003

Liu, J. F., Lee, C. W., Tsai, M. H., Tang, C. H., Chen, P. C., Lin, L. W., et al. (2018). Thrombospondin 2 promotes tumor metastasis by inducing matrix metalloproteinase-13 production in lung cancer cells. *Biochem. Pharmacol.* 155, 537–546. doi: 10.1016/j.bcp.2018.07.024

Liu, Q., Zheng, C., Shen, H., Zhou, Z., and Lei, Y. (2015). MicroRNAs-mRNAs Expression Profile and Their Potential Role in Malignant Transformation of Human Bronchial Epithelial Cells Induced by Cadmium. *Biomed. Res. Int.* 2015, 902025. doi: 10.1155/2015/902025

Liu, Q.H., and Ma, L.S. (2018). Knockdown of thrombospondin 2 inhibits metastasis through modulation of PI3K signaling pathway in uveal melanoma cell line M23. *Eur. Rev. Med. Pharmacol. Sci*. 22, 6230–6238. doi: 10.26355/eurrev_201810_16029

Lombard, D. B., Cierpicki, T., and Grembecka, J. (2019). Combined MAPK Pathway and HDAC Inhibition Breaks Melanoma. *Cancer Discov.* 9, 469–471. doi: 10.1158/2159-8290.CD-19-0069

Martens-Uzunova, E. S., Bottcher, R., Croce, C. M., Jenster, G., Visakorpi, T., and G. A. (2014). Calin: long noncoding RNA in prostate, bladder, and kidney cancer. *Eur. Urol.* 65 (6), 1140–1151. doi: 10.1016/j.eururo.2013.12.003

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Nezu, Y., Hagiwara, K., Yamamoto, Y., Fujiwara, T., Matsuo, K., Yoshida, A., et al. (2016). miR-135b, a key regulator of malignancy, is linked to poor prognosis in human myxoid liposarcoma. *Oncogene*. 35, 6177–6188. doi: 10.1038/onc.2016.157

Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., et al. (2016). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res*. 44, D231–238. doi: 10.1093/nar/gkv1270

Qi, H. (2016). T follicular helper cells in space-time. *Nat. Rev. Immunol.* 16 (10), 612–625. doi: 10.1038/nri.2016.94

Qian, Z., Zhang, G., Song, G., Shi, J., Gong, L., Mou, Y., et al. (2017). Integrated analysis of genes associated with poor prognosis of patients with colorectal cancer liver metastasis. *Oncotarget*. 8, 25500–25512. doi: 10.18632/oncotarget.16064

Roberts, D. D. (2008). Thrombospondins: from structure to therapeutics. *Cell. Mol. Life Sci.* 65, 669–671. doi: 10.1007/s00018-007-7483-2

Rodina, A., Wang, T., Yan, P., Gomes, E. D., Dunphy, M. P., Pillarsetty, N., et al. (2016). The epichaperome is an integrated chaperome network that facilitates tumour survival. *Nature* 538 (7625), 397–401. doi: 10.1038/nature19807

Rupaimoole, R., Calin, G. A., Lopez-Berestein, G., and Sood, A. K. (2016). miRNA deregulation in cancer cells and the tumor microenvironment. *Cancer Discov.* 6 (3), 235–246. doi: 10.1158/2159-8290.CD-15-0893

Schadendorf, D., et al. (2018). Melanoma. *Lancet* 392, 971–984. doi: 10.1016/S0140-6736(18)31559-9

Schadendorf, D., Hauschild, A., Santinami, M., Atkinson, V., Mandalà, M., Chiarion-Sileni, V., et al. (2019). Patient-reported outcomes in patients with resected, high-risk melanoma with BRAFV600E or BRAFV600K mutations treated with adjuvant dabrafenib plus trametinib (COMBI-AD): a randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.*20, 701–710. doi: 10.1016/S1470-2045(18)30940-9

Schmitt, A. M., and Chang, H. Y. (2016) Long noncoding RNAs in cancer pathways. *Cancer Cell* 29 (4), 452–463. doi: 10.1016/j.ccell.2016.03.010

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Sia, D., Villanueva, A., Friedman, S. L., and Llovet, J. M. (2017). Liver cancer cell of origin, molecular class, and effects on patient prognosis. *Gastroenterology* 152, 745–761. doi: 10.1053/j.gastro.2016.11.048

Slavin, S., Yeh, C. R., Da, J., Yu, S., Miyamoto, H., Messing, E. M., et al. (2014). Estrogen receptor alpha in cancer-associated fibroblasts suppresses prostate cancer invasion via modulation of thrombospondin 2 and matrix metalloproteinase 3. *Carcinogenesis* 35 (6), 1301–1309. doi: 10.1093/carcin/bgt488

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, D607–d613. doi: 10.1093/nar/gky1131

Tangye, S. G., Ma, C. S., Brink, R., and Deenick, E. K. (2013). The good, the bad and the ugly - TFH cells in human health and disease. *Nat. Rev. Immunol.* 13 (6), 412–426. doi: 10.1038/nri3447

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Vinuesa, C. G., Linterman, M. A., Yu, D., and MacLennan, I. C. (2016). Follicular helper T cells. *Annu. Rev. Immunol.* 34, 335–368. doi: 10.1146/annurev-immunol-041015-055605

Volders, P. J., Helsens, K., Wang, X., Menten, B., Martens, L., Gevaert, K., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41 (Database issue), D246–D251. doi: 10.1093/nar/gks915

Wang, Y., Hou, J., He, D., Sun, M., Zhang, P., Yu, Y., et al. (2016). The Emerging Function and Mechanism of ceRNAs in Cancer. *Trends Genet*. 32, 211–224. doi: 10.1016/j.tig.2016.02.001

Wei, W.F., Zhou, C.F., Wu, X.G., He, L.N., Wu, L.F., Chen, X.J., et al. (2017). MicroRNA-221-3p, a TWIST2 target, promotes cervical cancer metastasis by directly targeting THBS2. *Cell Death Dis*. 8, 3220. doi: 10.1038/s41419-017-0077-5

Yamaguchi, K., Mishima, K., Ohmura, H., Hanamura, F., Ito, M., Nakano, M., et al. (2018). Activation of central/effector memory T cells and T-helper 1 polarization in malignant melanoma patients treated with anti-programmed death-1 antibody. *Cancer Sci.* 109 (10), 3032–3042. doi: 10.1111/cas.13758

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* 47, D721–d728. doi: 10.1093/nar/gky900

Zhu, J., Yamane, H., and Paul, W. E. (2010). Differentiation of effector CD4 T cell populations (*). *Annu. Rev. Immunol.* 28, 445–489. doi: 10.1146/annurev-immunol-030409-101212

Zhuo, C., Li, X., Zhuang, H., Tian, S., Cui, H., Jiang, R., et al. (2016). Elevated THBS2, COL1A2, and SPP1 Expression Levels as Predictors of Gastric Cancer Prognosis. *Cell Physiol. Biochem*. 40, 1316–1324. doi: 10.1159/000453184

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.

ORIGINAL RESEARCH
published: 25 October 2019
doi: 10.3389/fgene.2019.01021

Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes

Tianyi Zhao[1], Yang Hu[2], Tianyi Zang[1], and Yadong Wang[1]**

[1] Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China,[2] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China

It is estimated that the impact of related genes on the risk of Alzheimer's disease (AD) is nearly 70%. Identifying candidate causal genes can help treatment and diagnosis. The maturity of sequencing technology and the reduction of cost make genome-wide association study (GWAS) become an important means to find disease-related mutation sites. Because of linkage disequilibrium (LD), neither the gene regulated by SNP nor the specific SNP can be determined. Because GWAS is affected by sample size and interaction, we introduced empirical Bayes (EB) to make a meta-analysis of GWAS to greatly eliminate the bias caused by sample and the interaction of SNP. In addition, most SNPs are in the noncoding region, so it is not clear how they relate to phenotype. In this paper, expression quantitative trait locus (eQTL) studies and methylation quantitative trait locus (mQTL) studies are combined with GWAS to find the genes associated with Alzheimer disease in expression levels by pleiotropy. Summary data-based Mendelian randomization (SMR) is introduced to integrate GWAS and eQTL/mQTL data. Finally, we prioritized 274 significant SNPs, which belong to 20 genes by eQTL analysis and 379 significant SNPs, which belong to seven known genes by mQTL. Among them, 93 SNPs and 2 genes are overlapped. Finally, we did 10 case studies to prove the effectiveness of our method.

Keywords: Alzheimer's disease, Mendelian randomization, GWAS, eQTL, mQTL

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Rui Guo,
Harvard Medical School,
United States
Eunhee Choi,
Harvard Medical School,
United States

****Correspondence:***
Tianyi Zang
tianyi.zang@hit.edu.cn
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *22 April 2019*
Accepted: *24 September 2019*
Published: *25 October 2019*

Citation:
Zhao T, Hu Y, Zang T and Wang Y (2019) Integrate GWAS, eQTL, and mQTL Data to Identify Alzheimer's Disease-Related Genes. Front. Genet. 10:1021. doi: 10.3389/fgene.2019.01021

INTRODUCTION

It is estimated that the impact of related genes on the risk of AD is nearly 70%. Importantly, neuronal cell death precedes the appearance of cognitive symptoms for 10 years or more, suggesting that targeted treatment needs to be performed before symptoms appear. Therefore, the identification of AD biomarkers such as genes, RNAs (Jiang et al., 2015; Cheng et al., 2018; Cheng et al., 2019), proteins, and metabolites (Cheng et al., 2019) is critical for early detection and early intervention in AD. In addition, identifying candidate genes and loci can also help us understand the pathogenesis of AD and develop drugs.

Recently, Jansen et al. (Jansen et al., 2019) published his AD GWAS study on natural genetics. The sample size is more than eight times that of Lambert et al. (Lambert et al., 2013) in 2013. Due to the increase in the number of samples, they found nine AD risk loci more than in previous studies. Jansen et al. found that most of the AD-related DNA mutations were located in the noncoding part of the genome in regions that affected gene transcription. It means that combining GWAS data with transcriptional expression data will greatly advance AD research (Cheng et al., 2016).

However, GWAS still has certain limitations. The SNP is not necessarily the true pathogenic locus, but only related to the SNP that actually causes the disease due to the LD. GWAS usually analyzes the edge effects of individual loci while ignoring the interaction of multiple genes in complex diseases (Battle et al., 2014). Therefore, GWAS still cannot fully reveal the genetic susceptibility factors of complex diseases (Cheng et al., 2018). It is only an important part of exploring the genetic etiology of complex diseases (Cheng and Hu, 2018). Therefore, using GWAS data for research, we must first start with the expression of SNP, that is, combined with data affecting gene expression, which can weaken the impact of LD on significance. Then, the interaction of multiple genes is considered, that is, the statistical values of each SNP are revised within the whole genome.

It was found that about 80% of the genetic susceptibility loci detected by GWAS were located in the noncoding region of the genome, suggesting that the pathogenic loci may have regulatory functions on gene expression. An important role of large-scale eQTL research is to be able to prioritize SNP loci (Barral et al., 2012) in GWAS susceptible regions and to infer possible biological mechanisms through the influence of DNA polymers on biological characteristics. At present, many studies have used eQTL analysis as a very effective tool to explain the results of GWAS. Hormozdiari et al. (Hormozdiari et al., 2016) present a probabilistic method named eCAVIAR, which can detect target genes by colocalization of GWAS and eQTL signals. Xu et al. purposed a more powerful method based on PrediXcan and TWAS. It can integrate single set or multiple sets of eQTL data with GWAS.

mQTL is mainly based on the analysis of cis-mQTL, that is, using Beta value of methylation level of CpG locus near a gene as dependent variable, screening all SNP variations in the chromosomal region upstream and downstream of the gene as independent variable and regressing each SNP locus S and methylation level M in this region one by one, so as to obtain SNP loci significantly related to the methylation level of a gene. There is no doubt that methylation affects gene expression. This is very similar to eQTL, both of which can cause changes in expression through mutations in a single locus. Therefore, in recent years, more and more studies have been carried out to screen genes related to traits by combining mQTL with GWAS. Hägg et al. (Hägg et al., 2015) integrated GWAS, eQTL, and mQTL to find out genes which are related to obesity. Pharoah et al. (Pharoah et al., 2013) identified three new susceptibility loci for ovarian cancer by GWAS meta-analysis and verified the result by mQTL.

In our previous paper (Hu et al., 2018), we have identified some AD-related genes by GWAS and eQTL using SMR. There are three points to be improved. Firstly, mQTL should be included to verify and improve our result. Secondly, we used several eQTL datasets in that paper, whereas a meta-analysis method should be used to integrate the datasets, which can improve the accuracy of eQTL's statistical results. Finally, GWAS datasets should also be integrated into one dataset so that can overcome the difference of statistical power caused by sample size.

METHODS

SMR

Since Zhu et al. proposed "SMR" in 2016, it has become a common way to identify the genes whose expression levels are associated with a complex trait because of pleiotropy. Using GWAS and eQTL data, SMR could screen trait-related genes. After two years, they applied SMR to mQTL data. They found 7,858 DNAm sites which are related to 14 complex traits.

The basic idea of this method is as follows. First, let y be the phenotype, which is the outcome variable. x is the gene expression, which is the exposure factor. z is the gene mutation, which is the instrumental variable. Then, b_{xy} is the effect of x on y, b_{zx} is the effect of z on x, and b_{zy} is the effect of z on y. The definition of b_{xy} is $b_{xy} = b_{zy}/b_{zx}$, which means the effect of gene expression on phenotype without confounding factors. This idea is based on the Mendelian randomization (Cheng et al., 2018; Cheng et al., 2019).

Figure 1 is a hypothetical model of a mediation mechanism tested in SMR. The blue line represents causal relationship. Methylation will cause SNP. Both SNP and methylation can affect the change of transcription. The change of transcription will cause the difference of trait. The red line denotes the relationship data represents. mQTL denotes the relationship between methylation and SNP. eQTL denotes the relationship between transcription and SNP. GWAS denotes the relationship between SNP and trait.

Based on this hypothesis, many researchers have found the genes which are related to certain traits. Diseases like bone mineral density (BMD) (Meng et al., 2018), amyotrophic lateral sclerosis (ALS) (Du et al., 2017), and neuroticism (Fan et al., 2017) have been found some potential related genes by SMR. Other traits like height, BMI (Yengo et al., 2018), and obesity (Liu et al., 2018) have also researched by SMR.

Eb-GWAS

Due to the complex linkage effects and statistical errors of the samples, the contribution of GWAS to biological research is reduced. GWAS may associate common diseases with thousands of DNA mutations, that is, every DNA region that happens to be active in diseased tissues may be associated with disease (Jiang et al., 2013). Many GWAS matches are not specifically biologically related to disease and, therefore, cannot be used as effective drug targets. In fact, these "peripheral" mutations are likely to affect the activity of "core" genes, which are more directly related to disease, through complex biochemical regulatory networks (Jiang et al., 2010).

As we discussed before in the introduction, the interaction of multiple genes is considered, that is, the statistical values of each SNP are revised within the whole genome. In this section, we will

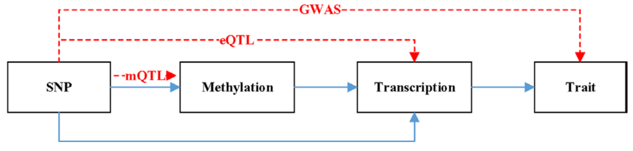


FIGURE 1 | A hypothetical model of a mediation mechanism.

process GWAS data in two steps: 1. meta-analysis, 2. using EB, revise the statistical value of each SNP within the whole genome.

Meta-Analysis

Since SE denotes the standard error of each SNP, it represents the reliability of Beta values. Then, weight of each Beta should be:

$$w_i = 1/SE_i^2 \tag{1}$$

SE_i denotes the standard error for study i, w_i denotes the weight of Beta.

Then, the Beta after meta-analysis would be:

$$\beta = \sum_i \beta_i w_i / \sum_i w_i \tag{2}$$

β_i denotes effect size estimate for study i.

Then, we could use the weight of each Beta to calculate the result of meta-analysis.

$$SE = \sqrt{1/\sum_i w_i} \tag{3}$$

Finally, the overall Z-score could be obtained by the original equation.

$$Z = \beta / SE \tag{4}$$

Eb-GWAS

After meta-analysis, we could summary several GWAS datasets into one dataset. Then, we used EB to integrate all the Z scores in the whole genomic level. As we know that the SNP could interact with each other, the Z score of all SNP should have some relationship and obey normal distribution.

The overall Z-score we obtained before obeying normal distribution with standard deviation is 1. Then,

$$\hat{Z}_i \,|\, Z_i \overset{ind}{\sim} N(\hat{Z}_i, 1) \tag{5}$$

$\hat{Z}_i$ denotes the Z score we obtained. It is a value with bias. Z_i denotes the real Z score.

Real Z score obeys normal distribution:

$$Z \overset{ind}{\sim} N(\theta, \sigma^2) \tag{6}$$

Then, the marginal distribution of $\hat{Z}_i$ is

$$\hat{Z} \overset{ind}{\sim} N(\theta, \sigma^2 + 1) \tag{7}$$

Moreover, the posterior distribution should be:

$$Z_i \,|\, \hat{Z}_i \overset{ind}{\sim} N(\theta + B(\hat{Z}_i - \theta), B \tag{8}$$

$$B = \frac{\sigma^2}{1 + \sigma^2} \tag{9}$$

Then, we could know that $E(\hat{Z}_i) = \theta$, so the mean of $\hat{Z}_i$ can be used to estimate θ.

$$\hat{\theta} = mean(\hat{Z}_i) = \bar{\bar{Z}}_i \tag{10}$$

$$\frac{\sum_i^N (\hat{Z}_i - \bar{\bar{Z}}_i)^2}{\sigma^2 + 1} = \frac{S}{\sigma^2 + 1} \sim \chi^2(N-1) \tag{11}$$

Then,

$$\frac{\sigma^2 + 1}{S} \sim inverse - \chi^2(N-1) \tag{12}$$

From the properties of inverse chi-square distribution,

$$E(\frac{\sigma^2 + 1}{S}) \sim \frac{1}{N-3} \tag{13}$$

Then,

$$E(\frac{N-3}{S}) = \frac{1}{\sigma^2 + 1} = 1 - B \tag{14}$$

Therefore, the EB estimation of B is

$$B = 1 - \frac{(N-3)}{S} \tag{15}$$

Finally, we can put the (Hu et al., 2018) into (Battle et al., 2014)

$$Z_i = \bar{\bar{Z}} + (1 - \frac{(N-3)}{S})(\hat{Z}_i - Z) \tag{16}$$

Then, we have done the meta-analysis and revised the statistical value of each SNP within the whole genome.

Dataset

As shown in **Table 1** we obtained five GWAS datasets, three eQTL dataset, and three mQTL datasets. All the eQTL and mQTL are from brain tissue. Yang Jian et al. have already meta-analysis the eQTL and mQTL datasets. Therefore, we used the data they processed.

For GWAS dataset, Scelsi M A et al. obtained the data from 1,517 Caucasian ADNI subjects. Lambert JC et al.'s dataset is

TABLE 1 | Datasets used in this paper.

Data	Name	Reference
GWAS	ADNI_DPS_GWAS	Scelsi et al. (2018)
	ADNI_amyloid_GWAS	(include three datasets)
	ADNI_hippo_GWAS	
	IGAP_stage_1	Lambert et al. (2013)
	UK_Biobank	Marioni et al. (2018)
eQTL	GTEx-brain eQTL	GTEx Consortium (2017)
	CMC	Fromer et al. (2016)
	ROSMAP	Ng et al. (2017)
mQTL	ROSMAP	Ng et al. (2017)
	Human fetal brain	Hannon et al. (2016)
	Frontal cortex	Jaffe et al. (2016)

consisted of 17,008 Alzheimer's disease cases and 37,154 controls. Marioni R E et al. obtained data from 314,278 participants.

For eQTL dataset, SNPs within 1Mb distance from each probe are available in these three datasets. After meta-analysis, the estimated effective sample size n = 1194.

For mQTL dataset, 5kb, 500kb, and 20kb are the available distance for the three datasets, respectively. After meta-analysis, the estimated effective sample size n = 1160.

RESULTS

Results of GWAS Meta-Analysis

We did a meta-analysis of five groups of GWAS data and integrated them into a GWAS file.

The blue block in **Figure 2** is P value density of GWAS after meta-analysis. The red block in **Figure 2** is P value density of GWAS after EB. As we can see in **Figure 2**, the distribution approximates uniform distribution. After using EB in all SNPs in whole dataset, the P value of the final GWAS data approximates the normal distribution.

Results of SMR

GWAS included 1,474,846 SNPs, mQTL included 6,966,746, and eQTL included 1,067,443 SNPs. There are 149,326 SNPs occur in both GWAS and eQTL and 408,896 SNPs occur in both GWAS and mQTL. Therefore, we use SMR to test these repeated SNPs in data sets.

Note that some SNPs are marked by multiple probes, so one SNP may significant in more than one gene. One SNP may affect expression of multiple genes.

In **Figures 3** and **4**, we can see that SNPs' P value in GWAS are not related to eQTL and mQTL. It means that only few significant SNPs in GWAS have significance in eQTL and mQTL. Anyway, the points near the upper right corner in the images mean that the difference in expression level caused by these SNPs is related to AD and SMR can help us detect these SNPs.

We set a threshold as 0.05/(number of probers). For eQTL data, the threshold is 0.05/8362 = 5.98e-06. For mQTL data, the threshold is 0.05/97263 = 5.14e-07. The numbers of SNPs and genes identified by the two experiments are shown in **Table 2**.

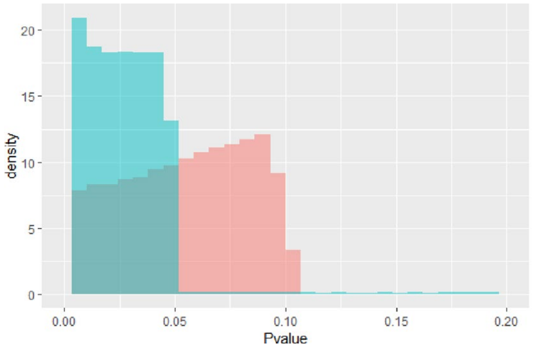


FIGURE 2 | Pvalue density of genome-wide association study (GWAS).

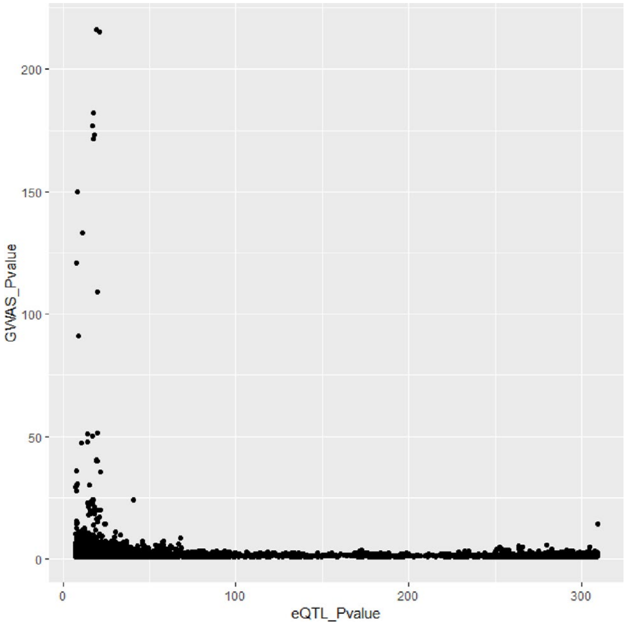


FIGURE 3 | Duplicated SNPs' P value in genome-wide association study (GWAS) and eQTL.

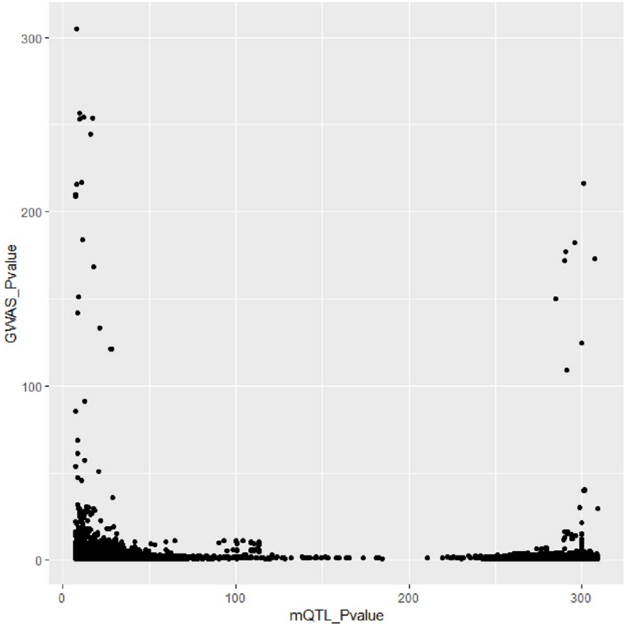


FIGURE 4 | Duplicated SNPs' P value in genome-wide association study (GWAS) and mQTL.

Figure 5 shows all the SNPs' P value. The red points are the P value of GWAS SNPs. The blue points are the P value of eQTL SNPs and the green points are the P value of mQTL SNPs. There is a black line in the first picture. The line is the significant threshold of P value. It is -log10(5*10-8). The SNPs of eQTL and mQTL are already screened so each SNP's P value is less than 5*10-8.

TABLE 2 | The results of summary data-based Mendelian randomization (SMR).

Dataset	Number of SNPs	Number of Genes
GWAS&eQTL	274	20
GWAS&mQTL	379	7
Overlapped	93	2

Figure 6 shows the result of SMR by two different datasets. The first graph is the result of GWAS and eQTL and the second one is the result of GWAS and mQTL. The black line in the two graphs is significant threshold, respectively. As we can see, only few of SNPs can pass the SMR test. Some of them are not very significant in GWAS, but combined with eQTL or mQTL, they would be significant.

As we can see in **Table 3**, HLA-DQA1 and HLA-DRB5 are selected in both eQTL and mQTL datasets. The HLA complex is located in the 21.31 region (6p21.31) on the short arm of chromosome 6 and is composed of 3.6 million base pairs. It is the region with the highest gene density and the most polymorphic region in human chromosomes. Known as "chemical fingerprints in humans". Due to the complexity of HLA, the methylation level and expression level differ greatly.

Case Study

In this section, we want to confirm whether the 25 AD-related genes we found have been reported by others. In order to be precise, we only use the literature that got AD-related genes by biological experiments, rather than the bioinformatics method or GWAS method.

Zhu et al. (2017) found four CR1 SNPs showed significant associations with the Aβ deposition at the baseline level.

James et al. (2018) gathered 71 cognitively healthy women's the volumes of total gray matter, cerebrocor-tical gray matter, and subcortical gray matter by structural magnetic resonance imaging

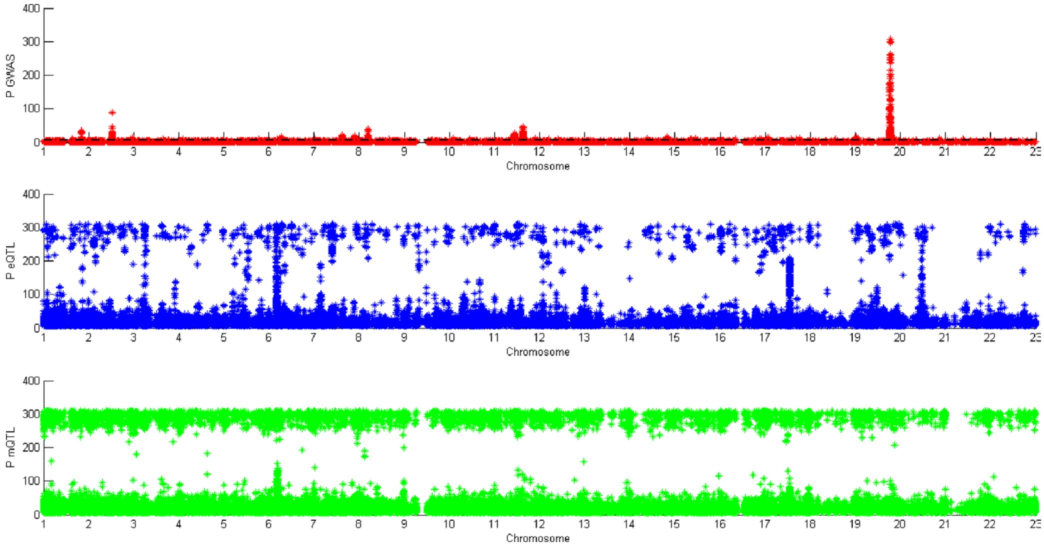


FIGURE 5 | P value of genome-wide association study (GWAS), eQTL, and mQTL.

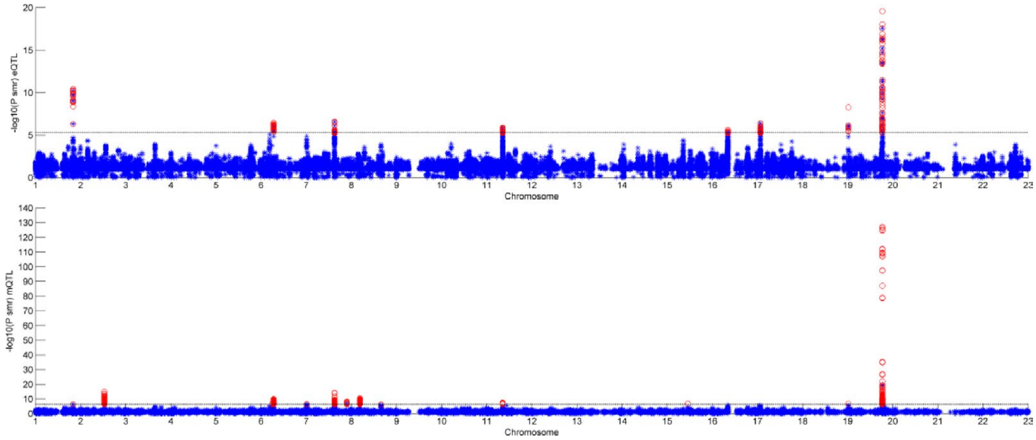


FIGURE 6 | Result of summary data-based Mendelian randomization (SMR).

TABLE 3 | The candidate genes selected by summary data-based Mendelian randomization (SMR).

	Gene	Number of SNPs
eQTL	*CR1*	20
	HLA-DRB1	69
	HLA-DQA1	39
	HLA-DRB5	8
	HLA-DQB1	3
	HLA-DQB1-AS1	1
	RP11-385F7.1	36
	ZSCAN21	8
	PILRB	5
	PILRA	5
	MTCH2	20
	KAT8	20
	AC012146.7	23
	ZNF232	4
	POLR2E	7
	PVR	12
	CTB-171A8.1	24
	CEACAM19	11
	TOMM40	23
	ZNF296	6
mQTL	*BIN1*	11
	HLA-DRB5	15
	HLA-DRB1	16
	EPHA1-AS1	3
	FAM63B	2
	APOC1	12
	EXOC3L2	24

(sMRI) scan and found that the protective effect of DRB1*13:02 is related to successful elimination of specific pathogens that would ultimately cause gradual brain atrophy.

Yu et al. (2015) found that BIN1 was associated with Aβ load and brain DNA methylation in HLA-DRB5 was associated with pathological AD by 447 participants

Lee et al. (2018) used non-Hispanic Caucasians with neuroimaging and found that HLA-DQB1 is significantly associated with entorhinal cortical thickness by controlling for multiple testing.

Yoshino et al. (2016) found that SNCA mRNA expression in 50 AD subjects was significantly higher than that in control subjects. Therefore, they inferred mRNA expression and methylation of SNCA intron 1 are altered in AD, whereas ZSCAN21 at upstream of these CpG site were reported to bind at intron 1.

Rathore et al. (2018) noted that both TREM2 and PILRB function as activating receptors and signal through DAP12. A reduction of PILRA inhibitory signals in R78 carriers could allow more microglial activation via PILRB/DAP12 signaling and reinforce the cellular mechanisms by which TREM2 is believed to protect from AD incidence.

Ruggiero et al. (2017) did biological experiments on mice and found that MTCH2 is a critical player in neuronal cell biology, controlling mitochondria metabolism, motility, and calcium buffering to regulate hippocampal-dependent cognitive functions.

De Jager et al. (2014) used a collection of 708 prospectively collected autopsied brains to assess the methylation state of the brain's DNA in relation to AD and found two SNPs associated with POLR2E are related to AD in methylation levels.

Roses et al. (2010) identified polymorphic poly-T variant rs10524523 in transposase of TOMM40 gene, which can be used to estimate the starting age of LOAD with APOE ε3 carriers.

Prendecki et al. (2018) recruited 230 individuals and found that APOC1 and TOMM40 rs2075650 polymorphisms may be independent risk factors of developing AD, whose major variants are accompanied by disruption of biothiols metabolism and inefficient removal of DNA oxidation.

We found 10 of 25 genes are reported to be related to AD by biological experiments. Some literary works may found that the other 15 genes are related to AD via other methods, but we would not discuss in this paper. This case study verified the effectiveness of our method and we hope the other 15 genes could be verified by biological experiments in future.

CONCLUSION

AD brings great burden to patients and society and identifying AD-related genes can help us known the machanism of AD then diagnose and treatment. In this paper, we used SMR to find AD-related genes by GWAS, eQTL, and mQTL. There are some overlaps between GWAS and the other two datasets, which means that some SNPs are related to AD due to the change of expression level. SMR is a method which can identify the genes whose expression levels are associated with a complex trait because of pleiotropy.

Due to the LD and interaction between genes, GWAS data has bias. In order to overcome these, we did meta-analysis on five GWAS datasets and then used EB to revise the Z-score of each SNPs in whole-SNP level.

Finally, we found 653 SNPs reached the threshold of significance and they are associated with 25 genes. Ninety-three of SNPs are significant in both GWAS&eQTL and GWAS&mQTL tests. We did 10 case studies at last, which means that the 10 of 25 genes we identified have been verified to correlated to AD by biological experiments in existing literary works.

DATA DEPOSITION

eQTL and mQTL Data

The direct link for accessing eQTL and mQTL data is as follows (origin from PMID: 29891976).

1) eQTL data: https://cnsgenomics.com/data/SMR/Brain-eMeta.tar.gz
2) mQTL data: https://cnsgenomics.com/data/SMR/Brain-mMeta.tar.gz

GWAS Dataset 1,2,3

GWAS dataset 1,2,3 are from paper PMID:29860282. The direct link is for accessing them is as following.

1) https://www.ebi.ac.uk/gwas/studies/GCST006134 & ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ScelsiMA_29860282_GCST006134
2) https://www.ebi.ac.uk/gwas/studies/GCST006136 & ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ScelsiMA_29860282_GCST006135
3) https://www.ebi.ac.uk/gwas/studies/GCST006135 & ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/ScelsiMA_29860282_GCST006136

GWAS Data 4

GWAS data 4 is from PMID: 24162737. The direct link is for accessing it is as following:

http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php

GWAS Data 5

GWAS data 5 is from PMID: 29777097. The direct link is for accessing it is as following:

http://datashare.is.ed.ac.uk/download/DS_10283_3364.zip

All code could be downloaded by

https://github.com/zty2009/Integrate-GWAS-eQTL-and-mQTL-data-to-identify-Alzheimer-s-Disease-related-genes

AUTHOR CONTRIBUTIONS

TZang and YW are the corresponding authors. They help to revise and support data for this data. TZhao and YH are the co-first authors. They wrote the code and write the paper.

FUNDING

This work was supported by the National Natural Science Foundation of China (No: 61571152 and 61502125), the National High-tech R&D Program of China (863 Program) [Nos: 2014AA021505, 2015AA020101, 2015AA020108], the National Science and Technology Major Project [Nos: 2013ZX03005012 and 2016YFC1202302], the Heilongjiang Postdoctoral Fund (Grant No. LBH-Z15179), and the China Postdoctoral Science Foundation (Grant No. 2016M590291).

REFERENCES

Barral, S., Bird, T., Goate, A., Farlow, M., Diaz-Arrastia, R., Bennett, D., et al. (2012). Genotype patterns at PICALM, CR1, BIN1, CLU, and APOE genes are associated with episodic memory. *Neurology* 78, 1464–1471. doi: 10.1212/WNL.0b013e3182553c48

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* 24, 14–24. doi: 10.1101/gr.155192.113

Cheng, L., and Hu, Y. (2018). Human Disease System Biology. *Curr. Gene. Ther.* 18, 255–256. doi: 10.2174/1566523218666181010101114

Cheng, L., Sun, J., Xu, W. Y., Dong, L. X., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 1–9. doi: 10.1038/srep34820

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018). Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 9, 657. doi: 10.3389/fgene.2018.00657

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018). InfAcrOnt: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19, 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J. W., Tan, R. J., et al. (2019). Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 10, 10. doi: 10.3389/fgene.2019.00094

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). 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, G. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204. doi: 10.1038/nature24277

De Jager, P. L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L. C., Yu, L., et al. (2014). Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nat. Neurosci.* 17, 1156. doi: 10.1038/nn.3786

Du, Y., Yan, W., Guo, X., Hao, J., Wang, W., He, A., et al. (2017). and Pathways Associated with Amyotrophic Lateral Sclerosis. *Cell. Mol. Neurobiol.* 38, 1–5. doi: 10.1007/s10571-017-0512-2

Fan, Q., Wang, W., Hao, J., He, A., Wen, Y., Guo, X., et al. (2017). Integrating genome-wide association study and expression quantitative trait loci data identifies multiple genes and gene set associated with neuroticism. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* 78, 149–152. doi: 10.1016/j.pnpbp.2017.05.017

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* 19, 1442. doi: 10.1038/nn.4399

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature* 550 (7675), 204.

Hägg, S., Ganna, A., Van Der Laan, S. W., Esko, T., Pers, T. H., Locke, A. E., et al. (2015). Gene-based meta-analysis of genome-wide association studies implicates new loci involved in obesity. *Hum. Mol. Genet.* 24, 6849–6860. doi: 10.1093/hmg/ddv379

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., et al. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* 19, 48. doi: 10.1038/nn.4182

Hormozdiari, F., Vandebunt, M., Segrè, A., Li, X., Joo, J. W., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* 99, 1245–1260. doi: 10.1016/j.ajhg.2016.10.003

Hu, Y., Zhao, T., Zang, T., Zhang, Y., and Cheng, L. (2018). Identification of Alzheimer's disease-related genes based on data integration method. *Front. Genet.* 9, 703. doi: 10.3389/fgene.2018.00703

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., et al. (2016). genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* 19, 40. doi: 10.1038/nn.4181

James, L. M., Christova, P., Lewis, S. M., Engdahl, B. E., Georgopoulos, A., and Georgopoulos, A. P. (2018). Protective effect of human leukocyte antigen (HLA) Allele DRB1* 13: 02 on age-related brain gray matter volume reduction in healthy women. *EBioMedicine* 29, 31–37. doi: 10.1016/j.ebiom.2018.02.005

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* 51, 404–413. doi: 10.1038/s41588-018-0311-9

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. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/IJDMB.2013.056078

Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., et al. (2015). LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genomics* 16 Suppl 3, S2. doi: 10.1186/1471-2164-16-S3-S2

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45, 1452. doi: 10.1038/ng.2802

Lee, Y., Han, S., Kim, D., Kim, D., Horgousluoglu, E., Risacher, S. L., et al. (2018). Genetic variation affecting exon skipping contributes to brain structural atrophy in Alzheimer's disease. *AMIA Summits on Translat. Sci. Proc.* 2017, 124.

Liu, L., Fan, Q., Zhang, F., Guo, X., Liang, X., Du, Y., et al. (2018). A Genomewide Integrative Analysis of GWAS and eQTLs Data Identifies Multiple Genes and Gene Sets Associated with Obesity. *Biomed. Res. Int.* 2018. 1–5 doi: 10.1155/2018/3848560

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., et al. (2018). GWAS on family history of Alzheimer's disease. *Transl. Psychiatry* 8 (1), 99. doi: 10.1038/s41398-018-0150-6

Meng, X. H., Chen, X. D., Greenbaum, J., Zeng, Q., You, S. L., Xiao, H. M., et al. (2018). Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone* 113, 41–48. doi: 10.1016/j.bone.2018.05.012

Ng, B., White, C. C., Klein, H.-U., Sieberts, S. K., McCabe, C., Patrick, E., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* 20, 1418. doi: 10.1038/nn.4632

Pharoah, P. D., Tsai, Y.-Y., Ramus, S. J., Phelan, C. M., Goode, E. L., Lawrenson, K., et al. (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat. Genet.* 45, 362. doi: 10.1038/ng.2564

Prendecki, M., Florczak-Wyspianska, J., Kowalska, M., Ilkowski, J., Grzelak, T., Bialas, K., et al. (2018). Biothiols and oxidative stress markers and polymorphisms of TOMM40 and APOC1 genes in Alzheimer's disease patients. *Oncotarget* 9 (81), 35207. doi: 10.18632/oncotarget.26184

Rathore, N., Ramani, S. R., Pantua, H., Payandeh, J., Bhangale, T., Wuster, A., et al. (2018). Paired immunoglobulin-like type 2 receptor alpha G78R variant alters ligand binding and confers protection to Alzheimer's disease. *PLoS Genet.* 14 (11), e1007427. doi: 10.1371/journal.pgen.1007427

Roses, A., Lutz, M., Amrine-Madsen, H., Saunders, A., Crenshaw, D., Sundseth, S., et al. (2010). A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. *Pharmacogenomics J.* 10, 375. doi: 10.1038/tpj.2009.69

Ruggiero, A., Aloni, E., Korkotian, E., Zaltsman, Y., Oni-Biton, E., Kuperman, Y., et al. (2017). Loss of forebrain MTCH2 decreases mitochondria motility and calcium handling and impairs hippocampal-dependent cognitive functions. *Sci. Rep.* 7, 44401. doi: 10.1038/srep44401

Scelsi, M. A., Khan, R. R., Lorenzi, M., Christopher, L., Greicius, M. D., Schott, J. M., et al. (2018). Genetic study of multimodal imaging Alzheimer's disease progression score implicates novel loci. *Brain* 141, 2167–2180. doi: 10.1093/brain/awy141

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., et al. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ~700,000 individuals of European ancestry. *Hum. Mol. Genet.* 27 (20), 3641–3649. doi: 10.1101/274654

Yoshino, Y., Mori, T., Yoshida, T., Yamazaki, K., Ozaki, Y., Sao, T., et al. (2016). Elevated mRNA expression and low methylation of SNCA in Japanese Alzheimer's disease subjects. *J. Alzheimer's Dis.* 54, 1349–1357. doi: 10.3233/JAD-160430

Yu, L., Chibnik, L. B., Srivastava, G. P., Pochet, N., Yang, J., Xu, J., et al. (2015). Association of Brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease. *JAMA Neurol.* 72, 15–24. doi: 10.1001/jamaneurol.2014.3049

Zhu, X.-C., Wang, H.-F., Jiang, T., Lu, H., Tan, M.-S., Tan, C.-C., et al. (2017). Initiative, Effect of CR1 genetic variants on cerebrospinal fluid and neuroimaging biomarkers in healthy, mild cognitive impairment and Alzheimer's disease cohorts. *Mol. Neurobiol.* 54, 551–562. doi: 10.1007/s12035-015-9638-8

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.

ORIGINAL RESEARCH
published: 29 October 2019
doi: 10.3389/fgene.2019.01020

A New Algorithm for Identifying Genome Rearrangements in the Mammalian Evolution

*Juan Wang[1], Bo Cui[1], Yulan Zhao[1] and Maozu Guo[2,3]**

[1] School of Computer Science, Inner Mongolia University, Hohhot, China, [2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [3] Beijing University of Civil Engineering and Architecture, Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing, China

Genome rearrangements are the evolutionary events on level of genomes. It is a global view on evolution research of species to analyze the genome rearrangements. We introduce a new method called RGRPT (recovering the genome rearrangements based on phylogenetic tree) used to identify the genome rearrangements. We test the RGRPT using simulated data. The results of experiments show that RGRPT have high sensitivity and specificity compared with other tools when to predict rearrangement events. We use RGRPT to predict the rearrangement events of six mammalian genomes (human, chimpanzee, rhesus macaque, mouse, rat, and dog). RGRPT has recognized a total of 1,157 rearrangement events for them at 10 kb resolution, including 858 reversals, 16 translocations, 249 transpositions, and 34 fusions/fissions. And RGRPT has recognized 475 rearrangement events for them at 50 kb resolution, including 332 reversals, 13 translocations, 94 transpositions, and 36 fusions/fissions. The code source of RGRPT is available from https://github.com/wangjuanimu/data-of-genome-rearrangement.

Keywords: genome rearrangements, mammal, phylogenetic tree, evolution, algorithm

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Yungang Xu,
University of Texas Health Science Center at Houston,
United States
Zhen Tian,
Zhengzhou University, China
Wei Lan,
Guangxi University, China

****Correspondence:***
Maozu Guo
guomaozu@bucea.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 July 2019*
Accepted: *24 September 2019*
Published: *29 October 2019*

Citation:
Wang J, Cui B, Zhao Y and Guo M (2019) A New Algorithm for Identifying Genome Rearrangements in the Mammalian Evolution. Front. Genet. 10:1020. doi: 10.3389/fgene.2019.01020

INTRODUCTION

The rapid development of sequencing technologies makes the phylogenetic analysis from the level of whole genome possible. A studied genome is represented as a line of conserved segments (called syntenic blocks). The genome rearrangements of species are changes of syntenic block orderings and losing of sequence blocks. These events include reversal, translocation, transposition, fusion, fission, and so on (Xu et al., 2017; Cheng et al., 2019; Dong et al., 2018). The research on genome rearrangements is mainly three aspects.

One is the computation of evolutionary distance between two species by considering genome rearrangements. Researchers have proposed a lot of metric for measuring the dissimilarity of evolution between species and a large amount of algorithms for computing the metrics. The breakpoint distance is the minimum rearrangement operations transforming one genome to the other genome, which is computed by means of breakpoint graph (Blanchette et al., 1997; Sankoff and Blanchette, 1998). There are lots of algorithms for computing breakpoint distance. In 1995, Hannenhalli and Pevzner put forward an algorithm with O(n^5) time complexity to compute the breakpoint distance just considering reversal events (Hannenhalli and Pevzner, 1999). Later, Kaplan improved the algorithm to time complexity O(n^5) (Kaplan et al., 2000). In 1996, Hannenhalli designed an algorithm with O(n^3) time complexity to compute it by considering translocation events (Hannenhalli, 1995). In 2001, Zhu et al. improved the algorithm to time complexity O(n^2logn) (Zhu and Ma, 2002). And then Zhu et al. devised an algorithm with O(n^2) time complexity (Liu et al., 2004). The DCJ distance is introduced by Yancopoulos et al. (Sophia et al., 2005), which uses the double cut and join (DCJ for short) operation to model rearrangement events, such as reversal, translocation, transposition, fusion, and fission in an unified way. Yancopoulos et al. first propose a method to compute the DCJ distance by considering only translocations and reversals on linear chromosomes (Sophia et al., 2005). Paper (Lu et al., 2006) has proposed an $O(n^2)$ time algorithm to compute the distance by considering the fusions and fissions between circular unsigned chromosomes. Unimog (Hilker et al., 2012) is software for computing DCJ distance which implements lots of algorithms (Erdös et al., 2011; Jakub et al., 2011). SoRT is a tool to compute breakpoint distance and the DCJ distance for linear/circular multi-chromosomal gene orders (Yen-Lin et al., 2010). SCJ distance (Feijão and Meidanis, 2011) is defined using the single cut and join (SCJ for short) operations, which is in analogy to DCJ measure. The distance can be computed by a speedily computable.

Two is the reconstruction of the ancestral gene orders by using the genomes of extant species. Ma et al. (Ma et al., 2006) use maximum parsimony principle to recover reliably ancestral genomes starting from phylogenetic tree and adjacent genes in genome and make the probabilistic reconstruction accuracy analysis for the six mammalian genome (human, mouse, rat, dog, opossum, and chicken) based on the improved Jukes–Cantor model. PMAG utilized the Bayesian theorem in the probabilistic framework to infer ancestral genomes (Yang et al., 2014). Multiple Genome Rearrangements (MGR) recovers the ancestral genome by minimizing the rearrangement distance (Bourque and Pevzner, 2002). Multiple Genome Rearrangements and Ancestors (MGRA) is developed to reconstruct ancestral genomes based on multiple breakpoint graphs and is used to analyze rearrangement evolutionary events of seven mammalian genomes (human, chimpanzee, macaque, mouse, rat, dog, and opossum) (Alekseyev and Pevzner, 2009). Decostar (Duchemin et al., 2017) is a software which reconstructs neighborhood relations of ancestral genes aiming at reconstructing the organization of ancestral genomes.

Three is the recognition of the rearrangement events of existing species. Efficient Method to Recover Ancestral Events (EMRAE) is an algorithm which can recognize rearrangement events in evolution described by phylogenetic tree by means of adjacent genes in genomes (Zhao and Bourque, 2009).

MATERIALS AND METHODS

Preliminaries

A genome is composed of several chromosomes, and each chromosome is an ordering of syntenic blocks. For convenience, each syntenic block is recorded by an integer, so a chromosome is represented by a signed permutation $X{=}c_1c_2\cdots g_n$, where $c_i(1{\le}i{\le}n)$ is an integer representing a syntenic block, its sign is assigned with the orientation that is either positive (recorded by c_i) or negative (recorded by $-c_i$). The chromosome $X{=}c_1c_2\cdots c_n$ is the same as $-X = -c_n - c_{n-1}\cdots - c_1$.

A reversal $r\ (i, j)\ (i \le j)$ converts chromosome $X{=}c_1c_2\cdots c_n$ into a new chromosome $X'{=}c_1c_2\cdots{-c_j}{-c_{j-1}}\cdots{-c_{i+1}}{-c_i}c_{j+1}\cdots c_n$, where the reversal is from c_i to c_j.

A translocation event breaks two chromosomes into four segments and then reconnects them into two new chromosomes. Given two chromosomes $X = X_1X_2$ and $Y = Y_1Y_2$, where $X_1{=}x_1x_2\cdots x_{i-1}, X_2{=}x_ix_{i+1}\cdots x_m, Y_1{=}y_1y_2\cdots y_{j-1}$, and $Y_2{=}y_jy_{j+1}\cdots y_n$, a translocation is represented by $tl(i,j)$. X_1 and Y_1 are exchanged to form two new chromosomes $X'{=}Y_1X_2$ and $Y'{=}X_1Y_2$, or X_1 and Y_2 are exchanged to form two new chromosomes $X" = -Y_2X_2$ and $Y" = X_1 - Y_1$.

A transposition event is to exchange two adjacent fragments on one chromosome into a new chromosome. A transposition is represented by $tp(i, j, k)$, i.e., the fragment $c_i\cdots c_j$ of one chromosome inserted into after c_k. If c_k is on the same chromosome ($k > j$ or $k < i$), then the transposition $tp(i, j, k)$ is called intra-chromosomal; otherwise, it is inter-chromosomal. Given a chromosome $X{=}c_1c_2\cdots c_ic_{i+1}\cdots c_{j-1}c_j\cdots c_k\cdots c_n$ and an intra-chromosomal transposition, X is converted into $X'{=}c_1c_2\cdots c_kc_ic_{i+1}\cdots c_jc_{k+1}\cdots c_n$.

A fusion event is to connect two chromosomes into a new chromosome. The fusion acting on chromosomes X_1 and X_2 is represented by $f\ u(X_1, X_2)$ and forming a new chromosome X_1X_2 or $X_1{-}X_2$. A fission is to split a chromosome into two new chromosomes. A fission acting on the chromosome $X = X_1X_2$ is represented by $f\,i(X)$ and forming two new chromosomes X_1 and X_2 (where X_1 and X_2 are non-empty segments).

An adjacency $a(c_i,c_{i+1})$ of genome X is two adjacent integers in one chromosome of X. $a(c_i,c_{i+1})$ is the same as $a(-c_{i+1},-c_i)$. For example, all adjacencies on chromosome X = 1,234 are a(1, 2), a(2, 3), and a(3, 4). For a set of genomes S, an adjacency a is effective w.r.t. S if it belongs to at least one genome and not all genomes. For example, two uni-chromosomal genomes G_1 and G_2, the chromosome X = 1,234 of G_1 and the chromosome Y = 1 – 3 – 24 of G_2, then all effective adjacencies w.r.t. G_1 and G_2 are a(1, 2), a(2, 3), a(3, 4), a(1, −3), and a(−2, 4).

EMRAE

Given a phylogenetic tree T describing the evolution of the genomes G, EMRAE first computes all effective adjacencies w.r.t. G. Then, it predicts the rearrangement events for each edge of T by means of inference rules (will be introduced in the following).

Figure 1 shows a reversal r(2, 3) during the evolution from A to B, where A and B are two uni-chromosomal genomes, and the chromosomes are X = 1,234 and Y = 1 – 3 – 24, respectively. The set of genomes will be divided into two subsets recorded by S_A and S_B after removing the edge e from T. Suppose there is not any rearrangement events inside S_A and S_B. Then, adjacencies a(1, 2) and a(3, 4) can be found in each genome of S_A and not in any one genome of S_B; a(1,−3) and a(−2,4) can be

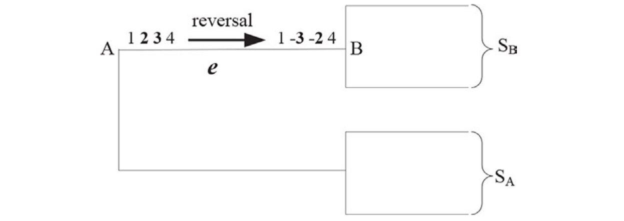


FIGURE 1 | A reversal ***r*** (2, 3) during the evolution from *A* to *B*; *S*\s\do5 **(A)** and *S*\s\do5 **(B)** are two subsets of all leaves species divided by the edge *e*.

found in each genome of S_B and not in any one genome of S_A. In turn, we can utilize the four adjacencies $a(1, 2)$, $a(3, 4)$, $a(1, -3)$, and $a(-2,4)$ to identify a reversal $r(2, 3)$ occurring on the edge e. The EMRAE method infers the rearrangement events by means of the similar rules.

Let $e = (A, B)$ be an edge of T, $G=\{G_1,G_2,\cdots,G_m\}$the genomes of leaves, and $a_1,a_2,\cdots a_i$ the children of A and $b_1,b_2,\cdots b_j$ the children of B. EMRAE first selects a number of adjacencies as candidate adjacencies $Ca(e,A)$ for edge e and node A according the following steps.

1. Find the adjacencies are in each genome of S_A and not in any one genome of S_B, then put them to $Ca(e, A)$;
2. If A is an internal node, find all edges connected with A except e and record them with $e_1,e_2,\cdots,e_k$. For each $e_i=(u_i,A)(1\le i\le k)$, G can be divided into two parts after removing e_i, S_{ui} is the part not including A.
   a. Find the adjacencies that are in one genome of each S_{ui} $(1 \le i \le k)$ and not in any one genome of S_B, then put them to $Ca(e,A)$;
   b. Compute $Ca(e_i, u_i)$ and $Ca(e_i,u)(1\le i\le k)$. For each one $Ca(e_i, u_i)$, find the adjacency a_1 from $Ca(e_i, u_i)$, such that a_1 is not overlap gene with any one adjacency in $Ca(e_i, u)$, a_1 has overlap gene with one adjacency a_2 in each $Ca(e_j,u_j)(1\le j\ne i\le k)$, and a_2 has overlap gene with at least one adjacency in $Ca(e_j, u)$, then put a\s\do5(1) to $Ca(e, u)$.

EMRAE then infers rearrangement from $Ca(e, A)$ and $Ca(e, B)$ for edge $e = (A, B)$ with the help of inference rules in the following section. From the definitions of genome rearrangements, we find that each genome rearrangement can change several adjacencies. For example, each reversal $r(i, j)(i \le j)$ can change two adjacencies $a_1=a(c_{i-1},c_i)$ and $a_2=a(c_j,c_{j+1})$ into $b_1 = a(c_i{-}_1, -c_j)$ and $b_2=a(-c_i,c_{j+1})$. Based on those facts, we obtain the inference rules introduced in the following section.

Inference Rule

Let $e = (A,B)$ be an edge of the phylogenetic tree T. Given adjacencies $a_1 = a\,(c_1{-}_1, c_i)$, $a_2 = a\,(c_{j,}\,c_{j+1})$ in $Ca(e,A)$ and $b_1=a(c_{i-1},-c_j)$, $b_2=a(-c_i,c_{j+1})$ in $Ca(e,B)$, EMRAE infers a reversal $r(i,j)$ from A to B if all genomes are uni-chromosomal or a_1, a_2 are in the same chromosome in S_A and b_1, and b_2 are in the same chromosome in S_B. Otherwise, we infer a translocation $tl(i, j)$. Similarly, given adjacencies $a_1=a(c_{i-1},c_i)$, $a_2=a(c_jc_{j+1})$ in $Ca(e,A)$ and $b_1=a(c_{i+1},c_{j+1})$, $b_2=a(c_j,c_i)$ in $Ca(e,B)$, EMRAE infers a translocation $tl(i,j)$, or a reversal for a_1, a_2 in $Ca(e,A)$ and adjacencies b_1, b_2 in $Ca(e,B)$.

Assume that there are adjacencies $a_1=a(c_{i-1},c_i)$, $a_2=a(c_j,c_{j+1})$, and $a_3=a(c_k,c_{k+1})$ in $Ca(e,A)$ and $b_1=a(c_{i-1},c_{j+1})$, $b_2=a(c_k,c_i)$, and $b_3=a(c_j,c_{k+1})$ in $Ca(e,B)$. EMRAE can predict a transposition $tp(i,j,k)$ during the evolution from A to B if all genomes are uni-chromosomal. Otherwise, suppose m genomes in S_A have a_1 and a_2, then EMRAE can predict a transposition $tp(i,j,k)$ if there are at least $m/2$ genomes such that the four integers of a_1 and a_2 on the same chromosome, or there are at least $m/2$ genomes such that the four integers of a_2 and a_3 on the same chromosome.

Assume that there is $a=a(c_i,c_j)$ in $Ca(e,A)$. EMRAE can predict a fission that splits the adjacency $a=a(c_i,c_j)$ if a is sign-compatible for each genome G_k in S_B. The fusion from A to B can be seen as a fission from B to A.

Recovering the Genome Rearrangements Based on Phylogenetic Tree

EMRAE can not identify the rearrangement occurring in the frontier of genomes. We take **Figure 2**, for example, where species A, B, and C are uni-chromosomal genomes A = 1,234, $B = -2 - 134$, and C = 1,234. A reversal r(1,2) has occurred in the evolution from A to B. EMRAE can compute the candidate adjacencies $a(-1,3)$ for $Ca(e_1,B)$ and $a(2,3)$ for $Ca(e_1,A)$. So, EMRAE can not infer the reversal r(1,2) on the edge e_1 according to the candidate adjacencies.

We improve EMRAE so that the improved method (called RGRPT) is able to infer the rearrangement events occurring in the frontier region. The inference rule of RGRPT is the same as that of EMRAE. The difference between RGRPT and EMRAE is that they have different candidate adjacencies. RGRPT puts 0 to the head and tail for each chromosome, so there will be added a lot of adjacencies for each genome. For example, considering the uni-chromosomal genomes X = 1,234 and $Y = -2\ -134$, the two additional candidate adjacencies $a(0,1)$ and $a(0,-2)$ are added.

RGRPT adds candidate adjacencies in the step b of EMRAE. For each one $Ca(e_i,u_i)$ and an adjacency a_1 from $Ca(e_i,u_i)$, if there is an adjacency a_2 in each $Ca(e_j,u_j)(1\le j\ne i\le k)$ such that a_1 with a_2 has overlap gene, then put a_1 to $Ca(e,u)$.

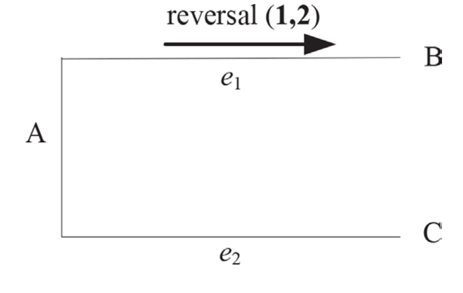


FIGURE 2 | The tree topology with two taxa (**B** and **C**).

RESULTS

All of the experiments were performed on a computer with Intel Vostro 14 2.0 GHz CPU, 4 GB RAM, and 500 GB Hard Disk Drives (HDD). The operating system was Win10 64 bit with Java 1.6 installed. RGRPT was written in Java.

We tested RGRPT with both simulated data and the practical data (i.e., real biological data) introduced by the following section.

Simulated Data

Here, we start with an uni-chromosomal genome as the ancestor, and it evolves along the phylogenetic tree with *n* taxa whose topology sees the **Figure 3**.

We generate two simulated data sets in order to test the affectivity of RGRPT. One of them is created from the phylogeny only with reversals events. The other data set is generated from the phylogeny with kinds of events, including reversals, translocation, transposition, fusion, and fission, and the quantity of those events is in a certain ratio. The two data sets can test the ability of methods to recover the simple and the complex evolution histories. First data set is created just using reversal events. Since the reversal on only one gene is rare (Korbel et al., 2007), we set the ratio of reversal on one gene and on more than one gene as 1:3. The number of leaves is from 3 to 10 with step 1. For each number of leaves, the ancestor genome with *m* gene, where *m* from 50 to 150 with step 10. Each edge will happen *k* reverse, where *k* is random integer number from 3 to 10. So, there are 11 groups data for each leaf number. Sensibility is the percentage of correctly predicted events in all practical events. Specificity is the percentage of correctly predicted events in all predicted events. We compute the sensibility and specificity for RGRPT and EMRAE for each group data. **Table 1** shows the average sensitivity and specificity for each leaf number. The second column of the table records the number of all events, and its last row records the average values.

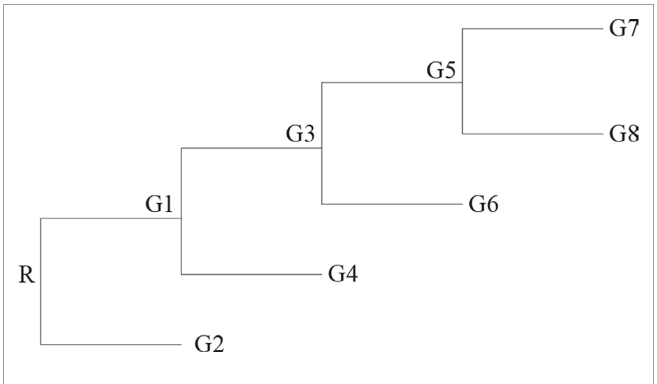


FIGURE 3 | The topology used to generate the simulation data.

Table 1 shows that RGRPT achieves higher sensibility than EMRAE, and RGRPT achieves comparable specificity with EMRAE. Obviously, RGRPT can distinguish more actually occurred events than EMRAE. So, the experimental results show that the RGRPT is more efficient than EMRAE for predicting reversal events.

Second data set is generated by using all events, i.e., reversal, translocation, transposition, fusion, and fission. The reversals are generally more than the other rearrangement events. The fusions and the fissions are very rare; so, we record the number of the two events together. Here, we set the ratio of those events as 10:2:2:0.1. The ancestor genome has 5 chromosomes and each chromosome with 100 genes. The ancestor genome evolves along the topology with four leaves (see **Figure 3**). Each edge happen *k* events, where *k* is random number from 1 to μ and μ is 6, 12, 18, and 24. For each μ, it runs 10 times; so, we can obtain 10 groups data for each μ. **Table 2** shows the average of 10 groups data for each μ. This table indicates that the RGRPT is more efficient than EMRAE for predicting all events.

Practical Data

The practical data is from the paper (Zhao and Bourque, 2009). It contains six mammalian genomes, i.e., human, chimpanzee, rhesus monkey, mouse, voles, and dog. The data are created from two different levels of resolution 10 kb and 50 kb. **Figure 4** is the tree describing the phylogeny of species. The results are shown in **Tables 3** and **4**. EM and RG represent EMRAE and RGRPT respectively, and Rev, Tloc, Tran, Fus, and Fis represent reversal,

TABLE 1 | Results of EMRAE and recovering the genome rearrangements based on phylogenetic tree algorithms in predicting reversal events.

Leaves	Reversal	Sensibility		Specificity	
		EMRAE	RGRPT	EMRAE	RGRPT
3	24	64%	76%	89%	90%
4	39	65%	76%	94%	94%
5	45	61%	72%	92%	93%
6	59	57%	66%	90%	90%
7	69	54%	65%	92%	91%
8	79	59%	80%	92%	92%
9	92	55%	63%	90%	90%
10	104	55%	62%	89%	89%
Mean		58.7%	70%	91%	91.1%

TABLE 2 | Results of EMRAE and recovering the genome rearrangements based on phylogenetic tree algorithms in predicting all events.

Events of each edge	All events	Sensibility		Specificity	
		EMRAE	RGRPT	EMRAE	RGRPT
6	19	75.8%	85.7%	95.8%	96.2%
12	29	74.2%	80.3%	97%	96.5%
18	38	53.5%	58.1%	95.4%	96.7%
24	50	47.7%	50.5%	94.9%	94.1%
Mean		62.8%	68.7%	95.8%	95.9%

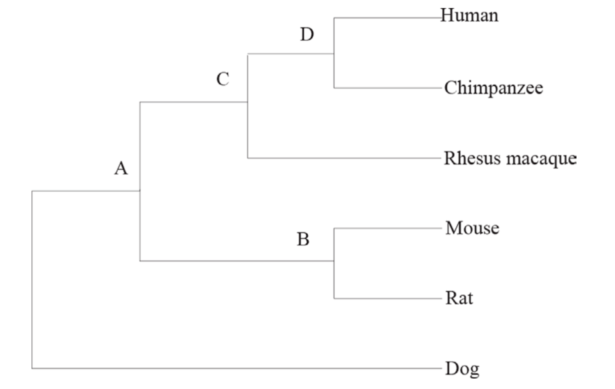


FIGURE 4 | The tree describing the phylogeny of mammalian species.

translocation, transposition, fusion, and fission, respectively. Each row in the table records the ancestor rearrangement events of the edge. For example, the values in the human row are the rearrangement events from D to human; the values in MR row are the rearrangement events from A and B.

At 10 kb resolution, the RGRPT algorithm predicts 1,157 ancestor rearrangement events, including 858 reversals, 16 translocations, 249 transpositions, and 34 fusions and fissions. It identifies 48 rearrangement events more than the EMRAE. The reversal events are in the majority in all predicted events. At 50 kb resolution, the RGRPT algorithm predicts 475 ancestor rearrangement events, including 332 reversals, 13 translocations, 94 transpositions, and 36 fusion and fissions. RGRPT identifies 21 rearrangement events more than EMRAE algorithm. The rearrangement events identified in the rat

TABLE 3 | Genome rearrangement predictions of EMRAE and recovering the genome rearrangements based on phylogenetic tree at 10 kb resolution.

Species	Rev		Tloc		Tran		Fus/Fis		Total events	
	EM	RG	EM	RG	EM	RG	EM	RG	EM	RG
Human	12	13	0	0	4	5	0	0	16	18
HC	29	32	0	0	15	15	0	1	44	48
HCP	83	84	0	0	8	10	2	8	93	102
Chimp	17	19	0	0	7	8	1	1	25	28
Rhesus	49	50	0	0	40	42	1	2	90	94
Mouse	90	95	3	3	10	13	5	5	108	116
Rat	227	233	0	0	127	129	3	3	357	365
MR	140	143	2	3	9	10	0	0	151	156
Dog	184	189	10	10	17	17	14	14	225	230
Total	831	858	15	16	237	249	26	34	1,109	1,157

TABLE 4 | Genome rearrangement predictions of EMRAE and recovering the genome rearrangements based on phylogenetic tree at 50 kb resolution.

Species	Rev		Tloc		Tran		Fus/Fis		Total events	
	EM	RG	EM	RG	EM	RG	EM	RG	EM	RG
Human	2	2	0	0	1	1	0	0	3	3
HC	19	19	0	0	4	4	1	1	24	24
HCP	27	29	0	0	5	6	2	6	34	41
Chimp	17	19	0	0	7	8	1	1	25	28
Rhesus	22	23	0	0	6	7	1	3	29	33
Mouse	25	27	3	3	0	0	5	6	33	36
Rat	128	131	0	0	65	65	5	5	198	201
MR	41	42	2	2	2	2	0	0	45	46
Dog	46	47	7	8	8	8	13	14	74	77
Total	322	332	12	13	92	94	28	36	454	475

edge are mostly in all edges either at 10 kb resolution or at 50 kb resolution. The syntenic blocks of genomes at 10 kb resolution are more than the syntenic blocks of genomes at 50 kb resolution. The fact reduces the recognized rearrangement events at 10 kb resolution that are more than the recognized rearrangement events at 50 kb resolution. Experiments show that RGRPT can recover more ancestor events than EMRAE.

DISCUSSION

This paper proposes a new method, RGRPT, to infer ancestor rearrangement events. RGRPT takes a phylogenetic tree describing the evolution of species and the genomes of species as input. Experiments on the simulated data and practical data show that RGRPT is more efficient than EMRAE and can recover more ancestor rearrangement events than EMRAE. RGRPT provides a method for us to research the genome rearrangement of species. We can use RGRPT to recognize the ancestral genome rearrangement for the evolution of other species in future (Tian et al., 2018).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/wangjuanimu/data-of-genome-rearrangement.

AUTHOR CONTRIBUTIONS

JW proposed and implemented the RGRPT method. JW and BC designed all experiments. All authors participated in the designing the algorithm and writing the paper.

FUNDING

The work was supported by the National Natural Science Foundation of China (61661040, 61661039, 61571163, 61532014, 61671189, 91735306, 61751104); the National Key Research and Development Plan Task of China (Grant No. 2016YFC0901902).

REFERENCES

Alekseyev, M. A., and Pevzner, P. A. (2009). Breakpoint graphs and ancestral genome reconstructions. *Genome Res.* 19 (5), 943–957.

Blanchette, M., Bourque, G., and Sankoff, D. (1997). Breakpoint phylogenies. *Genome Inform. Ser. Workshop Genome Inform.* 8, 25–34.

Bourque, G., and Pevzner, P. A. (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.* 11 (1), 26–36.

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019). Metsigdis: a manually curated resource for the metabolic signatures of diseases. *Briefings Bioinf.* doi: 10.1093/bib/bbx103

Dong, S., Zhao, C., Fei, C., Liu, Y., Zhang, S., Hong, W., et al. (2018). The complete mitochondrial genome of the early flowering plant nymphaea colorata is highly repetitive with low recombination. *Bmc Genomics* 19 (1), 614–626.

Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Brard, S., Chauve, C., et al. (2017). Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biol. Evol.* 9 (5), 1312–1319.

Erdös, P. L., Soukup, L., and Stoye, J. (2011). Balanced vertices in trees and a simpler algorithm to compute the genomic distance. *Appl. Math. Lett.* 24 (1), 82–86.

Feijão, P., and Meidanis, J. (2011). Scj:a breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8 (5), 1318–1329.

Hannenhalli, S. (1995). Polynomial-time algorithm for computing translocation distance between genomes. *Discrete Appl. Math.* 71 (1–3), 137–151.

Hannenhalli, S., and Pevzner, P. A. (1999). Transforming cabbage into turnip:polynomial algorithm for sorting signed permutations by reversals. *J. Acm* 46 (1), 1–27.

Hilker, R., Sickinger, C., Pedersen, C. N., and Stoye, J. (2012). Unimog–a unifying framework for genomic distance calculation and sorting based on dcj. *Bioinformatics* 28 (19), 2509.

Jakub, K., Robert, W., Braga, M. D. V., and Jens, S. (2011). Restricted dcj model: rearrangement problems with chromosome reincorporation. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 18 (9), 1231–1241.

Kaplan, H., Shamir, R., and Tarjan, R. E. (2000). Faster and simpler algorithm for sorting signed permutations by reversals. *SIAM J. Comput.* 29 (3), 880–892.

Korbel, J. O., Urban A. E., Affourtit J. P., Godwin B., Grubert F., Simons J. F. et al. (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318 (5849), 420–426.

Liu, X., Zhu, D., Ma, S., Li, Z., and Wang, L. (2004). An o(n2) algorithm for sorting oriented genomes by translocations. *Chin. J. Comput.* 27 (10), 1354–1360.

Lu, C. L., Huang, Y. L., Wang, T. C., and Chiu, H. T. (2006). Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. *Bmc Bioinf.* 7 (1), 295.

Ma, J., Zhang, L., Suh, B., and e. a. Raney, B. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16 (12), 1557–1565.

Sankoff, D., and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.* 5, 555–570.

Sophia, Y., Oliver, A., and Richard, F. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* 21 (16), 3340–3346.

Tian, Z., Teng, Z., Cheng, S., and Guo, M. (2018). Computational drug repositioning using meta-path-based semantic network analysis. *BMC Syst. Biol.* 12 (S9), 134.

Xu, Y., Wang, Y., Luo, J., Zhao, W., and Zhou, X. (2017) Deep learning of the splicing (epi)genetic code reveals a novel candidate mechanism linking histone modifications to esc fate decision. *Nucleic Acids Res.* 45 (21), 12100–12112.

Yang, N., Hu, F., Zhou, L., and Tang, J. (2014). Reconstruction of ancestral gene orders using probabilistic and gene encoding approaches. *PLoS One* 9 (10), e108796.

Yen-Lin, H., Chen-Cheng, H., Chuan Yi, T., and Chin Lung, L. (2010). Sort2: a tool for sorting genomes and reconstructing phylogenetic trees by reversals, generalized transpositions and translocations. *Nucleic Acids Res.* 38 (Web Server issue), W221–W227.

Zhao, H., and Bourque, G. (2009). Recovering genome rearrangements in the mammalian phylogeny. *Genome Res.* 19 (5), 934–942.

Zhu, D., and Ma, S. (2002). An improved algorithm for the translocation sorting problem of genomes. *Chin. J. Comput.* 25 (2), 189–196.

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.

ORIGINAL RESEARCH
published: 12 November 2019
doi: 10.3389/fgene.2019.01136

Integrating the Ribonucleic Acid Sequencing Data From Various Studies for Exploring the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids and Their Functions

Zhijie Han[1,2]*, Jiao Hua*[3]*, Weiwei Xue*[2] *and Feng Zhu*[1,2]*

[1] *College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China,* [2] *School of Pharmaceutical Sciences, Chongqing University, Chongqing, China,* [3] *School of Mathematics, Harbin Institute of Technology, Harbin, China*

Multiple sclerosis (MS) is a chronic fatal central nervous system (CNS) disease involving in complex immunity dysfunction. Recently, long noncoding RNAs (lncRNAs) were discovered as the important regulatory factors for the pathogenesis of MS. However, these findings often cannot be repeated and confirmed by the subsequent studies. We considered that the small-scale samples or the heterogeneity among various tissues may result in the divergence of the results. Currently, RNA-seq has become a powerful approach to quantify the abundances of lncRNA transcripts. Therefore, we comprehensively collected the MS-related RNA-seq data from a variety of previous studies, and integrated these data using an expression-based meta-analysis to identify the differentially expressed lncRNA between MS patients and controls in whole samples and subgroups. Then, we performed the Jensen-Shannon (JS) divergence and cluster analysis to explore the heterogeneity and expression specificity among various tissues. Finally, we investigated the potential function of identified lncRNAs for MS using weighted gene co-expression network analysis (WGCNA) and gene set enrichment analysis (GSEA), and 5,420 MS-related lncRNAs specifically expressed in the brain tissue were identified. The subgroup analysis found a small heterogeneity of the lncRNA expression profiles between brain and blood tissues. The results of WGCNA and GSEA showed that a potential important function of lncRNAs in MS may be involved in the regulation of ribonucleoproteins and tumor necrosis factor cytokines receptors. In summary, this study provided a strategy to explore disease-related lncRNAs on genome-wide scale, and our findings will be benefit to improve the understanding of MS pathogenesis.

Keywords: ribonucleic acid sequencing, multiple sclerosis, long non-coding ribonucleic acids, meta-analysis, function analysis

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University,
China

Reviewed by:
Guiyou Liu,
Chinese Academy of Sciences,
China
Zhi-Liang Ji,
Xiamen University, China

****Correspondence:***
Feng Zhu
zhufeng@zju.edu.cn;
prof.zhufeng@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *22 July 2019*
Accepted: *18 October 2019*
Published: *12 November 2019*

Citation:
Han Z, Hua J, Xue W and Zhu F (2019) Integrating the Ribonucleic Acid Sequencing Data From Various Studies for Exploring the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids and Their Functions. Front. Genet. 10:1136. doi: 10.3389/fgene.2019.01136

INTRODUCTION

Multiple sclerosis (MS) is a chronic fatal neurodegenerative disease involving in complex immunity [central nervous system (CNS)] (Sospedra and Martin, 2005; Frohman et al., 2006; Li et al., 2018). Based on the 2014 statistics of the Atlas of MS investigation, the estimated number of the people afflicted with the MS worldwide has reached approximately 2.3 million (Browne et al., 2014). Although much remains unknown about the molecular etiology of MS, more and more studies showed that the dysregulation of transcriptional processes could potentially contribute to the pathogenesis of MS (Li et al., 2017; Selmaj et al., 2017; Angerer et al., 2018; Cheng et al., 2018; Han et al., 2018b; Zhang et al., 2019).

Recently, long noncoding RNA (lncRNA), one of the non-protein-coding genes whose transcripts are longer than 200 nucleotides, has been discovered as the important regulatory factor of immune system and pathogenesis of CNS disorders including MS (Gomez et al., 2013; Ng et al., 2013; Dong et al., 2015; Cheng et al., 2016; Santoro et al., 2016; Zhang et al., 2016; Chen et al., 2017; Eftekharian et al., 2017; He et al., 2017; Cheng et al., 2018; Yin et al., 2019). However, for MS, these results often cannot be repeated and confirmed by subsequent study. For example, multiple variants of the lncRNA antisense non-coding RNA in the INK4 locus (*ANRIL*) are found significantly associated with the risk of MS through the haplotype analysis of blood samples (Rezazadeh et al., 2018). But following study reveals that the function of *ANRIL* does not contribute the pathogenesis of MS in blood, cortex, and cerebellum tissues (Pahlevan Kakhki et al., 2018). Study showed a significant upregulation of lncRNA *MALAT1* in MS blood tissues (Cardamone et al., 2019), while the expression of *MALAT1* was found markedly decreased in MS brain by the subsequent study (Masoumi et al., 2019). Moreover, another study found that *MALAT1* is not significantly differentially expressed between MS patients and controls (Gharesouran et al., 2019). We considered that the small-scale samples or the heterogeneity among various tissues may result in the divergence of the results.

Currently, specifically for lncRNAs, using RNA-seq data to quantify abundance of the transcripts has become very powerful approach compared with the traditional ones (e.g., gene microarray) (Wang et al., 2009). Particularly, almost all of the expression of the known lncRNA transcripts can be measured using RNA-seq data, but this proportion is just approximately 0.1 to 10.6% by the method of probe re-annotation using various types of microarrays (Du et al., 2013; Fang et al., 2018; Yang et al., 2019). Moreover, lncRNA abundance quantification using RNA-seq data also shows higher accuracy based on its deep read coverage, while the re-annotation approach only requires the sequence match of 1 to 4 probes when quantifies lncRNA abundance (Du et al., 2013; Gellert et al., 2013; Li et al., 2019). A previous study reported that by paying attention to some aspect of library and sequencing process [i.e., poly-A tail selection, paired-end sequencing, and sequencing of double-stranded complementary DNA (cDNA)], the lncRNAs are more easily and more accurately identified through RNA-seq (Ilott and Ponting, 2013).

In this study, we thus selected all MS-related RNA-seq data in a variety of studies by searching three authoritative public databases: GEO DataSets (Barrett et al., 2013), EBI-EMBL ArrayExpress (Athar et al., 2019), and DDBJ Sequence Read Archive (Ogasawara et al., 2013) using the keyword "multiple sclerosis." Then, we used these RNA-seq data to perform expression quantification of the lncRNA in each of the selected studies. Next, we integrated the lncRNA expression results of all selected studies by an expression-based meta-analysis to identify the significantly differentially expressed lncRNAs between MS patients and controls. Further, we explored their heterogeneity and expression specificity among various tissues. After that, the weighted gene co-expression network analysis (WGCNA) was performed using the expression data of lncRNAs and protein-coding genes to identify the significant modules for MS. The expression of the protein-coding genes was calculated using the same approach on lncRNA. Finally, we conducted gene set enrichment analysis (GSEA) on the co-expressed protein-coding genes in each significant module to infer the function of the differentially expressed lncRNAs potentially contributing to the pathogenesis of MS.

MATERIALS AND METHODS

Selection of the Multiple Sclerosis-Related Ribonucleic Acid Sequencing Datasets and Studies

We used the keyword "multiple sclerosis" to search all the possible MS-related RNA-seq datasets in three authoritative databases: GEO DataSets (Barrett et al., 2013), EBI-EMBL ArrayExpress (Athar et al., 2019), and DDBJ Sequence Read Archive (Ogasawara et al., 2013). The search was performed before the last update of the databases on May 16 2019. Then, we selected the suitable datasets using four criteria: 1) the organism in the dataset is the human being; 2) the study in the dataset is designed using the case-control method; 3) the dataset has provided the FASTQ data; (4) the FASTQ data in the dataset is not generated by metagenome, whole genome, or whole exome sequencing. Finally, the studies from these datasets based on various tissues were selected. **Figure 1** showed the workflow.

Quantification of Long Noncoding Ribonucleic Acid Sequencing Abundance Using Ribonucleic Acid Sequencing Sequencing Data

We first downloaded the sequence data of these studies by *Prefetch* and converted them into FASTQ files using *fastq-dump* tool of the SRA Toolkit software (Leinonen et al., 2011). Next, we downloaded the reference sequences of lncRNA and protein-coding transcripts in FASTA format from NONCODE (version 5) (Fang et al., 2018) and Ensembl (release 91) (Aken et al., 2017), respectively, and further merged the two FASTA format files. Particularly, NONCODE is one of the most complete and well-annotated databases of the noncoding RNAs, and we obtained a total of 172,216 transcript sequences of 96,308 human

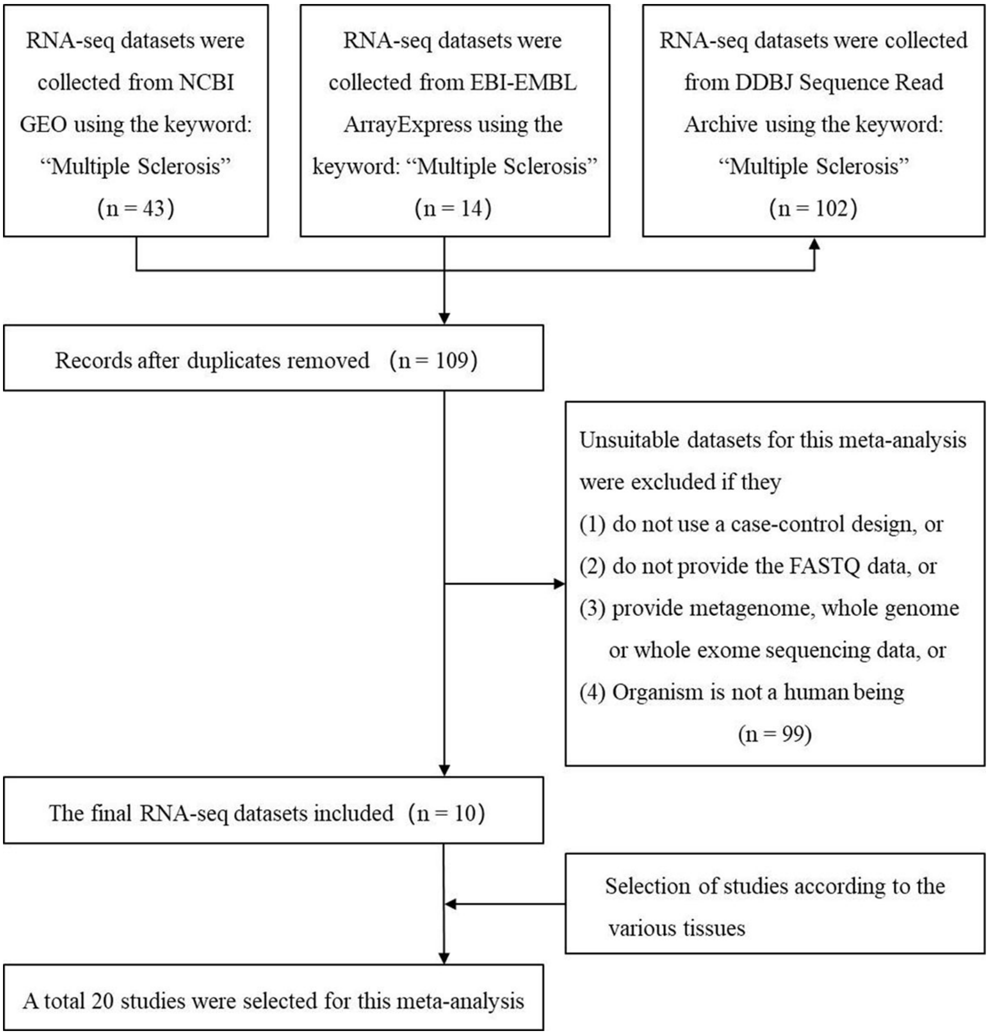

RNA-seq datasets were collected from NCBI GEO using the keyword: "Multiple Sclerosis" (n = 43)
RNA-seq datasets were collected from EBI-EMBL ArrayExpress using the keyword: "Multiple Sclerosis" (n = 14)
RNA-seq datasets were collected from DDBJ Sequence Read Archive using the keyword: "Multiple Sclerosis" (n = 102)
Records after duplicates removed (n = 109)
Unsuitable datasets for this meta-analysis were excluded if they
(1) do not use a case-control design, or
(2) do not provide the FASTQ data, or
(3) provide metagenome, whole genome or whole exome sequencing data, or
(4) Organism is not a human being
(n = 99)
The final RNA-seq datasets included (n = 10)
Selection of studies according to the various tissues
A total 20 studies were selected for this meta-analysis

FIGURE 1 | The flow chart of selecting the RNA sequencing (RNA-seq) datasets and studies which are used to identify the multiple sclerosis-related long noncoding RNAs.

lncRNA genes from it. Ensembl aggregated the cDNA data from National Center for Biotechnology Information (Sayers et al., 2019), UniProt (UniProt, 2015), Genome Reference Consortium (Church et al., 2011), and UCSC Genome Browser (Kent et al., 2002) databases. After removing the pseudogenes, we obtained a total of 160,040 transcript sequences of 22,810 human protein-coding genes from it. Then, we performed the quantification of the lncRNA and protein-coding transcripts simultaneously by mapping the RNA-seq reads of each study to the merged reference sequence (pseudoalignment) and calculating the count values using *Kallisto* software (Bray et al., 2016). *Kallisto* is a fast and highly accurate quantification tool for transcript abundance through k-mer lookup technique. Here, the merged reference sequences have been processed into a transcriptome index to conduct the pseudoalignment which has the same effect as the reads alignment to a given reference genome in the traditional transcript-level RNA-seq processing but can substantially reduce calculation time. For the paired-end sequencing samples, the arguments were set to defaults, i.e., the number of bootstrap samples (-b) equals 0 and the number of threads (-t) equals 1. For the single-end sequencing samples, besides these default parameter settings, we set the estimated average fragment length (-l) and the standard deviation of fragment length (-s) to 200 and 20, respectively, according to Kallisto's recommended parameters. Finally, based on the annotation file "Transcript2Gene," we integrated transcript-level count values of lncRNAs to calculate their corresponding gene-level count values using the R package "tximport" (Soneson et al., 2015).

Heterogeneity Test and Meta-Analysis

To identify the significantly differentially expressed lncRNAs between MS patients and controls, we calculated and integrated the results of each study by a meta-analysis. These analyses were conducted using R package "MetaOmics," which is a comprehensive analytical pipeline to meta-analyze multiple transcriptomic studies (Ma et al., 2019). This meta-analysis includes a normalization process same as the edgeR's strategy and a "AW-Fisher" method to integrate data (Bullard et al., 2010; Robinson et al., 2010; Ma et al., 2019). First, we calculated the two parameters, I^2 and P value, to measure the lcnRNA expression heterogeneity by the Cochran's Q Statistics, which is based on a chi-square test with $k - 1$ degrees of freedom (k equals to the number of studies used for the meta-analysis). According to the previous studies, the heterogeneity was considered as statistically significant when $I^2 > 50\%$ and $P < 0.01$ (Han et al., 2015; Li et al., 2016; Liu et al., 2017; Han et al., 2018a; Xue et al., 2018). Then, the meta-analysis was performed for each of these lncRNAs based on their count values. Particularly, the random effect model (REM) and fixed effect model (FEM) were used, respectively, for the lncRNAs with a significant heterogeneity or not. Using the REM in meta-analysis can reduce bias of the results (Kim et al., 2015; Szajewska and Kolodziej, 2015). We calculated standardized mean difference (SMD) with its 95% confidence interval (CI) to identify the differentially expressed lncRNA between the MS patients and controls (95% CI of SMD does not include zero, FDR adjusted $P < 0.05$). The SMD is given by the mean difference between case and control divided by the standard deviation and applies to meta-analysis when the outcome is continuous variable (e.g., expression level). Moreover, since all these samples can be split into brain and blood, we performed the meta-analysis for the two subgroups, and explored the differential expression pattern of the MS-related lncRNAs between brain and blood.

In addition, we further explored the specific target genes of the lncRNAs using LncRNA2Target v2.0 database which is authoritative source including 152,137 lncRNA-target relationships confirmed by the knockdown or overexpression analysis and binding experimental technologies, and provides web interface for searching the targets by a particular lncRNA (Cheng et al., 2019).

Tissue Specificity Analysis of the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids

We explored the tissue expression specificity of the significantly differentially expressed lncRNAs in MS, which was important aspects of neurological disease research (usually, specifically expressed in CNS system) (Fatica and Bozzoni, 2014; He et al., 2017; Tang et al., 2019b). For this purpose, lncRNA expression data were first downloaded from the NONCODE, which were involved in primary human tissue/cell line (e.g., brain, heart, breast, lung, liver, foreskin, lung, lymph node, colon, skeletal muscle, leukocyte, HeLa cells, and fibroblasts, etc.). Then, we extracted the expression data of various tissues by the corresponding differentially expressed lncRNAs in brain, blood, and whole sample, respectively, and stored them in three independent sets. Further, based on these data, we used the Jensen-Shannon (JS) divergence, an entropy-based approach, to calculate a tissue specificity score of the differentially expressed lncRNAs according to previous study (Cabili et al., 2011). Briefly, the lncRNA expression vectors were converted to abundance density, and the distance between two tissue expression patterns was defined as the square root of JS divergence. The tissue specificity of a lncRNA expression pattern was measured through the distance between expression patterns across various tissues and predefined extreme pattern in which the lncRNA is uniquely expressed in one tissue (1 minus the distance). Thus, the metric of tissue specificity ranged from 0 to 1. The nearer the score to one, the stronger the tissue specificity becomes. Finally, using the same data, we performed the cluster analysis with Manhattan distance for differentially expressed lncRNAs in brain, blood and whole sample by R package "gplots."

Inferring the Functions of Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids by Weighted Gene Co-Expression Network Analysis

To infer the potential biological functions of these significantly differentially expressed lncRNAs in MS, we used WGCNA approach to determine the co-expression profile of these MS-related lncRNAs and protein-coding genes, and further performed the GSEA by the co-expressed protein-coding genes. First, in the same way used for identifying MS-related lncRNAs, we quantified the abundance of the protein-coding genes and identified the significantly differentially expressed genes by a meta-analysis. Second, we constructed the co-expression network by integrating the count values of the differentially expressed lncRNAs and protein-coding genes using the R package "WGCNA" (Langfelder and Horvath, 2008). Particularly: 1) we conducted the sample clustering to check if there were any outlier samples using "hclust" function of R package "WGCNA"; 2) after quality control, we used "pickSoftThreshold" function of R package "WGCNA" to calculate the satisfactory soft threshold power β for ensuring the scale-free topology characteristics of the co-expression network; 3) based on the β value, we applied the Pearson's method to calculate an adjacency matrix which includes the weighted correlation of all gene pairs; 4) by adjacency matrix, we used the dynamic cut-tree algorithm to construct a hierarchical clustering dendrogram and identified the co-expression modules where genes have high topological overlap with each other. Finally, we assessed the significance of the modules for MS by measuring two indices. Particularly, one of the indices is correlation between module membership (i.e., intramodular connectivity) and gene significance for MS. High correlation means that the hub genes (i.e., the genes with high connectivity in a co-expression module) of the corresponding module also tend to be highly correlated with disease states (MS or healthy) (Langfelder and Horvath, 2008). The other is the average correlation of the genes in each module with disease states. This was also applied to assess association of each module with the platforms and the tissue types, respectively.

Pathway Analysis of the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids by Gene Set Enrichment Analysis

Based on the two indices of module significance, we selected the most significant modules of disease states to investigate the lncRNA functions in MS by GSEA. We first extracted the ID numbers of the protein-coding genes co-expressed with lncRNAs in the modules. Then, we downloaded the signaling pathway data from two common databases, Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). GO is a public resource of data on the gene functions in the biological process, molecular function, and cellular component (The Gene Ontology, 2017), and KEGG is comprehensive database which integrates the information of genes involved in signaling pathways, cellular processes, human diseases, etc. (Kanehisa et al., 2017). Finally, we used the co-expressed protein-coding genes and the signaling pathway data to conduct the GSEA of the most significant modules using R package "clusterProfiler" (Yu et al., 2012). The adjusted P value calculated by the multiple testing (Benjamini-Hochberg method) was set at less than 0.05 as the threshold of significance.

RESULTS AND DISCUSSION

Results of Study Selection and Long Noncoding Ribonucleic Acid Abundance Quantification

Using keyword search and quality filtering, we identified ten MS-related RNA-seq datasets including: GSE60424, GSE66573, GSE66763, GSE89843, GSE100297, GSE120411, GSE111972, GSE123496, GSE77598, and SRP132699 from three authoritative databases. We found that the library preparation and sequencing methods in most of these datasets meet one/multiple requirements for improving the lncRNAs identification (i.e., poly-A tail selection, paired-end sequencing, and sequencing of double-stranded cDNA). Then, after the investigating the source of samples, we found that these datasets are involved in eight brain tissues (optic chiasm, corpus callosum, occipital cortex, astrocytes, frontal cortex, hippocampus, internal capsule, parietal cortex) and seven blood tissues (B cell, T cell, monocyte, platelets, neutrophils, natural killer cell, and whole blood). According to the various tissues, we selected a total of 20 studies (207 MS cases and 348 controls) for the following analysis. The detailed information of each study was shown in **Table 1**. Finally, we downloaded RNA-seq data of the samples in each study, and used them to measure lncRNA expression (count values) using *Kallisto* (Bray et al., 2016) and R package "tximport" (Soneson et al., 2015). In total, lncRNA abundance in 555 samples was quantified.

Heterogeneity Test and Meta-Analysis

Based on the count values of the 96,308 lncRNAs in 20 studies, the meta-analysis was performed to calculate SMD value with its 95% CI for each lncRNA using REM/FEM. Heterogeneity test showed that only about 2.90% lncRNAs have the significant heterogeneity ($I^2 > 50\%$ and $P < 0.01$). Therefore, the homogeneous unbiased results could be identified in >97% lncRNAs by FEM. For the remaining lncRNAs of significant heterogeneity, REM could reduce resulting bias. In total, 5,420 lncRNAs were identified significantly differentially expressed between MS cases and controls, which included 368 downregulated and 5,052 upregulated lncRNAs (shown in **Figure 2A** and **Supplementary Table S1**). For example, the **Figure 2B** exhibited the meta-analysis results of the lncRNA NONHSAG108980.1 which shows the

TABLE 1 | Summary of the 20 selected studies for the meta-analysis. NK, natural killer cell.

Study Number	Dataset	Tissue	Year	No. of cases	No. of controls	Sequencing platform	RNA-seq library type		
							Poly-A tail select	Sequencing of double-stranded cDNA	Read type
1	GSE60424	B-cells	2014	6	4	Illumina HiScanSQ	Yes	Not described	Paired-end
2	GSE60424	Monocytes	2014	6	4	Illumina HiScanSQ	Yes	Not described	Paired-end
3	GSE60424	Neutrophils	2014	6	4	Illumina HiScanSQ	Yes	Not described	Paired-end
4	GSE60424	NK	2014	3	4	Illumina HiScanSQ	Yes	Not described	Paired-end
5	GSE60424	T-cells	2014	12	8	Illumina HiScanSQ	Yes	Not described	Paired-end
6	GSE60424	Whole blood	2014	6	4	Illumina HiScanSQ	Yes	Not described	Paired-end
7	GSE66573	Whole blood	2015	6	8	Illumina HiSeq 2500	Yes	Yes	Paired-end
8	GSE66763	T-cells	2015	10	6	Illumina HiSeq 2500	Not described	Not described	Paired-end
9	GSE77598	Monocytes	2016	5	3	Illumina HiSeq 2000	Not described	Not described	Paired-end
10	GSE89843	Platelets	2017	58	234	Illumina HiSeq 2500	Yes	Yes	Single-end
11	GSE100297	Optic chiasm	2017	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end
12	GSE111972	Corpus callosum	2018	10	11	Illumina NextSeq 500	Yes	Not described	Single-end
13	GSE111972	Occipital cortex	2018	5	5	Illumina NextSeq 500	Yes	Not described	Single-end
14	GSE120411	Astrocytes	2018	24	18	Illumina HiSeq 2500	Yes	Not described	Single-end
15	SRP132699	Monocytes	2018	20	5	Illumina HiSeq 2500	Not described	Not described	Single-end
16	GSE123496	Corpus callosum	2019	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end
17	GSE123496	Frontal cortex	2019	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end
18	GSE123496	Hippocampus	2019	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end
19	GSE123496	Internal capsule	2019	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end
20	GSE123496	Parietal cortex	2019	5	5	Illumina HiSeq 3000	Yes	Yes	Single-end

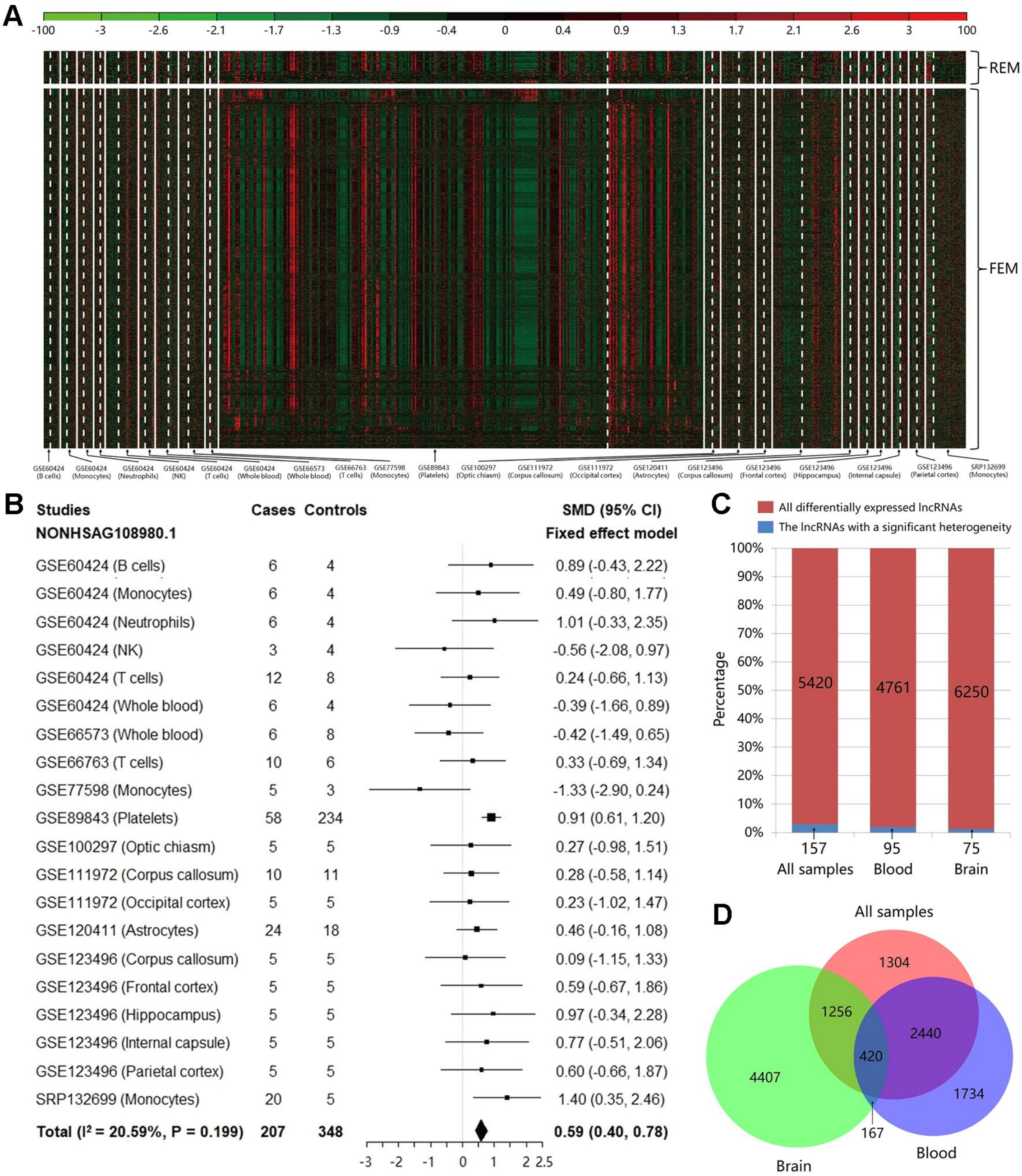


FIGURE 2 | The results of heterogeneity test and meta-analysis for all samples and subgroups. **(A)** The expression level of the significantly differentially expressed long noncoding RNAs (lncRNAs) in each study after meta-analysis. The random effect model was used for 157 lncRNAs with a significant heterogeneity, while the fixed effect model was used for 5,263 non-heterogeneous lncRNAs. The details can be clearly viewed by enlarging the electronic version. **(B)** The forest plot for the meta-analysis of the lncRNA NONHSAG108980.1 which is the most significant result associated with an increased risk of MS (SMD = 0.59, 95% CI = 0.40–0.78, P = 1.89×10^{-9}). **(C)** The bar plot showing the results of heterogeneity test in each group. For all samples, the proportion of lncRNAs with a significant heterogeneity is not high (about 2.90%), and this percentage is further decreased to about 1.99 and 1.20% in blood and brain, respectively. **(D)** The Venn diagram exhibiting the overlap among the significantly differentially expressed lncRNAs that are identified using brain tissues, blood tissues, and all samples.

most significant association with an increased risk of MS (SMD = 0.59, 95% CI = 0.40–0.78, P = 1.89×10^{-9}). Then, to investigate the heterogeneity of the lncRNA expression profile in various tissues, we split the samples into brain and blood tissue, and performed the heterogeneity test and meta-analysis for subgroups. We found that not only the proportion of lncRNAs with a significant heterogeneity was not high for the whole samples, but also this percentage is further reduced to about 1.99 and 1.20% in blood and brain, respectively (**Figure 2C**). Finally, we explored the difference of the differentially expressed lncRNAs identified in various tissues. We found that there was the higher specificity for these lncRNAs identified in brain compared with them identified in blood. Particularly, about 60.06% of the 5,420 differentially expressed lncRNAs can also be identified in the blood, while percentage is only 26.82% in brain (**Figure 2D**). Moreover, the total number of upregulated lncRNAs is far more than that of the downregulated ones in the blood (**Supplementary Table S2**) and the brain (**Supplementary Table S3**), which indicated that MS risk was related to lncRNA overexpression.

In addition, previous studies found that lncRNAs were modestly evolutionarily conserved in sequence (Guttman et al., 2009; Iyer et al., 2015). Therefore, we explored the conservation in sequence of these differentially expressed lncRNAs using conservation constrain search in NONCODE which contains the conservation information of lncRNAs in 13 common model organisms (i.e., human, chimp, gorilla, orangutan, rhesus, mouse, rat, cow, pig, opossum, platypus, chicken, and zebrafish). The results showed that only 0.11% of the differential lncRNAs were conserved in sequence among all these 13 organisms, while this percentage is increased to 28.5% in primates (human, chimp, gorilla, orangutan, and rhesus).

Tissue Specificity Analysis of the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids

Using expression data of NONCODE database, we performed the JS divergence metric and the cluster analysis to explore the tissue specificity of MS-related lncRNAs. The results of JS divergence metric showed that the MS-related lncRNA had high tissue specificity when used the brain, blood and whole samples (**Figure 3A**). For cluster analysis, relied on the same data, we further compared the expression patterns of these differentially expressed lncRNAs in various human tissues and cell lines. We found that the differentially expressed lncRNAs identified based on whole sample were highly specifically expressed in brain tissue (**Figure 3B**). Similarly, we observed a significant brain-specific expression for the differentially expressed lncRNAs identified based on brain sample (**Figure 3C**). Interestingly, although the differentially expressed lncRNAs were identified from blood sample, their expressions were still highly specific in brain tissue (**Figure 3D**). These results are consistent with the findings of the previous step and our recently published study (Han et al., 2019), which suggest that MS possesses the characteristics of the CNS disorder in lncRNA dysregulation.

Inferring the Functions of Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids by Weighted Gene Co-Expression Network Analysis

After abundance quantification together with meta-analysis, we identified 2,051 protein-coding genes significantly differentially expressed between MS patients and controls (**Supplementary Table S4**). Then, we combined the count values of 2,051 differentially expressed protein-coding genes and 5,420 MS-related lncRNAs to perform the WGCNA. By quality control, we removed three outlier samples whose minimum cluster size less than 5 and cutting height less than 4.0×10^6 (**Supplementary Figure S1**). The satisfactory soft threshold power β was set as 9 when the model fitting index R^2 equals 0.8 and the mean connectivity is close to 0 simultaneously (**Supplementary Figure S2**). Finally, we constructed a co-expression network which includes 1,938 protein-coding genes and 5,022 lncRNAs, and according to the interconnectedness of gene pairs, they were clustered into 15 modules in network (MEyellow, MEturquoise, MEblue, MEsalmon, MEred, MEpurple, MEpink, MEmagenta, MEgreen, MEmidnightblue, MEcyan, MEtan, MEgreenyellow, MEbrown, and MEblack) (**Figure 4A**). Moreover, to assess the significance of these modules for MS, we calculated two types of correlations as the index. The results of the average correlation of the genes in each module with the disease states showed that MEyellow is the most associated module with MS ($r = 0.33$, $P = 5\times10^{-15}$), and the following three are MEred ($r = 0.32$, $P = 2\times10^{-14}$), MEpink ($r = -0.28$, $P = 2\times10^{-11}$), and MEbrown ($r = 0.24$, $P = 9\times10^{-9}$). This was also applied to assess the association of each module with the platforms and the tissue types, respectively. Consistently, we found that the MEred ($r = 0.71$, $P = 2\times10^{-85}$), MEbrown ($r = 0.52$, $P = 1\times10^{-39}$), and MEyellow ($r = 0.38$, $P = 2\times10^{-20}$) were most significantly associated with the tissue types. While there is no module strongly associated with platforms (**Figure 4B**). These findings are generally consistent with the result of the correlation between the module membership and the gene significance for MS. For example, MEyellow and MEred are the top two module with the high average correlation of genes with disease states, and they also show a high correlation between module membership and gene significance ($cor = 0.43$, $P = 4.6\times10^{-15}$ and $cor = 0.50$, $P = 2.6\times10^{-19}$, respectively) (**Figures 4C**, **D**). On the contrary, MEcyan shows a very low level both for the two types of correlations ($r = -0.058$, $P = 0.2$ and $cor = 0.038$, $P = 0.8$) (**Figure 4E**).

In addition, we also perform a WGCNA with the satisfactory soft threshold power β = 9 using all the quantified genes. We found that these genes are clustered into 119 modules in the network, and about 82.2% differentially expressed genes are clustered into 16 of the 119 modules (including a gray one). We also found that these modules show low/modest association with MS (the correlation coefficients are < 0.19). These results reflect the similar distribution of the differentially expressed genes between using all and filtering genes in this WGCNA, and imply that the extra genes may mask the association of the differentially expressed genes with MS.

Pathway Analysis of the Multiple Sclerosis-Related Long Noncoding Ribonucleic Acids by Gene Set Enrichment Analysis

To explore the function of lncRNAs in MS, we performed GSEA in the four most significant modules for MS based on the two types of correlations, i.e., MEyellow ($r = 0.33$, $P = 5\times10^{-15}$ and $cor = 0.43$, $P = 4.6\times10^{-15}$), MEred ($r = 0.32$, $P = 2\times10^{-14}$ and $cor = 0.50$, $P = 2.6\times10^{-19}$), MEpink ($r = -0.28$, $P = 2\times10^{-11}$ and $cor = 0.63$, $P = 3.5\times10^{-14}$), and MEbrown (r

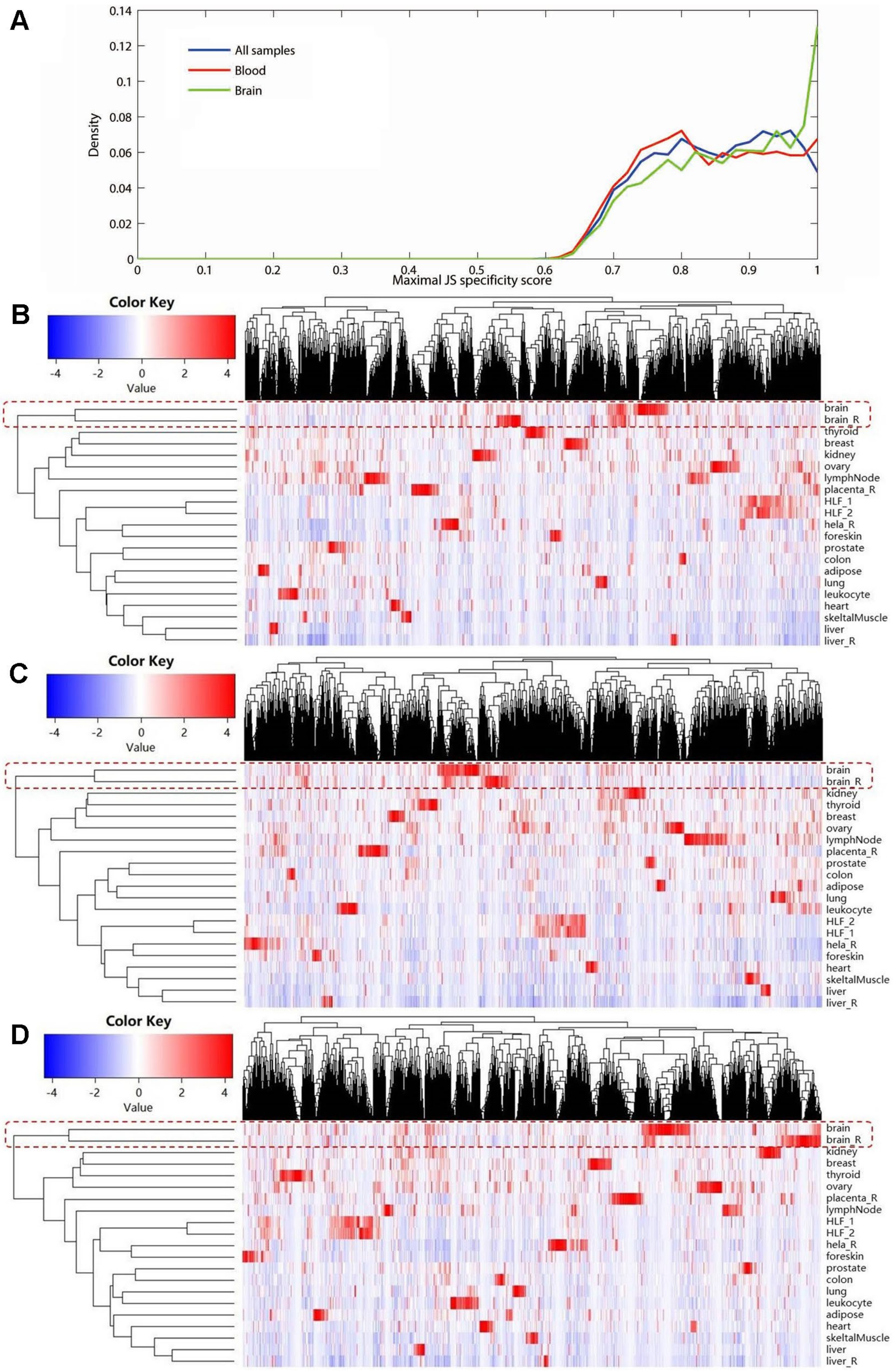


FIGURE 3 | Continued

FIGURE 3 | The tissue specificity of the multiple sclerosis-related long noncoding RNAs (lncRNAs) based on expression data from NONCODE database. **(A)** Tissue specific expression measured by Jensen-Shannon divergence. The distributions of the maximal tissue specificity scores showed the high tissue specificity of the differentially expressed lncRNAs identified using whole (blue), brain (green), and blood sample (red), respectively. The **(B)** to **(D)** showed the hierarchical clustering heatmap for expression of these lncRNAs in primary human tissues and cell lines. These differentially expressed lncRNAs identified using whole **(B)**, brain **(C)**, and blood sample **(D)** are all highly specifically expressed in brain tissue. The Manhattan distance was used to perform all of the three cluster analyses.

= 0.24, P = 9×10^{-9} and cor = 0.32, P = 4.7×10^{-9}). We found no significantly enriched pathway related to the MEred. Based on the result of LncRNA2Target, we identified that two differentially expressed lncRNAs in MEred could target the MS-related genes. Particularly, two target genes (CDH1 and CDH2) of the lncRNA NONHSAG081583.2 encoded cadherin protein which is the most abundant adhesion molecules participating in nerve conduction in synaptic junctions and the proinflammatory cytokines in MS can downregulate its expression (Minagar et al., 2003; Tian et al., 2009). The lncRNA NONHSAG000840.2 targets a MS-related gene NOTCH2, and reducing NOTCH2 in the proinflammatory monocytes can increase the frequency of the nonclassical monocytes and neutralizing antidrug antibody induction in IFN-β treated MS patients (Adriani et al., 2018). For MEbrown, the co-expressed protein-coding genes were mainly involved in leukocytes and interleukin-related immune response (**Figure 5A** and **Supplementary Table S5**), which was similar to the finding of our recent study (Han et al., 2019). Many genomic variants in the human leukocyte antigen complexes and interleukin receptor were identified significantly associated with susceptibility of MS (Rubio et al., 2002; Teutsch et al., 2003; Lundmark et al., 2007; Hollenbach and Oksenberg, 2015; Tang et al., 2019a). The protein-coding genes in MEpink are mainly associated with intercellular junction

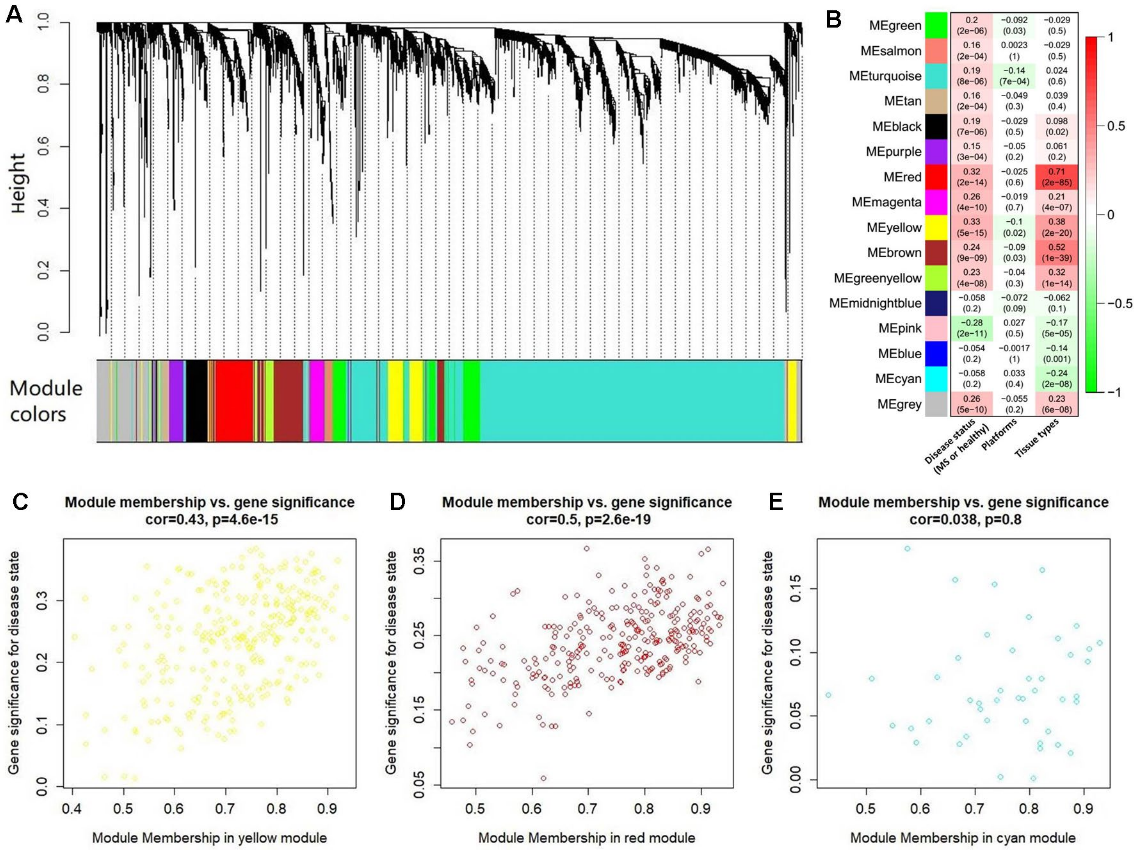


FIGURE 4 | The co-expression network analysis of the differentially expressed long noncoding RNAs (lncRNAs) and protein-coding genes. **(A)** The clustering dendrogram of these co-expressed lncRNAs and protein-coding genes. There are 15 clustered modules in the hierarchical clustering dendrogram which is constructed by a dynamic cut-tree algorithm. These clustered modules are marked as 15 different colors, respectively, i.e., yellow, turquoise, tan, salmon, red, purple, pink, midnight blue, magenta, green yellow, green, cyan, brown, blue, and black. **(B)** The heatmap for the association of each module with the disease states, platforms, and tissue types. Each cell represents a module, and contains the correlation r and corresponding P value (in brackets). Panels **(C)** to **(E)** show the results of correlation between the module membership and the gene significance in MEyellow, MEred, and MEcyan, respectively. The results of other modules were described in **Supplementary Figure S3**.

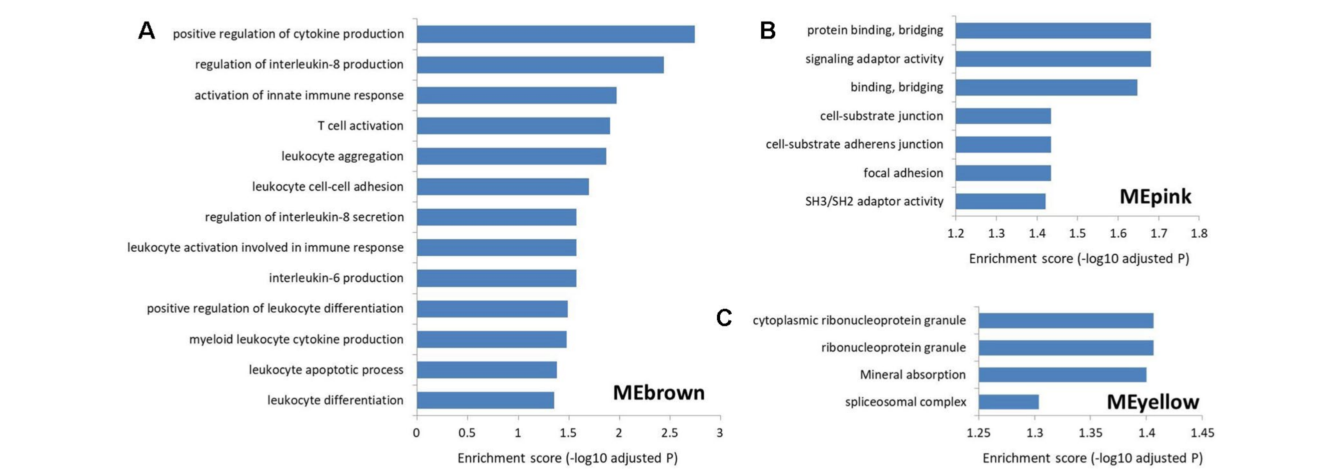


FIGURE 5 | The Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway enrichment in the three most significant modules for multiple sclerosis. **(A)** The enrichment for MEbrown. The protein-coding genes co-expressed with the MS-related lncRNAs in this module are mainly involved in leukocyte and interleukin-related immune response. **(B)** The enrichment for MEpink. The co-expressed protein-coding genes in this module are mainly associated with the intercellular junction and signaling transmission. **(C)** The enrichment for MEyellow. The co-expressed protein-coding genes in this module are mainly related to the ribonucleoprotein.

and signaling transmission (**Figure 5B**). Previous studies found that the defect of axon-glial signaling transmission caused by the oligodendrocyte gap junction loss and disconnection contributes to MS pathogenesis (Brand-Schieber et al., 2005; Markoullis et al., 2012; Markoullis et al., 2014). The results of LncRNA2Target showed that lncRNA NONHSAG049754.2 in MEyellow targets the MS-related gene TNFRSF10A. This gene encodes the receptor of tumor necrosis factor (TNF) cytokines which plays a important role in inflammation regulations and is related to susceptibility of developing MS (De-la-Torre et al., 2019). The protein-coding genes in the MEyellow are related to ribonucleoprotein (**Figure 5C**). Ribonucleoprotein is a kind of ribonucleic acid-binding protein which participates in the mRNA splicing (Guthrie, 1991). Previous study showed that as an important autoantigen in the neuroimmune disease, the ribonucleoprotein significantly more often interact with the autoantibodies in MS cerebrospinal fluids compared with controls (Sueoka et al., 2004; Yukitake et al., 2008). The following studies further identified a ribonucleoprotein-related lncRNA, TNF-α, and heterogeneous nuclear ribonucleoprotein L, which was significantly upregulated and produced transcriptional activating complexes to promote TNF-α expression by cooperating with ribonucleoprotein in the circulating blood cells of MS (Li et al., 2014; Eftekharian et al., 2017). Given that MEyellow is the most significant module for MS, we inferred that one of the key mechanisms of lncRNAs in MS is associated with the regulation of ribonucleoprotein and TNF cytokines receptor.

CONCLUSIONS

In this study, we comprehensively collected MS-related RNA-seq data from a variety of studies, and integrated these data by an expression-based meta-analysis to assess the affection of lncRNAs on the MS pathogenesis on genome scale. We identified a total of 5,420 lncRNAs significantly differentially expressed between MS patients and controls. Then, the subgroup analysis found a small heterogeneity of the lncRNA expression profile between the brain and blood tissues. Further, the specificity analysis of multiple tissues showed that the differentially expressed lncRNAs (including identified using brain, blood, and whole sample) are highly specifically expressed in brain tissue. Finally, the result of GSEA and WGCNA demonstrated that the potential important function of lncRNAs in MS may be involved in the regulation of ribonucleoprotein and TNF cytokines receptor. All in all, we performed a strategy to resolve the inconsistent MS-related lncRNA findings in previous studies, and explore the functions of these lncRNAs in MS. The findings of this study will be benefit to improve the understanding of the pathogenesis of MS.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/gds, https://www.ebi.ac.uk/arrayexpress, https://ddbj.nig.ac.jp/DRASearch.

AUTHOR CONTRIBUTIONS

ZH and FZ designed the research. ZH, FZ, JH, and WX collected the data. ZH performed the research, analyzed data, and wrote the paper. FZ reviewed and modified the manuscript. All authors discussed the results, and contributed to the final manuscript. All authors read and approved the final manuscript.

FUNDING

This work was funded by National Key R&D Program of China (2018YFC0910500), National Natural Science Foundation of China (81872798), Fundamental Research Fund for Central Universities (10611CDJXZ238826, 2018QNA7023, 2018CDQYSG0007 & CDJZR14468801), and Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztzx120003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01136/full#supplementary-material

REFERENCES

Adriani, M., Nytrova, P., Mbogning, C., Hassler, S., Medek, K., Jensen, P. E. H., et al. (2018). Monocyte NOTCH2 expression predicts IFN-beta immunogenicity in multiple sclerosis patients. *JCI Insight* 3. doi: 10.1172/jci.insight.99274

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., et al. (2017). Ensembl 2017. *Nucleic Acids Res.* 45, D635–D642. doi: 10.1093/nar/gkw1104

Angerer, I. C., Hecker, M., Koczan, D., Roch, L., Friess, J., Ruge, A., et al. (2018). Transcriptome profiling of peripheral blood immune cell populations in multiple sclerosis patients before and during treatment with a sphingosine-1-phosphate receptor modulator. *CNS Neurosci. Ther.* 24, 193–201. doi: 10.1111/cns.12793

Athar, A., Fullgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* 47, D711–D715. doi: 10.1093/nar/gky964

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Brand-Schieber, E., Werner, P., Iacobas, D. A., Iacobas, S., Beelitz, M., Lowery, S. L., et al. (2005). Connexin43, the major gap junction protein of astrocytes, is down-regulated in inflamed white matter in an animal model of multiple sclerosis. *J. Neurosci. Res.* 80, 798–808. doi: 10.1002/jnr.20474

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Browne, P., Chandraratna, D., Angood, C., Tremlett, H., Baker, C., Taylor, B. V., et al. (2014). Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity. *Neurology* 83, 1022–1024. doi: 10.1212/WNL.0000000000000768

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinf.* 11, 94. doi: 10.1186/1471-2105-11-94

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi: 10.1101/gad.17446611

Cardamone, G., Paraboschi, E. M., Solda, G., Cantoni, C., Supino, D., Piccio, L., et al. (2019). Not only cancer: the long non-coding RNA MALAT1 affects the repertoire of alternatively spliced transcripts and circular RNAs in multiple sclerosis. *Hum. Mol. Genet.* 28, 1414–1428. doi: 10.1093/hmg/ddy438

Chen, Y. G., Satpathy, A. T., and Chang, H. Y. (2017). Gene regulation in the immune system by long noncoding RNAs. *Nat. Immunol.* 18, 962–972. doi: 10.1038/ni.3771

Cheng, L., and Hu, Y. (2018). Human disease system biology. *Curr. Gene Ther*. doi: 10.2174/1566523218666181101143116

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). 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

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., et al. (2011). Modernizing reference genome assemblies. *PloS Biol.* 9, e1001091. doi: 10.1371/journal.pbio.1001091

De-la-Torre, A., Silva-Aldana, C. T., Munoz-Ortiz, J., Pineros-Hernandez, L. B., Otero, O., Dominguez, A., et al. (2019). Uveitis and multiple sclerosis: potential common causal mutations. *Mol. Neurobiol*. doi: 10.1007/s12035-019-1630-2

Dong, X., Chen, K., Cuevas-Diaz Duran, R., You, Y., Sloan, S. A., Zhang, Y., et al. (2015). Comprehensive identification of long non-coding RNAs in purified cell types from the brain reveals functional LncRNA in OPC fate determination. *PloS Genet.* 11, e1005669. doi: 10.1371/journal.pgen.1005669

Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., et al. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913. doi: 10.1038/nsmb.2591

Eftekharian, M. M., Ghafouri-Fard, S., Soudyab, M., Omrani, M. D., Rahimi, M., Sayad, A., et al. (2017). Expression analysis of long non-coding RNAs in the blood of multiple sclerosis patients. *J. Mol. Neurosci.* 63, 333–341. doi: 10.1007/s12031-017-0982-1

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46, D308–D314. doi: 10.1093/nar/gkx1107

Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606

Frohman, E. M., Racke, M. K., and Raine, C. S. (2006). Multiple sclerosis–the plaque and its pathogenesis. *N Engl. J. Med.* 354, 942–955. doi: 10.1056/NEJMra052130

Gellert, P., Ponomareva, Y., Braun, T., and Uchida, S. (2013). Noncoder: a web interface for exon array-based detection of long non-coding RNAs. *Nucleic Acids Res.* 41, e20. doi: 10.1093/nar/gks877

Gharesouran, J., Taheri, M., Sayad, A., Ghafouri-Fard, S., Mazdeh, M., and Omrani, M. D. (2019). A novel regulatory function of long non-coding RNAs at different levels of gene expression in multiple sclerosis. *J. Mol. Neurosci.* 67, 434–440. doi: 10.1007/s12031-018-1248-2

Gomez, J. A., Wapinski, O. L., Yang, Y. W., Bureau, J. F., Gopinath, S., Monack, D. M., et al. (2013). The NeST long ncRNA controls microbial susceptibility and epigenetic activation of the interferon-gamma locus. *Cell* 152, 743–754. doi: 10.1016/j.cell.2013.01.015

Guthrie, C. (1991). Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. *Science* 253, 157–163. doi: 10.1126/science.1853200

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223–227. doi: 10.1038/nature07672

Han, Z., Jiang, Q., Zhang, T., Wu, X., Ma, R., Wang, J., et al. (2015). Analyzing large-scale samples confirms the association between the rs1051730 polymorphism and lung cancer susceptibility. *Sci. Rep.* 5, 15642. doi: 10.1038/srep15642

Han, Z., Qu, J., Zhao, J., and Zou, X. (2018a). Analyzing 74,248 samples confirms the association between CLU rs11136000 polymorphism and Alzheimer's disease in caucasian but not Chinese population. *Sci. Rep.* 8, 11062. doi: 10.1038/s41598-018-29450-2

Han, Z., Qu, J., Zhao, J., and Zou, X. (2018b). Genetic variant rs755622 regulates expression of the multiple sclerosis severity modifier D-dopachrome tautomerase in a sex-specific Way. *BioMed. Res. Int.* 2018, 8285653. doi: 10.1155/2018/8285653

Han, Z., Xue, W., Tao, L., Lou, Y., Qiu, Y., and Zhu, F. (2019). Genome-wide identification and analysis of the eQTL lncRNAs in multiple sclerosis based on RNA-seq data. *Brief Bioinform*. doi: 10.1093/bib/bbz036

He, D., Wang, J., Lu, Y., Deng, Y., Zhao, C., Xu, L., et al. (2017). lncRNA functional networks in oligodendrocytes reveal stage-specific myelination control by an lncOL1/Suz12 complex in the CNS. *Neuron* 93, 362–378. doi: 10.1016/j.neuron.2016.11.044

Hollenbach, J. A., and Oksenberg, J. R. (2015). The immunogenetics of multiple sclerosis: a comprehensive review. *J. Autoimmun.* 64, 13–25. doi: 10.1016/j.jaut.2015.06.010

Ilott, N. E., and Ponting, C. P. (2013). Predicting long non-coding RNAs using RNA sequencing. *Methods* 63, 50–59. doi: 10.1016/j.ymeth.2013.03.019

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208. doi: 10.1038/ng.3192

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, D353–D361. doi: 10.1093/nar/gkw1092

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Kim, H., Kim, J. H., Kim, S. Y., Jo, D., Park, H. J., Kim, J., et al. (2015). Meta-analysis of large-scale toxicogenomic data finds neuronal regeneration related protein and cathepsin D to be novel biomarkers of drug-induced toxicity. *PloS One* 10, e0136698. doi: 10.1371/journal.pone.0136698

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Leinonen, R., Sugawara, H., and Shumway, M.International Nucleotide Sequence Database, C. (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., et al. (2017). NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.* 45, W162–W170. doi: 10.1093/nar/gkx449

Li, Y., Song, D., Jiang, Y., Wang, J., Feng, R., Zhang, L., et al. (2016). CR1 rs3818361 Polymorphism contributes to Alzheimer's disease susceptibility in Chinese population. *Mol. Neurobiol.* 53, 4054–4059. doi: 10.1007/s12035-015-9343-7

Li, Y. H., Li, X. X., Hong, J. J., Wang, Y. X., Fu, J. B., Yang, H., et al. (2019). Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform*. doi: 10.1093/bib/bby130

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res.* 46, D1121–D1127. doi: 10.1093/nar/gkx1076

Li, Z., Chao, T. C., Chang, K. Y., Lin, N., Patil, V. S., Shimizu, C., et al. (2014). The long noncoding RNA THRIL regulates TNFalpha expression through its interaction with hnRNPL. *Proc. Natl. Acad. Sci. U.S.A.* 111, 1002–1007. doi: 10.1073/pnas.1313768111

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM rs3851179 variant confers susceptibility to Alzheimer's disease in Chinese population. *Mol. Neurobiol.* 54, 3131–3136. doi: 10.1007/s12035-016-9886-2

Lundmark, F., Duvefelt, K., Iacobaeus, E., Kockum, I., Wallstrom, E., Khademi, M., et al. (2007). Variation in interleukin 7 receptor alpha chain (IL7R) influences risk of multiple sclerosis. *Nat. Genet.* 39, 1108–1113. doi: 10.1038/ng2106

Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., et al. (2019). MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics* 35, 1597–1599. doi: 10.1093/bioinformatics/bty825

Markoullis, K., Sargiannidou, I., Schiza, N., Hadjisavvas, A., Roncaroli, F., Reynolds, R., et al. (2012). Gap junction pathology in multiple sclerosis lesions and normal-appearing white matter. *Acta Neuropathol.* 123, 873–886. doi: 10.1007/s00401-012-0978-4

Markoullis, K., Sargiannidou, I., Schiza, N., Roncaroli, F., Reynolds, R., and Kleopa, K. A. (2014). Oligodendrocyte gap junction loss and disconnection from reactive astrocytes in multiple sclerosis gray matter. *J. Neuropathol. Exp. Neurol.* 73, 865–879. doi: 10.1097/NEN.0000000000000106

Masoumi, F., Ghorbani, S., Talebi, F., Branton, W. G., Rajaei, S., Power, C., et al. (2019). Malat1 long noncoding RNA regulates inflammation and leukocyte differentiation in experimental autoimmune encephalomyelitis. *J. Neuroimmunol.* 328, 50–59. doi: 10.1016/j.jneuroim.2018.11.013

Minagar, A., Ostanin, D., Long, A. C., Jennings, M., Kelley, R. E., Sasaki, M., et al. (2003). Serum from patients with multiple sclerosis downregulates occludin and VE-cadherin expression in cultured endothelial cells. *Mult. Scler* 9, 235–238. doi: 10.1191/1352458503ms916oa

Ng, S. Y., Lin, L., Soh, B. S., and Stanton, L. W. (2013). Long noncoding RNAs in development and disease of the central nervous system. *Trends Genet.* 29, 461–468. doi: 10.1016/j.tig.2013.03.002

Ogasawara, O., Mashima, J., Kodama, Y., Kaminuma, E., Nakamura, Y., Okubo, K., et al. (2013). DDBJ new system and service refactoring. *Nucleic Acids Res.* 41, D25–D29. doi: 10.1093/nar/gks1152

Pahlevan Kakhki, M., Nikravesh, A., Shirvani Farsani, Z., Sahraian, M. A., and Behmanesh, M. (2018). HOTAIR but not ANRIL long non-coding RNA contributes to the pathogenesis of multiple sclerosis. *Immunology* 153, 479–487. doi: 10.1111/imm.12850

Rezazadeh, M., Gharesouran, J., Moradi, M., Noroozi, R., Omrani, M. D., Taheri, M., et al. (2018). Association study of ANRIL genetic variants and multiple sclerosis. *J. Mol. Neurosci.* 65, 54–59. doi: 10.1007/s12031-018-1069-3

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rubio, J. P., Bahlo, M., Butzkueven, H., van Der Mei, I. A., Sale, M. M., Dickinson, J. L., et al. (2002). Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. *Am. J. Hum. Genet.* 70, 1125–1137. doi: 10.1086/339932

Santoro, M., Nociti, V., Lucchini, M., De Fino, C., Losavio, F. A., and Mirabella, M. (2016). Expression profile of long non-coding RNAs in serum of patients with multiple sclerosis. *J. Mol. Neurosci.* 59, 18–23. doi: 10.1007/s12031-016-0741-8

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., et al. (2019). Database resources of the national center for biotechnology Information. *Nucleic Acids Res.* 47, D23–D28. doi: 10.1093/nar/gky1069

Selmaj, I., Cichalewska, M., Namiecinska, M., Galazka, G., Horzelski, W., Selmaj, K. W., et al. (2017). Global exosome transcriptome profiling reveals biomarkers for multiple sclerosis. *Ann. Neurol.* 81, 703–717. doi: 10.1002/ana.24931

Soneson, C., Love, M. I., and Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* 4, 1521. doi: 10.12688/f1000research.7563.2

Sospedra, M., and Martin, R. (2005). Immunology of multiple sclerosis. *Annu. Rev. Immunol.* 23, 683–747. doi: 10.1146/annurev.immunol.23.021704.115707

Sueoka, E., Yukitake, M., Iwanaga, K., Sueoka, N., Aihara, T., and Kuroda, Y. (2004). Autoantibodies against heterogeneous nuclear ribonucleoprotein B1 in CSF of MS patients. *Ann. Neurol.* 56, 778–786. doi: 10.1002/ana.20276

Szajewska, H., and Kolodziej, M. (2015). Systematic review with meta-analysis: *Saccharomyces boulardii* in the prevention of antibiotic-associated diarrhoea. *Aliment Pharmacol. Ther.* 42, 793–801. doi: 10.1111/apt.13344

Tang, J., Fu, J., Wang, Y., Li, B., Li, Y., Yang, Q., et al. (2019a). ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform*. doi: 10.1093/bib/bby127

Tang, J., Fu, J., Wang, Y., Luo, Y., Yang, Q., Li, B., et al. (2019b). Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains. *Mol. Cell Proteomics* 18, 1683–1699. doi: 10.1074/mcp.RA118.001169

Teutsch, S. M., Booth, D. R., Bennetts, B. H., Heard, R. N., and Stewart, G. J. (2003). Identification of 11 novel and common single nucleotide polymorphisms in the interleukin-7 receptor-alpha gene and their associations with multiple sclerosis. *Eur. J. Hum. Genet.* 11, 509–515. doi: 10.1038/sj.ejhg.5200994

The Gene Ontology, C. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Tian, L., Rauvala, H., and Gahmberg, C. G. (2009). Neuronal regulation of immune responses in the central nervous system. *Trends Immunol.* 30, 91–99. doi: 10.1016/j.it.2008.11.002

UniProt, C. (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212. doi: 10.1093/nar/gku989

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Xue, W., Yang, F., Wang, P., Zheng, G., Chen, Y., Yao, X., et al. (2018). What contributes to serotonin-norepinephrine reuptake iinhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation. *ACS Chem. Neurosci.* 9, 1128–1140. doi: 10.1021/acschemneuro.7b00490

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2019). Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data. *Brief Bioinform*. doi: 10.1093/bib/bbz049

Yin, J., Sun, W., Li, F., Hong, J., Li, X., Zhou, Y., et al. (2019). VARIDT 1.0: variability of drug transporter database. *Nucleic Acids Res*. doi: 10.1093/nar/gkz779

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

Yukitake, M., Sueoka, E., Sueoka-Aragane, N., Sato, A., Ohashi, H., Yakushiji, Y., et al. (2008). Significantly increased antibody response to heterogeneous nuclear ribonucleoproteins in cerebrospinal fluid of multiple sclerosis patients but not in patients with human T-lymphotropic virus type I-associated myelopathy/tropical spastic paraparesis. *J. Neurovirol.* 14, 130–135. doi: 10.1080/13550280701883840

Zhang, F., Gao, C., Ma, X. F., Peng, X. L., Zhang, R. X., Kong, D. X., et al. (2016). Expression profile of long noncoding RNAs in peripheral blood mononuclear cells from multiple sclerosis patients. *CNS Neurosci. Ther.* 22, 298–305. doi: 10.1111/cns.12498

Zhang, M., Chang, Y.-C., Shankara, S., Jacobs, A., Godin, J., Klinger, K., et al. (2019). *Characterization of the Peripheral Blood Transcriptome in Alemtuzumab-Treated Relapsing-Remitting Multiple Sclerosis Patients From the CARE-MS I and II Studies (P4. 2-041)*. AAN Enterprises.

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.

ORIGINAL RESEARCH
published: 22 November 2019
doi: 10.3389/fgene.2019.01184

CircSLNN: Identifying RBP-Binding Sites on circRNAs *via* Sequence Labeling Neural Networks

Yuqi Ju[1†], *Liangliang Yuan*[1†], *Yang Yang*[1,2,3*] *and Hai Zhao*[1,2,3]

[1] Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, [2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China, [3] Brain Science and Technology Research Center, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Xiaofeng Song,
Nanjing University of Aeronautics and Astronautics, China
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Yang Yang
yangyang@cs.sjtu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 August 2019*
Accepted: *25 October 2019*
Published: *22 November 2019*

Citation:
Ju Y, Yuan L, Yang Y and Zhao H (2019) CircSLNN: Identifying RBP-Binding Sites on circRNAs via Sequence Labeling Neural Networks. Front. Genet. 10:1184. doi: 10.3389/fgene.2019.01184

The interactions between RNAs and RNA binding proteins (RBPs) are crucial for understanding post-transcriptional regulation mechanisms. A lot of computational tools have been developed to automatically predict the binding relationship between RNAs and RBPs. However, most of the methods can only predict the presence or absence of binding sites for a sequence fragment, without providing specific information on the position or length of the binding sites. Besides, the existing tools focus on the interaction between RBPs and linear RNAs, while the binding sites on circular RNAs (circRNAs) have been rarely studied. In this study, we model the prediction of binding sites on RNAs as a sequence labeling problem, and propose a new model called circSLNN to identify the specific location of RBP-binding sites on circRNAs. CircSLNN is driven by pretrained RNA embedding vectors and a composite labeling model. On our constructed circRNA datasets, our model has an average F_1 score of 0.790. We assess the performance on full-length RNA sequences, the proposed model outperforms previous classification-based models by a large margin.

Keywords: RNA–protein binding sites, sequence labeling, convolutional neural network, bidirectional LSTM neural network, deep learning

INTRODUCTION

Benefitting from the rapid development of high-throughput experimental technologies, transcriptome, proteome, epigenome and other omics data have accumulated in an unprecedented speed. The multi-omics data have enabled large-scale studies on gene regulation at different levels. Especially, the interactions between RNAs and RNA binding proteins (RBPs) are crucial for understanding post-transcriptional regulation mechanisms (Filipowicz et al., 2008). The RNA–RBP-interactions play important roles in protein synthesis, gene fusion, alternative mRNA processing, etc. (Bolognani and Perrone-Bizzozero, 2008). The aberrant expression of RBPs and disruption of RNA–RBP-interactions are closely related to various diseases of human beings (Khalil and Rinn, 2011). In the early stage of RNA–RBP-interaction studies, the recognition of binding sites mainly relies on the analysis of RNA–protein complexes *via* biophysical methods. As the experimental process is costly and laborious, it is increasingly important to develop automatic tools to predict binding sites.

As for protein–protein-interactions, both structures and amino acid sequences are commonly used for identifying binding sites, including POCKET (Liu and Hu, 2011), Fpocket (Le Guilloux et al., 2009) LIGSITE (Hendlich et al., 1997), etc. The structural feature-based prediction methods exploit protein 3D structures and appropriate geometries to locate potential binding regions. Most structure-based methods assume that proteins bound to the same ligand have similar overall structure and biochemistry characteristics, while some researchers found that proteins having the same binding site may have diverse sequences or structures (Muppirala et al., 2011). Sequence-based methods usually utilize amino acid composition, function domain, secondary structure and solvent accessibility information (Shen et al., 2007).

Due to the lack of solved structures for RNA-protein complexes, most of the existing studies have turned to sequence information and machine learning methods for predicting RBP-binding sites on RNAs, like support vector machines (SVMs) (Kumar et al., 2008) and random forest (RF) (Liu et al., 2010). Moreover, deep learning models have emerged in this field (Alipanahi et al., 2015; Pan and Shen, 2017). Deep learning is a data-driven approach that allows automatic learning of the advanced features from data without the need for domain knowledge, by stacking multiple layers of neural networks (LeCun et al., 2015). Compared to traditional machine learning models, it does not require feature engineering and can achieve better performance. A few deep learning methods, including convolutional neural network (CNN) and recurrent neural network (RNN), have been developed to predict RBP-binding sites (Pan and Shen, 2017; Pan et al., 2018).

Although researchers have made some progress in predicting RNA–protein binding sites, current mainstream prediction methods have some limitations.

First, most prediction methods simplify the prediction task as a binary classification problem, i.e. they assign a positive/negative label to a segment of RNA, where the positive label denotes the presence of a binding site. Actually, binding sites on RNAs are sequence fragments that range from tens to hundreds of nucleotides in length. Thus, the prediction based on fixed-length fragments may be inaccurate, as it only yields approximate locations of binding sites and could not specify the length that the sites span.

Second, most of the existing methods predict the interaction between linear RNAs and RBPs, while circular RNAs (circRNAs) have been rarely studied. CircRNAs play an important role in gene regulation, and they also play crucial roles in the development of many complex diseases (Fan et al., 2018). Thanks to the advances of new sequencing technology, circRNAs have been identified on the whole genome scale (Song et al., 2016). Moreover, the interplay between circRNAs and proteins or microRNAs has attracted more and more research interests from biomedical field, resulting in large-scale data of circRNA–RBP interactions using high-throughput experiments, like CLIP-Seq (Dudekula et al., 2016). Thus, the models for predicting binding sites on circRNAs are in great demand.

In this study, we propose a sequence labeling neural network model to predict circRNA–protein binding sites, called circSLNN, which is composed of a long-short-term memory (LSTM) network, a convolutional neural network (CNN) and a conditional random field (CRF) model. Instead of performing a binary classification on the whole fragment, it assigns a label (bound or unbound) to each position on the fragment. Compared with traditional classifiers, it can not only predict whether the input segment is bound to a given RBP, but also predict the specific location of binding sites on the segment. Besides, in order to fully utilize the sequence information of circRNAs, we propose to use RNA embeddings learned *via* a similar word embedding algorithm for processing natural languages, where the corpus is extracted from the whole human genome. To the best of our knowledge, this is the first predictor for RNA–protein binding sites using a sequence labeling scheme. The contributions of this study are listed in the following.

1. We construct the sequence labeling network of LSTM-CNN-CRF for predicting RBP-binding sites on RNA sequences. Compared to previous methods, it has the advantage in identifying location and length of binding sites.
2. We apply RNA embeddings to the prediction of RNA–RBP interaction, and demonstrate the effectiveness of continuous dense feature vectors trained by word embedding and whole-genome corpus.
3. We propose a predictor, circSLNN, trained on circRNA binding sites, which may help researchers reveal the interaction mechanisms of circRNAs and proteins.

RELATED WORK

Prediction Based on Traditional Machine Learning Methods

The prediction of molecular interactions has been a hot topic in bioinformatics over the past decades. Especially, the protein–protein-interactions (PPIs) have been well-studied due to the abundant information that can be utilized in the prediction, e.g. amino acid sequences, function domains, gene ontology annotation (Ashburner et al., 2000). The machine learning-based predictors usually consist of two parts, i.e. the feature extraction and classification. Similar to PPI, the prediction of RNA–RBP-interaction is a typical machine learning problem. However, due to the lack of functional annotation of RNAs, the feature extraction mainly relies on RNA sequences or secondary structures. For some types of RNAs, like circRNAs which have constrained structures, i.e. covalently closed continuous loops, the effective feature extraction from sequences are more important.

Traditional feature representation of RNA sequences include *k*-tuple composition, pseudo *k*-tuple composition (PseKNC) (Chen et al., 2013), etc. The features are discrete vectors, working with shallow learning models. For instance, Muppirala et al. (2011) used the SVMs and random forest methods to predict the RNA–RBP-interactions. As the rise of deep learning, sequence encoding schemes and deep neural networks have been emerging and achieved better prediction performance.

Prediction Based on Deep Neural Networks

DeepBind (Alipanahi et al., 2015) is a pioneer work in developing deep learning models for RNA–RBP-interactions. The model is based on a convolutional neural network, which not only improves prediction accuracy but also reveals new sequence patterns at the binding area. Later, Pan et al. released a series of computational tools, including iDeep (Pan and Shen, 2017), iDeepS (Pan et al., 2018) and iDeepE (Pan and Shen, 2018), which have different feature representation and model architecture. iDeep utilizes five different information sources, i.e. secondary structure information, motif information for describing the conserved region of sequences, CLIP co-binding, region type, and sequence information, to extract high-level abstraction features *via* deep learning models. Especially, the sequence information is processed by a CNN (Krizhevsky et al., 2012), while other four data sources are processed by deep belief networks (Zou and Conzen, 2004). Compared with iDeep, iDeepS reduces the types of data sources and only retains sequence information and secondary structure information. The authors added bi-directional long short-term memory (BiLSTM) (Schuster and Paliwal, 1997) to integrate the data, which better reserves contextual information based on relative position relationship of nucleotides.

Generally, the performance of deep learning-based methods depends on informative feature representation and powerful model architecture. In this study, we explore both the two parts to improve prediction accuracy.

MATERIALS AND METHODS

Data Source

To construct a predictor for circRNA–RBP-interactions, we collect a standard dataset of RBP-binding sites on circular RNAs from the circRNA Interactome database (Dudekula et al., 2016), which contains sequence information for more than 100,000 human circRNAs, as well as specific locations of binding sites for different RBPs. Each binding site is represented as an interval from the start index to the end index on the circRNAs. We extend 50-nt upstream and downstream respectively by taking the midpoint of each interval as the center. In this way, 101-nt fragments can be obtained as positive samples. Then we randomly extract 101-nt segments from the remaining fragments as negative samples. In order to avoid the issue caused by repeated sequences, we remove redundant sequences using CD-HIT (Li and Godzik, 2006). The positive-to-negative ratio is 1:1, and the training-to-test ratio is 5:1.

Then we generate standard labels for all samples. For positive samples, we label all the symbols within the binding sites as “I” and all the other locations as “O”, meanwhile we mark all symbols as “O” for negative samples. Here we use the IO tag scheme, where “I” is short for inside (a binding site) and “O” is short for outside, i.e. not a binding site. As it is known that, the BIO format (short for inside, outside, beginning) is a common tagging format in natural language. As there are a lot of adjacent labeling objects in text, it is hard to distinguish between different labeling objects using only the IO scheme. By contrast, in the sequence labeling problem of binding sites, the distribution of binding sites is extremely sparse, and usually binding segments are far from each other. Thus, we use the IO labeling scheme to reduce the types of labels and make the training model easier to converge.

Data Encoding

As mentioned in the *Related Work* section, feature representation can have a substantial impact on the performance for both shallow learning and deep learning models. To work with deep models, RNA sequences need to be encoded into numerical vectors, like one-hot vectors. In recent years, more and more studies on biological sequence analysis have adopted word embedding-based encoding schemes to replace one-hot encoding (Harris and Harris, 2010), as embedding vectors are continuous and high-dimensional, which may capture more context and semantic information in sequences. In our previous studies, we propose the RNA2Vec method to get RNA embeddings (Xiao et al., 2018). We regard 10-mer segments as words and train the word embeddings using Glove (Pennington et al., 2014).

Model Architecture

In this study, we design a sequence labeling model based on deep neural networks to predict RBP-binding sites on RNAs. We first feed the embedding vectors to a convolutional neural network (Krizhevsky et al., 2012) to extract local features, and then learn the long-distance dependency information among bases through a BiLSTM layer. Finally, the label identification of the entire RNA sequence is completed by the CRF layer (Lafferty et al., 2001). The network structure is shown in **Figure 1**.

CNN Layer

Convolutional neural network (CNN) (Krizhevsky et al., 2012) is a widely used deep learning architecture. CNN generates feature maps at different abstract levels by stacking convolutional layers. In circSLNN, the CNN serves as a feature extractor from the initial input vectors. As sequence labeling models predict a label for each symbol in the sequence, whereas the embedding vectors are trained for 10-mers, we adopt CNN to extract high-level features for each nucleotide in RNA sequences based on the embedding vectors of its surrounding 10-mers, i.e. a window centered by the nucleotide.

Specifically, for each individual nucleotide (except for the first 9 nucleotides), there are 10 fragments of length 10 containing it. Based on the vectors of the 10 fragments, we perform feature extraction *via* a one-dimensional CNN. Suppose the dimensionality of embedding vectors is m, then each nucleotide can be represented as a matrix of size 10×m, which is fed to the CNN. Before using CNN, we need to expand the 101-nt fragments to 110-nt (101 + 10 − 1), which is passed through a sliding window of size 10. Here we pad the matrix by zero vectors.

Let h_j be the size of the jth convolutional kernel, X_i be the matrix of the sliding window at the ith time step, which consists of the ith to the $(i + h_j - 1)$th columns of the original input. Thus, the features learned by the convolutional layer can be expressed in Eq. 1,

$$\begin{gathered} c_{ij} = f(w_j * X_{i:i+hj-1} + b_j) \\ i \in \{1,2,\ldots,N - h_j + 1\}, \mathrm{j} \in \{1,2,\ldots n\} \end{gathered} \tag{1}$$

where n is the number of filters, f(.) is the activation function, and w_j and b_j are the weight matrix and the offset, respectively.

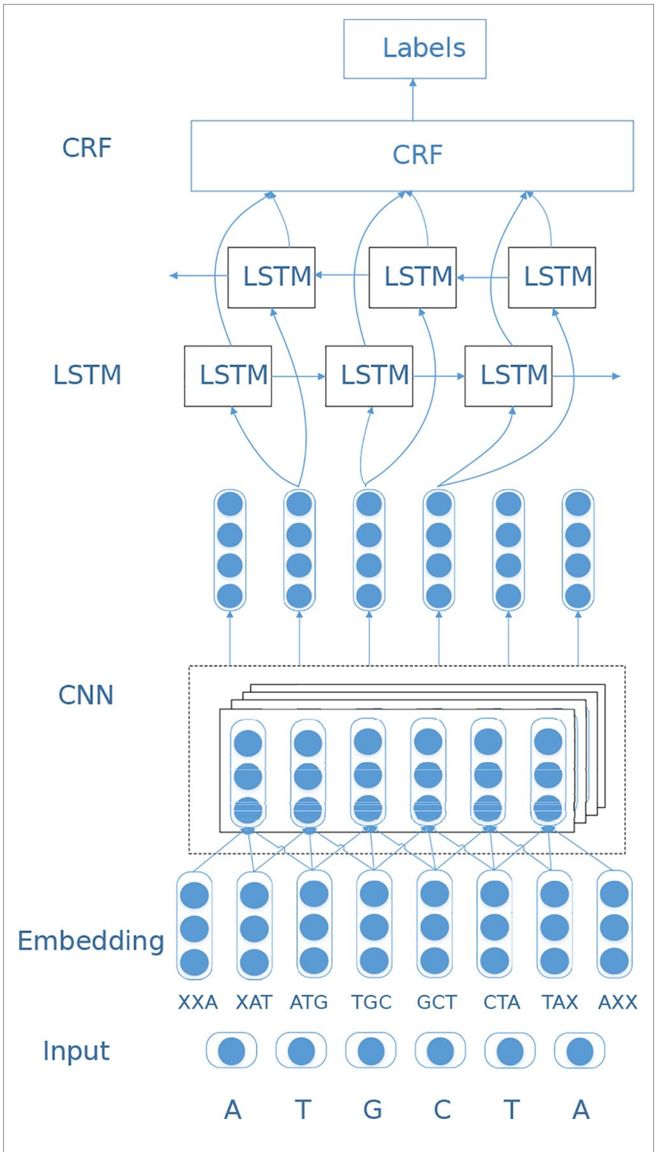


FIGURE 1 | The overall architecture of CircSLNN.

BiLSTM Layer

Till now, the mechanism of RNA–RBP-interaction has not been fully understood yet, and various factors impact the binding between RNAs and RBPs, include not only the local structural motifs and binding domains but also long-term dependencies of nucleotides. In our model, the CNN component serves as a feature extractor from raw input and learn the context information in local regions. To further exploit sequence information, we adopt bi-directional long short-term memory (BiLSTM) (Schuster and Paliwal, 1997) network. BiLSTM is a combination of forward LSTM and backward LSTM, which is a special type of recurrent neural network (RNN). It is often used to model context information in natural language processing tasks. BiLSTM was designed to learn the relationship between base before and after the current position, and to capture longer distance dependencies.

Let x_t be the input vector of the tth time step, and s_t and s'_t be the hidden states of the forward and backward calculations of the tth time step. Then the calculations of s_t and s'_t depend on $s_{t\text{-}1}$ and s'_{t+1}, respectively, as shown in Eqs. 2 and 3.

$$s_t = g(Ux_t + Ws_{t-1}) \tag{2}$$

$$s'_t = g(U'x_t + W's'_{t+1}) \tag{3}$$

where U and W are the weight matrices of the input and hidden states in the forward pass. U' and W' are the weight matrices of the input and hidden states in the backward pass.

The final output o_t of step t is a combination of a forward hidden layer and a backward hidden layer, defined as follows.

$$o_t = h(Vs_t + V's'_t) \tag{4}$$

where V and V' are the weight matrices of the hidden layers to the output layer in forward pass and backward pass, respectively.

CRF Layer

As mentioned in the *CNN Layer* and *BiLSTM Layer* sections, CNN and RNN have their respective advantages. The hybrid CNN-RNN architecture has been proposed in previous studies and achieved much better performance than using CNN or RNN alone. For instance, both CRIP (Zhang et al., 2018) and iDeepS (Pan et al., 2018) are hybrid CNN-RNN models, and both use LSTM for classification. CRIP feeds the outputs for all time-steps of the LSTM to a fully-connected layer and get the decision result, while iDeepS uses the output of the last time-step for classification. Actually, based on the output on each time-step of LSTM, it is straightforward to get the sequence labeling results. However, the raw outputs without any constraint are often meaningless, e.g. OIOI … OOI, as it is known that binding sites are continuous regions on RNA sequences. In order to avoid such cases, we add a conditional random field (CRF) layer to process the output of BiLSTM. The purpose of the CRF layer is to predict the probability of the entire sequence rather than the probability of each individual tag. The CRF layer can add some constraints to the predicted labels to ensure that the output labels are legal. During the data training process, these constraints can be automatically learned through the CRF layer, so the probability of occurrence of illegal sequences in the prediction phase will be greatly reduced. Specifically, the CRF layer calculates the conditional probability shown in Eq. 5

$$P(y_1,\ldots,y_n \mid x_1,\ldots,x_n) = \mathrm{P}(y_1,\ldots,y_n \mid x), x = (x_1,\ldots,x_n) \tag{5}$$

where $P(y|x)$ is the probability that the prediction label is y if the input is x, where x_i is the output of ith time-step by the LSTM layer.

In order to estimate the probability, CRF makes two assumptions. First, the distribution is an exponential family distribution. Second, the association between the outputs occurs only at adjacent locations, and the association is exponentially additive. This allows the probability to be calculated by the probability density function as shown in Eq. 6.

$$f(y_1,\ldots,y_n;x)=h(y_1;x)+g(y_{1,}\ y_2;x)+h(y_2;x)+$$
$$g(y2,y3;x)+h(y3;x)+\cdots+g(y_{n-1},y_{\mathrm{n}};x)+h(y_{\mathrm{n}};x) \tag{6}$$

where *f*, *g*, *h* are probability density functions and can be considered as scoring functions. The overall score *f* of all tags can be broken down into the sum of the score *h* of each individual tag and the score *g* of each pair of adjacent tags. Since LSTM is capable to learn the mapping from input *x* and its output *y*, we assume that the function *g* is independent of *x* and the final probability distribution can be formulated in Eq. 7,

$$P(y_1,\ldots,y_n\,|\,\mathrm{x})=\frac{1}{Z(x)}exp(h(y_1;x)+\sum_{k=1}^{n-1}[g(y_k,y_{k+1})+h(y_{k+1};x)]) \tag{7}$$

where the single-label scoring function $h(y_i;\ x)$ is fitted by the BiLSTM layer, thus completing the construction of the CRF layer.

EXPERIMENTAL RESULTS

Experimental Settings

In circSLNN, the number of convolution kernels in the CNN layer is 128, the convolution window size is 10, the hidden layer size of the BiLSTM layer is 256, and the activation function used by the middle layer is ReLU. The optimization algorithm is RMSProp, with batch size 512 and epoch number 20, using the early stopping mode. The performance metrics include precision, recall and F_1, which are computed based on the labels of individual nucleotides.

Prediction Performance of circSLNN

We perform experiments on all 37 datasets described in the **Data Source** section. For each dataset, we perform a 6-fold cross-validation. The original datasets are divided into 6 folds with approximately equal size (5 folds for training and validation, and one fold for test). The accuracies shown in **Table 1** are averaged over 6 times of independant test.

As can be seen, circSLNN achieves high prediction accuracy for most RBPs. The F_1 scores are higher than 0.8 on 24 out of the 37 datasets, showing the effectiveness of the sequence labeling model.

TABLE 1 | Prediction accuracies on 37 different protein datasets.

Protein	Precision	Recall	F_1-Measure
AGO1	0.820	0.853	0.836
AGO2	0.804	0.429	0.559
AGO3	0.840	0.773	0.805
ALKBH5	0.908	0.928	0.918
AUF1	0.908	0.938	0.923
C17ORF85	0.889	0.926	0.907
C22ORF28	0.847	0.828	0.838
CAPRIN1	0.881	0.789	0.833
DGCR8	0.794	0.863	0.827
EIF4A3	0.520	0.749	0.614
EWSR1	0.892	0.912	0.902
FMRP	0.473	0.679	0.557
FOX2	0.999	0.925	0.961
FUS	0.583	0.566	0.575
FXR1	0.958	0.951	0.955
FXR2	0.799	0.825	0.812
HNRNPC	0.841	0.892	0.866
HUR	0.542	0.609	0.573
IGF2BP1	0.522	0.716	0.604
IGF2BP2	0.691	0.660	0.675
IGF2BP3	0.533	0.618	0.572
LIN28A	0.543	0.702	0.613
LIN28B	0.764	0.636	0.694
METTL3	0.774	0.806	0.790
MOV10	0.805	0.808	0.806
PTB	0.609	0.597	0.603
PUM2	0.910	0.988	0.948
QKI	0.982	0.971	0.976
SFRS1	0.797	0.704	0.748
TAF15	0.916	0.968	0.941
TDP43	0.864	0.760	0.809
TIA1	0.915	0.863	0.888
TIAL1	0.836	0.824	0.829
TNRC6	0.952	0.841	0.893
U2AF65	0.848	0.796	0.821
WTAP	0.976	0.953	0.964
ZC3H78	0.848	0.790	0.818
Average	0.794	0.795	0.790

Data Encoding Analysis

In circSLNN, the inputs are pretrained embedding vectors for *k*-mers, while most of the existing methods for predicting RBP-binding sites use one-hot encoding, e.g. iDeep and DeepBind. In order to investigate the impact of encoding scheme on model performance, we compare one-hot and our embedding vectors on the same datasets. We randomly choose 5 RBPs. **Figure 2** depicts the comparison results.

Apparently, the pretrained embedding vectors perform much better than the one-hot vectors. The average F_1 score is increased by 0.087. This result suggests that the word embedding encoding method can effectively extract the feature information of RNA sequences from the human genome database, and can effectively improve the performance of the binding site predictor.

The Role of CNN Layer

Compared to ordinary text sequence labeling tasks, we introduce the CNN layer to extract local features from RNA sequences. The purpose of the CNN layer is to characterize the local sequence pattern surrounding the base to be labeled, and encode each individual base with richer information. Here we assess the contribution of CNN by removing it from the model. The inputs of the LSTM-CRF model are the pretrained *k*-mer embedding vectors. Specifically, for each base, we choose the embedding vector of the fragment that centered by the base as its feature

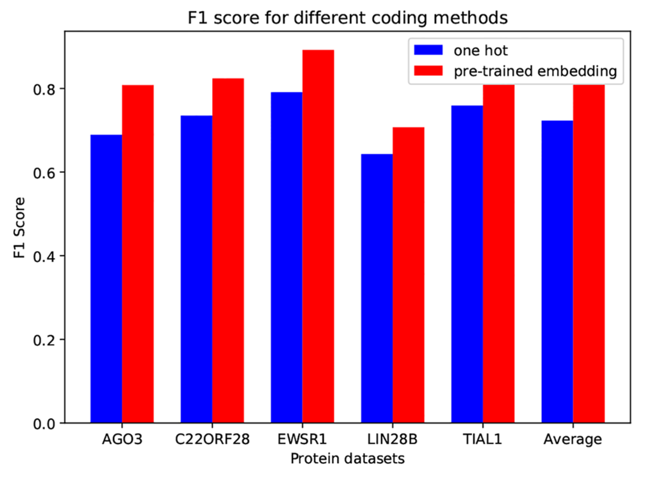


FIGURE 2 | F_1 Score for Different Coding Methods.

vector. The following training on LSTM and CRF is the same as circSLNN. We compare the performance of the two methods on five randomly selected data sets, as shown in **Figure 3**.

As can be seen, the average F_1 is increased by 0.021 by introducing CNN layer. Although the overall improvement seems not significant, we find that CNN has larger contribution for the difficult datasets, e.g. HUR and LIN288, compared with easy datasets, indicating the importance of further feature learning from raw inputs.

Comparison of Different Sequence Labeling Schemes

The sequence labeling scheme used in this study is IO tag, not the BIO or BME (BME is short for begin, middle and end) that commonly used in text labeling tasks (Carpenter, 2009), as binding sites generally span tens of bases in length, whereas common text labeling objects only consist of several words, such as a typical place name in the named entity recognition mission (NER), 'Shanghai Jiao Tong University'. In order to assess the performance of these three tag systems, we conduct experiments on five randomly selected protein datasets, as shown in **Figure 4**.

As can be seen, the IO tag system outperforms BIO and BME by a large margin. BIO and BME have close performance. We find that the B-coded labeling systems can hardly find tag B in the test set, i.e. their results contain only tag I and tag O. The reason is that the B tag is extremely sparse due to the long binding sites, which leads to an imbalanced distribution of tags, and it is very hard to recognize tag B.

Investigation on Positive-to-Negative Data Ratio

In our experiments, the positive-to-negative ratio for all datasets is 1:1, which is the same as previous studies (Pan and Shen, 2017), (Zhang et al., 2018). However, the length of human circRNAs could be tens of thousands bases, including 1–5 exons (Memczak et al., 2013), while the binding sites are small regions and very sparse on the sequences. That is to say, the true ratio between positive and negative data is very small, leading to an extremely imbalanced problem, thus most studies adopt a sampling strategy to control the ratio. Here, to get closer to the actual situation, we compare the performance of circSLNN under different positive-to-negative ratios, i.e. 1:1, 1:2, and 1:4. The results are shown in **Figure 5**.

Note that although adding negative samples results into data imbalance, the increase in data volume is beneficial for training the model. As shown in **Figure 5**, the accuracies on some datasets, e.g. LIN28B, LIN28B, and TDP43, have even been increased by using expanded negative set. Generally, the performance of circSLNN has little variance when expanding negative set several times, showing the model robustness.

Comparison With the Existing Methods on Sequence Labeling for Full-Length circRNAS

In order to assess the performance of circSLNN in real cases, we conduct experiments on full-length circRNAs instead of sampled

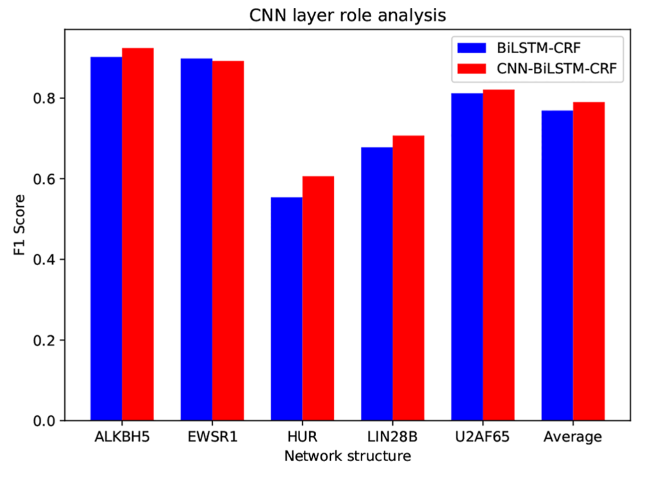


FIGURE 3 | Performance comparison between models with and without the CNN layer.

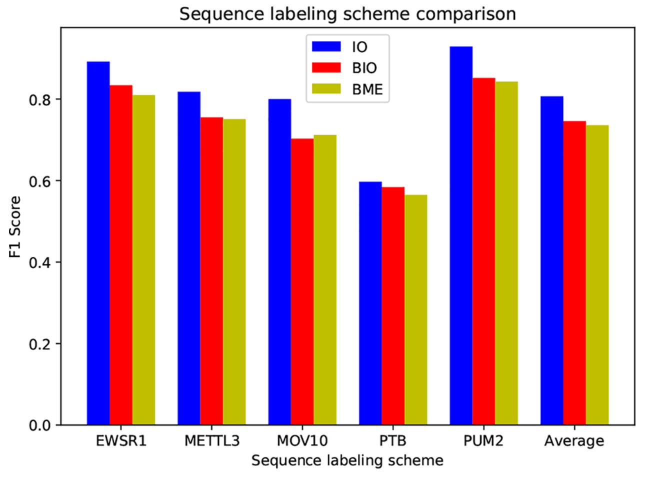


FIGURE 4 | Performance comparison on three sequence labeling schemes.

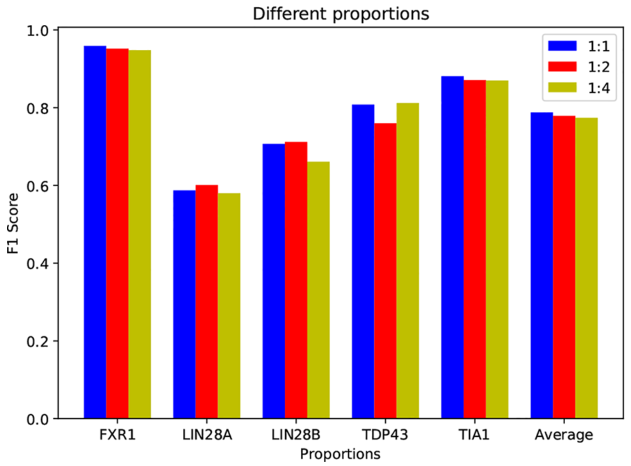


FIGURE 5 | Performance on datasets with different positive-to-negative data ratios.

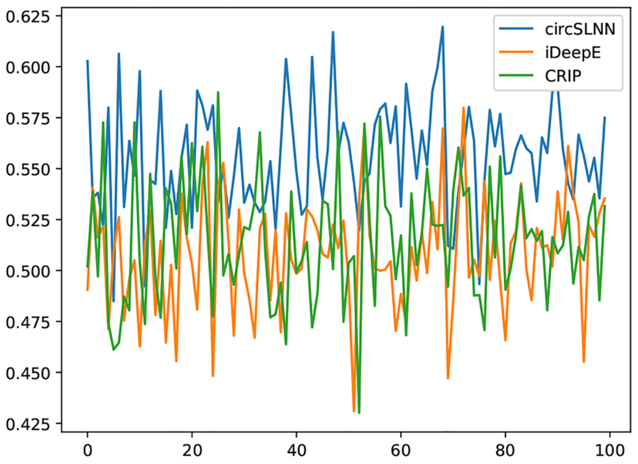


FIGURE 6 | F_1 score on 100 full-length RNAs.

segments in the datasets, and compare it with the state-of-the-art predictors for RNA–RBP binding sites.

To the best of our knowledge, circSLNN is the first sequence labeling model for identifying RBP-binding sites on circRNAs. Therefore, for the convenience of comparison, we need to process the output of the existing classification models, i.e. converting the labels for segments into labels for individual nucleotides. Specifically, for a full-length RNA, we divide it from beginning to end into 101-nt fragments. For each fragment, the circSLNN model is used to predict whether each base belongs to the binding site. If it belongs, it is marked as 1; otherwise, it is marked as 0. For the classification model, whether the fragment belongs to the binding site is predicted. If the fragment is predicted as positive, then all the bases in the sequence are labeled by 1, otherwise all bases are labeled by 0. In this way, we obtain the label sequences of full-length RNAs predicted by two different models. By comparing the predicted sequence labels with the actual labels, we can calculate the F_1 score.

We collect a dataset of 100 full-length circRNAs that are bound to different RBPs. They are first segmented into 101-nt segments, and then fed to the classification models and sequence labeling model, respectively, to predict the binding sites. F_1 scores are computed based on individual bases. The results are shown in **Figure 6**.

As can be seen from the results, circSLNN achieves the highest F_1 on almost all circRNAs in the dataset. The average F_1 score of circSLNN reaches 0.568, while the average F_1 scores of iDeepE (Pan and Shen, 2018) and CRIP (Zhang et al., 2018) are 0.504 and 0.494, respectively. This suggests that the sequence labeling model can more accurately identify the position of the binding site, which is important for further verification of the interaction regions using biological experiments.

Despite the advantages over other methods, we can find that the overall accuracy is much lower than that computed on the short segments (the average F_1 of 37 test sets is 0.790 as shown in **Table 1**). It is mainly due to the extremely imbalanced class distribution in this new test set. In training sets, the positive-to-negative ratio is 1:1, while when the full-length circRNAs are segmented, most of them contain no binding site at all. Although the model can handle imbalanced distribution to some extent as described in the *Investigation on Positive-to-Negative Data Ratio* section, the performance decreases greatly when the data set is severely imbalanced.

DISCUSSION

This study aims to develop a machine learning model for identifying RBP-binding sites on RNAs. The existing prediction methods consider this problem as a classification problem, which divide RNA sequences into fragments and predict whether or not binding sites exist in the fragments. To further predict the location and length of binding sites, we propose a sequence labeling model, circSLNN, which assigns a label to each base in fragments instead of the whole fragments, so as to provide more information of the binding regions. Besides, considering the lack of tools designed for circRNAs, circSLNN is specially trained by circRNA datasets. Although trained on circRNAs, circSLNN provides a general sequence labeling framework that can be applied to all types of RNAs.

Despite the enhancement of performance, this study is still a preliminary exploration on characterizing binding sites on circRNAs. The first limitation lies in the input features. As it is known that the interaction between RNAs and other molecules has complex mechanisms, especially the circRNAs that have not been well studies, the prediction of circSLNN is based only on circRNA sequences, which is a very limited information source. One future research direction is to incorporate more biological properties or domain knowledge related to circRNAs.

Second, although we have used a hybrid neural network, the proposed model structure is relatively simple. In recent years, not only new embedding training methods but also deep architecture have emerged in the field of natural language processing (Devlin et al., 2018), (Peters et al., 2018), which have achieved substantial improvement on a variety of tasks. Many of them could be adapted to biological sequence analysis, thus our network structure still has a lot of room for improvement.

Third, because the lengths of circular RNA sequences vary greatly, ranging from a few hundred to several millions, which seriously affects the training of the model. Most of the predictors including circSLNN are trained on short segments of RNAs, which may lose some information of whole RNAs and lead to high false-positive-rate. Better predictions based on full-length RNAs or longer segments are the focus of our future work.

CONCLUSION

This study proposes a sequence labeling neural network for predicting RBP-binding sites on circRNAs, called circSLNN. To fully exploit sequence information, we train continuous embedding vectors for 10-mers of RNAs using the whole human genome sequences, and we construct a hybrid CNN–LSTM–CRF network to perform the sequence labeling task. The purpose of using a hybrid model is to combine the advantages of two deep architectures and to obtain better high-level abstract feature representations for classification. We train circSLNN on 37 datasets of circRNA fragments, and the average F_1 score is 0.790. The experimental results show that it is feasible to use the sequence labeling method for identifying binding sites on circRNAs. Both the RNA fragment embedding vectors and the hybrid architecture contribute to improved performance. Compared with the classification model, it can more accurately label the position of the binding site on the full-length RNAs. The proposed model will help researchers study the circRNA–RBP-interactions and reveal regulatory functions of circRNAs.

DATA AVAILABILITY STATEMENT

All datasets generated/analyzed for this study are available at https://github.com/JuYuqi/circSLNN.

AUTHOR CONTRIBUTIONS

YJ, LY, YY and HZ designed the model. YJ and LY implemented the model and performed the experiments. YJ, LY, YY and HZ analyzed the results and drafted the article. YY and HZ supervised this work.

FUNDING

This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011) and the National Natural Science Foundation of China (No. 61972251).

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831. doi: 10.1038/nbt.3300

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., and Cherry, J. M. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25. doi: 10.1038/75556

Bolognani, F., and Perrone-Bizzozero, N. I. (2008). Rna–protein interactions and control of mrna stability in neurons. *J. Neurosci. Res.* 86, 481–489. doi: 10.1002/jnr.21473

Carpenter, B. (2009). Coding chunkers as taggers: Io, bio, bmewo, and bmewo+. *LingPipe Blog*. Available at: lingpipe-blog. com/2009/10/14.

Chen, W., Feng, P.-M., Lin, H., and Chou, K.-C. (2013). irspot-psednc: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* 41, e68. doi: 10.1093/nar/gks1450

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). *Bert: pre-training of deep bidirectional transformers for language understanding*. arXiv: Computation and Language.

Dudekula, D. B., Panda, A. C., Grammatikakis, I., De, S., Abdelmohsen, K., and Gorospe, M. (2016). Circinteractome: a web tool for exploring circular rnas and their interacting proteins and micrornas. *RNA Biol.* 13, 34–42. doi: 10.1080/15476286.2015.1128065

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. doi: 10.1093/database/bay044

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by micrornas: are the answers in sight? *Nat. Rev. Genet.* 9, 102. doi: 10.1038/nrg2290

Harris, D., and Harris, S. (2010). *Digital design and computer architecture*. Morgan Kaufmann.

Hendlich, M., Rippmann, F., and Barnickel, G. (1997). Ligsite: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graphics Model.* 15, 359–363. doi: 10.1016/S1093-3263(98)00002-3

Khalil, A. M., and Rinn, J. L. (2011). "Rna–protein interactions in human health and disease", in *Seminars in cell & developmental biology*, vol. 22. (Elsevier), 359–365. doi: 10.1016/j.semcdb.2011.02.016

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems* 141 (5), 1097–1105.

Kumar, M., Gromiha, M. M., and Raghava, G. (2008). Prediction of rna binding sites in a protein using svm and pssm profile. *Proteins* 71, 189–194. doi: 10.1002/prot.21677

Lafferty, J., McCallum, A., and Pereira, F. C. (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", in *Proceeding of the 2001 international conference on machine learning*, 282–289.

Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf.* 10, 168. doi: 10.1186/1471-2105-10-168

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436. doi: 10.1038/nature14539

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Liu, R., and Hu, J. (2011). Hemebind: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinf.* 12, 207. doi: 10.1186/1471-2105-12-207

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., and Chen, L. (2010). Prediction of protein–rna binding sites by a random forest method with combined features. *Bioinformatics* 26, 1616–1622. doi: 10.1093/bioinformatics/btq253

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

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting rna-protein interactions using only sequence information. *BMC Bioinf.* 12, 489. doi: 10.1186/1471-2105-12-489

Pan, X., and Shen, H.-B. (2017). Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC Bioinf.* 18, 136. doi: 10.1186/s12859-017-1561-8

Pan, X., and Shen, H.-B. (2018). Predicting rna–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* 34, 3427–3436. doi: 10.1093/bioinformatics/bty364

Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of rna-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* 19, 511. doi: 10.1186/s12864-018-4889-1

Pennington, J., Socher, R., and Manning, C. (2014). "Glove: global vectors for word representation", in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543. doi: 10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., and Lee, K., et al. (2018). "Deep contextualized word representations" in *North american chapter of the association for computational linguistics*, 2227–2237.

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi: 10.1109/78.650093

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., and Chen, K. (2007). Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci.* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Song, X., Zhang, N., Han, P., Moon, B.-S., Lai, R. K., and Wang, K. (2016). Circular rna profile in gliomas revealed by identification tool uroborus. *Nucleic Acids Res.* 44, e87–e87. doi: 10.1093/nar/gkw075

Xiao, Y., Cai, J., Yang, Y., Zhao, H., and Shen, H. (2018). "Prediction of microrna subcellular localization by using a sequence-to-sequence model", in *Proceedings of the 2018 International Conference On Data Mining (ICDM)* 1332–1337. doi: 10.1109/ICDM.2018.00181

Zhang, K., Pan, X., Yang, Y., and Shen, H.-B. (2018). Predicting circrna-rbp interaction sites using a codon-based encoding and hybrid deep neural networks. *bioRxiv*, 499012. doi: 10.1101/499012

Zou, M., and Conzen, S. D. (2004). A new dynamic bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* 21, 71–79. doi: 10.1093/bioinformatics/bth463

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: 06 December 2019
doi: 10.3389/fgene.2019.01261

Frin: An Efficient Method for Representing Genome Evolutionary History

*Yan Hong and Juan Wang**

School of Computer Science, Inner Mongolia University, Hohhot, China

Phylogenetic analysis is important in understanding the process of biological evolution, and phylogenetic trees are used to represent the evolutionary history. Each taxon in a phylogenetic tree has not more than one parent, so phylogenetic trees cannot express the complex evolutionary information implicit in phylogeny. Phylogenetic networks can be used to express genome evolutionary histories. Therefore, it is great significance to research the construction of phylogenetic networks. Cass algorithm is an efficient method for constructing phylogenetic networks because it can construct a much simpler network. However, Cass relies heavily on the order of input data, i.e. different networks can be constructed for the same dataset with different input orders. Based on the frequency and incompatibility degree of taxa, we propose an efficiently improved algorithm of Cass, called as Frin. The experimental results show that the networks constructed by Frin are not only simpler than those constructed by other methods, but Frin can also construct more consistent phylogenetic networks when the treated data have different input orders. Furthermore, the phylogenetic network constructed by Frin is closer to the original information described by phylogenetic trees. Frin has been built as a Java software package and is freely available at https://github.com/wangjuanimu/Frin.

Keywords: evolution, phylogenetic network, incompatibility degree, frequency, genome

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Junwei Luo,
Henan Polytechnic University, China
Junwei Han,
Harbin Medical University, China
Dong Chen,
Heilongjiang Institute of Technology, China

****Correspondence:***
Juan Wang
wangjuan@imu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *18 October 2019*
Accepted: *14 November 2019*
Published: *06 December 2019*

Citation:
Hong Y and Wang J (2019) Frin: An Efficient Method for Representing Genome Evolutionary History. Front. Genet. 10:1261. doi: 10.3389/fgene.2019.01261

INTRODUCTION

Studying the evolution of species is helpful for humans to reveal biological secrets, prevent, and treat diseases. The purpose of phylogenetic analysis is to reveal the evolutionary relationships between different species or taxa and study the evolution of life on Earth (Huson and Scornavacca, 2011). The evolutionary history is like the growth of trees, and all species can be traced back to a common ancestor. It makes sense to use trees to represent the evolutionary history, in which each node except the root has only one parent. There are a number of reticulate evolutionary events, such as reversal, translocation, and fusion, which have resulted in more than one parent of some taxa in the evolution (Gusfield et al., 2007a; Gusfield et al., 2007b; Kelk and Scornavacca, 2014; Wu, 2010; Van Iersel et al., 2017). Such a complex evolutionary history can be represented by the phylogenetic networks (Doolittle, 1999; Nakhle, 2010; Yu and Nakhleh, 2015; Huber et al., 2018). A network is a generalization of a tree in that it contains nodes with in-degree greater than one (Iersel et al., 2009). Phylogenetic networks are functionally classified into implicit networks and explicit networks (Huson et al., 2007; Huson and Rupp, 2008; Van Iersel et al., 2010). Implicit networks can be used to represent conflicting patterns due to the model misspecification. However, explicit networks can capture reticulate evolutionary events.

In recent years, a lot of work has been developed on the methods for constructing phylogenetic networks (Albrecht 2015; Albrecht et al., 2012; Bordewich et al., 2007; Francis et al., 2018; Gambette et al., 2017; Linz and Semple, 2009; Makarenkov et al., 2006; Mirzaei and Wu, 2016; Jansson and Sung, 2006). Cluster network method uses the network-popping algorithm to construct an implicit network, which can be drawn as a cladogram (Huson and Rupp, 2008). Galled network method uses the seed-growing algorithm to find the solution of RMCS (Restricted Maximum Compatible Subset) problem for input dataset, and then construct phylogenetic network (Huson et al., 2007). The relationships between phylogenetic trees and networks are the basis for the reconstruction and verification of phylogenetic networks. TCP algorithm solved the problem whether or not certain existing phylogenetic trees are displayed in a phylogenetic network (Gunawan et al., 2016; Gunawan et al., 2018). Cass is an efficient method to construct a phylogenetic network for any input trees, and is able to construct much simpler networks than other available methods (Van Iersel et al., 2010). But Cass usually constructs some different networks for the same dataset when it is input as different orders. The phylogenetic network constructed by Cass represents lots of redundant information except for the original information. Both factors considered it is obvious that Cass has poor practical application. Lnetwork improves the Cass by fixing the order of removed taxa in the construction process of phylogenetic networks. It saves the running time for us and reduces the dependence on the input data order (Wang et al., 2013a). BIMLR is also an improved algorithm of Cass by considering incompatibility of taxa in the construction process of phylogenetic network (Wang et al., 2013b). Such methods, including Cass, Lnetwork, and BIMLR, have the significant flexibility that they are not restricted to binary input trees and are not restricted to trees on the same taxa set. In addition, they can construct simpler networks for the same input than other methods, although they are relatively slow. Therefore, The above three methods are efficient and widely used in the construction of phylogenetic networks.

In this paper, we will introduce another improved Cass algorithm, Frin. It constructs phylogenetic networks with phylogenetic trees as input, just like Cass algorithm. Experiments show that Frin is less dependent on the input data order and runs faster than Cass. Moreover, Frin constructs a simpler network than other available methods.

PRELIMINARIES

Related Knowledge

Given a set of taxa X, a subset of X, excluding the empty set and the complete set, is called a cluster. A cluster C is non-trivial if it contains more than one element. If two clusters C_1' and C_2' are compatible if either $C_1' \bigcap C_2' = \phi$ or $C_1' \subset C_2'$ or $C_2' \subset C_1'$. Otherwise, they are incompatible. For a set of cluster Y on X, Y is said to be compatible if any one pair of clusters are compatible. An incompatible cluster set is represented by an incompatible graph $IG(Y) = (E, V)$, which consists of a node set and an edge set. The node set consists of all the non-trivial clusters in the Y and the edge set consists of edges connecting the incompatible clusters. The set of clusters represented by a rooted phylogenetic tree is compatible; on the contrary, any one compatible cluster set can be constructed into a rooted phylogenetic tree.

Supposed that $N = (V, E)$ is a network on taxa set X. $\delta^{-}(v)$ represents the in-degree of the node v. We introduce a concept used to describe the complexity of a network, which is called reticulation number. Reticulation number of a network is not necessarily equal to the number of reticulate nodes. It is defined as:

$$\sum_{v \in V,\, \delta^{-} > 0} (\delta^{-}(v) - 1) = |E| - |V| + 1$$

If each connected component of a network contains reticulation number at most k, then we call that it is a *level-k* network. A level-k network is called a simple level- $< k$ network, which does not contain cut nodes. A node is a cut node if its removal disconnects the graph.

Each phylogenetic tree T is uniquely defined by the set of clusters. For a phylogenetic tree, an edge $e=(u, v)$ represents the cluster containing those taxa that are descendants of v. Similarly, a phylogenetic network represents clusters in the soft-wired sense or in the hard-wired sense. For each reticulate node of the network N, we switch on its one incoming edge and switch off the others, and we called the network N represents the cluster C in the soft-wired sense if cluster C equals the set of all taxa that can be reached from v. On the other hand, if cluster C equals the set of taxa that are descendants of v, we said the edge (u, v) of a network represents the cluster C in the hard-wired sense. In this article, we research the representing in the soft-wired sense, whose pseudocode is shown by Algorithm 1.

ALGORITHM 1 | The clusters represented by a network in the soft-wired sense.

```
Input: a phylogenetic network (level-$k$)
Output: a cluster set $Y$
Begin
Y = null
$i$ = $k$-1; $j[k]$ = false
soft ($N$, $i$, $j$)
  for: $v \in V$ of $N$
    if $i$ < 0 then
      if $j$ = true then
        switch on the left incoming edge of each reticulate node and
        switch off the right one
      else
        switch off the left incoming edge of each reticulate node and
        switch on the right one
      end if
      for: $v \in V$ of $N$
        if out-degree($v$) = 0 then
          add a cluster represented by $v$ to $Y$
        else
          add clusters represented by the child of $v$ to $Y$
        end if
      end for
    else
      $j[i]$ ← true
      continue: soft ($N$, $i$-1, $j$)
      $j[i]$ ← false
```

```
            continue: soft (N, i-1, j)
        end if
    end for
return the cluster set Y
End
```

Cass, Lnetwork, BIMLR, and Frin all take the set of trees as the input when to construct a phylogenetic network. They first compute all clusters represented by input trees, and then construct a phylogenetic network representing those clusters. Assume that Y is the cluster set represented by the input file, N is a constructed network. Y' is the cluster set represented by the network, which are greater than or equal to the clusters in the Y. The clusters in Y-$Y'$are called the redundant clusters. Both the reticulation number and the number of redundant clusters describe the complexity of a network. The best phylogenetic network should contain fewer reticulation numbers and have fewer redundant clusters.

Suppose that N is a network on taxa set X, $e = (u, v)$ is an edge of N with parent node u and child node v. If each way from the root node to v passes through u, we called that u is the stable ancestor on v; otherwise, it is the unstable ancestor. For an edge $e = (u, v)$, let $\mathrm{P}(e) = \{x \in X | \ x$ is the stable ancestor on $v\}$, $\mathrm{Q}(e) = \{x \in X \mid x$ is the unstable ancestor on $v\}$, $\mathrm{S}(e) = \{x \in X \mid x$ is not a descendant of $v\}$. We call $\{\mathrm{P}(e), \mathrm{Q}(e), \mathrm{S}(e)\}$ the tripartition of e. $\Theta(N)$ represents all tripartition sets of network N. Given two networks N_1 and N_2, tripartition distance between them is computed by $|\Theta(N_1)\ \ \Theta(N_2)|/2$, of which Δ is symmetry variation. The tripartition distance measures the topology different between two phylogenetic networks. In this paper, we use the tripartition distance to measure the dissimilarity of the phylogenetic networks.

Cass Algorithm

We will have a brief description for Cass algorithm in the following. Given a set of clusters Y on taxa X, Cass algorithm is divided into four steps:

Step 1: Cass works out non-trivial connected component $Y_1,\ldots,Y_p$ of incompatibility graph $IG(\mathrm{Y})$. Then, Cass collapses the maximal ST-sets for each non-trivial connected component $Y_i$and gets Y_i'. Given a taxa set X and a subset $S \subset X$, each cluster $C \subset Y$ removes the elements of subset S, and the remaining cluster set Y' is called the restriction of Y to S, denoted by $Y|s$. The largest set of ST-set is called the maximal ST-set. Given $|\mathrm{S}|>1$, if S is compatible with each cluster of Y, and $Y|s$ are compatible, we called S is a strict tree set (ST-set) of Y.

Step 2: Cass (k) constructs simple level- $< k$ networks for each Y_i', which is crucial step of Cass algorithm. For each non-trivial connected component, Cass(k) loops all taxa and removes them from each cluster, and collapses all of the maximal ST-sets for the remaining cluster set. Cass(k) repeats above operations k times, until the remaining cluster set is compatible to construct phylogenetic trees. The removed taxon is added to the phylogenetic tree as children of reticulate nodes, which becomes a simple level- $< k$ network.

Step 3: For each $i \in \{1,\ldots,p\}$, Cass removes all clusters that are in C_i, adds a cluster X_i and each maximal subset $X \subset X_i$ that is not separated by C_i. All above set become cluster set C''. Then Cass constructs a rooted phylogenetic tree T for C'', which is the whole frame of the resulting network.

Step 4: Cass adds all the simple level- $< k$ networks constructed in step 2 to the rooted phylogenetic tree T by the method of ancestor nodes displacement.

When Cass starts constructing a simple level- $< k$ network, it does not know the number of network level. Thus, it first sets $k = 0$ and runs Cass(0),which constructs a simple level- < 0 network. If such a network exists, it outputs the result and halts. Otherwise, Cass continues to sets $k = k + 1$, and runs Cass(1), Cass(2),…, Cass(k), until the constructed network represents the given clusters sets the soft-wired sense. The process is very time-consuming, because Cass(k) loops over all taxa and repeatedly attempts to remove each taxon. The selection of removed taxa is highly uncertain, which makes the algorithm depend heavily on the order of input data, and it also reduces the speed of the construction.

METHOD

Given a set of clusters Y on taxa set X, the frequency of a taxon $x \in X$ is the number of clusters containing taxon x, denoted by $f(x)$. The number of edges of the graph $IG(Y)$ is called incompatibility degree of Y, denoted by $d(Y)$. The incompatibility degree of a taxon $x \in X$, denoted by $d(x)$, is the result of subtracting the incompatibility degree of $Y_{|\mathrm{X}|\{\mathrm{x}\}}$ from that of Y, i.e. $d(x) = d(Y) - d\ (Y_{|X|\{x\}})$. For example, given incompatible cluster set $Y = \{1, 2\}, \{2, 3\}$, we can get taxa frequency $f(1) = 1, f(2) = 2, f(3) = 1$ and taxa incompatibility degree $d(1) = 0$, $d(2) = 1$, $d(3) = 0$. Moreover, we know that only by removing taxa 2, the remaining clusters are compatible. Frequency and incompatibility degree of taxa contribute a lot to the compatibility of a cluster set, which will affect the construction of phylogenetic networks. The premise of constructing a network is to construct a phylogenetic tree for the compatible cluster set, which is the result by removing some taxa from the originally incompatible set of clusters. The key of Frin method lies in the addition of taxa removal rules, which makes the algorithm select removed taxa more efficiently. Frin chooses the removed taxa based on its frequency and incompatibility degree. Such choices make the remaining cluster set compatible as quickly as possible.

Frin constructs phylogenetic networks in four steps; steps 1, 3, and 4 are the same as Cass algorithm. Frin improves the step 2 of the Cass for the construction of simple level- $< k$ networks. Frin first find the non-trivial connected components of the incompatibility graph $IG(Y)$; next it constructs simple level- $< k$ network based on taxa frequency and incompatibility degree; then it constructs a unique phylogenetic trees for compatible clusters; finally it integrates simple level- $< k$ networks into the resulting phylogenetic networks. Frin (k) constructs a simple level- $< k$ network as follows.

For each taxon $x \in X'$, Frin(k) obtains the frequency and incompatibility degree, and then calculates the weighted value |equ_0013.eps| on the frequency and incompatibility degree, i.e. $s(x) = p \times f(x) + q \times d(x)$, where p and are q weight values of its frequency and incompatibility degree. All taxon $x \in X'$ are ordered according to the value of *s*. Frin(k) selects the taxon with the maximum *s* as the removed taxa each time, until the remaining cluster set is compatible to construct a phylogenetic tree. Then Frin(k) adds all the removed taxa to the tree as the child of reticulate nodes, and gets a resulting network representing all clusters. Here, we set the value of p and q, $0 < p \leq 1$, $0 \leq q <1$, $p + q = 1$, and step size is 0.1. Then we can get ten groups of p and q values, for each group of values, Frin(k) constructs only one network. To avoid the same network that can be constructed over and over again when it runs, we ignore constructing the same network as before by comparing the taxa removal process. Finally, Frin constructs one or more different networks, and records the network with less reticulation number and redundant clusters as the final phylogenetic network.

In addition, Frin sometimes adds dummy taxa to construct a network. The dummy taxa are removed before outputting the resulting network.

Example 3.1, given taxa set $X = \{1, 2, 3, 4, 5\}$ and the cluster set $Y = \{\{1, 2\}, \{1, 4\}, \{3, 4\}, \{1, 3, 4\}, \{4, 5\}, \{1, 2, 3, 4\}, \{2, 3\}, \{2, 3, 4\}, \{2, 3, 4, 5\}\}$, Frin constructs two different networks $N_1$and N_2 for Y, as shown in **Figure 1**. N_1 is a level-3 network with $r = 3$, $c = 3$ and N_2 is a level-3 network with $r = 3$, $c = 6$, where r is the reticulation number and c is the number of redundant clusters. The two networks have the same reticulation number, and N_1 has fewer redundant clusters than N_2. Therefore, Frin outputs N_1 as the final network. The example shows that Frin can construct several different networks for each input trees due to the coefficients' uncertainty of the taxa frequency and incompatibility degree. By comparing the number of reticulation nodes and redundant clusters, we select the optimal network from different networks as the output.

Example 3.2, we consider the taxa set $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and the cluster set $Y = \{\{7, 8, 9\}, \{2, 3, 4, 7, 8, 10\}, \{5, 6, 7, 8, 9\}, \{2, 3, 4, 5, 6, 7, 8, 9\}, \{2, 3, 4, 5, 6\}, \{2, 3, 4, 10\}, \{2, 3, 4, 5, 6, 7, 8, 10\}\}$. We take the cluster set Y for example to illustrate that the input data order has different influence degree on Frin, Cass, BIMLR and Lnetwork. Then we need to give all permutations of input data, and construct networks for each permutation. We represent the difference between the resulting networks by tripartition distance. For all permutations of the input data, Frin can construct the same network N_3, as shown in **Figure 2**. Cass constructs three different networks N_4, N_5, and N_6, and the minimum, maximum, and mean tripartition distance between them are 1.5, 2, and 1.67 respectively, as shown in **Figure 3 |** N_4, N_5 and $N_6$are the networks constructed by Cass for all permutations of input data in Example 3.2. BIMLR constructs three different networks N_7, N_8, and N_9, and the minimum, maximum and mean tripartition distance between them is 1, 3, and 2, as shown in **Figure 4**. Lnetwork also constructs three different networks N_{10}, N_{11}, and N_{12}, and the minimum, maximum, and mean tripartition distance between them is 1, 1.5, and 1.33, as shown in **Figure 5**. The example shows that Frin can construct more consistent networks than other methods for the same data with different input order, i.e. Frin reduces the influence of input data order. The conclusion will be demonstrated by the following section.

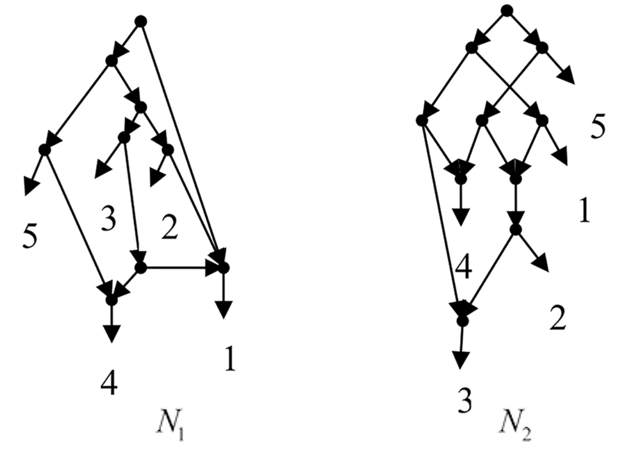


FIGURE 1 | Two networks N_1 and N_2 are constructed by Frin for the cluster set of Example 3.1.

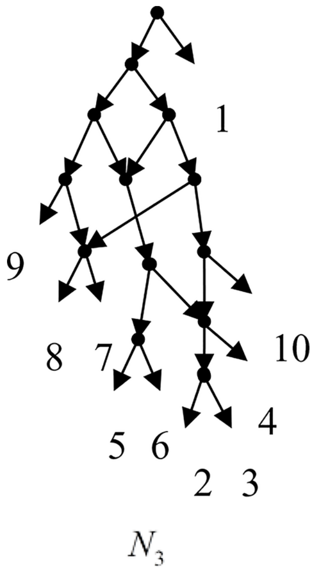


FIGURE 2 | N_3 is the network constructed by Frin for all permutations of input data in Example 3.2.

RESULTS

The experiments are performed on a personal computer with an Intel Core i5-4200U, 1.6GHz CPU, and 4GB RAM. All programs are written in Java.

We test the efficiencies of Frin, Cass, Lnetwork, and BIMLR on artificial and the practical dataset, which can be accessed from the website (https://sites.google.com/site/cassalgorithm/data-sets). The results are shown in **Tables 1–3**. On the one hand, we use practical data to test the influence of input data order on constructing network (see **Table 1**). On the other hand, we compared the network complexity, i.e. the level; the reticulation

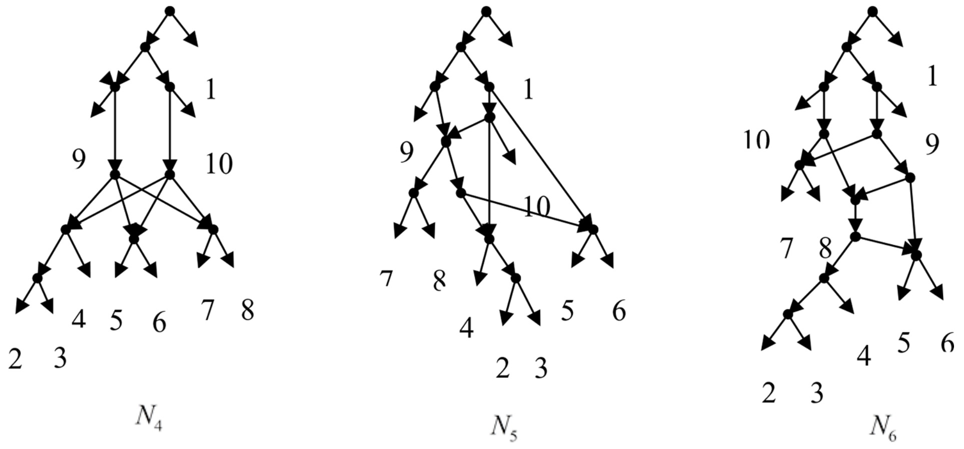


FIGURE 3 | N_4, N_5 and N_6 are the networks constructed by Cass for all permutations of input data in Example 3.2.

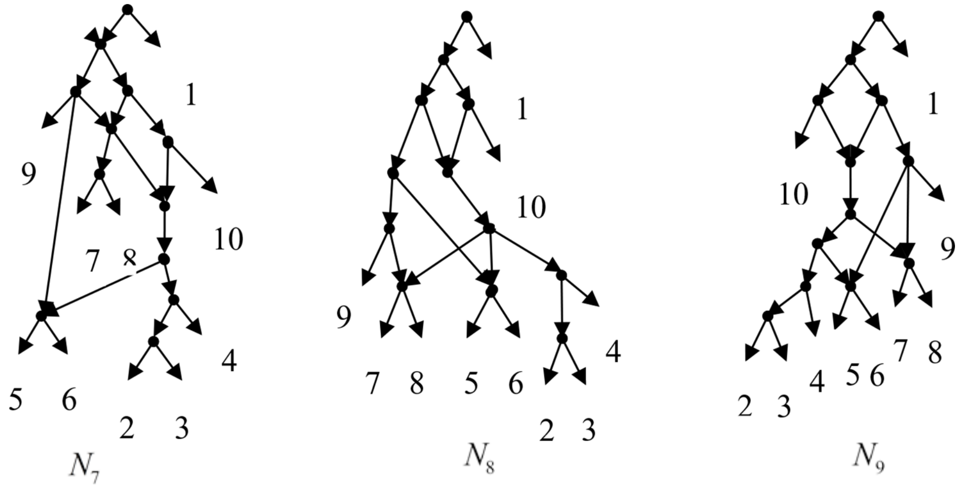


FIGURE 4 | N_7, N_8 and N_9 are the networks constructed by BIMLR for all permutations of input data in Example 3.2.

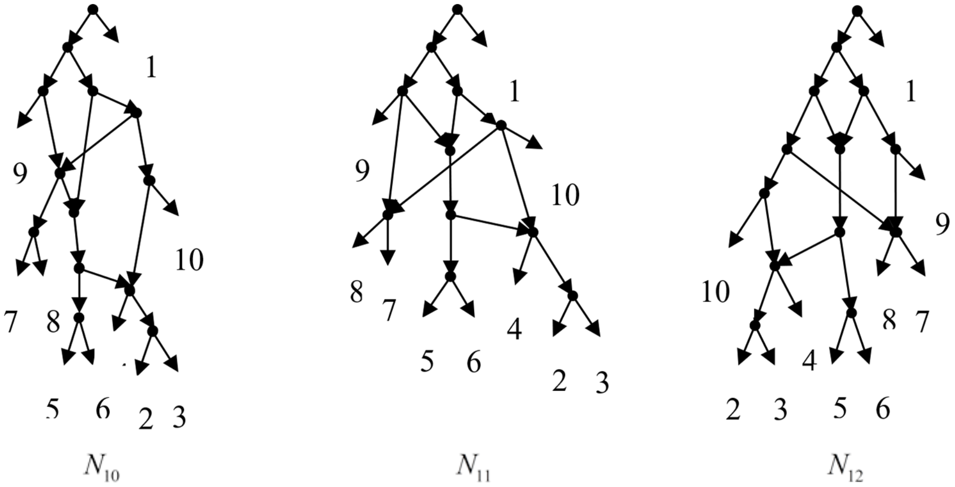


FIGURE 5 | N_{10}, N_{11} and N_{12} are the networks constructed by Lnetwork for all permutations of input data in Example 3.2.

TABLE 1 | The results of Frin, Cass, Lnetwork and BIMLR on practical datasets with clusters |*C*| and taxa |*X*| when input order is different.

Data		Firm				Cass				Lnetwork				BIMLR			
|*C*|	|*X*|	n	mean	min	max	n	mean	min	max	n	mean	min	max	n	mean	min	max
35	22	1	0	0	0	2	6.5	6.5	6.5	1	0	0	0	1	0	0	0
25	15	1	0	0	0	2	3	3	3	1	0	0	0	1	0	0	0
22	13	2	1.5	1.5	1.5	2	0.5	0.5	0.5	2	1	1	1	2	1.5	1.5	1.5
27	15	3	3.3	1	5	3	3	3	3	2	1	1	1	2	1	1	1
25	13	1	0	0	0	4	6.3	2	7.5	3	1.2	0.5	1.5	1	0	0	0
22	11	2	5.5	5.5	5.5	3	3	2.5	3.5	1	0	0	0	1	0	0	0
17	10	1	0	0	0	3	2	1.5	2.5	3	1.3	1	1.5	3	2	1	3
13	8	1	0	0	0	4	3.6	1.5	4	2	1	1	1	1	0	0	0
23	11	1	0	0	0	4	5.6	3	7.5	2	1	1	1	2	1	1	1
18	10	1	0	0	0	4	1.5	0.5	3	3	2.5	1.5	3.5	3	1.5	0.5	2.5
22	11	2	0.5	0.5	0.5	3	3.2	1.5	5	1	0	0	0	2	0.5	0.5	0.5
12	11	1	0	0	0	2	3	3	3	1	0	0	0	1	0	0	0
21	10	2	5.5	5.5	5.5	4	3.9	1.5	5.5	2	1.5	1.5	1.5	2	0.5	0.5	0.5
13	7	1	0	0	0	4	3.8	1.5	4	2	1	1	1	1	0	0	0
22	10	3	2.7	2	3.5	2	1.5	1.5	1.5	1	0	0	0	2	0.5	0.5	0.5
21.1	11.8	1.5	1.3	1.1	1.4	3.1	3.4	2.2	4.0	1.8	1.2	1.1	1.4	1.6	0.6	0.4	0.7

TABLE 2 | The results of Frin, Cass, Lnetwork and BIMLR on artificial datasets with clusters |*C*| and taxa |*X*|.

Data		Frin				Cass				Lnetwork				BIMLR			
|*C*|	|*X*|	t	k	r	c	t	k	r	c	t	k	r	c	t	k	r	c
86	37	14s	4	9	12	3s	3	8	27	4s	3	8	11	8s	3	8	23
38	20	33s	5	7	11	2s	4	6	25	25s	4	6	15	2s	4	6	25
43	22	1s	3	5	3	1s	2	4	12	1s	3	5	3	1s	3	5	11
72	27	32s	5	7	19	15s	5	7	43	3s	5	7	19	4s	5	7	29
52	22	27s	4	8	12	17s	4	7	33	3s	4	8	15	6s	4	8	15
79	27	3m54s	8	10	80	7m21s	6	8	89	47s	6	8	44	2m40s	8	10	52
38	16	1m44s	6	8	28	15s	5	7	50	4m22s	7	9	36	13s	6	8	25
41	16	2s	4	5	6	1s	4	5	29	1s	4	5	4	1s	4	5	7
12	8	1s	2	2	0	1s	2	2	2	1s	2	2	0	1s	2	2	0
45	20	1m51s	6	7	34	4h4m	6	7	66	35s	6	7	28	17s	6	7	47
22	11	44s	2	3	1	1s	2	3	5	1s	2	3	1	1s	2	3	4
17	10	1s	3	3	4	1s	3	3	8	1s	3	3	4	1s	3	3	7
46	16	6m8s	6	8	10	23s	5	7	34	7s	6	8	15	12s	6	8	22
22	11	41s	4	4	14	2s	4	4	23	3s	4	4	13	2s	5	5	21
22	10	54s	4	4	10	2s	4	4	21	6s	4	4	12	2s	5	5	19
42.3	18.2	1m2s	4.4	6	16	16m51s	3.9	5.5	31	24.9s	4.2	5.8	14.7	15.4s	4.4	6	20.5

TABLE 3 | The results of Frin, Cass, Lnetwork and BIMLR on practical datasets with clusters |*C*| and taxa |*X*|.

Data		Frin				Cass				Lnetwork				BIMLR			
|*C*|	|*X*|	t	k	r	c	t	k	r	c	t	k	r	c	t	k	r	c
14	4	1s	3	3	0	1s	3	3	0	1s	3	3	0	1s	3	3	0
30	5	1s	4	4	0	2s	4	4	0	2s	4	4	0	1s	4	4	0
62	6	6s	5	5	0	11s	5	5	0	6s	5	5	0	7s	5	5	0
42	10	1s	4	4	8	5s	4	4	34	1s	4	4	8	1s	4	4	8
39	11	23s	6	6	10	21s	5	5	7	13s	5	5	8	3s	5	5	8
61	11	23s	5	5	11	1m26s	5	5	48	5s	5	5	11	1s	5	5	11
75	30	1s	2	2	19	5s	2	2	122	1s	2	2	19	1s	2	2	19
180	51	8s	2	2	0	40s	2	2	0	4s	2	2	0	1s	2	2	0
70	56	1s	1	4	0	1s	1	4	0	1s	1	4	0	2s	1	4	0
270	76	1m7s	2	2	0	6m22s	2	2	0	12s	2	2	0	24s	2	2	0
404	122	4m1s	2	2	0	1h44m	2	2	0	27s	2	2	0	27s	2	2	0
113.4	34.7	43.7s	3.3	3.5	4.4	10m18s	3.2	3.5	10	6.6s	3.4	3.6	8.5	7.1s	3.2	3.5	4.2

number and the redundant cluster number, of four methods on artificial and practical data (see **Table 2** and **3**).

We get all permutations of input order for each data, and then construct networks for each permutation. Since the running time of the experiment is factorial, we choose small-scale data as the input. In order to measure the influence of input data order, we record the number of different resulting networks and compute the tripartition distance between them. We use the tripartition distances to measure the dissimilarity between the networks. The experimental result is shown in **Table 1**. Each dataset consists of cluster number |*C*| and taxa number |*X*|. The table records the number of different networks (n) and mean (mean), minimum (min), maximum (max) values of the tripartition distance, and the last row is the average of the corresponding columns. **Table 1** shows that the number of different networks constructed by Frin is less than other three methods for most data, and the tripartition distance between them is also smaller, especially compared with Cass algorithm. Hence, Frin constructs more consistent networks when the input data orders are different.

We test the complexity of the networks constructed by Frin, Cass, Lnetwork, and BIMLR, including the network level (k), the reticulation number (r) and the redundant cluster number (c), and as well as the running time (t) of those methods in h/m/s. The following tables show the results of experiment on artificial and practical data with the cluster number |*C*| and the taxa number |*X*|. The last row of the tables is the average of the corresponding columns. **Table 2** compares Frin with other three methods in several artificial datasets. It shows that Frin consumes less time for the same input data compared with Cass, and Frin has significantly fewer redundant clusters than Cass and BIMLR. **Table 3** compares the four methods in several practical datasets. It shows that the average reticulation number of Frin is slightly larger than the other methods, but it has fewer redundant clusters than Cass and Lnetwork in most cases. Thus, the network constructed by Frin is simpler than that constructed by other methods in the aspect of redundant clusters, and the execution time of Frin has also been greatly reduced compare with Cass, although it takes longer than the other two methods.

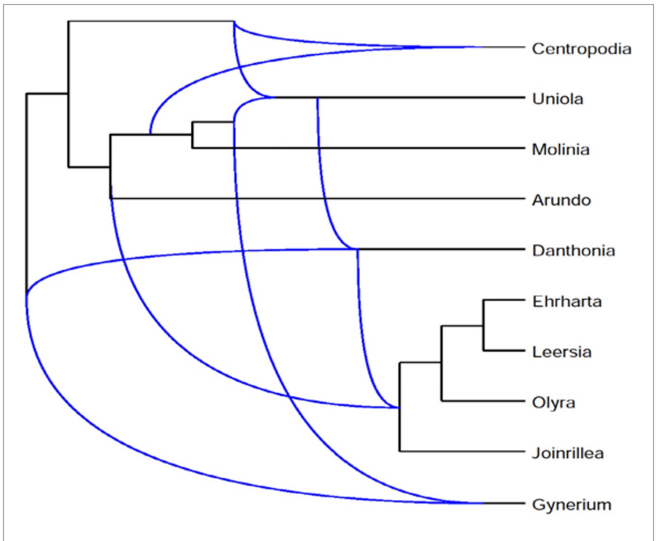


FIGURE 6 | Frin constructs a level-5 network with r = 5, c = 31 for the three gene trees of the *Poaceae* datasets.

We describe the application of Frin to the *Poaceae* dataset and also compare it with other programs. The dataset consists of three phylogenetic trees of the *Poaceae* family, which are based on sequences data for three difference gene loci, petD, ndhB, and rpl2. The gene sequences are downloaded from NCBI database. We do sequence alignment on the obtained sequence using Clustalx, and construct a phylogenetic tree using Phylip. Frin constructs a level-5 network with 10 taxa, 5 reticulations and 31 redundant clusters for the three gene trees of *poaceae* datasets. The resulting network is shown in **Figure 6** using Dendroscope3 (Huson et al., 2007; Vaughan, 2017). For the same input, BIMLR constructs a level-5 network with r = 5, c = 33 and Lnetwork constructs a level-5 network with r = 5, c = 37; while Cass algorithm cannot construct the network in a day. The result shows that the network constructed by Frin is the simplest. It illustrates that the network constructed by Frin which can describe real evolutionary history better than the other methods.

CONCLUSION

In this paper, we propose an efficient method called Frin to construct phylogenetic networks. In the process of construction, Frin considers the two factors that affect the compatibility of a cluster set, which are the frequency and incompatibility degree of taxa, respectively. Frin can construct several different networks, and select the simplest network from them as the resulting network. The experimental results show that Frin is an improved method. First, Frin can construct less different networks when the input data order is different than the other methods. Second, the networks constructed by Frin have less the number of redundant clusters than the other methods in the case of the level and the reticulation number of the networks not are increasing. Both facts indicate that Frin can better describe the biological evolutionary history.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in Github (https://github.com/wangjuanimu/Frin). The artificial and the practical datasets can be accessed from the Cass website (https://sites.google.com/site/cassalgorithm/data231 sets).

AUTHOR CONTRIBUTIONS

YH proposed the method and designed the experiments. YH and JW wrote the paper.

FUNDING

The work was supported by National Natural Science Foundation of China under Grant No. 61661040.

REFERENCES

Albrecht, B., Scornavacca, C., and Cenci, A. (2012). Fast computation of minimum hybridization networks. *Bioinformatics* 28 (2), 191–197. doi: 10.1093/bioinformatics/btr618

Albrecht, B. (2015). Computing all hybridization networks for multiple binary phylogenetic input trees. *BMC Bioinf.* 16 (1), 1–15. doi: 10.1186/s12859-015-0660-7

Bordewich, M., Linz, S., and John, K. S. (2007). A reduction algorithm for computing the hybridization number of two trees. *Evol. Bioinf.* 3, 117693430700300. doi: 10.1177/117693430700300017

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* 284 (5423), 2124–2128. doi: 10.1126/science.284.54232124

Francis, A., Huber, K. T., and Moulton, V. (2018). Tree-based unrooted phylogenetic networks. *Bull. Math. Biol.* 80 (2), 404–416. doi: 10.1007/s11538-017-0381-3

Gambette, P., Huber, K. T., and Kelk, S. (2017). On the challenge of reconstructing level-1 phylogenetic networks from triplets and clusters. *J. Math. Biol.* 74 (7), 1729–1751. doi: 10.1007/s00285-016-1068-3

Gunawan, A. D. M., Lu, B., and Zhang, L. (2016). A program for verification of phylogenetic network models. *Bioinformatics* 32 (17), i503–i510. doi: 10.1093/bioinformatics/btw467

Gunawan, A. D. M., Lu, B., and Zhang, L. (2018). Fast methods for solving the cluster containment problem for phylogenetic networks. *Bioinformatics* 34 (17), i680–i686. doi: 10.1093/bioinformatics/bty594

Gusfield, D., Bansal, V., and Bafna, V. (2007a). A decomposition theory for phylogenetic networks and incompatible characters. *J. Comput. Biol.* 14 (10), 1247–1272. doi: 10.1089/cmb.20060137

Gusfield, D., Hickerson, D., and Eddhu, S. (2007b). An efficiently computed lower bound on the number of recombination in phylogenetic networks: theory and empirical study. *Discrete Appl. Math.* 155 (6-7), 806–830. doi: 10.1016/j.dam.2005.05.044

Huber, K. T., van Iersel, L., and Moulton, V. (2017). Reconstructing phylogenetic level-1 networks from nondense binet and trinet sets. *Algorithm.* 77 (1), 173–200. doi: 10.1007/s00453-015-0069-8

Huson, D. H., and Rupp, R. (2008). Summarizing multiple gene trees using cluster networks. *Int. Workshop Algo. Bioinf.* 5251, 296–305. doi: 978-3-540-87361-7_25

Huson, D. H., and Scornavacca, C. (2011). A survey of combinatorial methods for phylogenetic networks. *Genome Biol. Evol.* 3, 23–35. doi: 10.1093/gbe/evq077

Huson, D. H., Rupp, R., Berry, V., Gambette, P., and Paul, C. (2007). Computing galled networks from real data. *Bioinformatics* 25 (12), i85–i93. doi: 10.1093/bioinformatics/btp217

Huson, D. H., Richter, D. C., and Rausch C. (2007). Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinf.* 8 (1), 460–460. doi: 10.1186/1471-2105-8-460

Iersel, L. V., Keijsper, J., and Kelk, S. (2009). Constructing Level-2 phylogenetic networks from triplets. *EEE/ACM Trans. Comput. Biol. Bioinform.* 6, 667–681. doi: 10.1109/TCBB.2009.22

Jansson, J., and Sung, W. K. (2006). Algorithms for combining rooted triplets into a galled phylogenetic network. *SIAM J. Comput.* 35, 1098–1121. doi: 10.1137/S0097539704446529

Kelk, S., and Scornavacca, C. (2014). Constructing minimal phylogenetic networks from softwired clusters is fixed parameter tractable. *Algorithm.* 68 (4), 886–915. doi: 10.1007/s00453-012-9708-5

Linz, S., and Semple, C. (2009). Hybridization in Nonbinary Trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 6 (1), 30–45. doi: 10.1109/TCBB.2008.86

Makarenkov, V., Kevorkov, D., and Legendre, P. (2006). Phylogenetic network construction approaches. *Appl. Mycol. Biotechnol.* 6 (06), 61–97. doi: 10.1016/S1874-5334(06)80006-7

Mirzaei S., and Wu, Y. (2016). Fast construction of near parsimonious hybridization networks for multiple phylogenetic trees. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 13 (3), 1–1. doi: 10.1109/TCBB.2015.2462336

Nakhleh, L. (2011). Evolutionary Phylogenetic Networks: Models and Issues.The Problem Solving Handbook for Computational Biology and Bioinformatics. Springer, pp.125-158.doi: 10.1007/978-0-387-09760-2_7

Van Iersel, L., Kelk, S., Rupp, R., and Huson, D. (2010). Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26 (12), i124–i131. doi: 10.1093/bioinformatics/btq202

Van Iersel, L., Kelk, S., and Stamoulis, G. (2017). On unrooted and root-uncertain variants of several well-known phylogenetic network problems. *Algorithm* 80, 2993–3022. doi: 10.1007/s00453-017-0366-5

Vaughan, T. G. (2017). IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics* 33 (15), 2392–2394. doi: 10.1093/bioinformatics/btx155

Wang, J., Guo, M., Liu, X., Liu, Y., Wang, C., Xing, L., et al. (2013a). LNETWORK: an efficient and effective method for constructing phylogenetic networks. *Bioinf.* 29 (18), 2269–2276. doi: 10.1093/bioinformatics/btt378

Wang, J., Guo, M., Xing, L., Che, K., Liu, X., and Wang, C. (2013b). BIMLR: a method for constructing rooted phylogenetic networks from rooted phylogenetic trees. *Gene* 527 (1), 344–351. doi: 10.1016/j.gene.2013.06.036

Wu, Y. (2010). Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics* 26 (12), i140–i148. doi: 10.1093/bioinformatics/btq198

Yu, Y., and Nakhleh, L. (2015). A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16 (10), S10. doi: 10.1186/1471-2164-16-S10-S10

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.

ORIGINAL RESEARCH
published: 09 January 2020
doi: 10.3389/fgene.2019.01310

Identification of Prognostic Dosage-Sensitive Genes in Colorectal Cancer Based on Multi-Omics

*Zhiqiang Chang, Xiuxiu Miao and Wenyuan Zhao**

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

Several studies have already identified the prognostic markers in colorectal cancer (CRC) based on somatic copy number alteration (SCNA). However, very little information is available regarding their value as a prognostic marker. Gene dosage effect is one important mechanism of copy number and dosage-sensitive genes are more likely to behave like driver genes. In this work, we propose a new pipeline to identify the dosage-sensitive prognostic genes in CRC. The RNAseq data, the somatic copy number of CRC from TCGA were assayed to screen out the SCNAs. Wilcoxon rank-sum test was used to identify the differentially expressed genes in alteration samples with |SCNA| > 0.3. Cox-regression was used to find the candidate prognostic genes. An iterative algorithm was built to identify the stable prognostic genes. Finally, the Pearson correlation coefficient was calculated between gene expression and SCNA as the dosage effect score. The cell line data from CCLE was used to test the consistency of the dosage effect. The differential co-expression network was built to discover their function in CRC. A total of six amplified genes (NDUFB4, WDR5B, IQCB1, KPNA1, GTF2E1, and SEC22A) were found to be associated with poor prognosis. They demonstrate a stable prognostic classification in more than 50% threshold of SCNA. The average dosage effect score was 0.5918 ± 0.066, 0.5978 ± 0.082 in TCGA and CCLE, respectively. They also show great stability in different data sets. In the differential co-expression network, these six genes have the top degree and are connected to the driver and tumor suppressor genes. Function enrichment analysis revealed that gene NDUFB4 and GTF2E1 affect cancer-related functions such as transmembrane transport and transformation factors. In conclusion, the pipeline for identifying the prognostic dosage-sensitive genes in CRC was proved to be stable and reliable.

Keywords: colorectal cancer, somatic copy number alteration, survival analysis, gene dosage effect, differential co-expression

OPEN ACCESS

Edited by:
Lei Deng,
Central South University,
China

Reviewed by:
Li Hongdong,
Gannan Medical University,
China
Xiaowei Chen,
Institute of Biophysics (CAS),
China

****Correspondence:***
Wenyuan Zhao
zhaowenyuan@ems.hrbmu.edu.cn

Specialty section:
This article was submitted to
Statistical Genetics and
Methodology,
a section of the journal
Frontiers in Genetics

Received: *22 September 2019*
Accepted: *27 November 2019*
Published: *09 January 2020*

Citation:
Chang Z, Miao X and Zhao W (2020)
Identification of Prognostic Dosage-
Sensitive Genes in Colorectal
Cancer Based on Multi-Omics.
Front. Genet. 10:1310.
doi: 10.3389/fgene.2019.01310

INTRODUCTION

Colorectal cancer (CRC), is the 3rd leading cause of cancer-associated deaths in the world (Siegel et al., 2019). Studies have shown that somatic copy number alteration (SCNA) is one of the most common and important structural mutations in CRC (Li et al., 2017; Oliveira et al., 2018). SCNA genes are usually considered as the driver gene for cancer development and an important factor for the progression of CRC (Wang et al., 2009; Rosenberg et al., 2018; Lee et al., 2019).

In addition to this few SCNA genes are also being considered as prognostic markers for CRC patients (Roy et al., 2016; Sefrioui et al., 2017). Previous research has shown that a high copy number of mitochondrial DNA can help in identifying the poor prognosis associated with advanced-stage CRC patients (Wang et al., 2016). However, the reason for this specific attribute is still unknown. SCNAs are generated by chromosomal rearrangement. Another important mechanism of SCNA influencing cancer progression is through the gene dosage effect (Harel and Lupski, 2018; Salpietro et al., 2018). For a gene in the region of SCNA, if its expression increases with amplification of the copy number and *vice versa*, this gene would be defined as dosage-sensitive gene. With respect to the unstable and complex nature of expression regulation, the DNA copy number is relatively more stable. Therefore, the copy number of dose-sensitive genes is more likely to be used as a driver gene in cancer. Some of the dosage-sensitive genes (DSGs) such as CD274/PD-L1 gene amplification (Lee et al., 2018b), fibroblast growth factor 1 amplification (Bae et al., 2019), RING-Finger Protein 6 amplification (Steinman et al., 1979), have been shown to be associated with poor prognosis, suggesting DSGs can also be considered as prognostic markers.

The amount of SCNA can be considered as one important indicator of cancer progression. Cancerous tissue may contain both tumor and non-tumor cells, and the copy number of DNA in all cells can be measured during detection. The copy number value obtained from the whole tissue sample with respect to the control sequence reflects the frequency of copy number alteration in the whole sample. This value is often in parts. However, identifying a threshold value of SCNA to be considered as pathogenic or mutant needs a thorough investigation. Jianxin Shi et al. identified significant CNVs using the FASST2 algorithm and selected the number of probes per fragment >5 and log2ratio greater than 0.3 as amplification gene (Shi et al., 2016). Villela et al. also used 0.3 as the SCNA threshold (Kostolansky et al., 1986; Villela et al., 2018). In addition, the copy number amplification or deletion of 0.5 (*i.e.* half amplification or deletion) is pathogenic (Birchler et al., 2001; Birchler and Veitia, 2012). These results suggest that different threshold values should be used as a measure of SCNA.

Due to the importance of DSGs and the fact that SCNA could be a prognostic marker of CRC, we hypothesize that the dosage-sensitive prognostic genes should also affect CRC progression. TCGA is a milestone project of cancer genome covering CNV, RNA-seq data, and patient-specific data of CRC. It can provide a possibility for relatively large-scale excavation of prognostic genes of CRC. In this paper, we have established a pipeline for screening prognosis sensitive genes in CRC, organically identified stable prognostic markers with dosage sensitivity of copy number in CRC, and verified their dosage sensitivity by cell line data. This analysis can help to further enhance our understanding of the value of the prognostic gene of SCNA and can lay a foundation for further analysis.

MATERIALS AND METHODS

Datasets and Processing

The data of CNA, RNA-seq data, and clinical data of CRC were downloaded from the TCGA database. By mapping the copy number probe across the reference genome of hg38, the SCNA at gene level was calculated using Gistic2 software (Mermel et al., 2011). The value of SCNA represents the portability of copy number alteration and the *q*-value for the genes in aberrant regions. The *q*-value > 0.1 and *q*-value < –0.1 were considered as copy number amplified and deleted, respectively. For each gene, the samples with SCNA value > = x (x represents the threshold of SCNA with a value >0) were identified as copy number amplification samples (CNAS), the samples with SCNA < = –x were identified as copy number deleted samples (CNDS), and the samples with | SCNA | < x were identified as copy number non-altered samples (CNNS). The location information of chromosomes was obtained from the HGNC database (Braschi et al., 2019). RNAseq FPKM data was downloaded from University of California Santa Cruz (UCSC, http://genome.ucsc.edu/), and more than 80% of genes with 0 value were filtered out. The test data-set was collected from the Cancer Cell Line Encyclopedia (CCLE; http://www.broadinstitute.org/ccle/home).

Filtering of Prognosis-Sensitive SCNA Genes

PSGs of SCNA were screened in five steps as described below:

Step 1: Set x (x > 0) as the threshold for SCNA, then the samples of CRC were classified into three groups, somatic copy number amplification samples (CNAS), somatic copy number deletions samples (CNDS), and somatic copy number non-alteration samples (CNNS). The number of CNAS or CNDS was more than or equal to 10. Wilcoxon rank-sum test was performed to identify differentially expressed genes between CNAS and CNNS and between CNDS and CNNS. The *p*-value was corrected by the Benjamini-Hochberg method. As there were very small differences in gene expression between SCNA and CNNS samples their false discovery rate (FDR) < 0.1 and *p* < 0.01, fold change >1.2 were considered as differential expression.

Step 2: In order to further screen the candidate genes on the basis of Step 1. We identified genes with expression up-regulation (*p*-value < 0.01 and FC > 1.2) and copy number amplification (SCNA > x) in CNAS, and the genes with expression down-regulation (*p*-value < 0.01 and FC < 1/1.2) and copy number deletion (SCNA < –x) in CNDS as candidates for the dosage-sensitive gene.

Step 3: The data of SCNA and survival time of all the samples for each abnormal candidate gene was analyzed by Cox regression and the genes with *p*-value < 0.05 were identified as candidate PSG.

Step 4: In order to further screen stable SCNA-PSGs, the SCNA threshold x was raised from 0.1 to 0.5 with 0.02 steps, and the cancer samples were divided into CNAS, CNDS, and CNNS. For each threshold of SCNA, the log-Rank test was used to assess the significance of overall survival times in CNAS vs. CNNS and CNDS vs. CNNS groups. The abnormal driver genes with the number more than 50% number of the thresholds were selected as a stable PSG.

Step 5: In order to further screen dosage-sensitive genes from stable PSGs in different SCNA threshold, the prognostic sensitive abnormal genes of DSGs were selected. Linear regression was applied to assess the dosage-sensitivity. The R-value represents the dosage-effect score. The genes with the *p*-value < 0.05 and *R* > = 0.3 were considered as prognostic dosage-sensitive genes (PDSGs).

Verification of DSGs in Cell Lines

In order to verify the stability of the dosage-sensitivity of PDSGs, the correlation coefficients between gene expression and copy number alteration were calculated with the RNA-seq of CRC and CNA at gene level downloaded from the CCLE database. These values were compared with the findings obtained from TCGA.

Building the Differential Co-Expression Network

In order to further identify the genes affected by PDSGs, Pearson correlation coefficients of these six PDSGs and other genes was calculated as co-expression values in CNAS or CNDS, CNNS. Gene pairs with correlation coefficients higher than 0.5 in one group and less than 0.1 in another group were screened as differentially co-expressing gene pairs. Network visualization tools were executed using Cytoscape (Shannon et al., 2003).

Analysis

All the analysis was performed in the R computing environment. Survival curves were estimated using the Kaplan-Meier method. Gene function enrichment was performed using the Cluster Profiler package (Yu et al., 2012).

RESULTS

PDSGs in CRC

A total of 448 CRC samples with SCNA and RNA-seq data were downloaded from The Cancer Genome Atlas (TCGA). The samples were screened for survival information. There were 22,752 genes, of these 17,442 were protein-coding and 14,688 were differentially expressed.

After applying FDR < 0.1 and FC > 1.2, 6,814 genes had up-regulated expression in CNAS. Twenty-five genes had a down-regulated expression in CNDS. Cox regression analysis was applied to calculate the correlation between SCNA and survival time. A total of 215 prognosis-sensitive genes (PSGs) significantly related to SCNA were obtained, of these 214 were amplified and one was deleted. Next, the 21 SCNA threshold value was raised from 0.1 to 0.5 at a step of 0.02. For each threshold, the samples were classified into CNNS, CNAS, CNDS group and logRank test between CNNS and CNAS, CNDS and CNNS was performed. As shown in **Figure 1**, 73.02% of genes didn't show any significant classification with any threshold. A total of 15 genes showed stable prognosis classification of patients in more than 10 threshold values, suggesting these 15 genes can be considered as stable markers for prognosis classification in CRC.

After further screening stable PSGs which are highly affected by copy number dosage effect, the Pearson correlation coefficient between copy number and corresponding expression value (FPKM) of these 15 genes was calculated. Finally, six genes (NDUFB4, WDR5B, IQCB1, KPNA1, and SEC22A) which are stable PSGs (**Figure 2**) were identified. The average dosage effect score was 0.5918 and the variance was 0.066.

Kaplan-Meier survival curve analysis revealed six (6) PDSGs with similar results in a different threshold of SCNA. In the 0.1 SCNA threshold value, genes GTF2E1, NDUFB4, IQCB1, KPNA 1 and WDR5B had a significant classification effect (**Figures 3A–C**). At the 0.3 threshold value of SCNA, all six genes had a similar and significant classification effect (**Figure 3D**). At the 0.5 threshold value, five genes (GTF2E1, NDUFB4, IQCB1, KPNA1, WDR5B) had similar classification effect (**Figures 3E, F**). Although the statistical significance of the two classifications (*p*-value = 0.087199 and *p*-value = 0.12643) in 0.5 SCNA threshold was not significant, their classification curves were distinctly separated. The non-significance can be primarily attributed to the very small number of samples with SCNA threshold >0.5.

Testing Dosage Effect of PDSGs in CCLE

In order to verify if the copy number of six PDSGs is dosage-sensitive in the data from cell lines with 53 cell line samples, the

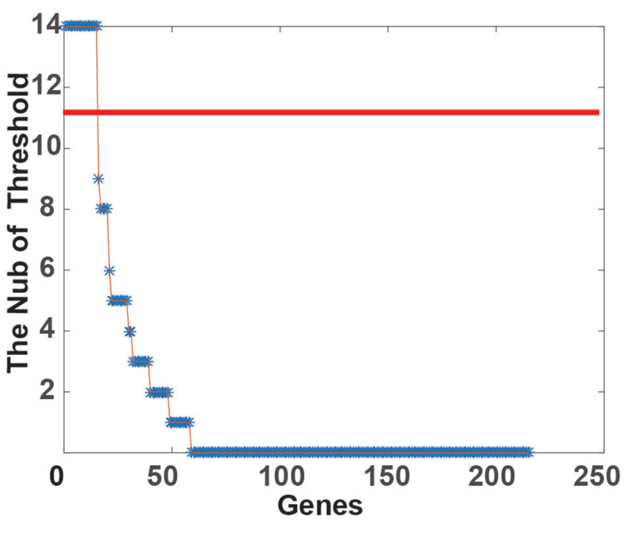


FIGURE 1 | Classification stability of gene prognosis. For each threshold of somatic copy number alteration (SCNA) (from 0.1 to 0.5, at 0.02 step), the *p*-value was calculated by the log-rank test in corresponding alteration and CNNS samples. The Number of Threshold will increase if the *p*-value < 0.05.

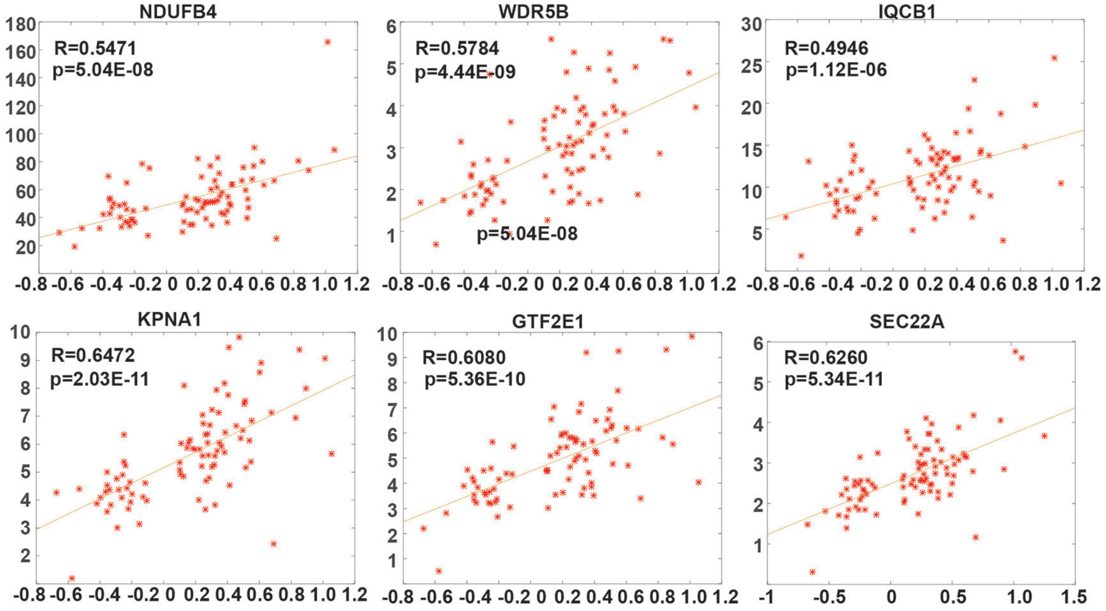


FIGURE 2 | The dosage sensitivity of six prognostic dosage-sensitive genes (PDSGs). The X-axis represents the somatic copy number alteration (SCNA) value and Y-axis represents the FPKM of genes.

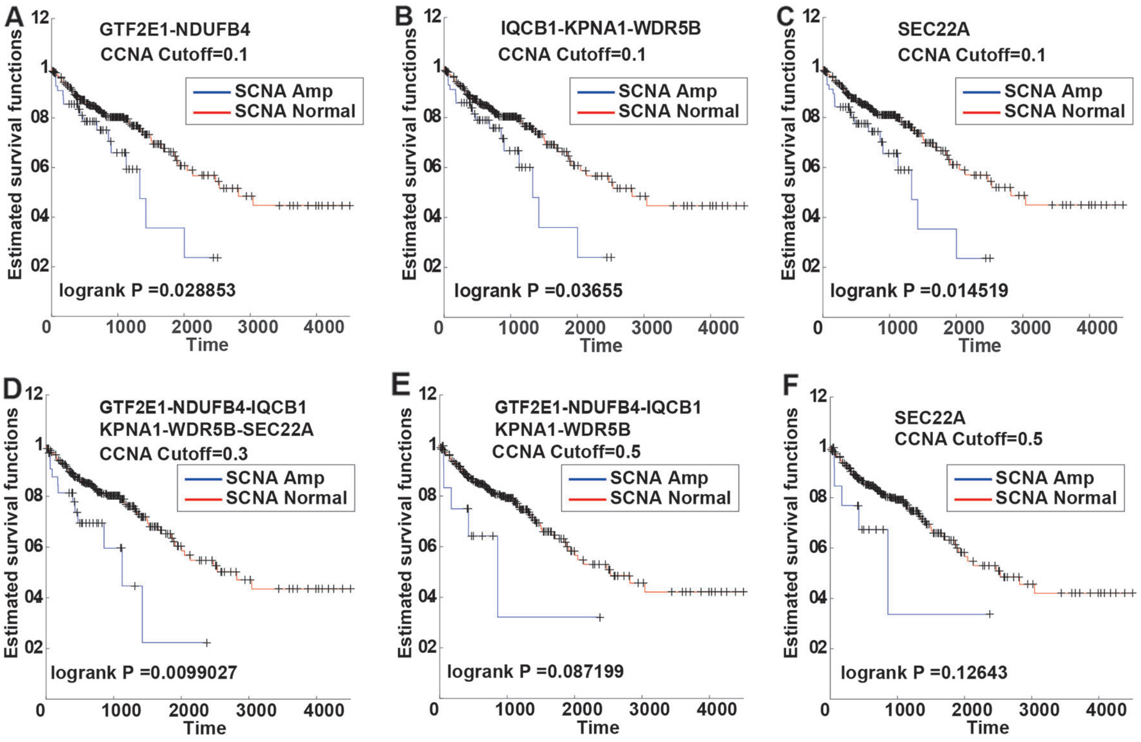


FIGURE 3 | The Kaplan-Meier curves of six PSDGs for samples in CNAS and CNNS. **(A–C)** with the somatic copy number alteration (SCNA) threshold 0.1, gene GTF2E1 and NDUFB4 had similar prognostic classification efficacy. **(D)** with the SCNA threshold 0.3, all six PSDGs have similar efficacy. **(E, F)** with the SCNA threshold 0.5, although the *p*-value was > 0.05, the two survival curves still separated from each other.

dosage effect score of these six PDSGs in CRC from CCLE was calculated. An average score of 0.5978 and variance was 0.082 consistent with the result from TCGA was obtained (**Figure 4A**). The Pearson correlation coefficient was 1, suggesting that the gene dosage effect is stable in CRC different data.

Six PDSGs Are Co-Alteration in CRC

Further to test similarity between survival curves of these six PDSGs, we mapped them to chromosomes and found that they all are located on 3q13.33–3q21.1. By computing the correlation coefficients between the copy number of two pairs of genes an average value of 0.9967 (**Figure 4B**) was observed. This indicates that these six PDSGs are highly consistent with each other during alteration.

Research have shown that heterogeneity of copy number alterations exists in ongoing unstable chromosome in COAD (Bolhaqueiro et al., 2019). There are some chromosomes fragile sites in genome, the genes in fragile sites may break when they fell external pressure. In order to determine the presence of breakpoints in the region near to 6PDSGs, they were mapped on the database of human chromosomes fragile sites (HumCFS, http://webs.iiitd.edu.in/raghava/humcfs/). As a result, FRA3D (3q25.32) and FRA3C were found to be near to six PDSGs. Correlation analysis of SCNA in six PDSGs and the genes in FRA3D and FRA3C was performed. Gene RSRC1 (R = 0.82), MLF1(R = 0.82) in FRA4D, and LPP (R = 0.80) in FRA3C had lowest relationship with PDSGs. Thus we infer that the breakpoints in fragile site may explain the reason for the nearby region and a similar SCNA value.

Building and Analysis of Differential Co-Expression Network With PDSGs

In order to further explore if these six PDSGs can also affect the expression of other genes in CRC, we screened genes with (R) > 0.5 and (R) < 0.1 in a different class of samples by calculating the differences of gene co-expression between CNAS and CNNS. A total of 234 co-expressed gene pairs were observed and 215 genes (**Figure 5A**) involved in differential co-expression networks were identified. The whole network constitutes a component suggesting that CRC is a disease involving multiple genes. Among these 194 gene pairs were co-expressed in alteration samples (R > 0.5), but not co-expressed in non-alteration samples (R < 0.1), while the other 40 pairs behaved in a reverse manner. In the network, gene NDUFB4, SEC22A had the highest degree (109 and 45 respectively) consisting of 15 co-linked genes. The genes CAPN14 and CMPK2 were affected by three PDSGs (NDUFB4, SEC22A, and IQCB1). This suggests that PDSGs are closely linked and interact with each other.

Each PDSG in the network was related to at least 13 genes and 22 genes were associated with more than one PDSG. We also found that several PDSGs-associated genes were also COAD-related. The co-expression of GTF2E1-WNT8B was activated in CNAS(R = 0.59). WNT8B one member of the WNT signal was differentially expressed in COAD (Neumann et al., 2014). In addition to this, after mapping the PDSG-related genes to the driver gene list from DriverDB (Liu et al., 2019), three genes (C8orf33, LAPTM4B, PTP4A3) were found (**Figure 5B**), and they all were co-expressed with gene NDUFB4 in CNAS but not in CNNS. Mapping of PDSG-related genes on the tumor suppressor database (TSGene, http://bioinfo.mc.vanderbilt.edu/TSGene/) revealed 16 TSGs (**Figure 5A**, Triangle). Among these, gene DCDC2, ISG15, RARRES3 can affect more than one PDSG. Gene RARRES3 has been shown to be mutated, differentially expressed and also inhibits metastasis in COAD (Lee et al., 2018a). ISG15 is shown to have significant differential expression in COAD (Yu et al., 2019; Zamanian-Azodi and Rezaei-Tavirani, 2019).

Further to explore the possible functions of these six PDSGs, linked genes were extracted and gene ontology function enrichment analysis was performed. Genes linked to gene NDUFB4 (**Figure 5C**) were mainly enriched in functions such as “transmembrane receptor,” “transmembrane transport,” “peptide receptor,” “G protein-coupled receptor,” “transforming growth factor.” Genes linked to gene GTF2E1

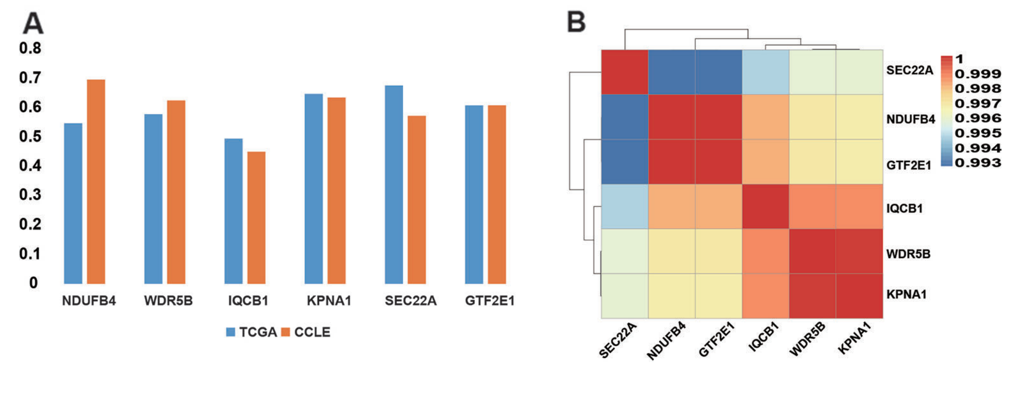


FIGURE 4 | The dosage-sensitive and the correlation scores of somatic copy number alteration (SCNA) of prognostic dosage-sensitive genes (PDSGs). **(A)** The correlation coefficient of SCNA and gene expression. Both results suggest strong concordance. **(B)** The heatmap of the SCNA of PDSG. All these six PDSGs show high co-alteration in colorectal cancer (CRC).

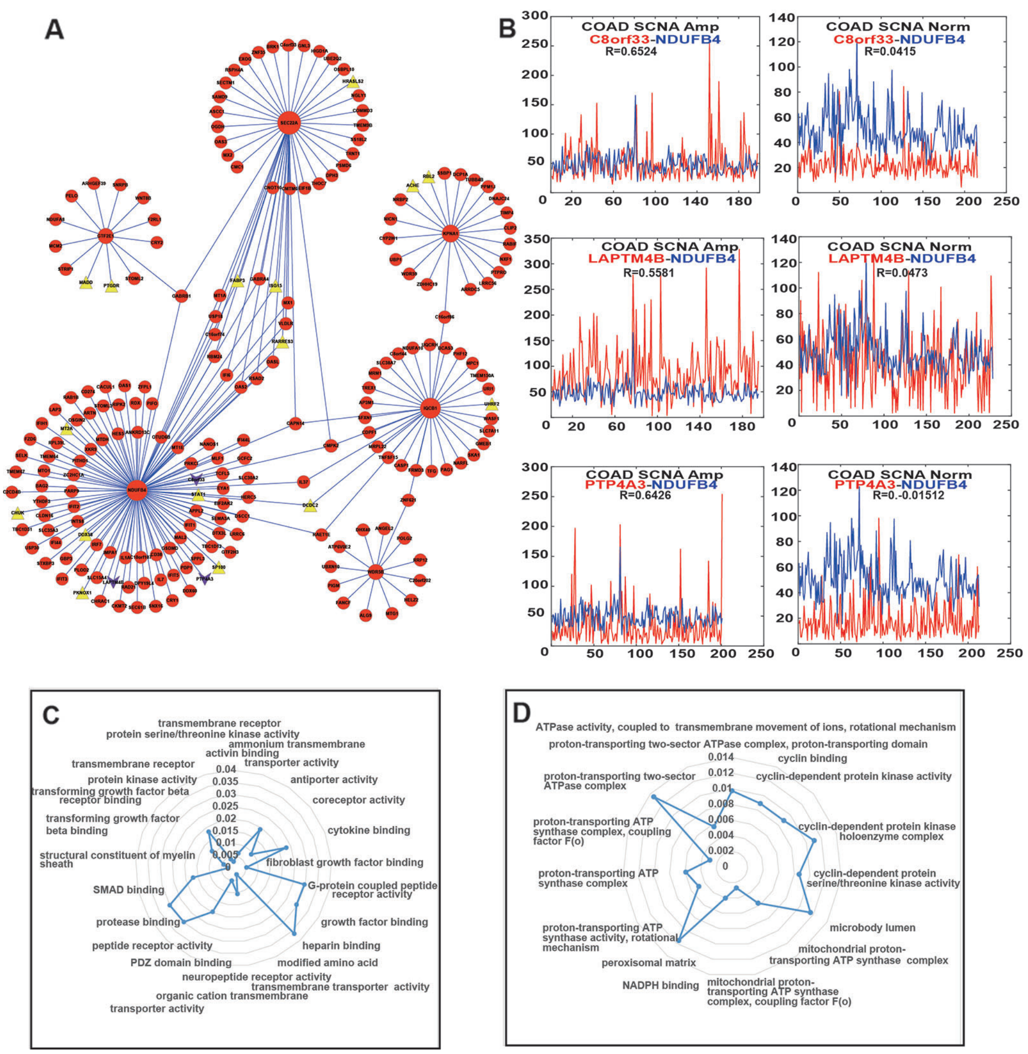


FIGURE 5 | Differential co-expression network and function of enrichment of prognostic dosage-sensitive genes (PDSGs). **(A)** Differential co-expression networks, Triangle represent tumor suppressor genes, lower triangular represent driver gene. Six PDSGs (NDUFB4, WDR5B, IQCB1, KPNA1, GTF2E1, and SEC22A) have the top degree. The edge represents co-expression of the adjacent genes above 0.5 in one group and below 0.1 in another group. **(B)**. The co-expression curve of gene **(C)** Normal and abstained function of gene NDUFB4 using Cluster Profiler R package **(D)** Normal and abstained function of gene GTF2E1 using Cluster Profiler R package.

were enriched (**Figure 5D**) in functions such as "cyclin-dependent protease," "ATP synthase transport proton-related functions." Previous studies have shown that transforming growth factor can also promote tumorigenesis (De Miranda et al., 2015; Yu et al., 2018; Kim et al., 2019). G-protein-coupled receptors (GPCRs) are a member of the largest cell surface molecule family involved in signal transduction and are considered as the key molecule in the growth and metastasis of tumors (Wielenga et al., 2015; Insel et al., 2018). Malignant cells often hijack the normal physiological functions of GPCRs to survive, proliferate independently, escape the epidemic system, increase blood supply, invading the surrounding tissues and spread to other organs.

DISCUSSION

In this manuscript, a series of screening methods were established to identify PDSGs in CRC. A total of six PDSGs identified in the present study not only have the robustness to different SCNA threshold in prognostic classification but also have the same dosage effect in CRC cell lines. This indicates that our screening pipeline is suitable, reasonable, and effective. The amplification of the copy number of these six PDSGs can lead to poor prognosis, indicating that the SCNA of genes could serve as an important prognostic marker in CRC.

In addition to the stable results, these PDSGs have been shown to be associated with CRC. Gene NDUFB4 encodes a non-catalytic subunit of the NADH. The NADH dehydrogenase complex I is overexpressed in incipient metastatic murine CRC cells (Marquez et al., 2019). Mutations in mitochondrial NADH dehydrogenase subunit 1 (mtND1) gene were found in CRC (Yusnita et al., 2010). WDR5B encodes a protein containing several WD40 repeats, and it is reported as an important target of miR-31. The knockout of microRNA-31 promotes the development of colitis-associated cancer (Liu et al., 2017). The protein encoded by gene SEC22A belongs to the member of the SEC22 family of vesicle trafficking proteins. It has a similarity to rat SEC22 and may act in the early stages of the secretory pathway, which is related to CRC (Jilling and Kirk, 1996; Baron et al., 2010).

Compared with the gene expression the DNA copy number often occurs in arm-level, *i.e.* the same segment tends to have the same copy number alteration (Roy et al., 2016; Xu et al., 2018). The results of this study not only support this opinion but also suggest that even in the same fragment the correlation between different samples is not always 1. There are some differences indicating that somatic alterations have some heterogeneity, and demonstrates the diversity of alteration in CRC. In addition, although chromosomes play a role through the dosage effect to some extent they may be affected by the regulation of gene expression. Six of the 15 genes obtained in this paper have a strong dosage effect suggesting that not all gene copy number amplification will lead to up-regulation of expression. The contribution is a combination of copy number and dosage effect. In future, if targeted drugs or therapies can be developed to reduce the copy number of these six PDSGs, patients with amplified copies of these six genes may receive a precise treatment. This is also an important starting point and foothold of this topic.

The ratio of amplified and non-amplified samples of CPCDGs gene is 1:11, which indicates that these prognostic markers are valuable only for patients with high SCNA. Therefore, SCNA can be an important part of precise medical treatment. Due to computational limitations, the minimum alteration sample selected in this paper is 10, which may reduce the excavation of alteration genes to a certain extent. However, it is believed that in the future, with the increase of the sample size, the increase of different DNA copy number alteration types in CRC will lead to the identification of much clinically relevant SCNA genes.

In summary, the findings of the present study suggest that PDSGs obtained from the analysis of CRC have good application value and can provide an important reference for the precise treatment of CRC.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

WZ designed and supervised the study and was a major contributor in editing the manuscript. ZC analyzed and interpreted the data and was a major contributor in writing the manuscript. XM performed analysis and contributed to the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by the Fundamental Research Funds for the Provincial Universities (31041180039).

REFERENCES

Bae, J. M., Wen, X., Kim, T. S., Kwak, Y., Cho, N. Y., Lee, H. S., et al. (2019). Fibroblast growth factor receptor 1 (FGFR1) amplification detected by droplet digital polymerase chain reaction (ddPCR) is a prognostic factor in colorectal cancers. *Cancer Res. Treat.* doi: 10.4143/crt.2019.062

Baron, S., Vangheluwe, P., Sepulveda, M. R., Wuytack, F., Raeymaekers, L., and Vanoevelen, J. (2010). The secretory pathway Ca(2+)-ATPase 1 is associated with cholesterol-rich microdomains of human colon adenocarcinoma cells. *Biochim. Biophys. Acta* 1798, 1512–1521. doi: 10.1016/j.bbamem.2010.03.023

Birchler, J. A., and Veitia, R. A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U.S.A.* 109, 14746–14753. doi: 10.1073/pnas.1207726109

Birchler, J. A., Bhadra, U., Bhadra, M. P., and Auger, D. L. (2001). Dosage-dependent gene regulation in multicellular eukaryotes: implications for dosage compensation, aneuploid syndromes, and quantitative traits. *Dev. Biol.* 234, 275–288. doi: 10.1006/dbio.2001.0262

Bolhaqueiro, A. C. F., Ponsioen, B., Bakker, B., Klaasen, S. J., Kucukkose, E., Van Jaarsveld, R. H., et al. (2019). Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids. *Nat. Genet.* 51, 824–834. doi: 10.1038/s41588-019-0399-6

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., et al. (2019). Genenames.org: the HGNC and VGNC resources in 2019. *Nucleic Acids Res.* 47, D786–D792. doi: 10.1093/nar/gky930

De Miranda, N. F., Van Dinther, M., Van Den Akker, B. E., Van Wezel, T., Ten Dijke, P., and Morreau, H. (2015). Transforming growth factor beta signaling in colorectal cancer cells with microsatellite instability despite biallelic mutations in TGFBR2. *Gastroenterology* 148, 1427–1437 e1428. doi: 10.1053/j.gastro.2015.02.052

Harel, T., and Lupski, J. R. (2018). Genomic disorders 20 years on-mechanisms for clinical manifestations. *Clin. Genet.* 93, 439–449. doi: 10.1111/cge.13146

Insel, P. A., Sriram, K., Wiley, S. Z., Wilderman, A., Katakia, T., Mccann, T., et al. (2018). GPCRomics: GPCR expression in cancer cells and tumors identifies new, potential biomarkers and therapeutic targets. *Front. Pharmacol.* 9, 431. doi: 10.3389/fphar.2018.00431

Jilling, T., and Kirk, K. L. (1996). Cyclic AMP and chloride-dependent regulation of the apical constitutive secretory pathway in colonic epithelial cells. *J. Biol. Chem.* 271, 4381–4387. doi: 10.1074/jbc.271.8.4381

Kim, Y. H., Lee, S. B., Shim, S., Kim, A., Park, J. H., Jang, W. S., et al. (2019). Hyaluronic acid synthase 2 promotes malignant phenotypes of colorectal cancer cells through transforming growth factor beta signaling. *Cancer Sci.* 110, 2226–2236. doi: 10.1111/cas.14070

Kostolansky, F., Styk, B., and Russ, G. (1986). A simple and rapid characterization of influenza virus isolates by monoclonal antibodies in radioimmunoassay. *Acta Virol.* 30, 267–270. doi: 10.1016/0168-1702(86)90085-7

Lee, J. H., An, C. H., Yoo, N. J., and Lee, S. H. (2018a). Mutational intratumoral heterogeneity of a putative tumor suppressor gene RARRES3 in colorectal cancers. *Pathol. Res. Pract.* 214, 601–602. doi: 10.1016/j.prp.2017.12.011

Lee, K. S., Kim, B. H., Oh, H. K., Kim, D. W., Kang, S. B., Kim, H., et al. (2018b). Programmed cell death ligand-1 protein expression and CD274/PD-L1 gene amplification in colorectal cancer: Implications for prognosis. *Cancer Sci.* 109, 2957–2969. doi: 10.1111/cas.13716

Lee, K. T., Vider, J., Tang, J. C., Gopalan, V., and Lam, A. K. (2019). GAEC1 drives colon cancer progression. *Mol. Carcinog.* 58, 1145–1154. doi: 10.1002/mc.22998

Li, J., Dittmar, R. L., Xia, S., Zhang, H., Du, M., Huang, C. C., et al. (2017). Cell-free DNA copy number variations in plasma from colorectal cancer patients. *Mol. Oncol.* 11, 1099–1111. doi: 10.1002/1878-0261.12077

Liu, Z., Bai, J., Zhang, L., Lou, F., Ke, F., Cai, W., et al. (2017). Conditional knockout of microRNA-31 promotes the development of colitis associated cancer. *Biochem. Biophys. Res. Commun.* 490, 62–68. doi: 10.1016/j.bbrc.2017.06.012

Liu, S. H., Shen, P. C., Chen, C. Y., Hsu, A. N., Cho, Y. C., Lai, Y. L., et al. (2019). DriverDBv3: a multi-omics database for cancer driver gene research. *Nucleic Acids Res.* doi: 10.1093/nar/gkz964

Marquez, J., Kratchmarova, I., Akimov, V., Unda, F., Ibarretxe, G., Clerigue, A. S., et al. (2019). NADH dehydrogenase complex I is overexpressed in incipient metastatic murine colon cancer cells. *Oncol. Rep.* 41, 742–752. doi: 10.3892/or.2018.6892

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, R41. doi: 10.1186/gb-2011-12-4-r41

Neumann, P. A., Koch, S., Hilgarth, R. S., Perez-Chanona, E., Denning, P., Jobin, C., et al. (2014). Gut commensal bacteria and regional Wnt gene expression in the proximal versus distal colon. *Am. J. Pathol.* 184, 592–599. doi: 10.1016/j.ajpath.2013.11.029

Oliveira, D. M., Santamaria, G., Laudanna, C., Migliozzi, S., Zoppoli, P., Quist, M., et al. (2018). Identification of copy number alterations in colon cancer from analysis of amplicon-based next generation sequencing data. *Oncotarget* 9, 20409–20425. doi: 10.18632/oncotarget.24912

Rosenberg, S., Ducray, F., Alentorn, A., Dehais, C., Elarouci, N., Kamoun, A., et al. (2018). Machine learning for better prognostic stratification and driver gene identification using somatic copy number variations in Anaplastic Oligodendroglioma. *Oncologist* 23, 1500–1510. doi: 10.1634/theoncologist.2017-0495

Roy, D. M., Walsh, L. A., Desrichard, A., Huse, J. T., Wu, W., Gao, J., et al. (2016). Integrated genomics for pinpointing survival loci within arm-level somatic copy number alterations. *Cancer Cell* 29, 737–750. doi: 10.1016/j.ccell.2016.03.025

Salpietro, V., Manole, A., Efthymiou, S., and Houlden, H. (2018). A Review of copy number variants in inherited neuropathies. *Curr. Genomics* 19, 412–419. doi: 10.2174/1389202919666180330153316

Sefrioui, D., Vermeulin, T., Blanchard, F., Chapusot, C., Beaussire, L., Armengol-Debeir, L., et al. (2017). Copy number variations in DCC/18q and ERBB2/17q are associated with disease-free survival in microsatellite stable colon cancer. *Int. J. Cancer* 140, 1653–1661. doi: 10.1002/ijc.30584

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, J., Zhou, W., Zhu, B., Hyland, P. L., Bennett, H., Xiao, Y., et al. (2016). Rare germline copy number variations and disease susceptibility in Familial Melanoma. *J. Invest. Dermatol.* 136, 2436–2443. doi: 10.1016/j.jid.2016.07.023

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34. doi: 10.3322/caac.21551

Steinman, G., Kleiner, G. J., and Greston, W. M. (1979). Spontaneous rupture of membranes. rapid strip tests for detection. *N. Y. State J. Med.* 79, 1849–1851. doi: 10.1056/NEJM197906283002627

Villela, D., Suemoto, C. K., Leite, R., Pasqualucci, C. A., Grinberg, L. T., Pearson, P., et al. (2018). Increased DNA copy number variation mosaicism in elderly human brain. *Neural Plast.* 2018, 2406170. doi: 10.1155/2018/2406170

Wang, X. S., Prensner, J. R., Chen, G., Cao, Q., Han, B., Dhanasekaran, S. M., et al. (2009). An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nat. Biotechnol.* 27, 1005–1011. doi: 10.1038/nbt.1584

Wang, Y., He, S., Zhu, X., Qiao, W., and Zhang, J. (2016). High copy number of mitochondrial DNA predicts poor prognosis in patients with advanced stage colon cancer. *Int. J. Biol. Markers* 31, e382–e388. doi: 10.5301 /jbm.5000211

Wielenga, M. C. B., Colak, S., Heijmans, J., Van Lidth De Jeude, J. F., Rodermond, H. M., Paton, J. C., et al. (2015). ER-Stress-Induced differentiation sensitizes colon cancer stem cells to chemotherapy. *Cell Rep.* 13, 489–494. doi: 10.1016/j.celrep.2015.09.016

Xu, J. F., Kang, Q., Ma, X. Y., Pan, Y. M., Yang, L., Jin, P., et al. (2018). A novel method to detect early colorectal cancer based on chromosome copy number variation in plasma. *Cell Physiol. Biochem.* 45, 1444–1454. doi: 10.1159/000487571

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

Yu, C. Y., Chang, W. C., Zheng, J. H., Hung, W. H., and Cho, E. C. (2018). Transforming growth factor alpha promotes tumorigenesis and regulates epithelial-mesenchymal transition modulation in colon cancer. *Biochem. Biophys. Res. Commun.* 506, 901–906. doi: 10.1016/j.bbrc.2018.10.137

Yu, Y., Blokhuis, B. R., Garssen, J., and Redegeld, F. A. (2019). A transcriptomic insight into the impact of colon cancer cells on mast cells. *Int. J. Mol. Sci.* 20 (7), 1689. doi: 10.3390/ijms20071689

Yusnita, Y., Norsiah, M. D., and Rahman, A. J. (2010). Mutations in mitochondrial NADH dehydrogenase subunit 1 (mtND1) gene in colorectal carcinoma. *Malays J. Pathol.* 32, 103–110.

Zamanian-Azodi, M., and Rezaei-Tavirani, M. (2019). Investigation of health benefits of cocoa in human colorectal cancer cell line, HT-29 through interactome analysis. *Gastroenterol. Hepatol. Bed. Bench* 12, 67–73.

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: 09 January 2020
doi: 10.3389/fgene.2019.01309

eQTLMAPT: Fast and Accurate eQTL Mediation Analysis With Efficient Permutation Testing Approaches

Tao Wang[1], Qidi Peng[1], Bo Liu[1], Xiaoli Liu[2], Yongzhuang Liu[1], Jiajie Peng[3] and Yadong Wang[1]**

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Department of Neurology, Zhejiang Hospital, Hangzhou, China, [3] School of Computer Science, Northwestern Polytechnical University, Xi'an, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University,
China

Reviewed by:
Juan Wang,
Inner Mongolia University,
China
Fan Yang,
Harvard Medical School,
United States

***Correspondence:**
Jiajie Peng
jiajiepeng@nwpu.edu.cn
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to
Statistical Genetics
and Methodology,
a section of the journal
Frontiers in Genetics

Received: 08 September 2019
Accepted: 27 November 2019
Published: 09 January 2020

Citation:
Wang T, Peng Q, Liu B, Liu X, Liu Y,
Peng J and Wang Y (2020)
eQTLMAPT: Fast and Accurate eQTL
Mediation Analysis With Efficient
Permutation Testing Approaches.
Front. Genet. 10:1309.
doi: 10.3389/fgene.2019.01309

Expression quantitative trait locus (eQTL) analyses are critical in understanding the complex functional regulatory natures of genetic variation and have been widely used in the interpretation of disease-associated variants identified by genome-wide association studies (GWAS). Emerging evidence has shown that *trans*-eQTL effects on remote gene expression could be mediated by local transcripts, which is known as the mediation effects. To discover the genome-wide eQTL mediation effects combing genomic and transcriptomic profiles, it is necessary to develop novel computational methods to rapidly scan large number of candidate associations while controlling for multiple testing appropriately. Here, we present eQTLMAPT, an R package aiming to perform eQTL mediation analysis with implementation of efficient permutation procedures in multiple testing correction. eQTLMAPT is advantageous in threefold. First, it accelerates mediation analysis by effectively pruning the permutation process through adaptive permutation scheme. Second, it can efficiently and accurately estimate the significance level of mediation effects by modeling the null distribution with generalized Pareto distribution (GPD) trained from a few permutation statistics. Third, eQTLMAPT provides flexible interfaces for users to combine various permutation schemes with different confounding adjustment methods. Experiments on real eQTL dataset demonstrate that eQTLMAPT provides higher resolution of estimated significance of mediation effects and is an order of magnitude faster than compared methods with similar accuracy.

Keywords: trans-eQTL, cis-eQTL, mediation analysis, multiple testing control, permutation test, gene regulation

INTRODUCTION

Understanding the complex functional natures of genome variants has been the focus of many studies in recent years, which provides us with advanced insights into phenotype variability and disease susceptibility (Cheng et al., 2017; Watanabe et al., 2017; Gallagher and Chen-Plotkin, 2018). Vast genome variants relevant to disease risks and other traits have been unequivocally identified by genome-wide association studies (GWAS) (Visscher et al., 2017). However, most of those traits-associated variants localize in non-coding regions, intergenic, or intronic regions, indicating that genomic variants are likely to be involved in gene regulation instead of exerting their effects through altering the protein sequence directly (Gallagher and Chen-Plotkin, 2018). To understand the complex regulatory natures of genomic variants, one of the fundamental tasks is to discover target genes which can be regulated by variants in the cell. The expression quantitative trait loci (eQTL) analysis has been proven a powerful tool in achieving this goal.

An eQTL is essentially a variant at a specific genome location with its genetic variance associates with gene expression variation in a population. Most eQTL mapping studies access the eQTL effects through association tests between the genotypes of a variant and expression profiles of a gene using regression models (Shabalin, 2012; Ongen et al., 2015). And eQTL summary statistics have been widely used in the interpretation of GWAS results and Mendelian randomization studies (Cheng et al., 2018b; Peng et al., 2019a). eQTLs can exert their regulatory effects on local gene transcriptions (*cis*-acting) and distant gene transcriptions (*trans*-acting), defined by the physical distance between an eQTL and a gene, usually using 1 Mb as a threshold or on different chromosomes for *trans*-acting associations (Ongen et al., 2015; GTEx Consortium, 2017). *cis*-acting or *trans*-acting may reflect different underlying regulation natures. For example, *cis*-eQTLs usually reside close to transcription starting sites (TSS) and might affect the gene expression directly through affecting transcription factor (TF) binding process (Nica and Dermitzakis, 2013). However, very little knowledge is known for *trans*-eQTLs due to multifaceted reasons. First, *trans*-acting effects are usually weaker than *cis*-acting, which requires a large sample size to detect the weak signals (Yao et al., 2017). Second, the number of *trans*-eQTL associations is an order of magnitude more than *cis*-eQTL associations, which brings heavy computational burdens. Third, the multiple testing problem in identifying *trans*-eQTLs results in stringent significance thresholds. And *trans*-eQTLs have been proven less replicable across studies (Innocenti et al., 2011). Therefore, most eQTL studies only focus on *cis*-eQTLs, and the mechanisms underling the regulatory effects of genetic variation on the expression of distant genes and genes in other chromosomes are largely unknown (Bryois et al., 2014).

Recent studies have shown that *trans*-eQTLs are likely involved in indirect regulations, where the *trans*-eGene can be mediated by the *cis*-eGene, which is known as the mediation effect (Pierce et al., 2014; Brynedal et al., 2017; Yang et al., 2017; Yao et al., 2017). These studies provide evidence of a *cis*-mediated mechanism that explains distal regulation of *trans*-eGenes by *trans*-eQTLs (Yao et al., 2017). Characterizing these regulatory relationships will allow us to better understand regulatory networks and the biological mechanisms underlying *trans*-eQTLs (Westra et al., 2013). To discover the mediation effect among *cis*-/*trans*-eQTL (L), *cis*-eGene (C) and *trans*-eGene (T), represented by a trio ($L \rightarrow C \rightarrow T$), a recently proposed work which aims to test the significance of the effect of *cis*-eGene on *trans*-eGene controlled by the genotype of L and confounders (Yang et al., 2017). Mathematically, by using a linear regression model, with the formula $T = a + \beta_1 C + \beta_2 G + \Gamma\mathrm{Cov} + \epsilon$, where G represents the genotype of L (see details in *Material and Methods*), the objective is to test the significance of β_1. In practice, this requires performing a large amount of association tests in order to scan all possible candidate trios due to related variants in linkage disequilibrium (LD). Thus, it will result in a large number of nominal statistics, i.e., P values, and multiple testing has to be considered in order to control the false discovery rate. A traditional solution is to use Bonferroni correction method, which multiplies the nominal P value with the total number of tests to get an adjusted P value. However, the Bonferroni method has been proven overly stringent in genomic area due to the fact that a large number of tests are not independent because of variants in LD, and this method will result in a lot of false negatives (Ongen et al., 2015).

To solve this problem, a commonly adopted strategy is to use the non-parametric permutation testing approach. The permutation test can be performed by the following steps: first, perform thousands of permutations on gene expression profiles by randomly exchanging sample IDs. Notably, to break the potential mediation effects from C to T while keeping the *cis*-eQTL and *trans*-eQTL associations, the sample ID rearrangement need to be performed within each genotype group (i.e., AA, AB, or BB) (Yang et al., 2017). Second, calculate a list of permutation statistics, under the null hypothesis of no association, by performing associations using genotypes and permuted expressions. Third, compare the nominal statistics with the distribution of permutation statistics to assess how likely the observed nominal association statistics originates from the null distribution. The permutation tests have been applied to multiple bioinformatics applications to control for multiple testing, for example, eQTL mapping (Ongen et al., 2015), allelic association analysis (Zhao et al., 2000), and biological network analyses (Wang et al., 2019). In the context of detecting mediation effect of *cis*-eGenes on *trans*-eGenes, a recently proposed algorithm named GMAC adopts permutation strategy to control for multiple testing (Yang et al., 2017). However, it suffers from a main drawback: it relies on performing a fixed number, usually thousands of permutations per trio, to balance the running time and P value resolution empirically estimated. For example, 10,000 permutations can derive P value at a resolution of 10^{-4} at the best circumstance. There is no efficient built-in permutation scheme, which makes its practical application very time-consuming and not accurate in estimating significance of mediation effects.

In this work, we present eQTLMAPT, an R package which improves upon GMAC (Yang et al., 2017) by implementing faster and more efficient permutation-based multiple testing correction approaches. Besides the traditional fixed permutation scheme, eQTLMAPT also provides 1) the adaptive permutation scheme which prunes the permutation process opportunely; 2) the approximation of the tail of null distribution using generalized Pareto distribution (GPD) model, which allows the user to accurately estimate adjusted P values at any significance level in a short running time; and 3) flexible choices of different confounding factors adjustment methods. In addition, eQTLMAPT provides flexible interfaces for users to combine different features and perform the proper permutation scheme based on their practical needs. Experiments on a real eQTL dataset demonstrate that eQTLMAPT is an order of magnitude faster than GMAC, and its estimated significance has a much higher resolution than the compared method.

MATERIAL AND METHODS

Overview

To efficiently identify *cis*-eGene mediators of *trans*-eQTLs in whole genome, we developed eQTLMAPT, an R package to perform mediation analysis with multiple permutation schemes and flexible covariate adjustment strategies. The core regression models we used in mediation analysis is similar to the model used in the recently proposed method, GMAC (Yang et al., 2017). The models can be formalized as Equations 1, 2, and 3, where G represents the genotype of single nucleotide polymorphism (SNP)L; C, and T represent gene expression levels of *cis*-eGene and *trans*-eGene, respectively; Cov represents covariates; and ϵ represents the error term following normal distribution. For the trio (L,C,T), we assume L is significantly associated with C and T by testing $\beta_1 \neq 0$ and $\beta_2 \neq 0$ in the linear models, with β estimated by least-squares fitting. The statistic of mediation analysis here is to test the mediation effect of *cis*-eGene C on *trans*-eGene T while controlling for the effects of eQTL L, covariants Cov. The null hypothesis is $H_0{:}\beta_3 = 0$.

$$C = a_1 + \beta_1 G + \Gamma_1 Cov + \varepsilon_1 \tag{1}$$

$$T = a_2 + \beta_2 G + \Gamma_2 Cov + \varepsilon_2 \tag{2}$$

$$T = a_3 + \beta_3 C + \beta_4 G + \Gamma_3 Cov + \varepsilon_3 \tag{3}$$

Our method can be separated into two main steps: first, we calculate the nominal association statistic, $z = \beta_3/se$, in Equation 3, where se represents the standard error of β_3. Second, to account for multiple testing in assessing the significance of the mediation effect, we perform within-genotype group permutations of *cis*-eGene transcripts C to empirically characterize the null distribution of mediation effects (i.e., the distribution of z scores expected under the null hypothesis of no mediation effect, denoted by vector Z_0). The purpose of within-genotype group permutation is to break the potential mediation effects from C to T within each genotype group (i.e., AA, AB, or BB) while keeping the *cis*-eQTL and *trans*-eQTL associations. The adjusted empirical P value of mediation test would finally be calculated by comparing the observed mediation statistic z with the permutation statistics Z_0 under the null.

To obtain the null distribution of mediation effects, i.e., Z_0, and provide users with flexible choices, we implemented three permutation schemes in our package: 1) fixed permutation scheme, which generates N permutation datasets (*Estimation of P Values Under Fixed Permutation Scheme*); 2) adaptive permutation scheme, which prunes the permutation process when there are too many null statistics better than the observed z statistic (*Calculate Empirical P Value Using Adaptive Permutation Scheme*); and 3) GPD approximation, which models the tail of the null distribution *via* a drastically reduced number of null statistics and estimates P value with higher resolution (*Model the Tail of the Null Distribution Using GPD*). To deal with complex hidden confounding effects, we also adopt an adaptive confounder adjustment method (Yang et al., 2017) and a fixed confounder adjustment method incorporating the three permutation schemes (*Confounding Factors Adjustment*).

Estimation of *P* Values Under Fixed Permutation Scheme

The associations of trios (L,C,T) we aim to test are not independent due to the fact that multiple SNPs are correlated because of LD. Traditional multiple testing correction methods like Bonferroni and Benjamini–Hochberg correction, which give a global significance threshold based on all nominal P values, prove to be overly stringent and may result in false negatives in such correlated genomic analyses. Thus, we adapt permutation-based testing approaches to assess the significance in association test for each trio (L, C, T) (Equation 3). Permutation test is a widely used non-parametric method in many bioinformatics applications. It generates a null statistic distribution by random permutations and then assesses how likely the observed statistic obtained in the nominal association originates from the null distribution.

Assume the nominal mediation statistic $z = \beta_3/se$ is assessed for a trio (L, C, T) by Equation 3, where se is the standard error of β_3. Given a fixed number of N, we perform N times permutations within-genotype groups for *cis*-eGene C by randomly permuting sample labels in each genotype group, i.e., AA, AB, and BB. It will generate N null mediation statistics, denoted by $Z_0 = \{z_0^1, z_0^2, \ldots, z_0^N\}$, where z_0^i is in absolute value, $i \in [1,N]$. If M null statistics in Z_0 are stronger than the observed statistic $|z|$, the empirical P value is assessed by Equation 4, where pseudo-count 1 is added to avoid meaningless denominator.

$$P_{\text{fixed}} = \frac{M+1}{N+1} \tag{4}$$

The strategy of fixed permutation scheme is direct, easy to implement, and adopted by most permutation testing approaches. However, the adjusted P value has lower bound limitation that $P_{fixed} \geq \frac{1}{N+1}$. That means we have to increase the fixed number of N to get precise P value estimates for strong mediation effects with smaller P values, which will tremendously increase the computational costs. For example, if the true P value is 10^{-6} for a trio, at least 1 million permutations should be performed to achieve the precise P value. But for most trios, with true P values larger than 10^{-3}, 1 million permutations would be a waste of resources because thousands of permutations could lead to precise P values. To solve this problem, we implemented an adaptive permutation strategy in eQTLMAPT to prune permutations once we observe too many null statistics stronger than the nominal statistic z of mediation analysis.

Calculate Empirical *P* Value Using Adaptive Permutation Scheme

The basic idea of adaptive permutation strategy is to perform more permutations for significant trios while decreasing the number of permutations for insignificant trios. This is because insignificant trios could be assessed with fewer permutations than significant ones. By setting a significance level, $\alpha = 0.05$ for example, and a maximum permutation times N, in case of indefinitely running the process, we define the pruning threshold $K = \alpha{*}N$, and usually $K << N$. For each trio (L,C,T), if we observe more than or equal to K null statistics that $|z_0^i| > |z|$ or we reach the maximum permutation upper bound N, the permutations process will be stopped. Suppose Γ times of permutations are executed in total and M null statistics are found to be stronger than the observed statistic $|z|$, the adjusted P value is given by Equation 5.

$$P_{adaptive} = \frac{\min(K+1, M+1)}{\min(\Gamma+1, N+1)} \tag{5}$$

For example, given $N = 10{,}000$ and $\alpha = 0.05$, then $K = 500$, and assume we have performed 800 times of permutation for a trio and find K null statistics stronger than nominal statistic z. Then, we stop performing further permutations and the final adjusted P value = 501/801. In this case, only 800 times permutations are needed instead of 10,000 times in the fixed permutation scheme. This strategy tremendously reduces the number of permutations required for insignificant trios; however, the lower bound of adjusted P value still exists, which is $1/(N + 1)$. To solve the lower bound problem, we approximate the tail of null statistics distribution by generalized Pareto distribution and estimate the small P values at any significance level without the limitation of lower bound.

Model the Tail of the Null Distribution Using GPD

It is critical to accurately estimate small P values especially in large-scale genomic analyses, where huge numbers of associations are simultaneously tested. To determine precise small P values at any significance level without performing all possible permutations, we implemented a P value approximation method based on GPD, which has been widely used in modeling extreme values (Knijnenburg et al., 2009). The basic methodology is to estimate the small permutation P values using extreme value theory by fitting extreme permutation values originating from the tail of null distribution with generalized Pareto distribution (Gumbel, 2012). And it has been proven that the GPD approximation method can lead to precise estimation of small P values using much fewer permutations compared with fixed number of permutation approach (Knijnenburg et al., 2009).

In our case, given permutation statistics set $Z_0 = \{ z_0^1, z_0^2, \ldots, z_0^N \}$ and nominal mediation statistic z of a trio (L,C,T), we suppose both z and $z_0^i \in Z_0$ are in absolute value, and elements in Z_0 are sorted in decreasing order, i.e., $z_0^i \geq z_0^j$, $i<j$. Define *Nexc* as the number of exceedances (extreme values), and $Y_0 = \{ z_0^1, z_0^2, \ldots, z_0^{N_{exc}} \}$,$Y_0 \subset Z_0$, and exceedance threshold $t = (z_0^{N_{exc}} + z_0^{N_{exc}+1})/2$, such that $z_0 > t$, if $z_0 \in Y_0$. Then, we calculate $z_0 - t$ for each element $z_0 \in Y_0$ to get a vector of exceedances $X_0 = \{ x_0^1, x_0^2, \ldots x_0^{N_{exc}} \}$, where $x_0^i = z_0^i - t, x_0^i \in X_0, z_0^i \in Y_0$. Next, exceedances in X_0 are used to fit the tail of the null distribution modeling by GPD. The GPD has cumulative distribution function (CDF) shown in Equation 6.

$$F(x) = \begin{cases} 1 - \left(1 - \frac{kx}{a}\right)^{\frac{1}{k}}, & k \neq 0 \\ 1 - e^{\frac{-x}{a}}, & k = 0 \end{cases} \tag{6}$$

The a and k are scale parameter and shape parameter, respectively, and the range of x requires $0 \leq x \leq \frac{a}{k}$ for $k > 0$, and $x \geq 0$ for $k \leq 0$. If x falls out of these ranges, the GPD estimated P values will be zeros, i.e., $k > 0, x > \frac{a}{k}$. Maximum likelihood (ML) is used to estimate the two parameters a and k in $F(x)$ given X_0. The goodness-of-fit test of the Anderson–Darling statistic is used to evaluate whether the exceedances follow the GPD (Choulakian and Stephens, 2001). Finally, the permutation test P value of the GPD approximation is computed as shown in Equation 7, where z represents the absolute value of the nominal mediation statistic.

$$P_{gpd} = \frac{N_{exc}}{N}(1 - F(z - t)) \tag{7}$$

N_{exc} is initialized as minimum value between 250 and number of permutation tests by default. If it fails to fit GPD (goodness-of-fit test $P \leq 0.05$), then iteratively reduce N_{exc} by 10 until a good fit is achieved. Besides, the GPD approximation can only be used when the nominal mediation statistic z is in the range of extreme permutation null statistics (tail of null distribution). For example, if z is in the middle of the null distribution, this method cannot be applied. To specify, let M be the number of permutation values that exceed the test statistic z, if $M < N{*}\alpha$, $\alpha = 0.01$ in default, GPD approximation will be performed; otherwise, fixed permutation scheme will be performed. The detailed methods have been described in Knijnenburg et al. (2009), and we implemented this method with R language in our package to accurately estimate the mediation significance with much fewer permutations.

Confounding Factors Adjustment

The presence of heterogeneous known or latent unmeasured covariates that affect genotype and phenotype (gene expression in our context) is a major source of bias in the mediation analysis, which needs to be adjusted. The common sources of covariates, such as batch effects, age, sex, postmortem interval (PMI), RNA integrity number (RIN), and population stratification, are associated with either samples or individuals. The latent unwanted covariates can be identified by methods like principal component analysis (PCA) (Abdi and Williams, 2010), surrogate variables analysis (SVA) (Leek et al., 2012), and probabilistic estimation of expression residuals (PEER) (Stegle et al., 2012).

In our package, we adopt two covariates adjustment strategies: fixed confounder adjustment strategy and adaptive confounder adjustment strategy. The first one is to directly pass the user-given PCs/SVs or PEER factors together with known covariates into the *Cov* variable in Equation 3 when performing mediation analysis. The second way is proposed in GMAC (Yang et al., 2017), which adaptively selects hidden covariates for each trio. In brief, this method first identifies a pool of hidden covariates, represented by H, which can be supplied by users or identified with PCA on expression profiles automatically [first 30 principal components (PCs) in default]. Then, for each trio (L, C,T), only a small number of PCs will be selected from H for adjustment based on the correlations between PCs and C,T. And experiments demonstrated that this adaptive covariates selection method improved power and precision in mediation analysis (Yang et al., 2017). Notably, both covariates adjustment strategies can be flexibly selected by users for each of the three permutation schemes introduced above.

ROSMAP Dataset and Preprocessing

ROSMAP Study and Dataset

The Religious Orders Study (ROS) (A Bennett et al., 2012a) and Memory and Aging Project (MAP) (A Bennett et al., 2012b) are two longitudinal cohort studies of aging and Alzheimer's disease (AD). We downloaded the gene expression, genotype, and clinical dataset of ROSMAP Study from Synapse platform (ID: syn3219045) with approval. RNA samples were obtained from the homogenate of the dorsolateral prefrontal cortex of 724 subjects and RNA sequencing (RNA-seq) data have been processed into read count table using standard pipeline (syn9702085) (Mostafavi et al., 2018). DNA samples were from whole blood and genotype profiles of 1,179 subjects were calculated from whole-genome sequencing (De Jager et al., 2018). Only neuropathologically healthy individuals (cogdx score ≤3, no Alzheimer's disease and no dementia) with both genotype data and RNA-seq data passing quality controls were used in eQTL analysis, which downsized the sample size to $N = 334$.

Genotype Processing

We applied PLINK2 (v1.9beta) (Chang et al., 2015) and in-house scripts to perform rigorous subject and SNP quality control (QC) for genotype dataset derived from WGS. To QC in SNP level, we removed SNPs with genotype call rate <95%, with Hardy–Weinberg equilibrium testing $P < 10^{-6}$, informative missingness test $P < 10^{-9}$, and with minor allele frequency (MAF) < 0.05 seperately. To QC in subject level, we removed subjects with call rate <95%, with outlying heterozygosity rate based on heterozygosity F score (beyond 4*sd from the mean F score), and with gender mismatch. We also performed IBS/IBD filtering: pairwise identity-by-state probabilities were computed for removing both individuals in each pair with IBD > 0.98 and one subject of each pair with IBD > 0.1875. To test for population substructure, we performed PCA using smartPCA in ENGINSOFT (Patterson et al., 2006).

Gene Expression Profiles Processing

Stringent quality controls and normalization steps were also performed for gene expression profiles. Gene read count derived from RNA-seq was normalized to TPM (transcripts per kilobase million) by scaling gene length (union of exon length) and sequencing depth. We removed samples with gender mismatch by checking gender-specific expression genes XIST and RPS_4Y_1. Sample outliers with problematic gene expression profiles were detected and removed based on hierarchical clustering (AC't Hoen et al., 2013). Genes with low expression were also removed by keeping genes with >0.1 TPM in at least 20% of samples and ≥6 reads in at least 20% samples. For normalization, gene expression values were quantile normalized after log10-transformed. SVA package was applied for removing batch effect and adjusting age, sex, RIN, PMI, and latent covariates. Residuals were outputted for downstream eQTL analysis.

eQTL Mapping and Mediation Analysis

MatrixEQTL (Shabalin, 2012) was used for *cis*/*trans*-eQTL mapping using additive linear model. In *cis*-eQTL analysis, variants (SNPs and indels) within 1 M upstream and downstream from the TSS were tested for association with gene expression traits. And variants beyond the ±1M window were associated with the gene expression traits in *trans*-acting manner. For *cis*-eQTL results, a significance level of false discovery rate (FDR) ≤0.05 was used. And for *trans*-eQTL results, we adopt a global significance level $P < 1 \times 10^{-8}$ because of the tremendous amount of *trans*-associations and weak *trans*-eQTL effects.

For biological discovery, mediation analyses with adaptive permutation scheme and GPD approximation (N = 10,000, α = 0.05) were applied for all candidate trios (L,C,T), where eQTL L was significantly associated with *cis*-eGene C (FDR ≤ 0.05; Equation 1) and *trans*-eGene T ($P < 1 \times 10^{-8}$; Equation 2). For performance comparison, mediation analyses were performed in multiple scenarios described in the *Results* section.

RESULTS

Candidate (*L*, *C*, *T*) Trios Detected in ROSMAP Dataset

After stringent quality controls for both RNA-seq and genotyping data (*ROSMAP Dataset and Preprocessing*), 26,662 gene transcripts and 6,736,714 variants (including SNPs and indels) of 334 subjects were left for eQTL analysis. We detected 3,195,073 significant *cis*-eQTL associations, representing 5,711 unique *cis*-eGenes and 60,758 unique *cis*-eQTLs, and 145,153 *trans*-eQTL associations, representing 1,382 *trans*-eGenes and 66,847 unique *trans*-eQTLs, under significance thresholds of FDR ≤ 0.05 (corresponding $P < 1 \times 10^{-3}$) and $P < 1 \times 10^{-8}$ for *cis*- and *trans*-eQTL associations, respectively. Seventy-five percent of *trans*-eQTLs were also identified as *cis*-eQTLs, which is similar to previous findings (Pierce et al., 2014; Yao et al., 2017). To detect the mediation effects, 999,725 candidate trios (L,C,T) representing 6,217 unique gene pairs (C,T) were derived from significant *cis*- and *trans*-eQTL associations. For multiple correlated variants linked to each gene pair, we used permutation schemes introduced in *Material and Methods* to control for multiple testing, and for genome-wide unique gene pairs, we used a FDR procedure to control for multiple testing.

Performance With Adaptive Permutation Scheme

We first compared adaptive permutation scheme implemented in our package with fixed permutation strategy which was commonly adopted by traditional methods, including GMAC (Yang et al., 2017). For each unique gene pair (C,T) from candidate trios, we selected the most significant *cis*-eQTL for *cis*-eGene C, resulting in 6,217 trios. Mediation analyses with fixed permutation scheme (with N = 10,000) and adaptive permutation scheme (with N = 10,000, α = 0.05) were both performed on those 6,217 trios. Empirical P values P_{fixed} and P_{adaptive} were shown in **Figure 1A**, with Pearson's correlation r = 0.999, indicating the two schemes have similar precision. While fixed scheme always executed 10,000 times of permutations for each tested trio, adaptive scheme significantly reduced the permutation times, as shown in the histogram in **Figure 1B**. For example, 68% trios executed less than 2,000 times of permutations. The total time used with adaptive scheme is less than one-third of that with fixed permutation strategy (floating bar plot in **Figure 1B**).

More Accurate *P* Values and Fewer Permutations with GPD Approximation

Using generalized Pareto distribution to model the tail of null distribution of permutation statistics could derive more precise empirical P values with fewer number of permutations compared with traditional fixed permutation strategy (Knijnenburg et al., 2009). To test the performance of the GPD approximation method implemented in eQTLMAPT, we first randomly selected 1,000 (L,C,T) trios with fixed permutation P values were less than or equal to 0.01 (N = 10,000). And then we rerun mediation analyses for those trios with GPD approximation under fixed permutation schemes with N = 1,000, 5,000, and 10,000. The reason that we only select trios with $P \leq 0.01$ is because only permutation P values at the tail of null distribution can be estimated by the GPD approximation method (see *Model the Tail of the Null Distribution Using GPD*). **Figures 2A–C** show the GPD estimated P values versus P values derived from the fixed permutation scheme (N = 10,000, 5,000, and 1,000, respectively), and we can see that GPD-estimated P values have higher resolution than fixed permutation scheme. For instance, GPD-estimated P values range from 10^{-2} to 10^{-8}, while fixed permutation-derived mediation P values range from 10^{-2} to 10^{-3}, when N is set to 1,000. And GPD-estimated P values are much smaller than fixed permutation-derived P values, which demonstrates that the GPD approximation method has the ability to detect mediation effect more accurately with higher significance resolution.

To prove the accuracy of the GPD approximation strategy, we first sampled 1,000 trios with P value *equal to* 0.01 under the fixed permutation scheme with N = 100. It is reasonable to suppose that the significance is likely to be underestimated because of the small N ($P_{\text{fixed}} \leq 0.01$). Then we rerun the mediation analyses for those 1,000 trios with N set to 10,000, where $P_{\text{fixed}} \leq 10^{-4}$. The density plot of P values of those 1,000 trios derived under the fixed permutation scheme (N = 10,000) was shown in **Figure 3A**, where two peaks around 10^{-2} and 10^{-3} were shown. The peak around 10^{-2} indicates some trios have true significance level around 10^{-2}. However, the larger peak centers around 10^{-3} indicate that the significance of a large number of tests is underestimated when N = 100. Then we asked whether using GPD approximation strategy can derive P values proxy for true P values even when N was still set to 100. We extracted trios with significance levels between (a,b) interval (shown in **Figure 3A**) and rerun mediation analyses with GPD approximation and N was still set to 100. The distribution of the GPD approximation-derived P values was shown as the boxplot in **Figure 3A**, which were centered around 10^{-3}, as expected.

The other advantage of using GPD approximation in mediation effect analysis is that with fewer permutations large amount of time cost can be avoided. To achieve a resolution of $P \leq 10^{-8}$, at least 10^{8} permutations should be performed under

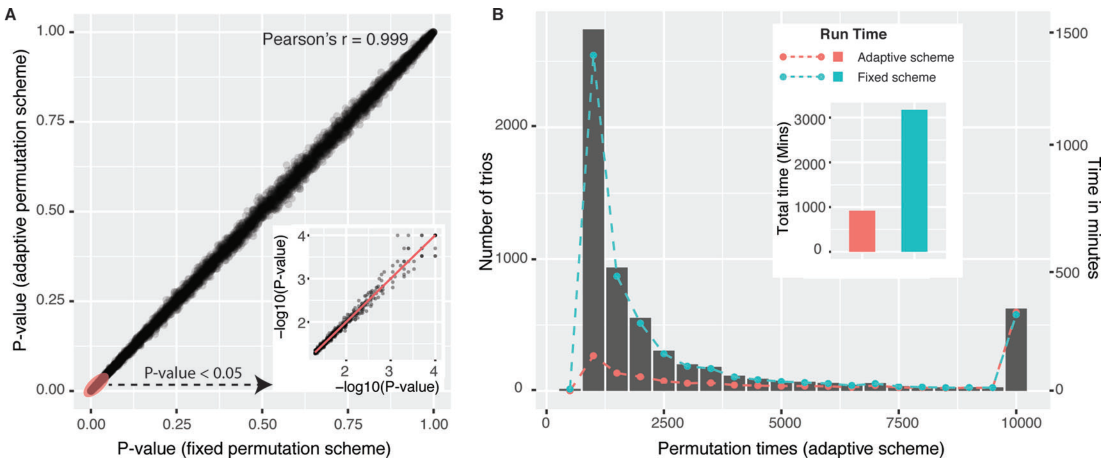


FIGURE 1 | Performance of mediation analysis with adaptive permutation scheme versus fixed permutation scheme. **(A)** Empirical P values of 6,217 (L,C,T) trios derived from adaptive scheme (y-axis) and fixed scheme (x-axis) were shown in Panel A, and the portion of P_{fixed} < 0.05 was enlarged in $-log_{10}$ scale. **(B)** Trios were grouped by permutation times (in adaptive scheme) and were shown in histogram (*left-side y*-axis). Running time of each group (*right-side y*-axis) using two permutation schemes was overlaid on the histogram with two *colored dash lines*, and the total running time was also shown in the *floating colored bar plot*. To be noted, all trios were executed 10,000 times of permutations in the fixed permutation scheme.

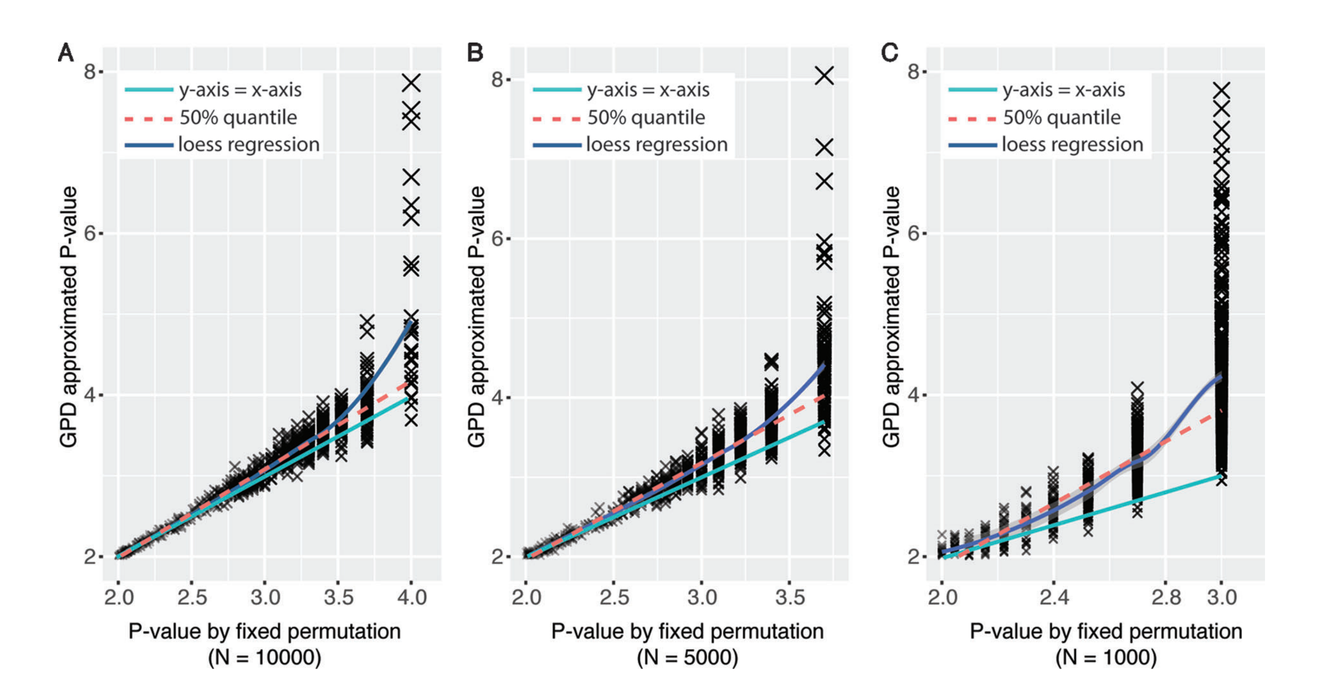


FIGURE 2 | Significance level in mediation analysis estimated under fixed permutation schemes with or without GPD approximation strategy. *X*-axis represents *P* values derived by different fixed permutation schemes (*N* = 10,000, 5,000, and 1,000, respectively) without GPD approximation. *Y*-axis represents *P* values derived with GPD approximation under certain fixed permutation scheme. *P* values were –log10-transformed.

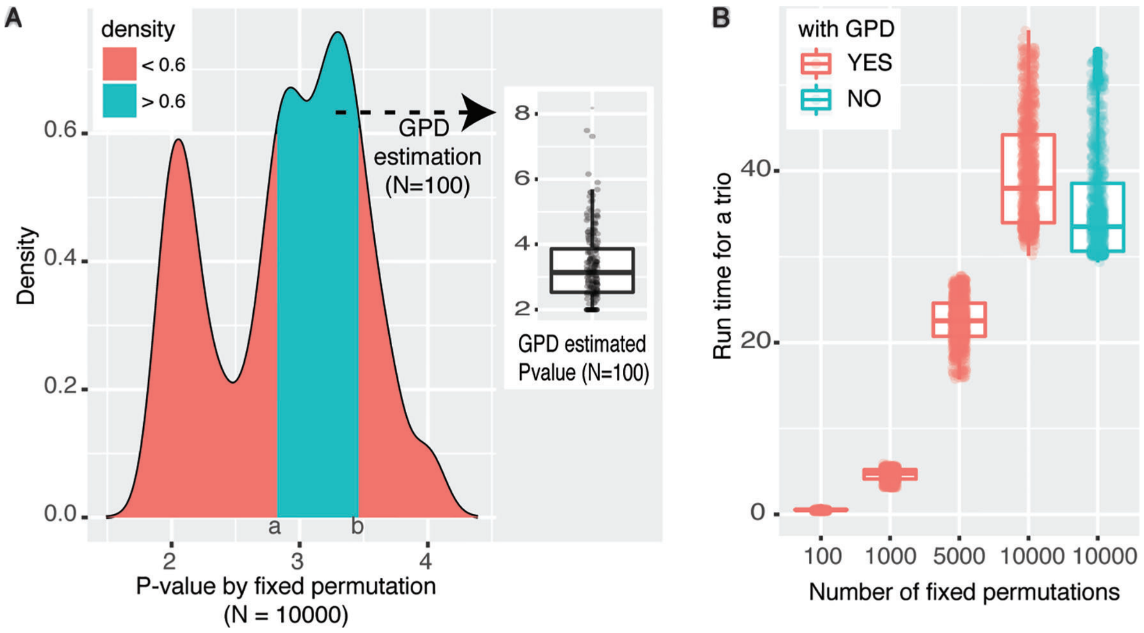


FIGURE 3 | Performance of eQTL mediation analysis with GPD approximation. **(A)** Density plot reflecting the distribution of empirical *P* values under fixed permutation scheme (*N* = 10,000) of 1,000 selected trios with P_{fixed} = 0.01 when *N* = 100. The *cyan area* was selected based on the density >0.6, and fixed permutation *P* values were around 10^{-3}, when *N* = 10,000. For trios covered by the *cyan area*, GPD-estimated *P* values (*N* = 100) were shown in the *floating boxplot*. **(B)** Time cost for analyzing the same set of trios under various permutation schemes. The *color legend* represents whether GPD estimation process is used. *P* values were –log10-transformed.

fixed permutation scheme, while the same resolution could be achieved with only 10^3 permutations with GPD estimation (see **Figure 2**). **Figure 3B** intuitively shows the time cost for analyzing the mediation effect of a trio under different permutation schemes. One hundred, 1,000, 5,000, and 10,000 permutations were performed in the mediation analysis of the same collection of trios. We can see that the run time is significantly correlated with permutation times. We also tested the time cost caused by the GPD estimation under 10,000 permutations (the two right-most boxplots in **Figure 3B**). We can see that the GPD estimation process only adds a few time cost burden than without GPD estimation, which shows the number of permutations are the most time-consuming. However, *P* value estimates have larger variance for small *N* and converge to the real P_{perm} when *N* is getting larger (Knijnenburg et al., 2009). Experimentally, we recommend users to use $N \leq 1{,}000$, and the larger *N* will result in more accurate estimated *P* values. In conclusion, by applying GPD approximation strategy, eQTLMAPT can accurately estimate the significance level with fewer permutation operations, which makes the mediation analysis much more efficient.

Discover *cis*-Mediators of *trans*-eQTLS Using ROSMAP Dataset

To test the speed and discovery performance, we compared eQTLMAPT, combining adaptive permutation scheme and GPD approximation strategy, with GMAC in the discovery of eQTL mediation effects using ROSMAP dataset. For each unique gene pair, we first selected the best trio showing the strongest mediation effect based on the nominal *P* value, resulting in 6,217 candidate trios. Then, we performed mediation analyses using eQTLMAPT and GMAC separately on those 6,217 candidate trios. Both methods adopt permutation tests to adjust *P* values for each trio, and FDR procedure described by Storey and Tibshirani (ST) (Storey and Tibshirani, 2003) to control for multiple testing of gene pairs. To make the comparison comparable, both methods applied the adaptive confounders selection strategy, taking all of the PCs derived from expression profiles as the selection pool of hidden confounders. And both methods adjusted the same fixed covariates (age, sex, RIN, PMI, and batch). We performed $N = 10{,}000$ permutations for GMAC and performed $N = 10{,}000$, 5,000, 1,000, and 500 permutations for eQTLMAPT, respectively. In our program, we set $\alpha = 0.05$ in adaptive permutation scheme.

Table 1 summarizes the performance between eQTLMAPT and GMAC. Both methods detected similar number of trios with suggestive mediation effects (permutation $P \leq 0.05$) and similar number of significant trios with FDR ≤ 0.25 (Storey and Tibshirani multiple-test controlling method). The Venn diagram in **Figure 4** demonstrated that most significant trios (with suggestive permutation $P \leq 0.05$ or FDR ≤ 0.25) detected by GMAC can be discovered by eQTLMAPT with $N = 10{,}000$, 5,000, 1,000, and even 500. For example, among the 113 significant trios with FDR ≤ 0.25 detected by GMAC, 110 (97%) can be discovered by eQTLMAPT with $N = 10{,}000$, and 104 (92%) can be discovered by eQTLMAPT with $N = 500$. With the similar ability in discovering significant trios, eQTLMAPT is about 90, 40, 8, and 4 times faster than GMAC when $N = 500$, 1,000, 5,000, and 10,000, respectively (**Table 1**). We also noticed that some significant trios detected by eQTLMAPT were missed by GMAC, which might be due to improved *P* value resolution. However, since there is no "true" set of trios with mediation effects, we are not able to compare the true positive rate and false positive rate. In summary, with similar discovery ability, eQTLMAPT is order of magnitudes faster than GMAC. The 519 trios intersected from the five compared strategies with suggestive permutation $P \leq 0.05$ were available in **Supplementary Table 1**.

TABLE 1 | Summary table of performance on speed and discoveries of eQTLMAPT and GMAC.

Software	No. of permutation	No. of trios (adjusted *P* ≤ 0.05)	No. of trios (FDR ≤ 0.25)	Time cost (mins)
GMAC	10,000	578	113	4,438
eQTLMAPT	10,000	580	118	1,131
	5,000	583	115	532
	1,000	577	108	101
	500	596	123	51

Enrichment Analysis for eQTLs Among GWAS SNPs

We first performed GWAS enrichment analyses for genome-wide significant *cis*-eQTLs (FDR ≤ 0.05) and *trans*-eQTLs ($P \leq 1 \times 10^{-8}$). From the NHGRI GWAS catalog (July 2019), 70,971 unique SNPs, reportedly associated with traits and genotyped in ROSMAP dataset, were downloaded (Welter et al., 2013). After pruning correlated SNPs in LD ($r^2 > 0.3$) using PLINK and ROSMAP genotype data, 30,894 independent trait-associated SNPs were left, of which, 16,398 SNPs had GWAS $P \leq 5 \times 10^{-8}$ and 14,496 SNPs had GWAS $P \leq 5 \times 10^{-8}$, respectively. Among SNPs with GWAS $P \leq 5 \times 10^{-8}$, 28% were *cis*-eQTLs compared with 18% in SNPs with GWAS $P \leq 5 \times 10^{-8}$ (Fisher's exact test OR = 1.75, with 95% CI = 1.66–1.85 and $P = 1.83 \times 10^{-93}$; **Figure 5A**). To be noted, the GWAS enrichment method was the same as described in previous work (Westra et al., 2013). In addition, we also observed GWAS enrichment for *trans*-eQTLs (Fisher's exact test OR = 2.58, with 95% CI = 1.8–3.76, and $P < 2.51 \times 10^{-8}$; **Figure 5B**). This demonstrated that SNPs known to be associated with traits were more likely to be *cis*/*trans*-eQTLs, which was consistent with previous findings (Fehrmann et al., 2011; Pierce et al., 2014).

Next, we performed GWAS enrichment analysis for eQTLs with significant mediation effects. Among the 999,725 candidate trios, 67,906 trios, representing 27,100 unique SNPs, showed suggestive mediation effects with permutation $P \leq 0.05$ under fixed permutation scheme ($N = 10{,}000$). Using the same GWAS enrichment method, we found GWAS SNPs were more likely to have mediation effects (Fisher's exact test OR = 4.19, with 95% CI = 2.16–8.9, and $P = 1.47 \times 10^{-6}$; **Figure 5C**), indicating that mediation analysis can help to explain GWAS findings.

Transcription Factors May Act as *cis*-Mediators

The 519 trios with suggestive permutation $P \leq 0.05$ (**Supplementary Table 1**) represent 351 unique *cis*-mediators (*cis*-eGenes). Among those *cis*-mediators, we found 14 are TFs, including ZNF488, ZSCAN26, ZNF254, TBX1, FOXS1, ZFP57,

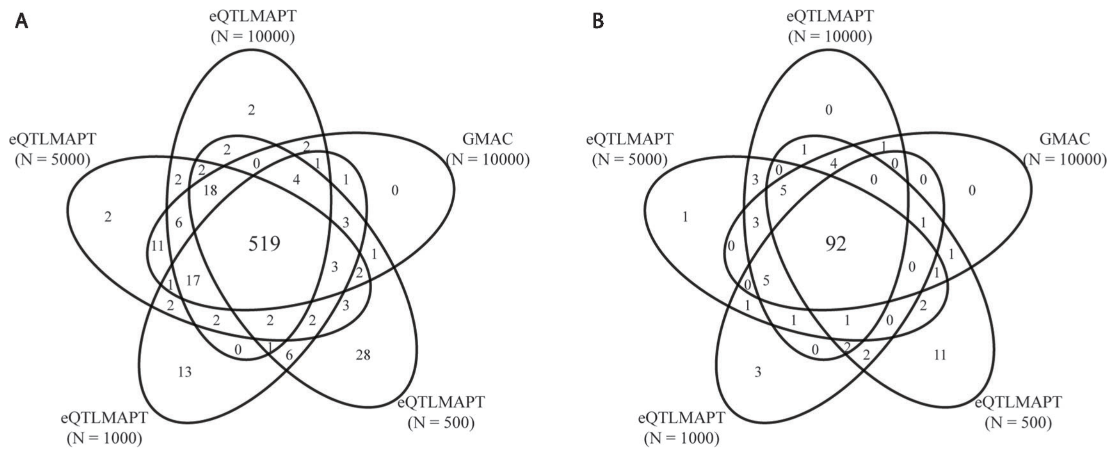
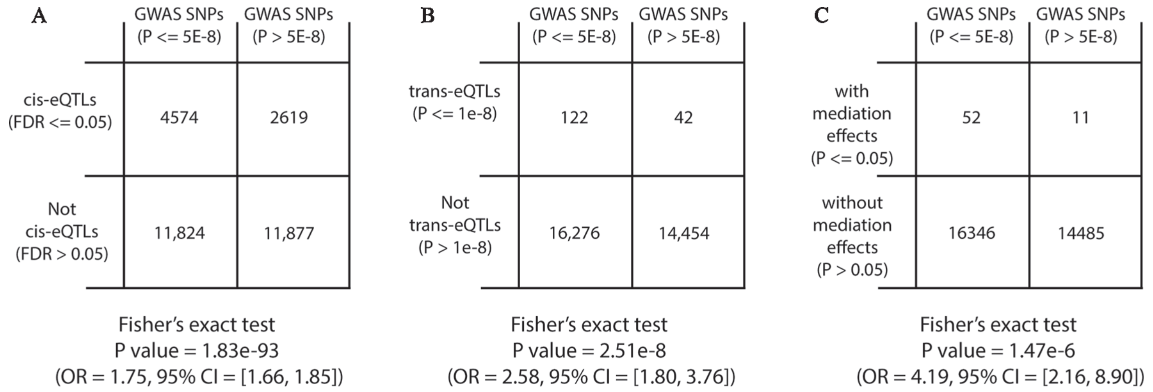


FIGURE 4 | Venn diagram of significant trios at suggestive permutation $P \leq 0.05$ **(A)** and FDR ≤ 0.25 **(B)** derived by GMAC and eQTLMAPT with different numbers of permutations.



A

	GWAS SNPs (P <= 5E-8)	GWAS SNPs (P > 5E-8)
cis-eQTLs (FDR <= 0.05)	4574	2619
Not cis-eQTLs (FDR > 0.05)	11,824	11,877

Fisher's exact test
P value = 1.83e-93
(OR = 1.75, 95% CI = [1.66, 1.85])

B

	GWAS SNPs (P <= 5E-8)	GWAS SNPs (P > 5E-8)
trans-eQTLs (P <= 1e-8)	122	42
Not trans-eQTLs (P > 1e-8)	16,276	14,454

Fisher's exact test
P value = 2.51e-8
(OR = 2.58, 95% CI = [1.80, 3.76])

C

	GWAS SNPs (P <= 5E-8)	GWAS SNPs (P > 5E-8)
with mediation effects (P <= 0.05)	52	11
without mediation effects (P > 0.05)	16346	14485

Fisher's exact test
P value = 1.47e-6
(OR = 4.19, 95% CI = [2.16, 8.90])

FIGURE 5 | Diagram of two-way contingency tables for Fisher's exact tests.

ZNF568, ZNF260, ZNF14, GTF2I, ZFX, CSDC2, GTF2IRD2B, and GTF2IRD2. For example, we observed the trio (rs77969091, TBX1, MSC), where TBX1 is the *cis*-eGene and MSC is the *trans*-eGene, and MSC has been predicted to be the target of the transcription factor TBX1 in brain tissue and central nervous system (Marbach et al., 2016). This indicates that *trans*-eQTLs can exert their effects on distant target genes through affecting TFs which act as mediators. However, we did not observe overrepresentation of TFs among *cis*-mediators (Fisher's exact test $P = 0.15$, compared with 1,665 TFs downloaded from HumanTFDB) (Hu et al., 2018).

DISCUSSION

There has been intense efforts to identify causal genes and other biomarkers such as RNA, protein, and microbiota underlying complex diseases (Cheng and Hu, 2018; Cheng et al., 2019). One of these efforts is to discover genes regulated by GWAS variants through eQTL analysis. However, less is known regarding how *trans*-eQTLs work on distant genes. The eQTL mediation analysis is a promising tool to uncover the mechanisms underlying *trans*-eQTLs. In order to discover the eQTL mediation effects in whole genome, millions of candidate associations of (eQTL, *cis*-eGene, *trans*-eGene) trios need to be tested, which requires the computational methods to control for multiple testing appropriately. In practice, there are hundreds of variants on average associated with eGenes in both *cis*- and *trans*-manner, which result in huge numbers of candidate trios. For example, in the ROSMAP dataset, nearly 1 million candidate trios need to be tested, which only represent 6,217 unique (*cis*-eGene, *trans*-eGene) pairs. To determine the genome-wide significance of a nominal testing statistics, we need to account for two multiple-testing levels: multiple genetic variants are tested per (*cis*-eGene, *trans*-eGene) pair, and multiple (*cis*-eGene, *trans*-eGene) pairs are tested genome-wide. We used permutation test to correct for the former and FDR estimation to control for the latter.

The traditional permutation scheme, which runs a fixed number of permutations, has to balance the time cost and the *P* value resolution, which is limited by a lower bound. And there is no efficient built-in permutation scheme in current tools aiming at analyzing eQTL mediation effect. To fill this gap, we present eQTLMAPT, which implements a fast and accurate eQTL analysis method with efficient permutation procedures to control for multiple testing. eQTLMAPT can correct for the multiple correlated variants tested *via* three different permutation schemes: the fixed permutation scheme, the adaptive permutation scheme, and the generalized Pareto distribution (GPD) approximation, which models the null distribution of no mediation effects using GPD trained from a few permutation statistics and could accurately estimate the adjusted *P* values without the limitation of lower bound. These strategies implemented in eQTLMAPT greatly accelerated the efficiency of multiple test controling in mediation analyses and provided users higher resolution of estimated significance which would help them distinguish the best signals.

In the analyses of the ROSMAP dataset, we detected 519 trios with suggestive mediation effects (permutation $P \leq 0.05$), representing 351 unique *cis*-eGenes. Among those *cis*-mediators, we found 14 are TFs, including ZNF488, ZSCAN26, ZNF254, TBX1, FOXS1, ZFP57, ZNF568, ZNF260, ZNF14, GTF2I, ZFX, CSDC2, GTF2IRD2B, and GTF2IRD2. This proves that TFs might play a role in the mediation effects. We also tried to replicate these significant trios with mediation effects in the GTEx dataset analyzed by Yang et al. (2017), and 70 trios, identified by gene pairs, can be replicated with mediation $P \leq 0.05$ in multiple tissues. For example, the gene pair (MZT2A, AC018804.6) was observed with mediation effects in multiple tissues including brain putamen, fibroblast, colon, esophagus, lung, muscle, pancreas, pituitary, skin, thyroid, and vagina. And the significance of the mediation effect can reach 2×10^{-7} in GTEx muscle tissue. This might suggest a common *trans*-eQTL regulatory mechanism across tissues.

There are some limitations of our method and discoveries in the ROSMAP dataset. The discovery of *trans*-eQTLs requires a large sample size because of smaller effect size of *trans*-eQTL associations. A small sample size might cause less replicable *trans*-eQTL signals across studies. The effective sample size of the ROSMAP dataset used in the discovery study is relatively small, which might be the reason that some trios were not able to be replicated in the GTEx dataset, whose sample size is also limited. Besides the transcription factors found in the *cis*-mediators, non-coding genes such as long non-coding RNA (lncRNA), microRNA, snRNA, antisense RNA, and pseudogene, were also detected. The top 3 gene classes are protein coding, pseudogene, and lncRNA genes. Although many studies have shown that non-coding RNAs play key roles in the complex regulatory networks in cell system, most of their functions are still missing (Cheng et al., 2018a; Cheng et al., 2018d; Peng et al., 2019b). Further computational methods and biological experiments are still needed to understand these unknown markers, such as using phynotypes, ontologies, deep learning methods, etc. (Cheng et al., 2016; Cheng et al., 2018c; Peng et al., 2019c; Peng et al., 2019d). In addition, since the gene expression is tissue-specific and cell type-specific, the mediation effects found in brain tissue might not show up in other tissues and cell types. Thus, with the development of single-cell RNA sequencing technologies, further studies should put more attention on cell type-specific mediation effects.

In conclusion, we present eQTLMAPT, an R package which aims to perform eQTL mediation analysis with efficient permutation procedures in multiple testing correction (**Supplementary Figure 1**). Experiments demonstrate that our method provides higher resolution in estimated significance and is an order of magnitude faster than the compared methods. Our method will be helpful in identifying mediation effects, which could allow us to better understand the biological mechanisms underlying *trans*-eQTLs and the regulatory network in the cell.

DATA AVAILABILITY STATEMENT

Genotype and RNA-seq data of ROSMAP study (in control use): Synapse platform (https://www.synapse.org/#!Synapse:syn3219045). Source code and comprehensive documentation of eQTLMAPT are freely available to download at https://github.com/QidiPeng/eQTLMAPT.

AUTHOR CONTRIBUTIONS

TW designed the study and co-implemented the R package, analyzed data, and wrote the paper. QP co-implemented the R package, performed dry experiments, and revised the paper. BL XL, and YL revised the paper and provided suggestions. JP and YW supervised the research, provided funding support, and revised the paper.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Nos. 2017YFC1201201 and 2017YFC0907503).

ACKNOWLEDGMENTS

We thank all the contributors in ROSMAP study. The results published here are in whole or in part based on data obtained from the AMP-AD Knowledge Portal (doi:10.7303/syn2580853). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01309/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Overview of functions implemented in eQTLMAPT.

REFERENCES

A Bennett, D., A Schneider, J., Arvanitakis, Z., and S Wilson, R. (2012a). Overview and findings from the religious orders study. *Curr. Alzheimer Res.* 9, 628–645. doi: 10.2174/156720512801322573

A Bennett, D., A Schneider, J., S Buchman, A., L Barnes, L., A Boyle, P., and S Wilson, R. (2012b). Overview and findings from the rush memory and aging project. *Curr. Alzheimer Res.* 9, 646–663. doi: 10.2174/156720512801322663

Abdi, H., and Williams, L. J. (2010). Principal component analysis Wiley. *Interdiscip. Rev. Comput. Stat.* 2, 433–459. doi: 10.1002/wics.101

AC't Hoen, P., Friedländer, M. R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S. Y., et al. (2013). Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nat. Biotechnol.* 31, 1015. doi: 10.1038/nbt.2702

Brynedal, B., Choi, J., Raj, T., Bjornson, R., Stranger, B. E., Neale, B. M., et al. (2017). Large-scale trans-eqtls affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *Am. J. Hum. Genet.* 100, 581–591. doi: 10.1016/j.ajhg.2017.02.004

Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., et al. (2014). Cis and trans effects of human genomic variants on gene expression. *PloS Genet.* 10, e1004461. doi: 10.1371/journal.pgen.1004461

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Cheng, L., and Hu, Y. (2018). Human disease system biology. *Curr. Gene Ther.* 18, 255–256. doi: 10.2174/1566523218666181010101114

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). Oahg: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2017). Metsigdis: a manually curated resource for the metabolic signatures of diseases. *Briefings Bioinf.* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, H., Wang, S., and Zhang, J. (2018b). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomisation study. *Front. Genet.* 9, 657. doi: 10.3389/fgene.2018.00657

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018c). Infacront: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19, 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018d). 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

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019). gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* doi: 10.1093/nar/gkz843

Choulakian, V., and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized pareto distribution. *Technometrics* 43, 478–484. doi: 10.1198/00401700152672573

De Jager, P. L., Ma, Y., McCabe, C., Xu, J., Vardarajan, B. N., Felsky, D., et al. (2018). A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Sci. Data* 5, 180142. doi: 10.1038/sdata.2018.142

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., et al. (2011). Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla. *PloS Genet.* 7, e1002197. doi: 10.1371/journal.pgen.1002197

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-gwas era: from association to function. *Am. J. Hum. Genet.* 102, 717–730. doi: 10.1016/j.ajhg.2018.04.002

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204. doi: 10.1038/nature24277

Gumbel, E. J. (2012). *Statistics of extremes* (Mineola, New York: Courier Corporation).

Hu, H., Miao, Y.-R., Jia, L.-H., Yu, Q.-Y., Zhang, Q., and Guo, A.-Y. (2018). Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* 47, D33–D38. doi: 10.1093/nar/gky822

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PloS Genet.* 7, e1002078. doi: 10.1371/journal.pgen.1002078

Knijnenburg, T. A., Wessels, L. F., Reinders, M. J., and Shmulevich, I. (2009). Fewer permutations, more accurate p-values. *Bioinformatics* 25, i161–i168. doi: 10.1093/bioinformatics/btp211

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., and Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 13, 366. doi: 10.1038/nmeth.3799

Mostafavi, S., Gaiteri, C., Sullivan, S. E., White, C. C., Tasaki, S., Xu, J., et al. (2018). A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease. *Nat. Neurosci.* 21, 811. doi: 10.1038/s41593-018-0154-9

Nica, A. C., and Dermitzakis, E. T. (2013). Expression quantitative trait loci: present and future. *Philos. Trans. R. Soc. B.: Biol. Sci.* 368, 20120362. doi: 10.1098/rstb.2012.0362

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., and Delaneau, O. (2015). Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics* 32, 1479–1485. doi: 10.1093/bioinformatics/btv722

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PloS Genet.* 2 (12), e190. doi: 10.1371/journal.pgen.0020190

Peng, J., Guan, J., and Shang, X. (2019a). Predicting parkinson's disease genes based on node2vec and autoencoder. *Front. Genet.* 10, 226. doi: 10.3389/fgene.2019.00226

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019b). A learning-based framework for mirna-disease association identification using neural networks. *Bioinformatics* 35 (21), 4364–4371. doi: 10.1093/bioinformatics/btz254

Peng, J., Lu, J., Hoh, D., Dina, A. S., Shang, X., Kramer, D. M., et al. (2019c). Identifying emerging phenomenon in long temporal phenotyping experiments. *Bioinformatics*, btz559. doi: 10.1093/bioinformatics/btz559

Peng, J., Wang, X., and Shang, X. (2019d). 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

Pierce, B. L., Tong, L., Chen, L. S., Rahaman, R., Argos, M., Jasmine, F., et al. (2014). Mediation analysis demonstrates that trans-eqtls are often explained by cis-mediation: a genome-wide analysis among 1,800 south asians. *PloS Genet.* 10, e1004818. doi: 10.1371/journal.pgen.1004818

Shabalin, A. A. (2012). Matrix eqtl: ultra fast eqtl analysis *via* large matrix operations. *Bioinformatics* 28, 1353–1358. doi: 10.1093/bioinformatics/bts163

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7, 500. doi: 10.1038/nprot.2011.457

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* 100, 9440– 9445. doi: 10.1073/pnas.1530509100

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

Wang, T., Peng, J., Peng, Q., Wang, Y., and Chen, J. (2019). Fsm: Fast and scalable network motif discovery for exploring higher-order network organizations. *Method* S1046–2023 (19), 30036–2. doi: 10.1016/j.ymeth.2019.07.008

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. *Nat. Commun.* 8, 1826. doi: 10.1038/s41467-017-01261-5

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2013). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Res.* 42, D1001–D100D, 1006. doi: 10.1093/nar/gkt1229

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eqtls as putative drivers of known disease associations. *Nat. Genet.* 45, 1238. doi: 10.1038/ng.2756

Yang, F., Wang, J., Pierce, B. L., Chen, L. S., Aguet, F., Ardlie, K. G., et al. (2017). Identifying cis-mediators for trans-eqtls across many human tissues using genomic mediation analysis. *Genome Res.* 27, 1859– 1871. doi: 10.1101/gr.216754.116

Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Liu, C., Freedman, J. E., et al. (2017). Dynamic role of trans regulation of gene expression in relation to complex traits. *Am. J. Hum. Genet.* 100, 571–580. doi: 10.1016/j.ajhg.2017.02.003

Zhao, J. H., Curtis, D., and Sham, P. C. (2000). Model-free analysis and permutation tests for allelic associations. *Hum. Heredity* 50, 133–139. doi: 10.1159/000022901

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: 09 January 2020
doi: 10.3389/fgene.2019.01243

GANsDTA: Predicting Drug-Target Binding Affinity Using GANs

Lingling Zhao [1], Junjie Wang [1], Long Pang [2], Yang Liu [1] and Jun Zhang [3]*

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Institute of Space Environment and Material Science, Harbin Institute of Technology, Harbin, China, [3] Department of Rehabilitation, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China

The computational prediction of interactions between drugs and targets is a standing challenge in drug discovery. State-of-the-art methods for drug-target interaction prediction are primarily based on supervised machine learning with known label information. However, in biomedicine, obtaining labeled training data is an expensive and a laborious process. This paper proposes a semi-supervised generative adversarial networks (GANs)-based method to predict binding affinity. Our method comprises two parts, two GANs for feature extraction and a regression network for prediction. The semi-supervised mechanism allows our model to learn proteins drugs features of both labeled and unlabeled data. We evaluate the performance of our method using multiple public datasets. Experimental results demonstrate that our method achieves competitive performance while utilizing freely available unlabeled data. Our results suggest that utilizing such unlabeled data can considerably help improve performance in various biomedical relation extraction processes, for example, Drug-Target interaction and protein-protein interaction, particularly when only limited labeled data are available in such tasks. To our best knowledge, this is the first semi-supervised GANs-based method to predict binding affinity.

Keywords: drug-target affinity prediction, deep learning, semi-supervised, generative adversarial networks, convolutional neural networks

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Hao Lin,
University of Electronic Science and Technology of China, China
Rubo Zhang,
Dalian Nationalities University, China

****Correspondence:***
Lingling Zhao
zhaoll@hit.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 August 2019*
Accepted: *11 November 2019*
Published: *09 January 2020*

Citation:
Zhao L, Wang J, Pang L, Liu Y and Zhang J (2020) GANsDTA: Predicting Drug-Target Binding Affinity Using GANs. Front. Genet. 10:1243. doi: 10.3389/fgene.2019.01243

INTRODUCTION

A basic task in the field of new drug design and development is to model the interaction between known drugs and target proteins and to identify drugs with a high affinity for specific disease proteins (Cheng et al., 2018a; Cheng et al., 2019b). However, this is a rather challenging and expensive process even when only approximately 97M compounds reported by the PubChem database (Bolton et al., 2008) and 12K drug entries reported by the DrugBank (Wishart et al., 2006 are considered. Computational methods, especially machine learning models, can considerably accelerate the drug development process and save costs by guiding biological experiments.

Drug-target interaction (DTI) prediction (Yamanishi et al., 2010; Liu et al., 2016; Nascimento et al., 2016; Keum and Nam, 2017) was modeled as a binary classification problem and solved by a few traditional machine learning methods in recent decades. These methods have achieved remarkable performancehowever, they still exhibit limitations because of their strong dependence on handcrafted features.

Apart from predicting DTI, the drug-target binding afi- nity (DTA)(Pahikkala et al., 2014; He et al., 2017) attracts more interest as it can indicate the strength of the interaction between a DT pair. Therefore, predicting DTA can considerably benefit drug discovery, because the searching space would be narrowed down by pruning those DT pairs with low binding affinity scores. Kronecker regularized least squares (KronRLS) Pahikkala et al. (2014) and boosting machines (SimBoost) He et al. (2017) are two state-of-the-art methods for both DTI and DTA prediction. KronRLS is a similarity-based method and can predict the interaction by evaluating the structure similarity among compounds and targets. On the contrary, SimBoost utilizes a gradient boosting machine and belongs to feature-based methods; its feature involves similarity matrices of the drugs and those of targets He et al. (2017). The similarity-based methods (Cheng et al., 2018b) generally rely on similarities to predict the interaction of DT, which inevitably leads to bias. For the feature-based methods, more information regarding the DT are involved; but expert knowledge and feature engineering are also required to construct appropriate features.

Deep learning can represent and recognize the hidden patterns in the data well, therefore, deep-learning based methods have been proposed to predict DTI or DTA utilizing deep neural networks (DNN) (Peng-Wei et al., 2016; Tian et al., 2016; Hamanaka et al., 2017), convolutional neural networks(CNN), (Jastrzebski et al., 2016; Gomez-Bombarelli et al., 2018) recurrent neural networks (RNNs) and stacked-autoencoders based architectures. These methods facilitate the learning of the 3D structures provided and the bimolecular interaction mechanism. However, on one hand, this indeed improves the prediction as more important structural information is exploited, on the other hand, when the 3D structure is the input, these methods depend considerably on the availability of the known 3D structure of the protein-ligand complex.

Another deep-learning based method, called DeepDTA, was implemented to predict the binding affinities with CNN using only 1D representation, that is, the sequences of the proteins and simplified molecular input line entry system(SMILES)of the compounds. In DeepDTA, two CNN blocks are employed as feature extractors, and a fully connected layer receives the output of the CNN blocks and outputs the final prediction results. DeepDTA utilizes the strong representation of CNN, while avoiding the dependence on the 3D structure information, which results in remarkable performance over the other traditional machine learning methods. However, similar to all the state-of-the-art methods for DTA prediction, DeepDTA is also primarily based on supervised machine learning with known labels information. It is known that creating large sets of training data is prohibitively expensive and laborious, particularly in biomedicine, as domain knowledge is required.

An unsupervised learning method, generative adversarial networks(GANs), devised by Goodfellow et al. in 2014 (Goodfellow et al., 2014) may address the challenge. The GANs architecture is characterized by two differentiable functions that play different roles in refining the system. One differentiable function is known as a generator and the other as a discriminator. The generator learns to produce data from a learned probability distribution. The discriminator determines if the produced data is valid by determining if the input comes from the generator or from the actual data set. GANs and its variants have achieved great success in many applications such as computer vision and natural language processing. Additionally, GANs are more attractive as they can learn representations by reusing parts of the generator and discriminator networks as feature extractors, which can be widely applied in many supervised classification or prediction tasks. On the other hand, there also exist some problems in GANs, for example, the better the discriminator is, the more serious the gradient of the generator disappears; the adversarial network may cause the collapse of the model during training, this also brings inconvenience in the practical application. In order to solve these problems, researchers continue to push forward new improvement methods, including least squares GAN(LSGAN) Mao et al. (2017), Wasserstein GAN(WGAN) Arjovsky et al. (2017) conditional GAN(CGAN) Mirza and Osindero (2014), information maximizing GAN(infoGAN) Chen et al. (2016), energy-based GAN(EBGAN) Zhao and Mathieu. (2016), boundary-seeking GAN(BEGAN) Hjelm R D (2017) and so on.

Owing to the unsupervised characteristics of GANs, in this paper, we propose a GANs-based method to predict binding affinity, called GANsDTA for short. Our method comprises two types of networks, two partial GANs for the feature extraction from the raw protein sequences and SMILES strings separately and a regression network using convolutional neural networks for prediction. The contributions of this paper mainly include: We proposed a semi-supervised framework for DTA prediction; we adopted GAN to extract features of protein sequence and compound SMILES in an unsupervised way. Therefore, the proposed model can accommodate unlabeled data for the training as feature extractor using GANs does not require labeled data. This semi-supervised mechanism enables more datasets even without labels available for our model to learn proteins drugs features, leading to better feature representation and prediction performance accordingly. To our best knowledge, this is the first semi-supervised GAN-based method to predict binding affinity. Our results suggest that utilizing such unlabeled data can considerably help improve performance in various biomedical relation extraction processes, particularly when only limited labeled data (e.g. 2000 samples or less) is available in such tasks.

MATERIALS AND METHODS

Data Sets

We evaluated our proposed method using two benchmark data sets, the Davis et al. (2011) and KIBA data set (Tang et al., 2014). **Table 1** and **Figure 1** provides the statistics of these two datasets.

TABLE 1 | Data set.

	Proteins	Compounds	Interactions
Davis	442	68	30056
KIBA	229	2111	118254

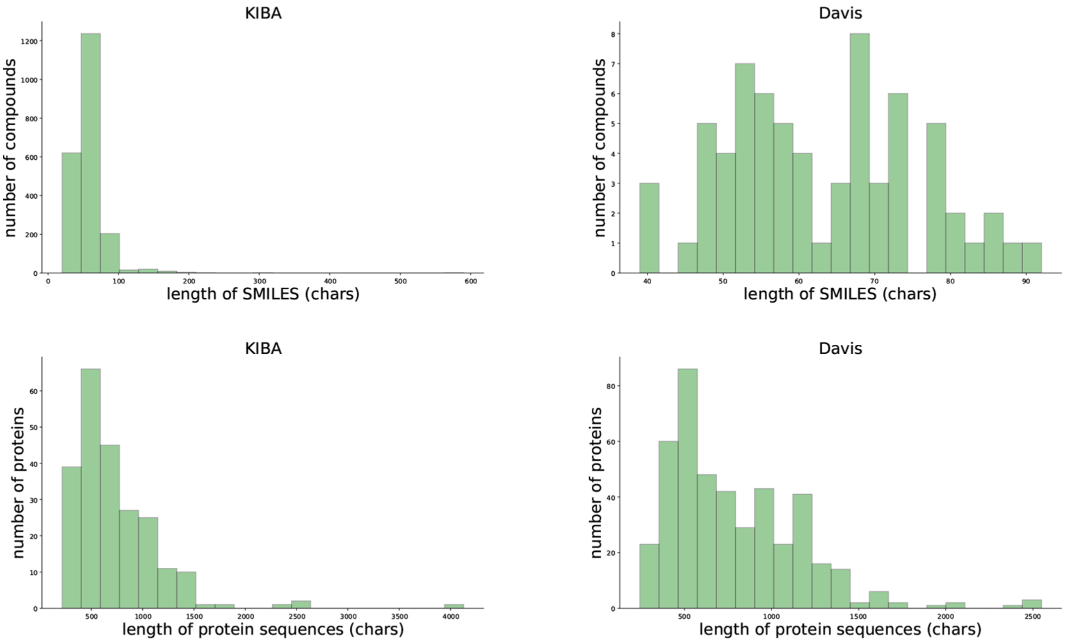


FIGURE 1 | Summary of the KIBA (left panel) and Davis (right panel) data sets.

Proposed Method

Overview of our Approach

Figure 2 provides an overview of the entire pipeline for our method for drug-target binding affinity prediction. Our approach comprises three elements: two feature extractors for protein sequence and compound, respectively, and a regressor for affinity value prediction. Each feature extractor is composed of a feature representation modular from GANs while the regressor is made up of a CNN. A two-round training pattern is employed. In the first training round, the feature extractors are trained in the context of GANs. First, fake samples are generated according to a given noise distribution by the generator of GANs, and then all the fake samples from the generator and the real samples from the available data sets are inputted to the discriminator network. In order to learn to distinguish real and fake sequences of proteins and SIMILES of compounds, the discriminator maps the input into a feature space by a local feature extractor, which promotes the sample classification. Thus, after the training of the whole GANs, a local feature extractor is obtained from the discriminator that can represent the characteristic of the input protein sequence or SMILE sequence. This trained local feature extractor is utilized as the feature representation of the proposed framework, followed by a regressor or classifier for prediction or classification task respectively. Finally, during the second round of training, with the labeled data (SIMILES and protein sequence) and fixed GANs-based feature extractor, the regressor is trained to minimize the loss function, leading to the optimal model parameters.

In the proposed method, the input proteins and drugs are treated as sequence representations. In particular, drugs are represented as SMILES strings – describing the chemical structure in short ASCII strings, and similarly, protein sequences are represented as a string of ASCII letters, which are the amino acids. Having the inputs as strings of text, the discriminator can learn the latent features of those sequences.

Feature Extracting Model

Goodfellow et al. (Goodfellow et al. (2014)) proposed a framework using a minimax game to train deep generative models, so called GANs. The GANs comprise two parts, a generator G and a discriminator D. The generator network G generates fake samples from the generator distribution P_G by transforming a noise variable $z{\sim}P_{noise}(z)$ into a sample $G(z)$. The discriminators are to differentiate these generated samples following distribution P_G from the true sample distribution P_{data}. G and D are trained by playing against each other which can be formulated by a minimax game as follows:

$$\begin{aligned}&\min_{G}\max_{D} V(D,G)\\&= \mathbb{E}_{x\sim P_{data}}[\log(x)] + \mathbb{E}_{z\sim Pnoise}[\log(1 - D(G(z)))] \qquad (1)\end{aligned}$$

Meanwhile, for a given generator G, the optimal discriminator is $D(x) = P_{data}(x)/\ (P_{data}(x)+P_G(x))$.

The GANs employed in our framework is depicted in **Figure 3** — in which the generator network is a four-layer fully connected network and considers a noise vector as input — and produce a sequence of proteins or SMILES. The

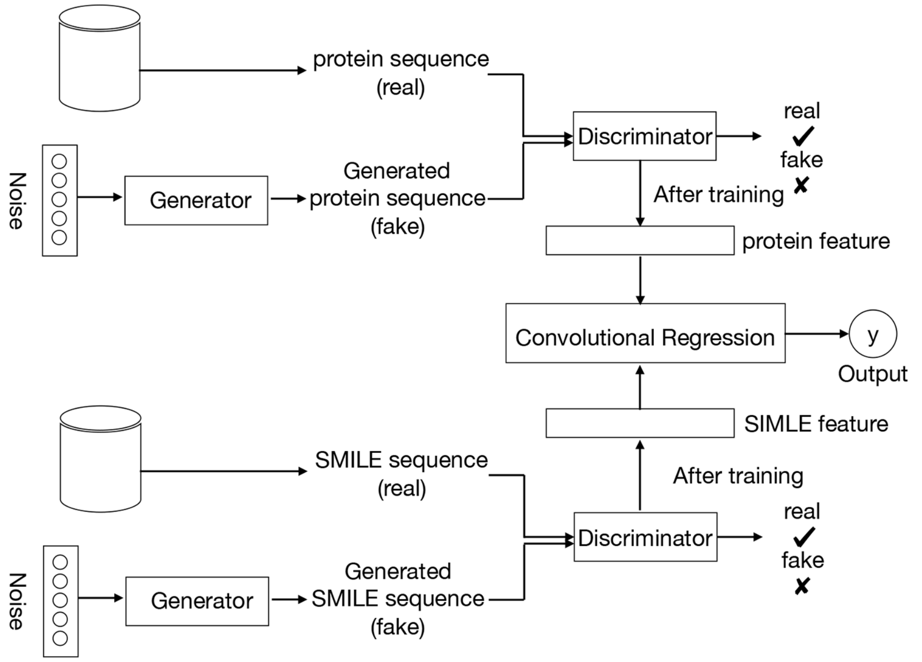


FIGURE 2 | Pipeline overview. We train the GANs on the unlabeled data set. Compound SMILES and protein sequences are encoded and two independent GANs are applied to generate the fake samples. The trained discriminator of the GANs can then be used to project the labeled data sets into a feature latent space. Based on this feature, we train a convolutional regression to predict the DT binding affinity.

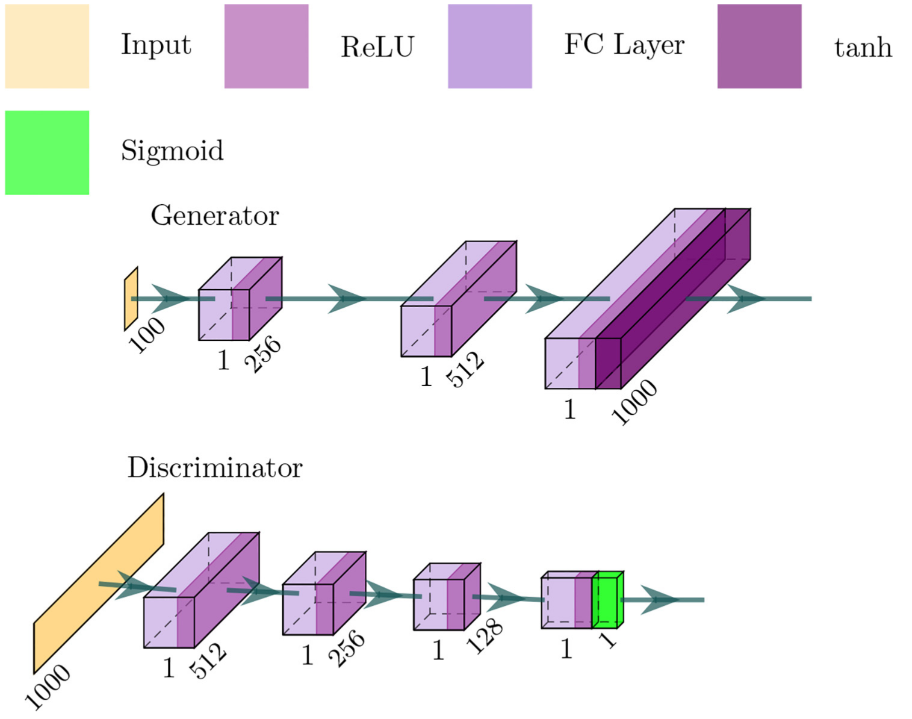


FIGURE 3 | Architecture of the generator and discriminator networks in the proposed method.

discriminator network is a three-layer fully connected network and the output is a probability value between 0 and 1, where 1 means that the input is real and 0 means that the input is fake.

Typically, the discriminator network can be decomposed into a feature extractor $F(\cdot;\varphi_f)$ and a sigmoid classification layer with weight vector ψ_l. Mathematically, given an input sequence s, we have

$$D(s) = \text{sigmoid}\ (\phi_l^T F(s;\phi_f)) = \text{sigmoid}\ (\phi_l^T f) \tag{2}$$

where $\phi = (\phi_f \phi_l)$ and $sigmoid(z) = 1/\ (1+e^{-z})$. $f = F\ (s;\phi_f)$ is the feature extractor of s in the last layer of D, which is to be leaked to the regression model.

Regression Model

To predict the binding affinity, we combine the intermediate features learned by the two GANs and then apply a few 1D convolution layers to learn the final regression output. The convolution regression model conducts convolution operations with the kernel size of 4 to acquire feature maps of the input information. The dimension of the first convolution layer is 16×4. All the convolution layers are connected to activation functions (ReLU function). The dimensions of the second and third, convolution layers are 32×4, and 48×4. The activation function of the output layer is a linear function (identity function, i.e., $y = x$) that obtains a continuous value. This network is trained by minimizing the loss function defined by the mean square error (MSE) between the outputs p of this network and depth values y included in the dataset:

$$MSE = \frac{1}{n}\sum_{k=1}^{n}(p_k - y_k)^2 \tag{3}$$

EXPERIMENTS AND RESULTS

We compared our proposed method with the state-of-the-art DTA prediction models using the Davis and Kiba datasets. For these two datasets, we used the same setting as DeepDTA, that is, 80% of data were split as training samples and 20% as testing samples. In addition, our model is trained by both the labeled and unlabeled instances. We apply the Adam optimizer with the initial learning rate of 0.0001 to optimize the parameters of the model. We manually tuned the hyperparameters based on the testing results on the validation set. The performance of the proposed model was measured by calculating the concordance index (CI) and mean squared error (MSE) metrics. CI evaluates the ranking performance of the models that output continuous values.

$$CI = \frac{1}{Z}\sum_{\delta_x > \delta_y} h\ (b_x - b_y) \tag{4}$$

where b_x is the prediction value for the larger affinity δ_x, b_y is the prediction value for the smaller affinity δ_y, Z is a normalization constant, and $h(m)$ is the step function.

$$h(m) = \begin{cases} 1; & if m > 0 \\ 0.5; & if m = 0 \\ 0; & if m < 0 \end{cases} \tag{5}$$

MSE is a common measure to quantify the difference between the predicted values p and the actual values, which is defined as follows:

We compared the predicted performance of our method with DeepDTA and two machine-learning-based KronRLS and SimBoost method. Both of our work and DeepDTA only utilize the information of protein sequence and SMILES of the compounds. The difference is that our method can extract features of proteins and compounds in an unsupervised manner. **Tables 2** and **3** present the MSE and CI values for different methods for Davis and KIBA datasets.

For the Davis dataset (**Table 2**), even the DeepDTA, with Simith–Waterman as the protein's representation form and drugs in the 1D strings, achieves the best CI score (0.886), slightly higher than our method - its MSE metric is much higher than our methods. Whereas another DeepDTA, CNN for protein and compound representation, achieves the best MSE with 0.261 as well as the lower CI than our method.

A similar performance is observed for the Kiba dataset (**Table 3**). In particular, DeepDTA is the best baseline in both measures, CI, at 0.863, and MSE, at 0.194, when both drugs and proteins are represented as 'words'. Regarding CI, the proposed GANsDTA exhibits a slight improvement. The best CI GANsDTA gained is 0.866.

To provide a better assessment of our model, we determined the performances of GANsDTA, DeepDTA with two CNN modules and two baseline methods with two different metrics: r_m^2 index and area under precision recall (AUPR) score as well. r_m^2

TABLE 2 | CI and MSE scores for the Davis dataset on the independent test for our method and other methods.

Method	Protein rep.	Compound rep.	CI	MSE
DeepDTA	Smith-Waterman	Pubchem-Sim	0.790	0.608
DeepDTA	Smith-Waterman	CNN	0.886	0.420
DeepDTA	CNN	Pubchem-Sim	0.835	0.419
DeepDTA	CNN	Pubchem-Sim	0.878	0.261
KronRLS	Smith-Waterman	Pubchem-Sim	0.871	0.379
SimBoost	Smith-Waterman	Pubchem-Sim	0.872	0.282
GANsDTA	GAN	GAN	**0.881**	0.276

Bolded texts mean the best results.

TABLE 3 | CI and MSE scores for the Kiba dataset on the independent test.

Method	Protein rep.	Compound rep.	CI	MSE
DeepDTA	Smith-Waterman	Pubchem-Sim	0.710	0.502
DeepDTA	Smith-Waterman	CNN	0.854	0.204
DeepDTA	CNN	Pubchem-Sim	0.718	0.571
DeepDTA	CNN	CNN	0.863	0.194
KronRLS	Smith-Waterman	Pubchem-Sim	0.782	0.411
SimBoost	Smith-Waterman	Pubchem-Sim	0.836	0.222
GANsDTA	GAN	GAN	**0.866**	0.224

Bolded texts mean the best results.

TABLE 4 | r_m^2 index and AUPR score for the Davis dataset."4 r_m^2 index and AUPR score for the Davis dataset."

Method	Protein rep.	Compound rep.	r_m^2	AUPR
DeepDTA	CNN	CNN	0.630	0.714
KronRLS	Smith-Waterman	Pubchem-Sim	0.407	0.661
SimBoost	Smith-Waterman	Pubchem-Sim	0.644	0.709
GANsDTA	GAN	GAN	0.653	0.691

TABLE 5 | The r_m^2 index and AUPR score for the KIBA dataset.

Method	Protein rep.	Compound rep.	r_m^2	AUPR
DeepDTA	CNN	CNN	0.673	0.788
KronRLS	Smith-Waterman	Pubchem-Sim	0.342	0.635
SimBoost	Smith-Waterman	Pubchem-Sim	0.629	0.760
GANsDTA	GAN	GAN	0.675	0.753

index is a metric which defines the possibility of an acceptable model. Generally, if the value of r_m^2 the index is greater than 0.5 on a test set, we consider this model to be acceptable. The metric is described in equation (6) where r^2 and r^0 are the squared correlation coefficients with and without intercept, respectively. The details of the formulation are explained in Pratim Roy et al. (2009); Roy et al. (2013).

$$r_m^2 = r^2 * \ (1 - \sqrt{r^2 - r_0^2}) \qquad (6)$$

The AUPR score is generally adopted for binary prediction. To measure AUPR based performances, the Davis and KIBA datasets should be converted into their binary forms *via* thresholding. For the Davis dataset we selected a pKd value of 7 as the threshold, while for KIBA dataset the threshold is 12.1, which is same as in the literature Öztürk et al. (2018).

Tables 4 and **5** list the r_m^2 index and AUPR score of GANsDTA and three baseline methods on the Davis and KIBA datasets, respectively. The results suggest that SimBoost, DeepDTA and GANsDTA are acceptable models for to predict affinity with result to r_m^2 value.

Figure 4 illustrates the predicted binding affinity values against the actual values for our GANsDTA on the Davis and KIBA datasets. Evidently, an ideal model is expected to enable predictions (p) equal to the measured (y) values. For GANsDTA, it can be observed that the density is high around the $p = y$ line, particularly for the KIBA dataset.

It can be observed that the proposed GANsDTA exhibits a similar performance to DeepDTA from **Tables 2**-**4**. For the Davis dataset, GANsDTA provides a slightly lower CI score (0.881) than the state-of-the-art DeepDTA with CNN the feature extraction (0.886), and a slightly higher MSE with 0.015. The reason is that the training for GANs is insufficient due to the small size of the Davis dataset which only includes 442 proteins, 68 compounds, and 30056 interactions. However, GANsDTA is still the second-best predictor. The other benchmark KIBA dataset includes 229 proteins, 2111 compounds, and 118254 interactions, enabling the GANs to be trained better, leading to better prediction accuracy. This indicates that GANsDTA is more suitable for the prediction task with a large dataset. In the future, more possible datasets (Cheng et al., 2018c; Cheng et al., 2019a) Cheng et al., 2016; Cheng et al., 2019a can be utilized to improve the training of GANsDTA.

CONCLUSION

Predicting drug-target binding affinity is challenging in drug discovery. The supervised-based methods heavily depend on labeled data, which are expensive and difficult to obtain on a large scale. In this paper, we propose a semi-supervised GAN-based method to estimate drug-target binding affinity, while effectively learning useful features from both labeled and unlabeled data. We use GANs to learn representations from

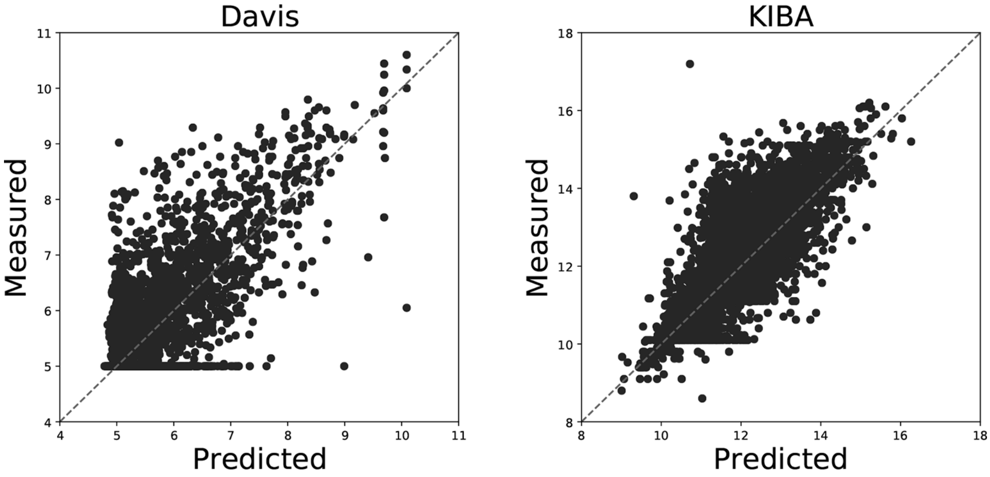


FIGURE 4 | Predictions from DeepDTA model with two CNN blocks against measured (real) binding affinity values for Davis (pKd) and KIBA (KIBA score) datasets.

the raw sequence data of proteins and drugs and convolutional regression when predicting the affinity. We compare the performance of the proposed model with the state-of-art deep-learning-based method as our baseline. By utilizing the unlabeled data, our model can achieve competitive performance while using freely available unlabeled data. However, because it is difficult to train GANs, this approach is not comparative in the scenarios of a small dataset, and the improved techniques for training GANs should be employed to enhance the adaptability of GANs.

DATA AVAILABILITY STATEMENT

The datasets KIBA and Davis for this study can be found in http://www.ebi.ac.uk/biostudies/studies/S-EPMC6129291?xr=true.

AUTHOR CONTRIBUTIONS

LZ, JW, and LP substantially contributed to the conception and design of the study, and acquisition of data. YL analyzed and interpreted the data. LZ, JW, and JZ drafted the article.

FUNDING

This work is supported by the National Natural Science Foundation of China (NSFC, Grant no.61305013 and 61872114).

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments.

REFERENCES

Arjovsky, M., Bottou, L., and Chintala, S. (2017). Wasserstein generative adversarial networks. *In International Conference on Machine Learning (ICML)* 2017.

Bolton, E. E., Thiessen., P. A., Wang, Y., and Bryant, S. H. (2008). Pubchem: integrated platform of small molecules and biological activities. *Annu. Rep. In Comput. Chem.* 4, 217–241. doi: 10.1016/s1574-1400(08)00012-1

Chen, X., Houttooft, R., and Duan, Y. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2172–2180.

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). Oahg: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function. *Bioinformatics* 34, 1953–, 1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018b). Infacront: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19, 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018c). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: A mendelian randomization study. *Front. Genet.* 9, 657 =. doi: 10.3389/fgene.2018.00657

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). 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

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019a). gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 1–7. doi: 10.1093/nar/gkz843

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019b). Metsigdis: a manually curated resource for the metabolic signatures of diseases. *Brief Bioinform.* 20, 203–209. doi: 10.1093/bib/bbx103

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046. doi: 10.1038/nbt.2017

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., et al. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276. doi: 10.1021/acscentsci.7b00572

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* 2672–2680.

Hamanaka, M., Taneishi, K., Iwata, H., Ye, J., Pei, J., Hou, J., et al. (2017). Cgbvs-dnn: Prediction of compound-protein interactions based on deep learning. *Mol. Inf.* 36, 1600045. doi: 10.1002/minf.201600045

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., and Ester, M. (2017). Simboost: a readacross approach for predicting drug-target binding affinities using gradient boosting machines. *J. Cheminf.* 9, 24. doi: 10.1186/s13321-017-0209-z

Hjelm, R. D., Che, T., and Jacob, A. P. (2017). Boundary-seeking generative adversarial networks. *arXiv* preprint arXiv:1702.08431.

Jastrzebski, S., Leśniak, D., and Czarnecki, W. M. (2016). Learning to SMILE(S). In: *International Conference on Learning Representation* (Workshop track).

Keum, J., and Nam, H. (2017). Self-blm: Prediction of drug-target interactions via self-training svm. *PloS One* 12, e0171839. doi: 10.1371/journal.pone.0171839

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016). Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PloS Comput. Biol.* 12, e1004760. doi: 10.1371/journal.pcbi.1004760

Mao, X., Xie, H., and Li, Q. (2017). Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017 pp. 2794–2802. doi: 10.1109/iccv.2017.304

Mirza, M., and Osindero., S. (2014). Conditional generative adversarial nets. *arXiv* preprint arXiv:1411.1784.

Nascimento, A. C., Prudêncio, R. B., and Costa, I. G. (2016). A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinf.* 17, 46. doi: 10.1186/s12859-016-0890-3

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* 34, i821–i829. doi: 10.1093/bioinformatics/bty593

Pahikkala, T., Airola, A., Pietilä, S., Shakyawar, S., Szwajda, A., Tang, J., et al. (2014). Toward more realistic drug-target interaction predictions. *Briefings In Bioinf.* 16, 325–337. doi: 10.1093/bib/bbu010

Peng-Wei, Chan, K. C., You, Z.-H., Chan, K. C. C., You, Z. H., et al. (2016). "Large-scale prediction of drug-target interactions from deep representations," International Joint Conference on Neural Networks (IJCNN) (IEEE), 1236–1243. doi: 10.1109/ijcnn.2016.7727339

Pratim Roy, P., Paul, S., Mitra, I., and Roy, K. (2009). On two novel parameters for validation of predictive qsar models. *Molecules* 14, 1660–, 1701. doi: 10.3390/molecules14051660

Roy, K., Chakraborty, P., Mitra, I., Ojha, P. K., Kar, S., and Das, R. N. (2013). Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: emphasis on scaling of response data. *J. Comput. Chem.* 34, 1071–, 1082. doi: 10.1002/jcc.23231

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inf. Modeling* 54, 735–743. doi: 10.1021/ci400709d

Tian, K., Shao, M., Wang, Y., Guan, J., and Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. *Methods* 110, 64–72. doi: 10.1016/j.ymeth.2016.06.024

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., et al. (2006). Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672. doi: 10.1093/nar/gkj067

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254. doi: 10.1093/bioinformatics/btq176

Zhao J, L. Y., and Mathieu, M. (2016). Energy-based generative adversarial network. *arXiv* preprint arXiv:1609.03126.

Conflict of Interest: The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 January 2020
doi: 10.3389/fgene.2019.01298

TriPCE: A Novel Tri-Clustering Algorithm for Identifying Pan-Cancer Epigenetic Patterns

*Yanglan Gan[1], Ning Li[1], Yongchang Xin[1] and Guobing Zou[2]**

[1] School of Computer Science and Technology, Donghua University, Shanghai, China, [2] School of Computer Engineering and Science, Shanghai University, Shanghai, China

Epigenetic alteration is a fundamental characteristic of nearly all human cancers. Tumor cells not only harbor genetic alterations, but also are regulated by diverse epigenetic modifications. Identification of epigenetic similarities across different cancer types is beneficial for the discovery of treatments that can be extended to different cancers. Nowadays, abundant epigenetic modification profiles have provided a great opportunity to achieve this goal. Here, we proposed a new approach TriPCE, introducing tri-clustering strategy to integrative pan-cancer epigenomic analysis. The method is able to identify coherent patterns of various epigenetic modifications across different cancer types. To validate its capability, we applied the proposed TriPCE to analyze six important epigenetic marks among seven cancer types, and identified significant cross-cancer epigenetic similarities. These results suggest that specific epigenetic patterns indeed exist among these investigated cancers. Furthermore, the gene functional analysis performed on the associated gene sets demonstrates strong relevance with cancer development and reveals consistent risk tendency among these investigated cancer types.

Keywords: epigenetic analysis, pattern discovery, tri-clustering, FP-growth algorithm, pan-cancer

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University,
China

Reviewed by:
Hui Liu,
Changzhou University,
China
Tianfan Fu,
Georgia Institute of Technology,
United States
Qingting Wei,
Nanchang University,
China

Correspondence:
Guobing Zou
gbzou@shu.edu.cn

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *24 August 2019*
Accepted: *25 November 2019*
Published: *15 January 2020*

Citation:
Gan Y, Li N, Xin Y and Zou G (2020)
TriPCE: A Novel Tri-Clustering
Algorithm for Identifying Pan-Cancer
Epigenetic Patterns.
Front. Genet. 10:1298.
doi: 10.3389/fgene.2019.01298

INTRODUCTION

Cancer genetics and epigenetics are closely linked in driving the cancer phenotype (Bailey et al., 2018). The vast majority of human cancers emerge from a gradual accumulation of somatic alterations and epigenetic abnormalities, which together lead to the malignant growth (Jones et al., 2016). Epigenetic changes can further enable tumor cells to escape from host immune surveillance and various treatments (You and Jones, 2012). Epigenetic abnormalities are usually observed as disrupted DNA methylation patterns (Chiappinelli et al., 2015), abnormal histone post translational modifications (Sawan and Herceg, 2010), and aberrant changes in chromatin organization (Allis and Jenuwein, 2016). How to identify epigenetic modification patterns that lead to the corresponding dysregulation in diverse cancers has become a critical research issue of cancer studies (Dawson, 2017; Kelly and Issa, 2017).

Great advancements have been made in delineating the underlying mechanisms of human cancers (Lawrence et al., 2014; Martincorena and Campbell, 2015). Extensive research has centered on the genetic aspect of cancers, such as how mutational activation and inactivation of cancer genes influence the cellular pathways (Vogelstein et al., 2013; Waddell et al., 2015). Recently, an increasing emphasis of drug discovery efforts has been targeting on the cancer epigenome (Flavahan et al., 2017). Many epigenome mapping projects have been gradually founded. The Cancer Genome Atlas Network (TCGA), BLUEPRINT, and the International Cancer Genome Consortium (ICGC) define the genome-wide distribution of epigenetic marks in many normal and cancerous tissues (Beck et al., 2012; Kundaje et al., 2015; Weinstein et al., 2015). Given the genome-wide distribution of epigenetic modifications of different cancers, it is urgent to decipher common epigenetic patterns across cancers and to understand the underlying mechanisms of tumorigenesis. Key epigenomic similarities shared by different cancer types would present an important opportunity to design effective cancer treatment strategies among cancers regardless of tissue or organ and enable the extension of effective treatments from one cancer type to another (Karlic et al., 2010; Gan et al., 2018).

To detect significant epigenetic patterns, existing computational methods mainly focus on identifying combinatorial states of different epigenetic marks. Specifically, CoSBI captures diverse histone modification patterns based on the correlations of different histone signals (Ucar et al., 2011). ChromHMM and HiHMM both apply a HMM model to annotate genomic sequences by the co-occurrence of multiple epigenetic marks (Ernst et al., 2011; Sohn et al., 2015). RFECS is developed mainly based on random forests (Rajagopal et al., 2013). IDEAS is able to jointly characterize epigenetic landscapes in many cell types and detect differential regulatory regions (Zhang et al., 2016). These methods have successfully identified the combinatorial epigenetic pattern in specific cell type. However, the relations among different cancer types still need to be investigated. Because DNA methylation in cancers has been addressed elsewhere (Kretzmer et al., 2015; Yang et al., 2016), here we only focus on the critical covalent histone modifications that are altered in various cancers, particularly the well-studied acetylation and methylation modifications.

In this paper, we proposed a tri-clustering approach, named TriPCE, for integrative pan-cancer epigenomic analysis. The method TriPCE adopts a tri-clustering strategy to identify the coherent patterns of various epigenetic modifications across different cancer types. We applied TriPCE to investigate six critical epigenetic marks among seven cancer types, and identified significant pan-cancer epigenetic modification patterns. The results reveal that there exists consistent epigenetic modification tendency among these cancer types. Meanwhile, the gene function analysis demonstrates that these associated genes are strongly relevant with the cancer cellular pathway.

MATERIALS AND METHODS

Datasets

To detect epigenetic similarities among different cancers, we analyzed the epigenome maps of seven cancer types, including A549, K562, HepG2, HCT116, Hela-S3, multiple myeloma-Cell Line, and sporadic Burkitt lymphoma-Cell Line. For the epigenetic marks, we first filtered out those marks that are not included in these seven cancer types, and then focused on six widely studied ones, including H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3. Meanwhile, the RNA expression profiles of these cancers were also collected. Totally, we obtained 42 epigenome maps and 7 RNA expression profiles for these cancers. The datasets were downloaded from the website of NIH Roadmap Epigenome Project.

General Scheme of the TriPCE Approach

We developed a tri-clustering approach TriPCE to dissect the pan-cancer epigenetic pattern. The method not only explicitly detects combinatorial states of various epigenetic marks in different genomic segments, but also mines similar epigenetic patterns across different cancer types. The proposed TriPCE model has three key components, as shown in **Figure 1**. Firstly, preprocess the modification data of various epigenetic marks in different cancer types. Secondly, identify bi-Clusters based on FP-growth algorithm for each epigenetic mark. Thirdly, mine tri-Clusters with coherent epigenetic modification patterns across different cancer types.

Step 1. Preprocess the epigenetic modification data of different cancer types. Firstly, the genome was divided into consecutive genomic segments, with a typical segment size of 200 bps (Gan et al., 2017). For each epigenetic modification map, we computed the summary tag count of every segment. Then, each segment is associated with the intensities of a set of epigenetic modifications in each cancer type. To deduce the impact of the noise resulting from spurious tag counts in the ChIP-seq experiments, raw sequence read counts of each epigenetic modification were further normalized by the total number of reads followed by arcsine transformation (Pinello et al., 2014). Finally, according to the genome annotation data, the epigenetic distribution in the promoter regions was extracted.

After the preprocessing step, we gained six epigenetic profiles of seven cancer types along the promoter regions. Let $G = \{g_1, g_2, \ldots, g_n\}$ be a set of n genes, let $T = \{t_1, t_2, \ldots, t_7\}$ be the investigated seven cancer types and let $E = \{e_1, e_2, \ldots, e_6\}$ be the six epigenetic marks. For each epigenetic mark, the epigenetic profiles of different cancer types in the promoter regions of these genes are organized as a matrix $D_k = T \times G = t_{i,j}^k$ (with $i \in [1,2\ldots,7]$, $j \in [1,2\ldots, n]$, $k \in [1,2\ldots,6]$), where rows correspond to the cancer types, and columns correspond to those genes, respectively. Each entry $t_{i,j}^k$ is a vector representing the epigenetic profile of e_k in the ith cancer along the promoter region of gene j.

Step 2. Identify bi-clusters based on FP-growth algorithm for each epigenetic mark. Given the preprocessed and reorganized epigenetic modification data matrix of each epigenetic mark, we first computed the Pearson correlation coefficients between the epigenetic profiles of any two cancer types at every promoter region, and then obtained a correlation coefficient matrix.

Specifically, for the promoter region g_i, we computed the Pearson correlation coefficients among the epigenetic modification distribution vectors of any different cancer types. If the calculated correlation coefficient is higher than a given threshold, the epigenetic modification trend in these two cancer

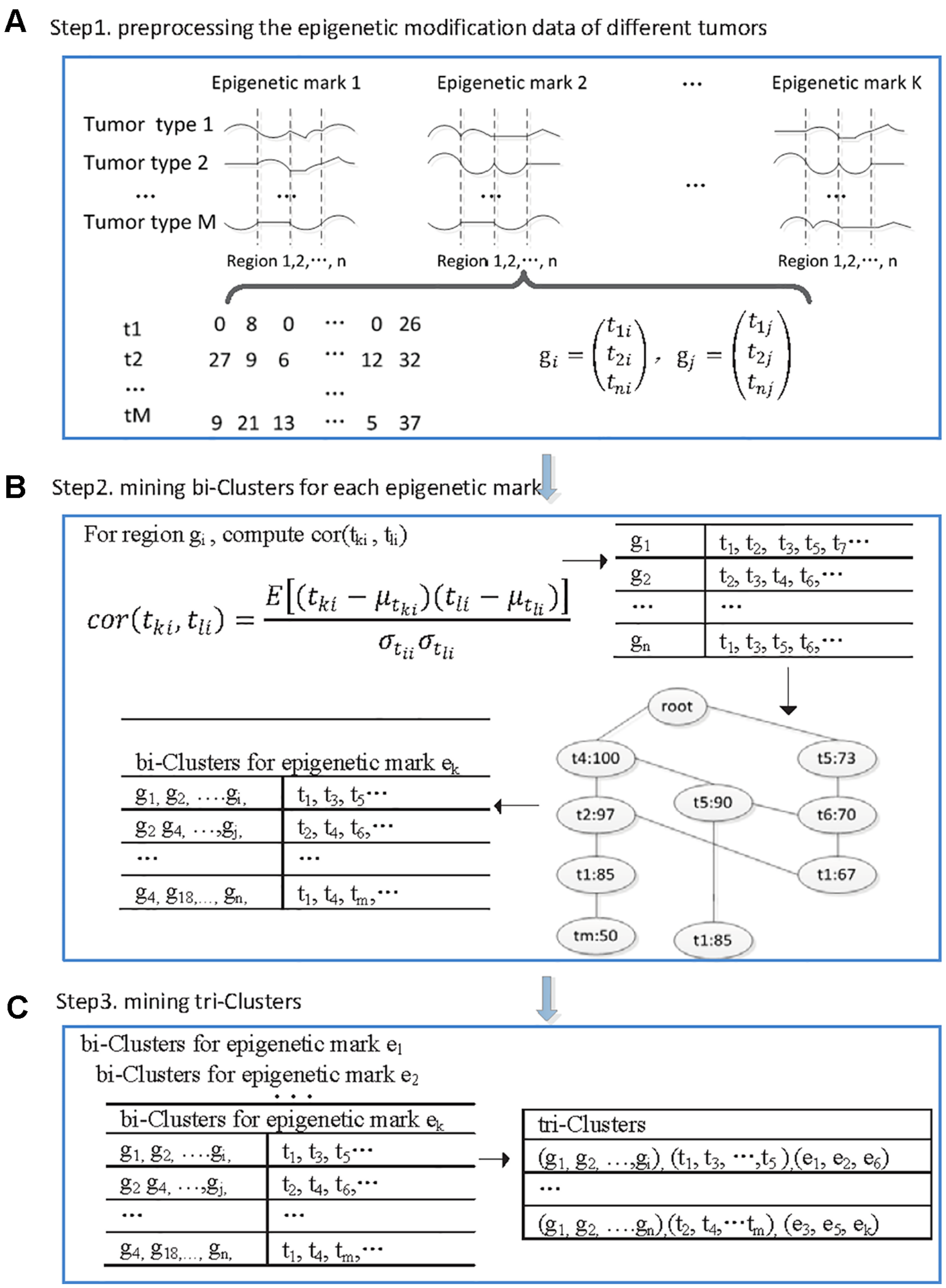

A
Step1. preprocessing the epigenetic modification data of different tumors
Epigenetic mark 1
Epigenetic mark 2
…
Epigenetic mark K
Tumor type 1
Tumor type 2
…
Tumor type M
Region 1,2,⋯, n
Region 1,2,⋯, n
Region 1,2,⋯, n
t1 0 8 0 ⋯ 0 26
t2 27 9 6 ⋯ 12 32
⋯ ⋯
tM 9 21 13 ⋯ 5 37
$g_i = \begin{pmatrix} t_{1i} \\ t_{2i} \\ t_{ni} \end{pmatrix}, \; g_j = \begin{pmatrix} t_{1j} \\ t_{2j} \\ t_{nj} \end{pmatrix}$
B
Step2. mining bi-Clusters for each epigenetic mark
For region g_i , compute cor(t_{ki} , t_{li})
$cor(t_{ki}, t_{li}) = \frac{E[(t_{ki} - \mu_{t_{ki}})(t_{li} - \mu_{t_{li}})]}{\sigma_{t_{ii}}\sigma_{t_{li}}}$
g_1 | t_1, t_2, t_3, t_5, t_7⋯
g_2 | t_2, t_3, t_4, t_6,⋯
⋯ | ⋯
g_n | t_1, t_3, t_5, t_6,⋯
root
t4:100
t5:73
t2:97
t5:90
t6:70
t1:85
t1:67
tm:50
t1:85
bi-Clusters for epigenetic mark e_k
$g_1, g_2, \dots g_i$, | t_1, t_3, t_5⋯
$g_2\ g_4, \dots, g_j$, | t_2, t_4, t_6,⋯
⋯ | ⋯
$g_4, g_{18}, \dots, g_n$, | t_1, t_4, t_m,⋯
C
Step3. mining tri-Clusters
bi-Clusters for epigenetic mark e_1
bi-Clusters for epigenetic mark e_2
• • •
bi-Clusters for epigenetic mark e_k
$g_1, g_2, \dots g_i$, | t_1, t_3, t_5⋯
$g_2\ g_4, \dots, g_j$, | t_2, t_4, t_6,⋯
⋯ | ⋯
$g_4, g_{18}, \dots, g_n$, | t_1, t_4, t_m,⋯
tri-Clusters
$(g_1, g_2, \dots, g_i), (t_1, t_3, \cdots, t_5), (e_1, e_2, e_6)$
⋯
$(g_1, g_2, \dots g_n), (t_2, t_4, \cdots t_m), (e_3, e_5, e_k)$

FIGURE 1 | The flowchart of the proposed TriPCE approach. **(A)** Preprocessing the epigenetic modification data of different cancer types. **(B)** For each epigenetic mark, identifying bi-Clusters based on the FP-growth algorithm. **(C)** Mining tri-Clusters with coherent epigenetic modification patterns across different cancer types.

types is regarded as coherent in this promoter region. Then, we added this cancer type to the corresponding itemset, which contains all the cancer types exhibiting similar epigenetic patterns in this region. Based on extensive experimental comparison, when the correlation coefficient threshold is set as 0.7, the identified epigenetic patterns are obviously coherent. For each epigenetic mark, we respectively constructed the corresponding similar itemsets for all promoter regions.

Based on the resulted itemset, we further identified the significant coherent epigenetic patterns using FP-growth algorithm (Han et al., 2004). FP-growth algorithm is a data mining method that was originally developed for frequent itemset mining in market basket analysis. Here, we adopted the FP-tree model to represent in a compact way all the cancer types with similar epigenetic patterns in different promoter regions. Then, it can be used to mine potential frequent itemsets and filter out most of the unrelated data. In this context, a typical frequent itemset represents a group of cancer types that share similar epigenetic patterns in abundant promoter regions. To gain the significant epigenetic states, we set the minimum support of genes as 10% of the investigated genes. For each frequent itemset, we then inversely identified the corresponding gene set and gained the bi-Cluster. The resulted bi-Cluster is in the form ("genomic regions," "cancer types"), representing the cancer types exhibit similar epigenetic patterns in these genes. Similarly, we obtained the corresponding bi-Cluster sets for all investigated epigenetic marks.

Step 3. Mine tri-Clusters with coherent epigenetic modification patterns across different cancer types. After obtaining the bi-Cluster sets for each epigenetic mark, we further mined the tri-Clusters. By enumerating the maximum subsets of different epigenetic marks, we obtained the tri-Clusters. In detail, we respectively computed the intersection of the bi-Cluster sets from two epigenetic marks e_k and e_l, which are kept with the epigenetic marks to get possible tri-Clusters. Further, by filtering out the candidates with the support lower than the predefined minimum support, we obtained the significant tri-Clusters. Iteratively, we continued the process with another epigenetic mark until all the epigenetic marks were analyzed. We tried all such paths and kept the maximal tri-Clusters only. Each tri-Cluster is represented as ("genomic regions," "cancer types," "epigenetic marks"), listing a gene set with similar trend of epigenetic modifications in different cancer types. The resulted tri-Clusters indicate that the conserved epigenetic signatures in these genomic regions are shared by multiple cancer types.

Functional Analysis of the Genes

From the identified tri-Clusters, we can obtain the gene sets associated with specific coherent epigenetic patterns. To investigate the potential functions of these genes, we performed the gene ontology (GO) enrichment analysis and pathway enrichment analysis *via* DAVID bioinformatics resources (Huang et al., 2007). The significant enrichment lists were obtained with P-value < 0.005.

RESULTS

Identifying Similar Epigenetic Patterns Across Different Cancer Types

We developed a tri-clustering approach, TriPCE, to capture similar epigenetic patterns among different cancer types. TriPCE was applied to the genome-wide epigenetic modification maps of seven cancer types, including A549, K562, HepG2, HCT116, Hela-S3, multiple myeloma-Cell Line, and sporadic Burkitt lymphoma-Cell Line. For each epigenetic mark, TriPCE first groups the promoter regions based on the epigenetic modification profiles among different cancer types. **Figure 2** shows a typical bi-Cluster of epigenetic mark H3K4me1, which contains abundant genes with similar modification pattern in four cancer types, including Hela-S3, HepG2, K562, and A549. From this figure, we observe that the epigenetic profiles of these genes are similar in these cancer types.

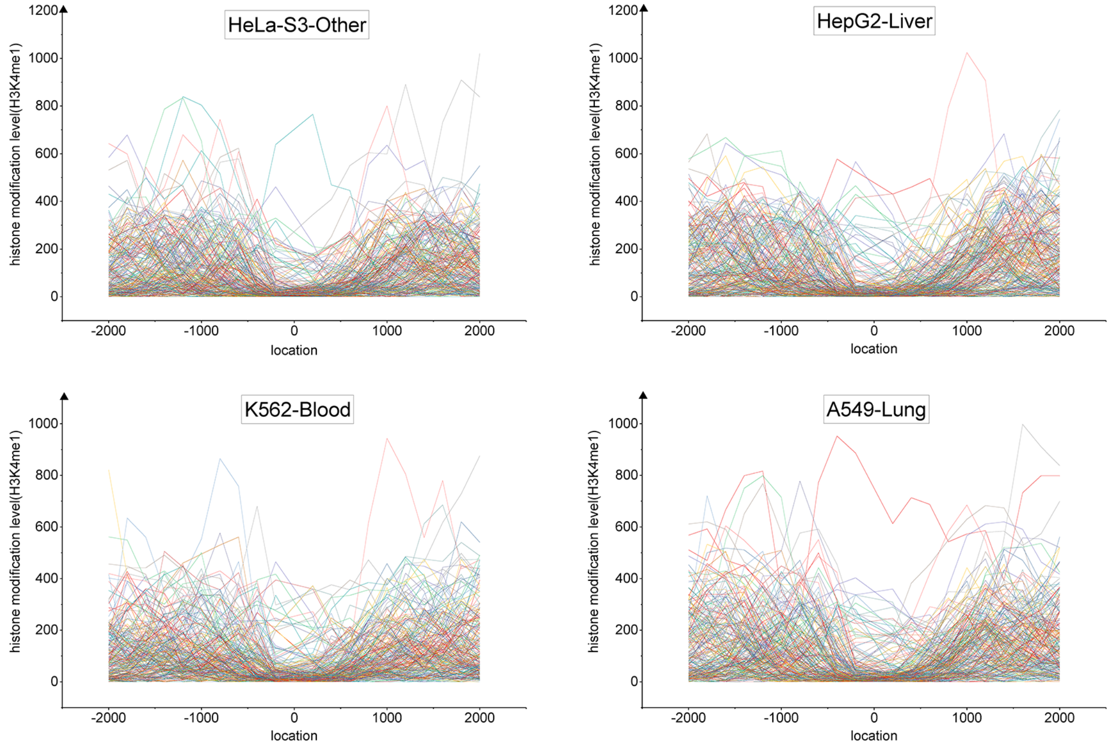


FIGURE 2 | The profiles of epigenetic mark H3K4me3 in a typical bi-Cluster exhibit a similar pattern in four cancer types, including Hela-S3, HepG2, K562 and A549.

Then, the epigenetic profile shared by a cluster of promoter regions in multiple cancer types is considered to be an epigenetic pattern. Meanwhile, different cancer types share similar epigenetic patterns. This result is consistent with previous finding that H3K9me3/me2 and H3K36me3/me2 frequently observed in breast cancer (Liu et al., 2009), esophageal cancer (Yang et al., 2000), MALT lymphoma (Vinatzer et al., 2008), and lung sarcomatoid carcinoma (Italiano et al., 2006). Based on the identified bi-Clusters of these investigated epigenetic marks, we noted that cancers (HepG2 and HCT116) are clustered together and share a larger number of epigenetic marks, implying that they share more similar epigenetic regulation mechanisms.

To identify the significant modification patterns, we set the minimal support of genes as 10% of the investigated genes. With diverse correlation coefficient thresholds, we respectively gained different numbers of bi-Clusters for epigenetic marks H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, and H3K27ac, among these cancer types, as shown in **Figure 3**. The comparison indicates that the similarities of these epigenetic marks are quite different. Under different threshold settings, the epigenetic mark H3K4me3 has a relatively small number of bi-Clusters, indicating that its profiles are less conserved and exhibit more variable patterns among these cancer types than other epigenetic marks. On the contrary, there are more similar epigenetic patterns of H3K4me1 and H3K27me3 among different cancer types (Baylin and Jones, 2016). The plasticity of epigenome depends on diverse environmental factors. Thus, it is not surprising that epigenotypes contribute to developmental human disorders and adult diseases (Brien et al., 2016). As the minimal support threshold slightly affects the trend among different epigenetic marks, we chose the bi-Clusters with threshold 0.7 for further analysis.

Identifying Coherent Patterns Among Different Epigenetic Marks

From the above results, we notice that there are obvious differences among the investigated epigenetic modifications. To identify the conserved epigenetic states and explore the similar patterns of these epigenetic modifications, we further clustered these epigenetic marks based on the detected bi-Clusters. By systematically computing the intersection of the bi-Cluster sets from different epigenetic marks, we kept the tri-Clusters with the support higher than the predefined minimum support. The identified tri-Clusters are represented as triples ("genomic regions," "cancer types," "epigenetic marks"). Each tri-Cluster represents that the promoter region of these genes exhibits similar epigenetic modification patterns in the related cancer types.

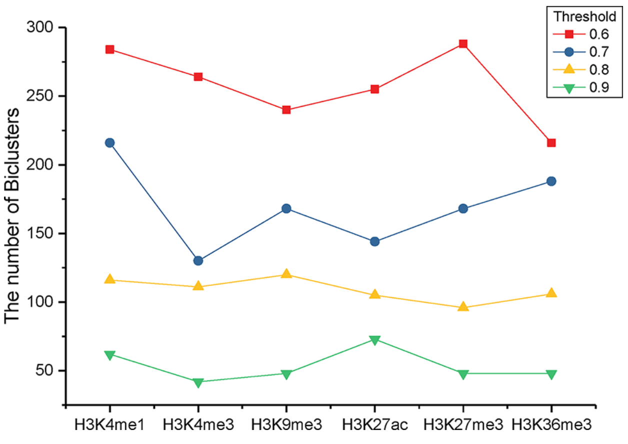


FIGURE 3 | The numbers of bi-Clusters with varied similarity thresholds for different epigenetic marks.

Applying TriPCE to the data set, we initially obtained 175 significant tri-Clusters. **Figure 4** shows the information of 15 typical clusters, including the epigenetic marks, the cancer types, and the supports of these tri-Clusters. The results indicate that specific genomic regions indeed share combinatorial epigenetic patterns across different cancer types. For example, the changing pattern of epigenetic modifications (H3K4me3, H3K9me3, H3K27me3, and H3K36me3) are shared by a large number of genes in cancer types A549, HepG2, and K562. On the contrary, some epigenetic modification patterns are only coherent in certain cancer types. Among these resulted clusters, we observe that the similar patterns of H3K36me3, H3K27ac, and H3kK27me3 exist in fewer cancer types, such as HepG2 and sporadic Burkitt lymphoma-Cell Line. Notably, these identified tri-Clusters reveal more information about the epigenetic patterns among these cancer types.

Analyzing the Potential Roles of Associated Genes

Based on the detected tri-Clusters, we further obtained those gene sets that exhibit coherent epigenetic patterns in different cancer types. Previous studies have shown that the modification intensities are significantly distinct between high-expression gene promoters and low-expression gene promoters, which suggests that these chromatin components have significant effect on gene regulation (Su et al., 2012). To investigate the potential functions of those genes in the cellular control pathways, we performed a systematic GO enrichment analysis using DAVID tools (https://david.ncifcrf.gov/). Then, for the associated gene sets in the identified tri-Clusters, we respectively summarized the key biological processes and pathways that they are involved in.

Overall, we found that those genes enriched in tri-Clusters exhibit an enrichment for cancer-related functions. **Table 1** lists the significant GO terms of a typical tri-Cluster (P-value < 0.005). In this tri-Cluster, the genes exhibit coherent modification patterns on epigenetic marks (H3K4me1, H3K4me3, H3K9me3, H3K27ac, and H3K27me3) in cancer types (HeLa-S3, HepG2, multiple myeloma-Cell Line, and sporadic Burkitt lymphoma-Cell Line). In the table, terms "positive regulation of cell proliferation" and "negative regulation of apoptotic process" are enriched in these gene sets. This result implies that the identified genes in this tri-Cluster are essential for cell proliferation and apoptotic process, which has been reported to be related to cancer development by

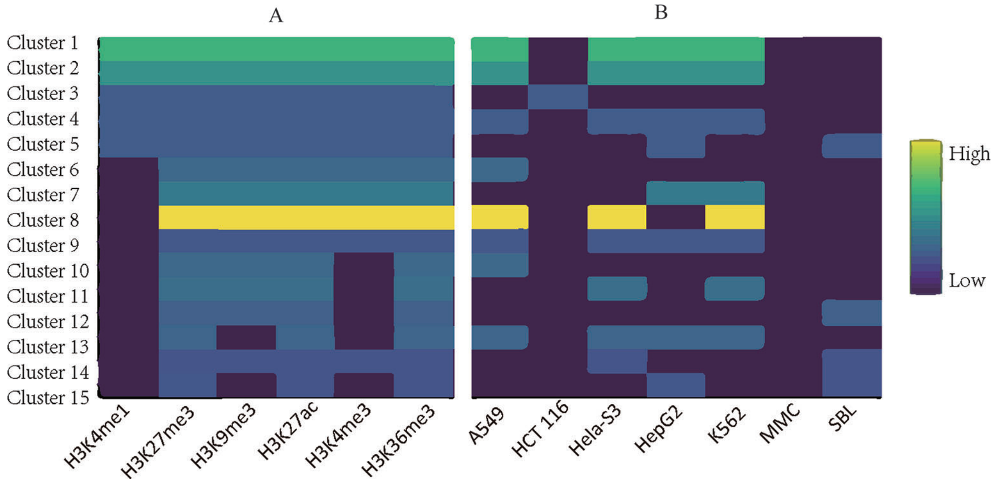


FIGURE 4 | Typical epigenetic tri-Clusters. **(A)** The epigenetic marks (column) in each cluster (row). **(B)** The cancer types (column) in each cluster (row). Fold enrichment was calculated as the ratio between the number of genes in the tri-Cluster to that of all genes.

TABLE 1 | Functional enrichment of genes in the identified tri-Clusters.

Term type	Term name	P-value	Term type	Term name	P-value
BP	Positive regulation of cell proliferation	2.84E-06	MF	Protein binding	1.10E-12
BP	Translational initiation	1.18E-05	MF	Poly(A) RNA binding	3.90E-10
BP	mRNA processing	2.72E-05	MF	RNA binding	2.13E-05
BP	Cell division	4.08E-05	MF	Glutathione binding	7.85E-04
BP	rRNA processing	2.70E-04	MF	Enzyme regulator activity	4.02E-03
BP	RNA splicing	4.04E-04	MF	Nucleosomal DNA binding	4.25E-03
BP	Positive regulation of gene expression, epigenetic	9.41E-04	MF	Translation initiation factor activity	4.30E-03
BP	Protein targeting to Golgi	8.87E-05	MF	Glutathione transferase activity	8.00E-03
BP	Nitrobenzene metabolic process	1.14E-04	MF	Protein binding, bridging	4.33E-03
BP	Xenobiotic catabolic process	1.13E-03	MF	ATP binding	4.57E-03
BP	mRNA splicing, *via* spliceosome	1.14E-03	CC	Nucleoplasm	6.18E-13
BP	Sister chromatid cohesion	2.13E-03	CC	Cytosol	3.96E-07
BP	SRP-dependent cotranslational protein targeting to membrane	1.06E-03	CC	Membrane	7.68E-06
BP	Negative regulation of transcription, DNA-templated	1.55E-03	CC	Nucleus	2.34E-04
BP	Negative regulation of apoptotic process	1.88E-03	CC	Cytoplasm	2.69E-04
BP	Nucleosome assembly	3.86E-03	KEGG	Glutathione metabolism	1.09E-03
BP	Glutathione derivative biosynthetic process	4.18E-03	KEGG	Systemic lupus erythematosus	1.93E-03

previous researches (Deng et al., 2016). Meanwhile, the term "positive regulation of gene expression" is also enriched in the gene set, further indicating that these genes might perform important regulation roles in these cancers.

DISCUSSION

Identifying epigenetic patterns is important to understand epigenetic mechanisms in various cancers. The detected patterns among different cancers could demonstrate critical cross-cancer similarities, which reveals some consistent clinical risk among different cancer types and further suggests strong clinical relevance. Our knowledge about the patterns of epigenetic modifications and the cause and consequence of them is still limited. Computational approach that exploits the complex epigenomic landscapes and discovers significant signatures out of them is required. Previous computational methods for analyzing epigenomes primarily focus on the combinatorial states of different epigenetic marks in a specific cell type. Differently, we developed a tri-clustering approach TriPCE for integrative pan-cancer epigenomic analysis. Based on the FP-tree structure, TriPCE can compactly represent all similar cancer types in the promoter regions for a specific epigenetic mark. Using the constructed FP-tree, the frequent patterns are then detected to yield the set of bi-Clusters of this epigenetic mark, indicating the similar epigenetic pattern in these cancer types along these genomic regions. TriPCE further mines the final tri-Clusters based on the bi-Clusters of all investigated epigenetic marks, explicitly detecting combinatorial epigenetic states in different genomic segments and similar epigenetic changes across different cancer types. In the proposed approach TriPCE, the tri-Cluster enumeration is an expensive operation. In the future we plan to develop heuristic techniques to efficiently prune the search space, and then improve the efficiency of mining the tri-Clusters. We applied TriPCE to uncover the similar patterns of six epigenetic marks among seven cancer types and successfully identified significant cross-cancer epigenetic modification similarities, which suggests that there exhibits consistent epigenetic modification tendency among these investigated cancer types. Furthermore, the gene functional analysis demonstrates that these associated genes are strongly relevant with the cancer cellular pathway.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

YG is responsible for the main idea, as well as the completion of the manuscript. NL and YX have developed the algorithm and performed data analysis. GZ has coordinated data preprocessing and supervised the effort. All authors have read and approved the final manuscript.

FUNDING

This work and the publication costs were supported in part by the National Natural Science Foundation of China (61772128, 61772367), National Key Research and Development Program of China (2016YFC0901704), Shanghai Natural Science Foundation (17ZR1400200,18ZR1414400), and the Fundamental Research Funds for the Central Universities (2232016A3-05),

ACKNOWLEDGMENTS

Authors are grateful to NIH Roadmap Epigenome Project and iHMS website for providing the epigenomic data to carry out this work. An earlier version of this paper was presented at the 2018 International Conference on Intelligent Computing (ICIC 2018).

REFERENCES

Allis, C. D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* 17, 487. doi: 10.1038/nrg.2016.59

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385. doi: 10.1016/j.cell.2018.02.060

Baylin, S. B., and Jones, P. A. (2016). Epigenetic determinants of cancer. *Cold Spring Harbor Perspect. In Biol.* 8, a019505. doi: 10.1101/cshperspect.a019505

Beck, S., Bernstein, B. E., Campbell, R. M., Costello, J. F., Dhanak, D., Ecker, J. R., et al. (2012). A blueprint for an international cancer epigenome consortium. a report from the aacr cancer epigenome task force. *Cancer Res.* 72, 6319–6324. doi: 10.1158/0008-5472.CAN-12-3658

Brien, G. L., Valerio, D. G., and Armstrong, S. A. (2016). Exploiting the epigenome to control cancer-promoting gene-expression programs. *Cancer Cell* 29, 464–476. doi: 10.1016/j.ccell.2016.03.007

Chiappinelli, K. B., Strissel, P. L., Desrichard, A., Li, H., Henke, C., Akman, B., et al. (2015). Inhibiting dna methylation causes an interferon response in cancer *via* dsrna including endogenous retroviruses. *Cell* 162, 974–986. doi: 10.1016/j.cell.2015.07.011

Dawson, M. A. (2017). The cancer epigenome: Concepts, challenges, and therapeutic opportunities. *Science* 355, 1147–1152. doi: 10.1126/science.aam7304

Deng, S. P., Zhu, L., and Huang, D. S. (2016). Predicting hub genes associated with cervical cancer through gene co-expression networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 13, 27–35. doi: 10.1109/TCBB.2015.2476790

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473, 43. doi: 10.1038/nature09906

Flavahan, W. A., Gaskell, E., and Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science* 357, eaal2380. doi: 10.1126/science.aal23800. 1126/science.aal2380

Gan, Y., Tao, H., Guan, J., and Zhou, S. (2017). ihms: a database integrating human histone modification data across developmental stages and tissues. *BMC Bioinf.* 18, 103. doi: 10.1186/s12859-017-1461-y

Gan, Y., Dong, Z., Zhang, X., and Zou, G. (2018). "Tri-clustering analysis for dissecting epigenetic patterns across multiple cancer types," in *International Conference on Intelligent Computing* (Springer), 330–336.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. knowl. Discovery* 8, 53–87. doi: 10.1023/B:DAMI.0000005258.31418.83

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., et al. (2007). David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* 35, W169–W175. doi: 10.1093/nar/gkm415

Italiano, A., Attias, R., Aurias, A., Pérot, G., Burel-Vandenbos, F., Otto, J., et al.(2006). Molecular cytogenetic characterization of a metastatic lungsarcomatoid carcinoma: 9p23 neocentromere and 9p23 p24 amplification including jak2 and jmjd2c. *Cancer Genet. Cytogenet.* 167, 122–130. doi: 10.1016/j.cancergencyto.2006.01.004

Jones, P. A., Issa, J. P. J., and Baylin, S. (2016). Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* 17, 630–641. doi: 10.1038/nrg.2016.93

Karlic, R., Chung, H. R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for geneexpression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2926–2931. doi: 10.1073/pnas.0909344107

Kelly, A. D., and Issa, J.-P. J. (2017). The promise of epigenetic therapy: reprogramming the cancer epigenome. *Curr. Opin. Genet. Dev.* 42, 68–77. doi: 10.1016/j.gde.2017.03.015

Kretzmer, H., Bernhart, S. H., Wang, W., Haake, A., Weniger, M. A., Bergmann, A. K., et al. (2015). Dna-methylome analysis in burkitt and follicular lymphomas identifies differentially methylated regions linked to somatic mutation and transcriptional control. *Nat. Genet.* 47, 1316–1325. doi: 10.1038/ng.3413

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravimoussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495. doi: 10.1038/nature12912

Liu, G., Bollig-Fischer, A., Kreike, B., van de Vijver, M. J., Abrams, J., Ethier, S. P., et al. (2009). Genomic amplification and oncogenic properties of the gasc1 histone demethylase gene in breast cancer. *Oncogene* 28, 4491. doi: 10.1038/onc.2009.297

Martincorena, I., and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science* 349, 1483–1489. doi: 10.1126/science.aab4082

Pinello, L., Xu, J., Orkin, S. H., and Yuan, G. C.(2014). Analysis of chromatin-state plasticity identifiescell-type-specific regulators of h3k27me3 patterns. *Proc. Natl. Acad. Sci. U.S.A.* 111, E344. doi: 10.1073/pnas.1322570111

Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., et al. (2013). Rfecs: a random-forest based algorithm for enhancer identification from chromatin state. *PloS Comput. Biol.* 9, e1002968. doi: 10.1371/journal.pcbi.1002968

Sawan, C., and Herceg, Z. (2010). Histone modifications and cancer. *Adv. In Genet.* 70, 57–85. doi: 10.1016/B978-0-12-380866-0.60003-4

Sohn, K.-A., Ho, J. W., Djordjevic, D., Jeong, H.-h., Park, P. J., and Kim, J. H. (2015). hihmm: Bayesian non-parametric joint inference of chromatin state maps. *Bioinformatics* 31, 2066–2074. doi: 10.1093/bioinformatics/btv117

Su, J., Liu, S., Wu, X., Lv, J., Liu, H., Zhang, R., et al. (2012). Revealing epigenetic patterns in gene regulation through integrative analysis of epigenetic interaction network. *Mol. Biol. Rep.* 39, 1701–1712. doi: 10.1007/s11033-011-0910-3

Ucar, D., Hu, Q., and Tan, K. (2011). Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. *Nucleic Acids Res.* 39, 4063–4075. doi: 10.1093/nar/gkr016

Vinatzer, U., Gollinger, M., Müllauer, L., Raderer, M., Chott, A., and Streubel, B. (2008). Mucosa-associated lymphoid tissue lymphoma: novel translocations including rearrangements of odz2, jmjd2c, and cnn3. *Clin. Cancer Res.* 14, 6426–6431. doi: 10.1158/1078-0432

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer Genome Landsc.. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Waddell, N., Pajic, M., Patch, A.-M., Chang, D. K., Kassahn, K. S., Bailey, P., et al. (2015). Whole genomes redefine the mutational Landscape of pancreatic cancer. *Nature* 518, 495. doi: 10.1038/nature14169

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2015). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Yang, Z.-Q., Imoto, I., Fukuda, Y., Pimkhaokham, A., Shimada, Y., Imamura, M., et al. (2000). Identification of a novel gene, gasc1, within an amplicon at 9p23–24 frequently detected in esophageal cancer cell lines. *Cancer Res.* 60, 4735–4739.

Yang, X., Lin, G., and Zhang, S.(2016). Comparative pan-cancer dna methylation analysis reveals cancer common and specific patterns. *Briefings Bioinf.* 18, 761. doi: 10.1093/bib/bbw063

You, J. S., and Jones, P. A. (2012). Cancer genetics and epigenetics: Two sides of the same coin? *Cancer Cell* 22, 9. doi: 10.1016/j.ccr.2012.06.008

Zhang, Y., An, L., Yue, F., and Hardison, R. C. (2016). Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* 44, 6721–6731. doi: 10.1093/nar/gkw278

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.2019.01379

Comprehensive Analysis of Copy Number Variations in Kidney Cancer by Single-Cell Exome Sequencing

Wenyang Zhou[1†], Fan Yang[2†], Zhaochun Xu[1†], Meng Luo[1], Pingping Wang[1], Yu Guo[1], Huan Nie[1], Lifen Yao[2*] and Qinghua Jiang[1*]*

[1] School of Life Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Department of Neurology, The First Affiliated Hospital of Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Hao Lin,
University of Electronic Science and Technology of China, China
Juan Wang,
Inner Mongolia University, China
Jingpu Zhang,
Henan University of Urban Construction, China

***Correspondence:**
Huan Nie
nh1212@hit.edu.cn
Lifen Yao
yaolifen_2015@sina.com
Qinghua Jiang
qhjiang@hit.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 September 2019*
Accepted: *17 December 2019*
Published: *23 January 2020*

Citation:
Zhou W, Yang F, Xu Z, Luo M, Wang P, Guo Y, Nie H, Yao L and Jiang Q (2020) Comprehensive Analysis of Copy Number Variations in Kidney Cancer by Single-Cell Exome Sequencing. Front. Genet. 10:1379. doi: 10.3389/fgene.2019.01379

Clear-cell renal cell carcinoma (ccRCC) is the most common and lethal subtype of kidney cancer. *VHL* and *PBRM1* are the top two significantly mutated genes in ccRCC specimens, while the genetic mechanism of the *VHL/PBRM1*-negative ccRCC remains to be elucidated. Here we carried out a comprehensive analysis of single-cell genomic copy number variations (CNVs) in *VHL/PBRM1*-negative ccRCC. Genomic CNVs were identified at the single-cell level, and the tumor cells showed widespread amplification and deletion across the whole genome. Functional enrichment analysis indicated that the amplified genes are significantly enriched in cancer-related signaling transduction pathways. Besides, receptor protein tyrosine kinase (RTK) genes also showed widespread copy number variations in cancer cells. Our studies indicated that the genomic CNVs in RTK genes and downstream signaling transduction pathways may be involved in *VHL/PBRM1*-negative ccRCC pathogenesis and progression, and highlighted the role of the comprehensive investigation of genomic CNVs at the single-cell level in both clarifying pathogenic mechanism and identifying potential therapeutic targets in cancers.

Keywords: copy number variations, single-cell exome sequencing, clear-cell renal cell carcinoma, receptor protein tyrosine kinase, signaling transduction pathway

INTRODUCTION

Renal cell carcinoma (RCC) is one kind of kidney cancer, accounting for nearly 300,000 new cancer cases per year worldwide (Hakimi et al., 2013). RCC includes several histological subtypes, among which clear cell renal cell carcinoma (ccRCC) is the most common and lethal one (Hakimi et al., 2016). Increasing studies have shown that the development of ccRCC seems to be shaped by chromosomal lesions and a number of somatic mutations (Sato et al., 2013). *VHL* and *PBRM1*, located within the chromosome 3p25 and 3p21 segments, are the top two significantly mutated genes in ccRCC (Sato et al., 2013). Nearly 90% of ccRCCs undertake the deletion on chromosome 3p, leading to a very high frequency of *VHL* inactivation (Gnarra et al., 1994). Moreover, *VHL* and *PBRM1* are mutated in about 50 and 41% of sporadic ccRCC, respectively (Kaelin, 2004; Varela et al., 2011). However, little is known about the genetic mechanisms in *VHL/PBRM1*-negative ccRCC.

Based on the next-generation sequencing technology, previous studies identified many driver mutations in ccRCC (Gnarra et al., 1994; Kaelin, 2004; Sato et al., 2013; Cheng et al., 2019). However, limited insights are available on the genomic diversity within tumor tissues (Wang et al., 2014). Generally, tumor tissues may contain cancer cells from multiple clones and noncancerous cells, which make it difficult to identify the mutations in each clone and detect the driver genes during the cancer progression (Xu et al., 2012; Casasent et al., 2018). Fortunately, single-cell DNA sequencing has been developed to meet this challenge, because it can provide unique insights into intratumor heterogeneity, development, and diversity of cancers at the single-cell level (Casasent et al., 2018). For example, Xu et al. (2012) carried out the single-cell exome sequencing on a ccRCC tumor and its adjacent normal tissue. They identified four genes (i.e., *AHNAK, SRGAP3, LRRK2,* and *USP6*) as potential driving factors for *VHL/PBRM1*-negative ccRCC development, which provided new insights into the pathogenesis of the ccRCC.

Genomic copy number variations (CNVs) play an important role in cancer progression, and emerging studies indicate that genomic CNVs are associated with the ccRCC (Gerlinger et al., 2014; Nouhaud et al., 2018) and other cancers (Waddell et al., 2015; Secrier et al., 2016; Hong et al., 2019). Xu et al. (Xu et al., 2012) performed a single-cell exome sequencing to elucidate the genetic mechanisms of the ccRCC by identifying the single nucleotide variants (SNVs). However, the authors did not examine whether the genomic copy number variations play a crucial role in ccRCC.

To further investigate the potential roles of CNVs in *VHL/PBRM1*-negative ccRCC, we performed a comprehensive single-cell CNV analysis based on a dataset provided by Xu et al., (2012). We delineated the genomic copy number variation landscape at the single-cell level and reclassified all single cells based on the single-cell genomic CNVs. We also identified several significantly amplified/deleted loci and genes in cancer cells. Finally, we further investigated the biological pathways which may be involved in the ccRCC pathogenesis.

METHODS

Datasets

The sample data and information used in our article came from a previous study, and the original sequencing data were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA050201.

Quality Control

Quality control of the sequencing data was performed using FastQC. The adapter and low-quality ends were trimmed from reads using Trim-Galore version 0.5.0 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Trimmed reads shorter than 20 bp were discarded.

Reads Mapping

The human reference genome sequence (Hg19) was used for mapping (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/). Short read pairs were mapped to the reference genome using Burrows-Wheeler Aligner (BWA) version 0.7.12-r1039 (Li and Durbin, 2009). In this process, we adopted the BWA-MEM algorithm and adjusted the main parameters, setting the minimum seed length to 19, the penalty for a mismatch to 4, and shorter split hits were marked as secondary. Then, Samtools was used to convert SAM files to compressed BAM files, sort the BAM files by chromosomal coordinates, and remove the PCR duplicates from BAM files.

Copy Number Variations Calling

In each cell, germline and somatic copy number variations were called by Control-FREEC version 11.5 (Boeva et al., 2012). Considering the exome enrichment during library construction, read counts were calculated by exome region. The target region file of exome capture was downloaded from the Agilent website (https://earray.chem.agilent.com/suredesign/index.htm). The germline CNVs were detected in each cell and bulk normal tissue, respectively. Somatic CNVs were detected only in single cells. Gene annotations were performed with Annovar software (Wang et al., 2010) and OAHG database (Cheng et al., 2016).

Dimensionality Reduction of Cells

T-distributed stochastic neighbor embedding (t-SNE) was performed based on the germline CNVs of target regions. Both 25 single cells and bulk normal tissue were projected to 2D space using the R package named "Rtsne."

Significantly Somatic Copy Number Variation Loci Analysis

Significantly amplified/deleted loci in tumor cells were identified using GISTIC2.0 (Mermel et al., 2011). GISTIC2.0 was run on an input defined as the $\log_2()$-1 of somatic copy number values, with confidence (-conf) threshold of 0.9. Considering for downstream analysis, *thresholds suggested by GISTIC2.0 for copy number variation were as follows*: if GISTIC score ≥0.9, it means amplification; 0.1 < GISTIC score <0.9, corresponding to gain; −1.3 < GISTIC score < −0.1, loss; GISTIC score ≤ −1.3, deletion.

Receptor Protein Tyrosine Kinase Gene Copy Number Profiling

To examine the landscape of copy number variations in RTK genes, we derived GISTIC-equivalent scores by dividing the germline copy numbers and classifying genes as amplified if score ≥ 0.9, deleted if score ≤ −1.3, gained if score > 0.1, and loss if the score < −0.1.

Function Analysis

The significantly amplified and deleted genes were identified according to significantly somatic CNV loci (q-value < 0.0001) in GISTIC2.0. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway function enrichment analysis was performed using the Carcinogenic Potency Database (CPDB) (Kamburov et al., 2013). In this study, the p-value threshold for KEGG enrichment analysis is 0.05.

RESULTS

Identification of Single-Cell Genomic Copy Number Variations in Kidney Cancer and Normal Cells

To identify genomic CNVs in ccRCC, we analyzed the sequencing data from a ccRCC patient, which includes 20 single-cell exome sequencing data from the tumor tissue, 5 single-cell exome sequencing data from the adjacent normal tissue, and a bulk exome sequencing data from the adjacent normal tissue. Trim-Galore was used to remove the low-quality and adapter segments and analyze the quality of sequencing reads. The cleaned reads were mapped to the reference genome with BWA software (Li and Durbin, 2009). The sequencing depth was more than 20X (29.68 ± 5.68) in all single cells. The genomic CNVs were called by using Control-FREEC (Boeva et al., 2012).

Germline CNVs were identified in all the samples. The comparison between cancer and normal cells revealed widespread amplification and deletion across the whole genome in tumor cells (**Figure 1A**). At the same time, some deleted loci were found both in normal and cancer cells, which may be caused by multiple displacement amplification (MDA) amplification (Yilmaz and Singh, 2012) or exome capture during DNA library preparation.

To remove the background mutations caused by germline or technology flaws, somatic CNVs were identified in all cells using bulk normal tissue as control. The somatic CNVs showed much more amplification than germline CNVs in the cancer cells remarkably (**Figure 1B**). Large-scale of somatic CNVs were found in the ccRCC single cells, which was consistent with the previous studies based on the bulk sequencing (Cancer Genome Atlas Research, N, 2013; Gerlinger et al., 2014; Nouhaud et al., 2018). What's more, single-cell sequencing data revealed the amplification of copy number showed a high degree of consistency, which suggests the amplification may play an important role in the progression of ccRCC. On the contrary, the deletion showed higher intratumor heterogeneity in the cancer cells.

Re-Classification of Kidney Cancer and Normal Cells Based on Single-Cell Copy Number Variations

Generally, surgically removed cancer tumors may contain both cancer and normal cells (Xu et al., 2012). To reclassify all the single cells accurately, the t-distributed stochastic neighbor embedding (t-SNE) was performed based on the cell copy number in exome target regions. The results of dimensionality reduction (**Figure 2**, **Supplementary Table S1**) showed that three cancer cells (CC-15, CC-17, and CC-20) clustered tightly

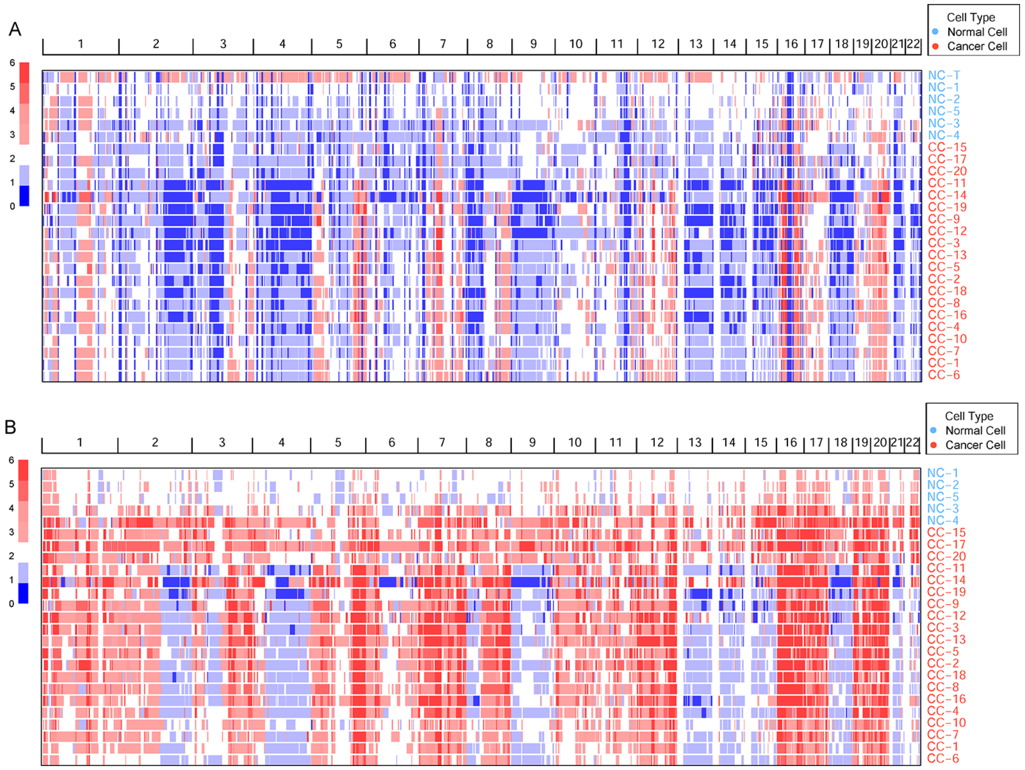


FIGURE 1 | The genomic copy number variations (CNVs) identified across all cells. **(A)** The germline CNVs in single cells and normal tissue. Genomic CNVs within the whole genome are shown, the color scale ranges from blue (deletion) to red (amplification) with estimated copy numbers shown. The cell names are marked by different cell types. **(B)** The somatic CNVs in single cells.

with the normal cells and tissue, suggesting that they probably were normal cells in the tumor tissue. These results were consistent with the previous findings which based on the single-cell SNVs (Xu et al., 2012). These three cells (CC-15, CC-17, and CC-20) were excluded from the cancer cell group in the downstream analysis. Focusing on the remaining cancer cells, we found no subpopulation of cancer cells within the cancer tissue.

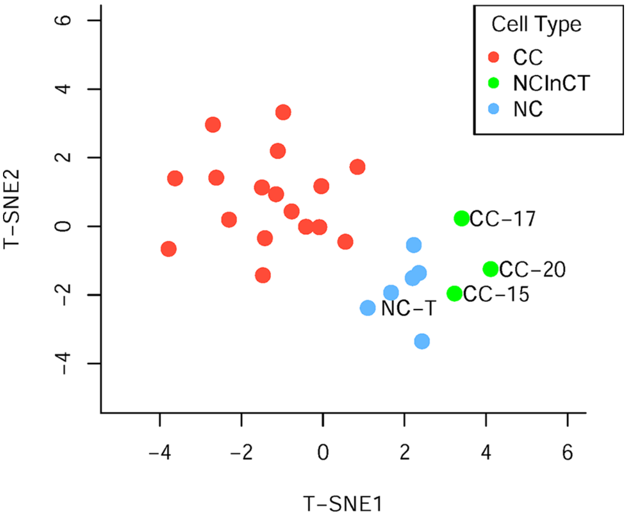


FIGURE 2 | Population analysis based on the germline copy number variations (CNVs). T-distributed stochastic neighbor embedding (T-SNE) analysis of cancer cell (red), normal cell (blue), and normal cell in cancer tissue (green) based on the germline CNVs.

According to the single-cell genomic CNVs, all the single cells can be reclassified into three groups, namely cancer cell (CC), normal cell (NC), and normal cell in cancer tissue (NCinCT). To address whether the genomic CNVs were significantly different between the three groups, we calculated the proportion of whole genome that covered with amplification (copy number ≥ 4) and loss (copy number = 0), respectively. The results (**Figure 3**) showed that there were more amplified loci in CC group than NC group ($P = 3\times10^{-4}$) and NCinCT group ($P = 1.8\times10^{-3}$). Besides, there was no significant difference between NC and NCinCT groups ($P = 0.79$). The lost loci also showed a similar result. Single-cell genomic CNVs indicated that the genome of cancer cells was in an extremely unstable state.

Loci Distribution of Significant Genomic Copy Number Variations in Kidney Cancer

To investigate the loci distribution of the significant genomic CNVs across all tumor single cells, GISTIC2.0 (Mermel et al., 2011) was used to identify the significant genomic CNVs loci based on the somatic CNVs in 17 cancer cells, but not including germline CNVs which are not involved in cancer development generally. The results indicated that copy numbers in the significant CNV loci have a high degree of consistency across all the cancer cells. Although lots of lost loci (more slight than deletion, −1.3 < GISTIC score < −0.1) were identified, there was no significantly deleted locus (GISTIC score ≤ −1.3) found in cancer cells, which was consistent with high heterogeneity of deletion region in our cancer cells.

Significantly amplified loci (**Figure 4**, **Supplementary Table S2A**) according to GISTIC2.0 (12q13.3, 12p13.31, 5q35.3, etc.; q-value < 0.05) comprised genes such as *IGFBP4*, *ERBB2*, *ERBB3*, *FGFR4*, *CDK2*, *FLT4*, and so on. The *IGFBP4* gene had been reported to be associated with several types of cancer (Hallberg

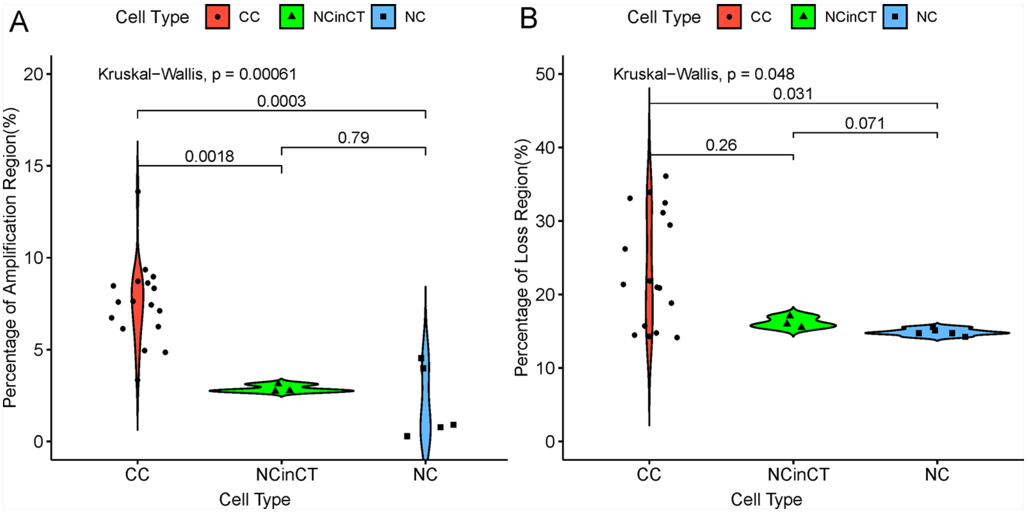


FIGURE 3 | The coverage of genomic copy number variations (CNV) regions in three cell types. **(A)** The percentage of amplification region (copy number ≥ 4) across the whole genome in different cell types. **(B)** The percentage of loss region (copy number = 0) across the whole genome in different cell types. In the two sub-graphs **(A)** and **(B)**, p-values between two groups (Wilcoxon signed-rank test) and all groups (Kruskal-Wallis test) were calculated.

et al., 2000; Romero et al., 2011; Yang et al., 2017), it can promote the RCC cell metastasis and activate Wnt/beta-catenin signaling pathway in humans (Ueno et al., 2011). *ERBB2* and *ERBB3* genes belong to the epidermal growth factor receptor (*EGFR*) family, and they had been identified as common driver genes of multiple cancer types by promoting solid tumor growth (Yarden, 2001; Henson et al., 2017; Oldrini et al., 2017). The amplification of *EGFR* also was found in other cancers, which contributed to the *EGFR* excessive activation (Sigismund et al., 2018). *FGFR4* gene belongs to the fibroblast growth factor receptor family, and the activation of *FGFR4* can promote cell growth and angiogenesis in cancer (Bai et al., 2015). *CDK2* gene is commonly excessive activation in human cancers, and dysfunction of *CDK2* can lead to uncontrolled cell growth (Mihara et al., 2001). *FLT4* gene, belonging to the vascular endothelial growth factor family, had been reported to regulate cancer cell survival and proliferation (Varney and Singh, 2015).

While the top significantly deleted loci (**Figure 4**, **Supplementary Table S2B**) (1q21.3, 1p35.2, 16q24.3, 3p14.1, etc.; q-value < 0.05) showed loss of *Chmp1A*, *CADM2*, *PRAP1*, and *ULK1* genes. *Chmp1A* and *CADM2*, belonging to cell adhesion molecules family, had been found to be a tumor suppressor gene in RCC. The overexpression of *Chmp1A* and *CADM2* significantly suppressed cancer growth and invasion (You et al., 2012; He et al., 2013). *PARP1* gene played an important role in DNA repair and cell apoptosis (Tulin, 2011), the cell with *PARP1* deficiency show resistance to DNA damage-induced programmed cell death and increased cancer risk (Schiewer and Knudsen, 2014). *ULK1* was an initiate autophagy gene, and the down-regulation of *ULK1* had been found in cancer (Zhang et al., 2017). *ULK1* may play a pivotal role in cancer by promoting cell death (Chen et al., 2014).

The genes in significantly amplified loci include a number of known driver genes, which may promote the cancer progression by the up-regulation of cell growth and cell cycle. Significantly deleted loci include some tumor suppressor genes and autophagy genes. The inactivation of these genes leads to uncontrolled tumor growth, which may contribute to the *VHL/PBRM1*-negative ccRCC pathogenesis and progression

Functional Analysis of Significant Genomic Copy Number Variations in Kidney Cancer

To better understand the potential biological and functional characteristics of the significantly amplified and deleted genes in cancer cells, biological function pathways in ccRCC had been further investigated. The KEGG functional enrichment analysis was performed using the CPDB Database based on the amplified and deleted genes, respectively. The amplified genes showed significant enrichment (p-value < 0.05) for signal transduction, metabolism, cell cycle, immunity, and other cancer-related pathways (**Figure 5**, **Supplementary Table S3**). In contrast to amplified genes, deleted genes only showed significant enrichment for the fatty acid elongation pathway (p-value = 7.6×10^{-3}).

The most notable result is that a large portion of enrichment pathways belong to the signaling transduction pathway. Both of the HIF-1 (Posadas et al., 2013), ErbB (Liu et al., 2015), PI3K-Akt (Linehan et al., 2010; Sato et al., 2013; Guo et al., 2015), Ras (de Araujo Junior et al., 2015; Chen et al., 2018), Rap1 (Chen et al., 2018), and MAPK signaling pathway (Liu et al., 2015) had been

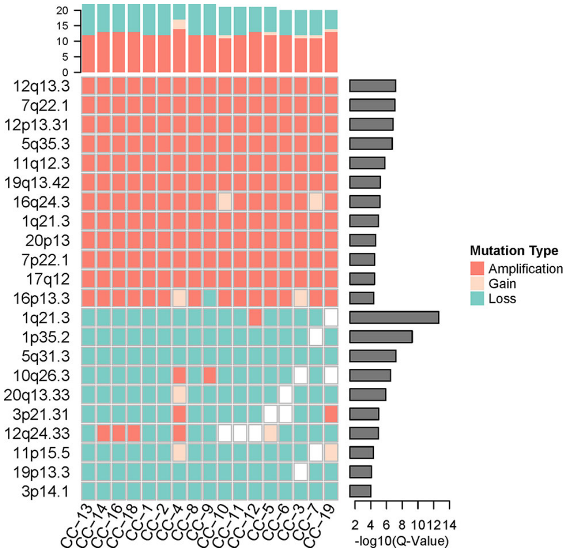


FIGURE 4 | The significant genomic copy number variation (CNV) loci in cancer cells. All CNV types in each cancer cell were counted for the top frequency histogram, and q-value for each significant genomic CNV loci was shown on the right. Only the loci with q-value < 0.0001 were shown.

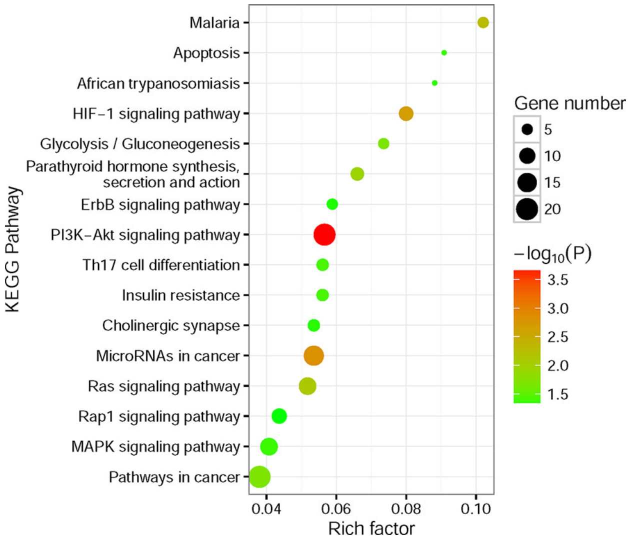


FIGURE 5 | Kyoto Encyclopedia of Genes and Genomes (KEGG) functional enrichment analysis for significantly amplified genes. The size of the point means the gene number both in our amplified gene set and KEGG pathway terms. The color of point means enrichment significance ($-\log_{10}P$). The pathways were sorted by rich factor (the ratio of significantly amplified gene number in this pathway term to gene number in this pathway term).

found involved in the pathogenesis of RCC. What's more, the results also showed that Th17 cells (Li et al., 2015) and microRNAs (Gowrishankar et al., 2014) seem to have a connection with the ccRCC pathogenesis. Interestingly, the fatty acid elongation pathway was significantly deleted in ccRCC, which may account for the fact that ccRCC tumors are lipid-laden (Hakimi et al., 2016).

Receptor Protein Tyrosine Kinase Genes Show Widespread Copy Number Variations in Cancer Cells

Since lots of cancer-related signaling transduction pathways showed significantly amplified in cancer cells, we then examined the copy number variations in their upstream RTK genes (Robinson et al., 2000; Secrier et al., 2016) to investigate possible reasons for the negative results that tumor did not appear known driver mutations in *VHL* and *PBRM1*.

The single cancer cells show widespread amplification and deletion on multiple RTKs compared with the normal cells, the NC and NCinCT groups show similar RTK gene profile. There were some RTK genes (*EPHB6*, *EPHA1*, *EPHB3*, *FGFR4*, *PDGFRB*, and *FLT4*) showing amplification in cancer cells. On the contrary, *EPHB2*, *ERBB4*, *FGFR1*, *PDGFRA*, *KDR*, and *FLT1* genes showed deletion in cancer cells (**Figure 6**). Genomic copy number is varied across these RTKs and downstream pathways, indicating that the genomic CNVs in RTKs and downstream

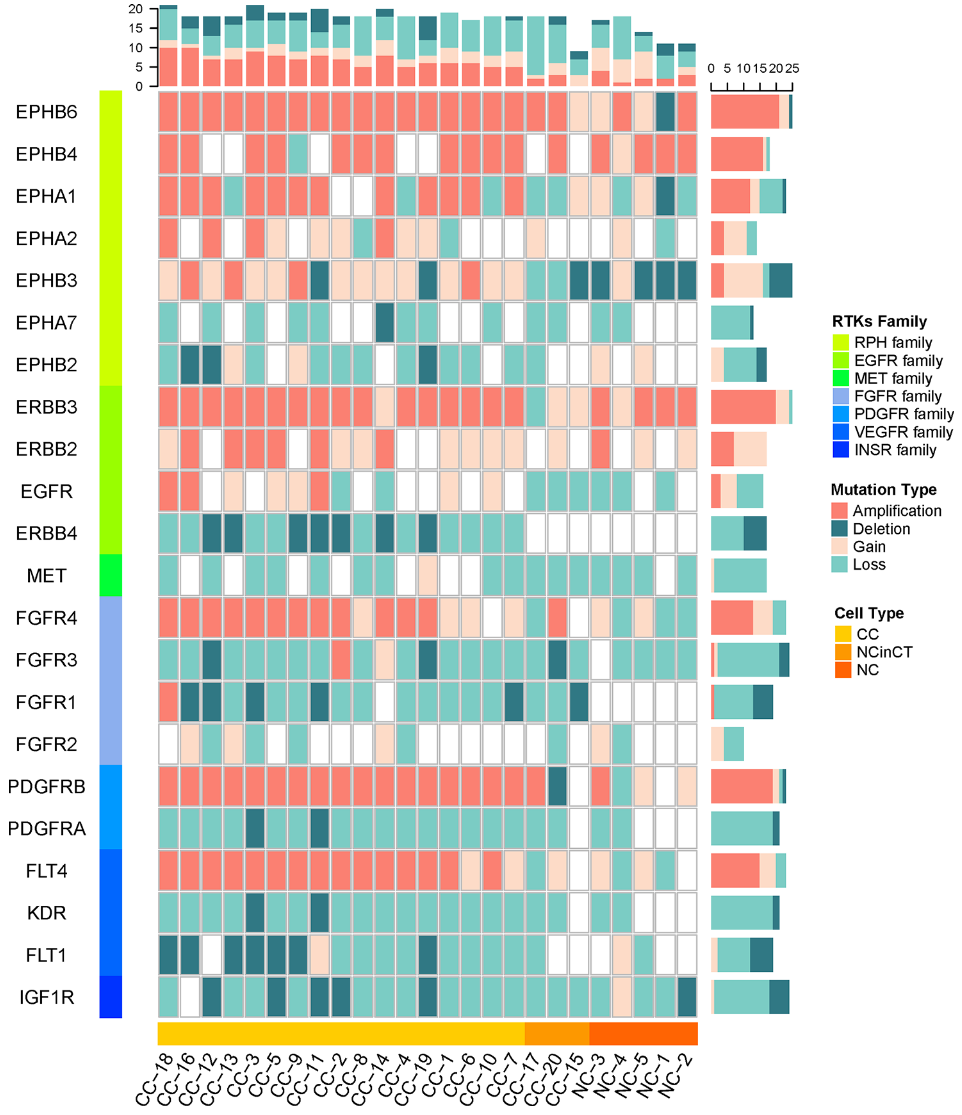


FIGURE 6 | The copy number of receptor protein tyrosine kinase (RTK) genes in all single cells. The copy number variations (CNVs) on RTK genes in both tumor and normal cells were shown. The RTKs family and cell types were shown at the left and bottom of the plot. The mutation types in each cell and gene were counted for the top and right frequency histograms, respectively.

signaling transduction pathways may have important roles in the pathogenesis and progression of the *VHL/PBRM1*-negative ccRCC.

DISCUSSION

Previous studies have shown that *VHL* and *PBRM1* are the top two significantly mutated genes in ccRCC. However, the pathogenesis in *VHL/PBRM1*-negative ccRCC is still unclear. Our comprehensive analysis of CNVs in 25 single cells from a ccRCC patient provided new insights into the pathogenesis of the ccRCC. We reclassified all the single cells and identified pathological mutations in *VHL/PBRM1*-negative ccRCC cells. Similar to the genomic CNVs in other cancers, the pathogenesis in *VHL/PBRM1*-negative ccRCC seems to be shaped by the accumulation of amplification in driver genes (*IGFBP4*, *ERBB2*, *ERBB3*, *FGFR4*, *CDK2*, and *FLT4*), the loss of function in tumor suppressor genes (*Chmp1A*, *CADM2*) and autophagy genes (*PRAP1*, *ULK1*).

Pathway analysis of these significantly amplified and deleted genes identified several signaling transduction pathways, including HIF-1, ErbB, PI3K-Akt, Ras, Rap1, and MAPK signaling pathways, were affected by genomic amplification. At the same time, RTK genes showed widespread copy number variations in cancer cells specifically. Mutations on RTKs may take part in the overactivity of downstream signaling transduction pathways, leading to the uncontrolled growth of ccRCC cells.

Overall, our single-cell analysis of the copy number in *VHL/PBRM1*-negative ccRCC revealed that the genomic CNVs in RTKs may cooperate with downstream signaling transduction pathways to take part in *VHL/PBRM1*-negative ccRCC pathogenesis and progression. Clinically, our findings may provide more effective targeted therapeutic approaches for patients with *VHL/PBRM1*-negative ccRCC. However, because of the small number of cells and the high intratumor heterogeneity, our findings need to be verified in larger cohorts.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The original sequencing data can be downloaded from NCBI (http://www.ncbi.nlm.nih.gov/sra) under the accession number SRA050201.

AUTHOR CONTRIBUTIONS

HN, LY, and QJ designed the experiments. PW obtained data from NCBI. WZ and ML analyzed the data. FY, ZX, and YG wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This work is supported by the National Natural Science Foundation of China [61571152, 61822108], the National Science and Technology Major Project of China [2016YFC1202302] and the Natural Science Foundation of Heilongjiang Province [F2015006].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01379/full#supplementary-material

TABLE S1 | The results of dimensionality reduction based on the germline CNVs. The name and coordinate in 2D space of all single cells were shown in this table.

TABLE S2 | The results of significantly amplified (**Table S2A**) and deleted (**Table S2B**) loci according to GISTIC2.0. The cytoband name, q-value and gene names of each amplification/deletion loci were shown in this table.

TABLE S3 | The results of KEGG enrichment analysis based on significantly amplified (**Table S3A**) and deleted (**Table S3B**) genes according to the CPDB database. The pathway name, p-value and gene sets of each pathway were shown in this table.

REFERENCES

Bai, Y. P., Shang, K., Chen, H., Ding, F., Wang, Z., Liang, C., et al. (2015). FGF-1/-3/FGFR4 signaling in cancer-associated fibroblasts promotes tumor progression in colon cancer through Erk and MMP-7. *Cancer Sci.* 106 (10), 1278–1287. doi: 10.1111/cas.12745

Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappo, J., Schleiermacher, G., et al. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28 (3), 423–425. doi: 10.1093/bioinformatics/btr670

Cancer Genome Atlas Research, N. (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499 (7456), 43–49. doi: 10.1038/nature12222

Casasent, A. K., Schalck, A., Gao, R., Sei, E., Long, A., Pangburn, W., et al. (2018). Multiclonal invasion in breast tumors identified by topographic single cell sequencing. *Cell* 172 (1-2), 205–217 e212. doi: 10.1016/j.cell.2017.12.007

Chen, Y., He, J., Tian, M., Zhang, S. Y., Guo, M. R., Kasimu, R., et al. (2014). UNC51-like kinase 1, autophagic regulator and cancer therapeutic target. *Cell Prolif.* 47 (6), 494–505. doi: 10.1111/cpr.12145

Chen, Y. L., Ge, G. J., Qi, C., Wang, H., Wang, H. L., Li, L. Y., et al. (2018). A five-gene signature may predict sunitinib sensitivity and serve as prognostic biomarkers for renal cell carcinoma. *J. Cell Physiol.* 233 (10), 6649–6660. doi: 10.1002/jcp.26441

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47 (D1), D140–D144. doi: 10.1093/nar/gky1051

de Araujo Junior, R. F., Leitao Oliveira, A. L., de Melo Silveira, R. F., de Oliveira Rocha, H. A., de Franca Cavalcanti, P., and de Araujo, A. A. (2015). Telmisartan induces apoptosis and regulates Bcl-2 in human renal cancer cells. *Exp. Biol. Med. (Maywood)* 240 (1), 34–44. doi: 10.1177/1535370214546267

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., et al. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* 46 (3), 225–233. doi: 10.1038/ng.2891

Gnarra, J. R., Tory, K., Weng, Y., Schmidt, L., Wei, M. H., Li, H., et al. (1994). Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat. Genet.* 7 (1), 85–90. doi: 10.1038/ng0594-85

Gowrishankar, B., Ibragimova, I., Zhou, Y., Slifker, M. J., Devarajan, K., Al-Saleem, T., et al. (2014). MicroRNA expression signatures of stage, grade, and progression in clear cell RCC. *Cancer Biol. Ther.* 15 (3), 329–341. doi: 10.4161/cbt.27314

Guo, H., German, P., Bai, S., Barnes, S., Guo, W., Qi, X., et al. (2015). The PI3K/AKT pathway and renal cell carcinoma. *J. Genet. Genomics* 42 (7), 343–353. doi: 10.1016/j.jgg.2015.03.003

Hakimi, A. A., Pham, C. G., and Hsieh, J. J. (2013). A clear picture of renal cell carcinoma. *Nat. Genet.* 45 (8), 849–850. doi: 10.1038/ng.2708

Hakimi, A. A., Reznik, E., Lee, C. H., Creighton, C. J., Brannon, A. R., Luna, A., et al. (2016). An integrated metabolic atlas of clear cell renal cell carcinoma. *Cancer Cell* 29 (1), 104–116. doi: 10.1016/j.ccell.2015.12.004

Hallberg, L. M., Ikeno, Y., Englander, E., and Greeley, G. H.Jr. (2000). Effects of aging and caloric restriction on IGF-I, IGF-I receptor, IGFBP-3 and IGFBP-4 gene expression in the rat stomach and colon. *Regul. Pept.* 89 (1-3), 37–44. doi: 10.1016/s0167-0115(00)00095-1

He, W., Li, X., Xu, S., Ai, J., Gong, Y., Gregg, J. L., et al. (2013). Aberrant methylation and loss of CADM2 tumor suppressor expression is associated with human renal cell carcinoma tumor progression. *Biochem. Biophys. Res. Commun.* 435 (4), 526–532. doi: 10.1016/j.bbrc.2013.04.074

Henson, E., Chen, Y., and Gibson, S. (2017). EGFR family members' regulation of autophagy is at a crossroads of cell survival and death in cancer. *Cancers (Basel)* 9 (4), 27–40. doi: 10.3390/cancers9040027

Hong, X., Qiao, S., Li, F., Wang, W., Jiang, R., Wu, H., et al. (2019). Whole-genome sequencing reveals distinct genetic bases for insulinomas and non-functional pancreatic neuroendocrine tumours: leading to a new classification system. *Gut* 0, 1–11. doi: 10.1136/gutjnl-2018-317233

Kaelin, W. G. Jr. (2004). The von Hippel-Lindau tumor suppressor gene and kidney cancer. *Clin. Cancer Res.* 10 (18 Pt 2), 6290S–6295S. doi: 10.1158/1078-0432.CCR-sup-040025

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* 41 (Database issue), D793–D800. doi: 10.1093/nar/gks1055

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, L., Yang, C., Zhao, Z., Xu, B., Zheng, M., Zhang, C., et al. (2015). Skewed T-helper (Th)1/2- and Th17/T regulatorycell balances in patients with renal cell carcinoma. *Mol. Med. Rep.* 11 (2), 947–953. doi: 10.3892/mmr.2014.2778

Linehan, W. M., Srinivasan, R., and Schmidt, L. S. (2010). The genetic basis of kidney cancer: a metabolic disease. *Nat. Rev. Urol.* 7 (5), 277–285. doi: 10.1038/nrurol.2010.47

Liu, X., Wang, J., and Sun, G. (2015). Identification of key genes and pathways in renal cell carcinoma through expression profiling data. *Kidney Blood Press Res.* 40 (3), 288–297. doi: 10.1159/000368504

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12 (4), R41. doi: 10.1186/gb-2011-12-4-r41

Mihara, M., Shintani, S., Nakahara, Y., Kiyota, A., Ueyama, Y., Matsumura, T., et al. (2001). Overexpression of CDK2 is a prognostic indicator of oral cancer progression. *Jpn. J. Cancer Res.* 92 (3), 352–360. doi: 10.1111/j.1349-7006.2001.tb01102.x

Nouhaud, F. X., Blanchard, F., Sesboue, R., Flaman, J. M., Sabourin, J. C., Pfister, C., et al. (2018). Clinical relevance of gene copy number variation in metastatic clear cell renal cell carcinoma. *Clin. Genitourin Cancer* 16 (4), e795–e805. doi: 10.1016/j.clgc.2018.02.013

Oldrini, B., Hsieh, W. Y., Erdjument-Bromage, H., Codega, P., Carro, M. S., Curiel-Garcia, A., et al. (2017). EGFR feedback-inhibition by Ran-binding protein 6 is disrupted in cancer. *Nat. Commun.* 8 (1), 2035. doi: 10.1038/s41467-017-02185-w

Posadas, E. M., Limvorasak, S., Sharma, S., and Figlin, R. A. (2013). Targeting angiogenesis in renal cell carcinoma. *Exp. Opin. Pharmacother.* 14 (16), 2221–2236. doi: 10.1517/14656566.2013.832202

Robinson, D. R., Wu, Y. M., and Lin, S. F. (2000). The protein tyrosine kinase family of the human genome. *Oncogene* 19 (49), 5548–5557. doi: 10.1038/sj.onc.1203957

Romero, D., O'Neill, C., Terzic, A., Contois, L., Young, K., Conley, B. A., et al. (2011). Endoglin regulates cancer-stromal cell interactions in prostate tumors. *Cancer Res.* 71 (10), 3482–3493. doi: 10.1158/0008-5472.CAN-10-2665

Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., et al. (2013). Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat. Genet.* 45 (8), 860–867. doi: 10.1038/ng.2699

Schiewer, M. J., and Knudsen, K. E. (2014). Transcriptional roles of PARP1 in cancer. *Mol. Cancer Res.* 12 (8), 1069–1080. doi: 10.1158/1541-7786.MCR-13-0672

Secrier, M., Li, X., de Silva, N., Eldridge, M. D., Contino, G., Bornschein, J., et al. (2016). Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nat. Genet.* 48 (10), 1131–1141. doi: 10.1038/ng.3659

Sigismund, S., Avanzato, D., and Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. *Mol. Oncol.* 12 (1), 3–20. doi: 10.1002/1878-0261.12155

Tulin, A. (2011). Re-evaluating PARP1 inhibitor in cancer. *Nat. Biotechnol.* 29 (12), 1078–1079. doi: 10.1038/nbt.2058

Ueno, K., Hirata, H., Majid, S., Tabatabai, Z. L., Hinoda, Y., and Dahiya, R. (2011). IGFBP-4 activates the Wnt/beta-catenin signaling pathway and induces M-CAM expression in human renal cell carcinoma. *Int. J. Cancer* 129 (10), 2360–2369. doi: 10.1002/ijc.25899

Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C. K., Stephens, P., et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* 469 (7331), 539–542. doi: 10.1038/nature09639

Varney, M. L., and Singh, R. K. (2015). VEGF-C-VEGFR3/Flt4 axis regulates mammary tumor growth and metastasis in an autocrine manner. *Am. J. Cancer Res.* 5 (2), 616–628.

Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., Kassahn, K. S., Bailey, P., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 518 (7540), 495–501. doi: 10.1038/nature14169

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38 (16), e164. doi: 10.1093/nar/gkq603

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512 (7513), 155–160. doi: 10.1038/nature13600

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., et al. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148 (5), 886–895. doi: 10.1016/j.cell.2012.02.025

Yang, B., Zhang, L., Cao, Y., Chen, S., Cao, J., Wu, D., et al. (2017). Overexpression of lncRNA IGFBP4-1 reprograms energy metabolism to promote lung cancer progression. *Mol. Cancer* 16 (1), 154. doi: 10.1186/s12943-017-0722-8

Yarden, Y. (2001). The EGFR family and its ligands in human cancer. signalling mechanisms and therapeutic opportunities. *Eur. J. Cancer* 37 Suppl 4, S3–S8. doi: 10.1016/s0959-8049(01)00230-1

Yilmaz, S., and Singh, A. K. (2012). Single cell genome sequencing. *Curr. Opin. Biotechnol.* 23 (3), 437–443. doi: 10.1016/j.copbio.2011.11.018

You, Z., Xin, Y., Liu, Y., Sun, J., Zhou, G., Gao, H., et al. (2012). Chmp1A acts as a tumor suppressor gene that inhibits proliferation of renal cell carcinoma. *Cancer Lett.* 319 (2), 190–196. doi: 10.1016/j.canlet.2012.01.010

Zhang, L., Fu, L., Zhang, S., Zhang, J., Zhao, Y., Zheng, Y., et al. (2017). Discovery of a small molecule targeting ULK1-modulated cell death of triple negative breast cancer *in vitro* and *in vivo*. *Chem. Sci.* 8 (4), 2687–2701. doi: 10.1039/c6sc05368h

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.2019.01336

Integrating Multi-Omics Data to Identify Novel Disease Genes and Single-Neucleotide Polymorphisms

Sheng Zhao[1], *Huijie Jiang*[1*], *Zong-Hui Liang*[2*] and *Hong Ju*[3*]

[1] Department of Radiology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Department of Radiology, Jian'an District Centre Hospital of Fudan University, Shanghai, China, [3] Department of Information Engineering, Heilongjiang Biological Science and Technology Career Academy, Harbin, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Tianyi Zhao,
Harvard University, United States
Hao Lin,
University of Electronic Science and Technology of China, China

***Correspondence:**
Huijie Jiang
jianghuijie@hrbmu.edu.cn
Zong-Hui Liang
liangzh@vip.163.com
Hong Ju
hongju.hit@hotmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *15 October 2019*
Accepted: *06 December 2019*
Published: *24 January 2020*

Citation:
Zhao S, Jiang H, Liang Z-H and Ju H (2020) Integrating Multi-Omics Data to Identify Novel Disease Genes and Single-Neucleotide Polymorphisms. Front. Genet. 10:1336. doi: 10.3389/fgene.2019.01336

Stroke ranks the second leading cause of death among people over the age of 60 in the world. Stroke is widely regarded as a complex disease that is affected by genetic and environmental factors. Evidence from twin and family studies suggests that genetic factors may play an important role in its pathogenesis. Therefore, research on the genetic association of susceptibility genes can help understand the mechanism of stroke. Genome-wide association study (GWAS) has found a large number of stroke-related loci, but their mechanism is unknown. In order to explore the function of single-nucleotide polymorphisms (SNPs) at the molecular level, in this paper, we integrated 8 GWAS datasets with brain expression quantitative trait loci (eQTL) dataset to identify SNPs and genes which are related to four types of stroke (ischemic stroke, large artery stroke, cardioembolic stroke, small vessel stroke). Thirty-eight SNPs which can affect 14 genes expression are found to be associated with stroke. Among these 14 genes, 10 genes expression are associated with ischemic stroke, one gene for large artery stroke, six genes for cardioembolic stroke and eight genes for small vessel stroke. To explore the effects of environmental factors on stroke, we identified methylation susceptibility loci associated with stroke using methylation quantitative trait loci (MQTL). Thirty-one of these 38 SNPs are at greater risk of methylation and can significantly change gene expression level. Overall, the genetic pathogenesis of stroke is explored from locus to gene, gene to gene expression and gene expression to phenotype.

Keywords: stroke, genome-wide association study, expression quantitative trait loci, mQTL, SMR, single-nucleotide polymorphisms

INTRODUCTION

Stroke is a major cerebrovascular disease caused by a transient or permanent decrease of local cerebral blood flow. It is characterized by arterial obstruction (Krishnamurthi et al., 2018), so it is also called cerebral infarction (Dargazanli et al., 2018). According to the World Health Organization, stroke affects more than 15 million people worldwide and directly kills about 5.7 million people. It also causes approximately 5 million people to have a lifelong disability, while about 4.3 million people died due to disability. At present, thrombolytic therapy (Castellanos et al., 2018) (recombinant tissue plasminogen activator) is the only acute treatment for ischemic stroke with a narrow time window (3–4.5 hours). Therefore, only 3.4%–5.2% of patients were treated within the short time window. Researchers have been focusing on how to improve the clinical diagnosis and treatment of cerebral infarction beyond the time window of thrombolysis (Feil et al., 2019).

The occurrence and development of ischemic stroke is affected by a variety of risk factors, such as family history of stroke (Zheng et al., 2019), history of heart disease (Beck et al., 2018), history of diabetes (Zou et al., 2018), history of hypertension, etc. According to the investigation and analysis of Li et al. (2019), the prevalence rate of the family with a family history of stroke is 10.52%. In recent years, a number of genetic association studies have suggested that there are multiple genetic risk factors for ischemic stroke, and multiple risk loci were found to affect the susceptibility to ischemic stroke.

Cacabelos et al. (2018) and Yee et al. (2019) showed that the C7673T polymorphism of APOB gene was significantly associated with the risk of ischemic stroke. Chen et al. (2019), Nordestgaard et al. (2018) confirmed that the polymorphism of ϵ 2,ϵ3,ϵ4 of APOE gene was associated with ischemic stroke. APOB gene and APOE gene are both known ischemic stroke susceptibility genes because of blood lipid level. In addition, many studies have shown that the SG13S114 (rs10507391) polymorphism of ALOX5AP gene and SG13S32 (rs9551963) polymorphism are associated with susceptibility to ischemic stroke. Zheng et al., (2018) found that carriers of SG13S114 polymorphism TT/TA genotype of ALOX5AP gene had a higher risk of acute cerebral infarction. Naderi et al. (2019) showed that SG13S114 polymorphism of ALOX5AP gene was associated with acute cerebral infarction. Previous genetic studies have found that some ischemic stroke susceptibility genes on chromosome 14, such as GCH1 gene (Wei et al., 2018), MEG3 gene (Han et al., 2018), MMP-14 gene (Elgebaly et al., 2019), PRKCH gene (Krupinski et al., 2018), are associated with the risk of ischemic stroke.

Genome-wide association study (GWAS) reveals candidate loci, susceptible genes and their loci related to the occurrence, development and treatment of diseases by genome-wide high-density genetic markers (Pei Li and Wang, 2015; Cheng et al., 2019a; Cheng et al., 2019b). Since 2009, GWAS has been widely used to explore and excavate candidate gene loci related to new types of stroke. GWAS is generally believed to be able to identify some previously undetected or identified biological markers related to stroke (Ye et al., 2018; Cheng et al., 2019c), and because of its large sample size, it can minimize false positive results. The National Institute of Neurological Diseases (NIND) has conducted the largest and most comprehensive GWAS to explore the genetic loci of stroke and its subtypes. The results supported the previously established genetic association of ischemic stroke. New loci on chromosome 1p13 (such as rs12122341 of TSPAN2 gene) have been found to be associated with ischemic stroke. Although GWAS has many advantages and is widely used, it is still very hard to understand the role of nucleotide polymorphism (SNP) loci in diseases from the huge results of GWAS.

Therefore, recently many researchers have tried to integrate GWAS with expression quantitative trait loci (eQTL) to mine the disease-related genes (Cheng et al., 2018a; Cheng et al., 2018b). Since eQTL conveys gene expression information and GWAS conveys disease-related SNPs information, combining the two datasets, we could know the loci which are associated with diseases because of affecting other genes expression. Zhao et al. (2019) found many Alzheimer's disease-related genes and SNPs by GWAS and eQTL. Asthma-related genes were identified by Li et al. (2015). by integrating GWAS and eQTL. Systematic integration of Brain eQTL and GWAS were done by Luo et al. (2015) and they identified ZNF323 as a novel Schizophrenia risk gene.

Zhu et al. (2016) generalized Mendelian randomization to SMR. SMR is used to test the association between a trait and the expression level of each gene across the whole genome using summary data from GWAS and eQTL studies. SMR is a common tool to identify the genes whose expression levels are associated with a complex trait because of pleiotropy. Twenty-eight GWAS datasets are used by Pavlides et al. (2016) to find genes whose expression levels were associated with complex phenotype. Bone mineral density (BMD)-related genes are studied by Meng et al. (2018) using SMR. SMR is also used to identify genes and pathways for Amyotrophic Lateral Sclerosis by Du et al. (2017). Fan et al. (2017) found 6 genes are associated with neuroticism by SMR. Liu et al. (2018) used SMR on doing research on Obesity and found 20 BMI associated genes. Veturi and Ritchie (2018) compared two popular methods: MP and SMR by different datasets. Though these scholars' researches, we could judge that SMR is an effective tool. In this paper, summary-level data mendelian randomization (SMR) is used to integrate GWAS and eQTL datasets. In this way, the most functionally relevant genes at the loci identified in GWAS for stroke are found.

METHODS

Work Frame

As shown in **Figure 1**, since GWAS has identified SNPs which are related to stroke, and eQTL has identified SNPs which can affect genes expression, SMR is used to identify SNPs that can change gene expression and this should be the reason that they are associated with stroke. Therefore, firstly, we should obtain GWAS and eQTL data. Then, we checked the overlap between these two datasets. Finally, SMR is used to screen SNPs.

SMR

z in summary data level Mendelian Randomization (SMR) is a genetic variant (SNP), x is the expression level of a gene and y denotes the trait, then the two-step least-squares estimate of the effect of x on y from an MR analysis is:

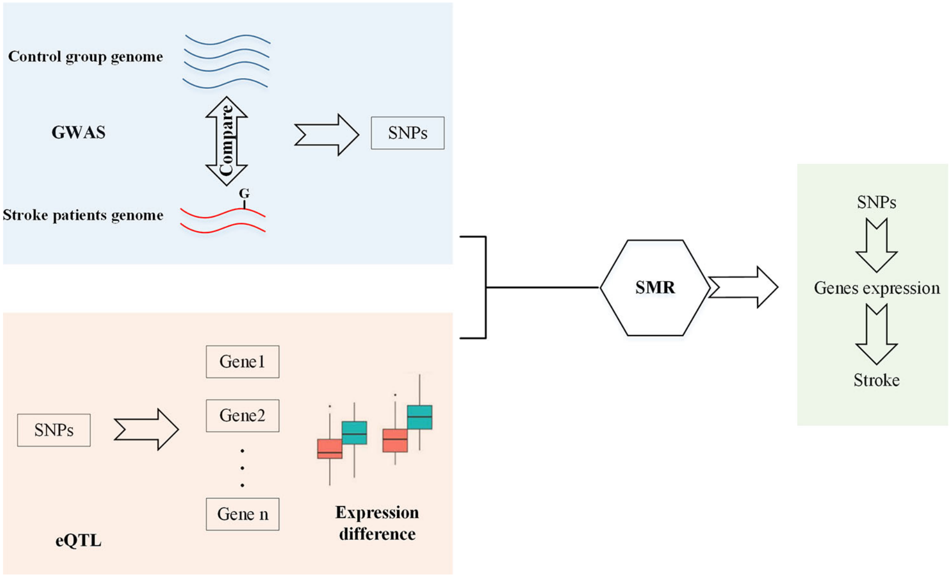


FIGURE 1 | Workflow of SMR.

$$\hat{b}_{xy} = \hat{b}_{zy}/\hat{b}_{zx} \tag{1}$$

$\hat{b}_{zy}$ and $\hat{b}_{zx}$ are the least-squares estimates of y and x on z, respectively. Then, $\hat{b}_{xy}$ denotes the effect size of x on y without confounding from non-genetic factors. The variance of $\hat{b}_{xy}$ is:

$$T_{MR} = \hat{b}^2_{xy}/\mathrm{var}(\hat{b}_{xy}) \tag{2}$$

Here, T_{MR} obeys a chi-square distribution with a degree of freedom of 1. As we can see in equation (Dargazanli et al., 2018), MR requires genotype, gene expression and phenotype to be measured on the same sample. However, Zhu et al. have proved that the power of detecting $\hat{b}_{xy}$ can be greatly increased using a two-sample MR analysis. Therefore, the T_{MR} can be replaced by T_{SMR}.

$$T_{SMR} = \hat{b}^2_{xy}/\mathrm{var}(\hat{b}_{xy}) \approx \frac{z^2_{zy}z^2_{zx}}{z^2_{zy} + z^2_{zx}} \tag{3}$$

z_{zy} is the z statistics from GWAS and z_{zx} is the z statistics from eQTL.

RESULTS

Data Description

GWAS

We used the data from Malik et al.'s research. Eight GWAS datasets are used. **Table 1** shows the detailed information about these data.

We collected GWAS data for four different types of stroke (ischemic stroke, large artery stroke, cardioembolic stroke, small vessel stroke).

Figure 2 shows P value of SNPs in GWAS1 and GWAS2. The SNPs are almost same in these GWAS dataset, but difference races cause the difference of P value. We could know different races have different stroke susceptibility genes.

eQTL

eQTL data is from a meta-analysis of GTEx brain (Consortium G, 2017), CMC (Fromer et al., 2016), and ROSMAP (Ng et al., 2017). All the data are from brain. Only SNPs within 1Mb distance from each probe are available. The estimated effective n is 1,194.

TABLE 1 | GWAS data description.

Dataset	Disease	Sample
GWAS 1	ischemic stroke	Europeans (40,585 cases; 406,111 controls)
GWAS 2	ischemic stroke	trans-ethnic meta-analysis (67,162 cases; 454,450 controls)
GWAS 3	large artery stroke	Europeans (40,585 cases; 406,111 controls)
GWAS 4	large artery stroke	trans-ethnic meta-analysis (67,162 cases; 454,450 controls)
GWAS 5	cardioembolic stroke	Europeans (40,585 cases; 406,111 controls)
GWAS 6	cardioembolic stroke	trans-ethnic meta-analysis (67,162 cases; 454,450 controls)
GWAS 7	small vessel stroke	Europeans (40,585 cases; 406,111 controls)
GWAS 8	small vessel stroke	trans-ethnic meta-analysis (67,162 cases; 454,450 controls)

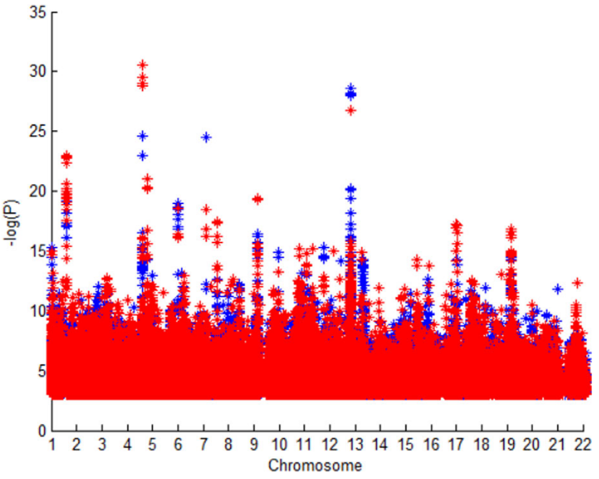


FIGURE 2 | P value of SNPs in GWAS1 and GWAS2.

mQTL

mQTL used in this paper is a set of brain data from a meta-analysis of ROSMAP (Ng et al., 2017), Hannon et al. (2016) and Jaffe et al. (2016). In the ROSMAP data, only SNPs within 5Kb of each DNA methylation probe are available. In the Hannon et al. data, only SNPs within 500Kb distance from each probe and with PmQTL < 1.0e-10 are available. In the Jaffe et al. data, only SNPs within 20Kb distance from each probe and with FDR < 0.1 are available. The estimated effective n is 1,160.

Four Kinds of Stroke

Ischemic stroke is a kind of stroke which caused by arterial obstruction. It accounts for approximately 85% of the total. large artery stroke and cardioembolic stroke are the subgroup of this kind of this stroke.

Large artery stroke is caused by blood clots (thrombus) which are formed in the neck or cerebral arteries. There may be accumulation of fatty deposits (often referred to as plaques) in these arteries.

Cardioembolic stroke is caused by blood clots that reach the brain and blocks the blood vessels. A common cause is the formation of blood clots in the two upper atrial rhythm abnormalities of the heart (atrial fibrillation).

Small vessel stroke is actually a transient stroke symptom that usually lasts only a few minutes. small vessel stroke is caused by transient blood supply to specific parts of the brain and does not cause significant persistent effects on patients. However, it is generally believed that the risk of stroke after small vessel stroke is higher.

SNPs and Genes for Ischemic Stroke

10 SNPs which change six genes expression are screened by Europeans dataset and 11 SNPs which change five genes expression are screened by trans-ethnic dataset.

As we can see in **Table 2**, HSD17B12 is overlapped in the two tests. Moreno et al. (2018) found upregulation of HSD17B12 is associated ischemic stroke using 82 cases and 67 controls. ALDH2 is generally considered as a gene (Guo et al., 2013) which can protect against ischemic stroke, because overexpression of ALDH2 rescued neuronal survival against 4-HNE treatment in PC12 cells (Lee et al., 2012). These two genes show the accuracy of our results.

TABLE 2 | SMR results of ischemic stroke.

SNP	P-value	Gene
Europeans dataset		
rs9651613	4.17E-06	HSD17B12
rs648997	5.72E-06	ALDH2
rs11065976	6.36E-06	ALDH2
rs4286007	6.70E-06	CKAP2
rs847892	7.79E-06	ALDH2
rs66480035	7.97E-06	ALDH2
rs532436	7.99E-06	SURF1
rs487399	8.21E-06	CEP192
rs11618716	8.80E-06	CKAP2
rs11620062	9.24E-06	CKAP2
Trans-ethnic dataset		
rs9651613	3.58E-07	HSD17B12
rs10838185	5.14E-06	HSD17B12
rs6599175	5.42E-06	ULK4
rs6801343	5.55E-06	ULK4
rs9874975	5.70E-06	ULK4
rs12774577	7.99E-06	C10orf32
rs10400343	8.38E-06	HSD17B12
rs3087681	8.47E-06	C10orf32
rs2371623	8.81E-06	ULK4
rs9825741	9.00E-06	ULK4
rs11191606	9.04E-06	C10orf32

SNPs and Genes for Large Artery Stroke

None SNP is screened by Europeans dataset for large artery stroke. Three SNPs which correspond one gene 'C3orf18' are screened by trans-ethnic dataset.

Phenotypes for C3orf18 Gene include Decreased homologous recombination repair frequency, Decreased ionizing radiation sensitivity, Upregulation of Wnt pathway, Increased vaccinia virus (VACV) infection, Mildly decreased CFP-tsO45G cell surface transport. It is considered to be associated with cognitive function measurement.

SNPs and Genes for Cardioembolic Stroke

11 SNPs are significant in Europeans dataset and trans-ethnic dataset. rs3807989 is screened more than one time in Europeans dataset because it can affect more than one gene expression. Both CAV1 and CAV2's expression can be changed by this SNP.

As we can see in **Table 3**, 6 genes and 3 genes are screened by SMR in Europeans dataset and Trans-ethnic dataset, respectively. Three of them are overlapped.

SNPs and Genes for Small Vessel Stroke

13 SNPs and 4 SNPs are significant in Europeans dataset and trans-ethnic dataset, respectively. None of these SNPs or their corresponding genes are overlapped in these two tests. As we can see in **Table 4**, although no overlap is found between these two

TABLE 3 | SMR results of cardioembolic stroke.

SNP	P-value	Gene
Europeans dataset		
rs3807989	2.03E-05	CAV1
rs532436	4.03E-05	SURF1
rs72790984	4.68E-05	PLEKHH2
rs11773845	4.96E-05	CAV1
rs4745721	4.96E-05	ECD
rs1997571	5.62E-05	CAV1
rs507666	6.08E-05	SURF1
rs1997572	6.20E-05	CAV1
rs9313620	6.27E-05	BNIP1
rs76192127	6.35E-05	ECD
rs3807989	6.58E-05	CAV2
rs2519093	7.40E-05	SURF1
rs600038	9.67E-05	SURF1
Trans-ethnic dataset		
rs4745721	2.21E-05	ECD
rs76192127	2.87E-05	ECD
rs532436	3.37E-05	SURF1
rs507666	4.15E-05	SURF1
rs616154	5.26E-05	SURF1
rs72790984	5.60E-05	PLEKHH2
rs2519093	5.92E-05	SURF1
rs72790983	6.37E-05	PLEKHH2
rs559723	7.26E-05	SURF1
rs183153921	7.34E-05	ECD
rs3878005	9.62E-05	ECD

TABLE 4 | SMR results of small vessel stroke.

SNP	P-value	Gene
Europeans dataset		
rs3807989	2.03E-05	CAV1
rs532436	4.03E-05	SURF1
rs72790984	4.68E-05	PLEKHH2
rs11773845	4.96E-05	CAV1
rs4745721	4.96E-05	ECD
rs1997571	5.62E-05	CAV1
rs507666	6.08E-05	SURF1
rs1997572	6.20E-05	CAV1
rs9313620	6.27E-05	BNIP1
rs76192127	6.35E-05	ECD
rs3807989	6.58E-05	CAV2
rs2519093	7.40E-05	SURF1
rs600038	9.67E-05	SURF1
Trans-ethnic dataset		
rs2501966	3.53E-06	CENPQ
rs6599175	4.49E-06	ULK4
rs2501965	4.77E-06	CENPQ
rs9874975	6.07E-06	ULK4

tests, some genes are overlapped between cardioembolic stroke and small vessel stroke.

SNPs Changes Gene Expression Level by Methylation

Since both genetic and environmental factors are key to cause stroke, while methylation plays an important role in the interaction between environmental factors and genetic expression, we assumed that some of the SNPs identified above are at greater risk of methylation and can change gene expression levels.

Therefore, we integrated the SNPs found above with mQTL data for research. Thirty-eight unique SNPs are found in four different types of stroke. Thirty-one of these 38 SNPs are significant in mQTL dataset. We draw the P value of these 31 SNPs as **Figure 2**. As shown in **Figure 3**, most of these SNPs are associated with several genes expression. In addition, most of SNPs have a quite low P value, which means that they can significant change the expression of genes.

Case Study

ULK4

Guo et al. (2016) have found that genetic variants in LRP1 and ULK4 are associated with acute aortic dissections. In their paper, they also mentioned that ULK4 may contribute stroke.

CAV1

Shyu et al. (2017) discussed association of eNOS and CAV1 gene polymorphisms with susceptibility risk of large artery atherosclerotic stroke. A tendency toward an increased LAA stroke risk was significant in carriers with the eNOS Glu298Asp variant in conjunction with the G14713 A and T29107A polymorphisms of the CAV1 (aOR = 2.03, P-trend = 0.002).

CAV2

Jolobe (2012) found that recurrent stroke is because of a novel voltage sensor mutation in CAV2. They compared stroke mouse and normal mouse to obtain this conclusion.

CONCLUSIONS

Stroke is the primary cause of disability in adults, which constitutes a serious public health burden. Stroke is generally believed to be caused by genetic and environmental factors. Therefore, in this paper, we identified stroke-related genes and loci from both genetic and environmental aspects.

GWAS identified a large number of stroke-related SNPs, which were difficult to explain. We tried to identify the pathogenesis of significant SNPs by combining SMR with eQTL data. Since eQTL shows the SNPs that can significantly change genes expression and GWAS shows the SNPs that are significant related to stroke, we combined these two data to identify the genes whose expression levels are associated with stroke because of pleiotropy.

38 SNPs which cause changes in 14 genes expression were found by 8 GWAS data and brain eQTL. Those 8 GWAS data are from two different races sample and include four types of stroke (ischemic stroke, large artery stroke, cardioembolic stroke, small vessel stroke). CAV1, SURF1, PLEKHH2, ECD, BNIP1, CAV2 are found to be associated with cardioembolic stroke and Small vessel stroke in Europeans. ULK4 is a susceptibility gene for ischemic stroke and small vessel stroke.

Since methylation (Lv et al., 2019) plays an important role in the interaction between environmental factors and genetic expression, we tried to find out whether 38 SNPs are affected by methylation and lead to the changes in other genes expression levels. Thirty-one of these 38 SNPs are significant in mQTL data and most of them can affect more than one gene expression.

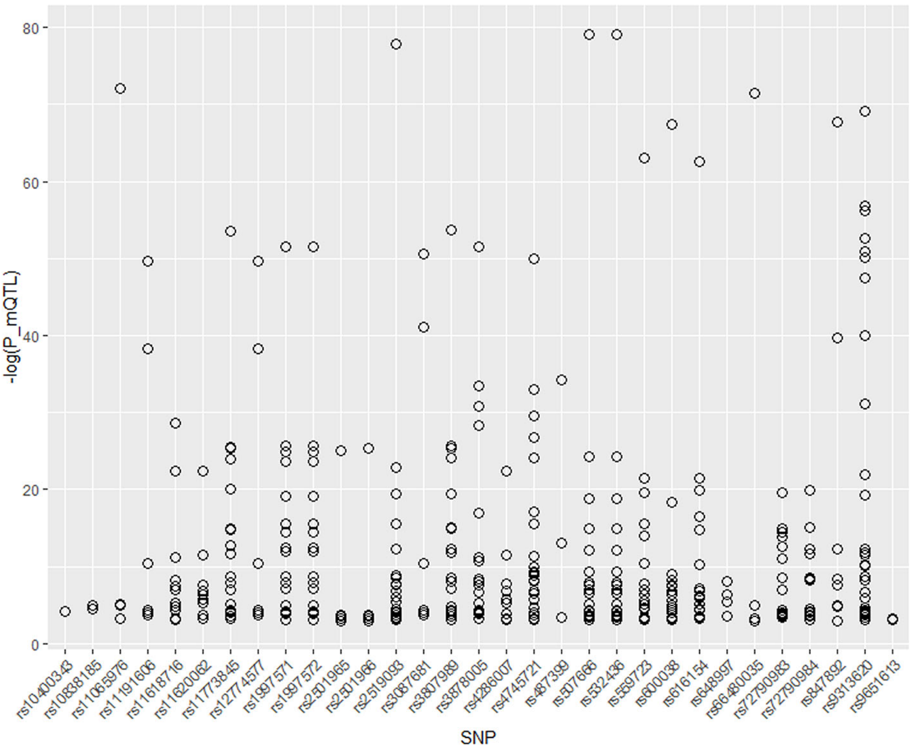


FIGURE 3 | P value of 31 significant SNPs in mQTL.

Overall, integrating GWAS with eQTL, we found 38 SNPs and 14 genes are related to stroke by SMR. Thirty-one of 38 SNPs are at high risk of methylation which can also cause changes in gene expression. These findings serve as a guide to understanding the pathogenesis of stroke at the molecular level.

DATA AVAILABILITY STATEMENT

All the datasets used in this paper could be downloaded from GWAS: ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST006908/MEGASTROKE.2.AIS.EU

R.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST005843/MEGASTROKE.2.AIS.TR

ANS.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST006907/MEGASTROKE.3.LAS.EU

R.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST005840/MEGASTROKE.3.LAS.TR

ANS.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST005842/MEGASTROKE.4.CES.TR

ANS.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST006910/MEGASTROKE.4.CES.EU

R.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST005841/MEGASTROKE.5.SVS.TR

ANS.out ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/MalikR_29531354_GCST006909/MEGASTROKE.5.SVS.EU

R.out eQTL: https://cnsgenomics.com/software/smr/#eQTLsummarydata

mQTL: https://cnsgenomics.com/software/smr/#mQTLsummarydata.

AUTHOR CONTRIBUTIONS

HuJ, Z-HL, and HoJ conceived and designed the experiments. SZ analyzed data. SZ, HuJ, Z-HL, and HoJ wrote this manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by grants from the National Natural Science Foundation of China (81671760 and 81873910), Scientific Research Transformation Special Fund of Heilongjiang Academy of Medical Sciences (2018415);Scientific Research Project of Health and Family Planning Commission of Heilongjiang Province (201812 and 201622), National Natural Science Foundation of China (81871423), and Shanghai Municipal Commission of Health and Family Planning (20160064).

REFERENCES

Beck, J. D., Moss, K. L., Morelli, T., and Offenbacher, S. (2018). Periodontal profile class is associated with prevalent diabetes, coronary heart disease, stroke, and systemic markers of C-reactive protein and interleukin-6. *J. periodontol.* 89 (2), 157–165. doi: 10.1002/jper.17-0426

Cacabelos, R., Lombardi, V., Fernández-Novoa, L., Carrera, I., Cacabelos, P., Corzo, L., et al. (2018). "Basic and Clinical Studies With Marine LipoFishins and Vegetal Favalins in Neurodegeneration and Age-Related Disorders," in *Studies in Natural Products Chemistry*, vol. 59. (Netherlands: Elsevier), 195–225.

Castellanos, M., van Eendenburg, C., Gubern, C., Kádár, E., Huguet, G., Puig, J., et al. (2018). Low levels of caveolin-1 predict symptomatic bleeding after thrombolytic therapy in patients with acute ischemic stroke. *Stroke* 49 (6), 1525–1527. doi: 10.1161/strokeaha.118.020683

Chen, J.-X., Liu, J., Hu, F., Bi, Y., Li, M., and Zhao, L. (2019). Genetic variants on chromosome 9p21 confer risks of cerebral infarction in the Chinese population: a meta-analysis. *Int. J. immunopathol. Pharmacol.* 33, 2058738419847852. doi: 10.1177/2058738419847852

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34 (11), 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018b). Exposing the causal effect of C-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. In Genet.* 9, 657. doi: 10.3389/fgene.2018.00657

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019a). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48 (D1), D554–D560 doi: 10.1093/nar/gkz843

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019b). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47 (D1), D140–D144. doi: 10.1093/nar/gky1051

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J., Tan, R., et al. (2019c). Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. In Genet.* 10, 94. doi: 10.3389/fgene.2019.00094

Consortium G (2017). Genetic effects on gene expression across human tissues. *Nature* 550 (7675), 204. doi: 10.1038/nature24277.

Dargazanli, C., Fahed, R., Blanc, R., Gory, B., Labreuche, J., Duhamel, A., et al. (2018). Modified thrombolysis in cerebral infarction 2c/thrombolysis in cerebral infarction 3 reperfusion should be the aim of mechanical thrombectomy: insights from the ASTER Trial (Contact Aspiration Versus Stent Retriever for Successful Revascularization). *Stroke* 49 (5), 1189–1196. doi: 10.1161/strokeaha.118.020700

Du, Y., Yan, W., Guo, X., Hao, J., Wang, W., He, A., et al. (2017). A genome-wide expression association analysis identifies genes and pathways associated with amyotrophic lateral sclerosis. *Cell. Mol. Neurobiol.* 38 (3), 1–5. doi: 10.1007/s10571-017-0512-2

Elgebaly, M. M., Arreguin, J., and Storke, N. (2019). Targets, treatments, and outcomes updates in diabetic stroke. *J. Stroke Cerebrovasc. Dis.* 28 (6), 1413–1420 doi: 10.1016/j.jstrokecerebrovasdis.2019.02.005

Fan, Q., Wang, W., Hao, J., He, A., Wen, Y., Guo, X., et al.(2017). Integrating genome-wide association study and expression quantitative trait loci data identifies multiple genes and gene set associated with neuroticism. *Prog. In Neuropsychopharmacol. Biol. Psychiatry* 28 (6), 1413–1420. doi: 10.1016/j.pnpbp.2017.05.017

Feil, K., Reidler, P., Kunz, W. G., Küpper, C., Heinrich, J., Laub, C., et al. (2019). Addressing a real life problem: treatment with intravenous thrombolysis and mechanical thrombectomy in acute stroke patients with an extended time window beyond 4.5 hours based on computed tomography perfusion imaging. *Eur. J. Neurol.* 27 (1), 168–174 doi: 10.7861/clinmedicine.17-2-161

Fromer, M., Roussos, P., Sieberts, S. K., Johnson, J. S., Kavanagh, D. H., Perumal, T. M., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* 19 (11), 1442. doi: 10.1038/nn.4399

Guo, J.-M., Liu, A.-J., Zang, P., Dong, W.-Z., Ying, L., Wang, W., et al. (2013). ALDH2 protects against stroke by clearing 4-HNE. *Cell Res.* 23 (7), 915. doi: 10.1038/cr.2013.69

Guo, D-c, Grove, M. L., Prakash, S. K., Eriksson, P., Hostetler, E. M., LeMaire, S. A., et al. (2016). Genetic variants in LRP1 and ULK4 are associated with acute aortic dissections. *Am. J. Hum. Genet.* 99 (3), 762–769. doi: 10.1016/j.ajhg.2016.06.034

Han, X., Zheng, Z., Wang, C., and Wang, L. (2018). Association between MEG3/miR-181b polymorphisms and risk of ischemic stroke. *Lipids In Health Dis.* 17 ((1)), 292. doi: 10.1186/s12944-018-0941-z

Hannon, E., Spiers, H., Viana, J., Pidsley, R., Burrage, J., Murphy, T. M., et al. (2016). Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* 19 (1), 48. doi: 10.1038/nn.4182

Jaffe, A. E., Gao, Y., Deep-Soboslay, A., Tao, R., Hyde, T. M., Weinberger, D. R., et al. (2016). Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat. Neurosci.* 19 (1), 40. doi: 10.1038/nn.4181

Jolobe, O. M. (2012). Stroke and familial hemiplegic migraine. *Lancet Neurol.* 11 (6), 484. doi: 10.1016/s1474-4422(12)70123-0

Krishnamurthi, R. V., Barker-Collo, S., Parag, V., Parmar, P., Witt, E., Jones, A., et al. (2018). Stroke incidence by major pathological type and ischemic subtypes in the Auckland regional community stroke studies: changes between 2002 and 2011. *Stroke* 49 (1), 3–10. doi: 10.1161/strokeaha.117.019358

Krupinski, J., Carrera, C., Muiño, E., Torres, N., Al-Baradie, R., Cullell, N., et al. (2018). DNA methylation in stroke. Update of latest advances. *Comput. Struct. Biotechnol. J.* 16, 1–5. doi: 10.1016/j.csbj.2017.12.001

Lee, W.-C., Wong, H.-Y., Chai, Y.-Y., Shi, C.-W., Amino, N., Kikuchi, S., et al. (2012). Lipid peroxidation dysregulation in ischemic stroke: plasma 4-HNE as a potential biomarker? *Biochem. Biophys. Res. Commun.* 425 (4), 842–847. doi: 10.1016/j.bbrc.2012.08.002

Li, X., Hastie, A. T., Hawkins, G. A., Moore, W. C., Ampleford, E. J., Milosevic, J., et al. (2015). eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Allergy* 70 (10), 1309–1318. doi: 10.1111/all.12683

Li, W., Wang, D., Wang, X., Gong, Y., Cao, S., Yin, X., et al. (2019). The association of metabolic syndrome components and diabetes mellitus: evidence from China National Stroke Screening and Prevention Project. *BMC Public Health* 19 (1), 192. doi: 10.1186/s12889-019-6415-z

Liu, L., Fan, Q., Zhang, F., Guo, X., Liang, X., Du, Y., et al. (2018). A genomewide integrative analysis of GWAS and eQTLs Data identifies multiple genes and gene sets associated with obesity. *BioMed. Res. Int.* 2018, 3848560. doi: 10.1155/2018/3848560

Luo, X.-J., Mattheisen, M., Li, M., Huang, L., Rietschel, M., Børglum, A. D., et al. (2015). Systematic integration of brain eQTL and GWAS identifies ZNF323 as a novel schizophrenia risk gene and suggests recent positive selection based on compensatory advantage on pulmonary function. *Schizophr. Bull.* 41 (6), 1294–1308. doi: 10.1093/schbul/sbv017

Lv, H., Zhang, Z. M., Li, S. H., Tan, J. X., Chen, W., and Lin, H. (2019). Evaluation of different computational methods on 5-methylcytosine sites identification. *Briefings In Bioinf.* doi: 10.1093/bib/bbz048

Meng, X. H., Chen, X. D., Greenbaum, J., Zeng, Q., You, S. L., Xiao, H. M., et al. (2018). Integration of summary data from GWAS and eQTL studies identified novel causal BMD genes with functional predictions. *Bone* 113, 41–48. doi: 10.1016/j.bone.2018.05.012

Moreno-Ramírez, C. E., Gutiérrez-Garzón, E., Barreto, G. E., and Forero, D. A. (2018). Genome-wide expression profiles for ischemic stroke: a meta-analysis. *J. Stroke Cerebrovasc. Dis.* 27 (11), 3336–3341. doi: 10.1016/j.jstrokecerebrovasdis.2018.07.035

Naderi, N., Yousefi, H., Mollazadeh, S., Seyed Mikaeili, A., Keshavarz Norouzpour, M., Jazebi, M., et al. (2019). Inflammatory and immune response genes: a genetic analysis of inhibitor development in Iranian hemophilia A patients. *Pediatr. Hematol. Oncol.* 36 (1), 28–39. doi: 10.1080/08880018.2019.1585503

Ng, B., White, C. C., Klein, H.-U., Sieberts, S. K., McCabe, C., Patrick, E., et al. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* 20 (10), 1418. doi: 10.1038/nn.4632

Nordestgaard, L. T., Tybjærg-Hansen, A., Rasmussen, K. L., Nordestgaard, B. G., and Frikke-Schmidt, R. (2018). Genetic variation in clusterin and risk of dementia and ischemic vascular disease in the general population: cohort studies and meta-analyses of 362,338 individuals. *BMC Med.* 16 (1), 39. doi: 10.1016/j.atherosclerosis.2018.06.075

Pavlides, J. M. W., Zhu, Z., Gratten, J., Mcrae, A. F., Wray, N. R., and Yang, J. (2016). Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* 8 (1), 84. doi: 10.1186/s13073-016-0338-4

Pei Li, M. G., and Wang, C. (2015). Xiaoyan Liu, Quan Zou: An overview of SNP interactions in genome-wide association studies. *Briefings In Funct. Genomics* 14 (2), 143–155. doi: 10.1093/bfgp/elu036

Shyu, H.-Y., Chen, M.-H., Hsieh, Y.-H., Shieh, J.-C., Yen, L.-R., Wang, H.-W., et al. (2017). Association of eNOS and Cav-1 gene polymorphisms with susceptibility risk of large artery atherosclerotic stroke. *PloS One* 12 (3), e0174110. doi: 10.1371/journal.pone.0174110

Sun, W., Han, Y., Yang, S., Zhuang, H., Zhang, J., Cheng, L., et al. (2019). The assessment of interleukin-18 on the risk of coronary heart disease. *Med. Chem.* doi: 10.2174/1573406415666191004115128

Veturi, Y., and Ritchie, M. D. (2018). How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures? *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 23, 228–239. doi: 10.1142/9789813235533_0021

Wei, J., Zhang, Y., Li, Z., Wang, X., Chen, L., Du, J., et al. (2018). GCH1 attenuates cardiac autonomic nervous remodeling in canines with atrial-tachypacing *via* tetrahydrobiopterin pathway regulated by microRNA-206. *Pacing Clin. Electrophysiol.* 41 (5), 459–471. doi: 10.1111/pace.13289

Ye, Z., Zhang, H., Sun, L., Cai, H., Hao, Y., Xu, Z., et al. (2018). GWAS-supported CRP gene polymorphisms and functional outcome of large artery atherosclerotic stroke in Han Chinese. *Neuromol. Med.* 20 (2), 225–232. doi: 10.1007/s12017-018-8485-y

Yee, J., Kim, W., Chang, B. C., Chung, J. E., Lee, K. E., and Gwak, H. S. (2019) APOB gene polymorphisms may affect the risk of minor or minimal bleeding complications in patients on warfarin maintaining therapeutic INR. *Eur. J. Hum. Genet.* 27 (10), 1542–1549. doi: 10.1038/s41431-019-0450-1

Zhao, T., Hu, Y., Zang, T., Wang, Y., and Integrate, G. W. A. S. (2019). eQTL, and mQTL data to identify alzheimer's disease-related genes. *Front. In Genet.* 10, 1021. doi: 10.3389/fgene.2019.01021

Zheng, Z., Liu, S., Wang, C., and Han, X. (2018). A functional polymorphism rs145204276 in the promoter of long noncoding RNA GAS5 is associated with an increased risk of ischemic stroke. *J. Stroke Cerebrovasc. Dis.* 27 (12), 3535–3541. doi: 10.1016/j.jstrokecerebrovasdis.2018.08.016

Zheng, X., Zeng, N., Wang, A., Zhu, Z., Peng, H., Zhong, C., et al. (2019). Family history of stroke and death or vascular events within one year after ischemic stroke. *Neurol. Res.* 41 (5), 466–472. doi: 10.1080/01616412.2019.1577342

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* 48 (5), 481. doi: 10.1038/ng.3538

Zhuang, H., Zhang, Y., Yang, S., Cheng, L., and Liu, S. L. (2019). A mendelian randomization study of infant length and type 2 diabetes mellitus risk. *Curr. Gene Ther.* 19 (4), 224–231(8) doi: 10.2174/1566523219666190925115535

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Front. In Genet.* 9, 515. doi: 10.3389/fgene.2018.00515

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: 05 February 2020
doi: 10.3389/fgene.2020.00029

CHG: A Systematically Integrated Database of Cancer Hallmark Genes

Denan Zhang[1†], Diwei Huo[2†], Hongbo Xie[1†], Lingxiang Wu[1†], Juan Zhang[1], Lei Liu[1], Qing Jin[1] and Xiujie Chen[1]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] The 2nd Affiliated Hospital of Harbin Medical University, Harbin, China

Background: The analysis of cancer diversity based on a logical framework of hallmarks has greatly improved our understanding of the occurrence, development and metastasis of various cancers.

Methods: We designed Cancer Hallmark Genes (CHG) database which focuses on integrating hallmark genes in a systematic, standard way and annotates the potential roles of the hallmark genes in cancer processes. Following the conceptual criteria description of hallmark function the keywords for each hallmark were manually selected from the literature. Candidate hallmark genes collected were derived from 301 pathways of KEGG database by Lucene and manually corrected.

Results: Based on the variation data, we finally identified the hallmark genes of various types of cancer and constructed CHG. And we also analyzed the relationships among hallmarks and potential characteristics and relationships of hallmark genes based on the topological structures of their networks. We manually confirm the hallmark gene identified by CHG based on literature and database. We also predicted the prognosis of breast cancer, glioblastoma multiforme and kidney papillary cell carcinoma patients based on CHG data.

Conclusions: In summary, CHG, which was constructed based on a hallmark feature set, provides a new perspective for analyzing the diversity and development of cancers.

Keywords: Hallmark genes, mutation, methylation, copy number variation, annotating Hallmark features, database

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Jiajie Peng,
Northwestern Polytechnical University, China
Xuekun Ren,
Harbin Institute of Technology, China
Qiang Lei,
Harbin Institute of Technology, China

****Correspondence:***
Xiujie Chen
chenxiujie@ems.hrbmu.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *01 July 2019*
Accepted: *09 January 2020*
Published: *05 February 2020*

Citation:
Zhang D, Huo D, Xie H, Wu L, Zhang J, Liu L, Jin Q and Chen X (2020) CHG: A Systematically Integrated Database of Cancer Hallmark Genes. Front. Genet. 11:29. doi: 10.3389/fgene.2020.00029

INTRODUCTION

In 2000, Weinberg et al. (2000) first proposed six hallmarks of cancer, including Sustaining Proliferative Signaling (SPS), Evading Growth Suppressors (EGS), Resisting Cell Death (RCD), Enabling Replicative Immortality (ERI), Inducing Angiogenesis (IA), and Activating Invasion and Metastasis (AIM), which provided a logical framework for conceptualizing a variety of neoplastic diseases. In 2011, they added another four hallmarks to more fully capture the features of cancers, including Genome Instability and Mutation (GIM), Tumor-Promoting Inflammation (TPI), Reprogramming Energy Metabolism (REM), and Evading Immune Destruction (EID) (Hanahan and Weinberg, 2011). The hallmarks of cancer capture the most essential phenotypic characteristics of malignant transformation and progression, but numerous factors involved in this multistep process are still unknown to date. It is undoubtedly that the framework constructed by hallmarks has greatly improved the analysis on diversity of cancers. Balázs Győrffy et al. reviewed the available techniques that are capable of and appropriate for determining the characteristic features of each hallmark (Menyhart et al., 2016). Hallmark capabilities are regulated by partially redundant signaling pathways, and the significance of these pathways depends on the tumor's underlying molecular features. Recently, many studies have focused on the integration of various cancer-related pathways or genes for analysis, and they have found some significant results. In 2011, Jie Li et al. identified high-quality breast cancer prognostic markers and metastasis network modules by integrating hallmark-related genes from GO terms (Li et al., 2010). In 2013, Naif Zaman et al. predicted breast cancer subtype-specific drug targets by exploring the modules (including apoptosis, cell proliferation and cell cycle) in a signaling network assessment of mutations and copy number variations (CNVs) (Zaman et al., 2013). These researches strongly emphasized the importance of constructing gene sets for hallmarks. Moreover, the advantages of the analysis based on a hallmark framework are notable: 1) It reduces feature dimension of cancer (more attention will be focused on the significant genes in each hallmark rather than on all genes, which will reduce the large number of passenger genes analyzed). 2) It is explicable (the results of analysis are depicted more easily). 3) It provides a potential avenue for exploring the mechanism of carcinogenesis. However, the overlap rate of the hallmark genes in current studies is low because the studies use different extraction methods. Furthermore, no gene sets have been systematically collected for the different hallmarks thus far, which makes it difficult to clarify the gene alteration features (including mutations, DNA methylations and CNVs) in each hallmark (Wang et al., 2015).

To address this problem, we established a database called Cancer Hallmark Genes in (CHG), which provides gene sets for the ten hallmarks and the corresponding statistical analysis results, including the frequency of different mutation types (e.g., missense, deletion, insertion), methylation and CNV (e.g., loss or gain) for each gene. To maximize the usage of our database, we collected a total of 22697 samples from TCGA and analyzed the variations of mutation, CNV, and methylation of hallmark genes across 34 cancer types.

Furthermore, we analyzed the relationship among ten hallmarks by Fisher's exact test and unsupervised hierarchical clustering (method 2). Eventually, the hallmarks were clustered into four classes: 1) Reprogramming Energy Metabolism (REM). 2) Activating Invasion and Metastasis (AIM), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI), and Sustaining Proliferative Signaling (SPS). 3) Genome Instability and Mutation (GIM). 4) Tumor-Promoting Inflammation (TPI), Evading Immune Destruction (EID), Resisting Cell Death (RCD), and Inducing Angiogenesis (IA).

Even though the hallmark genes identified in the database came from the confirmed literature and databases, we manually confirmed the top 10 altered (mutation, methylation, CNV) genes of each hallmark to further ensure the accuracy of the data. In addition, we also used several of cancers as examples for further analysis with the CHG data to demonstrate the value of this database at a practical level.

The CHG database is freely available at our website: http://www.bio-bigdata.com/CHG/index.html.

MATERIALS AND METHODS

Data for Hallmarks

In this work, 301 pathways were downloaded from KEGG (version 78.0) (Kanehisa et al., 2017). This data was used for Lucene search and extraction of pathway genes. Gene variant data (7,075 samples of mutation in 34 cancers, 6,177 samples of methylation in 20 cancers, 9445 samples of CNV in 33 cancers) from TCGA (Stratton et al., 2013) were downloaded, where the methylated data was selected as JHU_USC (HumanMethylation 450) and BI (Genome_Wide_SNP_6) was selected for CNV data. These data were used to calculate the frequency of gene variation, and the proportion of different types of variation. The data in this article across DNA methylation, mutation and CNV were from the same samples of TCGA database. In the TCGA database, there are strict rules for the sequencing, processing and analysis, etc. of the samples data and provide standardized data downloading. Human protein-protein interaction data was downloaded from HPRD (Keshava Prasad et al., 2009), STRING (Szklarczyk et al., 2011), BioGRID (Chatraryamontri et al., 2013) and HTRIdb (Bovolenta et al., 2012). Human gene regulation data was downloaded from HTRIdb. These data were used to integrate an integrated gene interaction network. The cDNA data (GRCh38 version and GRCh37 version) was downloaded from Ensembl (Flicek et al., 2014). This data was used for the processing of CNV data (**Supplementary Table 3**).

The Construction Process of the CHG Database

Following the conceptual criteria description of hallmark function in the article "Hallmarks of Cancer: The Next Generation," published in Cell in 2011, we searched the relevant literature in PubMed, and screened the high-frequency descriptive vocabulary appearing in the abstract of the literature as the key words of the corresponding Hallmark. The core idea of our CHG database is to transform the conceptual description of Hallmark features into real biological processes and their corresponding entities. So, we built a process that consists of three main steps (**Figure 1**).

First, we identify the Hallmark description keyword. This step is to materialize the conceptual description of the Hallmark feature. The relevant literature is determined by searching the Hallmark feature description in the literature, and the specific descriptors associated with each Hallmark feature are determined by identifying the high frequency vocabulary in the relevant document abstract. In this step, we manually confirmed the results from the literature scan. In addition to determining that the identified keywords are related to the Hallmark feature, some of the words without more information such as "cancer"

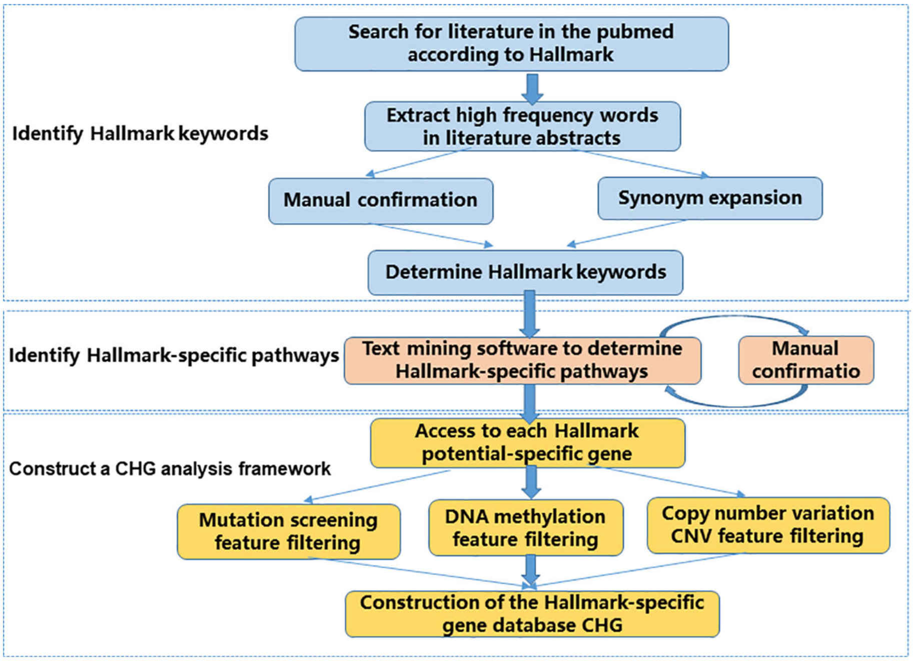

Search for literature in the pubmed according to Hallmark
Identify Hallmark keywords
Extract high frequency words in literature abstracts
Manual confirmation
Synonym expansion
Determine Hallmark keywords
Identify Hallmark-specific pathways
Text mining software to determine Hallmark-specific pathways
Manual confirmatio
Access to each Hallmark potential-specific gene
Construct a CHG analysis framework
Mutation screening feature filtering
DNA methylation feature filtering
Copy number variation CNV feature filtering
Construction of the Hallmark-specific gene database CHG

FIGURE 1 | CHG construction flow chart. The CHG database uses a process consisting of three main steps to transform a conceptual description of Hallmark features into real biological processes and their corresponding entities.

and "tumor" are not directly provided to vocabulary. At the same time, we also further enrich the identified Hallmark description keywords through synonym expansion, for example, "apoptosis" and "cell death" (**Supplementary Table 1**).

Second, we use a text mining software package Lucene to identify the Hallmark-specific pathways in the literature and KEGG database based on the Hallmark description keywords identified in the previous step. The result of the identification is manually confirmed again. The manual confirmation step does not add any subjective results, and only in the case of certainty, significant unrelated results due to software recognition errors are removed (**Supplementary Tables 1** , **2**).

Finally, genes with potential specificity in the potential Hallmark-specific pathway were screened from gene mutation level, epigenetic level, and CNV level to construct CHG.

Cancer Type-Specific Variant Gene

Based on the variation data in TCGA (Montenegro et al., 2015), we calculated the variations of mutation, methylation and CNV for these hallmark genes in different types of cancers. Mutation, CNV, and methylation signatures were used as part of the filtration function in the Hallmark-specific gene screening process in our construction of the CHG database. This is because the relationship between these features and cancer has been confirmed in extensive and in-depth discussions in many previous studies (Kan et al., 2010; Kandoth et al., 2013; Laddha et al., 2014; Wu et al., 2017; Bouras et al., 2019; Sina et al., 2019; Tate et al., 2019). The variations in the characteristics of these different types of cancer not only provide more detailed information for analysis based on the hallmarks but also can be used as a "fingerprint" of cancer type or progression, and this cancer classification can be used as further guidance in prognosis and clinical treatment (**Supplementary Table 3**).

Gene Mutation

Based on the somatic mutation (level 2) data for the 34 types of cancers in TCGA, the frequency of each mutated gene was calculated in specific cancers(Chung et al., 2016). To account for the specific action of different somatic mutations in different types or periods of cancers, we mainly studied the following six types of somatic mutations: insertion (INS), deletion (DEL), missense mutations (SNP_mis), nonsense mutations (SNP_non), splice site mutations (SNP_spl), and gene silencing (SNP_sil) (Hu et al., 2018). The proportion of mutation types in each type of cancer was also statistically analyzed (Kan et al., 2010; Kandoth et al., 2013).

DNA Methylation

We carried out the following calculations for the level 3 data from 20 human tumors derived from TCGA that simultaneously contained both cancer and control samples (Bouras et al., 2019; Sina et al., 2019):

a. Calculate the methylation beta value of each sample (including cancer and normal samples). For genes with multiple methylation sites, the average beta value represents the gene methylation values. The average beta value of the gene in all normal samples was calculated as the methylation level of the control group (Tate et al., 2019);

b. When the gene methylation absolute beta value between the cancer and control groups was more than 0.5, it was called a methylation altered gene. We calculated the occurrence frequency of methylation variation and the corresponding beta value of each gene (Tate et al., 2019).
c. If the gene's methylated beta value was greater than 0.8 in the cancer samples, it was labeled as H (high), whereas when the methylated beta value was less than 0.2, it was labeled as L (low). We calculated the proportion of genes belonging to H or L (Tate et al., 2019).

Copy Number Variation

We analyzed gene segments for the CNV based on level 3 data derived from TCGA and cDNA data from Ensembl in 33 human tumors that simultaneously contained both cancer and control samples. For each pair of samples, if the CNV occurred in only one sample, the default value of the segment in any other sample was 0. Based on experience, we chose 0.2 and -0.2 as the thresholds for altered CNV genes; we marked the gene as a "gain" when the segment value was greater than 0.2 in the cancer samples and as a "loss" when the segment value was less than -0.2 (Laddha et al., 2014). We counted the frequency of CNV in the genes and the proportion of genes belonging to the "gain" and "loss" categories.

Analysis of Relationships of Hallmarks

We analyzed the relationships among the ten hallmarks by Fisher's exact test and unsupervised hierarchical clustering (Tan et al., 2011; Hashemi et al., 2013). We compared the relationship between the specific gene sets of two hallmarks to the final recognition of the overall relationships among the 10 hallmarks. We separately calculated the number of genes belonging to two hallmarks, only one hallmark and all hallmarks. Based on the null hypothesis of independence between any two hallmarks, we calculated the similarity through Fisher's exact test. Finally, we carried out hierarchical clustering with the 1-P value as the similarity score.

RESULTS

The Features of Hallmark Genes Across Cancers

Genome variation is a common phenomenon in cancer, and it is essential to understanding the internal mechanism and prognosis of the tumor in terms of whether the hallmark-related genes have a generally or specifically altered pattern. To this end, we processed the somatic mutation data, methylation data and copy number variant data for 34 cancers in TCGA and analyzed the frequency of somatic mutations, methylation and CNVs in different cancer types (**Table 1**).

To promote the analysis of carcinogenesis, we mapped the driven mutation, methylation and CNV gene data from TCGA into hallmarks to analyze the altered percentages of all hallmark genes. We found that, among all hallmark genes, 97.39% of the genes were altered by mutation, 33.44% were regulated by methylation, and 84.88% were influenced by CNV (**Figure 2**). In each hallmark, the ratio of genes altered by mutation, methylation and CNV was more than 95% (**Table 2**). These results indicate that the genomic changes in cancer are widespread.

We counted the number of hallmark genes that are mutated, differentially methylated and copied in 34 different cancer types (**Figure 3**). The results showed that the difference among the number of mutated genes in different cancer types is large, and there is a 9-fold difference between the maximum and the minimum number of mutated genes, with 2644 in LIHC (liver hepatocellular carcinoma) and 281 in LAML (acute myeloid leukemia). The largest number of differentially methylated genes is 490 in BRCA (breast invasive carcinoma), and the smallest number is 34 in LUAD (lung adenocarcinoma). The largest number of differentially CNV genes is 1972 in OV (ovarian serous cystadenocarcinoma), and the smallest number is 267 in THYM (thymoma).

We also found that different types of cancer have different alteration characteristics. As shown in **Figure 3**, some cancers, such as SKCM (skin cutaneous melanoma), ESCA (esophageal

TABLE 1 | Numbers of pathways and genes of 10 hallmarks.

Hallmarks of cancer	Num. of pathway	Num. of genes
AIM	9	1,101
ERI	4	302
EGS	4	678
RCD	24	1,150
SPS	27	1,263
EID	15	591
TPI	12	619
GIM	10	221
IA	3	483
REM	9	440

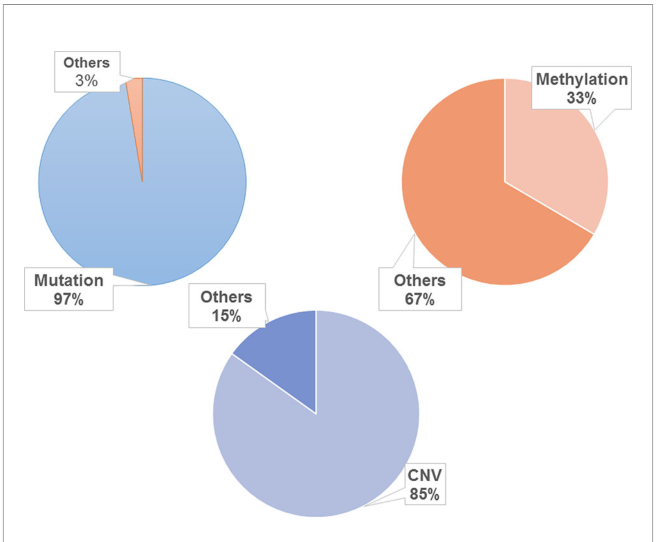


FIGURE 2 | Distribution of genomic changes in 10 hallmarks. The frequency of mutation is about 97.39%, the frequency of methylation is about 33.44% and the frequency of CNV is about 84.88%.

TABLE 2 | Ratio of altered Genes in hallmarks.

Hallmarks	Num. of driven Mutation genes	Num. of driven Methylation genes	Num. of driven CNV genes	alteration genes/all driven genes	Ratio of altered Genes
AIM	1,098	334	1,003	1,098/1,101	99.73%
ERI	301	88	277	301/302	99.67%
EGS	617	234	491	645/678	95.13%
RCD	1147	349	1,025	1,147/1,150	99.74%
SPS	1261	356	1,160	1,261/1,263	99.84%
EID	583	258	506	583/591	98.65%
TPI	614	230	537	614/619	99.19%
GIM	220	73	187	220/221	99.55%
IA	482	198	427	482/483	99.79%
REM	438	95	402	438/440	99.55%

For each hallmark, the ratio of genes altered by mutation, methylation, and CNV were more than 95%.

carcinoma), LIHC (liver hepatocellular carcinoma), mainly reflect the mutation pattern of the genome, and this is a common pattern in most cancers. Some cancers, such as PCPG (pheochromocytoma and paraganglioma), LAML (acute myeloid leukemia), and OV (ovarian serous cystadenocarcinoma), mainly reflect a pattern of CNV variation, which suggests that we should analyze the specific alteration patterns in specific cancers when uncovering the functional importance of the genomic alterations and the underlying mechanisms that drive cancer development, progression and metastasis in different cancer types.

Network of Hallmark Genes

The potential characteristics and relationships of hallmark genes can be effectively revealed based on the topological structures of their networks. Since the hallmark genes were identified from qualitative analysis without any relevant interaction information, we mapped these hallmark genes onto the integrated protein regulatory network to collect data on the interaction and regulation relationships between the hallmark genes and the extract interactions between the hallmark genes, which resulted in the construction of 10 hallmark subnetworks. The average degree of the integrated protein interactions is 36 and 54 in the regulation network and the entire hallmark network (constructed by all the hallmark interaction genes), respectively. This indicates that the interaction between hallmarks is higher than the average level of integrated protein interactions and shows that hallmark networks are more closely linked. On average, for the 10 hallmark subnetworks, 94% of the hallmark genes were involved in the network (**Supplementary Figure 1**). We performed an analysis of the 10 subnetworks and calculated the degree, betweenness and clustering coefficient of all nodes. We found that, in addition to the GIM network in **Figure 4**, the gene interactions inside each hallmark subnetwork were more closely related than the interactions between the 10 hallmark subnetworks. This result may be due to GIM as the basis of other hallmarks; genetic diversity of GIM will lead to in other hallmark features (Hanahan and Weinberg, 2011). At the same time, we also analyzed the correlation between the degree and number of genes in each subnetwork. The results showed that genes with large degrees often also have larger betweenness, as there was a positive correlation between these variables (**Supplementary Figure 1**).

Relationship of Hallmarks

Ten types of hallmarks described different aspects of the tumor characteristics, but there were few relationships mentioned between these characteristics on a pan-cancer scale. To this end, we analyzed the relationship among the hallmarks and divided the ten hallmarks into four classes (**Figure 5**). Interestingly, we found two classes with only one hallmark, namely, *Reprogramming Energy Metabolism (REM)* and *Genome Instability and Mutation (GIM)*. This result is reasonable, as both of these hallmarks are clearly different from the other hallmarks in terms of their mechanisms. As we know, almost all types of cancers are caused by DNA mutation or genome structure alterations and are followed by the appearance of other hallmarks.

In addition, the similarity among the hallmarks Activating Invasion and Metastasis (AIM), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI) and Sustaining Proliferative Signaling (SPS) is prominent. Many of the

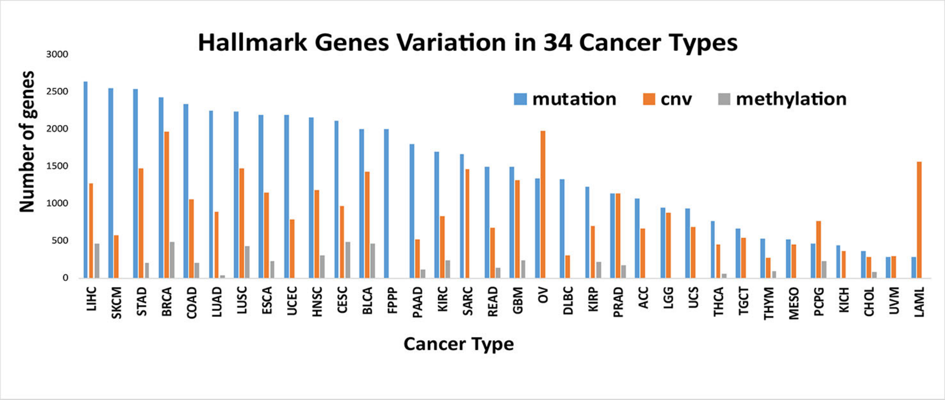


FIGURE 3 | Number of variant genes of Hallmarks in different cancer types. The number of hallmark genes with mutated, differentially methylated and copied in 34 different cancer types. It is showed that different types of cancer have different alteration characteristics.

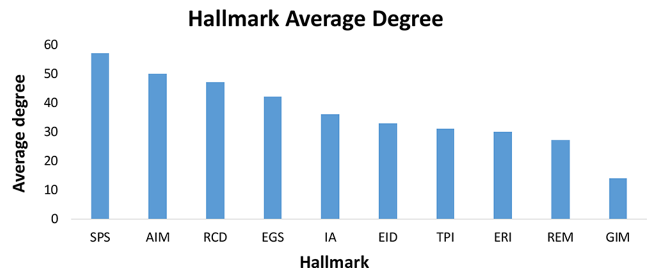


FIGURE 4 | The average degree of ten hallmarks. In addition to the GIM network, the gene interactions inside each hallmark subnetwork were more closely related than the interactions between the 10 hallmark subnetworks.

hallmarks in this set are related to the preliminary stage of cancers (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2011). One confusing inclusion in the set is AIM, which is a hallmark that is considered to be related to the end stage of cancers. However, recent research has also found that AIM occurs in early cancers as well (Hanahan and Weinberg, 2011).

The last class includes *Tumor-Promoting Inflammation (TPI), Evading Immune Destruction (EID), Resisting Cell Death (RCD),* and *Inducing Angiogenesis (IA)*. Noticeably, tumor-promoting inflammation may activate the response of immune system, and many recent studies have focused on the relationship between inflammation and the immune system in cancers (Grivennikov et al., 2010; Tan et al., 2011; Elinav et al., 2013; Hashemi et al., 2013).

In addition, we further analyzed the patterns of characteristic variation of the hallmark genes (**Figure 6**) in 34 different cancers (**Supplementary Table 3**). We looked at the top 10 altered features (e.g., mutation, CNV or methylation) of each hallmark gene as the Typical Characteristics of the Hallmark Gene (TCHG, **Supplementary Table 4**). In heat map analysis, we can clearly find major differences between the TCHGs as altered patterns in different types of cancer. In fact, these features can be used as simple markers for distinguishing cancer types.

Validation of CHG Data

Although the hallmark-related genes identified in the database came from the confirmed literature and databases, we manually

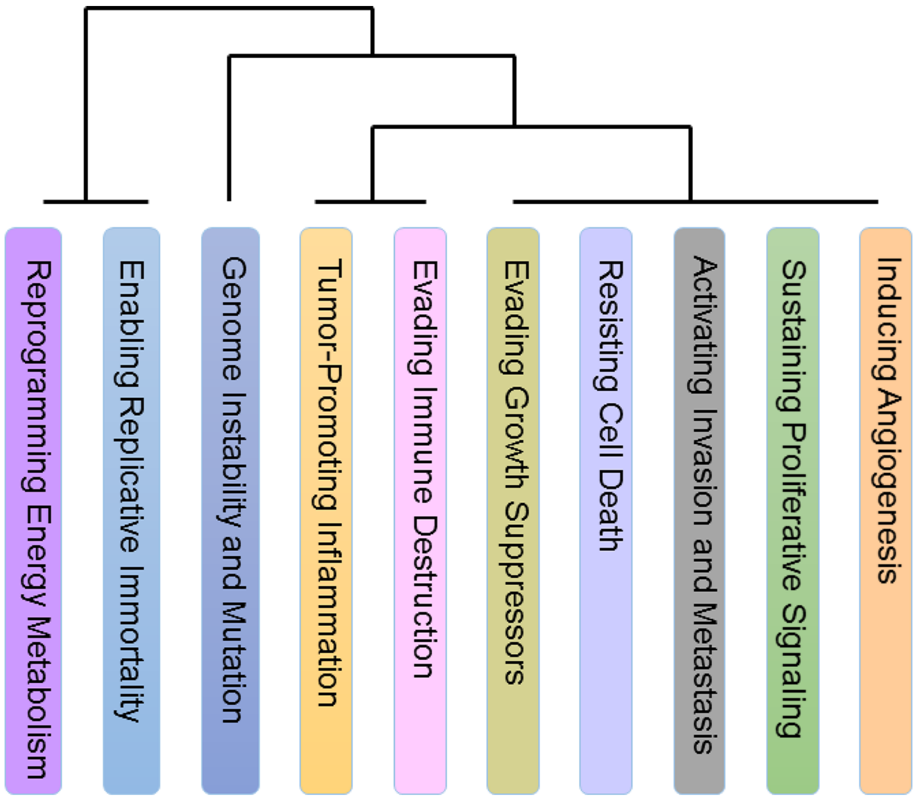

Reprogramming Energy Metabolism
Enabling Replicative Immortality
Genome Instability and Mutation
Tumor-Promoting Inflammation
Evading Immune Destruction
Evading Growth Suppressors
Resisting Cell Death
Activating Invasion and Metastasis
Sustaining Proliferative Signaling
Inducing Angiogenesis

FIGURE 5 | Relationship among ten hallmarks. The relationship among the hallmarks on a pan-cancer scale. There are two classes with only one hallmark, Reprogramming Energy Metabolism (REM) and Genome Instability and Mutation (GIM) and both of these hallmarks are clearly different from the other hallmarks in terms of their mechanisms. In addition, the similarity among the hallmarks Activating Invasion and Metastasis (AIM), Evading Growth Suppressors (EGS), Enabling Replicative Immortality (ERI), and Sustaining Proliferative Signaling (SPS) is prominent. Many of the hallmarks in this set are related to the preliminary stage of cancers. The last class includes Tumor-Promoting Inflammation (TPI), Evading Immune Destruction (EID), Resisting Cell Death (RCD), and Inducing Angiogenesis (IA). Noticeably, tumor-promoting inflammation may activate the response of immune system, and many recent studies have focused on the relationship between inflammation and the immune system in cancers.

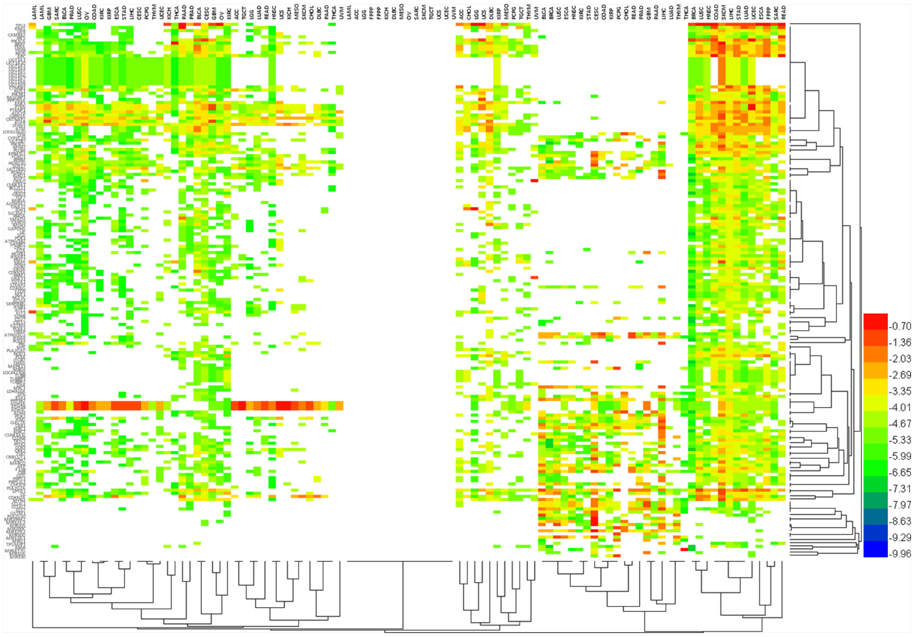


FIGURE 6 | The pattern of characteristic variation of Hallmark genes in 34 different cancers. Heat map shows major differences between the altered features (e.g., mutation, CNV or methylation) of each hallmark gene as altered patterns in different types of cancer. In fact, these features can be used as simple markers for distinguishing cancer types.

confirmed the TCHG to further ensure the accuracy of the data. Considering the very large dataset that we had to confirm, we have currently verified only the top 10 altered (mutation, methylation, CNV) genes of each hallmark. Over 92% of the typical characteristic genes have explanations of their specific hallmark functions in the literature, which demonstrates the accuracy and precision of the CHG data on a theoretical level (**Supplementary Table 4**).

In addition, we compared the results of this study with existing Sanger Cancer Gene Census databases (Futreal et al., 2004). The Sanger Cancer Gene Census database not only describes the genomic features of cancer-related genes themselves, but also includes information on tissue distribution, mutation information and protein structure. We also compared 699 cancer-related genes identified in the Sanger Cancer Gene Census database with the Typical Characteristics of the Hallmark Gene (TCHG) we identified. Of the 139 Hallmark-related TCHG genes we identified, 69 were also included in the Sanger database, accounting for 49.7%. These results also confirm the accuracy of our results. For other genes that are not included in the Sanger database, we also confirm their important role in cancer-related biological processes through literature verification, such as ETS1 (Watabe et al., 1998; Fujimoto et al., 2004; Zhang et al., 2014; Li et al., 2015) and RHOA (Lee et al., 2015; Zeng et al., 2015; Sun et al., 2016) in hallmark "Activating Invasion and Metastasis".

CHG Case Study

In addition, we used breast cancer data that was labeled as recurrent or not recurrent as samples for further analysis based on the CHG data. These analyses can be used as an example of the applications of the CHG database and can also prove the value of this database at a practical level. We performed a significant enrichment analysis of the differentially expressed genes based on data from 159 breast cancer patients from GEO with a significance level of $p < 0.01$. The sample group and the control group were patient data with and without recurrence, respectively. In particular, these differentially expressed genes were filtered by hallmark genes from the CHG database before performing the enrichment analysis. We found that these genes were enriched in 2 out of the 10 hallmarks, corresponding to the hallmarks whose main functions include *Genome Instability and Mutation (GIM)* and *Tumor-Promoting Inflammation (TPI)* (**Table 3**). It is well known that tumor development is jointly promoted by cell-intrinsic and cell-extrinsic factors. The hallmarks in **Table 3** include risk factors for tumor recurrence that are both extracellular (*Tumor-Promoting Inflammation*) and intracellular (*Genome Instability and Mutation*). These results not only expressed the theoretical interpretation of the enrichment analysis but also reflected the significance of the hallmark genes in the CHG database.

TABLE 3 | Hallmark function of differentially expressed genes based on 137 breast cancer data.

Hallmark	P-value
Genome Instability and Mutation	0.000121
Tumor-Promoting Inflammation	0.004591

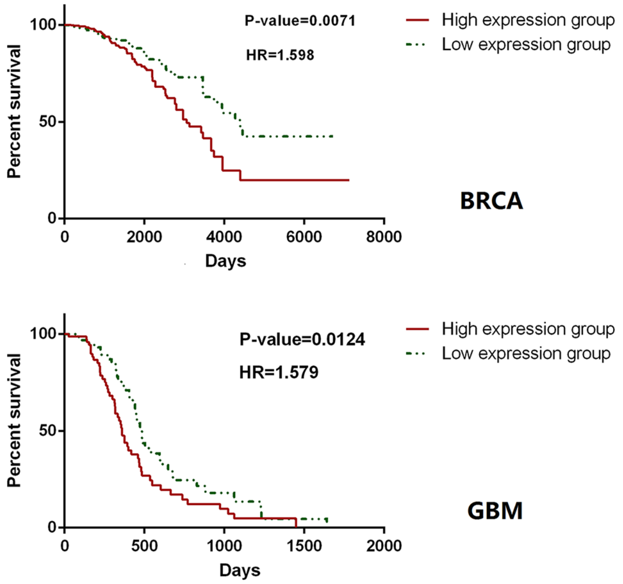


FIGURE 7 | Hallmark genes could clearly distinguish the length of the survival time in the prognosis. In a survival analysis of 1,183 breast cancer patients (up) and 156 glioblastoma multiforme patients (down), only the expression level of hallmark genes could clearly distinguish the length of the survival time in the prognosis.

The accuracy and specificity of the hallmark genes identified in CHG can also be confirmed by our analysis of the survival data for cancer patients. The survival analysis based on TCGA data was carried out with only hallmark genes as a single block, and it showed that patient groups with differentially expressed (compared to the average expression level) hallmark markers could clearly distinguish the prognosis of patients with high statistical significance. Similar results have been found in many types of cancer. For instance, in a survival analysis of 1183 breast cancer patients and 156 glioblastoma multiforme patients, only the expression level of hallmark genes could clearly distinguish the length of the survival time in the prognosis (**Figure 7**). In addition, the hallmark gene identified by CHG can also be used as a marker to determine the recurrence of cancer to some extent. An analysis of the survival data of 284 KIRP (kidney papillary cell carcinoma) patients with 27 recurrence cases in **Figure 8** shows that the hallmark genes identified in CHG have good sensitivity for distinguishing cancer recurrence. These results fully showed that the variation characteristics of the hallmark-related genes in CHG were representative, and they could be directly applied to rapid qualitative analysis.

DISCUSSION

Since Weinberg et al. firstly established the hallmarks for cancer in 2000, many studies have focused on the analysis of cancer based on a framework constructed by these hallmarks. In addition, in 2011, the number of hallmarks increased to ten, which indicates that the features of cancer may be exceedingly complex. Perhaps unsurprisingly, in 2013, another hallmark, *Aberrant Alternative Splicing*, was proposed by Michael Ladomery (Ladomery, 2013). It has been reported that the vast majority of human genes, possibly over 94%, are alternatively spliced (Pan et al., 2008). In 2015, MF Montenegro et al. targeted the epigenetic machinery of cancer cells and noted that there was increasing evidence linking the aberrant regulation of methylation to carcinogenesis (Montenegro et al., 2015), which implied that it may be a potential hallmark for cancer. In 2015, Mamatha Bhat et al. published a review about the translation machinery in cancer. They mentioned that translation played a major role in the regulation of gene expression, and the dysregulation of this process is considered a hallmark of cancer.

The CHG database that we constructed is based on the ten hallmarks that Weinberg proposed in 2011. As a specifically designed framework constructed from a hallmark database, CHG can provide a new perspective for an analysis of the diversity and development of cancers as well as a convenient method for in-depth data mining. The CHG database focused on integrating

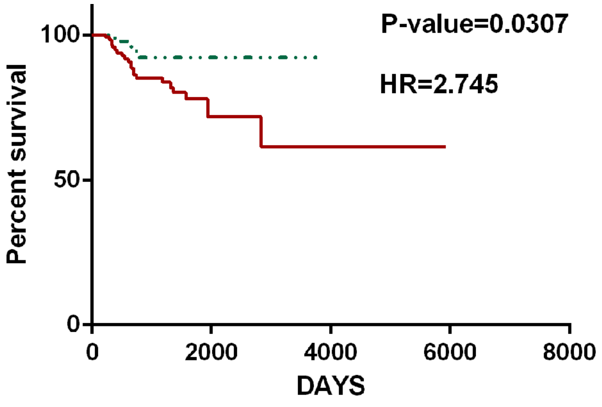


FIGURE 8 | CHG hallmark genes can be used as a marker to determine the recurrence of cancer. An analysis of the survival data of 284 KIRP (kidney papillary cell carcinoma) patients with 27 recurrence cases shows that the hallmark genes identified in CHG have good sensitivity for distinguishing cancer recurrence.

hallmark genes, annotating the potential roles of hallmark features in human cancer processes, and evaluating the relationships of the ten hallmarks by constructing hallmark networks and calculating the degree and distance between genes belonging to each network. Even though the hallmark-related genes identified in the database have been confirmed by consensus from the literature and databases, we manually confirmed the top 10 altered (mutation, methylation, CNV) genes in each hallmark to further ensure the accuracy of our data.

According to our plan, CHG database will be updated regularly every year to supplement the new findings in hallmark field or revise the existing results. We will also follow up the study of cancer hallmarks, the update of important data source (such as revision of TCGA or KEGG) and improve the practicality of CHG database in mechanism interpretation and clinical aspects. All of old version database would also be maintained and access to downloaded. The difference of each version of database would be listed.

Furthermore, over the past decade, analysis based on the integration of multiple datasets has become quite prevalent. In 2013, Du et al. (Du et al., 2013) analyzed clinically relevant long noncoding RNAs in human cancer by integrating SCNA (somatic copy number alteration), lncRNA and clinical data. In 2014, Wu et al., (2014) predicted disease-causing nonsynonymous single nucleotide variants by integrating multiple genomic datasets. Sanchez et al., (2014) integrated an analysis of Chip-Seq and RNA-Seq data to unveil an lncRNA tumor suppressor signature. Many studies, such as the work of Peng et al., have determined that miRNAs are a widely regulated regulatory mechanism in cancer (Peng et al., 2019b). Hence, it is worthwhile to integrate non-coding RNA (including miRNA, lncRNA, etc.) (Cheng et al., 2016; Cheng et al., 2019), fusion genes and drug information into a database. We have set out to construct a network that is comprised of these non-coding RNAs, genes and drugs. We hope that the next step will be to provide an online analysis tool (such as Peng et al., 2019a; Peng et al., 2019c) to provide further personalized analysis. We will gather these resources into the database in the next version, and we anticipate that the database will help promote the analysis of cancer and the identification of valuable drug targets.

DATA AVAILABILITY STATEMENT

The CHG database is freely available at our website: http://www.bio-bigdata.com/CHG/index.html.

AUTHOR CONTRIBUTIONS

DZ, DH, HX, and LW contributed equally to this work and should be considered co-first authors. JZ, LL, and HX collected data and conducts calculation and analysis. DH, QJ, and XC analyzed the results. DZ and LW wrote the paper. All authors reviewed the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China [Grant No. 61671191, 61971166, 61701142, 61802090].

ACKNOWLEDGMENTS

We are very grateful to Professor Xiujie Chen and Professor Hongbo Zhou for their suggestions and comments to this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00029/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Topological characteristics of hallmark gene networks.

SUPPLEMENTARY TABLE 1 | Characteristic pathways of different hallmarker with mapping keys.

SUPPLEMENTARY TABLE 2 | Genesets of 10 hallmarker.

SUPPLEMENTARY TABLE 3 | Specific genes in 34 cancer type of mutation.

SUPPLEMENTARY TABLE 4 | Literature validation of TOP10 altered (mutation, methylation, CNV) genes (TCHGs) of each hallmark.

REFERENCES

Bouras, E., Karakioulaki, M., Bougioukas, K. I., Aivaliotis, M., Tzimagiorgis, G., and Chourdakis, M. (2019). Gene promoter methylation and cancer: An umbrella review. *Gene* 710, 333–340. doi: 10.1016/j.gene.2019.06.023

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13, 405. doi: 10.1186/1471-2164-13-405

Chatraryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., et al. (2013). The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* 41, 816–823. doi: 10.1093/nar/gks1158

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). OAHG: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). 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

Chung, I. F., Chen, C.-Y., Su, S.-C., Li, C.-Y., Wu, K.-J., Wang, H.-W., et al. (2016). DriverDBv2: a database for human cancer driver gene research. *Nucleic Acids Res.* 44, D975–D979. doi: 10.1093/nar/gkv1314

Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., et al. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat. Struct. Mol. Biol.* 20, 908–913. doi: 10.1038/nsmb.2591

Elinav, E., Nowarski, R., Thaiss, C. A., Hu, B., Jin, C., and Flavell, R. A. (2013). Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms. *Nat. Rev. Cancer* 13, 759–771. doi: 10.1038/nrc3611

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2014). Ensembl 2014. *Nucleic Acids Res.* 42, D749–D755. doi: 10.1093/nar/gkt1196

Fujimoto, J., Aoki, I., Toyoki, H., Khatun, S., Sato, E., Sakaguchi, H., et al. (2004). Clinical implications of expression of ETS-1 related to angiogenesis in metastatic lesions of ovarian cancers. *Oncology* 66, 420–428. doi: 10.1159/000079491

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., et al. (2004). A census of human cancer genes. *Nat. Rev. Cancer* 4, 177–183. doi: 10.1038/nrc1299

Grivennikov, S. I., Greten, F. R., and Karin, M. (2010). Immunity, inflammation, and cancer. *Cell* 140, 883–899. doi: 10.1016/j.cell.2010.01.025

Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Hashemi, J., Fotouhi, O., Sulaiman, L., Kjellman, M., Hoog, A., Zedenius, J., et al. (2013). Copy number alterations in small intestinal neuroendocrine tumors determined by array comparative genomic hybridization. *BMC Cancer* 13, 505. doi: 10.1186/1471-2407-13-505

Hu, Y., Zhao, T., Zang, T., Zhang, Y., and Cheng, L. (2018). Identification of Alzheimer's Disease-Related Genes Based on Data Integration Method. *Front. Genet.* 9, 703. doi: 10.3389/fgene.2018.00703

Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., et al. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* 466, 869–873. doi: 10.1038/nature09208

Kandoth, C., Mclellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., et al. (2013). Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339. doi: 10.1038/nature12634

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, D353–D361. doi: 10.1093/nar/gkw1092

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

Laddha, S. V., Ganesan, S., Chan, C. S., and White, E. (2014). Mutational landscape of the essential autophagy gene BECN1 in human cancers. *Mol. Cancer Res.* 12, 485–490. doi: 10.1158/1541-7786.MCR-13-0614

Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* 2013, 463786. doi: 10.1155/2013/463786

Lee, H. K., Choung, H. W., Yang, Y. I., Yoon, H. J., Park, I. A., and Park, J. C. (2015). ODAM inhibits RhoA-dependent invasion in breast cancer. *Cell Biochem. Funct.* 33, 451–461. doi: 10.1002/cbf.3132

Li, J., Lenferink, A. E., 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

Li, A. X., Xin, W. Q., and Ma, C. G. (2015). Fentanyl inhibits the invasion and migration of colorectal cancer cells *via* inhibiting the negative regulation of Ets-1 on BANCR. *Biochem. Biophys. Res. Commun.* 465, 594–600. doi: 10.1016/j.bbrc.2015.08.068

Menyhart, O., Harami-Papp, H., Sukumar, S., Schafer, R., Magnani, L., De Barrios, O., et al. (2016). Guidelines for the selection of functional assays to evaluate the hallmarks of cancer. *Biochim. Biophys. Acta* 1866, 300–319. doi: 10.1016/j.bbcan.2016.10.002

Montenegro, M. F., Sanchez-Del-Campo, L., Fernandez-Perez, M. P., Saez-Ayala, M., Cabezas-Herrera, J., and Rodriguez-Lopez, J. N. (2015). Targeting the epigenetic machinery of cancer cells. *Oncogene* 34, 135–143. doi: 10.1038/onc.2013.605

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259

Peng, J., Guan, J., and Shang, X. (2019a). Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder. *Front. Genet.* 10, 226. doi: 10.3389/fgene.2019.00226

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019b). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 35 (21), 4364–4371. doi: 10.1101/276048

Peng, J., Wang, X., and Shang, X. (2019c). 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

Sanchez, Y., Segura, V., Marin-Bejar, O., Athie, A., Marchese, F. P., Gonzalez, J., et al. (2014). Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. *Nat. Commun.* 5, 5812. doi: 10.1038/ncomms6812

Sina, A. A., Carrascosa, L. G., and Trau, M. (2019). DNA Methylation-Based Point-of-Care Cancer Detection: Challenges and Possibilities. *Trends Mol. Med.* 25 (11), 955–966. doi: 10.1016/j.molmed.2019.05.014

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2013). The Cancer Genome Atlas. *Science* 320, 1958.

Sun, K., Duan, X., Cai, H., Liu, X., Yang, Y., Li, M., et al. (2016). Curcumin inhibits LPA-induced invasion by attenuating RhoA/ROCK/MMPs pathway in MCF7 breast cancer cells. *Clin. Exp. Med.* 16, 37–47. doi: 10.1007/s10238-015-0336-7

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568. doi: 10.1093/nar/gkq973

Tan, D. S., Iravani, M., Mccluggage, W. G., Lambros, M. B., Milanezi, F., Mackay, A., et al. (2011). Genomic analysis reveals the molecular heterogeneity of ovarian clear cell carcinomas. *Clin. Cancer Res.* 17, 1521–1534. doi: 10.1158/1078-0432.CCR-10-1688

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015

Wang, E., Zaman, N., Mcgee, S., Milanese, J. S., Masoudi-Nejad, A., and O'connor-Mccourt, M. (2015). Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.* 30, 4–12. doi: 10.1016/j.semcancer.2014.04.002

Watabe, T., Yoshida, K., Shindoh, M., Kaya, M., Fujikawa, K., Sato, H., et al. (1998). The Ets-1 and Ets-2 transcription factors activate the promoters for invasion-associated urokinase and collagenase genes in response to epidermal growth factor. *Int. J. Cancer* 77, 128–137. doi: 10.1002/(SICI)1097-0215(19980703)77:1<128::AID-IJC20>3.0.CO;2-9

Wu, J., Li, Y., and Jiang, R. (2014). Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. *PloS Genet.* 10, e1004237. doi: 10.1371/journal.pgen.1004237

Wu, P., Li, T., Li, R., Jia, L., Zhu, P., Liu, Y., et al. (2017). 3D genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations. *Nat. Commun.* 8, 1937. doi: 10.1038/s41467-017-01793-w

Zaman, N., Li, L., Jaramillo, M. L., Sun, Z., Tibiche, C., Banville, M., et al. (2013). Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Rep.* 5, 216–223. doi: 10.1016/j.celrep.2013.08.028

Zeng, Y., Xie, H., Qiao, Y., Wang, J., Zhu, X., He, G., et al. (2015). Formin-like2 regulates Rho/ROCK pathway to promote actin assembly and cell invasion of colorectal cancer. *Cancer Sci.* 106, 1385–1393. doi: 10.1111/cas.12768

Zhang, D., Wang, G., and Wang, Y. (2014). Transcriptional regulation prediction of antiestrogen resistance in breast cancer based on RNA polymerase II binding data. *BMC Bioinf.* 15 Suppl 2, S10. doi: 10.1186/1471-2105-15-S2-S10

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.2019.01374

A Pipeline for Reconstructing Somatic Copy Number Alternation's Subclonal Population-Based Next-Generation Sequencing Data

*Yanshuo Chu, Chenxi Nie and Yadong Wang**

Center of Bioinfomatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

State-of-the-art next-generation sequencing (NGS)-based subclonal reconstruction methods perform poorly on somatic copy number alternations (SCNAs), due to not only it needs to simultaneously estimate the subclonal population frequency and the absolute copy number for each SCNA, but also there exist complex bias and noise in the tumor and its paired normal sequencing data. Both existing NGS-based SCNA detection methods and SCNA's subclonal population frequency inferring tools use the read count on radio (RCR) of tumor to its paired normal as the key feature of tumor sequencing data; however, the sequencing error and bias have great impact on RCR, which leads to a large number of redundant SCNA segments that make the subsequent process of SCNA's subclonal population frequency inferring and subclonal reconstruction time-consuming and inaccurate. We perform a mathematical analysis of the solution number of SCNA's subclonal frequency, and we propose a computational algorithm to reduce the impact of false breakpoints based on it. We construct a new probability model that incorporates the RCR bias correction algorithm, and by stringing it with the false breakpoint filtering algorithm, we construct a whole SCNA's subclonal population reconstruction pipeline. The experimental result shows that our pipeline outperforms the existing subclonal reconstruction programs both on simulated data and TCGA data. Source code is publicly available as a Python package at https://github.com/dustincys/msphy-SCNAClonal.

Keywords: somatic copy number alternation, subclonal reconstruction, subclonal frequency, absolute copy number, bias correction

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Junwei Han,
Harbin Medical University, China
Juan Wang,
Inner Mongolia University, China

****Correspondence:***
Yadong Wang
ydwang@hit.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *16 August 2019*
Accepted: *16 December 2019*
Published: *27 February 2020*

Citation:
Chu Y, Nie C and Wang Y (2020) A Pipeline for Reconstructing Somatic Copy Number Alternation's Subclonal Population-Based Next-Generation Sequencing Data. Front. Genet. 10:1374. doi: 10.3389/fgene.2019.01374

INTRODUCTION

Tumor heterogeneity introduces challenges in cancer tissue diagnosis and subsequent treatment (Nowell, 1976). Tumor heterogeneity cannot be inferred by the properties of biomolecular through the ontology or pathway analysis (Cheng et al., 2017; Cheng et al., 2018c), but could be inferred by measuring the quantity of biomoleculars (Cheng et al., 2018b; Cheng et al., 2018d; Cheng et al., 2019). To decipher cell composition in bulk cells, somatic copy number alternations (SCNAs), most commonly found in tumor cells (Beroukhim et al., 2010), are utilized as the representative to determine tumor subclonal populations in a tumor–normal tissue paired manner (Oesper et al., 2013; Li and Xie, 2015).

The benefit of using SCNA to conduct subclonal reconstruction is that the WGS data doesn't have to be deeply sequenced (Li and Xie, 2015), because SCNA affects large, multi-kilobase-sized or megabase-sized regions of the genome, which allows the average copy number of these regions to be accurately estimated with whole genome sequencing (WGS) (Deshwar et al., 2015).

SCNA's subclonal reconstruction algorithms attempt to infer the population structure of heterozygous tumors based on the subclonal population frequency of SCNA (Deshwar et al., 2015). However, the cellular prevalence and the absolute copy number are intertwined and next-generation sequencing (NGS)-based subclonal reconstruction needs to simultaneously estimate population frequency and the absolute copy number for each SCNA. The solution space of subclonal frequency of SCNA remains poorly understood, and there might exist multiple solutions for subclonal frequency for some SCNAs (Oesper et al., 2013), which makes the infinite site assumptions (ISAs) (Kimura, 1969; Hudson, 1983; Jiao et al., 2014) invalid. ISA is the commonly accepted and powerful assumption, which posits that each mutation occurs only once in the evolutionary history of the tumor.

To infer the SCNA's subclonal population frequency based on NGS data, the location of SCNAs in the genome needs to be obtained first. The SCNA breakpoints are detected through multiple bin-merging processes, during which rcr of tumor to its paired normal is used as a key feature (Xi et al., 2010). However, the sequencing error and bias have great impact on RCR, which leads to false positive breakpoints and incorrect subclonal reconstruction (Please refer to **Figures S2** and **S3**, **Tables S2** and **S3** in the **Supplementary**). The higher sensitivity the SCNA detection tools show, the more prone to the sequencing error the tools would be. For example, BIC-seq (Xi et al., 2010) first splits whole genome into small bins, then uses the Bayesian Information Criterion as the bin merging and stopping criterion to detect SCNA breakpoints. When sensitivity parameter λ of BIC-seq is very high, the true positive rate and the false discovery rate will decrease simultaneously (Xi et al., 2010), which means the SCNA regions will be separated into small fragments by the false positive breakpoints (Xi et al., 2010). The choice of parameter λ is equivalent to setting type I error; in other words, when performing the loop of combining windows, two neighboring windows that should be combined are left separated apart. Since the reconstruction algorithm of subclone depends on the proportion of subclone populations of somatic mutation to define mutation set and its subpopulation (Deshwar et al., 2015) (Please refer to **Figure S4** for the definition of subpopulation and subclonal population), in order to more precisely estimate the subclonal population ratio of every SCNA fragment, we need to choose a smaller λ to ensure the high true positive rate of breakpoints, so as to more accurately estimate the subclonal population frequency. However, the false positive breakpoints split the SCNA regions into many small SCNA fragments, which violates ISA and results in many redundant input data and causes the subclone reconstruction process to be extremely slow and time consuming.

Existing (NGS) based subclonal reconstruction methods, such as ThetA (Oesper et al., 2013) and Mixclone (Li and Xie, 2015), use expectation maximation (EM) or maximum likelihood method (MLM) to infer the subclonal frequency and the absolute copy number of every input data. To reduce the searching space, MixClone assumes that the number of subclonal population is less than 3, and this number (1 or 2) needs to be predefined. During the maximization step of the EM process, MixClone assumes the subclonal frequencies of all the subclonal population only equal to several combinations of discrete values to further reduce the searching space. Thus, MixClone's accuracy is compromised for speed of computation. On the other side, Theta (Oesper et al., 2013) does not make any compromise on searching space. Thus, Theta is extremely time consuming while search optimal subclonal frequency in (0,1) for every input data, which makes it unable to perform subclonal reconstruction for more than three subclonal populations.

With the ever increasing data of biotechnology comes the chance of developing computational toolkit (Cheng et al., 2016; Cheng et al., 2018a; Cheng et al., 2019) to find out the pathogeny of diseases; in this article, we provide a pipeline for reconstructing SCNA's subclonal population-based NGS data. We first perform a mathematical analysis of the solution number of SCNA's subclonal frequency, propose and prove the theorem of solution number of SCNA's subclonal frequency, and present a method to filter out false SCNA breakpoints based on it. Then we propose a probability model that incorporates rcr bias correction algorithm we previously developed, and we construct an SCNA's subclonal population reconstruction pipeline by stringing it with the false breakpoint filtering algorithm. We model the read depth of tumor sample as a Poisson distribution with the expected tumor read count proportional to the absolute copy number and subclonal frequency. We use the tree-structured stick breaking Dirichlet process (Prescott Adams et al., 2010) to generate the tree structure of tumor's evolutionary history, and use the Markov Chain Monte Carlo (MCMC) to obtain the result of subclonal reconstruction. The experimental result shows that our pipeline outperforms the existing subclonal reconstruction programs both on simulated data and TCGA data.

MATERIALS AND METHODS

Solution Space of SCNA's Subclonal Population Frequency

The RCR and the b-allele frequency (BAF) of the heterozygous single nucleotide polymorphism (SNP) locus in the SCNA segment are commonly used as input for the sequencing data-based SCNA's copy number and subclonal frequency inferring tools (Wang et al., 2007; Oesper et al., 2013; Li and Xie, 2015). Since the number of reads mapped in certain genome region is proportional to the copy number of this region, the RCR is set to be proportional to $\frac{\bar{C}_j}{2}$ by existing tools (Oesper et al., 2013; Li and Xie, 2015), where $\frac{\bar{C}_j}{2}$ denotes its average copy number of the jth SCNA segment. Let ϕ_j denote the subclonal population cellular prevalence of the jth SCNA segment; C_j^{T} denote its absolute copy number; μ_{jk}^{T} represent the BAF of the kth heterozygous SNP locus in the *j*th SCNA segment; $\bar{\mu}_j$ represent the average BAF of the *k*th heterozygous SNP locus in the *j*th SCNA segment. Then we have the following equation set

$$\begin{cases} \bar{C}_j = \phi_j {\star} C_j^T + (1-\phi_j){\star}2, \\ \bar{C}_j = \frac{1}{\bar{\mu}_{jk}}\left[\phi_j {\star} C_j^T {\star} \mu_{jk}^T + (1-\phi_j){\star}2{\star}\frac{1}{2}\right], \quad k = 1,\ldots,K_j\,. \end{cases} \tag{1}$$

where K_j is the total number of heterozygous SNP loci in the *j*th SCNA segment. Since the B allele locates either in paternal or maternal haploid, both μ_{jk}^{T} and $(1 \quad - \mu_{jk}^{\mathrm{T}})$ could possibly be the BAF value in the same SCNA fragment and both $\bar{\mu}_{jk}$ and $(1 - \bar{\mu}_{jk})$ could possibly be the average BAF value in the same SCNA fragment. To reduce the complexity, we use $\hat{\mu}_{jk}^{\mathrm{T}}$ to denote the smaller one of μ_{jk}^{T} and $(1 \quad - \mu_{jk}^{\mathrm{T}})$; $\widehat{\bar{\mu}}_{jk}$ to denote the smaller one of μ_{jk}^{T} and $(1 - \bar{\mu}_{jk})$. Here we give a theorem to help answer the solution space of equation set 1 and we prove it in the **Supporting Information**.

THEOREM 1. *Given $\bar{C}_j$ and $\{\widehat{\bar{\mu}}_{jk}\}_{k=1}^{K_j}$ and let $\xi = \frac{C_j^{\mathrm{T}}\hat{\mu}_{jk}^{\mathrm{T}}-1}{C_j^{\mathrm{T}}-2}$, we have the following conclusions:*

1. *If $\bar{C}_j < 2$, there is only one solution ϕ_j in Equation set 1.*
2. *If $\bar{C}_j > 2$ and $\bar{C}_j < \frac{1}{\widehat{\bar{\mu}}_{jk}}$ there is only one solution of ϕ_j in Equation set 1.*
3. *If $\bar{C}_j > 2$ and $\bar{C}_j \geq \frac{1}{\widehat{\bar{\mu}}_{jk}}$, there are infinite solutions of ϕ_j in Equation set 1.*
4. *If $\bar{C}_j > 2$ and $\bar{C}_j \geq \frac{1}{\widehat{\bar{\mu}}_{jk}}$, there are multiple solutions of ϕ_j in Equation set 1 on the curves of the family of function $\widehat{\bar{\mu}}_{jk} = \xi(1 - \frac{2}{\bar{C}_j}) + \frac{1}{\bar{C}_j}$, under the restriction of maximum absolute copy number C_{max}. Suppose segment $s_{j\,'}$ and $s_{j''}$ are the two solutions for given $\bar{C}_j$ and $\{\widehat{\bar{\mu}}_{jk}\}_{k=1}^{K_j}$, then $\frac{C_{j'}^{\mathrm{T}}\hat{\mu}_{j'k}^{\mathrm{T}}-1}{C_{j'}^{\mathrm{T}}-2} = \frac{C_{j''}^{\mathrm{T}}\hat{\mu}_{j''k}^{\mathrm{T}}-1}{C_{j''}^{\mathrm{T}}-2}$. The multiple solution area would be $\bar{C}_j \in (2, \min(C_{j'}, C_{j''})$ and $\widehat{\bar{\mu}}_{jk} \in (\min(\hat{\mu}_{j'k}^{\mathrm{T}}, \hat{\mu}_{j''k}^{\mathrm{T}}), 2)$.*

As shown in **Figure 1**, given the observation value $\bar{C}_j$ and $\widehat{\bar{\mu}}_{jk}$ and maximum copy number $C_{\max} = 15$, only 7/43 of the curves of the family of function $\widehat{\bar{\mu}}_{jk} = \xi(1 - \frac{2}{\bar{C}_j}) + \frac{1}{\bar{C}_j}$ present multiple ϕ_j solutions (Please refer to **Table S1** for the detail information of multi-solution range).

The Algorithm of Filtering Out False Positive SCNA Breakpoints

We assume that there are no two adjacent SCNAs that present the same $\bar{C}_j$ and $\widehat{\bar{\mu}}_{jk}$ and meanwhile the different ϕ_j and C_j^{T} according to Theorem 1. We use the same method described in Li and Xie (2015) to model the read count ratio of tumor and its paired normal. Based on the Lander–Waterman model (Lander and Waterman, 1988), the probability of sampling a read from a given segment depends on three main factors: 1) its copy number, 2) its total genomic length, and 3) its mappability, which depends on factors such as repetitive sequence and GC content (Li and Xie, 2015). For each segment *j*, we associate a coefficient *j*) to account for the effect of its mappability and genomic length. Thus, the expected tumor read counts mapped to segment *j*, which is denoted as λ_j, are proportional to $\bar{C}_j\theta_j$. For example, for segment *x* and segment *y*, we have

$$\frac{\lambda_x}{\lambda_y} = \frac{\bar{C}_x\theta_x}{\bar{C}_y\theta_y} \tag{2}$$

Because the mappability coefficients matter only in a relative sense, we take $\theta_x/\theta_y = D_x^{\mathrm{N}}/D_y^{\mathrm{N}}$, as these segments should have the same sequence properties between the normal and tumor samples. Thus, Equation 2 is transformed into

$$\log(\lambda_x/D_x^{\mathrm{N}}) - \log(\lambda_y/D_y^{\mathrm{N}}) = \frac{\bar{C}_x}{\bar{C}_y}\,. \tag{3}$$

However, our previous study (Chu et al., 2017a) has shown the RCR of tumor to its paired normal presents a log-linear GC content bias, and has described a bias correction software "Pre-SCNAClonal" (Chu et al., 2017a) to correct this bias specifically. Let $\widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}$ denote the corrected read count ratio of tumor sample and its paired normal, and let $\Phi()$ denote the bias correction process. Then we have $\widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}} = \Phi(D_i^{\mathrm{S}}/D_i^{\mathrm{N}})$ and

$$\log\left(\widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}\right) - \log\left(\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}}\right) = \log\frac{\bar{C}_i}{\bar{C}_j}\,. \tag{4}$$

Then we use the following steps to filter out false positive SCNA breakpoints.

1. First, BIC-Seq with a small λ is used to detect SCNA breakpoints. Then the whole genome is separated into SCNA fragments by these breakpoints. We use $\{s_j\}_{j=1}^{J}$ to denote this SCNA fragment set.
2. Next, Pre-SCNAClonal (Chu et al., 2017a) is used to correct the bias of RCR.
3. Next, the hierarchical clustering algorithm is used to cluster $\{s_j\}_{j=1}^{J}$ based on $\log(\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}})$ of every segment with the maximum amount of cluster predefined as $C_{max} * \tau$, where τ is the number of subclonal populations. Suppose in this step, there are *N* clusters obtained by the hierarchical clustering algorithm. We denote the *n*th cluster as $\mathbb{S}_n$ where $n = 1, 2,\ldots, N$. For convenience, we call this step the aggregation step.
4. Next, the MeanShift algorithm is used to perform an unsupervised cluster search on $\cup_{s_j\in\mathbb{S}_n}\{\widehat{\bar{\mu}}_{jk}\}_{k=1}^{K_j}$, where $\mathbb{S}_n$ is obtained by step 3. Assume there are M_n BAF clusters detected in $\cup_{s_j\in\mathbb{S}_n}\{\widehat{\bar{\mu}}_{jk}\}_{k=1}^{K_j}$, and we use $\Psi(\widehat{\bar{\mu}}_{jk}) \in \{1,\ldots,M_n\}$ to represent the cluster index. Then for every $s_j \in \mathbb{S}_n$ we define the BAF cluster of s_j to be the BAF cluster of $\{\widehat{\bar{\mu}}_{jk}\}_{k=1}^{K_j}$ with the largest number. Then each $\mathbb{S}_n$ is split into subclusters $\{\mathbb{S}_{n,m}\}_{m=1}^{M_n}$ based on the BAF cluster of each s_j. For convenience, we call this step the decomposition step.
5. For each $\mathbb{S}_{n,m}$, $n = 1,2,\ldots,N$, $m = 1,2,..,M_n$, we merge two adjacent SCNA fragments, which are on the same chromosome and the distance between them is less than a predefined threshold ρ.

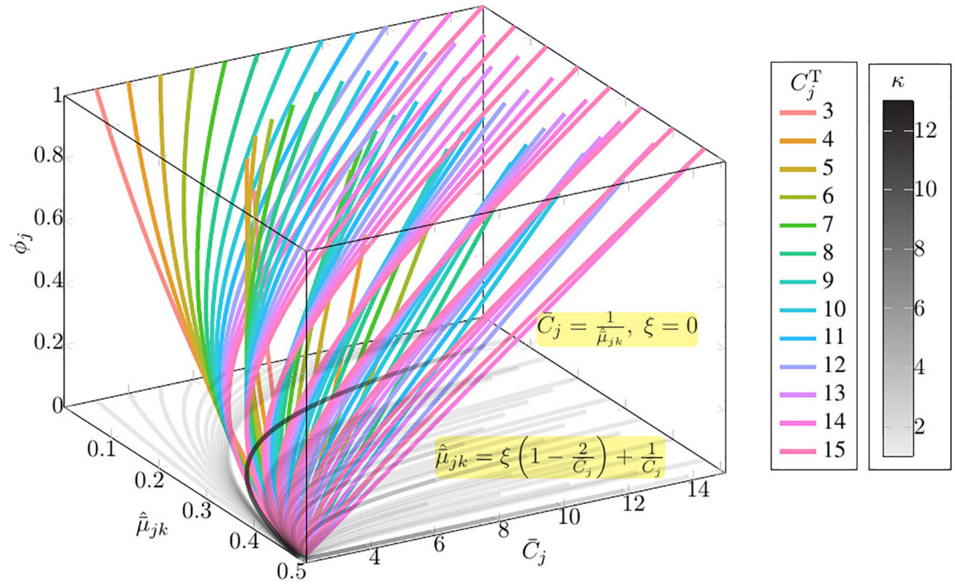


FIGURE 1 | The solution space of Equation set 1 given the observation value $\bar{C}_j$ and $\hat{\bar{\mu}}_{jk}$ and maximum copy number $C_{\max} = 15$. In this figure, κ denotes the number of solutions; $\xi = \frac{C_j^T \hat{\mu}_{jk}^T - 1}{C_j^T - 2}$, where C_j^T is the absolute copy number of SCNA in the *j*th segment s_j, $\hat{\mu}_{jk}^T$ is the normalized BAF of tumor reads mapped at the *k*th heterozygous SNP loci in the *j*th segments s_j; $\hat{\bar{\mu}}_{jk}$ denotes the normalized average tumor reads mapped at the *k*th heterozygous SNP loci in the *j*th segments s_j; $\bar{C}_j$ denotes the average copy number of segment s_j; ϕ_j denotes the subclonal frequency of segment s_j.

The space complexity of the algorithm of filtering out false positive SCNA breakpoints is $o(J^2)$. The computational complexity of "MeanShift" and "hierarchical clustering" are $o(\sum_{n=1}^{N}(I_n \star \sum_{s_j \in \mathbb{S}_n} K_j)^2)$ and $o(J^3)$, where I_n is the number of iterations for $\mathbb{S}_n$. Thus. the time complexity of the algorithm of filtering out false positive SCNA breakpoints is $o(J^3 + \sum_{n=1}^{N}(I_n \star \sum_{s_j \in \mathbb{S}_n} K_j)^2)$. The detail validation of this algorithm are described in Section 4 in the **Supplementary** (Please refer to **Figures S5–S8** for the results).

Normal Segments Detection Method

The task of normal segments detection is to find out all the segments that $\bar{C}_j = 2$, since the copy number C_j^{N} in s_j in normal sample equals 2, normally. A cancer genome differs from the reference genome by gains and losses of segments, or intervals, of the reference genome (Oesper et al., 2013).

However, due to two different sequencing processes and the coverage may not exactly be the same for tumor and its paired normal, $\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}}$ does not always equal to 1 for the normal segments (Li and Xie, 2015). In this paper, we use the same normal segments detection method described in our previous work (Chu et al., 2017a), which utilizes BAF information to detect normal segments.

Equation set 1 implies following conclusion

$$\begin{aligned} &\phi_j = 0 \quad \text{or} \quad C_j^{\mathrm{T}} = 2 \Leftrightarrow \bar{C}_j = 2, \\ &\phi_j = 0 \quad \text{or} \quad C_j^{\mathrm{T}} = 0 \quad \text{or} \quad \mu_{jk}^{\mathrm{T}} = \tfrac{1}{2} \Leftrightarrow \bar{\mu}_{jk}^{\mathrm{T}} = \tfrac{1}{2}\,. \end{aligned} \tag{5}$$

We detect the normal segments $\mathbb{N}_{t_m}$ from $\mathbb{S}_{t_m}$ according to Equation 5 by the following two steps. First, we filter out all the segments $s_j \in \mathbb{S}_{t_m}$ with $\bar{\mu}_{jk}^{\mathrm{T}} \neq \frac{1}{2}$ for $k = 1, \ldots, K_{s_j}$. In the remaining segments, the possible C_j^{T} could be any one in {0, 2, 4,…}, since all the possible genotypes G_{jk}^{T} of allele at the *k*th site for $\mu_{jk}^{\mathrm{T}} = \frac{1}{2}$ could be any one in {∅, PM, PPMM,…}. Next, we obtain all the normal segments $\mathbb{N}_{t_m}$ from these segments by selecting the segments with the read depth d_{jk}^{S} at the *k*th heterozygous SNP site equal to the coverage of the aligned WGS data of the tumor sample.

The Probability Model of Subclonal Population Frequency

Figure 2 shows the probabilistic graphical model of SCNA's subclonal population frequency. In this figure, *S* denotes the set of all the SCNA segments; ℕ denotes the set of segments that contain no SCNA. We use the same method described in Li's study (Li and Xie, 2015) to set the probability of BAF to obey binomial distribution

$$b_{jk}^{\mathrm{S}} | d_{jk}^{\mathrm{S}}, \mu_{jk}^{\mathrm{T}}, \phi_j \quad \sim \mathrm{Binomial}\left(d_{jk}^{\mathrm{S}}, \widehat{\bar{\mu}}_{jk}\right), \tag{6}$$

where b_{jk}^{S} denotes the number of tumor reads that contain B allele at the *k*th heterogeneous SNP locus and d_{jk}^{S} denotes the total number of tumor reads mapped at this locus. In this figure, G_{jk}^{T} denote the allele's genotype at the *k*th heterogeneous snp locus in segment s_j.

According to Equation 4, we have the expected tumor read counts mapped to segment *j*

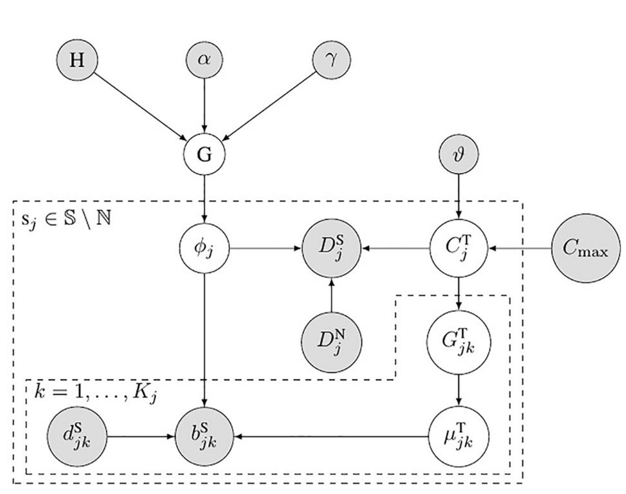


FIGURE 2 | Bayesian network model for subclonal population frequency. In this figure, G denotes the tree-structured Dirichlet process; H denotes the base distribution; α and γ are the scaling parameters of G; ϕ_j denotes the subclonal frequency of SCNA in segment s_j; D_j^S denotes the number of tumor reads mapped in segment s_j, while D_j^N denotes the number of normal reads mapped in segment s_j; C_j^T denotes the absolute copy number of SCNA in segment s_j; ϑ denotes the geometric mean of the read count ratio of all the baseline segments ℕ; $C_{\max}$ is the maximum absolute copy number pre-defined; G_{jk}^T denotes the tumor genotype of the *k*th heterozygous SNP loci in the *j*th segments s_j; u_{jk}^T denotes the tumor BAF of the *k*th heterozygous SNP loci in the *j*th segments s_j;b_{jk}^S and d_{jk}^S denote the number of B-allele and the total allele at the *k*th heterozygous SNP loci in the *j*th segments s_j.

$$\lambda_j = \Phi^{-1}\left(\frac{\bar{C}_j}{\bar{C}_i} \times \widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}\right) \times D_j^{\mathrm{N}} \tag{7}$$

where $\Phi^{-1}()$ denotes the reverse process of bias correction. Let $|\mathbb{N}|$ denote the number of baseline segments (Li and Xie, 2015) (in which the absolute copy number $C_j^{\mathrm{T}} = 2$). We use the average of read count's log ratio of all the baseline segments $\vartheta = \sqrt[-|\mathbb{N}|]{\prod_{s_i \in \mathbb{N}} \widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}}$ to calculate the expectation of tumor read count, and model the tumor read count as a Poisson distribution

$$D_j^{\mathrm{S}} | D_j^{\mathrm{N}}, C_j^{\mathrm{T}}, \phi_j \quad \sim Poisson\left(\Phi^{-1}\left(\frac{\bar{C}_j}{2} \times \vartheta\right) \times D_j^{\mathrm{N}}\right) \tag{8}$$

It could be deduced from the first equation in Equation set 1 that $\bar{C}_j > 2 \Leftrightarrow C_j^{\mathrm{T}} > 2$. Therefore, we may conclude that $\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}} > \vartheta \Leftrightarrow C_j^{\mathrm{T}} > 2$, since $\bar{C}_i$ must equal 2 if s_i contains no SCNA. We set C_j^{T} obeys the categorical distribution

$$C_j^{\mathrm{T}} \sim \text{Categorical}(\varsigma(\vartheta)), \tag{9}$$

where function ς (ϑ) denotes C_j^{T}'s range; ς (ϑ) = {0, 1, 2} if $\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}} < \vartheta$; ς (ϑ) = {2, 3,…, $C_{\max}$} if $\widehat{D_j^{\mathrm{S}}/D_j^{\mathrm{N}}} > \vartheta$.

The subclonal population frequency of certain mutation equals the sum of all its subpopulation frequencies (for details, refer to **Figure S1** in the **Supplementary**), and all the subpopulation frequencies in the tumor sample sums to 1. Therefore, all the subpopulation frequencies in the tumor sample obey the Dirichlet distribution, and this Dirichlet distribution obeys the tree-structured Dirichlet process (DP) (Prescott Adams et al., 2010). Suppose there are *P* subpopulations in a tumor sample; let $x_1, \ldots, x_p$ denote all the subpopulation frequencies

$$x_1, \ldots, x_P \quad \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_P), \tag{10}$$

where $\alpha_1, \ldots, \alpha_p$ are the concentration parameters. In this paper, we set $\alpha_1 = \ldots = \alpha_p = 1$, then Equation 10 is transformed into a uniform distribution of $(p - 1)$-dimension simplex. Therefore, the prior probability of subclonal frequency ϕ_j equals the probability of the tree structure. In **Figure 2**, *G* denotes the tree-structured DP; H denotes the base distribution; α and γ are the scaling parameters of *G*.

We use MCMC to obtain the prior distribution of ϕ_j since the probability of tree-structured DP cannot be explicitly expressed. We use the slice sampling method described in Prescott's study (Prescott Adams et al., 2010) to generate tree structure. The complete posterior probability of the subclonal population frequencies of all the SCNA segments

$$\begin{aligned}
&\Pr\left(\{\phi_j\}_{s_j \in \mathbb{S}\setminus\mathbb{N}} | \{D_J^S\}_{s_j \in \mathbb{S}\setminus\mathbb{N}}, \left\{\{b_{jk}^S\}_{k=1}^{K_j}\right\}_{sj \in \mathbb{S}\setminus\mathbb{N}}, \mathfrak{T}\right) \\
&\propto \Pr\left(\{D_J^S\}_{S_j \in \mathbb{S}\setminus\mathbb{N}}, \left\{\{b_{jk}^S\}_{k=1}^{K_j}\right\}_{S_j \in \mathbb{S}\setminus\mathbb{N}} | \{\phi_j\}_{S_j \in \mathbb{S}\,\}\mathbb{N}}\right) \\
&\times \Pr\left(\{\phi_j\}_{S_j \in \mathbb{S}\setminus\mathbb{N}}\right) \\
&= \prod_{N \in \mathfrak{T}} \sum_{C_j^T \in \{0,1\ldots \; C_{\max}\}} \sum_{G_{jk}^T \in \zeta\left(C_j^T\right)} \sum_{\mu_{jk}^T \in \eta\left(G_{jk}^T\right)} \prod_{Sj \in N} \\
&\left[\frac{1}{D_j^S\,!} \times \left(\Phi^{-1}\left(\frac{\bar{C}_j}{2} \times \sqrt[|\mathbb{N}|]{\prod_{s_i \in \mathbb{N}} \widehat{D_i^S/D_i^N}}\right) \times D_j^N\right)^{D_j^S} \times \right. \\
&\qquad e^{-\phi^{-1}}\left(\frac{\bar{C}_j}{2} \times \sqrt[|\mathbb{N}|]{\prod_{s_i \in \mathbb{N}} \widehat{D_i^S/D_i^N}}\right) \times D_j^N \\
&\left. \times \prod_{k=1}^{K_j} \begin{pmatrix} d_{jk}^S \\ b_{jk}^S \end{pmatrix} \hat{\bar{\mu}}_{jk}^{b_{jk}^S} \left(1 - \hat{\bar{\mu}}_{jk}\right)^{\left(d_{jk}^S - b_{jk}^S\right)}\right].
\end{aligned} \tag{11}$$

where $\mathfrak{T}$ denotes the tree structure, and N denotes a node in $\mathfrak{T}$. We select the tree structure with maximum posterior probability

$$\mathfrak{T}_{\max} = \underset{\mathfrak{T}^{(i)}}{\arg\max \Pr} \left(\{D_j^{\mathrm{S}}\}_{S_j \in \mathbb{S}\setminus\mathbb{N}}, \left\{\{b_{jk}^{\mathrm{S}}\}_{k=1}^{K_j}\right\}_{S_j \in \mathbb{S}\setminus\mathbb{N}} \middle| \{\phi_j\}_{S_j \in \mathbb{S}\setminus\mathbb{N}}^{(i)}, \mathfrak{T}^{(i)}\right), \tag{12}$$

where $\mathfrak{T}^{(i)}$ and $\{\phi_j\}_{s_j \in \mathbb{S}\setminus\mathbb{N}}^{(i)}$ denote tree structure and subclonal population frequencies of the *i*th sampling process. The absolute copy number of the *i*th sampling process is

$$\left\{C_j^T\right\}_{S_j\in\mathbb{S}\setminus\mathbb{N}}^{(i)}=\bigcup_{N\in\mathfrak{T}^{(i)}}\underset{\left\{C_j^T\right\}_{S_j\in N}}{\arg\max}\prod_{S_j\in\mathrm{N}}\left[\frac{1}{D_J^S!}\left(\Phi^{-1}\left(\frac{\bar{C}_j}{2}\sqrt[|\mathbb{N}|]{\prod_{s_i\in\mathbb{N}}\widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}}\right)\times D_j^{\mathrm{N}}\right)^{D_j^{\mathrm{S}}}\times\right.$$
$$\left.e^{-\Phi^{-1}}\left(\frac{\bar{C}_j}{2}\times\sqrt[|\mathbb{N}|]{\prod_{s_i\in\mathbb{N}}\widehat{D_i^{\mathrm{S}}/D_i^{\mathrm{N}}}}\right)\times D_j^{\mathrm{N}}\times\prod_{k=1}^{K_j}\binom{d_{jk}^{\mathrm{S}}}{b_{jk}^{\mathrm{S}}}\widehat{\bar{\mu}}_{jk}^{b_{jk}^{\mathrm{S}}}\left(1-\widehat{\bar{\mu}}_{jk}\right)^{\left(d_{jk}^{\mathrm{S}}-b_{jk}^{\mathrm{S}}\right)}\right], \tag{13}$$

where $\{C_j^{\mathrm{T}}\}_{s_j\in\mathbb{S}\setminus\mathbb{N}}^{(i)}$ are absolute copy numbers with the maximum posterior probability if the i'-th sampling process is the solution of Equation 12.

The Pipeline for Reconstructing SCNA's Subclonal Population-Based NGS Data

As shown in **Figure 3**, the pipeline consists of five models. The tumor and its paired normal sequence alignment sequencing data in BAM format are used as input of the pipeline. The SCNA segments are detected by BIC-seq (Xi et al., 2010), then the bias of read count ratio is corrected by the correction model (Chu et al., 2017a) we previously proposed. We filter out the false positive breakpoints by the algorithm we proposed in this paper, then we use the probability model of subclonal population frequency proposed in this paper to infer the subclonal frequency of each SCNA segment. Finally, we use the tree structure learning algorithm (Prescott Adams et al., 2010) to reconstruct the SCNA's subclonal population.

RESULTS

In this section, we evaluate the performance of probabilistic model on both simulated and real datasets and compare its performance with existing tools. Existing tools such as Mixclone (Li and Xie, 2015) and TheatA (Oesper et al., 2013) could not calculate the subclonal frequencies of more than three subclonal populations. Therefore, we use the simulated data, which contain more than three subclonal populations and TCGA benchmark data together to evaluate our model.

Results From Simulated Data

We use Pysubsim-tree (Chu et al., 2017b) to simulate a tumor's NGS read alignment data from Chromosome 21 with the evolution history configuration shown in **Figure 4** and the acquired SCNA's configuration listed in **Table 1**. In **Figure 4**, each circle represents a subpopulation; the squares with character a, b, c, d, e, and f represent five SCNAs; the number on the right side of the circle is the frequency of the subpopulation.

We set the first 50 cycles of the MCMC sampling process as burn-in and use the result of the following 300 cycles to calculate the probability of the evolutionary relationship between subpopulations. We set α = 1.0, γ = 1.0, H to be the uniform distribution. **Figures 5A**, **B** are the dot-plots of the distribution of the output of subclonal population frequency model. **Figure 5C** shows the partial order plot (Jiao et al., 2014) of the evolutionary relationship obtained by the model proposed in this paper. The arrows in this figure denote the direct evolutionary relationship of the two subpopulations. The width of the arrow denotes the probability of this evolutionary relationship present in the 300 cycles of the MCMC process. Suppose $\{\mathfrak{T}_i\}_{i=1}^{I}$ denotes all the trees obtained in all the cycles of the MCMC process, $\overrightarrow{ab}$ denotes the evolutionary relationship from subpopulation a to b. Then the probability of this evolutionary relationship is

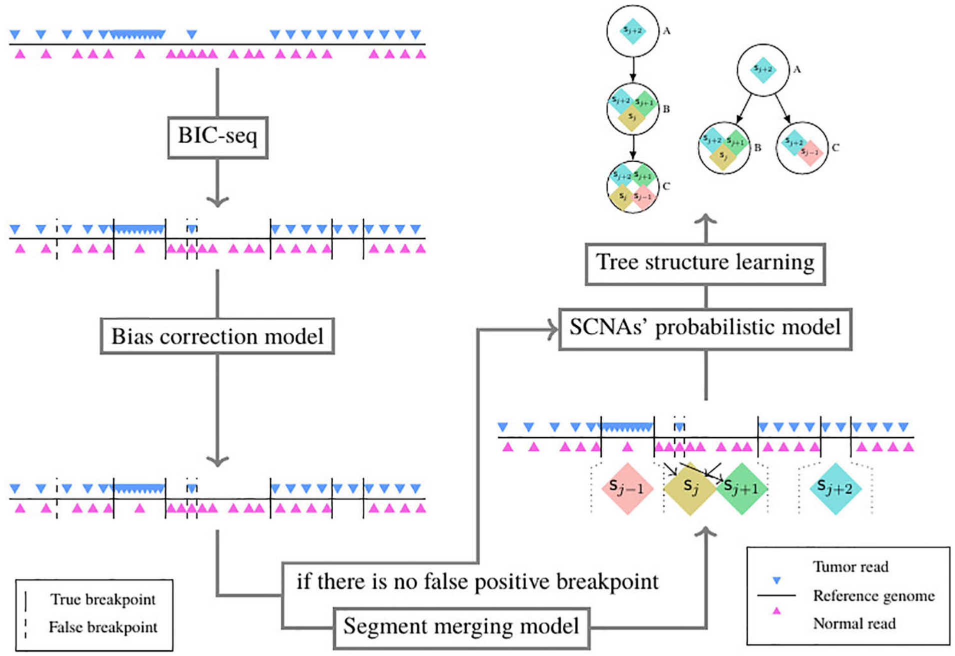


FIGURE 3 | The structure of the whole NGS data-based SCNAs' subclonal reconstruction pipeline.

$$\Pr\left(\overrightarrow{\mathrm{ab}}\right) = \frac{1}{I}\left|\left\{\mathfrak{T}_i \middle| \overrightarrow{\mathrm{ab}} \in \mathfrak{T}_i, i = 1,\ldots,I\right\}\right|. \tag{14}$$

According to Theorem 1, a and e have only one solution of ϕ_j while the others are not. The distribution of absolute copy numbers shown in **Figure 5A** is consistent with Theorem 1. The distribution of e's subclonal frequency is quite scattered in **Figure 5B** because the small subclonal frequency and the absolute copy number of e (closed to normal) cause the coverage to decrease by 5%, which is almost the same as the noise. The subclonal frequencies of other SCNAs are highly distributed at the positions of subclonal frequencies listed in **Table 1**. Each SCNA's absolute copy number and subclonal frequency with the maximum posterior probability are listed in **Table 2**. The subclonal frequencies of b and c are not correct because they have multiple solutions of subclonal frequencies according to Theorem 1, while the others are correct. The distribution of absolute copy number and subclonal frequency in **Figure 5** and the result listed in **Table 2** show that our SCNA probability model could correctly calculate the subclonal frequency of SCNA.

Results From Breast Cancer Sequencing Data

We use the ngs data "HCC1954-spiked1-n25t35s40" and "HCC1954-spiked1-n25t55s20" (denoted as "n25t35s40" and "n25t55s20" for convenience) of Cancer Genome Atlas (TCGA) Benchmark 4 dataset, which is publicly available at the National Cancer Institute GDC Data Portal (https://gdc.cancer.gov/resources-tcga-users/tcga-mutation-calling-benchmark-4-files) to further validate the subclonal frequency model proposed in this paper. HCC1954 is an immortal cell line derived from an invasive ductal carcinoma of the breast diagnosed in a 61-year-old woman (Bignell et al., 2007). "G15512.HCC1954.1" is the NGS data of this cell line, which contains one subclonal population with purity 0.99; however, this data has no ground truth of absolute copy number of the SCNA regions. "HCC1954-spiked1-n25t35s40" is generated by merging 35% of "G15512.HCC1954.1" with 25% of its paired normal NGS data and 40% of "G15512.HCC1954.1" with some SCNAs randomly spiked in it. Therefore, there are two subclonal populations in the tumor sample "HCC1954-spiked1-n25t35s40," and their subclonal frequencies are 75% and 40%, respectively. The ISA is invalid since each subclonal population contains multiple SCNAs; thus, we set the prior probability of tree structure to obey uniform distribution, and thus Equation 11 could be rewritten as follows:

$$\begin{aligned}
&\Pr\left(\phi_j\left\{D_j^S\right\}_{S_j\in\mathbb{S}\backslash\mathbb{N}}, \left\{b_{jk}^S\right\}_{k=1}^{k_j}, \mathfrak{T}\right) \propto \Pr\left(\left\{D_j^S\right\}_{S_j\in\mathbb{S}\backslash\mathbb{N}}\left\{b_{jk}^S\right\}_{k=1}^{k_j}, \mathfrak{T}|\phi_j\right)\\
&= \prod_{s_j\in\mathbb{S}\backslash\mathbb{N}} \sum_{C_j^T\in\{0,1\ldots C_{\max}\}} \left| \left[\frac{1}{D_j^S\,!} \times \left(\Phi^{-1}\left(\frac{\bar{C}_j}{2} \times \sqrt[|\mathbb{N}|]{\prod_{s_i\in\mathbb{N}} \widehat{D_i^S/D_i^N}}\right) \times D_j^N\right)^{D_j^S} \times \right.\right.\\
&\qquad e^{-\Phi^{-1}}\left(\sqrt[|\mathbb{N}|]{\prod_{s_i\in\mathbb{N}} \widehat{D_i^S/D_i^N}}\right) \times D_j^N \times\\
&\qquad \left.\prod_{k=1}^{K_j} \sum_{G_{jk}^T\in\zeta\left(C_j^T\right)} \sum_{\mu_{jk\in}^T\ \eta\left(G_{jk}^T\right)} \binom{d_{jk}^S}{b_{jk}^S} \widehat{\mu}_{jk}^{b_{jk}^S}\left(1-\widehat{\mu}\right)^{\left(d_{jk}^S-b_{jk}^S\right)}\right]
\end{aligned} \tag{15}$$

Figure 6 shows the subclonal frequencies obtained by the model proposed in this paper. In this figure, "P" denotes the parent subclonal population (subclonal frequency 75%) and "C" denotes the child subclonal population (subclonal frequency 40%). As shown in **Figure 6**, the subclonal frequencies of these two population obtained by the model proposed in this paper are 72% and 42% for sample "n25t35s40" and 77% and 25% for sample "n25t55s20," which are the most closed to the fact in comparison with MixClone and ThetA.

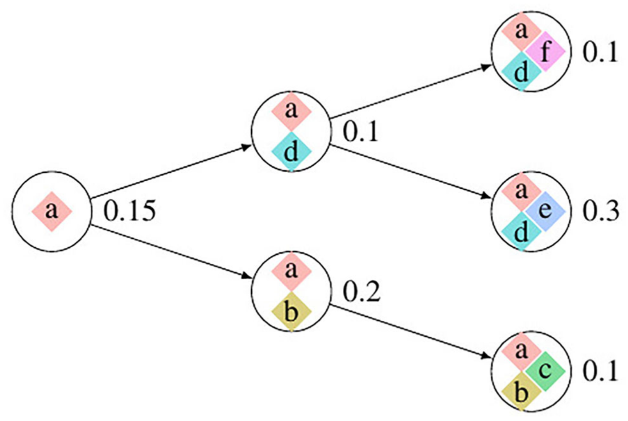


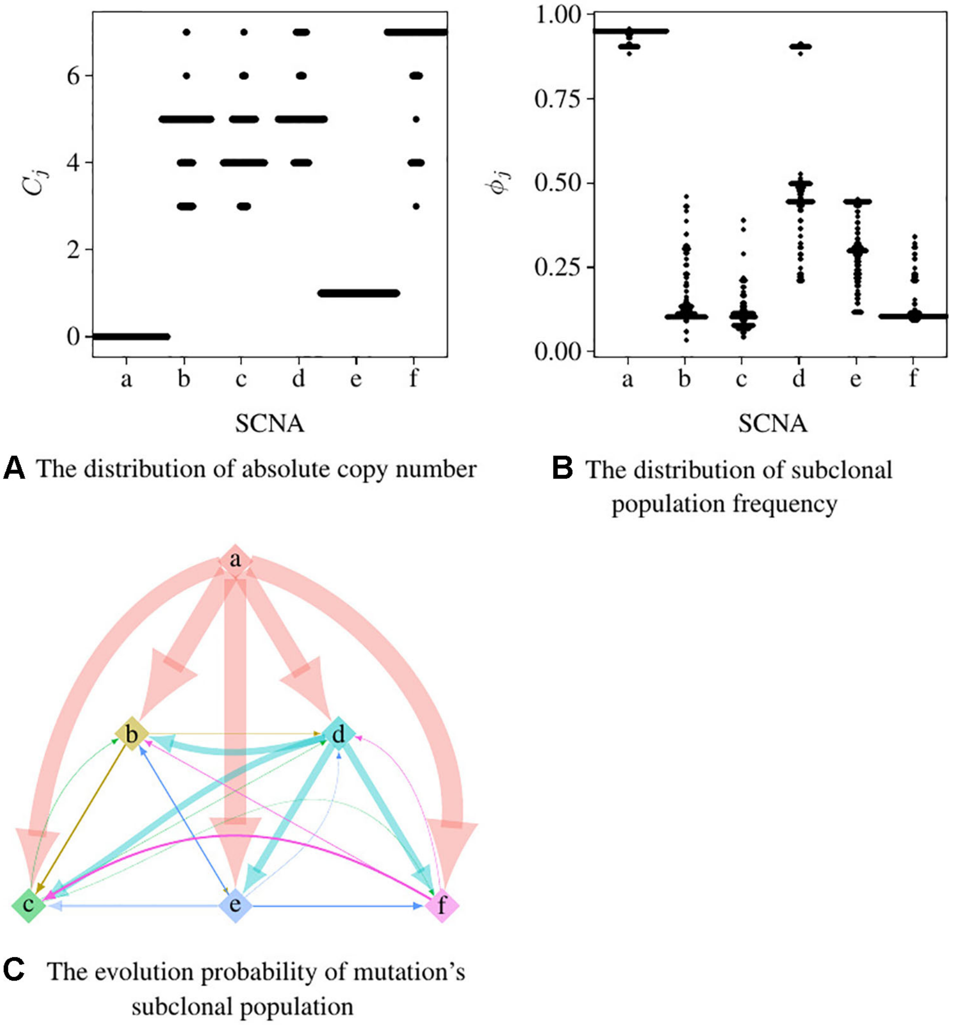
FIGURE 4 | The evolution process of subclonal population in the simulation data. In this figure, each circle denotes a subpopulation; the number on the left is its frequency; each square inside the circle denotes an SCNA; each arrow points an offspring subpopulation.

DISCUSSION

Generally, SCNAs with larger subclonal population frequency could relatively be more precisely located. However, due to the

TABLE 1 | The SCNA's configuration for each subpopulation of the simulation data.

SCNA	Chrom	Position	Length	C_j^T	G_j	ϕ_j
a	chr21	17478172	500000	0	Ø	0.95
b	chr21	27485802	500000	3	PPM	0.03
c	chr21	30959067	500000	4	PPPM	0.01
d	chr21	35841868	500000	5	PMMMM	0.05
e	chr21	43277023	500000	1	M	0.03
f	chr21	25056314	500000	7	MPPPPPP	0.01



A The distribution of absolute copy number

B The distribution of subclonal population frequency

C The evolution probability of mutation's subclonal population

FIGURE 5 | The result of subclonal reconstruction based on simulation data. **(A, B)** Dot-plots of the distribution of absolute copy number and subclonal frequency inferred by the 300 cycles of MCMC process. **(C)** The partial plot of the subclonal frequency.

TABLE 2 | The results of subclonal population frequency inferring based on simulation data.

	a	b	c	d	e	f
C_j^T result	0	7	5	5	1	7
C_j^T fact	0	3	4	5	1	7
ϕ_j result	0.950	0.106	0.075	0.501	0.304	0.106
ϕ_j fact	0.95	0.30	0.10	0.50	0.30	0.10

twice sequencing procedures of tumor and its paired normal, the read information of the genomic regions with the same copy number in tumor sample is not exactly the same as its paired normal's. Moreover, the lower read coverage of NGS makes the noise/error more likely to be mistaken for an SCNA. As shown in **Figure 7**, the number of SCNA breakpoints obtained by SCNA detection tool is proportional to the subclonal population frequency. If there exists a large proportion of false negative breakpoints, it will cause the read count in the segments incapable to reveal the copy number property, then it will affect all the read count-based SCNA analysis tools. On the other hand, if there exists a large proportion of false positive breakpoints, the segment clustering step of filtering out the false positive breakpoints could reduce the data size and make the read count information more robust to noise by merging the SCNA segments with the same absolute copy number and subclonal population frequency. As shown in Theorem 1, the SCNA segments with the same RCR and average B-allele frequency are indistinguishable to the NGS-based SCNA analysis tools. Merging two non-adjacent SCNA segments with the same NGS properties could not affect the result of the NGS-based SCNA analysis tools.

Tree-Structured Stick Breaking (TSSB) process (Prescott Adams et al., 2010) could learn the tree structure of the hierarchical data. A tree structure space could be generated

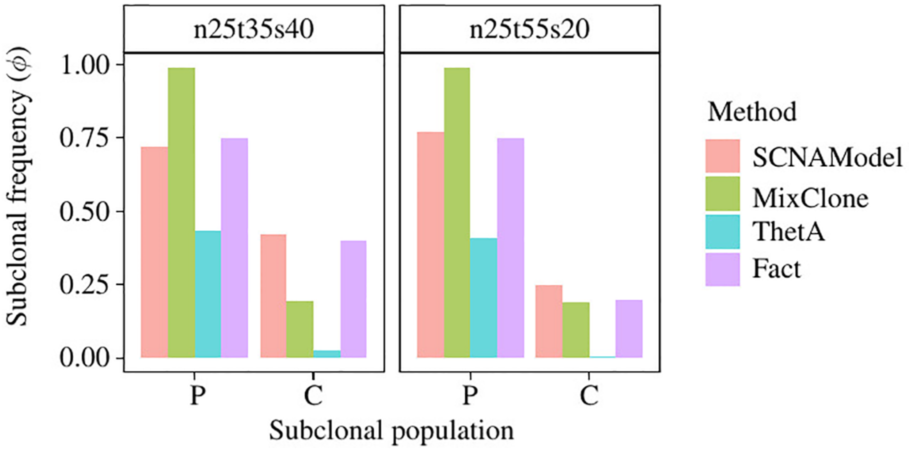


FIGURE 6 | The subclonal proportion of SCNAs in HCC1954 data. In this figure, SCNAModel is the subclonal frequency inferring model proposed in this paper.

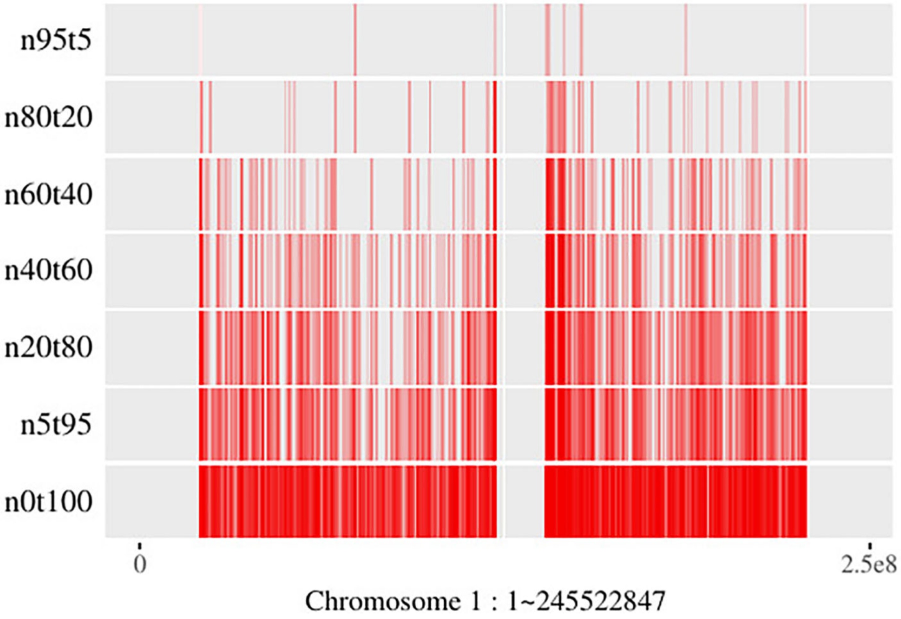


FIGURE 7 | Breakpoints distribution on chromosome 1 of mixed "HCC1954" samples. Here the "n5t95" to "n95t5" respectively denote the tumor sample from "HCC1954.mix1.n5t95" to "HCC1954.mix1.n95t5." "n0t100" denotes the tumor sample; "HCC1954" contains no normal contamination. Each of these samples contains one tumor subclone. All the breakpoints are obtained by BIC-seq (Xi et al., 2010).

by intertwining two DP; then as described in Prescott's paper (Prescott Adams et al., 2010), one can imagine throwing a dart (data) on the tree space and considering which node the dart hits. If we know subclonal number L in advance, then we could generate the tree structure in two steps. Step 1: generate a tree using all the data; Step 2: sort nodes by the sum of the size of the genome region hit, then find out the top L nodes and throw the rest of the darts (data not in the L nodes) into these L nodes randomly. **Figure 7** shows that subclonal frequency affects the number of breakpoints; thus, there might present false positive or false negative breakpoints in the result of the SCNA detection tool. The false positive breakpoints could be filtered out by the algorithm in this paper. Even if there exist false breakpoints, the redundant data that contains the same SCNA might hit the same node in the tree space generated by the TSSB process. Thus, the redundant data affects the time and space consumption, but could not affect the result of subclonal reconstruction theoretically.

CONCLUSION

In this paper, we first perform a mathematical analysis of the solution space of SCNA's subclonal frequency. Then based on the mathematical analysis, we propose an algorithm to filter out the false breakpoints and we construct a new probability model to reconstruct SCNA's subclonal population, which incorporates the algorithms of RCR bias correction we previously proposed. We use the tree-structured stick breaking DP (Prescott Adams et al., 2010) to generate the tree structure space of tumor's evolutionary history. In the probability model, the BAF of the heterozygous SNP locus in the SCNA segment is modeled as a binomial distribution and the read depth of tumor sampling data is modeled as a Poisson distribution with respect to the potential bias in RCR. We generate the distribution of subclonal frequency from the distribution of subpopulation frequency, which is drawn from the tree structure space. By stringing the model with the false breakpoint filtering algorithm, we construct a whole SCNA's subclonal population reconstruction pipeline, which is capable of inferring SCNA's absolute copy number and its subclonal population frequency and its evolutionary process while there are a lot of false positive SCNA breakpoints and the RCR presents bias. The results show that the model proposed in this paper could more accurately estimate the absolute copy number of SCNA segments and their subclonal population frequencies in comparison with existing methods both on simulated data and TCGA data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://gdc.cancer.gov/resources-tcga-users/tcga-mutation-calling-benchmark-4-files.

AUTHOR CONTRIBUTIONS

YC: Coming up with the theories and all the mathematical equations in this paper and implemented the initial version of P-SCNAClonal, the initial version of this paper. CN: Debugging of the initial version of P-SCNAClonal, experiments and result collecting, completed this paper with the result section. YW: Providing the basic idea and funding support.

FUNDING

This work was supported by funding from the National Key R&D Program of China (No: 2016YFC1202302 and 2017YFSF090117) and the National Nature Science Foundation of China (Grant No. 61822108 and 61571152).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01374/full#supplementary-material

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

Bignell, G. R., Santarius, T., Pole, J. C., Butler, A. P., Perry, J., Pleasance, E., et al. (2007). Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.* 17, 000–000. doi: 10.1101/gr.6522707

Cheng, L., Sun, J., Xu, W., Dong, L., Hu, Y., and Zhou, M. (2016). Oahg: an integrated resource for annotating human genes with multi-level ontologies. *Sci. Rep.* 6, 34820. doi: 10.1038/srep34820

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2017). Metsigdis: a manually curated resource for the metabolic signatures of diseases. *Briefings In Bioinf.* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018a). Dincrna: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncrna function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002

Cheng, L., Jiang, H., Wang, S., and Zhang, J. (2018b). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a mendelian randomisation study. *Front. In Genet.* 9, 657. doi: 10.3389/fgene.2018.00657

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., et al. (2018c). Infacront: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19, 919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2018d). 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

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019). gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* D554–560. doi: 10.1093/nar/gkz843

Chu, Y., Teng, M., Wang, Z., Wang, Y., and Wang, Y. (2017a).Pre-scnaclonal: Efficient gc bias correction for scna based tumor subclonal populations inferring, in: Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on (IEEE). pp. 262–265. doi: 10.1109/BIBM.2017.8217660

Chu, Y., Wang, L., Wang, R., Teng, M., and Wang, Y. (2017b).Pysubsim-tree: A package for simulating tumor genomes according to tumor evolution history, in: Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on (IEEE). 48 (D1), 2195–2197. doi: 10.1109/BIBM.2017.8217998

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., and Morris, Q. (2015). Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 16, 1. doi: 10.1186/s13059-015-0602-8

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul Biol.* 23, 183–201. doi: 10.1016/0040-5809(83)90013-8

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinf.* 15, 35. doi: 10.1186/1471-2105-15-35

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61, 893.

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231–239. doi: 10.1016/0888-7543(88)90007-9

Li, Y., and Xie, X. (2015). Mixclone: a mixture model for inferring tumor subclonal populations. *BMC Genomics* 16, S1. doi: 10.1186/1471-2164-16-S2-S1

Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science* 194, 23–28. doi: 10.1126/science.959840

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol.* 14, 1. doi: 10.1186/gb-2013-14-7-r80

Prescott Adams, R., Ghahramani, Z., and Jordan, M. I. (2010). Tree-structured stick breaking processes for hierarchical data. *arXiv preprint arXiv*. 1006.1062, 1–16.

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

Xi, R., Luquette, J., Hadjipanayis, A., Kim, T.-M., and Park, P. J. (2010). Bic-seq: a fast algorithm for detection of copy number alterations based on high-throughput sequencing data. *Genome Biol.* 11, 1. doi: 10.1186/1465-6906-11-S1-O10

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.

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fgene.2020.00209

A Deep Neural Network for Identifying DNA N4-Methylcytosine Sites

Feng Zeng[1]*, *Guanyun Fang*[1] and *Lan Yao*[2]*

[1] School of Computer Science and Engineering, Central South University, Changsha, China, [2] College of Mathematics and Econometrics, Hunan University, Changsha, China

Motivation: N4-methylcytosine (4mC) plays an important role in host defense and transcriptional regulation. Accurate identification of 4mc sites provides a more comprehensive understanding of its biological effects. At present, the traditional machine learning algorithms are used in the research on 4mC sites prediction, but the complexity of the algorithms is relatively high, which is not suitable for the processing of large data sets, and the accuracy of prediction needs to be improved. Therefore, it is necessary to develop a new and effective method to accurately identify 4mC sites.

Results: In this work, we found a large number of 4mC sites and non 4mC sites of *Caenorhabditis elegans* (*C. elegans*) from the latest MethSMRT website, which greatly expanded the dataset of *C. elegans*, and developed a hybrid deep neural network framework named 4mcDeep-CBI, aiming to identify 4mC sites. In order to obtain the high latitude information of the feature, we input the preliminary extracted features into the Convolutional Neural Network (CNN) and Bidirectional Long Short Term Memory network (BLSTM) to generate advanced features. Taking the advanced features as algorithm input, we have proposed an integrated algorithm to improve feature representation. Experimental results on large new dataset show that the proposed predictor is able to achieve generally better performance in identifying 4mC sites as compared to the state-of-art predictor. Notably, this is the first study of identifying 4mC sites using deep neural network. Moreover, our model runs much faster than the state-of-art predictor.

Keywords: N4-methylcytosine, machine learning, deep neural network, CNN, BLSTM, integrated algorithm

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Himel Mallick,
Merck, United States
Lin Meng,
Ritsumeikan University, Japan

Correspondence:
Feng Zeng
fengzeng@csu.edu.cn
Lan Yao
yao@hnu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *22 November 2019*
Accepted: *21 February 2020*
Published: *06 March 2020*

Citation:
Zeng F, Fang G and Yao L (2020) A Deep Neural Network for Identifying DNA N4-Methylcytosine Sites. Front. Genet. 11:209. doi: 10.3389/fgene.2020.00209

1. INTRODUCTION

DNA methylation is a form of chemical modification of DNA, which alters genetic performance without altering the DNA sequence. Numerous studies have shown that DNA methylation can cause changes in chromatin structure, DNA conformation, DNA stability, and DNA-protein interactions, thereby controlling gene expression (Wang and Qiu, 2012). In many species, the N-methylation would inhibit Watson-Crick hydrogen bond formation with guanosine (Fazakerley et al., 1987). The differential susceptibility of foreign DNA and self-DNA suggests that some process, such as cytosine methylation, may be affording protection to nuclear DNA (Carpenter et al., 2012). DNA methylation guided by specific methyltransferase enzymes occurs in both prokaryotes and eukaryotes. These modifications can label genomic regions to control various processes including base pairing, duplex stability, replication, repair, transcription, nucleosome localization, X chromosome inactivation, imprinting and epigenetic memory (Iyer et al., 2011; Allis and Jenuwein, 2016; O'Brown and Greer, 2016). The most widespread DNA methylation modifications are N6-methyladenine (6mA), 5-methylcytosine (5mC) and N4-methylcytosine (4mC) that have been detected in both prokaryotic and eukaryotic genomes (Fu et al., 2015; Blow et al., 2016; Chen et al., 2017). These modifications are catalyzed by specific DNA methyltransferases (DNMTs) that transfer a methyl group to specific exocyclic amino groups (He et al., 2018). In eukaryotes, 5mC is the most common DNA modification, which is essential for gene regulation, transposon suppression and gene imprinting (Suzuki and Bird, 2008). While 6mA and 4mC are very small, they can only be detected in eukaryotes by high sensitivity techniques. In prokaryotes, 6mA and 4mC are the majority, mainly used to distinguish host DNA from exogenous pathogenic DNA (Heyn and Esteller, 2015), and 4mc controls DNA replication and corrects DNA replication errors (Cheng et al., 1995; Wei et al., 2018). Moreover, 4mC as part of a restriction-modification (R-M) system prevents restriction enzymes from degrading host DNA (Schweizer et al., 2008; Wei et al., 2018).

Although extensive studies have been conducted on modifications of 5mC and 6ma, studies on 4mC are relatively limited due to the lack of effective experimental methods and large amounts of data. Single-molecule real-time sequencing (SMRT) technology can detect 4mC, 5mc, and 6mA base modifications (Ecker, 2010; Flusberg et al., 2010; Clark et al., 2013; Davis et al., 2013). However, SMRT sequencing is costly and is not conducive to the analysis of various species. Recently, Yu et al. (2015) proposed a method for the determination of methylcytosine in genomic DNA by 4 mC-Tet-assisted bisulfite sequencing, which can accurately generate a genome-wide, single-base resolution map of 4mC, and finally identify the 4mC motif associated with the bacterial R-M system. Biological experiments are laborious and expensive when performing genome-wide testing. Therefore, it is necessary to develop a calculation method for identifying 4mC sites.

So far, there are only four methods for identifying the 4mC sites, all of which adopt the SVM model, including iDNA4mC, 4mCPred, 4mcPred-SVM and 4mcPred-IFL. The four predictors are designed to predict 4mC sites directly from sequences. The first 4mC site predictor, called iDNA4mC (Chen et al., 2017), encodes DNA sequences using nucleotide chemistry properties and frequency and is tested across different species. The experimental results show that iDNA4mC has achieved initial results in identifying 4mC sites. However, the low predictive power is the main drawback of iDNA4mC. The second 4mC site predictor, called 4mCPred (He et al., 2018), proposes a new feature coding algorithm by combining position-specific trinucleotide propensity and electron-ion interaction pseudopotentials, which improves the accuracy of prediction. The third 4mC site predictor, called 4mcPred-SVM (Wei et al., 2018), proposes more useful sequence features in the predictor and improves the feature representation capability through a two-step feature selection method. However, the performance of the experiment did not improve much. Recently, Wei et al. (2019) proposed the fourth 4mC site predictor, called 4mcPred-IFL, which uses an iterative feature representation algorithm to learn probabilistic features from different sequential models and enhance feature representation in a supervised iterative manner. However, the complexity of 4mcPred-IFL is very high. When the data set is large, it takes a long time to obtain the results. Meanwhile, the prediction accuracy in 4mcPred-IFL can be improved further.

In this work, we developed a deep learning framework called 4mcDeep-CBI to identify the 4mC sites. Deep learning related methods are widely used in hot spots prediction of protein-protein interfaces (Pan et al., 2018; Wang et al., 2018; Deng et al., 2019; Liu et al., 2019), but we have not found any work with deep learning in 4mC sites prediction, and all previous studies have used SVM machine learning methods. This work is the first study of 4mC sites using deep learning. Especially, we have greatly expanded the dataset which is used to evaluated the prediction models of the 4mC sites. Experimental results demonstrate that 4mcDeep-CBI has better performance than other models. The contributions of our work can be summarized as follows.

(1) We have greatly expanded the dataset of *C. elegans*, and the number of samples was increased from 3,108 to 17,808, which is beneficial for subsequent research.
(2) we developed a deep learning framework to identify the 4mC sites. 3-CNN and BLSTM are used to extract deep information from the acquired features and to obtain advanced features. Experimental results show that advanced features have achieved better performance in identifying the 4mC sites.
(3) We finally take probability feature matrix obtained by the machine learning methods into the deep learning model, which further improve the prediction accuracy. In our experiment, compared with the state-of-art predictor, the proposed model has the accuracy increased from 87 to 93%.

2. MATERIALS AND METHODS

2.1. Datasets

We obtained samples genomes of *Caenorhabditis elegans* (*C. elegans*) from the latest MethSMRT website, found a lot of 4mC sites and non 4mC sites with the sequence lengths all of 41 bp. Each 4mC sequence sample has several indicators: position, coverage, IPDRatio (inter-pulse duration ratio), frac, fracLow, fracUp, identificationQv. In order to construct a reliable quality dataset, we did the following two steps. Firstly, as stated in the Methylome Analysis Technical Note, the Modification QV (modQV) score indicates that the IPD ratio is significantly different from the expected background. Since the modQV score of 30 is the default threshold for calling a position as modified, we removed the sample with the modQV score more than 30. Secondly, as elaborated in previous study (Chou et al., 2015), if training and testing are conducted through this biased dataset, the experimental results may have overestimated accuracy. To eliminate redundancy and minimize the bias, the CD-HIT software (Fu et al., 2012) with the cut off threshold set at 80% was used to remove those sequences with high sequence similarity. After the above two steps, we obtained 15, 639 samples in *C. elegans*.

We combine the new samples with the *C. elegans* benchmark dataset (Ye et al., 2017) that was used in the previous works to form a new data set with 18, 747 samples. Some of the new samples we extracted may be similar to the previous benchmark dataset. Therefore, we use the CD-HIT software to remove those samples with high sequence similarity. Finally, we get the new *C. elegans* dataset with 17, 808 samples which contains 111, 73 positive samples and 663, 5 negative samples. The positive samples are the sequences centroided with functional 4mC sites detected by the SMRT sequencing technology, while the negative samples are the sequences with the cytosines in the center but not detected as 4mC (Wei et al., 2019). The new dataset can be downloaded from our github, and the download link is given in section 3.

2.2. Model of 4mcDeep-CBI

2.2.1. Preliminary Feature Extraction

We use the eight features mentioned in Chen et al. (2017), He et al. (2018), Wei et al. (2018), and Wei et al. (2019) as preliminary features. These features are obtained by encoding the different sequence information by the feature representation algorithm of the sequence. These features are BKF (Binary and k-mer frequency), DBPF (Dinucleotide binary profile and frequency), KNN (K-Nearest Neighbor), PCP (Physical-Chemical Properties), MMI (Multivariate Mutual Information), PseDNC (Pseudo dinucleotide composition), PseEIIP (Electron-ion interaction pseudopotentials of trinucleotide) and RFHCP (Ring-function-hydrogen-chemical properties). The related feature extraction methods can be found in Wei et al. (2019).

2.2.2. 4mcDeep-CBI Network

As shown in **Figure 1**, 4mcDeep-CBI consists of 3-CNN layer, BLSTM layer, fully connected layer, and a sigmoid classifier. The input of 4mcDeep-CBI is one of eight preliminary features. First of all, the preliminary feature is used as the input to 3-CNN layer, which contains convolution layer, ReLU activation function and max pooling operation. Next, the output of 3-CNN layer will be imported to BLSTM layer to obtain an advanced feature. With the eight features as the inputs, we can get eight advanced features, respectively. Then, each advanced feature (matrix) will be further converted to one-dimensional feature (vector) using the flatten function, which will be finally connected to the fully connected layer. The last layer is the sigmoid layer, which is used to obtain advanced probability features and the prediction result of the first step. At last, we get an eight-dimensional feature, which will be the input of the integrated algorithm.

2.2.2.1. Convolutional neural network (CNN)

CNN has a powerful ability to extract abstract features, which is not only suitable for image processing, but also for natural language processing tasks. It consists of convolution, activation, and max-pool layers.

In the model design, since we have verified in experiment that the model with 3 CNN layers has the best performance, we employ 3-CNN as an advanced feature extractor, and the input is the preliminary feature extracted from DNA sequences. We first put the preliminary features into the 3-CNN layer, respectively, and set the weighting parameters of the convolution filter. Then, the convolution layer outputs the matrix inner product between the input preliminary feature and filters. After convolution, a rectified linear unit (ReLU) is applied to sparsify the output of the convolution layer. The Rectified Linear Unit (ReLU) (Nair et al., 2010) takes the output of a convolution layer and clamps all the negative values to zero to introduce non-linearity that can not only reduce the computational cost, but also avoid the phenomenon of vanishing gradient and over-fitting. Finally, a max pooling operation is used to reduce the dimensionality and over-fitting by taking the maximum value in a fixed-size sliding window. The output of the convolution module is represented by the following expression:

$$O_c = Pool\Big(ReLU\big(Conv(S)\big)\Big),$$

where O_c is the output tensor, S is the input preliminary feature of the sequence. For BKF as an example, the dimension of S is $1 \times 500 \times 1$ (input_shape). The nb_filter of 3-CNN are 16, 32, 64, respectively, and the filter_length of 3-CNN are all 8. The parameters of max pool is 2. Therefore, the dimension of O_c is $1\times223\times64$.

2.2.2.2. Long short term memory networks (LSTM)

LSTM is a recurrent neural network (RNN) architecture (an artificial neural network) published in 1997 (Hochreiter and Schmidhuber, 1997). Compered with traditional RNNs, LSTM network is well-suited to learn from experience to classify, process and predict time series, and it has advantages in dealing with long term dependency. Especially, Bidirectional LSTM can capture the bidirectional dependence of features and the outputs of individual directions are concatenated, which can well mine the deeper information in the features:

$$O_r = BiLSTM(\, O_c \,),$$

where O_r is the output of BLSTM layer and is also advanced feature of the sequence, O_c is the feature matrix of a sequence obtained by the 3-CNN layer. A LSTM contains a forget gate layer, an input gate layer and an output gate layer. When the LSTM traverses each element of the input, it first determines what information the forget gate layer is about to discard based on the previous input. The input gate layer then determines what information should be stored for the next layer and updates the current state value. Finally, the output gate layer will only output the part of our output that we determined (Pan and Shen, 2018).

2.3. Integrated Algorithm Model

In the integrated algorithm model, there are six machine learning algorithms involved, which are K-nearest neighbor algorithm, Logistic regression algorithm, Support vector machine algorithm, Naive Bayesian algorithm, Decision tree algorithm, and Random forest algorithm, respectively. With the 8-D advanced feature of the sequence as the input, we run these six different machine learning algorithms to predict the labels, and get the best

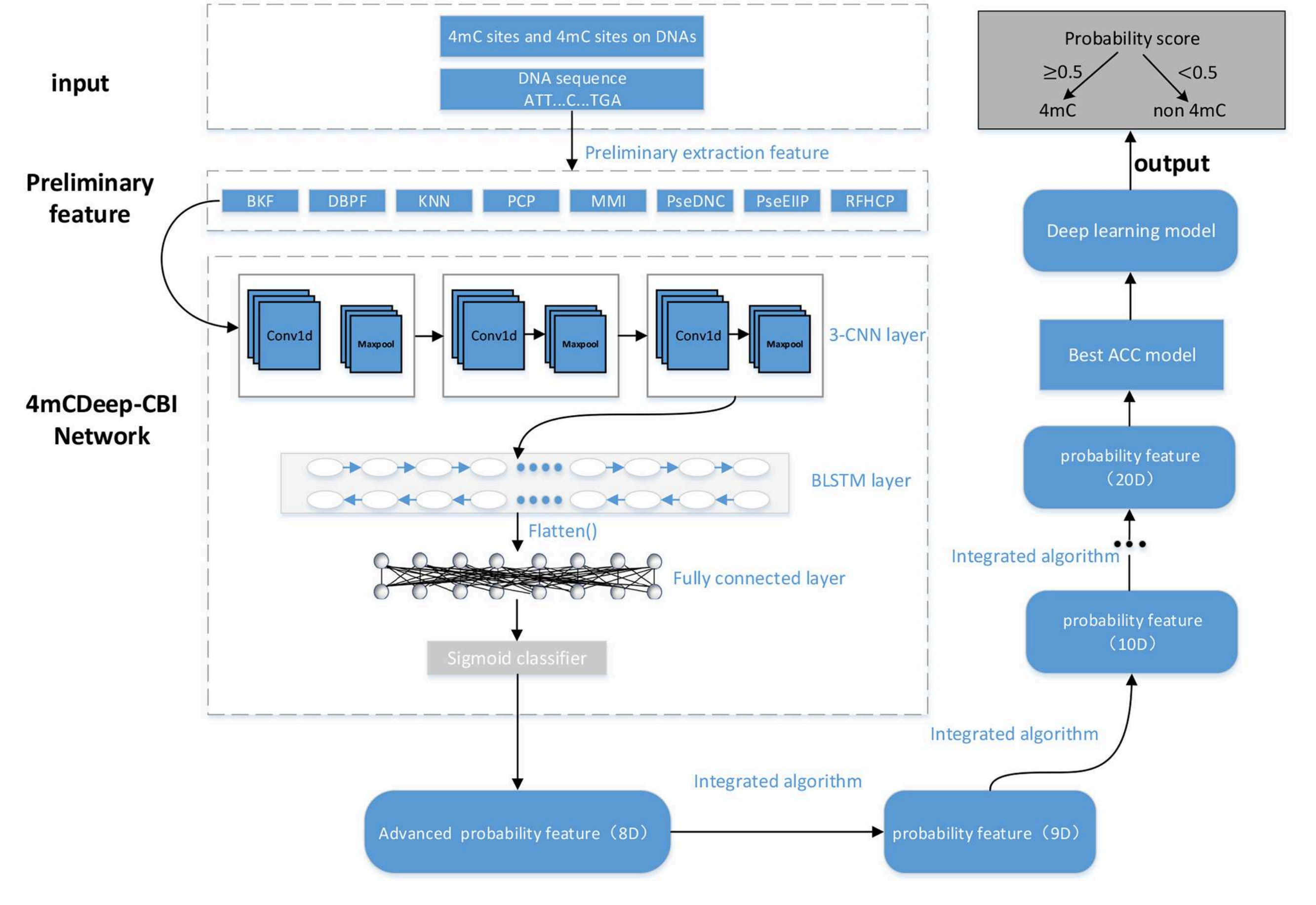


FIGURE 1 | A graphical illustration of the 4mcDeep-CBI model.

result. Then, the obtained probability value is combined with the previous 8-D advanced feature vector to form a new 9-D feature vector. Next, the 9-D feature are imported into the integrated algorithm model for the new iteration. This process will be repeated until performance reaches convergence. In each iteration, the multi-dimensional input features are trained, and the optimal algorithm is selected each time to obtain an one-dimensional probability feature, and then the input and output features are merged into a new feature vector which has one more dimension than the input and will be the new input for next iteration. For example, it is supposed that the vectors $f_1, f_2, \ldots, f_8$ are the advanced features obtained by previous processing, and with $(f_1, f_2, \ldots, f_8)$ as the algorithm input, we can get the result vector f_9. Then, $(f_1, f_2, \ldots, f_8, f_9)$ will be the algorithm input of the next iteration. If there are 5 iterations, we will get the result $(f_1, f_2, \ldots, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13})$ which will be the feature matrix for the following processing. In the experiment, after less than 10 iterations, the algorithm can reach the state of convergence, which can be shown in section 3.

2.4. Deep Learning Model

For the last part of 4mcDeep-CBI, a general neural network model is used to get the optimal solution. The neural network has 2–4 intermediate layers, each with a different activation function. In our experiment, we used two layers of intermediate layers, each using the ReLU function as the activation function, and finally used the sigmoid function as the output layer. We found that inputting the advanced feature matrix obtained by the integrated algorithm into the neural network model can further improve the accuracy.

2.5. Performance Evaluation

For performance evaluation, we used the following five generally-used metrics: Sensitivity (SN), Specificity (SP), Accuracy (ACC), Mathew's Correlation Coefficient (MCC) (Wei et al., 2019) and Area Under the ROC Curve (AUC). The definition of each evaluation metric is as follows:

$$
\begin{aligned}
SN &= \frac{TP}{TP + FN}, \\
SP &= \frac{TN}{TN + FP}, \\
ACC &= \frac{TP + TN}{TP + TN + FN + FP}, \\
MCC &= \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}},
\end{aligned}
$$

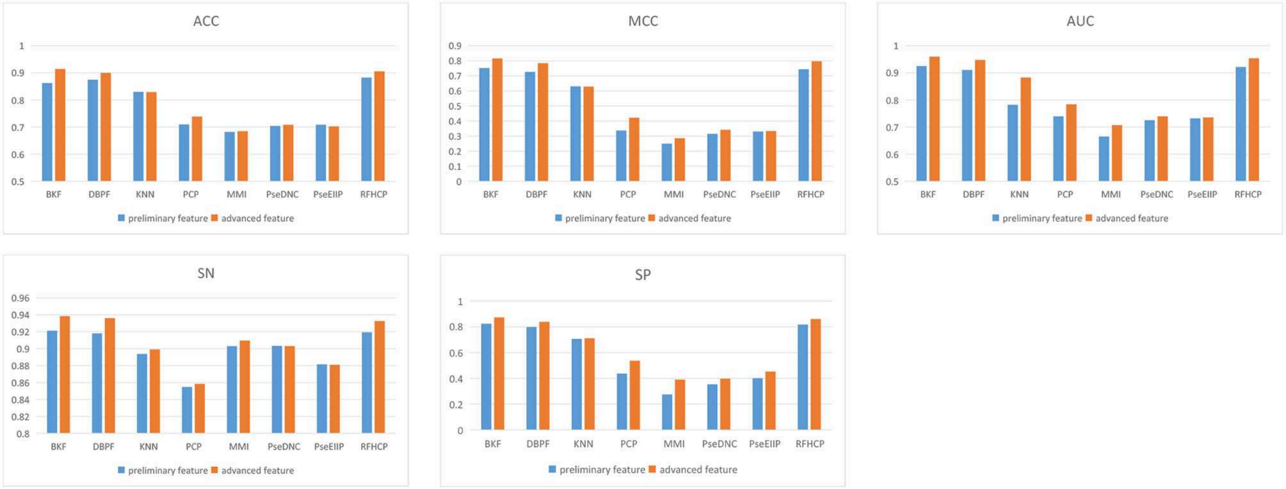


FIGURE 2 | Evaluate the performance of preliminary feature and advanced feature on the same data set.

where TP indicates that the actual result is a positive sample, and the predicted result is also a positive sample; TN indicates that the actual result is a negative sample, and the predicted result is also a negative sample; FP indicates that the actual result is a negative sample, and the predicted result is a positive sample (indicating that the negative sample is predicted incorrectly); FN indicates that the actual result is a positive sample, and the prediction result is a negative sample (indicating that the positive sample is predicted incorrectly).

The area under the ROC curve (AUC) is a comprehensive used metric. The abscissa of the ROC curve is the false positive rate and the ordinate is the positive rate. The AUC value is the enclosed area value of the ROC curve and the coordinate axis, and the value is between 0 and 1. The maximum value of AUC is 1, which means that the performance of the model is perfect, and all prediction results are correct. AUC value of 0 means that the model performance is very poor, and all prediction results are wrong.

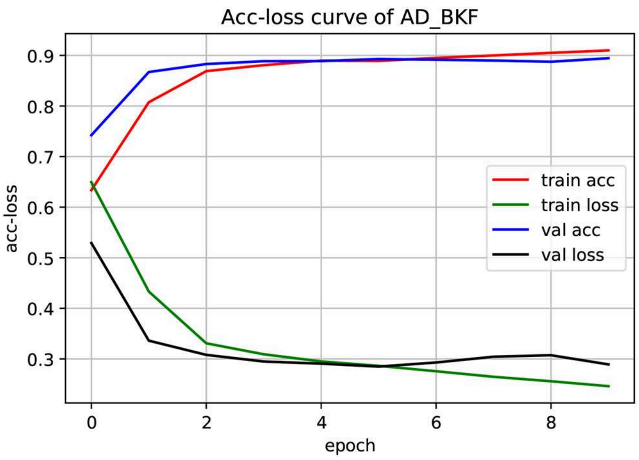


FIGURE 3 | Acc-loss curve of AD_BKF based on 3-CNN and BLSTM models. Where AD_BKF is a advanced feature of BKF.

3. RESULT AND DISCUSSION

We have done extensive experiments on the new dataset using the proposed predictor (4mcDeep-CBI) and the state-of-art predictor (4mcPred-IFL), respectively, then we make a performance comparison between two models. The dataset and code used in the experiment have been uploaded to our GitHub (https://github.com/mat310/4mcDeep), which is shared with other researchers. Due to limited space, part of experimental results are listed in **Supplementary Material**.

3.1. Performance of Different Features Used in Prediction

We put 8 preliminary features into the 3-CNN and BLSTM models to obtain advanced features. Then the advanced feature are sequentially passed through sigmoid classifier to obtain the prediction result of the first step. We performed different types of features for predictive performance analysis and compared the experimental results of 4mcPred-IFL with 4mcDeep-CBI. From **Figure 2**, we find that the predicted performance of the four features BKF, DBPF, KNN, and RFHCP ranks in the top four in the experimental results of both modes. In addition, the performance metrics of the eight characteristic experimental results have been improved in our model (The experimental results can be found in **Tables S1**, **S2**). **Figure 2** shows that our proposed model performs better than 4mcPred-IFL in the preliminary experimental results.

The experiment used a three-fold cross-validation. As shown in **Figure 3**, this is the acc-loss curve of AD_BKF during the preliminary experiment (acc-loss curves of other advanced feature can be found in **Figure S1**). Epoch refers to the number of times when all data were sent into the network to complete

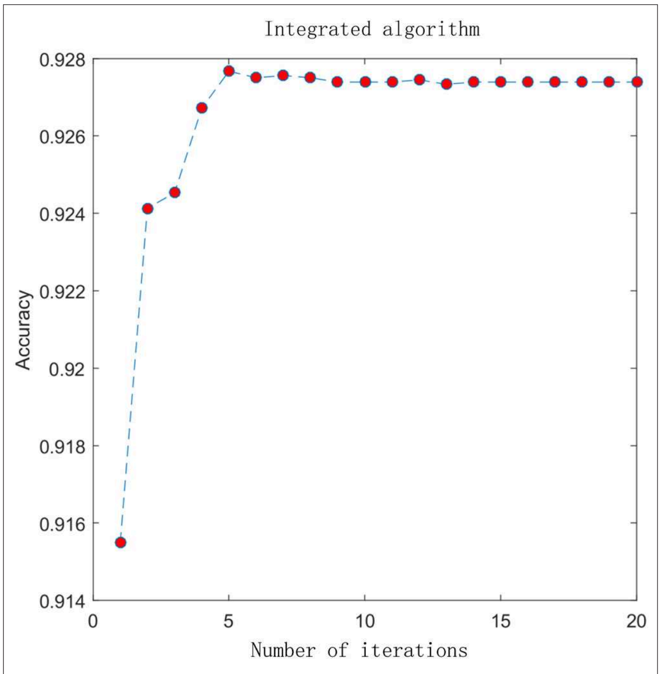


FIGURE 4 | Experimental result graph after using integrated algorithm.

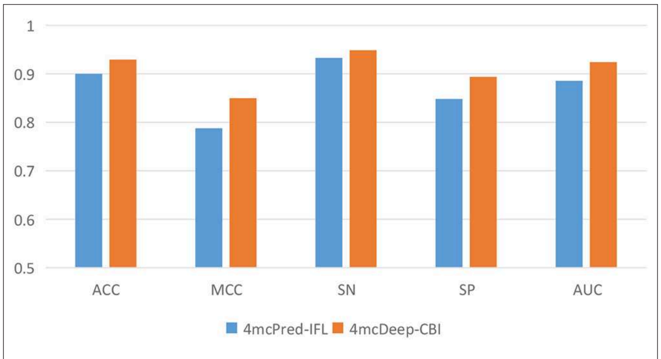


FIGURE 5 | Performance evaluation of our predictor and the state-of-the-art predictor on the same dataset.

one forward calculation and back propagation. As can be seen from the figure, with the increase of epoch value, the accuracy of the training set and verification set increased continuously, and finally converged at epoch = 5. The loss function values of the training set and verification set decreased continuously, and finally converged when epoch = 5. Therefore, we can set epoch = 5 to get the best experimental results. **Figure 3** illustrates that the prediction performance is continuously improved and there is no over-fitting during the experiment.

3.2. Performance of the Integrated Algorithm

In the previous section, we compared the experimental results of different advanced features. Here, we combine the advanced probability features obtained from the sigmoid classifier to form a matrix with 8-D probabilistic feature. This matrix is input into the integrated algorithm model and we get the experimental results. To visually analyze the results, we plot the ACC change with the increment of the feature size, which is shown in **Figure 4**. In the figure, the X-axis represents the number of iterations and the Y-axis represents the performance in terms of accuracy. Before performing the iterative operation, we have a matrix with 8-D probabilistic feature. As the number of iterations increases, performance increases rapidly from the beginning, reaching a maximum after 5 iterations when the feature size of the matrix is 13 and ACC is 0.9274, then gradually converge to a steady state. This suggests that the integrated algorithm model can improve feature representation and surely improve performance. 4mcPred-IFL adopted an iterative feature representation algorithm, which reached the maximum when the number of iterations was 30 and ACC was 0.9001, and then gradually converges to a stable state. The details can be found in **Figure S2**.

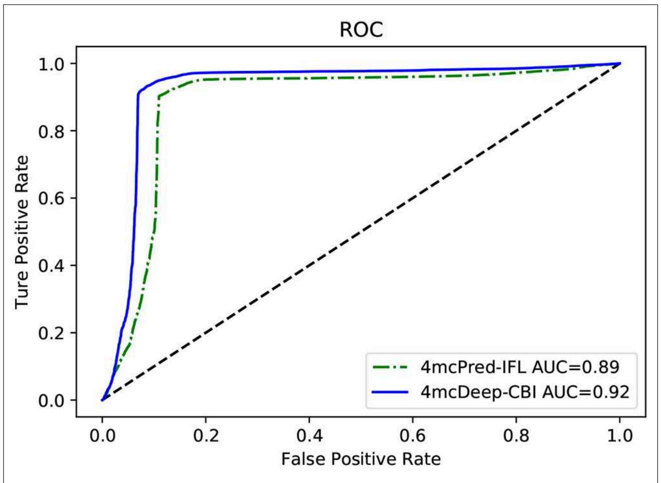


FIGURE 6 | ROC curves of our predictor and the state-of-the-art predictor on the same dataset.

3.3. 4mcDeep-CBI vs. State-of-Art Predictor on Performance

Our 4mcDeep-CBI model shows the best predictive performance, and we achieve ACC = 0.9294, MCC = 0.8498, SN = 0.9486, SP = 0.8938, AUC = 0.9242. To further evaluate the performance of our predictor 4mcDeep-CBI, we compared our predictor with the state-of-art predictor: 4mcPred-IFL. The performances of 4mcDeep-CBI and 4mcPred-IFL are depicted in **Figures 5**, **6**, respectively. **Figure 5** illustrates the performances in terms of ACC, MCC, SN, SP, and AUC, while **Figure 6** shows the ROC curves of 4mcDeep-CBI and 4mcPred-IFL. The details of their performances can be found in **Table S3**. It can be clearly seen that 4mcDeep-CBI achieved better performance than 4mcPred-IFL in all five metrics. Our predictor improves ACC by 3.26%. It is worth noting that our predictor increased the MCC by 7.88%. MCC is essentially a correlation coefficient between the actual classification and the prediction classification, and is a

TABLE 1 | Running time of the main modules of 4mcPred-IFL and 4mcDeep-CBI.

	Running_time (minute)		
Sample size	**SVM_10**	**SVM_50**	**4mcDeep-CBI**
1,000	31.3	9.2	3.1
4,000	1034.4	222.1	10.7
7,000	3123.6	698.4	19.8
10,000	6255.8	1365.6	24.5
13,000	9449.5	2173.2	35.1
16,000	15094.4	3261.3	48.2

TABLE 2 | ACC of 4mcDeep-CBI with 4 CNN layers under different parameters.

nb_filter	Filter_length	ACC (%)
4, 8, 16, 32	4, 4, 4, 4	90.02
4, 8, 16, 32	8, 8, 8, 8	89.46
4, 8, 16, 32	16, 16, 16, 16	88.70
8, 16, 32, 64	4, 4, 4, 4	90.17
8, 16, 32, 64	8, 8, 8, 8	90.02
8, 16, 32, 64	16, 16, 16, 16	89.25
16, 32, 64, 128	4, 4, 4, 4	89.78
16, 32, 64, 128	8, 8, 8, 8	89.37
16, 32, 64, 128	16, 16, 16, 16	89.18
32, 16, 8, 4	4, 4, 4, 4	89.36
32, 16, 8, 4	8, 8, 8, 8	89.29
32, 16, 8, 4	16, 16, 16, 16	88.31
64, 32, 16, 8	4, 4, 4, 4	89.89
64, 32, 16, 8	8, 8, 8, 8	88.72
64, 32, 16, 8	16, 16, 16, 16	87.97
128, 64, 32, 16	4, 4, 4, 4	90.03
128, 64, 32, 16	8, 8, 8, 8	89.96
128, 64, 32, 16	16, 16, 16, 16	89.09

relatively comprehensive metric. This shows that 4mcDeep-CBI is better than 4mcPred-IFL in terms of comprehensiveness and integrity.

The ROC curve between the different methods is shown in **Figure 6**. As can be seen from the figure, the ROC curve of 4mcDeep-CBI is closer to the upper left corner, and the area under the ROC curve is the largest, which is 4.35% larger than that of 4mcPred-IFL. In summary, the above results illustrate that the performance of 4mcDeep-CBI is better than 4mcPred-IFL, and 4mcDeep-CBI can effectively improve the accuracy of identifying 4mC sites.

3.4. 4mcDeep-CBI vs. State-of-Art Predictor on Running Time

The running time of the main modules of 4mcPred-IFL and 4mcDeep-CBI accounts for a large proportion in their respective models. Among them, the main module of 4mcPred-IFL refers to the preliminary experimental results obtained by putting the extracted preliminary features into the SVM model. The main module of the 4mcDeep-CBI model refers to the preliminary experimental results obtained by putting the extracted preliminary features into the deep learning model. In order to explore the operational efficiency of the model, we run the main modules of 4mcPred-IFL and 4mcDeep-CBI separately on the same server. The preliminary feature is BKF as an example. Experiments are carried out with different sample sizes. The results obtained are shown in **Table 1**. 4mcPred-IFL employed Sequential Forward Search (SFS) to determine the optimal feature subset. In **Table 1**, "SVM_10" refers to the distance of the SFS is 10, and "SVM_50" refers to the distance of the SFS is 50. The smaller the distance setting, the greater the possibility of better experimental results, and the longer the experiment runs. In addition, when the distance range from 10 to 50, the optimal subset of features can be obtained. As we can see in **Table 1**, our model runs much faster than the state-of-art predictor. After running 16,000 samples, 4mcDeep-CBI need 48.2 min only, but even if the distance is set to 50, 4mcPred-IFL takes 3261.3 min to run. The running time is more than 50 times slower than us. Moreover, as the number of samples increased, 4mcDeep-CBI grew more slowly than 4mcPred-IFL. There are at least two reasons: (1) The efficiency of 4mcpred-IFL using SFS method to obtain the optimal feature set is very slow. (2) There are two important parameters (the penalty parameter C and the kernel parameter γ) in the SVM model used by 4mcPred-IFL. Meanwhile, 4mcPred-IFL takes a lot of time to call SVM algorithm over and over again to optimize the penalty parameter C and the kernel parameter γ by using the grid search method. Consequently, the complexity of the 4mcpred-IFL model is much higher than our proposed model.

3.5. Impact of Different CNN Layers on 4mcDeep-CBI

In the proposed model 4mcDeep-CBI, we have three CNN layers which can efficiently extract the features from input data. In the experiment, with the CNN layers given, we obtain the accuracy of the 4mcDeep-CBI, and we make a performance comparison according to different CNN layers. For feature RFHCP, **Table 2** shows the experimental results of the 4mcDeep-CBI with 4 CNN layers. Parameters are set as batch_size = 32, 64, 128, 256; maxpool1D = 1, 2, 3; learning rate = 0.001, 0.005, 0.0001; dropout ratio = 0.1, 0.2, 0.5. It can be found from **Table 2** that the maximum ACC value is 90.17% when the 4mcDeep-CBI has 4 CNN layers. Similarly, we do experiments based on different (2, 3, 5, and 7) CNN layers. The experimental results are shown in **Figure 7**. As can be seen from **Figure 7**, maximum ACC value is 90.57% when the 4mcDeep-CBI has 3 CNN layers. For other features, the experiment has the same result. Therefore, the experiment verifies that 3-CNN layer model has the best performance, that is why we choose 3 CNN layers in the model design of the 4mcDeep-CBI.

4. CONCLUSION

In this paper, we propose a deep neural network named 4mcDeep-CBI, which can further boost the performance of identifying 4mC sites. Moreover, we found a large number of

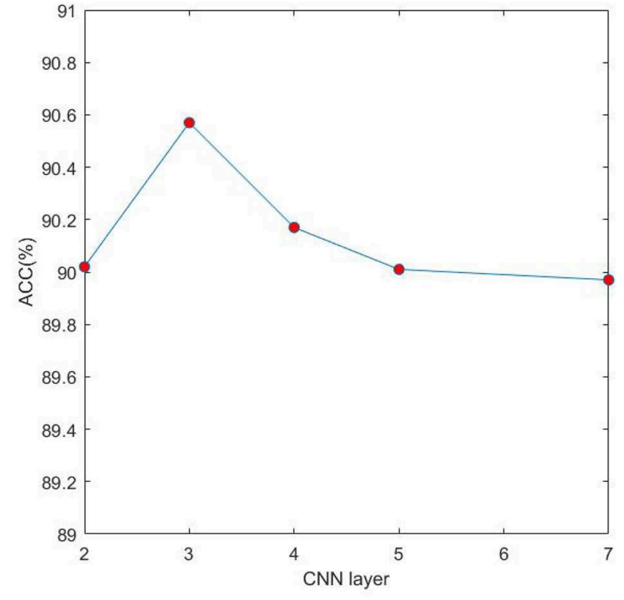


FIGURE 7 | Impact of different CNN layers on ACC.

4mC sites and non 4mC sites of *C. elegans* from the latest MethSMRT website, which greatly expanded the data set of *C. elegans*. The proposed model 4mcDeep-CBI uses 3-CNN and BLSTM modules to mine deep information of features to obtain advanced features. By experimental comparison with the state-of-art predictor, we found that our proposed framework performed better than the state-of-art predictor, and our model did not appear to have an over-fitting phenomenon. In addition, we have proposed an integrated algorithm to generate informative features. By analyzing the accuracy of the model during the iterative process, we find that the integrated algorithm is constantly improving the performance of the model. Finally, we evaluated our proposed 4mcDeep-CBI with the state-of-art predictor, and the results demonstrate that our model can achieve better performance in identifying 4mC sites and runs more efficiently. We hope that 4mcDeep-CBI can be an useful bioinformatics tool for identifying 4mC sites and promoting the DNA methylation analysis.

Deep learning is an important way of sequence analysis. For feature selection, we can use the most popular word embedding training method: Word2Vec algorithm, which can be combined with the secondary structure of DNA to predict 4mC sites. Moreover, the sequence length provided by the MethSMRT website is 41 bp, and we need longer DNA sequence fragments, such as 80, 100, and 150 bp to do further research.

DATA AVAILABILITY STATEMENT

The dataset and code used in the experiment have been uploaded to our GitHub (https://github.com/mat310/4mcDeep).

AUTHOR CONTRIBUTIONS

FZ and GF design the model, experiments, and wrote the paper. GF performed the experiments. LY analyzed the data, provided the suggestions to improve the performance, and contributed the materials and analysis tools.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments. This work is supported in part by the National Science Foundation of China (Grant No. 61502159).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00209/full#supplementary-material

REFERENCES

Allis, C. D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* 17, 487–500. doi: 10.1038/nrg.2016.59

Blow, M. J., et al. (2016). The epigenomic landscape of prokaryotes. *PLoS Genet.* 12:e1005854. doi: 10.1371/journal.pgen.1005854

Carpenter, M. A., Li, M., Rathore, A., Lackey, L., Law, E. K., Land, A. M., et al. (2012). Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J. Biochem.* 287, 34801–34808. doi: 10.1074/jbc.M112.385161

Chen, W., Yang, H., Feng, P., Ding, H., and Lin, H. (2017). iDNA4mC: identifying DNA N4-methylcytosine sites based on nucleotide chemical properties. *Bioinformatics* 33, 3518–3523. doi: 10.1093/bioinformatics/btx479

Cheng, X. (1995). DNA modification by methyltransferases. *Curr. Opin. Struct. Biol.* 5, 4–10. doi: 10.1016/0959-440X(95)80003-J

Chou, K. C. (2015). Impacts of bioinformatics to medicinal chemistry. *Med. Chem.*, 11, 218–234. doi: 10.2174/1573406411666141229162834

Clark, T. A., Lu, X., Luong, K., Dai, Q., Boitano, M., Turner, S. W., et al. (2013). Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* 11:4. doi: 10.1186/1741-7007-11-4

Davis, B. M., Chao, M. C., and Waldor, M. K. (2013). Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol* 16, 192–198. doi: 10.1016/j.mib.2013.01.011

Deng, L., Li, W., and Zhang, J. (2019). DAH2V: exploring meta-paths across multiple networks for lncRNA-disease association prediction. *IEEE/ACM Trans Comput Biol Bioinform.* doi: 10.1109/TCBB.2019.2946257. [Epub ahead of print].

Ecker, J. R. (2010). Zeroing in on DNA methylomes with no BS. *Nat. Methods* 7, 435–437. doi: 10.1038/nmeth0610-435

Fazakerley, G. V., Kraszewski, A., Teoule, R., and Guschlbauer, W. (1987). NMR and CD studies on an oligonucleotide containing NM-methylcytosine. *Nucl. Acids Res.* 15, 2191–201. doi: 10.1093/nar/15.5.2191

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., et al. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465. doi: 10.1038/nmeth.1459

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Fu, Y., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., et al. (2015). N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. *Cell* 161, 879–892. doi: 10.1016/j.cell.2015.04.010

He, W., Jia, C., and Zou, Q. (2018). 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 35, 593–601. doi: 10.1093/bioinformatics/bty668

Heyn, H., and Esteller, M. (2015). An adenine code for DNA: a second life for N6-methyladenine. *Cell*. 161, 710–713. doi: 10.1016/j.cell.2015.04.021

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Iyer, L. M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci*. 101, 25–104. doi: 10.1016/B978-0-12-387685-0.00002-0

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2019). DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *IEEE/ACM Trans. Comput. Biol. Bioinformat*. 48, 871–881. doi: 10.1093/nar/gkz1007

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa), 807–814.

O'Brown, Z. K., and Greer, E. L. (2016). "N6-Methyladenine: a conserved and dynamic DNA mark," in *DNA Methyltransferases - Role and Function*, eds R. Jurkowska and A. Jeltsch (Springer), 945, 213–46. doi: 10.1007/978-3-319-43624-1_10

Pan, X., and Shen, H. B. (2018). Predicting RNA-protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*34, 3427–3436. doi: 10.1093/bioinformatics/bty364

Pan, Y., Wang, Z., Zhan, W., and Deng, L. (2018). Computational identification of binding energy hot spots in protein-rna complexes using an ensemble approach. *Bioinformatics* 34, 1473–1480. doi: 10.1093/bioinformatics/btx822

Schweizer, H. P. (2008). Bacterial genetics: past achievements, present state of the field, and future challenges. *Biotechniques* 44, 633–641. doi: 10.2144/000112807

Suzuki, M. M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet*. 9, 465–476. doi: 10.1038/nrg2341

Wang W.-W., and Qiu L. H. (2012). Current review on DNA methylation in Ovarian cancer. *J. Int. Reproduct. Health Family Plan*. 9, 465–476.

Wang, H., Liu, C., and Deng, L. (2018). Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep*. 8:14285. doi: 10.1038/s41598-018-32511-1

Wei, L., Luan, S., Nagai, L. A. E., Su, R., and Zou, Q. (2018). Exploring sequence-based features for the improved prediction of DNA N4-methylcytosine sites in multiple species. *Bioinformatics* 35, 1326–1333. doi: 10.1093/bioinformatics/bty824

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Ye, P., Luan, Y., and Xie, X. (2017). MethSMRT: an integrative database for DNA N6-methyladenine and N4-methylcytosine generated by single-molecular real-time sequencing. *Nucl. Acids Res.* 45, 85–89. doi: 10.1093/nar/gkw950

Yu, M., Ji, L., Neumann, D. A., Chung, D. H., Groom, J., Westpheling, J., et al. (2015). Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. *Nucl. Acids Res*. 43:e148. doi: 10.1093/nar/gkv738

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: 15 April 2020
doi: 10.3389/fgene.2020.00272

GBDTL2E: Predicting lncRNA-EF Associations Using Diffusion and HeteSim Features Based on a Heterogeneous Network

Jiaqi Wang, Zhufang Kuang, Zhihao Ma and Genwei Han*

School of Computer and Information Engineering, Central South University of Forestry and Technology, Changsha, China

Interactions between genetic factors and environmental factors (EFs) play an important role in many diseases. Many diseases result from the interaction between genetics and EFs. The long non-coding RNA (lncRNA) is an important non-coding RNA that regulates life processes. The ability to predict the associations between lncRNAs and EFs is of important practical significance. However, the recent methods for predicting lncRNA-EF associations rarely use the topological information of heterogenous biological networks or simply treat all objects as the same type without considering the different and subtle semantic meanings of various paths in the heterogeneous network. In order to address this issue, a method based on the Gradient Boosting Decision Tree (GBDT) to predict the association between lncRNAs and EFs (GBDTL2E) is proposed in this paper. The innovation of the GBDTL2E integrates the structural information and heterogenous networks, combines the Hetesim features and the diffusion features based on multi-feature fusion, and uses the machine learning algorithm GBDT to predict the association between lncRNAs and EFs based on heterogeneous networks. The experimental results demonstrate that the proposed algorithm achieves a high performance.

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Jingpu Zhang,
Henan University of Urban Construction, China
Hui Liu,
Changzhou University, China

****Correspondence:***
Zhufang Kuang
zfkuangcn@163.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *23 November 2019*
Accepted: *06 March 2020*
Published: *15 April 2020*

Citation:
Wang J, Kuang Z, Ma Z and Han G (2020) GBDTL2E: Predicting lncRNA-EF Associations Using Diffusion and HeteSim Features Based on a Heterogeneous Network. Front. Genet. 11:272. doi: 10.3389/fgene.2020.00272

Keywords: long non-coding RNA, environmental factor, heterogenous network, HeteSim score, gradient boosting decision tree, random walk with restart

1. INTRODUCTION

The environment factor (EF) is a biological or non-biological factor that affects a living organism. Non-biological factors include physical factors, chemical factors, and social factors. Biological factors include parasites and viruses. Many studies have demonstrated that Gene-Environment (G–E) interactions play an important role in the etiology and progression of many complex diseases (Xu et al., 2019). Alzheimer's disease (AD), for example, is a disease that manifests as many intertwined factors, including environmental factors and the like (Eid et al., 2019). Moreover, fetal death and coronary-heart-disease (CHD) could also be caused by G–E interactions (Moreau et al., 2019).

According to the central law of molecular biology, genetic information is mainly saved in DNA sequences. Genetic information is transcribed from DNA into RNA, which is then translated into proteins. Genome sequence analysis shows that the protein-coding sequences account for about 2% of the human genome, and 98% are non-encoding protein sequences (Bertone et al., 2004). In biology, RNAs that do not code are called non-coding RNAs (ncRNAs). In ncRNAs, ncRNAs with a length between 200 and 100,000 nt are called Long non-coding RNAs (lncRNAs), and these play an important role in the understanding of life sciences (Deng et al., 2018). LncRNAs are significant in many aspects, such as in cellular biological processes, gene expression regulation at transcriptional and post-transcriptional levels, and others (Zhang Z. et al., 2019).

There are many studies on the biological mechanism and interaction between genes, microRNAs (miRNAs), lncRNAs, EFs, and diseases, such as the relationship between genes and diseases, miRNAs and diseases, lncRNAs and diseases, miRNAs and EFs, etc. Among them, microRNA (miRNA) is a kind of non-coding RNA that has only about 21–25 nucleotides (Deng et al., 2019b).

For the association between genes and diseases, a data synthesis platform based on gene variation and gene expression was established by Luo et al.. This method applies the method of network analysis to predict the interaction between genes and diseases (Luo Z. et al., 2018). The recent advances in predicting gene–disease associations have been reviewed by Opap and Mulder (2017). An understanding of the association between genetics and disease is an important step in understanding the etiology of diseases. There are many other studies about the association between genes and diseases. Due to the limitation of space, only a few studies have been introduced here.

For the association between miRNAs and diseases, KBMF-MDI was proposed by Lan et al. KBMF-MDI predicts the association between miRNAs and diseases based on their similarities to diseases (Lan et al., 2018), and this is a method that is based on the dynamic neighborhood regularized logical matrix factorization (DNRLMF-MDA) proposed by Yan et al. (2017). The IMCMDA (Chen et al., 2018) was subsequently proposed by Chen et al.. The IMCMDA is an inductive matrix filling model. A new computational model, called heterogeneous graph convolutional network (HGCNMDA) (Li et al., 2019), was presented by Li et al., and another method, the double Laplace regularization (DLRMC) matrix completion model, is proposed by Tang et al. (2019). Those studies have proven that the computational model could effectively predict the potential miRNA-disease associations and provide convenience for the verification experiment of biological researchers.

For the association between lncRNAs and diseases, a method to predict the association between human lncRNAs and diseases based on the random walk of the global network was proposed by Gu et al. (2017). The BRWLDA proposed by Yu et al. is a method to predict the lncRNA-disease associations based on the double random walk of heterogeneous networks (Yu et al., 2017). A global network-based framework named LncRDNetFlow (Zhang J. et al., 2019) was proposed by Zhang et al. LncRDNetFlow utilizes a flow propagation algorithm to predict lncRNA-disease associations. The calculation method LDASR was proposed by Guo et al. (2019). The LDASR analyzes the relationships between known lncRNAs and diseases to identify the relationships between lncRNAs and diseases. A bipartite graph network based on the known lncRNA-disease associations was constructed by Ping et al. (2018), and a bilateral sparse self-representation (TSSR) algorithm was proposed by Ou-Yang et al. (2019) to predict lncRNA-disease associations. A new method of lncRNA-disease-gene tripartite mapping (TPGLDA) was proposed by Ding et al. to predict the associations of lncRNA-disease, which combined the associations of gene-disease and lncRNA-disease (Ding et al., 2018). A new potential factor mixture model (LFMMs) estimation method was constructed by Caye et al. (2019), and the model is implemented in the updated version of the corresponding computer program. The ILDMSF is a novel framework that was proposed by Chen et al. (2020). Furthermore, a method named LDAH2V (Deng et al., 2019a) was proposed by Deng et al., and the HIN2Vec is used to calculate the meta-path and feature for each lncRNA-disease in the heterogeneous networks.

For the association between miRNAs and EFs, the MiREFRWR was proposed by Chen et al., and it uses the Random Walk with Restart algorithm in a complex network to predict interactions (Chen, 2016). The MEI-BRWMLL (Luo H. et al., 2018) method to reveal the relationships of miRNAs and EFs was proposed by FLuo et al.. In this approach, multi-label learning and double random walk are used to predict the associations between miRNAs and EFs. These studies provide directional guidance for the analysis of complex diseases and the association between miRNAs and EFs in clinical trials (Chen et al., 2012; Qiu et al., 2012).

With the application of computing technology in the field of biology, more and more public biological databases have also been established, such as HMDD (Huang et al., 2018), miR2Disease (Jiang et al., 2008), DrugCombDB (Liu et al., 2020), and gutMDisorder (Cheng et al., 2020).

The development of genomics and bioinformatics facilitated the identification of lncRNA. LncRNA has also been found to interact with various EFs, such as chemicals, smoking, and air pollution (Flynn and Chang, 2014). It has been found that these lncRNAs and EFs may be the cause of some diseases (Chen and Yan, 2013). However, compared with protein-coding genes and miRNAs, there are fewer methods using bioinformatics and computational methods to study the association between lncRNAs and EFs, and these are also less effective. Based on the restart random walk model, the RWREFD method and a lncRNA-EF associations database, LncEnvironmentDB, were designed by Zhou et al. (2014). A method based on a binary network and resource transfer algorithm to predict the associations of lncRNA-EF was designed by Zhou and Shi (2018). The KATZ measure and Gaussian interaction profile kernel similarity are used to predict new potential associations between lncRNAs and EFs, as proposed by Vural and Kaya (2018). Three computational models for predicting the relationship between lncRNAs and EFs using the similarity of gaussian interaction properties of lncRNAs and EFs were proposed by Xu (2018). They are the prediction methods of lncRNAs and EFs association based on the Laplacian regularized least square method, the KATZ method, and the double random walk algorithm. The above studies show that the computational approach can improve the speed and reduce the cost.

However, the aforementioned studies for predicting the association between disease-related lncRNAs and EFs usually use traditional similarity search methods, which focus on measuring the similarity between objects of the same type. Those existing methods to study the association between disease-related lncRNAs and EFs simply treat all objects as the same type without considering different subtle semantic meanings of different paths in the heterogeneous network. This will reduce the accuracy and persuasiveness of the results. In this paper, we have proposed a high-performance method to predict the correlation between lncRNAs and EFs based on heterogeneous networks. The proposed method integrates the structural information and heterogenous networks and combines the Hetesim features and the diffusion features as data features and uses the GBDT algorithm as a prediction model. The HeteSim features are a path-based measurement method in heterogeneous networks and can measure the relationship between objects of the same or different types. The Hetesim has not been used to predict the association between lncRNAs and EFs. It is the first time that the Hetesim is integrated as a fusion feature in the step of feature extraction for predicting the association between lncRNAs and EFs. The method GBDT is used in the proposed algorithm, which is an integrated learning method in machine learning, and has superior accuracy compared with other algorithms. It is also the first time that the integrated learning method GBDT is used to investigate the association between lncRNAs and EFs. From our perspective, on the one hand, our proposed method provides an efficient calculation method for mining the association between lncRNAs and EFs, which greatly saves manpower and material resources. On the other hand, it also helps biologists to explore the influence of environmental factors on diseases.

For the rest of the paper, the materials and methods have been presented in section 2, the experimental results and evaluates have been discussed in section 3, and, finally, we have concluded this paper in section 4.

2. MATERIALS AND METHODS

The data used in this experiment are downloaded from the DLREFD database (Sun et al., 2017). The data include 475 lncRNAs and 152 environmental factors. After the duplicate data are removed, the number of correlations between lncRNAs and EFs was 735. The set of lncRNAs and the set of EFs are shown in **Supplementary Material**.

A method based on the Gradient Boosting Decision Tree (GBDT) to predict the association between LncRNA and EFs (GBDTL2E) has been proposed in this section. The GDDTL2E integrates the structural information and heterogenous networks, combines the Hetesim features and the diffusion features based on multi-feature fusion, and uses the machine learning algorithm GBDT to predict the association. This mainly includes several steps: (1) according to the lncRNA-EF correlations dataset downloaded from the public database DLREFD, after the duplicate data are removed, the set of lncRNAs and EFs and the association matrix A of the lncRNA-EF correlations are obtained, respectively. Then, the gaussian interaction profile kernel similarity of lncRNA (KL) and the gaussian interaction profile kernel similarity of EFs (KE) are calculated, respectively. (2) The chemical structure similarity matrix E between EFs is calculated by using the published tool SimComp. (3) The lncRNA similar information (KL) is transformed by the logistic function to obtain lncRNA similarity information SL, and the chemical structure similarity matrix E and the gaussian interaction profile kernel similarity matrix (KE) are then used to construct a similarity matrix SE of EFs. (4) A global heterogeneous network is constructed by integrating the three subnets of association matrix A, similarity matrix SL of lncRNA, and similarity matrix SE of EFs to construct adjacency matrix G of the global heterogeneous network. On the heterogeneous network, the Random Walk with Restart (RWR) algorithm is used to calculate the diffusion score and obtain the diffusion features, and singular value decomposition (SVD) is used to reduce the dimension of the diffusion features. (5) The Hetesim feature (score) for the lncRNAs-EFs pair is calculated. (6) The feature data set is obtained by combining the diffusion feature and the HeteSim score. The obtained combined feature is used to train the Gradient Boosting Decision Tree (GBDT) for predicting the relationship between lncRNAs and EFs. **Figure 1** shows that the overview of the proposed method. Each step of GBDTL2E are described in the following section.

2.1. Calculate Gaussian Interaction Profile Kernel Similarity

In this section, the calculation of the gaussian interaction profile kernel similarity was presented first. The association matrix A of lncRNAs and EFs was obtained by the known lncRNA-EF correlations. The gaussian interaction profile kernel similarity matrix of lncRNA and the gaussian interaction profile kernel similarity matrix of EF were calculated. Let $A(l_i, e_j)$ indicate whether the lncRNA l_i is associated with e_j. Specifically, $A(l_i, e_j) = 1$ if there is an association between l_i and e_j; otherwise, $A(l_i, e_j) = 0$, which is given by

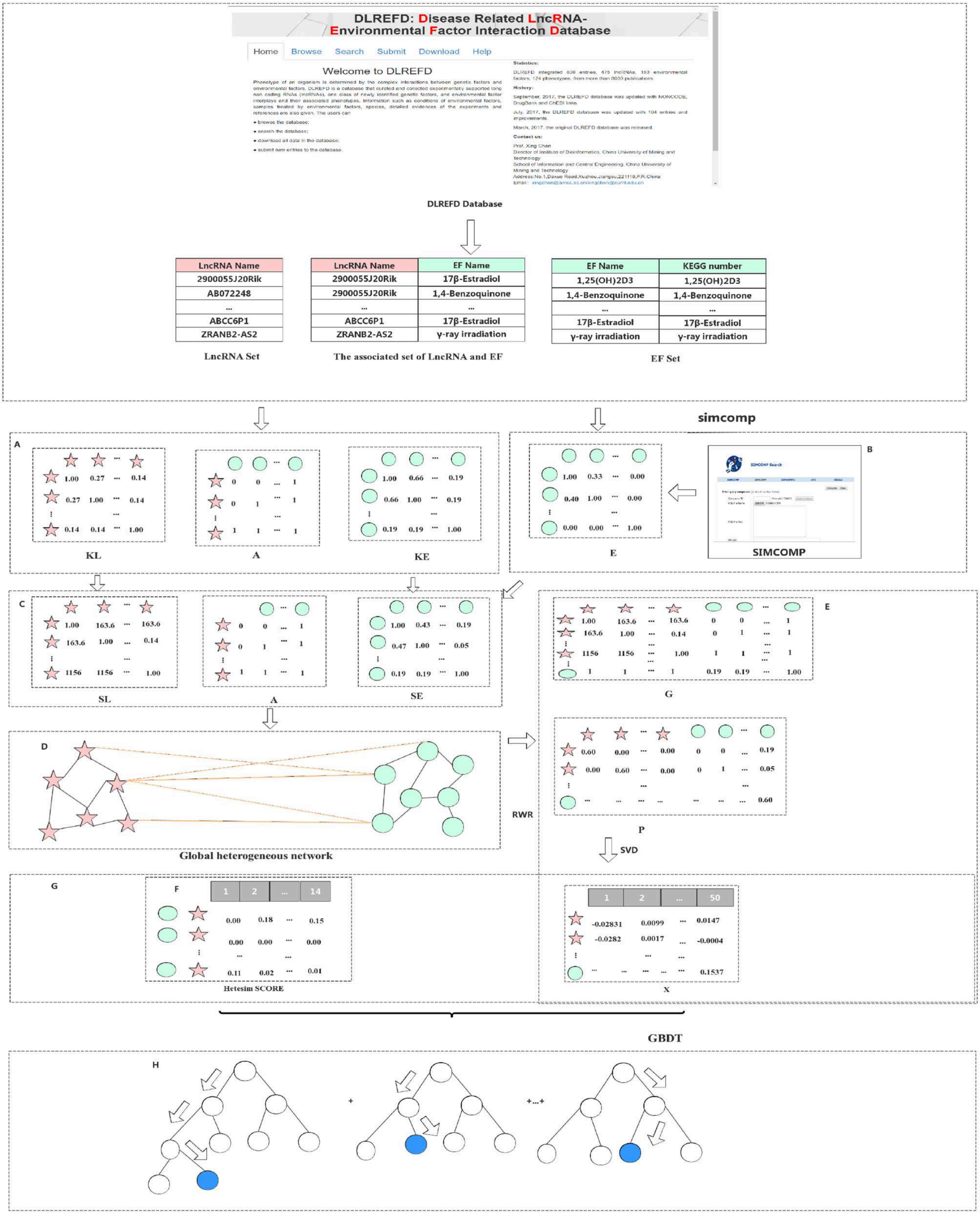
$$\mathrm{A}\left(l_i, e_j\right) = \begin{cases} 1 & l_i \text{ is associated with } e_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The gaussian interaction profile kernel similarity matrix KL of lncRNA was constructed. For a given lncRNA l_i, $IP(l_i)$ is defined as the i_{th} row of the adjacency matrix A. Then the gaussian interaction profile kernel similarity between lncRNA l_i and lncRNA l_j for each lncRNA pair is calculated, which can be written as

$$\mathrm{KL}\left(l_i, l_j\right) = \exp\left(-\gamma_l||\mathrm{IP}\left(l_i\right) - \mathrm{IP}\left(l_j\right)\|^2\right) \tag{2}$$

$$\gamma_l = \gamma_l'/\left(\frac{1}{nl}\sum_{i=1}^{nl}\| \mathrm{IP}\left(l_i\right)\right)\|^2\Bigg) \tag{3}$$

where γ_l is used to control the frequency band of Gaussian interaction profile kernel similarity. It represents the normalized frequency band of Gaussian interaction profile kernel similarity based on the new frequency band parameter γ_l'. Denote nl as the number of lncRNA. Denote KL as the gaussian interaction profile kernel similarity matrix of lncRNA, and denote KL $\left(l_i, l_j\right)$ as the gaussian interaction profile kernel similarity score of lncRNA l_i and lncRNA l_j.



DLREFD: Disease Related LncRNA-Environmental Factor Interaction Database

Home Browse Search Submit Download Help

Welcome to DLREFD

Phenotype of an organism is determined by the complex interactions between genetic factors and environmental factors. DLREFD is a database that curated and collected experimentally supported long non coding RNAs (lncRNAs), one class of newly identified genetic factors, and environmental factor interplays and their associated phenotypes. Information such as conditions of environmental factors, samples treated by environmental factors, species, detailed evidences of the experiments and references are also given. The users can

- browse the database;
- search the database;
- download all data in the database;
- submit new entries to the database.

Statistics:

DLREFD integrated 839 entries, 475 lncRNAs, 153 environmental factors, 124 phenotypes, from more than 8000 publications.

History:

September, 2017, the DLREFD database was updated with NONCODE, DrugBank and ChEBI links.

July, 2017, the DLREFD database was updated with 104 entries and improvements.

March, 2017, the original DLREFD database was released.

Contact us:

Prof. Xing Chen
Director of Institute of Bioinformatics, China University of Mining and Technology
School of Information and Control Engineering, China University of Mining and Technology
Address:No.1,Daxue Road,Xuzhou,Jiangsu,221116,P.R.China
Email: xingchen@amss.ac.cn/xingchen@cumt.edu.cn

DLREFD Database

LncRNA Name
2900055J20Rik
AB072248
...
ABCC6P1
ZRANB2-AS2

LncRNA Set

LncRNA Name	EF Name
2900055J20Rik	17β-Estradiol
2900055J20Rik	1,4-Benzoquinone
...	...
ABCC6P1	17β-Estradiol
ZRANB2-AS2	γ-ray irradiation

The associated set of LncRNA and EF

EF Name	KEGG number
1,25(OH)2D3	1,25(OH)2D3
1,4-Benzoquinone	1,4-Benzoquinone
...	...
17β-Estradiol	17β-Estradiol
γ-ray irradiation	γ-ray irradiation

EF Set

FIGURE 1 | Flowchart of our method: **(A)** Obtained the association matrix A; Calculated the gaussian interaction profile kernel similarity of lncRNA and EF respectively. **(B)** Calculated the chemical structure similarity matrix E. **(C)** Obtained lncRNA similarity information SL and construct a similarity matrix SE of EF. **(D)** Integrated three subnets A, SL, and SE to construct a global heterogeneous network. **(E)** Constructed the adjacency matrix G and obtain the diffusion feature. **(F)** Calculated the Hetesim score. **(G)** Combined the diffusion feature and the HeteSim score. **(H)** Trained the Gradient Boosting Decision Tree classifier (GBDT).

Similarly, the known lncRNA-EF correlations were used to construct the gaussian interaction profile kernel similarity matrix of EFs. For a given EF e_i, $IP'(e_i)$ is defined as the i_{th} column of the adjacency matrix A. KE represents the gaussian interaction profile kernel similarity matrix of environmental factors. Denote $\mathrm{KE}\left(e_i, e_j\right)$ as the gaussian interaction profile kernel similarity score of EFs e_i and e_j, which is given by

$$\mathrm{KE}\left(e_i, e_j\right) = \exp\left(-\gamma_e||\mathrm{IP}'\left(e_i\right) - \mathrm{IP}'\left(e_j\right)\|^2\right) \tag{4}$$

$$\gamma_e = \gamma_e'/\left(\frac{1}{ne}\sum_{i=1}^{ne}\|\mathrm{IP}'\left(e_i\right)\right)\|^2\Bigg) \tag{5}$$

where γ_e represents normalized gaussian interaction kernel similarity bandwidth based on the frequency width parameter γ_e'. Denote ne as the number of EFs.

2.2. Calculate Chemical Structure Similarity

In this section, the computation of the chemical structure similarity has been given. The chemical structural similarity matrix between EFs is calculated using the SimComp tool (Hattori et al., 2010). With the Kyoto Encyclopedia of Genes and Genomes (KEGG) database entry number corresponding to EFs in the DLREFD database as the parameter, the SimComp tool is used to calculate the chemical structure similarity score. By calling SimComp's API, the chemical structure similarity score $\mathrm{E}\left(e_i, e_j\right)$ of each pair of environmental factors e_i and e_j was calculated. SimComp (Similar Compound) is a kind of method based on a graph that is used to compare the chemical structure. It has been implemented in a KEGG system to search for similar chemical structures in a chemical structure database.

2.3. Obtain the Similarity Matrix

The structural information and heterogenous networks were integrated in the proposed GBDTL2E. The transformed similarity matrix SL and integrated similarity matrix calculation SE have been described in this section. The lncRNA similarity matrix KL was transformed by logistic function to obtain lncRNA similar matrix SL. The similarity matrix SE of EFs was constructed by using the chemical structure similarity matrix E of EFs and the gaussian interaction profile kernel similarity matrix KE of EFs, given by

$$\mathrm{SL}\left(l_i, l_j\right) = \frac{1}{1 + e^{c \cdot KL\left(l_i, l_j\right)} + v} \tag{6}$$

where $c = -15, v = \log(9999)$;

$$\mathrm{SE}\left(e_i, e_j\right) = \begin{cases} ew \cdot E\left(e_i, e_j\right) + (1 - ew) \cdot KE\left(e_i, e_j\right) & \mathrm{E}\left(e_i, e_j\right) \neq 0 \\ KE\left(e_i, e_j\right) & otherwise \end{cases} \tag{7}$$

where ew is the weight of correlation information of two EFs in SE.

2.4. Obtain Low-Dimensional Network Diffusion Features

In this section, the association matrix A of lncRNA-EF, the similarity matrix SL of lncRNA, and the similarity matrix SE of EFs were integrated to construct a global heterogeneous network. In heterogeneous networks, the Random Walk with Restart (RWR) is used to calculate the diffusion score and obtain the diffusion features. Due to the fact that the higher-dimensional features in model training are more susceptible to noise interference, the singular value decomposition (SVD) is used to reduce the dimension of the diffusion features. The details of each sub-steps were as follows.

2.4.1. Construct of Roaming Network

In this section, the roaming network was constructed firstly. The adjacency matrix G of the global heterogeneous network was obtained. The matrix G has $nl + ne$ dimensions, where nl is the number of lncRNA and ne is the number of EFs, respectively. G is given by

$$\mathrm{G} = \begin{bmatrix} SL & A \\ A^T & SE \end{bmatrix} \tag{8}$$

where A^{T} represents the transpose of A, and SL and SE are given by (6) and (7), respectively. T is the transition probability matrix of G, which is given by

$$\mathrm{T}(i,j) = \frac{G(i,j)}{\sum_{k=1}^{nl+ne} G(k,j)} \tag{9}$$

where $T(i,j)$ represents the probability of node i transferring to node j in the global network. For any two given nodes i and j in the wandering network, if $T(i,j)$ is not 0, there is an edge between them. If $T(i,j)$ is 0, and node i has no relationship with node j.

2.4.2. Obtain the Diffusion Features Using RWR

The RWR algorithm (Liu et al., 2016) is used to obtain the diffusion features of each node on the global network in this section. Based on the transition probability matrix T, the diffusion features of all nodes $\mathrm{P} = \left[P^i\right]$ were obtained by RWR, where $i \in \{1, 2, \ldots n\}$. P^i represents the diffusion features of node i, $n = nl + ne$, and $nl + ne$ is the total number of nodes in the global heterogeneous network. Starting from a node i in the global heterogeneous networks, each step prompted two choices: randomly select the neighboring node or return the starting node. The process of restarting the random walk is given by

$$P_{t+1}^i = (1 - r) * T * P_t^i + r * P_0^i \tag{10}$$

where r is the restart probability; P_t^i is an n-dimensional probability distribution vector of node i, and its j_{th} element represents the probability of accessing node j at step t, and $j \in \{1, 2, \ldots, n\}$. P_0^i represents the initial transition probability, which is given by

$$P_0^i = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n} \ldots \frac{1}{n}\right) \tag{11}$$

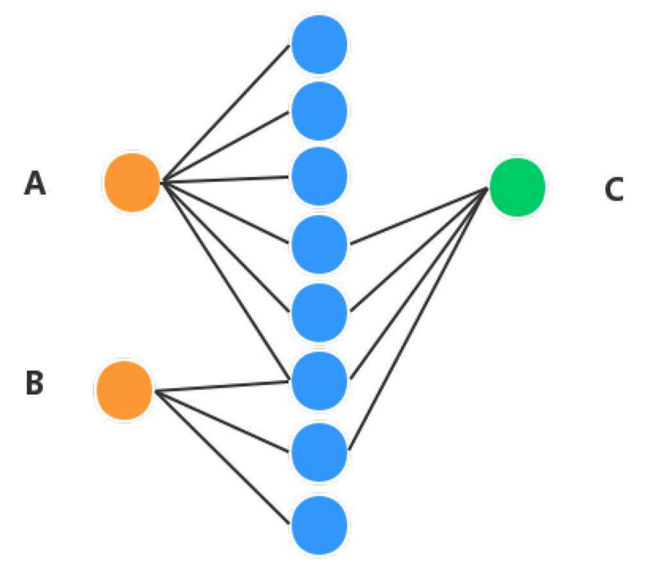


FIGURE 2 | Example of understanding HeteSim masure. Different color circles denote three different kinds of objects in the heterogeneous network. **(A–C)** represent three different nodes in the heterogeneous network.

The initial assumption is that the transition probability value of each node is $1/n$, and n is the total number of nodes. After several iterations, when $(P_{t+1} - P_t)$ is less than 10^{-10}, the final diffusion features were obtained.

2.4.3. Calculate Low-Dimensional Diffusion Features

The calculation of low-dimensional diffusion features has been given in this section following the diffusion features obtained by RWR. As the number of nodes increases, the diffusion state increases in dimension as well. Singular value decomposition (SVD) (Golub and Reinsch, 1971; Cho et al., 2015) is used to reduce the dimension of diffusion features. The high-dimensional diffusion feature matrix is decomposed:

$$\mathrm{P} = U\Sigma V^T \tag{12}$$

$$\mathrm{P} = U\Sigma^{1/2}\Sigma^{1/2}V^T \tag{13}$$

where U and V represent the left singular matrix and the right singular matrix, respectively. The U and V are units on an orthogonal matrix, Σ only has value on the diagonal, and the other elements are 0. We refer to these non-zero values as singular values and order these values in Σ from largest to smallest. Singular values can be thought of as representing values of a matrix, or as representing information about the matrix. The larger the singular value, the more information it represents. Therefore, in order to reduce the computation, we only need to take the first 50 maximum singular values, and we can basically restore the data itself. Therefore, we take the first 50 singular values and eigenvectors, which are given by

$$X = U_{n*d}\,(\Sigma_{d*d})^{1/2} \tag{14}$$

TABLE 1 | The paths from a lncRNA to an environmental factor in our heterogeneous network with a length of less than 5.

Id	Path	Meaning	Length
1	LLE	lncRNA-lncRNA-EF	2
2	LEE	lncRNA-EF-EF	2
3	LLLE	lncRNA-lncRNA-lncRNA-EF	3
4	LELE	lncRNA-EF-lncRNA-EF	3
5	LLEE	lncRNA-lncRNA-EF-EF	3
6	LEEE	lncRNA- EF-EF-EF	3
7	LLLLE	lncRNA-lncRNA-lncRNA-lncRNA-EF	4
8	LLLEE	lncRNA-lncRNA-lncRNA-EF-EF	4
9	LLELE	lncRNA-lncRNA-EF-lncRNA-EF	4
10	LLEEE	lncRNA-lncRNA-EF-EF-EF	4
11	LELLE	lncRNA-EF-lncRNA-lncRNA-EF	4
12	LELEE	lncRNA-EF-lncRNA- EF-EF	4
13	LEELE	lncRNA-EF-EF-lncRNA-EF	4
14	LEEEE	lncRNA-EF-EF-EF-EF	4

$$W = (\Sigma_{d*d})^{1/2}\,(V_{d*n})^T \tag{15}$$

where X is the low-dimensional node feature matrix derived from the high-dimensional diffusion feature. Each row of matrix X is the low-dimensional feature vector of each node in the network. W is the low-dimensional context eigenmatrix derived from the high-dimensional diffusion feature. Thus, we obtain the diffusion feature X after dimensionality reduction.

2.5. Calculate the Hetesim Score

In order to obtain high performance, apart from the diffusion feature obtained in the above section, the proposed method combines the Hetesim features and the diffusion features based on multi-feature fusion. Another important feature is that HeteSim (Shi et al., 2014) is used to calculate the relevance between objects in the heterogeneous network in this section. HeteSim is a path-based measure. For each pair object (of the same or different types) in the heterogeneous network, it could obtain one single score, which means their relatedness based on an arbitrary path. **Figure 2** illustrates a HeteSim score.

As we can see from **Figure 2**, the number of paths from A to C is three and the number of paths from B to C is two. The number of paths from A to C is larger than B to C, which might mean that A is closer to C than B. But, based on HeteSim, B is closer to C than A to C because there are two edges for B to C, which account for two-thirds of the edges starting from B to other objects. However, A only has a small part of the edges connected with C. In our proposed method, the HeteSim is used to measure the similarities between lncRNAs and EFs. Under the constraint of length less than five, there are 14 different paths from lncRNA to the EFs, as shown in **Table 1**.

The HeteSim score between lncRNA and EF is calculated:

Step (1): The transition probability matrix M_{LP} from lncRNA to EF, lncRNA to lncRNA, EF to lncRNA, and EF to EF in global heterogeneous networks are calculated.

The calculation formula of transfer probability matrix $M_{LP}(i,j)$ is given by

$$M_{LP}(i,j) = \frac{I_{LP}(i,j)}{\sum_{k=1} I_{LP}(i,k)} \tag{16}$$

where L and P represent two types of objects in the global heterogeneous network, and i and j represent two nodes in the global heterogeneous network. Matrix I is the incidence matrix of L and P. If both L and P are environmental factors, matrix I is matrix SE. If both L and P are lncRNAs, matrix I is matrix SL. If L and P are lncRNA and EFs respectively, then matrix I is matrix A. The four transfer probability matrices can be obtained as M_{LE}, M_{LL}, M_{EL}, and M_{EE} respectively.

Step (2): The $path = (h_1, h_2, \cdots, h_{m+1})$ is divided into two parts. When the path length m is even, divide the path into $path_L = \left(h_1, h_2, \cdots, h_{\text{mid}}\right)$ and $path_R = \left(h_{mid}, h_2, \cdots h_{m+1}\right)$, $mid = (m/2) + 1$; Otherwise, when the length of path m is odd, we need to take $mid = ((m+1)/2)$ and $mid = ((m+3)/2)$, respectively. Then, we can get different HeteSim scores when taking the two mid, and the final score is the average of the two HeteSim scores.

Step (3): The reachable probability matrix R_{path} under $path_L$ and $path_R$ is calculated. The reachable probability matrix R_{path_L} and R_{path_L} are given by

$$R_{path_L} = M_{h_1,h_2}, M_{h_2,h_3} \cdots M_{h_{mid-1},h_{mid}} \tag{17}$$

$$R_{path_R} = M_{h_{mid},h_{mid+1}}, M_{h_{mid+2},h_{,id+3}} \cdots M_{h_{m-1},h_m} \tag{18}$$

Step (4): The HeteSim score of path $path$ is calculated, which is given by:

$$\text{Hetesim} = \frac{R_{path_L}\left(R_{path_R^{-1}}\right)^T}{\left\|R_{path_L}\right\|_2 * \left\|R_{path_R^{-1}}\right\|_2} \tag{19}$$

where $path_R^{-1}$ is the reverse path of $path_R$. There are in total 14 different paths from a lncRNA to an EF under the constraint of length <5. So, we obtain 14-dimensional HeteSim features for each node in the heterogeneous networks.

2.6. Train the Gradient-Boosting Decision Tree Classifier

After the multi-features were combined, the Hetesim features and the diffusion features were obtained. The method for training the GBDT classifier model to predict the association between lncRNAs and EFs based on heterogeneous networks has been presented in this section. The 50-dimensional diffusion features and 14-dimensional HeteSim scores were combined to get the 64-dimensional features data set. The features of the data were used for training the Gradient Boosting Decision Tree (GBDT) (Friedman, 2001) classifier. The classifier was used to predict the correlation between lncRNAs and EFs.

GBDT is an effective machine learning method for classification and regression problems. GBDT is composed of multiple decision trees, and the final answer is obtained via the sum of the conclusion of all trees. GBDT generates a weak classifier in each iteration through multiple rounds of iteration. Each classifier is trained on the basis of the gradient (residual value) of the previous round of classifiers. The final total classifier is obtained by weighted summation of the weak classifier obtained in each round of training, which is the addition model. The model training steps have been presented:

Step (1): The initialization model is given by:

$$\Theta_0(x) = \frac{1}{2} * \log\left(\frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} 1 - y_i}\right) \tag{20}$$

where N is the number of training samples, and y_i is the real label. The loss function is given by:

$$\mathrm{L}\left(\mathrm{y}, \Theta_{m-1}\left(x_i\right)\right) = \log\left(1 + \exp\left(-\mathrm{y}\Theta_{m-1}\left(x_i\right)\right)\right) \tag{21}$$

where y is the real class label, and $\Theta_m\left(x\right)$ is the weak model in the m_{th} round.

Step (2): Cycle m in turn, where m = 1,2,...M

A: The calculation for the negative gradient of the loss function of the i_{th} sample in the m_{th} round is given by:

$$r_{m,i} = -\frac{\partial \mathrm{L}\left(\mathrm{y}_i, \Theta_{m-1}\left(x_i\right)\right)}{\partial \Theta_{m-1}\left(x_i\right)} = \frac{\mathrm{y}_i}{\left(1 + \exp\left(\mathrm{y}_i\right)\Theta\left(x_i\right)\right)} \tag{22}$$

where $i = 1, 2, \ldots N$.

B: Construct the m_{th} decision tree, and then get the corresponding leaf node area $R_{m,j}$, $wherej = 1, 2, ..., J$, and the J is the number of leaf nodes in the tree.

C: For the samples in each leaf node, we calculated the $c_{m,j}$, which minimizes the loss function, namely, the best output value of fitting the leaf node, given by:

$$c_{m,j} = \arg\min_{c} \sum_{x \in R_{m,j}} \log\left(1 + \exp\left(-\mathrm{y}_i\Theta\left(x_i\right) + \mathrm{c}\right)\right) \tag{23}$$

D: Update m_{th} weak model:

$$\Theta_m(x) = \Theta_{m-1}(x) + lr * \sum_{j=1}^{J} c_{m,j} I\left(x \in R_{m,j}\right) \tag{24}$$

where $I\left(x \in R_{m,j}\right)$ means that if x falls on a leaf node corresponding to $R_{m,j}$, then the corresponding term is 1, and lr means learning rate.

TABLE 2 | The experimental parameters of GBDTL2E.

Notation	Value	Definition
nl	475	The number of lncRNAs
ne	152	The number of EFs
n	627	The sum number of EFs and lncRNAs
γ_l'	1	The frequency band of gaussian interaction profile kernel similarity of lncRNA
γ_e'	1	The frequency band of gaussian interaction profile kernel similarity of EF
ew	0.7	The weight parameter of correlation information of two environmental factors in SE
m	5	The length constraint in Hetesim
d	50	The dimension of the low-dimensional diffusion features
r	0.5	The restart probability in the random walk with restart
N	600	The number of training samples
M	10	The number of training iterations

E: Judge whether m is greater than M. If m is less than M, then m=m+1 and jump to Step(1) for the next iterations. Otherwise, it means that m weak learners have been constructed, and we then jump to Step(3) to end the training.

Step (3): Obtain the final Strong Model:

$$\Theta(x) = \Theta_0(x) + lr * \sum_{m=1}^{M} \sum_{j=1}^{J} c_{m,j} I\left(x \in R_{m,j}\right) \tag{25}$$

2.7. GBDTL2E Algorithm

In this section, the proposed GBDTL2E algorithm to predict the association between lncRNAs and EFs based on heterogeneous networks has been described in Algorithm 1. From lines four to nine of Algorithm 1, the low-dimensional diffusion feature matrix X was obtained by using the random walk with restart algorithm and singular value decomposition. In lines 10–41 of Algorithm 1, the Hetesim score was obtained. In lines 42–58 of Algorithm 1, the training data is obtained and used to train the GBDT classifier. Furthermore, the final classification model is obtained.

3. RESULT AND DISCUSSION

3.1. Data Sets

We randomly selected 300 positive samples and 300 negative samples for training the model. Positive samples were that samples with a correlation between lncRNA and EF, while negative samples were samples without a correlation between lncRNA and EF. For objective performance evaluation, an independent test set was built by randomly selecting 300 positive samples and 300 negative samples. Note that all the positive and negative samples in these test sets were independently chosen and excluded from the training set.

Algorithm 1 GBDTL2E algorithm

Input: lncRNAs set, EFs set, The association matrix of the lncRNA-EFs A;

Output: The gaussian interaction profile kernel similarity matrices KL and KE. The chemical structural similarity matrix, E. The similarity matrices SL and SE.

```
1:  Construct the adjacency matrix $G$;
2:  Initialize the global transition probability matrix $T$;
3:  Initialize the transition probability vector for each node $P_0^i = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n} \cdots \frac{1}{n}\right)$
4:  while $P_{t+1}^i - P_t^i > 10^{-10}$ do:
5:      Obtain the updated probability vector:
6:      $P_{t+1}^i = (1-r) * T * P_t^i + r * P_0^i$;
7:  end while
8:  $\mathrm{P} = U_{n*d} \Sigma_{d*d} V_{d*n}^T$
9:  $X = U_{n*d} \Sigma_{d*d}^{1/2}$
10: Input L,P to caculate $M_{LP}(i,j)$
11: if $L \in EFs$ and $P \in EFs$ then
12:     $M_{LP}(i,j)$= $M_{EE}(i,j) = \frac{SE_{EE}(i,j)}{\sum_{k=1} SE_{EE}(i,k)}$
13: end if
14: if $L \in lncRNAs$ and $P \in EFs$ then
15:     $M_{LP}(i,j)$= $M_{LE}(i,j) = \frac{A_{LE}(i,j)}{\sum_{k=1} A_{LE}(i,k)}$
16: end if
17: if $L \in EFs$ and $P \in lncRNAs$ then
18:     $M_{LP}(i,j)$= $M_{EL}(i,j) = \frac{A_{EL}^T(i,j)}{\sum_{k=1} A_{EL}^T(i,k)}$
19: end if
20: if $L \in lncRNAs$ and $P \in lncRNAs$ then
21:     $M_{LP}(i,j)$= $M_{LL}(i,j) = \frac{SL_{LL}(i,j)}{\sum_{k=1} SL_{LL}(i,k)}$
22: end if
23: for $n = 1 \rightarrow 5$ do
24:     Divide the path into two parts.
25:     if $n\%2 == 0$ then
26:         $mid = (m/2) + 1$
27:         $path_L = \left(h_1, h_2, \cdots, h_{\mathrm{mid}}\right)$
28:         $path_R = \left(h_{mid}, h_2, \cdots h_{m+1}\right)$
29:     end if
30:     if $n\%2! = 0$ then
31:         $mid1 = ((m+1)/2)$
32:         $mid2 = ((m+3)/2)$
33:         $path_{L_1} = \left(h_1, h_2, \cdots, h_{\mathrm{mid1}}\right)$
34:         $path_{R_1} = \left(h_{mid1+1}, h_2, \cdots h_{m+1}\right)$
35:         $path_{L_2} = \left(h_1, \cdots, h_{\mathrm{mid2}}\right)$
36:         $path_{R_2} = \left(h_{mid2+1}, \cdots h_{m+1}\right)$
37:     end if
38:     $R_{path_L} = M_{h_1,h_2}, M_{h_2,h_3} \cdots M_{h_{mid-1},h_{mid}}$
39:     $R_{path_L} = M_{h_1,h_2}, M_{h_2,h_3} \cdots M_{h_{mid-1},h_{mid}}$
40:     Hetesim $= \frac{R_{path_L}\left(R_{path_R^{-1}}\right)^T}{\left\|R_{path_L}\right\|_2 * \left\|R_{path_R^{-1}}\right\|_2}$
41: end for
42: Combined with the diffusion feature and HeteSim score to get the data set
43: $\mathrm{D_{train}} = \left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_N, y_N\right)\right\}$, $\mathrm{D_{test}} = \left\{\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_N, y_N\right)\right\}$
```

```
Use $D_{train}$ to train the Gradient Boosting Decision Tree (GBDT).
Initialize the model as $\Theta_0(x)$
for $m = 1 \to M$ do
    for $i = 1 \to N$ do
        Calculate loss function: $L\left(y, \Theta_{m-1}(x_i)\right)$
        Calculate the residuals: $r_{m,i}$
    end for
    Construct the $m_{th}$ decision tree,
    Get the corresponding leaf node area $R_{m,j}, j = 1, 2, ..., J$
    for $J = 1 \to J$ do
        Calculate $c_{m,j}$
    end for
    Update weak model: $\Theta_m(x)$
end for
Get the strong model $\Theta_M(x)$
```

3.2. Performance Measures

The 10-fold cross-validation was used to measure the performance of the GBDTL2E. The GBDTL2E parameters used are listed in **Table 2**. The detailed process of 10-fold cross-validation has been described as: the training set was randomly divided into 10 groups of roughly the same size subsets. Each subset was used for validation data in turn, and the remaining nine subsets were used for training data. This process was repeated 10 times, and performance assessments were performed using average performance measures of more than 10 times. The experiment used a variety of methods to evaluate performance, including recall (REC), F1-score, accuracy (ACC), Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curves (AUC). They were defined:

$$Recall = \frac{TP}{TP + FN}, \tag{26}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{27}$$

$$F1 - Score = \frac{2 \times TP}{2TP + FP + FN}, \tag{28}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{29}$$

where *TP* and *FP* represent the numbers of correctly predicted positive and negative samples, and *FP* and *FN* represent the numbers of wrong predicted positive and negative samples, respectively. The AUC score is computed by varying the cutoff of the predicted scores from the smallest to the greatest value.

3.3. Performance Comparison With Existing Machine Learning Methods

In this section, the proposed GBDTL2E method was compared with the following schemes, which include the k-nearest neighbor algorithm (KNN) (Cover and Hart, 1967), random forest (RF) (Liaw et al., 2002), and support vector machine (SVM) (Burges, 1998). The 10-fold cross-validation was used by the four algorithms. For the KNN classifier, five nearest neighbors were used. The RF algorithm constructed multiple decision tree classifiers for training on a set of randomly selected benchmark samples to improve performance. For the SVM, we used the radial basis function (RBF) as the kernel function to optimize the penalty c and γ parameters. In addition, we set c and γ as 64 and 0.0001, respectively. **Table 3** and **Figure 3** show the predictive performance comparison of the machine learning approach used with other machine learning approaches. It can be seen that the method used in the present invention had the best performance. In order to further prove the performance of this model, we also compared the performances of these different machine learning methods on the independent test set. The ROC curve compared on the independent test set is shown in **Figure 4**. The AUC of GBDTL2E, KNN, RF, and SVM were 0.91, 0.82, 0.88, and 0.88, respectively. The results show that the performance using GBDT was better than that of other machine learning methods.

TABLE 3 | The performance comparison with other machine learning methods.

Method	ACC	RECALL	F1-score	MCC	AUC
KNN	0.953	0.937	0.952	0.907	0.985
RF	0.863	0.827	0.849	0.739	0.912
SVM	0.966	0.967	0.966	0.933	0.988
GBDTL2E	0.975	0.967	0.976	0.949	0.997

3.4. Performance Comparison With Different Topological Features

In order to verify the performance of combined diffusion and Hetesim features in GBDTL2E, we compared the performance by using two separate features and combined features in this section. **Figures 5**, **6** show the Performance comparison with different topological features, In the **Figure 5**, we denote the "Hete+Diff," "Hete," and "Diff" as the Hetesim and diffusion combined feature, HeteSim feature, and diffusion feature, respectively. As we can see from **Figure 5**, the Hetesim and diffusion combined features achieved higher performance than the two separate features. The results show that the combination of the two features can improve the prediction performance. **Figure 6** shows the ROC curve comparison with different feature groups, which is the method using GBDTL2E only with diffusion feature, using GBDTL2E only with HeteSim feature, and GBDTL2E with combined feature. We also used 10-fold cross validation to verify the influence of different feature groups on the experimental results. We can see, from **Figure 6**, that GBDTL2E with combined features can obtain higher performances than other two algorithms. The GBDTL2E with the Hetesim feature only could obtain a better performance than the GBDTL2E with the diffusion feature only.

3.5. Performance Comparison With Existing Methods

In this section, the GBDTL2E algorithm was compared with the existing methods for predicting associations between lncRNAs and EFs. However, there were a few studies that predicted new

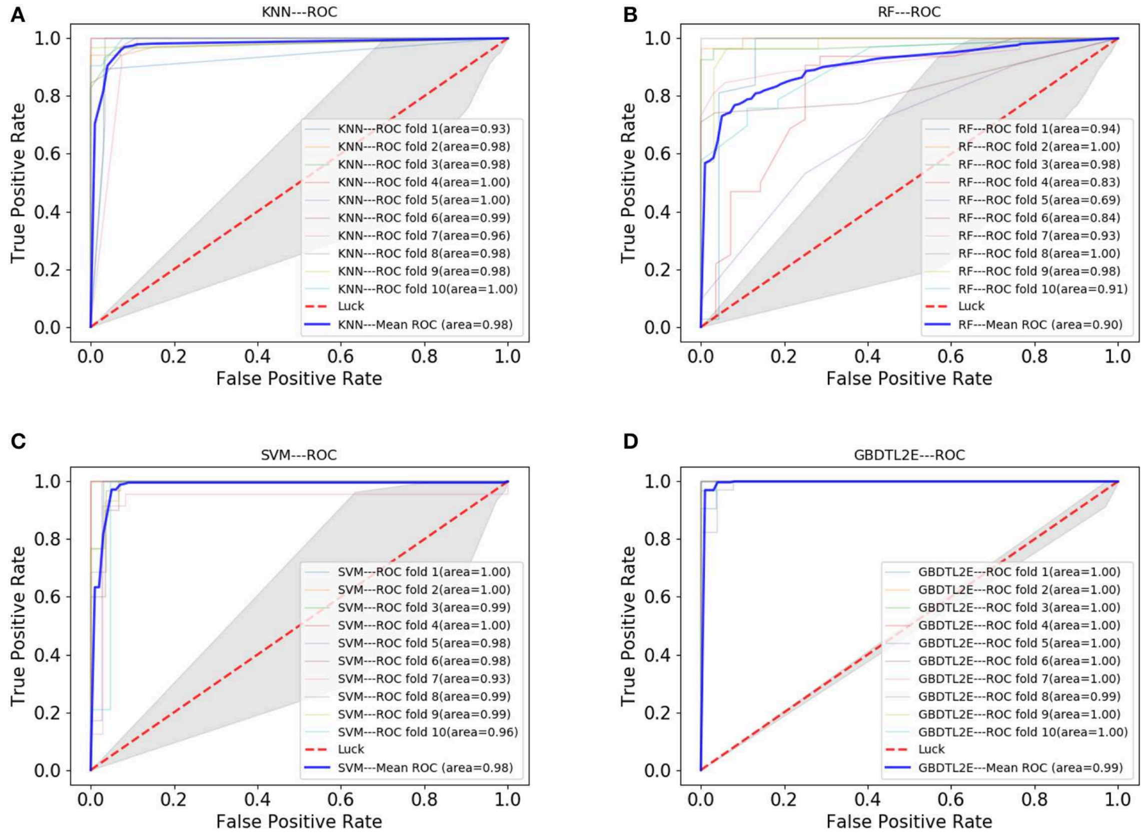


FIGURE 3 | The ROC curve comparison with other machine learning methods. **(A)** The ROC curve with using KNN. **(B)** The ROC curve with using RF. **(C)** The ROC curve with using SVM. **(D)** The ROC curve with using GBDT.

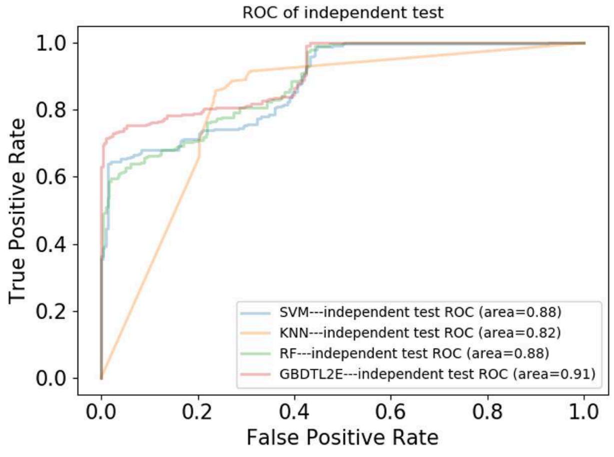


FIGURE 4 | The ROC curves comparison with other machine learning methods on independent dataset.

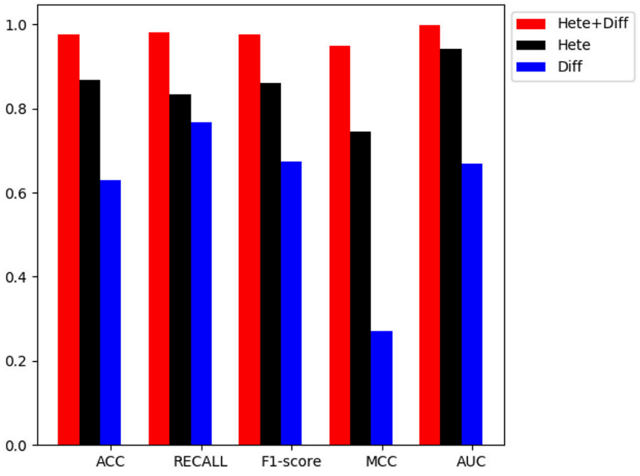


FIGURE 5 | The performance comparison of different feature groups (Diffusion, HeteSim and combined feature).

potential associations between lncRNAs and EFs. Three methods were chosen to compare with the proposed GBDTL2E method. These were KATZ (Vural and Kaya, 2018), MPALERLS (Xu, 2018), and BIRWAPALE (Xu, 2018).

- *KATZ*: The KATZ method, based on the KATZ, was used to find potential new associations between lncRNAs and EFs; it uses the DLREFD database as well and contains proven associations between lncRNAs and EFs. The KATZ and Gaussian interaction profile kernel similarity was used to predict new potential associations between lncRNAs and EFs. In this method, the parameters β and k are to 0.01 and 3, respectively.

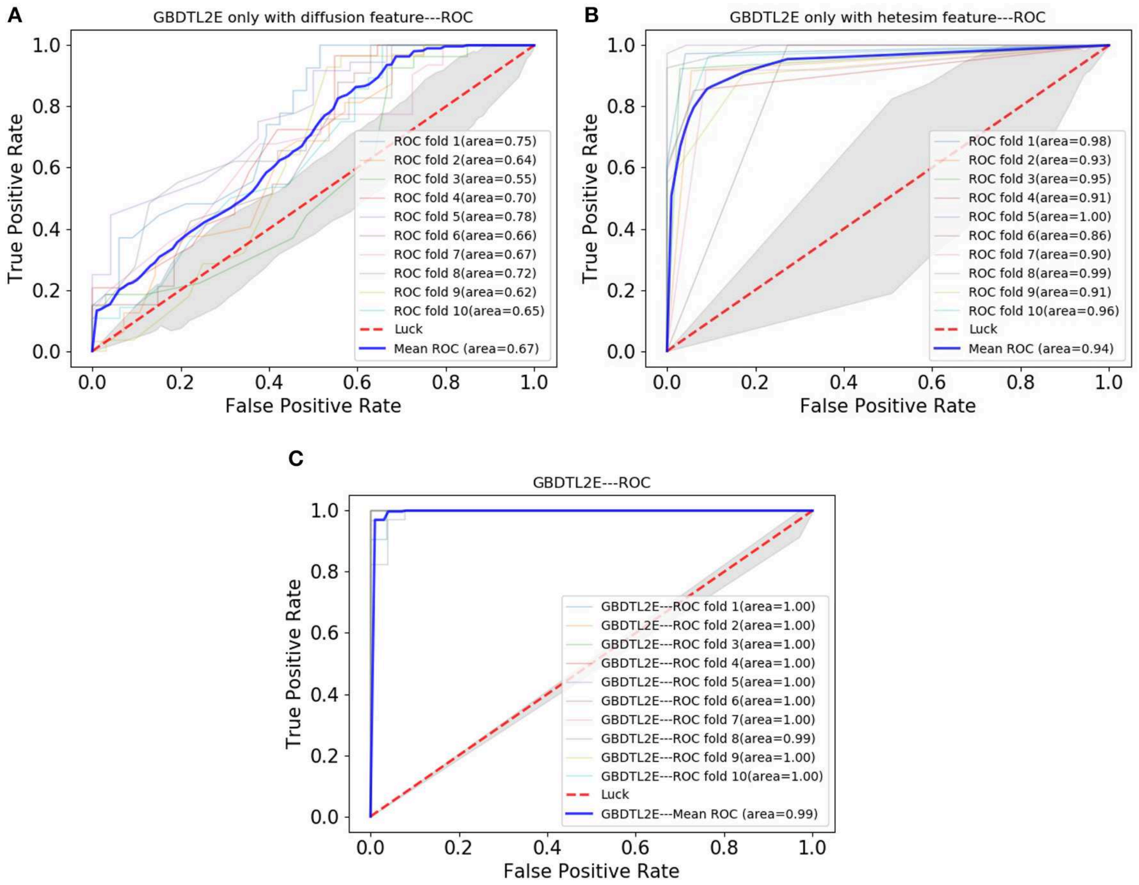


FIGURE 6 | The ROC curve comparison with different feature groups. **(A)** The ROC curve only with diffusion feature. **(B)** The ROC curve only with HeteSim feature. **(C)** The ROC curve with combined feature.

- *MPALERLS*: The MPALERLS method used the Laplace operator for regularization, built the cost function and minimized it, and finally obtained the optimal classifier of lncRNAs space and EFs space. Finally, the two optimal classifiers were transformed into a unified classifier to calculate the probability matrix of lncRNA-EFs association relation. They used the classifier to calculate the probability of lncRNA-EFs association relation and to rank the lncRNA-EF association according to the probability score. We set the weight of lncRNAs classifier and EFs classifier to 0.4 and 3, respectively.
- *BIRWAPALE*: The BIRWAPALE method is a double random walk algorithm on heterogeneous networks. Finally, the double random walk converged in the heterogeneous network, and the probability score of lncRNAs and EFs association relationship could be obtained. The parameters α, l, and r are set to 1, 2, and 3.

Figure 7 shows the comparison results. The experimental results show that the GBDTL2E algorithm can obtain a better performance than the other three algorithms. This was for several reasons: (1) Computing the HeteSim score of different paths from lncRNA to EFs in the heterogenous network to obtain the HeteSim features, and combining the HeteSim features and diffusion features as the data feature, could make better use of the topological characteristics of heterogeneous networks and thus obtain better performance. (2) The GBDT algorithm is an effective prediction model. As far as we know, we have been the first to apply both diffusion and HeteSim features to predict lncRNA-EFs interactions. As result show that, combine the diffusion and HeteSim features can further improve the performance.

3.6. Case Study

To further measure the performance of our proposed algorithm, we investigated an environmental factor "Cisplatin," which is an effective chemotherapy drug for many cancers (Florea and Büsselberg, 2011). The proven associations between "Cisplatin" and many lncRNAs have been discovered. In this study, we attempted to use our model to predict the association between "Cisplatin" and lncRNA. First, all associations between "Cisplatin" and lncRNA were deleted from the training set.

After processed by our algorithm, we sorted the correlation values between "Cisplatin" and ordered LncRNA from largest to smallest. We found that all the top 10 lncRNAs were related to "Cisplatin," and these lncRNAs are confirmed to be

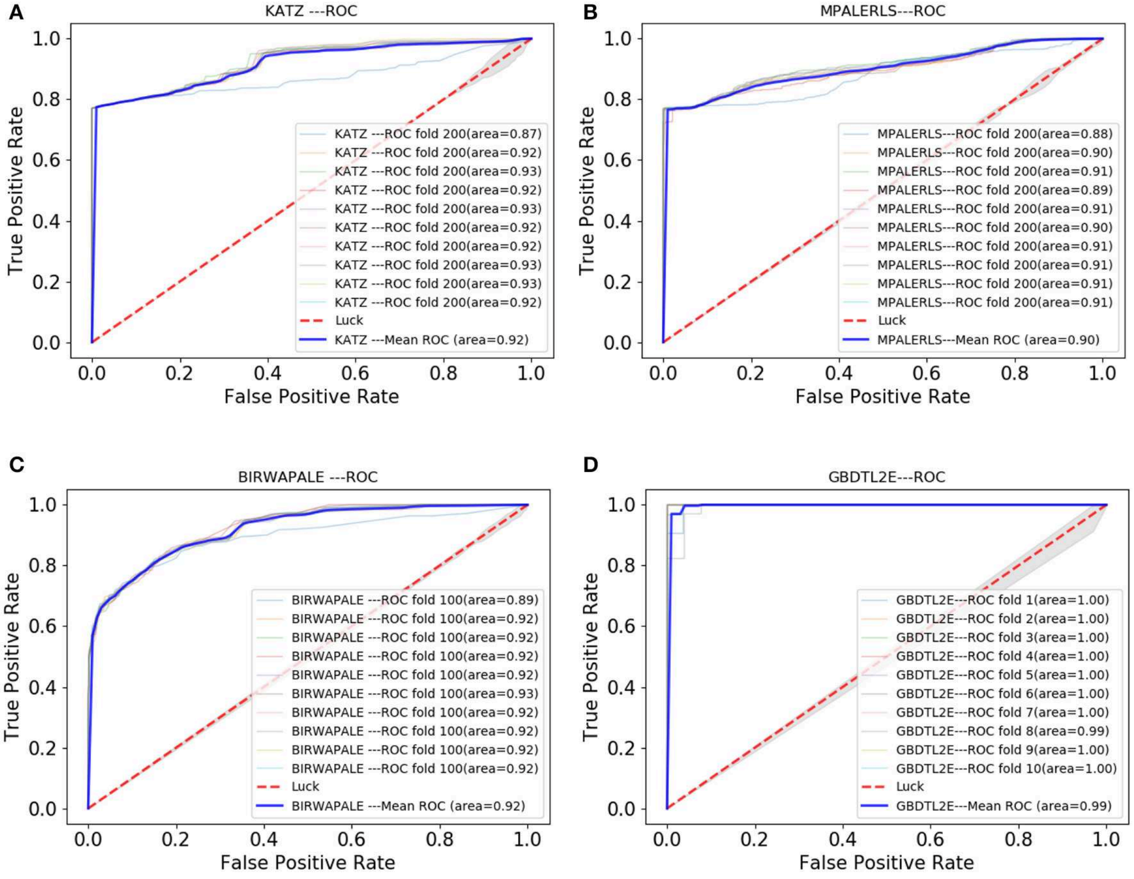


FIGURE 7 | The Roc curve comparison with existing method. **(A)** The ROC curve only of KATZ. **(B)** The ROC curve only of MPALERLS. **(C)** The ROC curve of BIRWAPALE. **(D)** The ROC curve of GBDTL2E.

TABLE 4 | The TOP 10 predicted lncRNAs related to cisplatin.

Number	LncRNA name	PubMedID
1	AK12669	23741487
2	AC015818.3	25250788
3	ABCC6P1	25250788
4	GABPB-AS1	24036268
5	CASC2	28495512
6	PSORS1C3	25250788
7	H19	28189050
8	AK125699	25250788
9	SRGAP3-AS2	25250788
10	XLOC_001406	25250788

related to “Cisplatin” in the DLREFD database. The 10 lncRNAs and their corresponding PUBMED reference ID are shown in **Table 4**.

4. CONCLUSIONS

Recent studies have shown that the interaction between lncRNA and EF is closely related to the production of diseases. As more and more computational methods are used to deal with biological problems, which can greatly save manpower, it is possible to use computational methods to predict the interaction between lncRNAs and EFs. In this paper, we proposed a method to predict the association between lncRNAs and EFs. The proposed method combined the Hetesim features and the diffusion features based on multi-feature fusion, and used the machine learning algorithm GBDT to predict the association between lncRNAs and EFs based on heterogeneous networks. The 10-fold cross validation was used to evaluate our method. We also compared our method with others. An environmental factor in the case study was also be used to compare our performance. The results show that the GDBTL2E can obtain high performance. In future, adding the expression profile of lncRNAs to further improve the performance will be investigated.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/zhufangkuang/DLREFD.

AUTHOR CONTRIBUTIONS

JW, ZK, ZM, and GH conceived this work and designed the experiments. JW and ZK carried out the experiments. ZM and

GH collected the data and analyzed the results. JW and ZK wrote, revised, and approved the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grants Nos. 61309027, 61702562, and 61702561, the Hunan Provincial Natural Science Foundation of China under Grants No. 2018JJ3888, the Scientific Research Fund of Hunan Provincial Education Department under Grant No. 18B197, the National Key R&D Program of China under Grant No. 2018YFB1700200, the Open Research Project of Key Laboratory of Intelligent Information Perception and Processing Technology (Hunan Province) under Grant No. 2017KF01, the Foundation Project of Hunan Internet of Things Society No. 2018-2, and the Hunan Key Laboratory of Intelligent Logistics Technology (2019TP1015).

ACKNOWLEDGMENTS

We would like to thank the Experimental Center of School of Computer and Information Engineering, Central South University of Forestry and Technology, for providing computing resources.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00272/full#supplementary-material

REFERENCES

Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., et al. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science* 306, 2242–2246. doi: 10.1126/science.1103388

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov*. 2, 121–167. doi: 10.1023/A:1009715923555

Caye, K., Jumentier, B., Lepeule, J., and François, O. (2019). LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Mol. Biol. Evol*. 36, 852–860. doi: 10.1093/molbev/msz008

Chen, Q., Lai, D., Lan, W., Wu, X., Chen, B., Chen, Y.-P. P., et al. (2020). ILDMSF: inferring associations between long non-coding RNA and disease based on multi-similarity fusion. *IEEE ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2936476

Chen, X. (2016). miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. *Mol. Biosyst*. 12, 624–633. doi: 10.1039/C5MB00697J

Chen, X., Liu, M.-X., Cui, Q.-H., and Yan, G.-Y. (2012). Prediction of disease-related interactions between microRNAs and environmental factors based on a semi-supervised classifier. *PLoS ONE* 7:e43425. doi: 10.1371/journal.pone.0043425

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

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

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res*. 48, D554–D560. doi: 10.1093/nar/gkz843

Cho, H., Berger, B., and Peng, J. (2015). "Diffusion component analysis: unraveling functional topology in biological networks," in *International Conference on Research in Computational Molecular Biology*, ed T. M. Przytycka (Cham: Springer International Publishing), 62–64. doi: 10.1007/978-3-319-16706-0_9

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theor*. 13, 21–27. doi: 10.1109/TIT.1967.1053964

Deng, L., Li, W., and Zhang, J. (2019a). LDAH2V: exploring meta-paths across multiple networks for lncRNA-disease association prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2946257

Deng, L., Wang, J., Xiao, Y., Wang, Z., and Liu, H. (2018). Accurate prediction of protein-lncRNA interactions by diffusion and HeteSim features across heterogeneous network. *BMC Bioinformatics* 19:370. doi: 10.1186/s12859-018-2390-0

Deng, L., Wang, J., and Zhang, J. (2019b). Predicting gene ontology function of human MicroRNAs by integrating multiple networks. *Front. Genet*. 10:3. doi: 10.3389/fgene.2019.00003

Ding, L., Wang, M., Sun, D., and Li, A. (2018). TPGLDA: novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph. *Sci. Rep*. 8:1065. doi: 10.1038/s41598-018-19357-3

Eid, A., Mhatre, I., and Richardson, J. R. (2019). Gene-environment interactions in Alzheimer's disease: a potential path to precision medicine. *Pharmacol. Ther*. 199, 173–187. doi: 10.1016/j.pharmthera.2019.03.005

Florea, A.-M., and Büsselberg, D. (2011). Cisplatin as an anti-tumor drug: cellular mechanisms of activity, drug resistance and induced side effects. *Cancers* 3, 1351–1371. doi: 10.3390/cancers3011351

Flynn, R. A., and Chang, H. Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell* 14, 752–761. doi: 10.1016/j.stem.2014.05.014

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat*. 29, 1189–1232. doi: 10.1214/aos/1013203451

Golub, G. H., and Reinsch, C. (1971). "Singular value decomposition and least squares solutions," in *Linear Algebra* (Berlin; Heidelberg: Springer), 134–151. doi: 10.1007/978-3-662-39778-7_10

Gu, C., Liao, B., Li, X., Cai, L., Li, Z., Li, K., et al. (2017). Global network random walk for predicting potential human lncRNA-disease associations. *Sci. Rep*. 7:12442. doi: 10.1038/s41598-017-12763-z

Guo, Z.-H., You, Z.-H., Wang, Y.-B., Yi, H.-C., and Chen, Z.-H. (2019). A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. *iScience* 19, 786–795. doi: 10.1016/j.isci.2019.08.030

Hattori, M., Tanaka, N., Kanehisa, M., and Goto, S. (2010). SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res*. 38(Suppl_2), W652–W656. doi: 10.1093/nar/gkq367

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2018). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res*. 47, D1013–D1017. doi: 10.1093/nar/gky1010

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(Suppl_1), D98–D104. doi: 10.1093/nar/gkn714

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. Bioinform*. 15, 1774–1782. doi: 10.1109/TCBB.2016.2586190

Li, C., Liu, H., Hu, Q., Que, J., and Yao, J. (2019). A novel computational model for predicting microRNA-disease associations based on heterogeneous graph convolutional networks. *Cells* 8:977. doi: 10.3390/cells8090977

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2/3, 18–22.

Liu, H., Zhang, W., Zou, B., Wang, J., Deng, Y., and Deng, L. (2020). Drugcombdb: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. *Nucleic Acids Res*. 48, D871–D881. doi: 10.1093/nar/gkz1007

Liu, Y., Zeng, X., He, Z., and Zou, Q. (2016). Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE ACM Trans. Comput. Biol. Bioinform*. 14, 905–915. doi: 10.1109/TCBB.2016.2550432

Luo, H., Lan, W., Chen, Q., Wang, Z., Liu, Z., Yue, X., et al. (2018). Inferring microRNA-environmental factor interactions based on multiple biological information fusion. *Molecules* 23, 2439. doi: 10.3390/molecules23102439

Luo, Z., Jegga, A. G., and Bezerra, J. A. (2018). Gene-disease associations identify a connectome with shared molecular pathways in human cholangiopathies. *Hepatology* 67, 676–689. doi: 10.1002/hep.29504

Moreau, J. L., Kesteven, S., Martin, E. M., Lau, K. S., Yam, M. X., O'Reilly, V. C., et al. (2019). Gene-environment interaction impacts on heart development and embryo survival. *Development* 146:dev172957. doi: 10.1242/dev.172957

Opap, K., and Mulder, N. (2017). Recent advances in predicting gene-disease associations. *F1000Res*. 6:578. doi: 10.12688/f1000research.10788.1

Ou-Yang, L., Huang, J., Zhang, X.-F., Li, Y.-R., Sun, Y., He, S., et al. (2019). LncRNA-disease association prediction using two-side sparse self-representation. *Front. Genet*. 10:476. doi: 10.3389/fgene.2019.00476

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. Bioinform*. 16, 688–693. doi: 10.1109/TCBB.2018.2827373

Qiu, C., Chen, G., and Cui, Q. (2012). Towards the understanding of microRNA and environmental factor interactions and their relationships to human diseases. *Sci. Rep*. 2:318. doi: 10.1038/srep00318

Shi, C., Kong, X., Huang, Y., Philip, S. Y., and Wu, B. (2014). Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Trans. Knowl. Data Eng*., 26, 2479–2492. doi: 10.1109/TKDE.2013.2297920

Sun, Y.-Z., Zhang, D.-H., Ming, Z., Li, J.-Q., and Chen, X. (2017). DLREFD: a database providing associations of long non-coding RNAs, environmental factors and phenotypes. *Database* 2017:bax084. doi: 10.1093/database/bax084

Tang, C., Zhou, H., Zheng, X., Zhang, Y., and Sha, X. (2019). Dual laplacian regularized matrix completion for microRNA-disease associations prediction. *RNA Biol*. 16, 601–611. doi: 10.1080/15476286.2019.1570811

Vural, H., and Kaya, M. (2018). Prediction of new potential associations between LncRNAs and environmental factors based on KATZ measure. *Comput. Biol. Med*. 102, 120–125. doi: 10.1016/j.compbiomed.2018.09.019

Xu, Y., Wu, M., Zhang, Q., and Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics* 111, 1115–1123. doi: 10.1016/j.ygeno.2018.07.006

Xu, Z. (2018). *Prediction of correlation between long non-coding RNA and environmental factors based on nuclear similarity of gaussian interaction attributes* (Master's thesis). South China University of Technology, Guangzhou, China.

Yan, C., Wang, J., Ni, P., Lan, W., Wu, F.-X., and Pan, Y. (2017). DNRLMF-MDA: predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE ACM Trans. Comput. Biol.Bioinform*. 16, 233–243. doi: 10.1109/TCBB.2017.2776101

Yu, G., Fu, G., Lu, C., Ren, Y., and Wang, J. (2017). BRWLDA: bi-random walks for predicting lncRNA-disease associations. *Oncotarget* 8:60429. doi: 10.18632/oncotarget.19588

Zhang, J., Zhang, Z., Chen, Z., and Deng, L. (2019). Integrating multiple heterogeneous networks for novel lncRNA-disease association inference. *IEEE ACM Trans. Comput. Biol.Bioinform*. 16, 396–406. doi: 10.1109/TCBB.2017.2701379

Zhang, Z., Zhang, J., Fan, C., Tang, Y., and Deng, L. (2019). KATZLGO: large-scale prediction of LncRNA functions by using the KATZ measure based on multiple networks. *IEEE ACM Trans. Comput. Biol.Bioinform*. 16, 407–416. doi: 10.1109/TCBB.2017.2704587

Zhou, J., and Shi, Y.-Y. (2018). A bipartite network and resource transfer-based approach to infer lncRNA-environmental factor associations. *IEEE ACM Trans. Comput. Biol.Bioinform*. 15, 753–759. doi: 10.1109/TCBB.2017.2695187

Zhou, M., Han, L., Zhang, J., Hao, D., Cai, Y., Wang, Z., et al. (2014). A computational frame and resource for understanding the lncRNA-environmental factor associations and prediction of environmental factors implicated in diseases. *Mol. Biosyst*. 10, 3264–3271. doi: 10.1039/C4MB00339J

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.

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: info@frontiersin.org | +41 21 510 17 00

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