ADVANCED INTERPRETABLE MACHINE LEARNING METHODS FOR CLINICAL NGS BIG DATA OF COMPLEX HEREDITARY DISEASES,

2nd Edition

EDITED BY: Yudong Cai, Tao Huang and Peilin Jia

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

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-862-5
DOI 10.3389/978-2-88966-862-5

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ADVANCED INTERPRETABLE MACHINE LEARNING METHODS FOR CLINICAL NGS BIG DATA OF COMPLEX HEREDITARY DISEASES, 2nd Edition

Topic Editors:
Yudong Cai, Shanghai University, China
Tao Huang, Chinese Academy of Sciences (CAS), China
Peilin Jia, University of Texas Health Science Center, United States

Publisher's note: *This is a 2nd edition due to an article retraction*

Citation: Cai, Y., Huang, T., Jia, P., eds. (2020). Advanced Interpretable Machine Learning Methods for Clinical NGS Big Data of Complex Hereditary Diseases, 2nd Edition. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-862-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 23 October 2020
doi: 10.3389/fgene.2020.600902

Editorial: Advanced Interpretable Machine Learning Methods for Clinical NGS Big Data of Complex Hereditary Diseases

Yudong Cai[1], *Tao Huang*[2]* *and Peilin Jia*[3]

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [3] University of Texas Health Science Center at Houston, Houston, TX, United States

Keywords: artificial intelligence, NGS - next generation sequencing, non-invasive prenatal testing (NIPT), WGS - whole-genome sequencing, WES - whole-exome sequencing

Editorial on the Research Topic

Advanced Interpretable Machine Learning Methods for Clinical NGS Big Data of Complex Hereditary Diseases

Next-generation sequencing (NGS) has revolutionized biomedical research, enabling genome-wide screening of genetic defects. NGS based tests have many applications in Non-Invasive Prenatal Testing (NIPT), early detection of diseases, targeted therapy of various cancers and etiology of rare diseases. There are numerous NGS based genetic test companies and associated data have been accumulated.

As the genomic data increases, it will be a challenge to identify genetic patterns with traditional sampling-based statistical methods. Therefore, advanced machine learning methods, such as deep learning, and Artificial Intelligence (AI) methods can be very beneficial. As an end-to-end method, the deep neural network can extract complex feature patterns automatically and construct prediction models with little manual feature engineering.

Another change the big data has caused is the comeback of instance-based methods or data-driven methods. Unlike the model-based learning or principle-driven methods, the instance-based learning, such as K nearest neighbors, is easy-to-use, easy-to-interpret and has high accuracy when the sample size is big enough to guarantee its performance and the system is too complex to build principle-driven models.

With clinical NGS big data, the genetic causes of various hereditary diseases can be revealed and the shared genetic relationships between diseases can be investigated. Some v very different diseases may share similar genetic causes and should be treated with similar approaches. Some similar diseases may have different genetic causes and should be treated accordingly. The integration of disease network and drug network will become important.

The interpretable model with simple rules is what we need most to transform information exacted from big data to the knowledge that we can master and apply in medical practice. A black box AI algorithm can't appease a worried patient. The interpretable model is not only good for genetic counseling but also essential for knowledge validation and formation. It can also check the correctness of the models and avoid misleading caused by the bias of big data.

The last but not the least change is that in clinical practice, the analysis methods for NGS panel data is quite different from the analysis methods for WGS/WES data which are widely used in the research community. Most research scientists have not faced such challenges and are not even aware of such problems. For clinical panels, we need to re-invent most NGS analysis methods and

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

****Correspondence:***
Tao Huang
tohuangtao@126.com

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

Received: *31 August 2020*
Accepted: *21 September 2020*
Published: *23 October 2020*

Citation:
Cai Y, Huang T and Jia P (2020)
Editorial: Advanced Interpretable
Machine Learning Methods for Clinical
NGS Big Data of Complex Hereditary
Diseases. Front. Genet. 11:600902.
doi: 10.3389/fgene.2020.600902

tools. Such work has mostly been done in industry and hospitals and requires additional research scientist input.

This Research Topic focuses on the challenges of clinical big data analysis in complex genetic diseases, by introducing the latest interpretable machine learning algorithms. There are 22 published articles.

Lv et al. developed a random forest-based sub-Golgi protein classifier rfGPT. The rfGPT used 2-gap dipeptide and split amino acid composition for the feature vectors and was combined with the synthetic minority over-sampling technique (SMOTE) and an analysis of variance (ANOVA) feature selection method. Its accuracy (ACC) was over 90%.

Zhang H. et al. investigated the lung adenocarcinoma (LUAD) and squamous cell lung carcinoma (SCLC) difference on multi-omics scale. With the Boruta method to remove irrelevant features and the MCFS (Monte Carlo Feature Selection) method to identify the significantly important features, they identified 113 key methylation features and 23 key gene expression features.

Wang Y. et al. identified 704 pathogenic genes, 3,848 pathogenic sites, and 2,075 standard phenotypes for underlying molecular perturbations and their phenotypic impact in 3,803 patients with the broad spectrum of intellectual disability (ID). They built the most comprehensive database of an ID phenotyped cohort to date: IDminer http://218.4.234.74:3100/IDminer/, which included the curated ID data and integrated IDpred tool for both clinical and experimental researchers.

Jin et al. studied the biological functions of LINC00356-miR-199a-3p-CDK1/CCNB1 axis in Hepatocellular carcinoma (HCC). Their results proved that LINC00346 could regulate the expression of CDK1/CCNB1 through the competitive adsorption of miR-199a-3p, thereby affecting the p53 signaling pathway and finally regulating the apoptosis, invasion and cell cycle of HCC cells.

Wang H. et al. analyzed the miRNA expression profiles and clinical data of esophageal carcinoma (EC) patients. They found that miR-29c-3p can target CCNA2 to mediate p53 signaling pathway, finally attributing to the inhibition of cell proliferation, migration and invasion, and making cells arrest in G0/G1 phase.

Zhang X. et al. investigated the effects of miR-221-3p in bone marrow mesenchymal stem cell (BMMSC)-derived microvesicles (MVs) on cell cycle, proliferation, and invasion of acute myelocytic leukemia (AML). They discovered that miR-221-3p in BMMSC-derived MVs can regulate AML cell cycle, cell proliferation, and invasion through targeting CDKN1C.

Cheng et al. analyzed the gene expression profiles of 2,343 tumor cells and 1,246 periphery cells. They applied computational methods to screen core biomarkers that can distinguish the discrepancy between Glioblastoma (GBM) tumor and environment (Cheng et al.). Thirty-one important genes were extracted that may be essential biomarkers for GBM tumor cells.

Liu B. et al. collected 10 patients with persistent atrial fibrillation, 10 patients with paroxysmal atrial fibrillation and 10 healthy individuals and did Methylation EPICBead Chip and RNA sequencing. By analyzing the methylation and gene expression data using machine learning-based feature selection method Boruta, they identified the key genes that were strongly associated with AF and found their interconnections.

Hu et al. applied bioinformatics methods for identifying the differentially expressed genes (DEGs) in the lung adenocarcinoma (LUAD) dataset, predicting where the potential target miRNA was expressed and exploring the corresponding downstream target mRNA. They found that exosome-derived miR-486-5p is responsible for cell cycle arrest as well as the inhibition of cell proliferation and metastasis in LUAD via targeting NEK2.

Li et al. proposed a novel method named faster randomized matrix completion for latent disease-lncRNA association prediction (FRMCLDA) by virtue of improved randomized partial SVD (rSVD-BKI) on a heterogeneous bilayer network. Case studies have shown that FRMCLDA is able to effectively predict latent lncRNAs correlated with three widespread malignancies: prostate cancer, colon cancer, and gastric cancer.

Yip et al. developed the Molecular Prognostic Indicators in Cirrhosis (MPIC) database as a representative example of a n omics database tailored for prognostic biomarker validation. MPIC assists cost-effective prognostic biomarker development by facilitating the process of validation and will transform the care of chronic diseases such as cirrhosis. MPIC is freely available at www.mpic-app.org.

Chen et al. presented a novel computational approach to identify potential distinctive features among bacterial subgroups based on a systematic dataset on the gut microbiome from approximately 1,500 human gut bacterial strains. They also established a group of quantitative rules for explaining such distinctions.

Yao et al. analyzed the gene expression profiles of two datasets: one training dataset that includes 144 COPD patients and 194 ILD patients, and one test dataset that includes 75 COPD patients and 61 ILD patients. They identified the 38-gene biomarker and built an SVM (support vector machine) classifier. Its accuracy, sensitivity, and specificity on training dataset evaluated by leave one out cross-validation were 0.905, 0.896, and 0.912, respectively. And on the independent test dataset, the accuracy, sensitivity, and specificity on were as great as and were 0.904, 0.933, and 0.869, respectively.

Xu et al. designed a new model called probability matrix factorization (PMFMDA) for discovering potential disease-related miRNAs. PMFMDA achieved reliable performance in the frameworks of global leave-one-out cross-validation (LOOCV) and 5-fold cross-validation (AUCs are 0.9237 and 0.9187, respectively) in the HMDD (V2.0) dataset, significantly outperforming a few state-of-the-art methods including CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA.

Huang et al. proposed an approach based on information entropy and machine learning for computationally identifying histone butyrylation sites. The proposed method achieved 0.92 of area under the receiver operating characteristic (ROC) curve over the training set by 3-fold cross-validation and 0.80 over the testing set by independent test.

Jiang et al. examined the transcriptional changes of Mycobacterium marinum (*M. marinum*), a pathogenic mycobacterial species closely related to *M. tb*, at different stages of resuscitation from hypoxia-induced dormancy. Their study provided valuable insight into the transcriptome changes of *M. marinum* upon resuscitation as well as gene module function of the bacteria during active metabolism and growth.

Zhou et al. enrolled a total of 564 lung adenocarcinoma patients. The relationship between CTTNB1 mutational status and clinicopathologic parameters, the rates of relapse-free survival (RFS) and overall survival (OS), and the mutational status of other genes commonly mutated in lung adenocarcinoma were analyzed. They found that Female patients and non-smokers are likely to harbor CTNNB1 mutation and primary lung adenocarcinoma with mutated CTNNB1 has a poor prognosis.

Wang C. et al. proposed a PU induction matrix completion algorithm based on heterogeneous information fusion (PUIMCHIF) to predict candidate genes involved in the pathogenicity of human diseases. The experimental results of the PUIMCHIF algorithm regarding the three indexes of precision, recall, and mean percentile ranking (MPR) were significantly better than those of other algorithms.

Zhang J. et al. analyzed the gene expression profiles of 156 KRAS mutation samples and other negative samples with two-stage feature selection approach. Forty-one predictive genes for KRAS mutation were identified and a KRAS mutation predictor was constructed. Its leave one out cross-validation MCC was 0.879.

Su et al. built three multivariable Cox models based on prognostic genes selected from the prognostic protein-coding genes (PCGs) and lncRNAs in gastric cancer. The performance of the three models based on features from only PCGs or lncRNAs or from all prognostic genes were systematically compared, which revealed that the features selected from all the prognostic genes showed higher performance than the features selected only from lncRNAs or PCGs.

Liu X. et al. analyzed the circulating tumor-derived DNAs (ctDNAs) fragment length distribution and found that ctDNA fragments were frequently shorter than the normal cell-free DNA (cfDNA). The findings of this study contributed to improving the detection of low-frequency tumor mutations.

Guo et al. conducted a linkage disequilibrium score regression analysis to confirm the strong genetic correlations between asthma, hay fever and eczema and integrated three distinct association analyses (metaCCA multi-trait association analysis, MAGMA genome-wide and MetaXcan transcriptome-wide gene-based tests) to identify shared risk genes based on the large-scale GWAS results in the GeneATLAS database. Their work may provide help on treatment of asthma, hay fever and eczema in clinical applications.

The 22 articles in this Research Topic only covered a small part of the advanced interpretable artificial intelligence applications in clinical NGS and panel data analysis. We hope more and more AI researchers will devote their time and effort into this field, accelerate the clinical applications of AI and eventually help patients.

AUTHOR CONTRIBUTIONS

YC, TH, and PJ wrote this editorial. All authors contributed to the article and approved the submitted version.

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 September 2019
doi: 10.3389/fbioe.2019.00215

A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features

Zhibin Lv[1†], *Shunshan Jin*[2†], *Hui Ding*[3] *and Quan Zou*[1,3*]

[1] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China, [2] Department of Neurology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China, [3] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China

To gain insight into the malfunction of the Golgi apparatus and its relationship to various genetic and neurodegenerative diseases, the identification of sub-Golgi proteins, both cis-Golgi and trans-Golgi proteins, is of great significance. In this study, a state-of-art random forests sub-Golgi protein classifier, rfGPT, was developed. The rfGPT used 2-gap dipeptide and split amino acid composition for the feature vectors and was combined with the synthetic minority over-sampling technique (SMOTE) and an analysis of variance (ANOVA) feature selection method. The rfGPT was trained on a sub-Golgi protein sequence data set (137 sequences), with sequence identity less than 25%. For the optimal rfGPT classifier with 93 features, the accuracy (ACC) was 90.5%; the Matthews correlation coefficient (MCC) was 0.811; the sensitivity (Sn) was 92.6%; and the specificity (Sp) was 88.4%. The independent testing scores for the rfGPT were ACC = 90.6%; MCC = 0.696; Sn = 96.1%; and Sp = 69.2%. Although the independent testing accuracy was 4.4% lower than that for the best reported sub-Golgi classifier trained on a data set with 40% sequence identity (304 sequences), the rfGPT is currently the top sub-Golgi protein predictor utilizing feature vectors without any position-specific scoring matrix and its derivative features. Therefore, the rfGPT is a more practical tool, because no sequence alignment is required with tens of millions of protein sequences. To date, the rfGPT is the Golgi classifier with the best independent testing scores, optimized by training on smaller benchmark data sets. Feature importance analysis proves that the non-polar and aliphatic residues composition, the (aromatic residues) + (non-polar, aliphatic residues) dipeptide and aromatic residues composition between NH2-termial and COOH-terminal of protein sequences are the three top biological features for distinguishing the sub-Golgi proteins.

Keywords: random forests, sub-Golgi protein classifier, ANOVA feature selection, split amino acid composition, k-gap dipeptide, synthetic minority over-sampling

OPEN ACCESS

Edited by:
Yudong Cai,
Shanghai University, China

Reviewed by:
Tao Zeng,
Shanghai Institutes for Biological Sciences (CAS), China
Zhiwen Yu,
South China University of Technology, China

****Correspondence:***
Quan Zou
zouquan@nclab.net

[†] These authors have contributed equally to this work

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

Received: *23 July 2019*
Accepted: *22 August 2019*
Published: *04 September 2019*

Citation:
Lv Z, Jin S, Ding H and Zou Q (2019) A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features. Front. Bioeng. Biotechnol. 7:215. doi: 10.3389/fbioe.2019.00215

INTRODUCTION

The Golgi apparatus (GA) is an important organelle in eukaryotic cells, because lipids and different types of proteins are modified, packaged, and transported in vesicles to different destinations (Rhee et al., 2005). The GA comprises three main parts (Xu and Esko, 2009): cis-Golgi, medial, and trans-Golgi. The cis-Golgi receives proteins and then delivers them to the medial section for protein biosynthesis. The trans-Golgi releases the biosynthesized proteins from the medial section. The proteins in the cis-region of the GA are called cis-Golgi proteins, whereas trans-Golgi proteins are in the trans-Golgi part (Pfeffer, 2001).

Malfunction of the GA can disrupt protein biosynthesis in the medial part, which can lead to neurodegenerative diseases, such Parkinson's (Fujita et al., 2006; Yang J. et al., 2016) and Alzheimer's (Gonatas et al., 1998; Yang et al., 2015). A key step in the understanding of GA function is to determine whether a protein is a sub-Golgi protein (cis-Golgi or trans-Golgi). Such determinations will improve comprehension of the mechanisms for GA dysfunction and provide clues for disease treatment and more effective drug research and development (Gunther et al., 2018).

In the past few years, several protein subcellular locations and protein type prediction tools, including sub-Golgi protein identification tools (Teasdale and Yuan, 2002; Van Dijk et al., 2008; Chou et al., 2010; Ding et al., 2011, 2013; Jiao et al., 2014; Lin et al., 2014; Nikolovski et al., 2014; Jiao and Du, 2016a,b; Yang R. et al., 2016; Ahmad et al., 2017; Wang et al., 2017; Rahman et al., 2018; Ahmad and Hayat, 2019; Wuritu et al., 2019), have been developed using various machine learning algorithms, including increment diversity Mahalanobis discriminant (IDMD) (Ding et al., 2011), support vector machine (SVM) (Ding et al., 2013, 2017; Jiao et al., 2014; Lin et al., 2014; Jiao and Du, 2016a,b), random forest (RF) (Ding et al., 2016a,b; Yang R. et al., 2016; Yu et al., 2017; Liu et al., 2018), and K nearest neighbor algorithm (KNN) (Ahmad et al., 2017; Ahmad and Hayat, 2019), among others. To generate feature vectors for sub-Golgi protein identification, protein amino acid composition (AAC) (Rahman et al., 2018), k-gapped dipeptide composition (k-gapDC) (Ding et al., 2011, 2013), pseudo amino acid composition (PseAAC) (Jiao et al., 2014; Liu et al., 2015), and protein sequences evolutionary information (e.g., position-specific scoring matrix, PSSM) and their derivative features (Yang et al., 2014; Jiao and Du, 2016a,b; Yang R. et al., 2016; Ahmad et al., 2017; Rahman et al., 2018) have been used. Because the extensively used training benchmark data sets (Ding et al., 2013; Yang R. et al., 2016) are unbalanced in sub-Golgi protein classes, a synthetic minority over-sampling technique (SMOTE) has been adopted to obtain class-balanced data sets for training (Yang R. et al., 2016; Ahmad et al., 2017; Wan et al., 2017; Rahman et al., 2018; Ahmad and Hayat, 2019). Diversified feature selection methods, including analysis of variance (ANOVA) (Ding et al., 2013; Jiao and Du, 2016a), minimal redundancy-maximal relevance (mRMR) (Jiao and Du, 2016b; Wang S. P. et al., 2018), maximum relevance-maximum distance (MRMD) (Zou et al., 2016a,b), RF/Wrapper (Pan et al., 2018; Rahman et al., 2018), multi-voting for feature selection (Ahmad and Hayat, 2019), and lasso (Liu et al., 2016), among others, have been used to remove redundant features and improve the prediction accuracy with as few features as possible (Yu et al., 2016; Zhu et al., 2017, 2018; Kuang et al., 2018; Wang H. et al., 2018).

Two widely used benchmark-training data sets have resulted in different optimization models with various independent testing prediction scores. For the benchmark data set of Ding (137 sequences with 25% sequence identity; Ding et al., 2013), Jiao and Du (2016b) applied 49-dimensional features of positional-specific physicochemical properties (PSPCP, a derived feature from PSSM) to train their best SVM model. They achieved jackknife cross-validation results with accuracy (ACC) of 91.2%; Matthew correlation coefficient (MCC) of 0.793; sensitivity (Sn) of 99.0%; and specificity (Sp) of 73.8%, whereas the independent prediction accuracy of their classifier was 87.1%. The best predictor built on the benchmark data set of Yang (304 sequences with 40% sequence identity) (Yang R. et al., 2016) was developed by Ahmad and Hayat (2019). They carefully selected 180-dimensional features from the combined features of split amino acid composition (SAAC), 3-gap dipeptide composition, and PSSM with its derivative features to obtain a designed KNN classifier with good jackknife cross-validation scores (ACC = 94.9%; MCC = 0.90; Sn = 97.2%; Sp = 92.6%) and good independent testing scores (ACC = 94.0%; MCC = 0.84; Sn = 81.5%; Sp = 96.9%).

To our best knowledge, all high-profile sub-Golgi protein predictors trained on either benchmark data sets are constructed on the basis of a PSSM and its derived feature vectors, whose acquisition requires the use of a position-specific iterative basic local alignment search tool to align sub-Golgi protein sequences with a protein database (Jiao and Du, 2016a,b; Rahman et al., 2018; Ahmad and Hayat, 2019). Then, a secondary data transformation is performed (Altschul et al., 1997) in which data are usually converted into a 20 by 20 matrix with average values in each feature dimension (Jiao and Du, 2016a,b; Yang R. et al., 2016; Ahmad et al., 2017; Rahman et al., 2018). The sequence alignment is typically time-consuming, particularly when the protein database for alignment is large and the computing power is limited.

In this paper, instead of using PSSM and its derived features, the focus was on constructing an efficient sub-Golgi protein RF classifier, namely rfGPT, based only on amino acid and dipeptide composition-based feature vectors. Related studies (Li et al., 2016; Luo et al., 2016; Tang et al., 2018; Zhang et al., 2018a,b) have demonstrated the effectiveness of composition and dipeptide and amino acid composition-based features for solving bioinformatics problems. The rfGPT with 55-dimensional features of 2-gap dipeptide composition attained better jackknife cross-validation scores (ACC = 91.1%; MCC = 0.823; Sn = 87.4%; Sp = 94.7%) and better independent testing results (ACC = 89.1%; MCC = 0.631; Sn = 53.8%; Sp = 98.0%) than those classifiers trained on the same data set (Ding et al., 2013; Jiao and Du, 2016a,b). Therefore, to date, the rfGPT is the best sub-Golgi predictor trained from the benchmark

Abbreviations: D/Dim, dimension; D0/D1/D2/D3, data sets; IDMD, increment diversity Mahalanobis discriminant; SVM, supporting vector machine; KNN, K-nearest neighbors; RF, random forests; 2-gapDC, 2-gap dipeptide composition; 3-gapDC, 3-gap dipeptide composition; DPDC, Dipeptide compostion; TPDC, Tripeptide composition; AAC, amino acid composition; SAAC, split amino acid composition; PseAAC, pseudo amino acid composition; PSPCP, positional-specific physicochemical properties derived feature from PSSM; PSSM, position-specific scoring matrix; PSSMDC, PSSM-Dipeptide Composition; BigramPSSM, Bi-gram features directly extracted from PSSM; EDPSSM, Evolutionary Difference PSSM; CSP, Common Spatial Patterns; SMOTE, synthetic minority over-sampling technique; ACC, accuracy; MCC, Matthew correlation coefficient; Sn, Sensitivity; Sp, Specificity.

data set of Ding via SMOTE (Ding et al., 2013). For further improvement of the rfGPT, 59 2-gap dipeptide composition features selected through ANOVA technology were fused with SAAC features to form 119 new dimensional features, which were then secondarily selected via ANOVA for rfGPT optimization. Ultimately, the rfGPT with 93 dimensional features [59 2-gap dipeptide composition (DC) sub-features plus 34 SAAC sub-features] was the best predictor, with jackknife cross-validation scores of ACC = 90.5%; MCC = 0.811; Sn = 92.6%; and Sp = 88.4%, and independent test scores of ACC = 90.6%; MCC = 0.696; Sn = 96.1%; and Sp = 69.2%.

MATERIALS AND METHODS

Data Sets

To train models for sub-Golgi protein identification, two benchmark-training data sets are widely used. One data set, D1 in this text, was constructed by Ding et al. (2013), and the other, D2 in this text, was constructed by Yang R. et al. (2016). Before D1 was developed, Ding et al. constructed a smaller data set (D0) which was used once and never used again (Ding et al., 2011). In this work, the data set D1 was downloaded from http://lin-group.cn/server/SubGolgi/data and used to train the sub-Golgi protein classifier. The D1 data set consisted of 137 Golgi-resident protein sequences, with 42 cis-Golgi and 95 trans-Golgi proteins. The D1 data set was selected for model training primarily because the sequence identity was <25%. Thus, the D1 data set contained less sequence noise and redundancy than the D2 data set.

For testing the optimized model, an independent data set D3 provided by Ding et al. (2013) was applied. The D3 data set has been adopted by most of the key researchers in previously reported sub-Golgi predictors (Ding et al., 2013; Jiao and Du, 2016b; Yang R. et al., 2016; Ahmad et al., 2017; Rahman et al., 2018; Ahmad and Hayat, 2019). The D3 data set is generally used only for independent testing and contains 64 test sequences, including 13 cis-Golgi and 51 trans-Golgi protein sequences. The D3 data set is available at http://lin-group.cn/server/SubGolgi/data.

Modeling Overview

The entire rfGPT modeling process is illustrated in **Figure 1**. Compared with previous predictors, the major difference of the rfGPT used in this study was that only extracted features

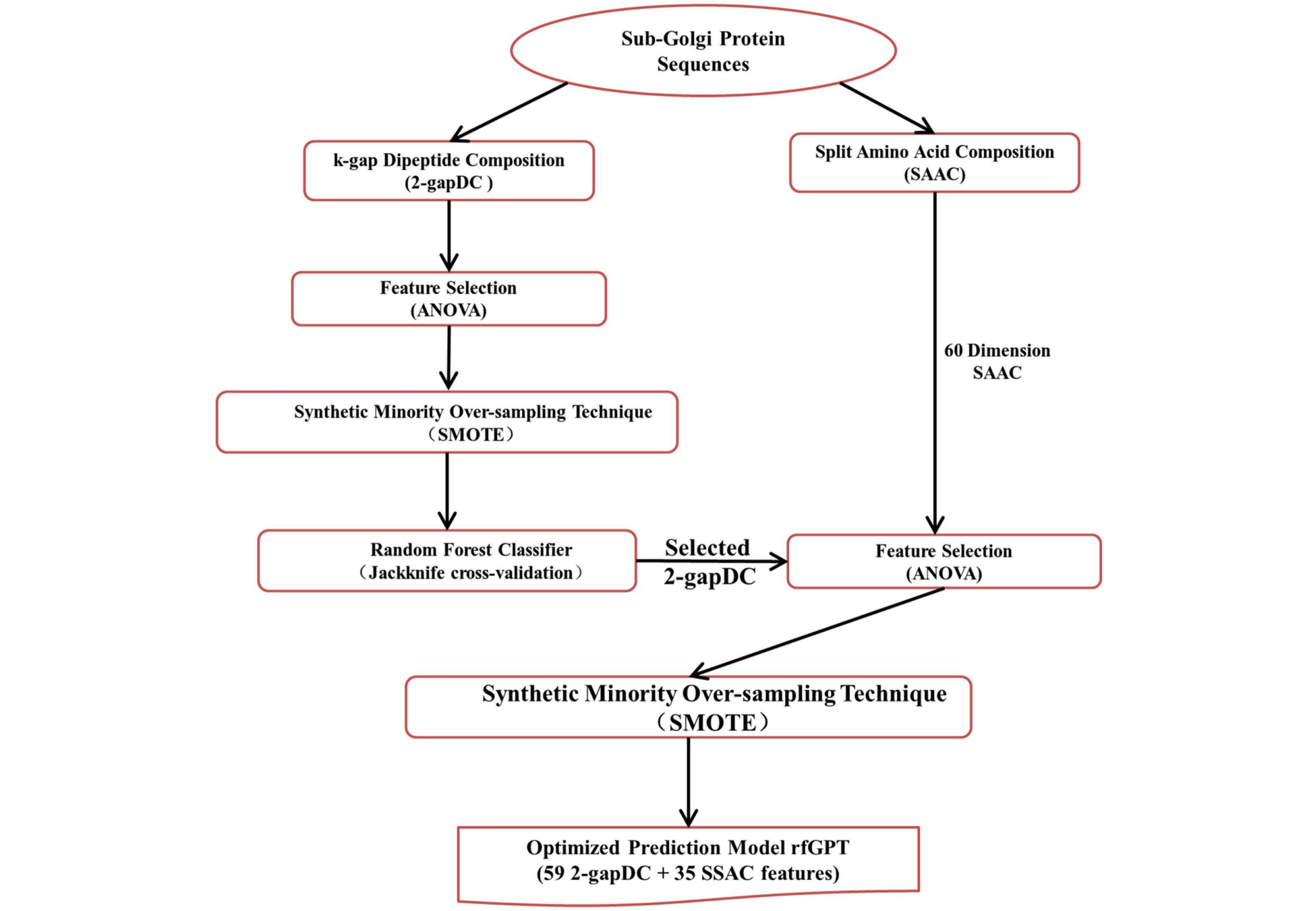

Sub-Golgi Protein Sequences
k-gap Dipeptide Composition (2-gapDC)
Split Amino Acid Composition (SAAC)
Feature Selection (ANOVA)
60 Dimension SAAC
Synthetic Minority Over-sampling Technique （SMOTE）
Random Forest Classifier （Jackknife cross-validation）
Selected 2-gapDC
Feature Selection (ANOVA)
Synthetic Minority Over-sampling Technique （SMOTE）
Optimized Prediction Model rfGPT (59 2-gapDC + 35 SSAC features)

FIGURE 1 | Modeling framework of the state-of-art random forests sub-Golgi protein classifier. ANOVA: analysis of variance.

from amino acid and dipeptide composition were used. In this study, the 2-gapped dipeptide composition profile and SAAC were adopted. Ding et al. (2013) verified the validity of the 2-gapped dipeptide composition profile for sub-Golgi prediction. The SAAC considers that the location of a Golgi protein is related to the composition of amino acid residues at the N-terminal and C-terminal of a protein sequence (Paulson and Colley, 1989). As shown in **Figure 1**, the 400 dimensions (400D) 2-gapDC features extracted from D1 were used to generate a class-balanced data set via ANOVA and SMOTE, which was then fed into a RF model for optimization and estimation by jackknife cross-validation and independent testing. In this step, an optimized prediction model was sought, whose selected features were then combined with the SAAC features as new features of a new model for further optimization. After the secondary feature selection via ANOVA and SMOTE, the new optimal model was evaluated through jackknife cross-validation and independent testing.

Feature Extraction

The methods for feature extraction used for sub-Golgi classification are divided into three categories: (1) amino acid and peptide composition and their derived features; (2) PSSM and its derived features; and (3) features combined with amino acid residue physical and chemical properties. In this research, the derived features of category 1 were adopted because they are simple and convenient for feature extraction, namely, to calculate the frequency of peptide and amino acid components. The following two AAC features were adopted.

k-Gapped Dipeptides Composition

In general, the composition of adjacent dipeptides can only reflect the short-range structure of the protein sequence. The dipeptide composition in the larger interval may better reflect the tertiary structure of the protein. In biology, interval residues are more important than adjacent residues. Especially in some common structures, such as helices and plates, two non-adjacent residues are joined by hydrogen bonds (Lin et al., 2015; Wang et al., 2019). The k-gap dipeptides composition (k-gapDC) is an indirect mathematical description of the biological significance, which has been extensively utilized for sub-Golgi protein classification and other bioinformatics fields (Xu et al., 2018; Agrawal et al., 2019; Akbar et al., 2019; Wang et al., 2019). For the k-gapDC, the frequency of a dipeptide separated by k positions is determined, which is then divided by the total number of k-gapped dipeptides; thus, a protein sequence is transformed into a 400D feature vector. The 2-gapDC features were utilized in this work.

Split Amino Acid Composition

It has been proved that the N-terminal and C-terminal of protein sequences can act as signal-anchor domains for subcellular locations, e.g., glycosyltransferases all have a short NH2-terminalcytoplasmic tail, a 16-20-amino acid signal-anchor domain, and an extended stem region which is followed by the large COOH-terminal catalytic domain (Paulson and Colley, 1989). Another example is that lysine at position 329 within a C-terminal dilysine motif is crucial for the endoplasmic reticulum localization of human SLC35B4 (Bazan et al., 2018). All of these inspire us to used split amino acid composition for sub-Golgi protein identification. The split amino acid composition was proposed by Chou (Chou and Shen, 2007), which converts variable-length protein sequences into fixed-length amino acids for feature representation. In SAAC, a protein sequence is initially segmented into different parts, and then the amino acid frequency of each independent part is calculated. In the current work, the protein sequences were split into three segments: 30 N-terminal residues, 30 C-terminal residues, and the intermediate-block residues, which are the sequences between N-terminal and C-terminal parts. A 60D feature vector was obtained from the SAAC instead of the traditional 20D amino acid component. The details of the SAAC feature extraction are described as follows. Considering the length of protein sequence L and the three segments [NSeg (N-terminal), ISeg (intermediate block), and CSeg (C-terminal)] with the lengths Xn, L – Xn – Xc, and Xc (Xn = Xc = 30), respectively, the SAAC feature vector $\left[f_1, f_2, \cdots, f_{60}\right]$ is generated by the following formulas:

- $f_i = \frac{N(AA_i)}{X_n},\ i = 1, 2, \ldots, 20$
- $f_i = \frac{N(AA_i)}{L - X_n - X_c},\ i = 21, 22, \ldots, 40$
- $f_i = \frac{N(AA_i)}{X_c},\ i = 41, 42, \ldots, 60$
  AA : amino acid residue;
  $N\left(AA\right)$: the numbers of AA in different segments.
  L: the length of protein sequence;
  X_n: the residues numbers of N-terminal segments;
  X_c: the residues numbers of C-terminal segments.
  f_i: the ith SAAC feature vector element, it is one of the 20 amino acid residue frequency in a segment.

Feature Selection

Feature selection is conducted to remove redundant information and to overcome over-fitting in machine learning modeling. A variety of feature selection techniques (Ding et al., 2013; Jiao et al., 2014; Zeng et al., 2015, 2016, 2018; Jiao and Du, 2016a,b; Yang R. et al., 2016; Ahmad et al., 2017; Rahman et al., 2018; Ahmad and Hayat, 2019; Liu Y. et al., 2019; Zhang X. et al., 2019) have been important for sub-Golgi protein identification and for other areas of bioinformatics. ANOVA ranks the importance of features in terms of the ratio of the variance of data within a category to the variance between categories. The larger the value of the ratio is, the more important the feature is. The details for the use of ANOVA as a feature selection technique have been presented previously (Ding et al., 2013; Jiao and Du, 2016a) and are not repeated here. In this study, the ANOVA module from the famous Scikit-learn machine learning tool kit was used for feature selection (https://scikit-learn.org/).

Synthetic Minority Over-sampling Technique

The D1 benchmark data set is imbalanced, with the cis-Golgi protein and trans-Golgi protein sequences ratio of 0.44. Such an imbalance has a significant impact on the acceptability of the application, because the classifiers can be overly suitable for the majority classes. In this case, the prediction accuracy may seem high, but the results may be unacceptable, as minority groups may be completely/partially ignored. To solve this problem, the very effective SMOTE was proposed by Chawla et al. (2002). SMOTE helps to balance unbalanced data sets by creating "synthetic" minority class examples rather than by oversampling with replacement, and is employed by various sub-Golgi classifiers trained on benchmark data set D2 (Yang R. et al., 2016; Ahmad et al., 2017; Rahman et al., 2018; Ahmad and Hayat, 2019). As this manuscript was prepared, the use of SMOTE with benchmark data set D1 had not yet been reported. In this research, the SMOTE module implemented was from http://imbalanced-learn.org.

Evaluation Metrics

Testing Methods

The jackknife cross-validation is a leave-one-out cross-validation method for testing the efficiency of protein classification (Chou and Shen, 2006) and is executed in the following steps. A training data set with T items is separated into two parts. For each run, one part consists of T−1 item for model training, and the remaining part contains one item for testing. This process is repeated T times, and all the items sampled in the training data set act as a testing sample only once. Jackknife cross-validation is a time-consuming method, particularly for large data sets, but the method is robust with small variance. In this article, the benchmark data set D1 collected by Ding et al. (2013) was used for the jackknife cross-validation.

In independent testing, a completely different data set from the training data set is used to evaluate the trained model. Once the model is built with the training data set, tests are performed on the independent data set to evaluate the model. In this article, the independent data set D3 collected by Ding et al. (2013) was used for model performance evaluation.

Performance Metrics

Four standard metrics were used to evaluate the proposed models: ACC, Sn, Sp, and MCC. The metrics are previously described (Wei et al., 2017a,b; Chen et al., 2018; Su et al., 2018; Feng et al., 2019; Zhang S. et al., 2019) and were calculated as follows:

- $ACC = \frac{TP+TN}{TP+TN+FP+FN}$
- $S_n = \frac{TP}{TP+FN}$
- $S_p = \frac{TN}{TN+FP}$
- $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TN+FN) \times (TP+FN) \times (TN+FP)}}$

where TP is a true positive, TN is a true negative, FP is a false positive, and FN is a false negative.

Classifier

Support vector machine (SVM) (Ding et al., 2011, 2013; Feng et al., 2013; Lin et al., 2014; Jiao and Du, 2016a,b; Zeng et al., 2017; Rahman et al., 2018; Chen et al., 2019; Dao et al., 2019; Liu B. et al., 2019), K-nearest neighbor (KNN) (Ahmad et al., 2017; Ahmad and Hayat, 2019), and random forests (RF) (Yang R. et al., 2016; Pan et al., 2017; Ru et al., 2019; Su et al., 2019; Zheng et al., 2019) classifiers have been used to identify sub-Golgi proteins and for other fields. In this study, RF was selected for modeling because it is a powerful machine-learning tool and facilitates analysis of feature importance. Previously, Yang R. et al. (2016) selected 55 features from composite features (3-gapDC + PSSM derived features) to optimize their random forest classifier. The jackknife cross-validation scores using data set D2 were ACC = 88.5%; MCC = 0.765; Sn = 88.9%; and Sp = 88.0%, and for the independent testing, the scores were ACC = 93.8%; MCC = 0.821; Sn = 92.3%; and Sp = 94.1% (Yang R. et al., 2016). However, those results are somewhat confusing, because other sub-Golgi predictors have lower independent test scores than those for the jackknife cross-validation. To date, no sub-Golgi RF predictor has been trained from benchmark data set D1. In this study, the random forest classification model in the Scikit-learn tool kit (https://scikit-learn.org/) was applied for the implementation, testing, and evaluation of the rfGPT classifier and for the analysis of feature importance.

RESULTS AND DISCUSSION

Performance of Random Forests Classifier Without Feature Selection

Table 1 shows the performance of the rfGPT using various extracted features. In the models with the SMOTE technique, the cross-validation scores improved remarkably for ACC, MCC, Sn, and Sp. For example, based on 460D SAAC + 2-gapDC features and SMOTE, the scores of the rfGPT were ACC = 90.5%; MCC = 0.817; Sn = 96.8%; and Sp = 84.2%, which were increases of 20, 132, 44, 2.2, and 171.6%, respectively, compared with the rfGPT without SMOTE. Although the SMOTE technique does

TABLE 1 | Jackknife cross-validation and independent testing results after training on the benchmark data set D1 without feature selection.

Feature(D)	SMOTE (Y/N)	Jackknife cross-validation				Independent testing			
		ACC	MCC	Sn	Sp	ACC	MCC	Sn	Sp
2-gapDC(400)	N	74.5%	0.326	94.7%	28.6%	79.7%	0.318	90.2%	38.5%
SAAC(60)	N	69.3%	0.073	97.9%	4.8%	78.1%	−0.07	98.0%	0.0%
2-gapDC+SAAC(460)	N	75.2%	0.351	94.7%	31.0%	79.7%	0.237	94.1%	23.1%
2-gapDC(400)	Y	86.3%	0.743	96.8%	75.8%	82.8%	0.351	98.0%	23.1%
SAAC(60)	Y	87.9%	0.763	93.7%	82.1%	81.2%	0.388	90.2%	46.2%
SAAC+2-gapDC(460)	Y	90.5%	0.817	96.8%	84.2%	81.2%	0.287	96.1%	23.1%

improve the recognition rate of minority classes, the accuracy of the independent testing for the rfGPT with diverse features ranged from 78.1 to 82.8%, with little improvement with SMOTE (**Table 1**). For the other metrics (MCC, Sn, Sp), the case was the same. Thus, other techniques are needed to improve the generalization prediction model. In this paper, to obtain a better rfGPT with fewer features, ANOVA feature selection was used to eliminate redundant features.

Classifier Optimizing via ANOVA Feature Selection

To obtain the optimized classifier, the ANOVA feature selection method was first conducted for 400 2-gapDC features. One hundred sub-data sets containing 1, 2, ... and 100 2-gapDC features generated separately after ANOVA feature selection were used for training 100 corresponding RF classifiers. For all 100 classifiers, jackknife cross-validation and independence testing were conducted. **Figure 2A** shows the accuracy of the cross-validation and independent tests of the 100 classifiers with varying numbers of features. Except for the models with nine and ten selected features, the average accuracy of the jackknife cross-validation of the other models was higher than that of the independent test results. Based on the jackknife cross-validation, the best-trained model with the highest accuracy was the classifier with 59 selected features (rfGPT_1), whereas the classifier with 55 selected features (rfGPT_2) had the highest independent testing accuracy results.

The performance scores of both classifiers are listed in **Table 2**. The jackknife cross-validation scores of rfGPT_2 (ACC = 91.1%; MCC = 0.823; Sn = 94.7%; Sp = 87.4%) were slightly lower than those of rfGPT_1 (ACC = 93.2%; MCC = 86.4%; Sn = 94.7%; Sp = 91.6%). However, rfGPT_2 had the better predictive performance on the independent test sets with scores of ACC = 89.1%; MCC = 0.631; Sn = 98%; and Sp = 53.8%, which were as much as 5.6, 35, 8.3, 10, and 16% larger than the corresponding values of rfGPT_1 (ACC = 84.4%; MCC = 0.466; Sn = 94.1%; Sp = 46.2%). The 89.1% independent testing accuracy of rfGPT_2 was an increase of 2.2% compared with the best SVM sub-Golgi classifier (Jiao and Du, 2016b) trained on the same benchmark data set (D1). The accuracy of 93.2% for rfGPT_1 and 91.1% for rfGPT_2 from the jackknife cross-validations was an increase of 9.0 and 6.5%, respectively, compared with that of the RF classifier obtained by Yang et al. which was trained on benchmark data set D2 (Yang R. et al., 2016).

For further optimization, the 59 2-gapDC features of rfGPT_1 obtained in the previous step were combined with 60 SAAC features to form 119-dimensional (2-gapDC + SAAC) composite features, and then ANOVA was used to construct 100 data sets with selected 1, 2, ... and 100 features for building 100 classifiers. The jackknife cross-validation and independent test results for these models are shown in **Figure 2B** and **Table 2**. For the cross-validation performance, classifier rfGPT_3 with 43 features was better than classifier rfGPT_4 with 93 features. However, for independent testing, the predictive metric of rfGPT_4 with ACC= 90.6%; MCC = 0.696; Sn = 96.1%; and Sp = 69.2% exceeded that of rfGPT_3 with ACC = 84.4%; MCC = 0.466;

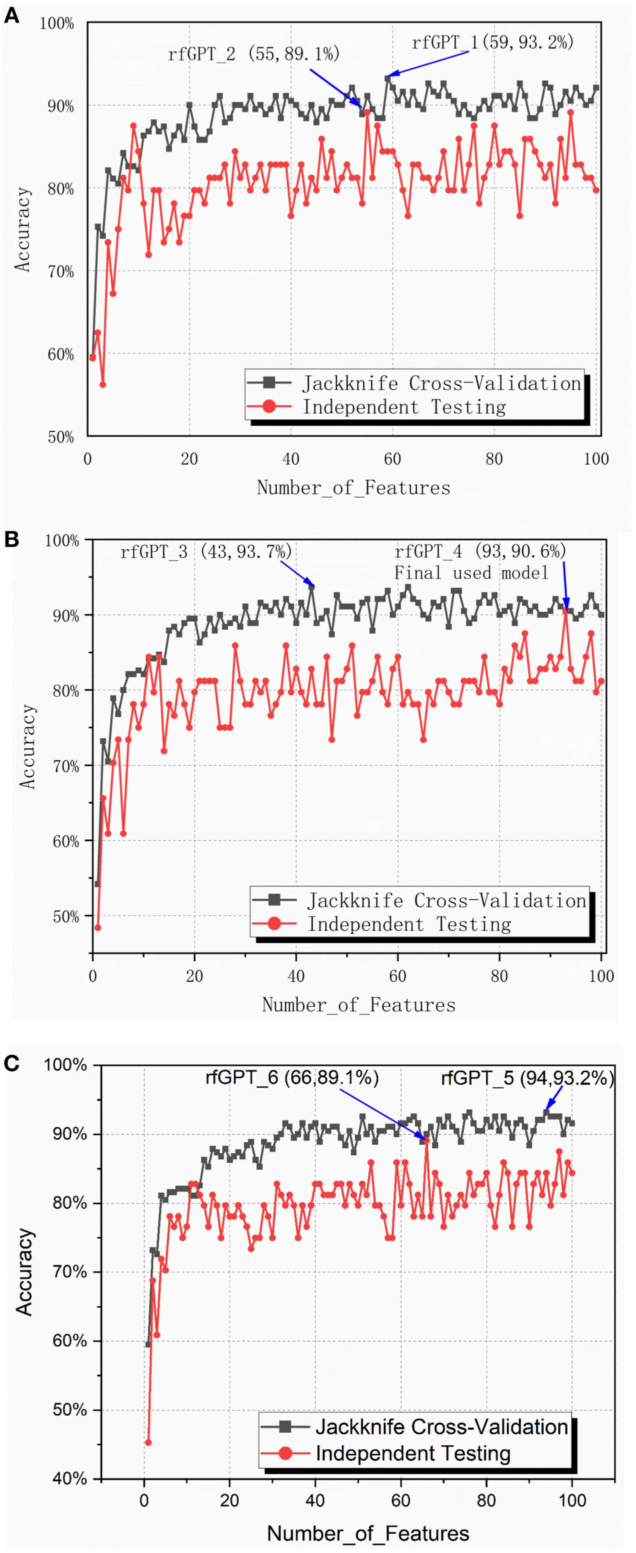


FIGURE 2 | Jackknife cross-validation and independent testing accuracy of the random forest classifier with the number of features varied: **(A)** 2-gap dipeptide composition (2-gapDC) features **(B)** 59 selected 2-gapDC features + 60 split amino acid composition (SAAC) features, and **(C)** 55 selected 2-gapDC features + 60 SAAC features.

TABLE 2 | The best evaluation scores from jackknife cross-validation and independent testing of different models with various feature types and feature numbers.

Classifier	Features(D)	Jackknife cross-validation				Independent testing			
		ACC	MCC	Sn	Sp	ACC	MCC	Sn	Sp
rfGPT_1	2-gapDC(59)	93.2%	0.864	94.7%	91.6%	84.4%	0.466	94.1%	46.2%
rfGPT_2	2-gapDC(55)	91.1%	0.823	94.7%	87.4%	89.1%	0.631	98.0%	53.8%
rfGPT_3	2-gapDC+SAAC(43)	93.7%	0.874	93.7%	93.7%	82.8%	0.484	88.2%	61.5%
rfGPT_4	2-gapDC+SAAC(93)	90.5%	0.811	92.6%	88.4%	90.6%	0.696	96.1%	69.2%
rfGPT_5	2-gapDC+SAAC(94)	93.2%	0.864	93.7%	92.7%	84.4%	0.546	88.2%	69.2%
rfGPT_6	2-gapDC+SAAC(66)	90.0%	0.800	89.5%	90.5%	89.1%	0.695	90.2%	84.6%

Sn = 88.2%; and Sp = 61.5%; the increases were 7.3%, 49, 8.3, 9.0, and 13%, respectively.

Optimization was also performed by combining the 55 2-gapDC features of rfGPT_2 with SAAC features to form 115-dimensional features for 100 new models with various features. The cross-validation and independent testing accuracy scores are revealed in **Figure 2C**. The scores for rfGPT_5 and rfGPT_6 are shown in **Table 2**. The independent accuracy of both models was inferior to that of rfGPT_4 (**Table 2**).

Because most cross-validation and independent testing scores of the classifier rfGPT_4 were superior to those of other models in **Table 2**, rfGPT_4 was designated as the final sub-Golgi model for prediction.

Feature Importance Analysis

To analyze the importance of the features selected for rfGPT_4, the feature importance function of the Scikit-learn RF model was exploited (**Figure 3**). As shown in **Figure 3A**, 59 2-gapDC features and 34 SAAC features were adopted in rfGPT_4, and their importance to the classification of Golgi proteins was 72.4 and 27.6%, respectively. **Figure 3B** shows the ranking of the 93 features by importance value and the cumulative importance score by importance value order. Among the combined features, the single feature importance was diverse and ranged from 0.16 to 3.64%. **Figure 3C** shows the importance order of the first 25 specific features, which accounted for 50% of the importance for the rfGPT. Only four of the top 25 features (which included 21 2-gapDC features and 4 SAAC features) had an importance value of more than 3% (**Figure 3C**).

To further analyze the feature bio-meaning, the feature importance values are assigned to different types of amino acid residues, that is aromatic residues, non-polar, and aliphatic residues, polar and non-charged residues, positively charged residues, and negatively charged residues. For instance, FP.gap2 feature as shown in **Figure 3C** means the composition frequency of dipeptide, which consists of F (phenylalanine) and P (proline) amino acid residence. The importance value 3.64% for FP.gap2 feature is divided by 2 to allocate 1.72% to aromatic residues type and non-polar and aliphatic residues type. Other features importance values are handled in the same way to assign importance value to five type amino acid residues (see **Table S1**). It finds out that the importance value of non-polar and aliphatic residues, aromatic residues, negatively charged residues, positively charged residues, polar, and non-charged residues are 30%, 24%, 21%, 13% and 12%, respectively. The non-polar and aliphatic property of amino acid residues plays the most critical role in sub-Golgi protein identification, and then the next is aromatic, negatively charged, positively charged, and polar and non-charged in turn. The importance values of the first three properties add up to 75%, so it concludes that to discriminate cis or trans sub-Golgi protein is mainly determined by the non-polar and aliphatic residues, aromatic residues, and negatively charged residues composition frequency.

For 2-gap DC features, the first three most important features are FP.gap2 (3.64%), IG.gap2 (3.50%), and GD.gap2 (3.44%), and five different residue types combined with each other generate 25 type dipeptides, whose feature importance values are listed in **Figure 3C** and **Table S2**. The (aromatic residues) + (non-polar, aliphatic residues) dipeptide, (non-polar, aliphatic residues) + (non-polar, aliphatic reduces residues) dipeptide and the (non-polar, aliphatic residues) + (aromatic residues) with the importance values as 8.54%, 8.18%, and 7.36%, respectively, are the top three important features for sub-Golgi classification.

For SAAC features, the protein sequence is segmented into three parts: N-terminal segment, C-terminal segment and the Interblock between N-terminal and C-terminal, whose amino acid composition frequency feature is labeled as Nterminal_A, Cterminal_A and InterTier_A (A represents one of the 20 amino acid residues; see **Figure 3C** and **Table S3**). The importance values of N-terminal features, C-terminal features, and Interblock features are 6.43%, 8.81%, and 12.37%, separately. The first three important values of 5 types residues of each block is aromatic residues of Interblock (5.05%), non-polar and aliphatic residues of C-terminal (3.13%), and negatively charged residues of N-terminal (3.00%). The D (aspartate) residues composition of N-terminal, as shown in **Figures 3C**, is the most important SAAC feature for sub-Golgi classification, but the aromatic residues composition frequency features of the Interblock seem even more important (see **Table S3**).

To sum up the above, the non-polar and aliphatic residues composition, the (aromatic residues) + (non-polar, aliphatic residues) dipeptide and aromatic residues composition between NH_2-termial and COOH-terminal of protein sequences are three top biological features for distinguishing the sub-Golgi proteins.

Metrics Comparison With Existing Predictors

Ten optimized sub-Golgi classifiers that have been developed are presented in **Table 3**. Three separate data sets (D0, D1,

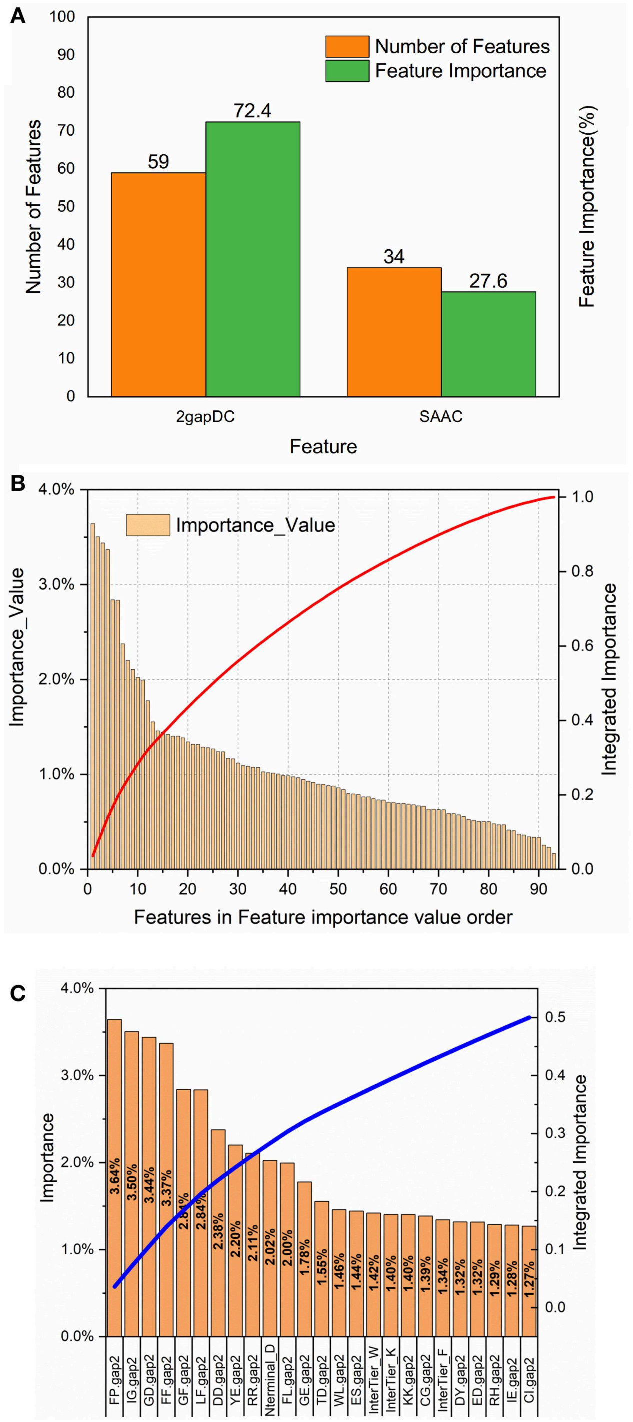


FIGURE 3 | Feature importance analysis of random forests sub-Golgi classifier, rfGPT _4: **(A)** importance of feature types **(B)** the ranking orders of 93 features for rfGPT_4 and their integrated importance (red line), and **(C)** the importance of the top 25 features, which accounted for 50% of the integrated importance (blue line). The A_1A_2.gap2 means the composition of dipeptide A1A2. A1 or A2 is one of the 20 amino acid residues. Nterminal_D means the composition of amino acid residues D (aspartate) in NH_2-terminal of protein sequence. InterTier_K, interTier_W, and interTier_F mean K(lysine), W(tryptophan), and F(phenylalanine) amino acid residues composition of the inter-tier between NH_2-terminal and COOH-terminal of protein sequence.

D2), and four machine learning algorithms (IDMD, SVM, KNN, RF) were exploited to train these sub-Golgi classifiers, and one common independent data set was used to evaluate the various sub-Golgi classifiers. A total of six classifiers adopted the PSSM and its derived features for sub-Golgi prediction. Ahmad et al. (2017), training on the D2 data set with 40% sequence identity, achieved the highest independent testing scores (ACC = 94.8%; MCC = 0.86; Sn = 93.9%; Sn = 94.0%) for a classifier; the KNN sub-Golgi classifier with 83 composited features. In contrast to the KNN sub-Golgi classifier of Ahmad et al. the ultimate classifier rfGPT_4 in this paper was trained on the benchmark data set D1 with 25% sequence identity and contained 93 features, without any PSSM and its derivative features. Therefore, the rfGPT_4 is more practical, because the time-consuming sequence alignment step to obtain the PSSM and its derivatives scores using the Position-Specific Iterative Basic Local Alignment Search Tool is avoided. In addition, rfGPT_4 is currently the model with the best independent testing scores for training on data set D1 and is a state-of-art sub-Golgi classifier with only dipeptide and amino acid composition features.

CONCLUSIONS

In this work, an optimized rfGPT classifier for sub-Golgi protein type (cis and trans) identification was developed. The rfGPT classifier was derived from a random forests machine-learning algorithm, followed by implementation of the SMOTE to overcome a severe imbalance in the training data set and selection of optimal-related features using an ANOVA feature selection technique. The independent testing scores (ACC = 90.6%; MCC = 0.696; Sn = 96.1%; Sp = 69.2%) of the rfGPT ranked it as the one of the top sub-Golgi predictors. The feature importance analysis proves that the non-polar and aliphatic residues composition, the (aromatic residues) + (non-polar, aliphatic residues) dipeptide and aromatic residues composition for block between NH_2-termial and COOH-terminal of protein sequence are the top biological features, which play the key role for sub-Golgi proteins identification.

As compared with previous reported sub-Golgi protein classifiers, the rfGPT is with only dipeptide and amino acid residue composition features, which exempted sequence alignment from the procedure. Also, the rfGPT adopted random forests algorithm is easier for feature analysis and for revealing the key bio-factors of sub-Golgi protein classification. However, the rfGPT had an independent prediction accuracy (from a training data set with 25% sequence identity) that was 4.4% lower than that for the best of the reported sub-Golgi protein identifiers (based on the 40% sequence identity data set) and rfGPT uses more features.

The expectation is to build a more general data set of Golgi protein sequences to train the rfGPT model and to realize a more advanced sub-Golgi classifier of the features. In the future, extreme learning (Li et al., 2019) and deep learning (Long et al., 2017; Yu et al., 2018; Lv et al., 2019; Wei et al., 2019; Zhang Z.

TABLE 3 | Jackknife cross-validation and independent testing scores list for reported sub-Golgi protein classifiers.

No.	Classifier (Reference)	Data Set	Features	Dim	Jackknife cross-validation				Independent testing			
					ACC	MCC	Sn	Sp	ACC	MCC	Sn	Sp
1	IDMD (Ding et al., 2011)	D0	2-gapDC	400	74.7%	0.495	79.6%	69.6%	/	/	/	/
2	SVM (Ding et al., 2013)	D1	2-gapDC	83	85.4%	0.652	90.5%	90.5%	85.9%	0.578	90.2%	69.2%
3	SVM (Jiao and Du, 2016a)	D1	PSPCP	59	86.9%	0.684	92.6%	73.8%	/	/	90.2%	69.2%
4	SVM (Jiao and Du, 2016b)	D1	PSPCP	49	91.2%	0.793	99.0%	73.8%	87.1%	/	/	/
5	SVM (Lin et al., 2014)	D1	TPDC	501	97.1%	0.949	100%	92.9%	/	/	/	/
6	SVM (Rahman et al., 2018)	D2	ACC +DPDC +TPDC +2-gapDC +PseAAC	2800	95.9%	0.920	95.9%	92.6%	93.8%	0.85	98.0%	84.6%
7	KNN (Ahmad et al., 2017)	D2	PseAAC +3-gapDC +Bigram-PSSM	83	94.9%	0.90	97.2%	92.6%	94.8%	0.86	93.9%	94.0%
8	KNN (Ahmad and Hayat, 2019)	D2	SAAC +PSSM +3-gapDC	180	98.2%	0.96	98.6%	97.7%	94%	0.84	96.9%	81.5%
9	RF (Yang R. et al., 2016)	D2	3-gapDC +CSP-PSSMDC +CSP-BigramPSSM +CSP-EDPSSM	55	88.5%	0.765	88.9%	88%	93.8%	0.821	94.1%	92.3%
10	RF (this work)	D1	2-gapDC+SAAC	93	90.5%	0.811	92.6%	88.4%	90.6%	0.696	96.1%	69.2%

et al., 2019; Zou et al., 2019) methods will be tested on this problem.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://lin-group.cn/server/subGolgi2.

AUTHOR CONTRIBUTIONS

ZL and SJ were responsible for experiments and manuscripts preparation. HD participated in discussions. QZ worked as supervisor for all procedures.

FUNDING

This work was supported by the National Key R&D Program of China (2018YFC0910405), the Natural Science Foundation of China (No. 61922020, No. 61771331), and the 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/fbioe.2019.00215/full#supplementary-material

REFERENCES

Agrawal, P., Kumar, S., Singh, A., Raghava, G. P. S., and Singh, I. K. (2019). NeuroPIpred: a tool to predict, design and scan insect neuropeptides. *Sci. Rep.* 9:12. doi: 10.1038/s41598-019-41538-x

Ahmad, J., and Hayat, M. (2019). MFSC: multi-voting based feature selection for classification of Golgi proteins by adopting the general form of Chou's PseAAC components. *J. Theoret. Biol.* 463, 99–109. doi: 10.1016/j.jtbi.2018.12.017

Ahmad, J., Javed, F., and Hayat, M. (2017). Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods. *Artif. Intell. Med.* 78, 14–22. doi: 10.1016/j.artmed.2017.05.001

Akbar, S., Hayat, M., Kabir, M., and Iqbal, M. (2019). iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. *Lett. Organic Chem.* 16, 294–302. doi: 10.2174/1570178615666180816101653

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Bazan, B., Wiktor, M., Maszczak-Seneczko, D., Olczak, T., Kaczmarek, B., and Olczak, M. (2018). Lysine at position 329 within a C-terminal dilysine motif is crucial for the ER localization of human SLC35B4. *PLoS ONE* 13:e0207521. doi: 10.1371/journal.pone.0207521

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, W., Feng, P., Liu, T., and Jin, D. (2018). Recent advances in machine learning methods for predicting heat shock proteins. *Curr. Drug. Metab.* 20:224–228. doi: 10.2174/1389200219666181031105916

Chen, W., Lv, H., Nie, F., and Lin, H. (2019). i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. *Bioinformatics* 35, 2796–2800. doi: 10.1093/bioinformatics/btz015

Chou, K.-C., and Shen, H.-B. (2006). Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. *Biochem. Biophys. Res. Commun.* 347, 150–157. doi: 10.1016/j.bbrc.2006.06.059

Chou, K.-C., and Shen, H.-B. (2007). Recent progress in protein subcellular location prediction. *Analyt. Biochem.* 370, 1–16. doi: 10.1016/j.ab.2007.07.006

Chou, W.-C., Yin, Y., and Xu, Y. (2010). GolgiP: prediction of Golgi-resident proteins in plants. *Bioinformatics* 26, 2464–2465. doi: 10.1093/bioinformatics/btq446

Dao, F. Y., Lv, H., Wang, F., Feng, C. Q., Ding, H., Chen, W., et al. (2019). Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics* 35, 2075–2083. doi: 10.1093/bioinformatics/bty943

Ding, H., Guo, S.-H., Deng, E.-Z., Yuan, L.-F., Guo, F.-B., Huang, J., et al. (2013). Prediction of Golgi-resident protein types by using feature selection technique. *Chemometr. Intell. Lab. Syst.* 124, 9–13. doi: 10.1016/j.chemolab.2013.03.005

Ding, H., Liu, L., Guo, F.-B., Huang, J., and Lin, H. (2011). Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition. *Protein Peptide Lett.* 18, 58–63. doi: 10.2174/092986611794328708

Ding, Y., Tang, J., and Guo, F. (2016a). Identification of Protein–Protein Interactions via a novel matrix-based sequence representation model with amino acid contact information. *Int. J. Molecul. Sci.* 17:1623. doi: 10.3390/ijms17101623

Ding, Y., Tang, J., and Guo, F. (2016b). Predicting protein-protein interactions via multivariate mutual information of protein sequences. *BMC Bioinform.* 17:398. doi: 10.1186/s12859-016-1253-9

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418–419, 546–560. doi: 10.1016/j.ins.2017.08.045

Feng, C. Q., Zhang, Z. Y., Zhu, X. J., Lin, Y., Chen, W., Tang, H., et al. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. *Bioinformatics* 35, 1469–1477. doi: 10.1093/bioinformatics/bty827

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, 118–125. doi: 10.1016/j.ab.2013.05.024

Fujita, Y., Ohama, E., Takatama, M., Al-Sarraj, S., and Okamoto, K. (2006). Fragmentation of Golgi apparatus of nigral neurons with α-synuclein-positive inclusions in patients with Parkinson's disease. *Acta Neuropathol.* 112, 261–265. doi: 10.1007/s00401-006-0114-4

Gonatas, N. K., Gonatas, J. O., and Stieber, A. (1998). The involvement of the Golgi apparatus in the pathogenesis of amyotrophic lateral sclerosis, Alzheimer's disease, and ricin intoxication. *Histochem. Cell Biol.* 109, 591–600. doi: 10.1007/s004180050257

Gunther, T., Tulipano, G., Dournaud, P., Bousquet, C., Csaba, Z., Kreienkamp, H. J., et al. (2018). International union of basic and clinical pharmacology. CV. Somatostatin receptors: structure, function, ligands, and new nomenclature. *Pharmacol. Rev.* 70, 763–835. doi: 10.1124/pr.117.015388

Jiao, Y., Du, P., and Su, X. (2014). "Predicting Golgi-resident proteins in plants by incorporating N-terminal transmembrane domain information in the general form of Chou's pseudoamino acid compositions," in: *2014 8th International Conference on Systems Biology (ISB)* (Qingdao), 226–229.

Jiao, Y.-S., and Du, P.-F. (2016a). Predicting Golgi-resident protein types using pseudo amino acid compositions: approaches with positional specific physicochemical properties. *J. Theoret. Biol.* 391, 35–42. doi: 10.1016/j.jtbi.2015.11.009

Jiao, Y.-S., and Du, P.-F. (2016b). Prediction of Golgi-resident protein types using general form of Chou's pseudo-amino acid compositions: approaches with minimal redundancy maximal relevance feature selection. *J. Theoret. Biol.* 402, 38–44. doi: 10.1016/j.jtbi.2016.04.032

Kuang, L., Yu, L., Huang, L., Wang, Y., Ma, P., Li, C., et al. (2018). A personalized QoS prediction approach for CPS service recommendation based on reputation and location-aware collaborative filtering. *Sensors* 18:1556. doi: 10.3390/s18051556

Li, D., Luo, L., Zhang, W., Liu, F., and Luo, F. (2016). A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinform.* 17:329. doi: 10.1186/s12859-016-1206-3

Li, Y., Niu, M., and Zou, Q. (2019). ELM-MHC: an improved MHC identification method with extreme learning machine algorithm. *J. Proteome Res.* 18, 1392–1401. doi: 10.1021/acs.jproteome.9b00012

Lin, H., Ding, H., and Chen, W. (2014). Prediction of golgi-resident protein types using computational method. *Bentham Sci.* 1, 174–193. doi: 10.2174/9781608058624114010011

Lin, H., Liu, W. X., He, J., Liu, X. H., Ding, H., and Chen, W. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. *Sci. Rep.* 5:16964. doi: 10.1038/srep16964

Liu, B., Chen, J., Guo, M., and Wang, X. (2019). Protein remote homology detection and fold recognition based on Sequence-Order Frequency Matrix. *IEEE/ACM Transact. Comput. Biol. Bioinform.* 16, 292–300. doi: 10.1109/TCBB.2017.2765331

Liu, B., Fang, Y., Huang, D.-S., and Chou, K.-C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformaitcs* 34, 33–40. doi: 10.1093/bioinformatics/btx579

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71. doi: 10.1093/nar/gkv458

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the Cancer Cell Line Encyclopedia. *Sci. Rep.* 6:22811. doi: 10.1038/srep22811

Liu, Y., Wang, X., and Liu, B. (2019). A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings Bioinform.* 20, 330–346. doi: 10.1093/bib/bbx126

Long, H. X., Wang, M., and Fu, H. Y. (2017). Deep convolutional neural networks for predicting hydroxyproline in proteins. *Curr. Bioinform.* 12, 233–238. doi: 10.2174/1574893612666170221152848

Luo, L., Li, D., Zhang, W., Tu, S., Zhu, X., and Tian, G. (2016). Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. *PLoS ONE* 11:e0153268. doi: 10.1371/journal.pone.0153268

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein function prediction: from traditional classifier to deep learning. *Proteomics* 19:1900119. doi: 10.1002/pmic.201900119

Nikolovski, N., Shliaha, P. V., Gatto, L., Dupree, P., and Lilley, K. S. (2014). Label-free protein quantification for plant golgi protein localization and abundance. *Plant Physiol.* 166, 1033–1043. doi: 10.1104/pp.114.245589

Pan, Y., Liu, D., and Deng, L. (2017). Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. *PLoS ONE* 12:e0179314. doi: 10.1371/journal.pone.0179314

Pan, Y. W., Zixiang Z. 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

Paulson, J. C., and Colley, K. J. (1989). Glycosyltransferases. Structure, localization, and control of cell type-specific glycosylation. *J Biol Chem.* 264, 17615–17618.

Pfeffer, S. R. (2001). Constructing a Golgi complex. *J. Cell Biol.* 155, 873–875. doi: 10.1083/jcb.200109095

Rahman, M. S., Rahman, M. K., Kaykobad, M., and Rahman, M. S. (2018). isGPT: an optimized model to identify sub-Golgi protein types using SVM and Random Forest based feature selection. *Artif. Intell. Med.* 84, 90–100. doi: 10.1016/j.artmed.2017.11.003

Rhee, S. W., Starr, T., Forsten-Williams, K., and Storrie, B. (2005). The steady-state distribution of glycosyltransferases between the golgi apparatus and the endoplasmic reticulum is approximately 90:10. *Traffic* 6, 978–990. doi: 10.1111/j.1600-0854.2005.00333.x

Ru, X. Q., Li, L. H., and Zou, Q. (2019). Incorporating distance-based top-n-gram and random forest to identify electron transport proteins. *J. Proteome Res.* 18, 2931–2939. doi: 10.1021/acs.jproteome.9b00250

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

Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2018). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE/ACM Transact. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/TCBB.2018.2858756

Tang, G., Shi, J., Wu, W., Yue, X., and Zhang, W. (2018). Sequence-based bacterial small RNAs prediction using ensemble learning strategies. *BMC Bioinform.* 19:503. doi: 10.1186/s12859-018-2535-1

Teasdale, R. D., and Yuan, Z. (2002). Prediction of Golgi Type II membrane proteins based on their transmembrane domains. *Bioinformatics* 18, 1109–1115. doi: 10.1093/bioinformatics/18.8.1109

Van Dijk, A. D. J., Van Der Krol, A. R., Ter Braak, C. J. F., Bosch, D., and Van Ham, R. C. H. J. (2008). Predicting sub-Golgi localization of type II membrane proteins. *Bioinformatics* 24, 1779–1786. doi: 10.1093/bioinformatics/btn309

Wan, S., Duan, Y., and Zou, Q. (2017). HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* 17:1700262. doi: 10.1002/pmic.201700262

Wang, H., Liu, C., and Deng, L. (2018a). Enhanced prediction of hot spots at protein-protein interfaces using extreme gradient boosting. *Sci. Rep.* 8:14285. doi: 10.1038/s41598-018-32511-1

Wang, S. P., Zhang, Q., Lu, J., and Cai, Y. D. (2018b). Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. *Curr. Bioinform.* 13, 3–13. doi: 10.2174/1574893611666160608075753

Wang, X., Li, H., Gao, P., Liu, Y., and Zeng, W. (2019). Combining support vector machine with dual g-gap dipeptides to discriminate between acidic and alkaline enzymes. *Lett. Organic Chem.* 16, 325–331. doi: 10.2174/1570178615666180925125912

Wang, Y., Ding, Y., Guo, F., Wei, L., and Tang, J. (2017). Improved detection of DNA-binding proteins via compression technology on PSSM information. *PLoS ONE* 12:e0185587. doi: 10.1371/journal.pone.0185587

Wei, L., Su, R., Wang, B., Li, X., Zou, Q., and Gao, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N 6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082

Wei, L., Wan, S., Guo, J., and Wong, K. K. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wuritu, Y., Xiao-Juan, Z., Jian, H., Hui, D., and Hao, L. (2019). A brief survey of machine learning methods in protein sub-golgi localization. *Curr. Bioinform.* 14, 234–240. doi: 10.2174/1574893613666181113131415

Xu, D., and Esko, J. D. (2009). A Golgi-on-a-chip for glycan synthesis. *Nat. Chem. Biol.* 5:612. doi: 10.1038/nchembio0909-612

Xu, L., Liang, G., Wang, L., and Liao, C. (2018). A novel hybrid sequence-based model for identifying anticancer peptides. *Genes* 9:158. doi: 10.3390/genes9030158

Yang, J., Grunewald, S., Xu, Y., and Wan, X. F. (2014). Quartet-based methods to reconstruct phylogenetic networks. *BMC Syst. Biol.* 8:21. doi: 10.1186/1752-0509-8-21

Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., et al. (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* 5:15145. doi: 10.1038/srep15145

Yang, J., Huang, T., Song, W. M., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016). Discover the network underlying the connections between aging and age-related diseases. *Sci. Rep.* 6:32566. doi: 10.1038/srep32566

Yang, R., Zhang, C., Gao, R., and Zhang, L. (2016). A novel feature extraction method with feature selection to identify golgi-resident protein types from imbalanced data. *Int. J. Molecul. Sci.* 17:218. doi: 10.3390/ijms17020218

Yu, L., Ma, X., Zhang, L., Zhang, J., and Gao, L. (2016). Prediction of new drug indications based on clinical data and network modularity. *Sci. Rep.* 6:32530. doi: 10.1038/srep32530

Yu, L., Su, R., Wang, B., Zhang, L., Zou, Y., Zhang, J., et al. (2017). Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *Ieee-Acm Transact. Comput. Biol. Bioinform.* 14, 966–977. doi: 10.1109/TCBB.2016.2550453

Yu, L., Sun, X., Tian, S. W., Shi, X. Y., and Yan, Y. L. (2018). Drug and nondrug classification based on deep learning with various feature selection strategies. *Curr. Bioinform.* 13, 253–259. doi: 10.2174/1574893612666170125124538

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2016). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Transact. Computat. Biol. Bioinform.* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Transact. Comput. Biol. Bioinform.* 14, 687–695.

Zeng, X., Yuan, S., Huang, X., and Zou, Q. (2015). Identification of cytokine via an improved genetic algorithm. *Front. Comput. Sci.* 9, 643–651. doi: 10.1007/s11704-014-4089-3

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

Zhang, S., Zhang, T., and Liu, C. (2019). Prediction of apoptosis protein subcellular localization via heterogeneous features and hierarchical extreme learning machine. *Sar Qsar Environ. Res.* 30, 209–228. doi: 10.1080/1062936X.2019.1576222

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018a). The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.neucom.2017.07.065

Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018b). SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. *PLoS Comput. Biol.* 14:e1006616. doi: 10.1371/journal.pcbi.1006616

Zhang, X., Zou, Q., Rodriguez-Paton, A., Zeng, C. B. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *IEEE/ACM Transact. Comput. Biol. Bioinform.* 16, 283–291. doi: 10.1109/TCBB.2017.2776280

Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q., et al. (2019). Deep learning in omics: a survey and guideline. *Brief. Funct. Genom.* 18, 41–57. doi: 10.1093/bfgp/ely030

Zheng, N., Wang, K., Zhan, W., and Deng, L. (2019). Targeting virus-host protein interactions: feature extraction and machine learning approaches. *Curr. Drug Metabol.* 20, 177–184. doi: 10.2174/1389200219666180829121038

Zhu, P. F., Xu, Q., Hu, Q. H., Zhang, C. Q., and Zhao, H. (2018). Multi-label feature selection with missing labels. *Pattern Recogn.* 74, 488–502. doi: 10.1016/j.patcog.2017.09.036

Zhu, P. F., Zhu, W. C., Hu, Q. H., Zhang, C. Q., and Zuo, W. M. (2017). Subspace clustering guided unsupervised feature selection. *Pattern Recogn.* 66, 364–374. doi: 10.1016/j.patcog.2017.01.016

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

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). 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., Zeng, J., Cao, L., and Ji, R. (2016b). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Conflict of Interest 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: 04 September 2019
doi: 10.3389/fgene.2019.00769

Inferring Latent Disease-lncRNA Associations by Faster Matrix Completion on a Heterogeneous Network

Wen Li[1], *Shulin Wang*[1]*, *Junlin Xu*[1], *Guo Mao*[1], *Geng Tian*[2] and *Jialiang Yang*[2]*

[1] *College of Computer Science and Electronic Engineering, Hunan University, Changsha, China,* [2] *Geneis Beijing Co., Ltd., Beijing, China*

OPEN ACCESS

Edited by:
Peilin Jia,
University of Texas Health Science Center, United States

Reviewed by:
Xin Zhou,
Stanford University, United States
Lu Zhang,
Hong Kong Baptist University, Hong Kong

***Correspondence:**
Shulin Wang
smartforesting@163.com
Jialiang Yang
yangjl@geneis.cn

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

Received: *21 May 2019*
Accepted: *19 July 2019*
Published: *04 September 2019*

Citation:
Li W, Wang S, Xu J, Mao G, Tian G and Yang J (2019) Inferring Latent Disease-lncRNA Associations by Faster Matrix Completion on a Heterogeneous Network. Front. Genet. 10:769. doi: 10.3389/fgene.2019.00769

Current studies have shown that long non-coding RNAs (lncRNAs) play a crucial role in a variety of fundamental biological processes related to complex human diseases. The prediction of latent disease-lncRNA associations can help to understand the pathogenesis of complex human diseases at the level of lncRNA, which also contributes to the detection of disease biomarkers, and the diagnosis, treatment, prognosis and prevention of disease. Nevertheless, it is still a challenging and urgent task to accurately identify latent disease-lncRNA association. Discovering latent links on the basis of biological experiments is time-consuming and wasteful, necessitating the development of computational prediction models. In this study, a computational prediction model has been remodeled as a matrix completion framework of the recommendation system by completing the unknown items in the rating matrix. A novel method named faster randomized matrix completion for latent disease-lncRNA association prediction (FRMCLDA) has been proposed by virtue of improved randomized partial SVD (rSVD-BKI) on a heterogeneous bilayer network. First, the correlated data source and experimentally validated information of diseases and lncRNAs are integrated to construct a heterogeneous bilayer network. Next, the integrated heterogeneous bilayer network can be formalized as a comprehensive adjacency matrix which includes lncRNA similarity matrix, disease similarity matrix, and disease-lncRNA association matrix where the uncertain disease-lncRNA associations are referred to as blank items. Then, a matrix approximate to the original adjacency matrix has been designed with predicted scores to retrieve the blank items. The construction of the approximate matrix could be equivalently resolved by the nuclear norm minimization. Finally, a faster singular value thresholding algorithm with a randomized partial SVD combing a new sub-space reuse technique has been utilized to complete the adjacency matrix. The results of leave-one-out cross-validation (LOOCV) experiments and 5-fold cross-validation (5-fold CV) experiments on three different benchmark databases have confirmed the availability and adaptability of FRMCLDA in inferring latent relationships of disease-lncRNA pairs, and in inferring lncRNAs correlated with novel diseases without any prior interaction information. Additionally, case studies have shown that FRMCLDA is able to effectively predict latent lncRNAs correlated with three widespread malignancies: prostate cancer, colon cancer, and gastric cancer.

Keywords: heterogeneous bilayer network, association prediction, matrix completion, faster SVT, randomized partial SVD, similarity measurements

INTRODUCTION

Long non-coding RNAs are RNA molecules whose transcripts are not less than 200 nucleotides, including intronic/exonic lncRNAs, antisense lncRNAs, overlapping lncRNA and long intergenic ncRNAs (lincRNAs). LncRNAs have long been considered as transcriptional noise, because of their absence in encoding proteins. Recently, it has been found that some lncRNAs regulate the expression of target genes after transcription, whose malfunction may lead to a number of diseases. For example, abnormal lncRNA expression may be involved in certain stages of cancer progression, which can serve as a potential biomarker for early tumor diagnosis (Zhou et al., 2015; Niknafs et al., 2016). In addition, lncRNAs are found able to interact with signaling pathways involved in the pathology of malignancy (Bian et al., 2015). However, studies on the prediction of relationships between lncRNAs and diseases are still limited in number. One key bottleneck is the high cost and labor-intensity of laboratory techniques in discovering the relationships between lncRNAs and diseases. To break the bottleneck, a lot of computational models have been proposed which can generally be divided into two major categories depending on the source of the interaction data: models for single-interaction data sources and models for multi-interaction data sources.

In the first major category, models for single-interaction data sources are based on diseases-lncRNAs interaction (association/link) data, which is unique known interaction information. According to its method, the model can be divided into two minor branches. The first minor branch is composed of machine-learning based models, in which the prediction of latent disease-lncRNA association takes experimentally validated disease-lncRNA associations as labeled data (training set) and unknown associations as unlabeled samples (invalidated relationship information). For example, a method named Laplacian Regularized Least Squares (LRLSLDA) was first proposed by Chen et al. to infer disease-lncRNA associations with a semi-supervised learning model (Chen et al., 2013). It is assumed that diseases with high semantic similarity are more likely to interact with lncRNAs with high functional similarity. LRLSLDA effectively predicts latent associations without negative samples, but it is difficult to select appropriate parameters and classifiers that optimize similarity measures for both lncRNAs and diseases. Inspired by the recommendation system, the authors consider disease-lncRNA association prediction as a recommendation task. A computational model named SIMCLDA is designed to predict latent disease-lncRNA relationships, taking advantage of the inductive matrix completion (IMC) method (Lu et al., 2018). The main idea of SIMCLDA is to extract informative feature vectors of lncRNAs and diseases to complete the association matrix. It is able to discover more accurate primary feature vectors and predict associations for novel lncRNAs and diseases.

Additionally, the second minor branch is composed of network-based models, random walk and a variety of propagation algorithms implemented on a heterogeneous network to infer latent disease-lncRNA associations. The heterogeneous network is constructed by integrating lncRNA-disease interaction network, disease similarity network and lncRNA similarity network. For instance, based on the hypothesis that functional lncRNAs are associated with diseases with similar phenotypes, a lncRNA functional similarity network (LFSN) is constructed and a novel computational framework RWRlncD is proposed for predicting latent disease-lncRNA associations through random walk with restart (Sun et al., 2014). However, the method of RWRlncD fails to infer related lncRNAs for novel diseases without prior interaction. Gu et al. put forward a global network-based random walk where negative samples are not required to predict latent disease-lncRNA relationships (Gu et al., 2017). Although this method can predict relationships related to isolated diseases or lncRNAs, it is prone to biased prediction. A computational method called BPLLDA is brought forward in a heterogeneous network on a basis of simple paths with finite length (Xiao et al., 2018). However, BPLLDA also has some limitations such as biased predictions. And the simplistic distance-decay function has yet to be improved by machine learning.

Due to the fact that known experimentally validated disease-lncRNA interactions are still rare, the second major category of computational models is models based on multi-interaction data sources proposed for association prediction. Multi-interaction data sources, such as lncRNA-gene interaction, lncRNAs-miRNAs interaction, disease-gene interaction and miRNA-disease interaction, are also included to infer latent disease-lncRNA associations. For example, a TPGLDA method is proposed to identify the underlying relationships by a tripartite graph of disease-lncRNA-gene and to develop an efficient resource allocation algorithm in the graph (Ding et al., 2018). TPGLDA effectively reduces the biased prediction in the resource allocation process, but it focuses on an unweighted tripartite graph and its accuracy for prediction needs to be improved. In this category, the weights of heterogeneous interaction data are difficult to determine, so the fusion of heterogeneous data is a challenging task.

Inferring latent disease-lncRNA association can also be modeled as a recommendation system which recommends top-ranked lncRNAs for given diseases. Based on matrix completion, the establishment of the disease-lncRNA recommendation system aims to complete unknown terms in the association matrix according to its ranked scores. Similar to the hypothesis in the user-item recommendation system that users with similar behaviors prefer similar items, the prediction of disease-lncRNA association assumes that phenotypically similar diseases tend to interact with functionally similar lncRNAs (Chen and Yan, 2013). In our study, we assume that semantically or functionally similar diseases tend to interact with similar lncRNAs (similar in sequence, expression profiles or function). Thus, theoretically the integration of lncRNA-lncRNA and disease-disease interaction will benefit the prediction of lncRNA-disease associations. Based on that, we proposed a computational model FRMCLDA similar to the recommendation system to infer latent associated lncRNAs for queried diseases, and solved it with faster randomized partial matrix completion (fSVT) algorithm (Feng et al., 2018). FRMCLDA consolidated disease integrated similarity network, lncRNA integrated similarity network and known disease-lncRNA interaction network to construct a heterogeneous bilayer network. Then a randomized SVD technique incorporating the block Krylov-subspace iteration (BKI) scheme (named rSVD-BKI algorithm) was proposed to complete large-scale matrix. In addition, a novel subspace reuse technique was integrated to accelerate matrix completion. Our method is based on semi-supervised machine learning, which does not need the information of negative samples. So, it generally belongs to the first category.

Our work main contributions are threefold: first, the integrated similarities for diseases and lncRNAs were properly calculated by different methods dealing with different types of data sources. The ratio of cosine similarity in integrated similarity was better determined by learning, which extracted similarity information based on known disease-lncRNA interaction. Therefore, FRMCLDA was able to offset biased predictions by similarity integration which was not entirely dependent on known interaction. Second, diseases and lncRNAs were mapped into the same network by constructing a heterogeneous bilayer network. FRMCLDA completed the disease-lncRNA interaction matrix by completing the adjacency matrix of the large-scale heterogeneous network. Thus, the similar information was included in the association prediction. Third, we took advantage of an effective fSVT algorithm which adopted rSVD-BKI and a novel subspace reuse technique to expeditiously approximate the dominant singular values and homogeneous singular vectors in an adaptive manner. Hence, the recommendation system could be extended for comprehensive adjacency matrices of heterogeneous bilayer networks.

For evaluating the performance of our method, cross validation experiments were performed on three benchmark databases, Dataset 1, Dataset 2 and Dataset 3. FRMCLDA obtained reliable AUCs of 0.92068, 0.91224 in global LOOCV and local LOOCV respectively in Dataset1, at least 5% higher than other comparison models. In Dataset 2 and Dataset 3, FRMCLDA achieved an AUC of 0.9182 and 0.8999 by global 5-fold CV, higher than other comparison methods. In addition, a case study on inferring latent lncRNAs associated with prostate cancer, colon cancer and gastric cancer in Dataset 3 were performed. In terms of the results, 16, 15, 16 out of the top 20 predicted lncRNAs associated with prostate cancer, colon cancer and gastric cancer respectively were confirmed by recent literature and public databases. The results show that FRMCLDA is able to effectively infer the associations between diseases and lncRNAs with higher accuracy than the other existing models.

MATERIALS AND METHODS

We denote a disease-lncRNA association matrix as $DL \in \mathbb{R}^{m \times n}$, the rows of which represent diseases and columns represent lncRNAs. The variable m is the number of diseases, and n is the number of lncRNAs. If disease d_i is associated with lncRNA l_j, the value of $DL(i, j)$ in the association matrix is 1. And if the link between d_i and l_j is unknown or uncertain, $DL(i, j)$ is 0. It is noted that the unlinked evidence between d_i and l_j is difficult to obtain. The known experimentally validated disease-lncRNA links can be retrieved from the public association database, based on which disease-lncRNA interaction matrix *DL* is established. If the number of nonzero elements is far smaller than that of zero elements, and the distribution of nonzero elements in the matrix is irregular, the matrix will be called sparse matrix. Generally, matrix *DL* is a sparse matrix, because, due to the insufficient number of studies, there are much more unknown associations (value 0) in matrix *DL* than known ones (value 1) (see **Table 1**). *LS* and *DS* denote lncRNA integrated similarity matrix and disease integrated similarity matrix respectively, which can be calculated through various biological data. It is assumed that the underlying determinants of the disease-lncRNA associations are closely related. Hence the number of independent factors is less than the number of lncRNAs or diseases. Accordingly, the matrix of known disease-lncRNA association is of low rank. That assumed, matrix completion can recover the unknown items of the disease-lncRNA interaction matrix by constructing a low-rank matrix which aims to approximates adjacency to matrix A.

TABLE 1 | Details of three benchmark datasets.

Datasets	Number of known associations	Number of lncRNAs	Number of diseases	Sparsity of the matrix DL	Weights in integrated Similarity
Dataset 1	352	156	190	$1.187*10^{-2}$	$w_l = 0.7$, $w_d = 0.9$
Dataset 2	540	115	178	$2.638*10^{-2}$	$w_l = 0.5$, $w_d = 0.7$
Dataset 3	621	258	226	$1.065*10^{-2}$	$w_l = 0.5$, $w_d = 0.5$

The sparsity is calculated by the ratio of existed known association number to the size of the matrix (all the possible association number).

Overview

In our work, a new method called FRMCLDA is proposed to infer latent disease-lncRNA associations on the basis of fast matrix completion. The mechanism of FRMCLDA is shown in **Figure 1**. Firstly, we obtaine known disease-lncRNA associations from the public databases. Secondly, we calculate the disease similarity and lncRNA similarity with different methods. Next, we construct a heterogeneous bilayer network with three networks, i.e., a disease similarity network, a lncRNA similarity network and a disease-lncRNA interaction network. Furthermore, we implement a faster matrix completion algorithm with an improved randomized partial SVD and a sub-space reuse technique to restore the adjacency matrix of heterogeneous bilayer network. Finally, we infer potential disease-lncRNA associations through the predicted scores.

Datasets and Data Preprocessing

All the known diseases-lncRNA interactions were obtained from three gold standard databases in three benchmark datasets respectively: MNDR database, Lnc2Cancer database and LncRNADisease database (Chen et al., 2013; Wang et al., 2013; Ning et al., 2016).

The known associations between lncRNA and disease in Dataset 1 were retrieved from the MNDR database in 2015. After removing all the duplicate records of lncRNAs and diseases, and what do not belong to human beings, and correcting the names of the lncRNAs (according to LncRNAdb, Lncipedia, NCBI and HGNC) and diseases (according to UMIS, MeSH and NCBI), we finalized 352 disease-lncRNA associations, including 156 lncRNAs and 190 diseases.

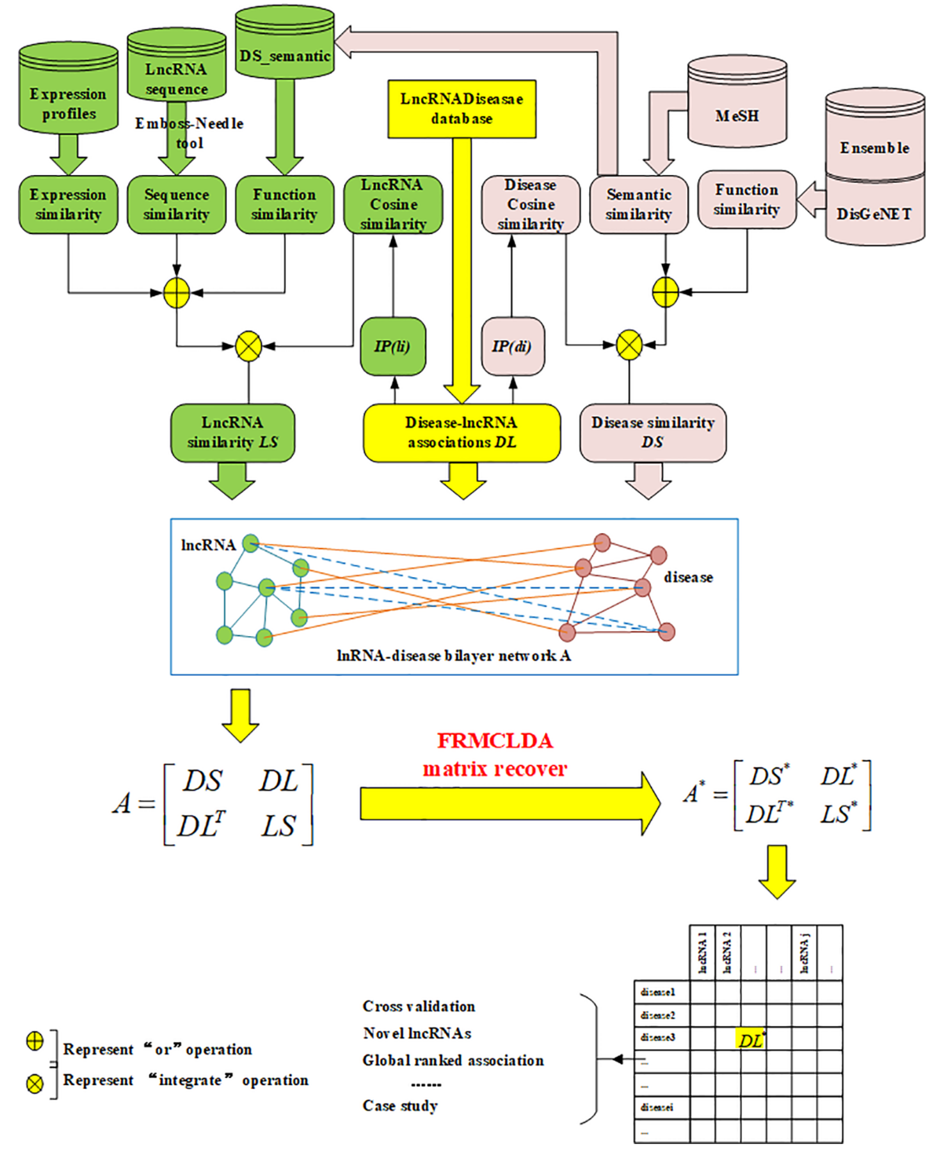


FIGURE 1 | Scheme of FRMCLDA to infer latent disease-related lncRNAs by matrix recovery.

In Dataset 2, the known associations between lncRNA and disease were obtained from the Lnc2Cancer database in 2016. After eliminating the duplicate associations on account of different evidences, we obtained 540 distinct known disease-lncRNA associations, including 115 lncRNAs and 178 diseases.

In Dataset 3, the known disease-lncRNA associations were downloaded from the manually curated LncRNADisease database (http://cmbi.bjmu.edu.cn/lncrnadisease) in 2015. In the same way as data preprocessing, we downloaded 621 known disease-lncRNA associations, including 248 lncRNAs and 226 diseases. The details of the three datasets are shown in **Table 1**.

Similarity Calculation

Diseases Similarity

In the three benchmark datasets, disease integrated similarities were calculated with three different similarity data sources.

a) *Disease semantic similarity:* In previous studies by Chen et al., a graph of directed acyclic (DAG) is utilized to label a disease, which includes overall relevant annotation labels acquired from the U.S. National Library of Medicine (MeSH) (Cai et al., 2008; Chen and Yan, 2013). It is assumed that diseases sharing larger common DAGs areas might have higher similarity scores. Therefore, the semantic similarity of diseases denoted as *DS_semantic* was calculated on the basis of DAG values by DOSim. DOSim is a package of R language for the semantic similarity calculation based on disease ontology (Wang et al., 2007).

b) *Disease functional similarity:* Disease functional similarity was calculated using the Jaccard similarity coefficient on account of gene-gene ontology relationships and disease-gene relationship, as reported in previous studies (Pinero et al., 2017; Lu et al., 2018). Disease functional similarity is denoted as *DS_jaccard*(d_i, d_j), and can be calculated by formula (1):

$$DS\_jaccard(d_i,d_j)=\frac{\left|GO_{d_i}\bigcap GO_{d_j}\right|}{\left|GO_{d_i}\bigcup GO_{d_j}\right|} \tag{1}$$

where $\boldsymbol{GO_{di}}$ represents the gene ontology terms related to disease $\boldsymbol{d_i}$, and the symbol |·| represents the number of items in a set.

c) *Disease cosine similarity:* Widely used in information retrieval and data mining, the cosine similarity is a popular method for calculating the similarity as the cosine of the angle between vectors. Here we used cosine similarity to extract disease feature information from the known interaction matrix *DL*. The disease cosine similarity denoted as *DS*_cos*ine*(d_i, d_j)can be calculated by formula (2):

$$DS\_\text{cos}ine(d_i,d_j)=\frac{IP(d_i)\cdot IP(d_j)}{\left\|IP(d_i)\right\|\left\|IP(d_j)\right\|} \tag{2}$$

where $IP(d_i)$ is the interaction profile of disease d_i, the *i-th* row vector of the interaction matrix *DL*. If disease d_i is associated with lncRNA l_k, the k-th element in $IP(d_i)$ is 1, otherwise the value is 0. The value 0 does not mean that association does not exist but means it is uncertain. $||IP(d_i)||$ is the 2-norm of $IP(d_i)$.

d) *Integrated disease similarity DS:* To illustrate the adaptability of our model to different similarity data, we adopted two different integrated disease similarities in three benchmark datasets. In Dataset 1 and Dataset 2, *DS* was calculated by: $DS = w_{d1} * DS\_semantic + (1 - w_{d1}) * DS\_\cos ine$ In Dataset 3, *DS* was calculated by: $DS = w_{d2} * DS\_jaccard + (1\text{-}w_{d2}) * DS\_\cos ine$.

LncRNAs Similarity

To calculate lncRNA integrated similarity, we adopted four different sources of similarity data: lncRNA sequence similarity, lncRNA expression similarity, lncRNA functional similarity and lncRNA cosine similarity.

a) *LncRNA sequence similarity:* Most of the RNA sequences of lncRNAs were downloaded mainly from the database LncRNADisease (http://www.cuilab.cn/lncrnadisease). The sequences not available in LncRNADisease were retrieved from the databases UCSC and LNCipedia. The sequence similarity between two lncRNAs were calculated with Needleman-Wunsch global alignment algorithm (Emboss-Needle tool) (Needleman and Wunsch, 1970; Rose and Eisenmenger, 1991). We set the parameters to default values. The Matrix file name was set to EDNAfull for nucleic, Gap opening penalty was set to 10 and Gap extension penalty was set to 0.5 for any sequences. LncRNA sequence similarity is defined as formula (3):

$$LS\_seq(l_i,l_j)=\frac{SW(l_i,l_j)}{\sqrt{SW(l_i,l_i)\cdot SW(l_j,l_j)}} \tag{3}$$

where $SW(l_i, l_j)$ is the alignment score calculated by Emboss-Needle, which is equal to the sum of the matches taken from the scoring matrix, minus penalties arising from opening and extending gaps in the aligned sequences.

b) *LncRNA expression similarity:* The expression profiles of lncRNA can be obtained from the dataset E-MTAB-513 in ArrayExpress (Parkinson et al., 2007; Derrien et al., 2012). Based on the previous literature, we normalized these expression data and calculated the lncRNA expression similarity *LS*_exp with the absolute Spearman correlation coefficient (Chen and Yan, 2013).

c) *LncRNA functional similarity:* Based on an accepted assumption that lncRNAs with similar functions have similar interaction patterns to those of diseases, the functional similarity of lncRNA can be obtained *via* computation of disease semantic similarity from a previous study by Sun et al., (2014). It is supposed that lncRNA l_i is correlated with a set of diseases $D_i = \{d_{i1}, d_{i2},\ldots, \mathrm{d}_{\mathrm{im}}\}$, and lncRNA l_j is correlated with a set of diseases $Dj = \{d_{j1}, d_{j2},\ldots d_{jn}\}$. Semantic similarity between d_{il} and Dj is calculated as formula (4):

$$DS\_semantic(d_{il}, D_j) = \max_{d \in D_j}(DS\_semantic(d_{il}, d)) \quad (4)$$

And then the functional similarity of lncRNAs can be computed as formula (5):

$$LS\_func(l_i, l_j) = \frac{\sum_{1\le l\le m} DS\_semantic(d_{il}, D_j) + \sum_{1\le k\le n} DS\_semantic(d_{jk}, D_i)}{m+n} \quad (5)$$

where $LS\_func(l_i, l_j)$ denotes the functional similarity between lncRNA l_i and lncRNA l_j.

d) *LncRNA cosine similarity:* In the same way, we used cosine similarity to extract lncRNA feature information from the known interaction matrix *DL*. lncRNA cosine similarity can be calculated as formula (6):

$$LS\_\cos ine(l_i, l_j) = \frac{IP(l_i)\cdot IP(l_j)}{\|IP(l_i)\|\|IP(l_j)\|} \quad (6)$$

where $IP(l_j)$ is resulted from the *j-th* column of the interaction matrix *DL*. $IP(l_j)$ is a vector which denotes the feature vector for lncRNA l_j.

e) *Integrated lncRNA similarity LS:* In three benchmark datasets, we adopted three different similarity computation methods to fully demonstrate the robustness of FRMCLDA. In Dataset 1, integrated lncRNA similarity *LS* was calculated as: $LS = w_{l1} * LS\_func + (1 - w_{l1}) * LS\_\cos ine$. In Dataset 2, *LS* was calculated as: $LS = w_{l2} * LS\_\exp + (1 - w_{l2}) * LS\_\cos ine$. In Dataset 3, *LS* was calculated as: $LS = w_{l3} * LS\_seq + (1 - w_{l3}) * LS\_\cos ine$.

Construction of the Heterogeneous Bilayer Network

Based on the integrated disease and lncRNA similarity matrices *DS* and *LS* calculated above, disease similarity network and lncRNA similarity network can be constructed. Let $D = \{d_1, d_2,\ldots, d_n\}$ represent the set of n diseases in the disease similarity network. The edge between disease d_i and d_j is weighted by integrated disease similarity $DS(i, j)$. Let $L = \{l_1, l_2,\ldots, l_m\}$ represent the set of m lncRNAs in the lncRNA similarity network. The edge between lncRNA l_i and l_j is weighted by integrated lncRNA similarity $LS(i, j)$. Besides, the disease-lncRNA interaction network can be modeled as $G(V, E)$, where $V(G) = \{D,L\}$, $E(G) \subseteq D \times L$, $E(G) = \{e_{ij}$, edge between disease d_i and lncRNA $l_j\}$. The edge e_{ij} is initialized to 1, if there exists a known link between disease d_i and lncRNA l_j, otherwise, e_{ij} is initialized to 0. *DL* is the adjacency matrix for the disease-lncRNA interaction network.

Finally, a heterogeneous bilayer network is constructed by connecting disease similarity network and lncRNA similarity network *via* disease-lncRNA association network, as shown in **Figure 1**. Accordingly, the adjacency matrix A of the heterogeneous bilayer network is defined as formula (7):

$$A = \begin{bmatrix} DS & DL \\ DL^T & LS \end{bmatrix} \quad (7)$$

where diagonal sub-matrices *DS* and *LS* are the adjacency matrix of the disease similarity network and the lncRNA similarity network. The off-diagonal sub-matrix *DL* is the adjacency matrix for the disease-lncRNA interaction network, DL^T is the transpose of *DL*. Usually, the interaction between lncRNAs and diseases is mutual, and values of the matrix *DL* are nonnegative, therefore the adjacent matrix A is meristic and positive semi-definite. The singular values of the adjacent matrix A are nonnegative real numbers and equivalent to the eigenvalues. In conclusion, the prediction of disease-lncRNA association can be remodeled as the matrix completion of the adjacency matrix A. If matrix A is only comprised of matrix *DL*, rather than the large-scale matrix of the heterogeneous network, then the completion based on rank minimization will not generate significant results. That is because all known disease-lncRNA associations are positive in matrix. Only restoring the matrix *DL* will result in an optimized solution to rank minimization problem, i.e., all-one matrix with rank 1.

Inferring Latent Associations by Faster Randomized Matrix Completion

Our goal is to restore the unknown entries of the adjacent matrix A by constructing a proximate matrix A^* with the same size $(m + n) \times (m + n)$. It is assumed that A have rank $r(r \ll (m + n))$. Φ is denoted as an index set for all the known entries of matrix A. The problem of matrix completion can be converted to solving the rank minimization problem by formula (8):

$$\min rank(A^*)$$

$$s.t. P_\Phi(A^*) = P_\Phi(A) \quad (8)$$

in which $P_\Phi(A)$ is denoted as an orthogonal projector onto the span of matrix A. Its value is 0 when the element (i, j) is not in the set Φ. Matrix *A* is the adjacency matrix of the heterogeneous network constructed in *Datasets and Data Processing*. However, the problem of rank minimization is generally considered as a NP-hard problem (Natarajan, 1995). An approach called relaxed convex optimization is widely used by minimizing the nuclear norm ($\|\cdot\|_*$) of the matrix, which is known to be solved by standard singular value threshold (SVT) algorithm (Candès and Recht, 2009). Therefore, the matrix completion can be resolved by a proximal optimization solution (Cai et al., 2008). Minimization of the nuclear norm can be resolved by formula (9):

$$\min \left\| A^{*} \right\|_{*}$$

$$s.t.P_{\Phi}(A^{*}) = P_{\Phi}(A) \tag{9}$$

Equation (9) can be solved by the iterative processes in formula (10) and (11):

$$\begin{cases} X^{(i)} = shrink(Y^{(i-1)}, \tau) \\ Y^{(i)} = Y^{(i-1)} + \delta P_{\Phi}(A - X^{(i)}) \end{cases} \tag{10}$$

$$shrink(Y^{(i)}, \tau) = \sum_{j=1}^{\sigma_j^{(i)} \geq \tau} (\sigma_j^{(i)} - \tau) u_j^{(i)} v_j^{(i)T} \tag{11}$$

where the function $shrink(Y^{(i)}, \tau)$ is a soft thresholding operator that computes the singular value of the matrix Y at level τ (Aken et al., 2016). δ is the iteration step length. $\sigma_j^{(i)}$ is one of the singular values of Yat the ith iteration. $u_j^{(i)}$ and $v_j^{(i)}$ are the corresponding left singular vector and right singular vector respectively. $Y^{(i)}$ is usually a relatively large matrix with high sparsity, and usually can be stored with a sparse matrix. Starting computation from $Y^{(0)}$=cδP_{Φ}(A), a series of $X^{(i)}$ and $Y_{(i)}$ can be generated through the linearized Bregman iteration.

In FRMCLDA, δ is set to $(\mathrm{m}+\mathrm{n})/\sqrt{|\Phi|}$ as assigned in previous literature (Li and Yu, 2017). We set $c = \left\lceil \tau / (\delta \left\| p_{\Phi}(A) \right\|) \right\rceil$, $\tau = \| P_{\phi}(A) \|_{F}\ (m+n)/\sqrt{|\Phi|}$ to balance the accuracy of approximation against the speed of convergence as suggested by Candès et al. (Candès and Recht, 2009). Although the SVT algorithm has high accuracy for both symmetric and real data matrix, the costs are large when executing SVD repeatedly at each iteration. So many improved methods like truncated singular value decomposition and randomized SVD have been proposed for accelerating SVT by keeping the cost of shrink (·) low throughout the iteration (Halko et al., 2010). In this study, FRMCLDA adopts a faster SVT algorithm, fSVT, based on partial and improved randomized SVD which exploits a sub-space reuse technique to extract key singular value and corresponding singular vector. The main concept of randomization is to determine the sub-spaces for obtaining dominant information and ignore insignificant information by random projection. Randomized SVD algorithms execute as many or fewer floating-point operations (*flops*) with the runtime benefit. Faster matrix completion even incorporates a block Krylov sub-space iteration rSVD-BKIr scheme and a novel sub-space reuse mechanism (reuse the orthogonal basis Q from the last round of iteration) (Musco and Musco, 2015). The fast matrix completion algorithm with rSVD-BKI has been proven the have the same reliability and accuracy as the original singular value thresholding algorithm, while at higher speed for large data matrix completion (Feng et al., 2018). FRMCLDA applies the faster singular value threshold (fSVT) algorithm for a similarly optimal low-rank approximation of the adjacency matrix, and prediction of latent links between lncRNAs and diseases in *LD*. Our faster randomized matrix completion method is illustrated in Algorithm 1 in the **Supplementary Materials S6**. The function *rSVD-BKI*(·) performs singular value decomposition and the details of realization can be found in an earlier study (Feng et al., 2018)

EXPERIMENTS AND RESULTS

We first put forward the evaluation metrics for the methods of association prediction. Second, we tested the effects of cosine similarity on diseases and lncRNAs and fine-tuned the weights of cosine similarity. Third, we implemented permutation test to assess the influence of different data sources on optimization procedure. Fourth, we recorded the time usage of FRMCLDA for different sizes of heterogeneous network. Fifth, we compared FRMCLDA with other existing methods by global LOOCV experiments, local LOOCV experiments and global 5-fold cross-validation experiments. Finally, we implemented case studies to validate the practicability of FRMCLDA.

Evaluation Metrics of Performance

In order to assess the performance of FRMCLDA in inferring latent disease-correlated lncRNAs, global LOOCV experiments, local LOOCV experiments and global 5-fold cross-validation experiments are implemented on three benchmark datasets. Under the framework of LOOCV, each known experimentally validated association is picked in turns as a test sample, and all the other known associations are considered as training samples. The test sample is sorted together with the candidate samples without known association evidence. The test sample whose rank exceed the given threshold would be considered as a successful prediction. The main difference between global and local LOOCV is whether to investigate all diseases simultaneously or only query one disease at a time to select candidate samples. That is to say, global LOOCV considered all the unknown associations as candidate samples, whereas local LOOCV only focused on one disease in the test sample and selected the corresponding unknown associations as candidate samples. In global 5-fold cross-validation, all of the known experimentally validated associations are divided into five uncrossed sets, whose size must be strictly equal. Each set of the five is taken in turns as the test sample, but the other 4 sets are served as training samples. After matrix completion is performed, the test samples are ranked together with candidate samples and then are sorted in the descending order of their predicted scores.

Furthermore, false negative (FN), false positive (FP), true negative (TN) and true positive (TP) are summarized based on the ranked results for each specific threshold. The receiver operating characteristic curves are made by plotting the true positive rate (TPR, recall) against false positive rate (FPR) based on varying thresholds. The precision-recall (PR) curve is also plotted to fully evaluate the performance of the prediction. The area under the ROC curve (AUC) and the area under the

PR curve (AUPR) are finally calculated to evaluate the overall performance of the prediction. An AUC value of 0.5 implies a random prediction and an AUC value of 1 implies a perfect prediction performance. Therefore, AUC and AUPR are used as primary evaluating measures.

Effects of Cosine Similarity on Diseases and lncRNAs

Both integrated disease similarity and integrated lncRNA similarity in three benchmark datasets are calculated with cosine similarity combined, which can extract feature information from the known interaction matrix. The weights of w_l and w_d in integrated similarity calculations can be fine-tuned by cross validation in three benchmark datasets separately. Let w_d and w_l vary from 0.1 to 1 at the increment of 0.1. According to AUC values of LOOCV based on Dataset1, FRMCLDA performed best when w_{d1} = 0.9 and w_{l1} = 0.7. Likewise, on Dataset2, we chose w_{d2} = 0.7 and w_{l2} = 0.5. On Dataset3, we chose w_{d3} = 0.5 and w_{l3} = 0.5. All can be seen in **Table 1**.

In Dataset2, we implemented 5-fold cross validation 20 times to test the effects of the cosine similarity on model performance. The four test settings were: 1) using cosine similarity both for integrated similarity of neither lncRNAs nor diseases; 2) only the lncRNA similarity integrating the cosine similarity; 3) only the disease similarity integrating the cosine similarity; 4) both lncRNA similarity and disease similarity integrating cosine similarity. The results can be seen in **Table 2**. When both similarities are calculated with cosine similarity combined, the AUC value (0.9145 ± 0.0013) achieves the best of four. Therefore, FRMCLDA performance can be improved by incorporating effective feature information extracted by cosine similarity from interaction profiles with fast matrix completion.

Influence of Different Data Sources on Optimization Procedure

To evaluate the influence of different data sources on the optimization procedure of matrix completion, we have implemented a permutation test on disease-lncRNA interaction matrix DL, lncRNA similarity matrix LS, and disease similarity matrix DS separately. Based on the LOOCV framework, we randomized each of the three matrices in turns, while keeping the other two matrices unchanged. We carried it out 20 times and recorded the average AUC value for each type of data source. Usually, if a matrix contributes more to the optimization procedure, the result of the permutation test based on it will be closer to the stochastic value. As shown in **Table 3**, the mean AUC based on randomized matrix DL is the lowest and close to 0.5, indicating that matrix DL has the greatest influence on the performance of our model. In the same way, it is concluded that matrix LS contributes more than DS to the performance of our model.

TABLE 2 | The effects of the cosine similarity on AUC by 5CV in dataset2.

No *LS*_cosine and *DS*_cosine	Only combing *LS_cosine* in LS	Only combing *DS_cosine* in DS	Combing *LS_cosine* in LS and *DS_cosine* in DS
0.7995 ± 0.0044	0.8705 ± 0.0050	0.8510 ± 0.0032	0.9145 ± 0.0013

TABLE 3 | The result of contribution test on performance of prediction by LOOCV in dataset 2.

Set the data source to random matrix	Average AUCs by 20 times randomization
lncRNA similarity matrix (LS)	0.8615 ± 0.0061
Disease similarity matrix (DS)	0.8081 ± 0.0059
disease-lncRNA association matrix (DL)	0.5332 ± 0.0174

Time Usage of FRMCLDA for Different Sizes of the Heterogeneous Network

In FRMCLDA model, we implement matrix completion through fSVT as proposed by a previous study (Feng et al., 2018). Algorithm fSVT used a block Krylov iteration approximation SVD method rSVD-BKIr and a sub-space reuse mechanism to replace the original exact SVD. Thus, the turn-around time of SVT is significantly reduced while the accuracy remains the same. As seen in algorithm 1 in **Supplementary file S6**, tolerance ε is the terminating condition. When mean absolute error (MAE) is greater than ε, the program will terminate. The value of power parameter p can be dynamically adjusted as the operation accuracy changes. Therefore, the parameter ε and p can decide the rounds of the iteration, which will greatly affect the main turn-around time of FRMCLDA. Therefore, we were not able to compare the running time with other of other methods because of different conditions. We just recorded the time usage of FRMCLDA for different sizes of heterogeneous network.

Here, we set $p = 2$ and ε = 0.4. We executed 20 times FRMCLDA in three benchmark datasets. Average time usage of FRMCLDA and standard deviations are shown in **Table 4**. In Dataset 1, 2 and 3, the CPU time reached 2.1758 ± 0.2826 s, 1.5367 ± 0.1799 s and 3.9016 ± 0.2703 s, respectively.

Comparison of Performance With Other Methods on Different Datasets

On Dataset 1, the performance of FRMCLDA is compared with four popular methods: LRLSLDA (Chen and Yan, 2013), KATZLDA (Chen, 2015), SIMCLDA (Lu et al., 2018) and

TABLE 4 | The time usage of FRMCLDA for different sizes of heterogeneous network.

	The size of heterogeneous network	CPU time (second)
Dataset1	156 × 190	2.1758 ± 0.2826
Dataset2	115 × 178	1.5367 ± 0.1799
Dataset3	258 ×226	3.9016 ± 0.2703

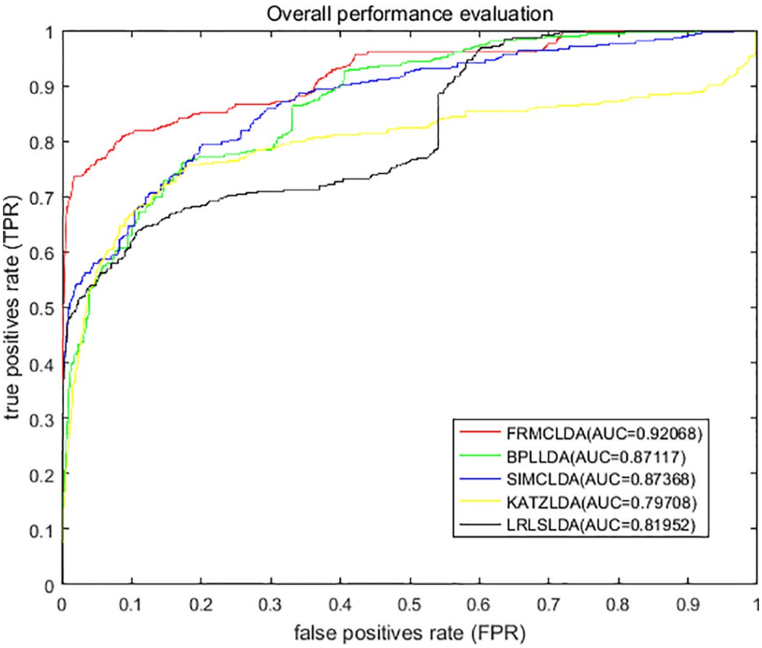


FIGURE 2 | Overall performance assessment of FRMCLDA, BPLLDA, SIMCLDA, KATZLDA and LRLSLDA in predicting disease-lncRNA relationships on Dataset 1 by global LOOCV.

BPLLDA (Xiao et al., 2018). The ROC curves of global LOOCV are shown in **Figure 2**. Obviously, FRMCLDA achieved an AUC of 0.92068, which outperformed LRLSLDA (0.81952), KATZLDA (0.79708), SIMCLDA (0.87368) and BPLLDA (0.87117) by 5% at least. Therefore, FRMCLDA is superior compared to other methods in predicting disease-lncRNA association.

One advantage of FRMCLDA is that it is able to infer latent correlated lncRNAs with queried diseases, even novel diseases. To show the performance of FRMCLDA in predicting novel disease-correlated lncRNAs, we implemented local LOOCV on Dataset 1. The results of FRMCLDA with local LOOCV on Dataset 1 were recorded in **Supplementary S1**. As shown in **Figure 3**, compared with three methods (LRLSLDA, BPLLDA, and GrwLDA) (Gu et al., 2017), the AUC of FRMCLDA was 0.91224, significantly higher than those of LRLSLDA (0.65812), BPLLDA (0.78528) and GrwLDA (0.65802) with increases of about 27.8%, 13.9%, and 27.86% respectively. The AUPR of FRMCLDA was 0.54644, significantly higher than those of LRLASLDA (0.12517), GrwLDA (0.1180) and BPLLDA (0.0753). In conclusion, FRMCLDA has been proven to be effective in inferring related lncRNAs with novel diseases in terms of AUC values and AUPR values. For example, we deleted all the known breast cancer-correlated associations, just as breast cancer was a novel disease. After matrix completion by FRMCLDA, we ranked all the candidate lncRNAs according to their scores. As can be seen in **Table 5**, all 14 deleted breast cancer-associated lncRNAs were finally successfully ranked out of top 20 of all the predicted lncRNAs.

The robustness of FRMCLDA was further validated by inferring latent associations on Dataset 2 and Dataset 3. We conducted 20 times global 5-fold cross-validation experiments to validate the precision of prediction by FRMCLDA on Dataset 2 and Dataset 3. The results of ROC curve, PR curve, precision-rank bars and recall-rank bars using different methods are shown in **Figure 4** and **Figure 5,** respectively. For example, as shown in **Figure 4**, after one time global 5-fold cross-validation on Dataset 2, FRMCLDA achieved an AUC of 0.91827, higher than SIMCLDA (0.88401) and KATZLDA (0.83693). The AUPR of FRMCLDA was 0.23794, also higher than those of SIMCLDA (0.1989) and KATZLDA (0.0635). Furthermore, on Dataset 2, the maximum precision reached by FRMCLDA is 0.88, which is higher than other methods, as shown in **Table 6**. On Dataset 3, after 20 times 5-fold CV, the average AUC of FRMCLDA is 0.8999 (±0.0049), which is superior to SIMCLDA 0.84694 (±0.0033) and KATZLDA 0.78561 (±0.0053). The AUPR of FRMCLDA is 0.1908 (±0.0033), higher than those of SIMCLDA 0.13717 (±0.0027) and KATZLDA 0.0293 (±0.0036), as shown in **Figure 5**. Furthermore, in terms of precision-rank bar and recall-rank bar, FRMCLDA boasts the best precision at every rank except for the top-20 rank in Dataset 3. In summary, FRMCLDA demonstrates high prediction accuracy on three different datasets.

Case Study

After performing cross validation to confirm the ability of FRMCLDA, we conducted a global prediction of potential related disease-lncRNA pairs. All known lncRNAs-diseases links were considered as training samples, while other unknown associations constituted the candidate samples. FRMCLDA can infer the latent correlated lncRNAs for all diseases simultaneously by faster random matrix completion. All candidate lncRNAs correlated with a queried disease were

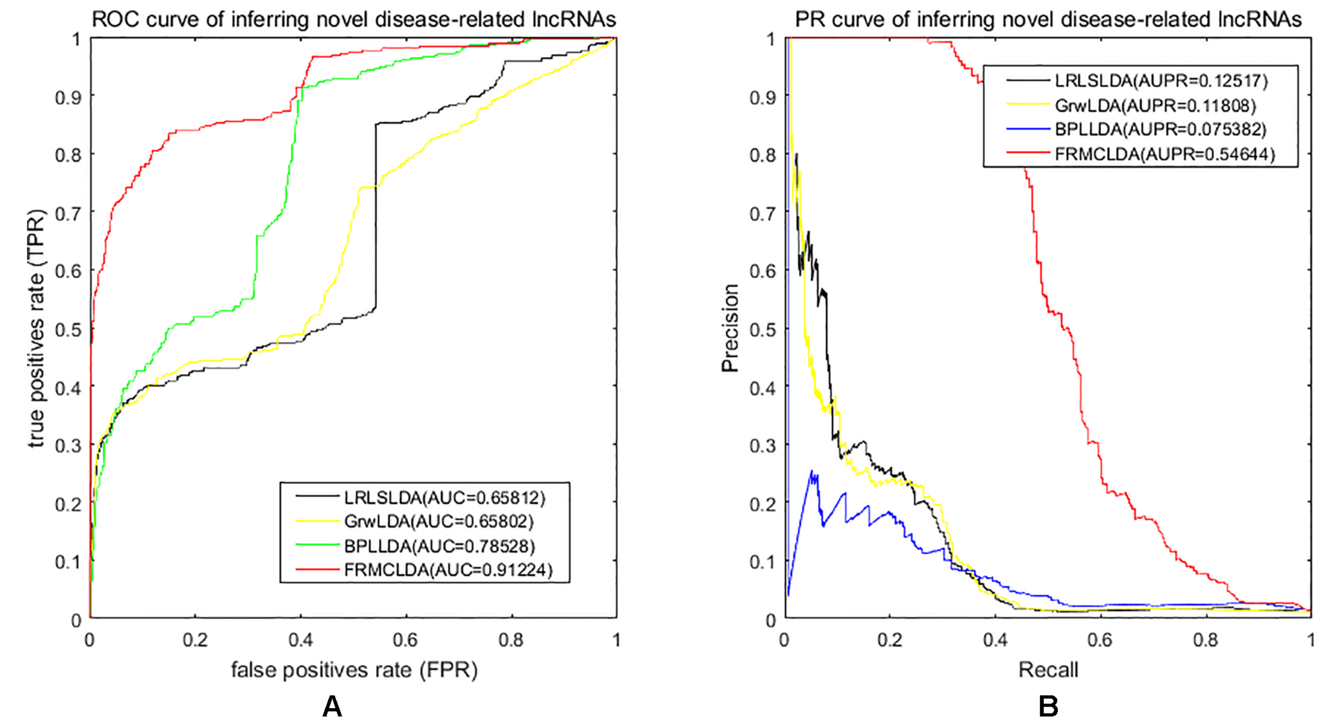


FIGURE 3 | Performance assessment of LRLSLDA, GrwLDA, BPLLDA and FRMCLDA in inferring novel disease-correlated lncRNAs on Dataset 1 by local LOOCV. **(A)** ROC curve of inferring novel disease-related lncRNAs. **(B)** PR curve of inferring novel disease-related lncRNAs.

TABLE 5 | Predicting novel disease-related lncRNAs by deleting known associations for each disease.

Known but deleted breast cancer-related lncRNAs	Rank number	Known but deleted breast cancer-related lncRNAs	Rank number
BCAR4	13	LSINCT5	7
BCYRN1	6	MALAT1	2
CDKN2B-AS1	4	MEG3	3
DSCAM-AS1	8	MIR31HG	14
GAS5	10	PINC	15
H19	1	PVT1	5
HOTAIR	9	SRA1	11

ranked according to predicted scores generated by FRMCLDA. The predicted and ranked lncRNAs (excluding known correlated lncRNAs) correlated with 226 diseases on Dataset 3 can be seen in **Supplementary S2**. To confirm whether top-ranked lncRNAs for queried diseases are real through public literature and three public databases (LncRNADisease, Lnc2Cancer and MNDR), we have conducted case studies on Dataset 3. Three databases are kept updated with new disease-lncRNA links verified by biological experiments in support of our validation. As shown by one of the prediction results in **Supplementary S2**, we take prostate cancer, colon cancer and gastric cancer and show the verification of top-20 predicted lncRNAs for each selected cancer.

Prostate cancer is one of the most common malignant tumors for males, accounting for about 13% of cancer-related death (Miller et al., 2016). In prostate cancer, the expression level of lncRNAs may be increasing or decreasing steadily (Smolle et al., 2017). Thus, it is justifiable to predict the possible links between lncRNAs and prostate cancer. Recent biological experiments have identified several lncRNAs associated with prostate cancer. For example, LncRNA H19 is down-regulated significantly in the cell line M12 of metastatic prostate carcinoma (Zhu et al., 2014). HOTAIR is found to be significantly regulated *via* genistein, and the expression of HOTAIR in castration-resistant PCa cell line is higher than that of standard prostate cell lines (Chiyomaru et al., 2013). MEG3 can enhance Bax, activate caspase 3 and inhibit the internal survival pathway of cells *in vivo* and *in vitro* through decreasing the bcl-2 protein expression (Zhang et al., 2003). MALAT1 is upregulated in prostate tumor tissue and cell line of human beings (Ren et al., 2013). It is reported that CBX7 and CDKN2B-AS1 levels are enhanced in prostate tumor tissues (Yap et al., 2010). PVT1 can accelerate the intrusion and transfer by prostate carcinoma *via* regulating EMT (Chang et al., 2018). Linc00963 is a new lncRNA which is involved in the transformation from the androgen-dependent stage to the androgen-independent stage of prostate carcinoma (Wang et al., 2014). In our work, FRMCLDA is performed to infer possible lncRNAs correlated with prostate cancer. Finally, 10 out of top-10 and 16 out of top-20 predicted prostate cancer-associated lncRNAs are verified on the three databases (LncRNADisease, MNDR, Lnc2Cancer) mentioned above. They are shown in **Table 7**.

Colon cancer is considered as one of the most widespread and deadly cancers in the world. Disorders of lncRNAs are associated

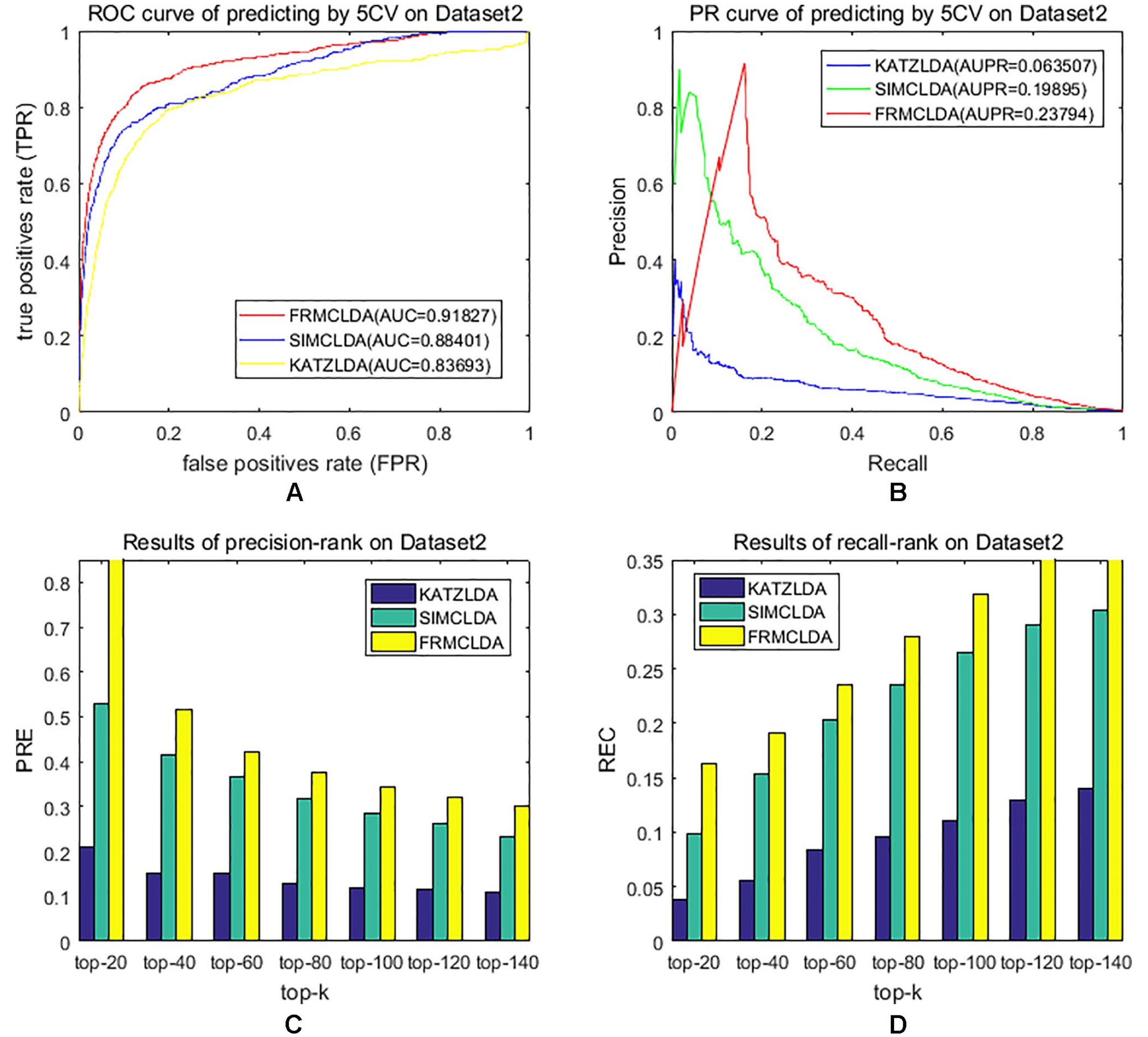


FIGURE 4 | Performance of FRMCLDA, KATZLDA and SIMCLDA on inferring lncRNAs by global 5-fold cross-validation on Dataset 2. **(A)** ROC curve of predicting disease-lncRNA associations. **(B)** PR curve of predicting disease-lncRNA associations. **(C)** Results of precision at every rank. **(D)** Results of recall at every rank.

with miscellaneous biological processes, including tumorigenesis (Ba-Alawi et al., 2016). For example, CDKN2B-AS1 up-regulates proliferation in HCT116 cells in a manner independent of the p15/p16-pRB pathway (Chiyomaru et al., 2013). The lower expression of GAS5 is highly related to big tumor volumes, low histological scores and late TNM stages. LncRNA Plasma UCA1 can be used as a potential biomarker for inchoate diagnosis and monitoring of colon cancer (Aken et al., 2016). It is found that overexpression of lncRNA TUG1 promotes colon cancer progression (Ba-Alawi et al., 2016). We utilize FRMCLDA to restore the possible colon cancer-correlated lncRNAs. The results suggest that, 9 out of the top 10 (9/10) and 15 out of the top 20 (15/20) predicted lncRNAs are confirmed by three databases mentioned before (LncRNADisease, MNDR, Lnc2Cancer), as shown in **Table 8**.

Gastric cancer is one of the cancers with the highest incidence and mortality in the world. Gastric cancer is a complicated disease, caused by an imbalance of the cancer-causing and cancer-suppressing pathways (Aken et al., 2016). An increasing number of studies show that lncRNAs may play an active role in primary processes of gastric cancer. FRMCLDA predicts the gastric cancer-associated lncRNAs, some of which are validated though the latest public literature and databases. For instance, the expression of GAS5 is found to be lowered in gastric tumors, contrary to the up-regulated expression of mir-23a (Ba-Alawi et al., 2016). It is suggested that a high

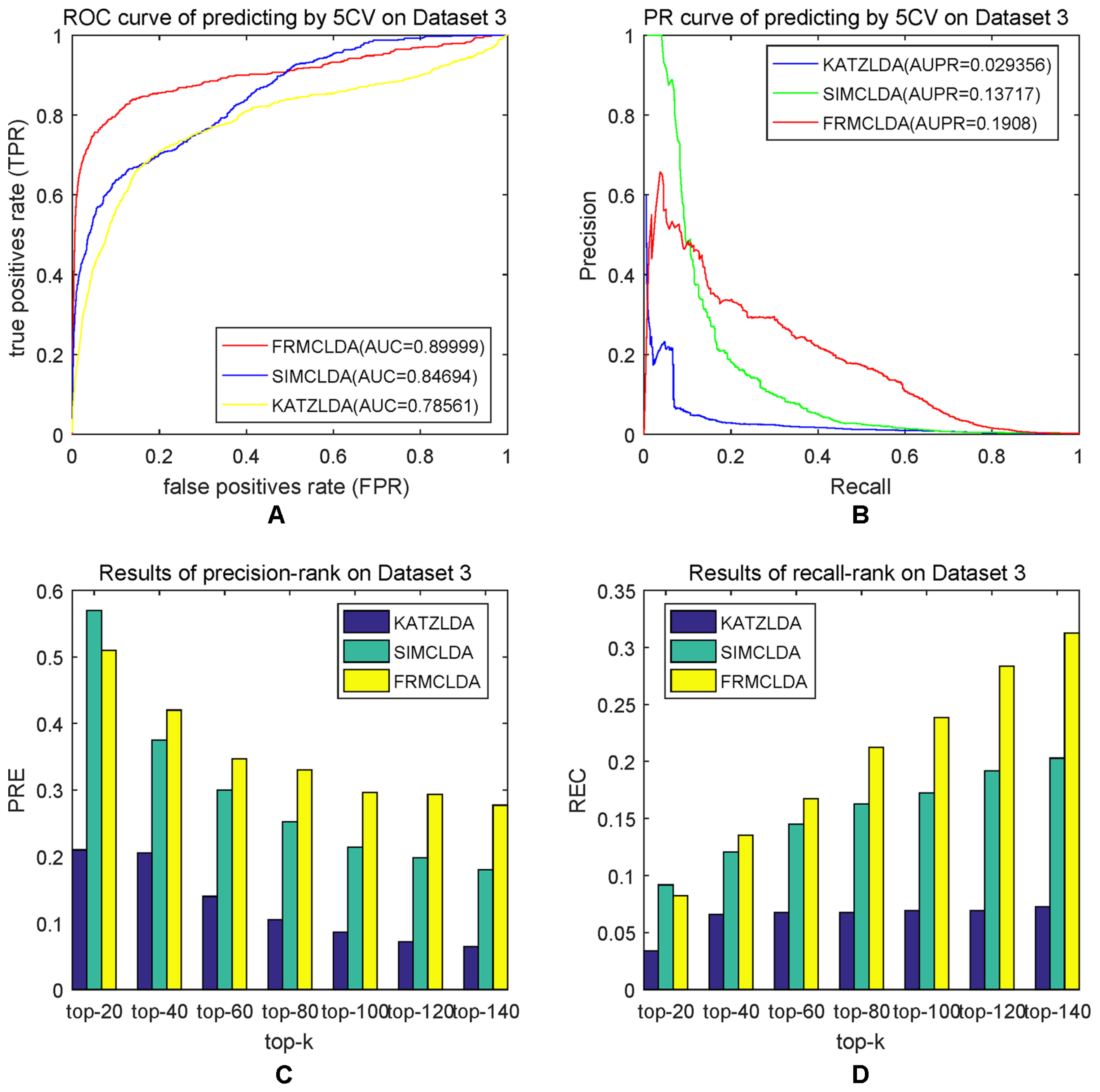


FIGURE 5 | Performance of FRMCLDA, KATZLDA and SIMCLDA on inferring lncRNAs by global 5-fold cross-validation on Dataset 3. **(A)** ROC curve of predicting disease-lncRNA associations. **(B)** PR curve of predicting disease-lncRNA associations. **(C)** Results of precision at every rank. **(D)** Results of recall at every rank.

TABLE 6 | Precision-rank on dataset 2.

	lncRNA	Top 20	Top 40	Top 60	Top 80	Top 100	Top 120	Top 140
precision	FRMCLDA	0.8800	0.5150	0.4233	0.3775	0.3440	0.3200	0.3029
	SIMCLDA	0.5300	0.4150	0.3667	0.3175	0.2860	0.2671	0.2343
	KATZLDA	0.2100	0.1500	0.1500	0.1300	0.1200	0.1167	0.1086
recall	FRMCLDA	0.1630	0.1707	0.2352	0.2796	0.3185	0.3556	0.3926
	SIMCLDA	0.0981	0.1537	0.2037	0.2352	0.2648	0.2907	0.3037
	KATZLDA	0.0389	0.0556	0.0833	0.0963	0.1111	0.1296	0.1407

TABLE 7 | The top-20 lncRNAs predicted for prostate cancer.

Rank	LncRNA	Pubmed ID	Rank	LncRNA	Pubmed ID
1	H19	24988946	11	IGF2-AS	27507663
2	HOTAIR	23936419	12	PCAT1	22664915
3	MEG3	14602737	13	LincRNA-p21	27976428
4	MALAT1	23845456	14	PTENpg1	not found
5	CDKN2B-AS1	20541999	15	PRNCR1	20874843
6	PVT1	23728290	16	SNHG16	not found
7	GAS5	22664915	17	MINA	not found
8	Linc00963	24691949	18	SRA1	16607388
9	C1QTNF9B-AS1	27507663	19	NEAT1	25415230
10	UCA1	27686228	20	LSINCT5	not found

TABLE 8 | The top-20 lncRNAs predicted for colon cancer.

Rank	Name of LncRNA	Pubmed ID	Rank	Name of LncRNA	Pubmed ID
1	CDKN2B-AS1	26708220	11	DRAIC	Not found
2	PVT1	25043044	12	IGF2-AS	Not found
3	GAS5	25326054	13	NPTN-IT1	23395002
4	LincRNA-p21	26656491	14	XIST	29679755
5	UCA1	26885155	15	PCAT29	Not found
6	KCNQ1OT1	16965397	16	LSINCT5	25526476
7	TUG1	27634385	17	anti-NOS2A	Not found
8	MINA	Not found	18	HIF1A-AS2	29278853
9	BCYRN1	29625226	19	SNHG16	24519959
10	MIAT	29686537	20	HIF1A-AS1	28946548

TABLE 9 | The top-20 lncRNAs predicted for gastric cancer.

Rank	Name of LncRNA	Pubmed ID	Rank	Name of LncRNA	Pubmed ID
1	GAS5	27827524	11	SNHG16	29081409
2	MALAT1	27486823	12	PTENpg1	25694351
3	LincRNA-p21	28969031	13	PCAT29	25700553
4	BCYRN1	29435146	14	XIST	29053187
5	KCNQ1OT1	Not found	15	BDNF-AS1	Not found
6	IGF2-AS	Not found	16	HIF1A-AS1	26722487
7	TUG1	27983921	17	HIF1A-AS2	25686741
8	NPTN-IT1	28951520	18	lncRNA-ATB	28115163
9	MIAT	29039602	19	HAR1B	Not found
10	DRAIC	25700553	20	CCAT2	29435046

level of MALAT1 could be a potential biomarker for distant metastasis of gastric cancer. Further studies have shown that lincrna-p21 knockout can promote the malignant behavior of gastric cancer cells according to overexpression assay (Aken et al., 2016). LncRNA SNHG16 is found to be highly expressed in gastric cancer and thus has become a novel target of clinical treatment for gastric cancer (Lian et al., 2017). The results show that, 8 out of the top-10 ranked lncRNAs and 16 out of the top-20 ranked lncRNAs are validated by FRMCLDA, as shown in **Table 9**.

The network of the top 50 ranked links with prostate cancer, colon cancer and gastric cancer on Dataset 3 by FRMCLDA is shown in **Figure 6**. We find that some top-ranked lncRNAs are associated with one or more diseases. The results in case studies for three selected cancers have shown an outstanding prediction performance of FRMCLDA. As stated, FRMCLDA is a comprehensive method which could infer latent disease-lncRNA link for overall diseases synchronously. As a result, we also prioritized overall candidate disease-lncRNA pairs (excluding known links) on Dataset 3 by their global scores assigned through FRMCLDA. The higher the global scores of the links, the more likely that links between them exist. For example, the predicted global score for GAS5 and gastric cancer ranks 6th out of all the 60,169 non-zero predicted results by FRMCLDA. This prediction was confirmed in the latest research by Sun M et al. (Aken et al., 2016). They verified that the reduced expression of GAS5 indicates poor prognoses and will lead to gastric cancer cells spreading. Therefore, top-priority prediction further proves the validity of FRMCLDA, and so do the other high-ranked links. The results of the global rank for all the predicted links are provided in **Supplementary S3**. We hope that the prediction results may help discovery of disease-related lncRNAs.

CONCLUSIONS

With the development of the next-generation sequencing in biomedical research, constructing a heterogeneous network on the basis of clinical NGS big data will benefit in prediction models of latent human disease-lncRNA associations. The prediction of disease-lncRNA links is very important in the biomedical field, among others. Construction of computational prediction models for new disease-lncRNA relationships will help understand the molecular mechanism of complicated human diseases at the level of lncRNA, and recognize the disease biomarker for diagnosis, treatment, prognosis and prevention of disease.

In this paper, we calculated the integrated similarities for diseases and lncRNAs using different methods and dealing with different types of data sources. We constructed a heterogeneous bilayer network by integrating similarity networks and interaction network. Then we utilized the algorithm fSVT to retrieve the unknown entries in adjacency matrix of the heterogeneous network. Theoretically FRMCLDA has a superior performance compared to other association prediction methods, because it takes account of all the predominant eigenvalues and the relevant eigenvectors of the matrix to be restored. In addition, FRMCLDA is able to process large scale matrices and execute proximate SVD rapidly at each SVT iteration by incorporating rSVD-BKI with a novel sub-space reuse technique. To assess the performances of FRMCLDA, experiments including global LOOCV and local LOOCV, global 5-fold CV and case studies are conducted. The experimental results show that the effectiveness of FRMCLDA is consistent with the theoretical estimation. Nevertheless, there are also a few limitations for FRMCLDA. First, if the adjacency matrix lacks low rank, then matrix completion with fSVT will lose its speed advantage. Second, the p value in power iteration can be adapted to guarantee the accuracy of SVD, but it can increase to

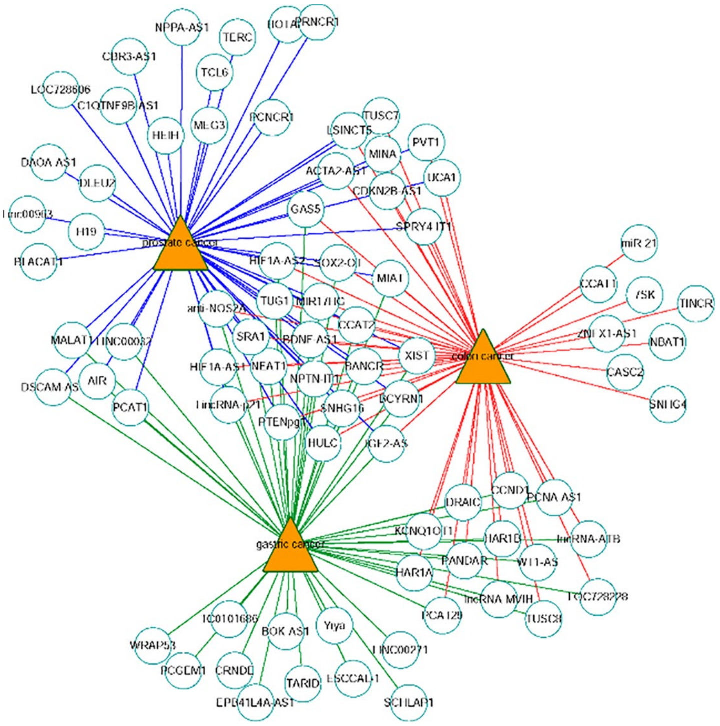


FIGURE 6 | Network of the top-50 predicted associations of prostate cancer, colon cancer and gastric cancer on Dataset 3. Circles and triangles represent lncRNAs and diseases, respectively.

several tens in some situations, which will lead to an increased running time and deprive the rSVD-BKI of advantages over the original SVD method. In conclusion, by expediting the matrix completion algorithm or properly extracting more effective features from lncRNAs and diseases, the performances of FRMCLDA can be further improved.

DATA AVAILABILITY

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

AUTHOR CONTRIBUTIONS

WL, SW and JY conceptualized the work and planned the procedure of experiments; JX and GM and GT implemented literature research; WL collected the data and analysed the results; WL and JY drafted the manuscript; all the authors have read and supported the final edition.

FUNDING

This work was supported by the National Nature Science Foundation of China (Grant Nos. 61672011 and 61474267, the National Key Research and Development Program (Grant Nos. 2017YFC1311003) and the Natural Science Foundation of Hunan, China (Grant No. 2018JJ2461).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00769/full#supplementary-material

REFERENCES

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The ensembl gene annotation system. *Database (Oxford)* 1–19 2016. doi: 10.1093/database/baw093

Ba-Alawi, W., Soufan, O., Essack, M., Kalnis, P., and Bajic, V. B. (2016). DASPfind: new efficient method to predict drug-target interactions. *J. Cheminf.* 8, 15. doi: 10.1186/s13321-016-0128-4

Bian, E. B., Li, J., Xie, Y. S., Zong, G., Li, J., and Zhao, B. (2015). LncRNAs: new players in gliomas, with special emphasis on the interaction of lncRNAs with EZH2. *J. Cell. Physiol.* 230, 496–503. doi: 10.1002/jcp.24549

Cai, J. F., Candes, E. J., and Shen, Z. (2008). A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* 20, 1956–1982. doi: 10.1137/080738970

Candès, E. J., and Recht, B. (2009). Exact matrix completion *via* convex optimization. *Found. Comput. Math.* 9, 717. doi: 10.1007/s10208-009-9045-5

Chang, Z., Cui, J., and Song, Y. (2018). Long noncoding RNA PVT1 promotes EMT *via* mediating microRNA-186 targeting of twist1 in prostate cancer. *Gene* 654, 36–42. doi: 10.1016/j.gene.2018.02.036

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

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

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

Chiyomaru, T., Yamamura, S., Fukuhara, S., Yoshino, H., Kinoshita, T., Majid, S., et al. (2013). Genistein inhibits prostate cancer cell growth by targeting miR-34a and oncogenic HOTAIR. *PLoS One* 8, e70372. doi: 10.1371/journal.pone.0070372

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22, 1775–1789. doi: 10.1101/gr.132159.111

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

Feng, X., Yu, W., and Li, Y. (2018). "Faster Matrix Completion Using Randomized SVD", in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)* (Volos, Greece: Greece, IEEE), 608–615. doi: 10.1109/ICTAI.2018.00098

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

Halko, N., Martinsson, P. G., and Tropp, J. A. (2010). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* 53, 217–288. doi: 10.1137/090771806

Li, Y., and Yu, W. (2017). A fast implementation of singular value thresholding algorithm using recycling rank revealing randomized singular value decomposition. *Computer Science: Numerical Analysis*.

Lian, D., Amin, B., Du, D., and Yan, W. (2017). Enhanced expression of the long non-coding RNA SNHG16 contributes to gastric cancer progression and metastasis. *Cancer Biomarker* 21, 151–160. doi: 10.3233/CBM-170462

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

Miller, K. D., Siegel, R. L., Lin, C. C., Mariotto, A. B., Kramer, J. L., Rowland, J. H., et al. (2016). Cancer treatment and survivorship statistics, 2016. *CA Cancer J. Clin.* 66, 271–289. doi: 10.3322/caac.21349

Musco, C., and Musco, C. (2015). Randomized block Krylov methods for stronger and faster approximate singular value decomposition, in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1, 1–9. (Montreal, Canada: MIT Press).

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.* 24, 227–234. doi: 10.1137/S0097539792240406

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453. doi: 10.1016/0022-2836(70)90057-4

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., et al. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nat. Commun.* 7, 12791. doi: 10.1038/ncomms12791

Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., et al. (2016). Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* 44, D980–D985. doi: 10.1093/nar/gkv1094

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.* 35, 747–750. doi: 10.1093/nar/gkl995

Pinero, J., Bravo, A., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., et al. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 45, D833–D839. doi: 10.1093/nar/gkw943

Ren, S., Liu, Y., Xu, W., Sun, Y., Lu, J., Wang, F., et al. (2013). Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J. Urol.* 190, 2278–2287. doi: 10.1016/j.juro.2013.07.001

Rose, J., and Eisenmenger, F. (1991). A fast unbiased comparison of protein structures by means of the Needleman-Wunsch algorithm. *J. Mol. Evol.* 32, 340–354. doi: 10.1007/BF02102193

Smolle, M. A., Bauernhofer, T., Pummer, K., Calin, G. A., and Pichler, M. (2017). Current insights into Long Non-Coding RNAs (LncRNAs) in prostate cancer. *Int. J. Mol. Sci.* 18. doi: 10.3390/ijms18020473

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

Wang, J. Z., Zhidian, D., Rapeeporn, P., Yu, P. S., and Chin-Fu, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, L., Han, S., Jin, G., Zhou, X., Li, M., Ying, X., et al. (2014). Linc00963: a novel, long non-coding RNA involved in the transition of prostate cancer from androgen-dependence to androgen-independence. *Int. J. Oncol.* 44, 2041–2049. doi: 10.3892/ijo.2014.2363

Wang, Y., Chen, L., Chen, B., Li, X., Kang, J., Fan, K., et al. (2013). Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis.* 4, e765. doi: 10.1038/cddis.2013.292

Xiao, X., Zhu, W., Liao, B., Xu, J., and Yang, J. (2018). BPLLDA: predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* 9, 411. doi: 10.3389/fgene.2018.00411

Yap, K. L., Li, S., Munoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., et al. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell.* 38, 662–674. doi: 10.1016/j.molcel.2010.03.021

Zhang, X., Zhou, Y., Mehta, K. R., Danila, D. C., Scolavino, S., Johnson, S. R., et al. (2003). A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. *J. Clin. Endocrinol. Metab.* 88, 5119–5126. doi: 10.1210/jc.2003-030222

Zhou, X., Yin, C., Dang, Y., Ye, F., and Zhang, G. (2015). Identification of the long non-coding RNA H19 in plasma as a novel biomarker for diagnosis of gastric cancer. *Sci. Rep.* 5, 11516. doi: 10.1038/srep11516

Zhu, M., Chen, Q., Liu, X., Sun, Q., Zhao, X., Deng, R., et al. (2014). lncRNA H19/miR-675 axis represses prostate cancer metastasis by targeting TGFBI. *FEBS J.* 281, 3766–3775. doi: 10.1111/febs.12902

Conflict of Interest Statement: JY and GT were employed by company Geneis Beijing Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflict of interest.

BRIEF RESEARCH REPORT
published: 18 September 2019
doi: 10.3389/fgene.2019.00830

MPIC: Molecular Prognostic Indicators in Cirrhosis Database for Clinical Context-Specific *in Silico* Prognostic Biomarker Validation

Shun H. Yip[1†], *Naoto Fujiwara*[1,2†], *Jason Burke*[3], *Anand Shetler*[1], *Celina Peralta*[1], *Tongqi Qian*[4], *Hiroki Hoshida*[1], *Shijia Zhu*[1*] *and Yujin Hoshida*[1*]

[1] Liver Tumor Translational Research Program, Simmons Comprehensive Cancer Center, Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, United States, [2] Department of Gastroenterology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, [3] Broad Institute of MIT and Harvard University, Cambridge, MA, United States, [4] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States

OPEN ACCESS

Edited by:
Yudong Cai,
Shanghai University,
China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Yu Wang,
Jilin University,
China

Correspondence:
Shijia Zhu
Shijia.Zhu@UTSouthwestern.edu
Yujin Hoshida
Yujin.Hoshida@UTSouthwestern.edu

†These authors have contributed equally to this work

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

Received: *20 June 2019*
Accepted: *12 August 2019*
Published: *18 September 2019*

Citation:
Yip SH, Fujiwara N, Burke J, Shetler A, Peralta C, Qian T, Hoshida H, Zhu S and Hoshida Y (2019) MPIC: Molecular Prognostic Indicators in Cirrhosis Database for Clinical Context-Specific in Silico Prognostic Biomarker Validation. Front. Genet. 10:830. doi: 10.3389/fgene.2019.00830

Prognostic biomarkers are vital in the management of progressive chronic diseases such as liver cirrhosis, affecting 1–2% of the global population and causing over 1 million deaths every year. Despite numerous candidate biomarkers in literature, the costly and lengthy process of validation hampers their clinical translation. Existing omics databases are not suitable for *in silico* validation due to the ignorance of critical factors, i.e., study design, clinical context of biomarker application, and statistical power. To address the unmet need, we have developed the Molecular Prognostic Indicators in Cirrhosis (MPIC) database as a representative example of an omics database tailored for prognostic biomarker validation. MPIC consists of (i) a molecular and clinical database structured by defined disease context and specific clinical outcome and annotated with employed study design and anticipated statistical power by disease domain experts, (ii) a bioinformatics analysis engine for user-provided gene-signature- or gene-based prognostic prediction, and (iii) a user interface for interactive exploration of relevant clinical cohort/scenario and assessment of significance and reliability of the result for prognostic prediction. MPIC assists cost-effective prognostic biomarker development by facilitating the process of validation, and will transform the care of chronic diseases such as cirrhosis. MPIC is freely available at www.mpic-app.org. The website is implemented in Java, Apache, and MySQL with all major browsers supported.

Keywords: prognostic prediction, study design, molecular signature, chronic disease, cirrhosis

INTRODUCTION

Management of chronic diseases is a considerable economic burden to the medical care systems. For example, progressive fibrosis in solid organs is one of the major life-limiting chronic disease conditions associated with at least one-third of deaths worldwide (Rockey et al., 2015). Liver cirrhosis is one of the major fibrotic conditions that costs >$12 billion even in the U.S. alone (Ge and Runyon, 2016; Fujiwara et al., 2018). Organ fibrosis progression generally takes decades and the rate of disease progression is highly variable across the patients. Therefore, prognostic prediction is critical to allocate limited medical resources to rapid progressors who most need intervention, while sparing the resources for slow progressors to maximize the cost-effectiveness of patient management. However, development of prognostic biomarker is extremely challenging as evidenced by the absence of clinically translated biomarkers despite years of research (Goossens et al., 2015). This is primarily due to requirement of lengthy and costly clinical validation of candidate biomarkers, which does not fit within the budget and time frame of typical clinical trial. A fast and cheap alternative strategy of prognostic biomarker validation is sorely needed.

Publicly available omics profiles of clinical specimens may provide the opportunity of *in silico* validation for candidate prognostic biomarkers and spare resources and efforts wasted for unsuccessful clinical trials. However, currently available databases do not meet the need because the following two critical issues for prognostic biomarker assessment are disregarded (Chen et al., 2014): (1) Study-design-related information is missing. Clinical prognostic information, defined as time to clinical event, is generally incomplete due to insufficient observation period and/or biases in patient enrollment and treatment and follow-up protocols. Therefore, observed prognostic association is vulnerable to flaws in study design that could lead to false positive or negative finding (Goossens et al., 2015). Clinical patient cohort can be assembled in either retrospective or prospective manner. A retrospective cohort is a collection of patients from previously performed clinical care, where patient inclusion/exclusion criteria cannot be optimized because the enrollment is already completed in the past. In contrast, a prospective cohort is a collection of patients from future clinical care, which can be enrolled based on pre-determined inclusion/exclusion criteria, although completion of patient enrollment and follow up will take long time and is costly. In reality, virtually most of omics data suffer from the issue of biased patient enrollment because of the use of "samples of convenience," i.e., readily available biospecimens retrospectively collected without predetermined intention of prognostic biomarker assessment (Simon et al., 2009). Thus, it is critical to annotate cohort/dataset for study design quality according to reporting guidelines to provide clue to reliability of observed prognostic association (Mcshane et al., 2006; Vandenbroucke et al., 2007); (2) Specific clinical context or scenario for biomarker application is missing. There is no clinical utility for a prognostic biomarker without specific indication of its use in real-world clinical practice, e.g., prediction of liver cancer development in Child-Pugh class A compensated viral cirrhosis patients monitored under biannual liver cancer screening, prediction of cancer-related death after 8-week cisplatin-based chemotherapy in stage III ovarian cancer.

To meet the unmet need by addressing the two major issues, we have developed Molecular Prognostic Indicators in Cirrhosis (MPIC) database as a proof of concept specifically designed for reliable prognostic assessment of candidate biomarkers using chronic fibrotic liver diseases as representative example. This scheme is readily applicable to other chronic diseases.

METHODOLOGY AND RESULTS

Genome-wide transcriptome datasets and associated clinical outcome data are from our previous and ongoing studies as well as private contribution. Although available data are still scarce, cohorts/outcomes from public databases such as NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) and EBI ArrayExpress (www.ebi.ac.uk/arrayexpress) are included.

The database currently contains 66 unique cohorts/outcomes of 5,540 subjects with unique clinical contexts, covering the major chronic liver diseases (i.e., viral or metabolic chronic hepatitis, cirrhosis, and cancer) for two types of outcome, time-to-event and binary outcomes (**Table 1**). The contents are curated and thoroughly annotated for study design by disease domain experts (NF and YH). The metadata include clinical demographics such as disease etiology, patient race/ethnicity, geographic region/country, median and interquartile range of clinical follow-up time, and % of patients who experienced clinical outcome of interest. Mode of patient enrollment is presented as prospective, retrospective to indicate the reliability of outcome association derived from the cohort. For instance, the analysis result from a prospective cohort can be reported as derived from "prospective-retrospective" study design, which indicates higher reliability compared to a result from retrospective study (Simon et al., 2009). Setting of patient enrollment is indicated as population-, community-, or hospital-based to explicitly indicate applicable clinical setting. Statistical power to detect certain magnitude of prognostic risk distinction is also provided to inform users about potential lack of statistical power for user-provided prognostic gene(s) at hazard ratios of 2.0 to 5.0 in Cox regression modeling, cutoffs often adopted to determine clinically meaningful prognostic risk distinction. Specific clinical contexts of biomarker application are unequivocally defined, and user can interactively find a clinical scenario of interest (see Step 1 in the next section).

MPIC consists of the following three components: (i) MySQL database of molecular profiles and clinical annotations for each specific clinical outcome in each patient cohort, (ii) bioinformatics data analysis engine developed based on GenePattern genomic analysis environment (Reich et al., 2006), and (iii) a user interface implemented using Java Grails, communicating with the database and analysis engine. Biostatistical analysis methods are implemented using the R statistical language (www.r-project.org).

TABLE 1 | Clinical demographics of subjects in MPIC database.

Clinical characteristic	
Age, median (IQR)	57 (50–65)
Sex, male no. (%)	4,035 (72.8)
Race/ethnicity, no. (%)	
Asian	3,369 (60.8)
Black	31 (0.6)
Caucasian	2,078 (37.5)
Hispanic	46 (0.8)
Disease etiology, no. (%)	
Hepatitis B	1,278 (23.0)
Hepatitis C	2.699 (48.7)
Alcohol	796 (14.4)
Non-alcoholic fatty liver disease	585 (10.6)
Observation time (yr), median (IQR)	2.9 (1.8–5.2)
Observation clinical events (%), median (IQR)	40 (31–55)

In MIPC, users can test their own candidate prognostic gene(s) for association with a specific clinical outcome in a patient cohort following the steps described below (**Figure 1**). MPIC helps circumvent the lengthy and costly process of biomarker validation by providing opportunity to quickly perform *in silico* assessment of candidate biomarkers without requiring any clinical and experimental resources.

Step 1: Select Patient Cohort and Clinical Outcome

Genome-wide molecular profiles of patient cohorts are hierarchically organized by disease condition (e.g., hepatocellular carcinoma, cirrhosis, alcoholic hepatitis), type of specimens (e.g., liver tissue, tumor tissue, serum), and clinical outcome (e.g., development of organ decompensation, diagnosis of stage I cancer within 2 years after surgical therapy, overall death). By selecting a patient cohort under a clinical outcome, a user can browse detailed metadata/annotations for the cohort. The cohort meta-data are summarized in **Supplementary Table 1**.

Step 2: Upload User-Defined Prognostic Gene or Molecular Signature

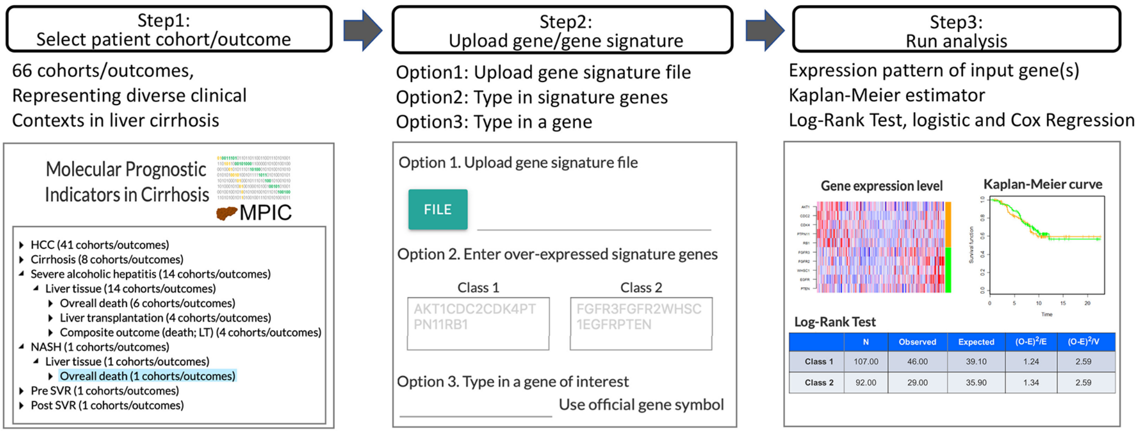
Subsequently, a user-defined prognostic molecular signature or gene is uploaded. A prognostic molecular signature is defined as two sets of genes, up- or down-regulated in association with the clinical outcome of interest, in official gene symbols. Alternatively, a single gene symbol can be provided to examine association of the gene's expression level with the clinical outcome of interest. MPIC currently supports only 2-class gene signature, i.e., two sets of genes overexpressed in association with either "Class 1" or "Class 2," corresponding to opposite clinical outcomes such as "poor survival" or "good survival," respectively.

Step 3: Patient Classification and Assessment of Prognostic Association

Using the user-defined molecular signature, each patient in the selected cohort is classified into either "Class 1" or "Class 2" subgroup (e.g., "poor survival" or "good survival" subgroup) by a nearest neighbor-based versatile class prediction algorithm, Nearest Template Prediction (NTP) using cosine distance as dissimilarity metric (Hoshida, 2010). Briefly, hypothetical representative "Class 1" and "Class 2" templates are defined as vectors with the same length with the user's input gene signature, where "Class 1" genes are set to 1 and "Class 2" genes are set to 0 for the "Class 1" template and vice versa for the "Class 2 template. Classification of each patient is performed based on proximity to either of the templates measured by cosine distance. Expression pattern of the user-provided molecular signature in the cohort is visualized as a heatmap of sample-wise Z-score for each gene. Alternatively, when a single gene symbol is provided as input, subjects are classified into high- or low-expression groups based on top quartile cut-off, and visualized as a bar graph. Association of the patient classification and time-to-event clinical outcome is evaluated by log-rank test and Cox regression and visualized as Kaplan-Meier curves. Correlation between each signature gene expression and selected time-to-event outcome is calculated as Cox score using the following equation adapted from previous study (Bair and Tibshirani, 2004):

$$d = \frac{\Sigma_{k=1}^{K}\left(\Sigma_{t_i=z_k} \boldsymbol{x_i} - \boldsymbol{d_k}\Sigma_{i\in R_k} \boldsymbol{x_i} / \boldsymbol{m_k}\right)}{\sqrt{\Sigma_{k=1}^{K}(\boldsymbol{d_k} / \boldsymbol{m_k})\Sigma_{i\in R_k}\left(\boldsymbol{x_i} - \Sigma_{i\in R_k} \boldsymbol{x_i} / \boldsymbol{m_k}\right)^2}} \qquad (1)$$

where i, sample index; k, unique death time indices $z_1{:}z_k$; x_i, transcript abundance in sample i, t_i, observation time; d_k, number of deaths at time z_k; m_k, number of samples in $R_k = i$: $t_i > z_k$. Statistical significance of the statistic is measured as false discovery rate based on random gene resampling-based ($n = 1{,}000$) nominal p-value and visualized as bar chart. Association with binary outcome is evaluated by 2 × 2 table statistics (e.g., sensitivity, specificity, positive/negative predictive values), Fisher's exact test, and logistic regression. Data analysis engine was developed based on GenePattern (Reich et al., 2006), which can be easily extended to incorporate more analytic pipelines towards more advanced requirements.


Step1:
Select patient cohort/outcome
66 cohorts/outcomes,
Representing diverse clinical
Contexts in liver cirrhosis
Molecular Prognostic
Indicators in Cirrhosis
MPIC
HCC (41 cohorts/outcomes)
Cirrhosis (8 cohorts/outcomes)
Severe alcoholic hepatitis (14 cohorts/outcomes)
Liver tissue (14 cohorts/outcomes)
Ovreall death (6 cohorts/outcomes)
Liver transplantation (4 cohorts/outcomes)
Composite outcome (death; LT) (4 cohorts/outcomes)
NASH (1 cohorts/outcomes)
Liver tissue (1 cohorts/outcomes)
Ovreall death (1 cohorts/outcomes)
Pre SVR (1 cohorts/outcomes)
Post SVR (1 cohorts/outcomes)
Step2:
Upload gene/gene signature
Option1: Upload gene signature file
Option2: Type in signature genes
Option3: Type in a gene
Option 1. Upload gene signature file
FILE
Option 2. Enter over-expressed signature genes
Class 1
Class 2
AKT1CDC2CDK4PTPN11RB1
FGFR3FGFR2WHSC1EGFRPTEN
Option 3. Type in a gene of interest
Use official gene symbol
Step3:
Run analysis
Expression pattern of input gene(s)
Kaplan-Meier estimator
Log-Rank Test, logistic and Cox Regression
Gene expression level
AKT1
CDC2
CDK4
PTPN11
RB1
FGFR3
FGFR2
WHSC1
EGFR
PTEN
Kaplan-Meier curve
Survival function
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Time
Log-Rank Test

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
Class 1	107.00	46.00	39.10	1.24	2.59
Class 2	92.00	29.00	35.90	1.34	2.59

FIGURE 1 | Workflow of MPIC for clinical context-specific *in silico* prognostic biomarker validation in cirrhosis.

Throughout the process, users do not have access to individual patient's molecular and clinical data. This is a logistical advantage that lowers the bar to deposit clinical outcome data by mitigating data contributors' concerns about sharing unpublished data, bleaching patient identity, and other regulatory issues. Besides ongoing regular expansion of cohort/dataset collection in the database, future developments will cover meta-analysis of prognostic associations derived from multiple patient cohorts for a molecular signature, multivariable analysis incorporating clinical prognostic factors, and comparison of prognostic performance across multiple molecular signatures.

DISCUSSION

Prognostic biomarker is the vital component in the management of patients with progressive and lethal chronic diseases. However, its development has been a daunting task due to the costly and lengthy process of clinical validation as evidenced by the scarce prognostic biomarker assays successfully translated to clinic. Currently available omics databases cannot accommodate the need because they disregard critical issues for clinical prognostic assessment such as study design, clinical context of biomarker use, setting of patient enrollment, statistical power, among many others.

To address the unmet need, we have developed a proof-of-concept database and web application, called MPIC. As opposed to biological hypothesis generation tools such as The Cancer Genome Atlas portal and associated databases, MPIC is specialized for prognostic biomarker validation using liver cirrhosis (cirrhosis) as a representative example that causes over one million deaths every year worldwide. It supports a quick go/no-go decision for prognostic biomarker candidates for further clinical development, avoids wasting cost and time for biomarker clinical trial, and enables revolutionarily more cost-effective prognostic biomarker development compared to the traditional strategy.

With this resource, we have successfully developed a prognostic assay implemented in FDA-approved clinical diagnostic platforms, supporting real-world clinical utility of our web application (initial discovery: (Hoshida et al., 2008), assay implementation and validation: (King et al., 2015; Nakagawa et al., 2016; Ono et al., 2017), incorporation in clinical trial as a companion biomarker: NCT02273362). Simulation-based analysis showed that personalized patient management with the prognostic assay is significantly cost-effective (Goossens et al., 2017), supporting that MPIC will have transformative biomedical impact on the dismal prognosis of cirrhosis patients. In the initial implementation, we primarily focused on gene expression datasets, but we will expand the database to cover other types of omics information such as non-coding RNA, epigenetic profiles, and DNA structural alterations. This scheme is readily applicable to other chronic diseases, and such an informatics resource will contribute to the substantial improvement of chronic disease management and patient prognosis.

DATA AVAILABILITY

MPIC database is freely available at www.mpic-app.org. Website implemented in Java, Apache, and MySQL with all major browsers supported.

AUTHOR CONTRIBUTIONS

YH and SZ conducted and designed this study. SY, JB, AS, and CP implemented the database and web application. NF, HH, and TQ performed the data curation. YH and SZ wrote the manuscript. All authors reviewed and approved the paper for publication.

FUNDING

This work has been supported by NIH/NIDDK (R01 DK099558, "Molecular Prognostic Indicators in Liver Cirrhosis and Cancer"), European Commission (ERC-2014-AdG-671231), Irma T. Hirschl Trust, and US Department of Defense (W81XWH-16-1-0363), Cancer Prevention and Research Institute of Texas (RR180016) to YH, and Uehara Memorial Foundation to NF.

ACKNOWLEDGMENTS

We thank Peter Carr, Ted Liefeld, and Michael Reich for technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00830/full#supplementary-material

REFERENCES

Bair, E., and Tibshirani, R. J. P. B. (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology* 2, e108. doi: 10.1371/journal.pbio.0020108

Chen, X., Sun, X., and Hoshida, Y. J. H. G. (2014). Survival analysis tools in genomics research. *Human Genom.* 8, 21. doi: 10.1186/PREACCEPT-1020290243146153

Fujiwara, N., Friedman, S. L., Goossens, N., and Hoshida, Y. J. J. O. H. (2018). Risk factors and prevention of hepatocellular carcinoma in the era of precision medicine. *J. Hepatol.* 68, 526–549. doi: 10.1016/j.jhep.2017.09.016

Ge, P. S., and Runyon, B. a. J. N. E. J. O. M. (2016). Treatment of patients with cirrhosis. *N. Engl. J. Med.* 375, 767–777. doi: 10.1056/NEJMra1504367

Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. J. T. C. R. (2015). Cancer biomarker discovery and validation. *Transl. Cancer Res.* 4, 256.

Goossens, N., Singal, A. G., King, L. Y., Andersson, K. L., Fuchs, B. C., Besa, C., et al. (2017). Cost-effectiveness of risk score–stratified hepatocellular carcinoma screening in patients with cirrhosis. *Clin. Transl. Gastroenterol.* 8, e101. doi: 10.1038/ctg.2017.26

Hoshida, Y., Villanueva, A., Kobayashi, M., Peix, J., Chiang, D. Y., Camargo, A., et al. (2008). Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *N. Engl. J. Med.* 359, 1995–2004. doi: 10.1056/NEJMoa0804525

Hoshida, Y. J. P. O. (2010). Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PloS One* 5, e15543. doi: 10.1371/journal.pone.0015543

King, L. Y., Canasto-Chibuque, C., Johnson, K. B., Yip, S., Chen, X., Kojima, K., et al. (2015). A genomic and clinical prognostic index for hepatitis C-related early-stage cirrhosis that predicts clinical deterioration. *Gut* 64, 1296–1302. doi: 10.1136/gutjnl-2014-307862

Mcshane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., Clark, G. M. J. B. C. R., et al. (2006). REporting recommendations for tumor MARKer prognostic studies (REMARK). *Breast Cancer Res. Treat* 100, 229–235. doi: 10.1007/s10549-006-9242-8

Nakagawa, S., Wei, L., Song, W. M., Higashi, T., Ghoshal, S., Kim, R. S., et al. (2016). Molecular liver cancer prevention in cirrhosis by organ transcriptome analysis and lysophosphatidic acid pathway inhibition. *Cancer Cell* 30, 879–890. doi: 10.1016/j.ccell.2016.11.004|

Ono, A., Goossens, N., Finn, R. S., Schmidt, W. N., Thung, S. N., Im, G. Y., et al. (2017). Persisting risk of hepatocellular carcinoma after hepatitis C virus cure monitored by a liver transcriptome signature. *Hepatol.* 66, 1344–1346. doi: 10.1002/hep.29203

Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P., and Mesirov, J. P. J. N. G. (2006). GenePattern 2.0. *Nature Genetics* 38, 500. doi: 10.1038/ng0506-500

Rockey, D. C., Bell, P. D., and Hill, J. a. J. N. E. J. O. M. (2015). Fibrosis—a common pathway to organ injury and failure. *N. Engl. J. Med.* 372, 1138–1149. doi: 10.1056/NEJMra1300575

Simon, R. M., Paik, S., and Hayes, D. F. J. J. O. T. N. C. I. (2009). Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Nat. Cancer Inst.* 101, 1446–1452. doi: 10.1093/jnci/djp335

Vandenbroucke, J. P., Von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., et al. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *PLoS Med.* 4, e297. doi: 10.1371/journal.pmed.0040297

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: 15 November 2019
doi: 10.3389/fgene.2019.01146

Identifying Microbiota Signature and Functional Rules Associated With Bacterial Subtypes in Human Intestine

Lijuan Chen[1], *Daojie Li*[1], *Ye Shao*[2], *Hui Wang*[1], *Yuqing Liu*[3] *and Yunhua Zhang*[3]*

[1] College of Animal Science and Technology, Anhui Agricultural University, Hefei, China, [2] School of Medicine, Huaqiao University, Quanzhou, China, [3] Anhui Province Key Laboratory of Farmland Ecological Conservation and Pollution Prevention, School of Resources and Environment, Anhui Agricultural University, Hefei, China

OPEN ACCESS

Edited by:
Peilin Jia,
University of Texas Health Science Center, United States

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Guang Wu,
Guangxi Academy of Sciences, China

***Correspondence:**
Yunhua Zhang
yunhua9681@163.com

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

Received: *17 July 2019*
Accepted: *21 October 2019*
Published: *15 November 2019*

Citation:
Chen L, Li D, Shao Y, Wang H, Liu Y and Zhang Y (2019) Identifying Microbiota Signature and Functional Rules Associated With Bacterial Subtypes in Human Intestine. Front. Genet. 10:1146. doi: 10.3389/fgene.2019.01146

Gut microbiomes are integral microflora located in the human intestine with particular symbiosis. Among all microorganisms in the human intestine, bacteria are the most significant subgroup that contains many unique and functional species. The distribution patterns of bacteria in the human intestine not only reflect the different microenvironments in different sections of the intestine but also indicate that bacteria may have unique biological functions corresponding to their proper regions of the intestine. However, describing the functional differences between the bacterial subgroups and their distributions in different individuals is difficult using traditional computational approaches. Here, we first attempted to introduce four effective sets of bacterial features from independent databases. We then presented a novel computational approach to identify potential distinctive features among bacterial subgroups based on a systematic dataset on the gut microbiome from approximately 1,500 human gut bacterial strains. We also established a group of quantitative rules for explaining such distinctions. Results may reveal the microstructural characteristics of the intestinal flora and deepen our understanding on the regulatory role of bacterial subgroups in the human intestine.

Keywords: gut microbiome, bacteria feature, pattern, rule, multi-class classification

INTRODUCTION

Gut microbiome refers to the integral microflora that is located in the human intestine and has symbiosis with human beings (Arumugam et al., 2011;Yatsunenko et al., 2012). According to recent publications, the identified microflora in the human intestine contains tens of trillions of microorganisms including bacteria, fungi, protists, archaea, and viruses (Yatsunenko et al., 2012). Among different subgroups of microorganisms, bacteria are the most significant subgroup that contains unique and functional species between 300 and 1000 (Barcenilla et al., 2000;Chadchan et al., 2019). More than 60% of all microorganisms can be clustered into different bacterial subgroups. In different sections of the human intestine, the species distributions of bacteria are quite different (Reichardt et al., 2014). For instance, in the gut, almost all the identified bacteria are anaerobes; however, in the cecum, aerobic bacteria, another subgroup of bacteria, are predominant (Wells et al., 1987; Kelly et al., 2004). Such distribution patterns of bacteria in the human intestine not only reflect the different microenvironments in different sections of the intestine but also indicate that bacteria may have their unique biological functions corresponding to their proper regions of the intestine. The symbiosis of human beings and bacterial subgroups/clusters maintains the stability of the intestinal microenvironment (Arumugam et al., 2011;Yatsunenko et al., 2012).

In general, the biological functions of symbiotic gut bacteria can be summarized into three major aspects: intestine immune regulation (Kelly et al., 2005), nutrition metabolism regulation (Ramakrishna, 2013), and regulation of gut–brain axis (Foster and McVey Neufeld, 2013; Plummer et al., 2013). First, the gut bacteria can initiate and activate the humoral and adaptive immune responses in the specific region of the gut (Slack et al., 2009; Bunker et al., 2015). As one of the major subgroups of immune response-associated processes in the intestinal immune system, cytokine-associated biological processes are important; different subgroups of gut bacteria have been confirmed to increase different subgroups of cytokines (Atarashi et al., 2013; Schirmer et al., 2016). In addition, most bacteria, such as filamentous bacteria, can activate the musical immune responses, indicating that different subgroups of bacteria can have different biological contributions to immune regulatory processes (Wu et al., 2010). Different subgroups of bacteria also contribute to the digestion and absorption of nutrients through specific nutrition-associated biological functions. For instance, saccharolytic fermentation is a specific fermentation process that helps synthesize unique subtypes of short-chain fatty acids, which are required by various organs, such as the brain, liver, and kidney, and cannot be synthesized independently (Miller and Wolin, 1979; Windey et al., 2012). Different subgroups of gut bacteria contribute to the manufacture of different nutrient subtypes (Windey et al., 2012). Thus, the collaborative contribution of different gut bacterial subgroups can maintain the nutrition supply and physical health of human beings. Importantly, the direct relationship between the gut bacteria and the central nervous system, known as the gut–brain axis, has been confirmed in recent studies (Ghaisas et al., 2016; Kohler et al., 2016). Early in 2004, an independent experiment confirmed that germ-free mice, which do not have gut microbiome, exhibited improved hypothalamic–pituitary axis response compared with normal controls (Riediger et al., 2004). This study directly confirms that the gut microbiomes have potential causal effects on the central nervous system.

Bacterial distribution in the human intestine is significantly diverse and exerts various biological effects on human health. However, describing the functional differences between the bacterial subgroups and their distributions in different individuals is difficult using traditional computational approaches. Therefore, we attempted to introduce four effective sets of features from four independent databases, namely, the Antibiotic Resistance Genes Database (ARGD) (Liu and Pop, 2009), the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013; Jia et al., 2017), the Virulence Factor Database (VFDB) (Liu et al., 2019), and Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa, 2002; Tanabe and Kanehisa, 2012). The combination of features may comprehensively describe the biological functions of different bacterial subgroups and screen their most critical differences. In the present study, using the dataset established by a systematic analysis on the gut microbiome from approximately 1500 human gut bacteria phyla (Zou et al., 2019), we presented a novel computational approach to identify the potential distinctive features among bacterial subgroups and established a group of quantitative rules for explaining such distinctions. We only focused on three bacterial subgroups, namely, Actinobacteria, Bacteroidetes, and Firmicutes, due to the quantitative characteristics of the sequencing data. Our results may reveal the microstructural characteristics of the intestinal flora and deepen our understanding on the regulatory role of bacterial subgroups in the human intestine.

MATERIALS AND METHODS

Datasets

We downloaded the functional annotations of human gut bacteria from the China National GeneBank under Project ID: CNP0000126 (https://db.cngb.org/search/project/CNP0000126/) (Zou et al., 2019). Each human gut bacteria were encoded with 342 Antibiotic Resistance Genes Database (ARDB) annotation features, 259 CARD annotation features, 243 KEGG annotation features, and 149 VFDB annotation features (a total of 993 features). We analyzed three human gut bacteria phyla with number of strains greater than 100, namely, 235 Actinobacteria, 447 Bacteroidetes, and 796 Firmicutes. Fusobacteria with six strains and Proteobacteria with 36 strains were excluded. The goal was to find the functional difference among different human gut bacterial phyla.

Features from different databases have their independent biological significance. The first database (ARDB) was built up to provide a basic summary for antibiotic resistance and facilitate the identification and annotation of novel drug resistance associated genes (Liu and Pop, 2009). Features in such database describes the gene ontology, COD&COG taxonomy, KEGG pathway information (McArthur et al., 2013; Jia et al., 2017), and mutation resistance information of all the annotated genes (Liu and Pop, 2009). Using such features, we can easily describe the biological functions of effective genes and the potential pathogenic effects of specific mutations, classifying mutant and wild-type genes into different types (Liu and Pop, 2009). As for the second database, CARD, it summarizes all the characterized, peer-reviewed resistance determinants and associated antibiotics based on Antibiotic Resistance Ontology (ARO) and AMR gene detection models (McArthur et al., 2013; Jia et al., 2017). Features of such database mainly focused on the description of drug resistance characteristics of different microbial strains (McArthur et al., 2013; Jia et al., 2017). Deferentially, the next database named as VFDB (Liu et al., 2019) turns out to be an integrated and comprehensive online resource for bacterial pathogenic analysis. Features from such databases describe the virulence factors and potential pathogens of various microbial types (Liu et al., 2019). As for the last database, as we have mentioned above, KEGG database (McArthur et al., 2013; Jia et al., 2017) mainly focuses on the functional description of potential microbial genes. Features of such database describe the unique functional characteristics.

Feature Ranking

Of the extracted 993 features from different sources, some features were redundant and not informative. To select the important features that contribute most to the classification tasks, we applied Monte Carlo feature selection (MCFS) (Cai et al., 2018; Chen et al., 2018a; Pan et al., 2018; Chen et al., 2019a; Chen et al., 2019c; Chen et al., 2019e; Li et al., 2019; Pan et al., 2019a; Pan et al., 2019b) to analyze these features and rank them according to their importance. MCFS is a supervised feature selection method based on multiple decision trees (Draminski et al., 2008). MCFS first generates *s* bootstrap sample sets and *m* feature subsets from the original data. A decision tree is grown for each combination of the bootstrap set and feature subset. Accordingly, *t*×*m* trees are constructed in total and used to calculate relative importance (RI) score for each feature with the assumption that the important features should be frequently involved in many growing decision trees. For each feature, RI score is calculated based on the following components: 1) number of splits involved in all nodes of *t*×*m* trees; 2) information gain by each split; and 3) classification accuracies of individual decision trees. Its calculation formula is as follows:

$$RI_g = \sum_{\tau=1}^{t\times m}(wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau))\left(\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau}\right)^v \qquad (1)$$

where $IG(n_g(\tau))$ stands for the gain information of node $n_g(\tau)$, (no.in $n_g(\tau)$) the number of samples in node $n_g(\tau)$, no.in τ the number of samples in tree τ, *wAcc* the weighted accuracy of decision tree τ. *u*, and *v* represent two regular factors, which were all set to one in this study. After obtaining the RI score of each feature, all features were ranked by the decreasing order of their RI scores. MCFS was implemented and downloaded at http://www.ipipan.eu/staff/m.draminski/mcfs.html.

Incremental Feature Selection

After ranking the input features by using MCFS, we determined whether all these features are necessary for classifying Actinobacteria, Bacteroidetes, and Firmicutes. We applied incremental feature selection (IFS) (Zhang et al., 2015a; Zhang et al., 2015b; Zhou et al., 2015; Chen et al., 2017b; Chen et al., 2017c; Liu et al., 2017; Chen et al., 2018b; Zhang et al., 2018; Chen et al., 2019d; Wang and Huang, 2019) with a classifier to the ranked features and selected the discriminate features with the best performance. Basing on the ranked features from MCFS, we constructed a series of feature subsets with step 1, e.g., the first feature subset has the top 1 feature, and the second subset has the top 1 and 2 features. For each feature subset, we trained a classifier on the samples consisting of features from the feature subset and evaluated the classification performance by 10-fold cross-validation. After running the process for all feature subsets, we selected the feature subset with the best performance (i.e., highest Matthews correlation coefficient); this feature subset was called the optimum feature subset.

Rule Learning

Many different supervised classifiers, including black-box and interpretable rule-based methods, exist. Black-box methods cannot explain their predictions in a manner that humans can understand, and rule-based methods can supply classification reasons in a way understandable to humans. In this study, we used an interpretable rule-based classification method with repeated incremental pruning to produce error reduction (RIPPER) (Cohen, 1995; Li et al., 2019; Pan et al., 2019a) (i.e., Jrip algorithm) to classify the samples from three bacterial groups, namely, Actinobacteria, Bacteroidetes, and Firmicutes. In addition, a rule usually consists of if-then statement; simply put, if conditions A and B are met, then we make a certain prediction of yes or no. RIPPER is a greedy method for learning classification rules. This method first generates a good rule covering some samples in the training set. These covered samples are removed, and the remaining training set is used for the next rule. This process of rule generation is repeated until all samples are covered by the learned rules or predefined stop conditions are met. Lastly, the learned rules are further pruned using reduced error pruning.

To quickly implement the RIPPER algorithm mentioned above, a tool "JRip" in Weka (Witten and Frank, 2005) was directly employed in this study. For convenience, its default parameters were used.

Performance Measurement

We used RIPPER as a multiclassification method to classify samples from Actinobacteria, Bacteroidetes, and Firmicutes. The 10-fold cross-validation was adopted for performance evaluation (Huang et al., 2009; Huang et al., 2010; Cai et al., 2012; Chen et al., 2013; Zhang et al., 2015a; Zhao et al., 2018; Zhang et al., 2019; Zhao et al., 2019), and the performance measurements should be appropriate for multiclass classification. Several measurements were employed in this task. They can be divided into two categories. The first measurement category was for each phylum, such as individual accuracy, precision, recall (same as individual accuracy), and Matthews correlation coefficient (MCC) (Matthews, 1975). The other measurement category fully evaluate the performance of the classification method, including overall accuracy and MCC in multi-class (Gorodkin, 2004), as detailed in previous works (Chen et al., 2017a; Li et al., 2018; Chen et al., 2019b; Chen et al., 2019c; Cui and Chen, 2019; Pan et al., 2019a; Pan et al., 2019b). Because MCC in multi-class is widely accepted to be a balanced measurement even if the dataset is of great imbalance, it was selected as the key measurement in our study.

RESULTS

In this study, we extracted 993 features to represent each sample. These features consist of 342 ARDB features, 259 CARD features, 243 KEGG features, and 149 VFDB features, wherein the names and values are given in **Supplementary Table S1**. Then, several advanced computational methods were adopted to analyze these features. The entire procedures are illustrated in **Figure 1**. Clearly, not all features have the same importance for distinguishing samples from different bacterial groups; as such, the features are ranked and selected using the RI scores from MCFS. The RI scores of individual features are given in **Supplementary Table S2**. A total of 432 of all 993 features have RI scores larger than zero and thus have discriminated ability for different bacterial groups. Other features were discarded in the following analysis.

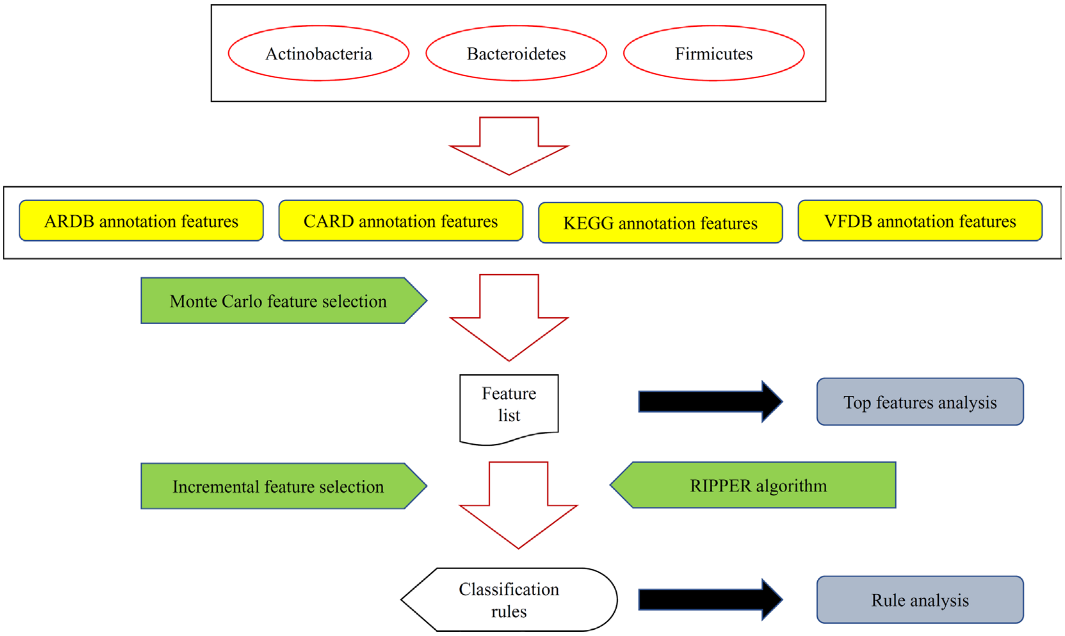


FIGURE 1 | A flow chart to illustrate the procedures of identifying microbiota signature and functional rules for bacterial subtypes in human intestine. Bacteria in three human gut bacteria phyla were represented by four types of features. These features were analyzed by the Monte Carlo feature selection method, resulting in a feature list. For some top features, an extensive analysis was performed. Furthermore, the incremental feature selection method, incorporating the rule learning algorithm (RIPPER algorithm), was applied on the feature list to construct optimal classification rules, which were also extensively analyzed.

To further select the optimum features from the 432 features, we used IFS with RIPPER for sample classification. RIPPER was trained and evaluated on the samples consisting of features from individual feature subsets by 10-fold cross-validation. As shown in **Figure 2**, among the top 432 features, the best MCC in multi-class of 0.998 and an overall accuracy of 0.999 were obtained when the top 153 features were used. The individual accuracy (recall), precision and MCC for each phylum are shown in **Figure 3**. It can be seen that each of these measurements was larger than 0.990, indicating the good performance of RIPPER on top 153 features. In particular, we obtained a high MCC in multi-class of 0.991 and an overall accuracy of 0.995 when only the top 25 features were used. The detailed predicted results were counted as a confusion map, as shown in **Figure 4**. Its performance on each phylum is shown in **Figure 3**, which was a little lower than that of the RIPPER with top 153 features; however, it was still very high. The corresponding performance of the RIPPER with the number of features ranging from 1 to 432 are shown in **Supplementary Table S3**. The results indicate that the interpretable rule-based method RIPPER is close to perfectly classify the samples from Actinobacteria, Bacteroidetes, and Firmicutes.

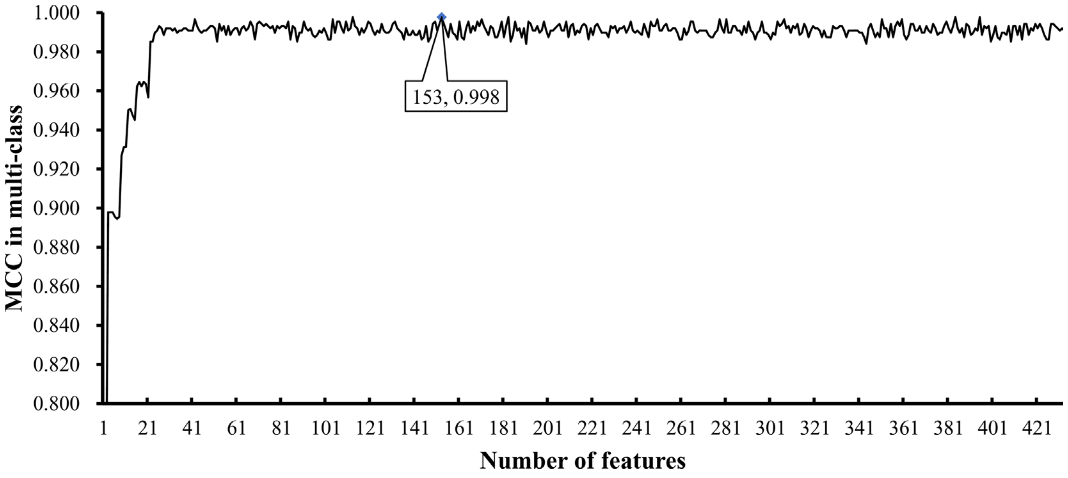


FIGURE 2 | Optimal performance of IFS with RIPPER algorithm. The RIPPER algorithm provided the highest MCC (0.998) when top 153 features were used.

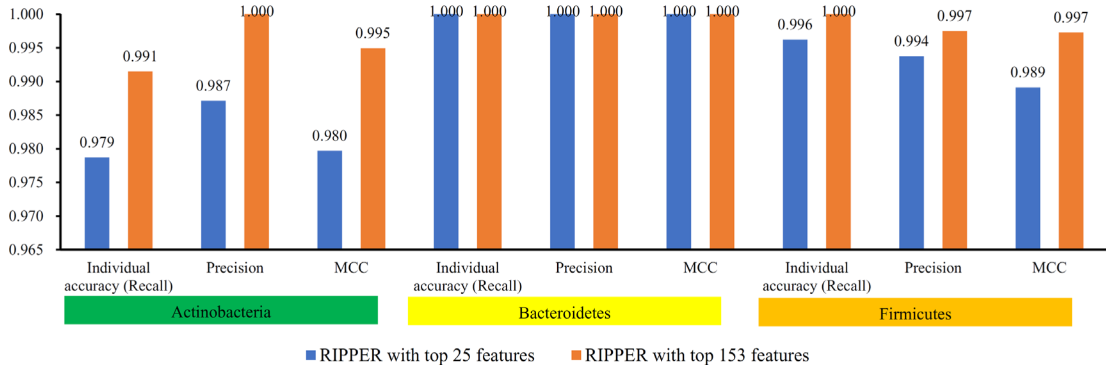


FIGURE 3 | Performance of RIPPER algorithm with top 25 and 153 features on each phylum. The RIPPER algorithm with top 153 features provided nearly perfect classification, while the RIPPER algorithm yielded a little lower performance.

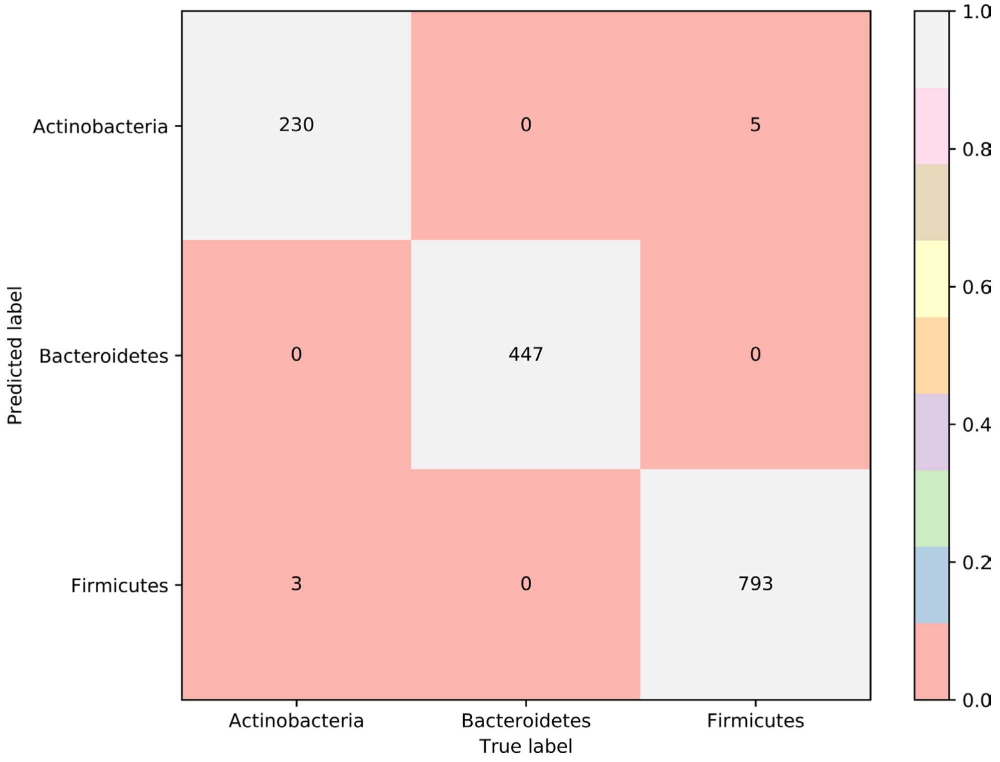


FIGURE 4 | Confusion matrix yielded by the RIPPER algorithm with top 25 features. The accuracy of Bacteroidetes reached 1.000, while those of two other phyla were higher than 0.970, indicating the high performance of RIPPER algorithm with top 25 features.

As mentioned above, RIPPER with top 25 features yielded quite high performance. To indicate the importance of these 25 features, we did the following test: 1000 feature subsets containing 25 features were randomly produced. RIPPER was trained on the samples represented by features from each of these feature subsets and evaluated by 10-fold cross-validation. Obtained MCCs in multi-class are illustrated in a box plot, as shown in **Figure 5**, in which the MCC in multi-class yielded by the RIPPER with top 25 features is also listed. It can be observed that all MCCs in multi-class on randomly produced feature subsets were lower than that yielded by the RIPPER with top 25 features. It is suggested that top 25 features were very important for identifying bacteria in different phyla. Therefore, we established five significant classification rules on all bacteria represented by top 25 features, as listed in **Table 1**, to elucidate how RIPPER can make accurate prediction. The details of these learned rules are discussed below. The results demonstrate the satisfactory discriminate powers of the five produced classification rules for different bacterial groups.

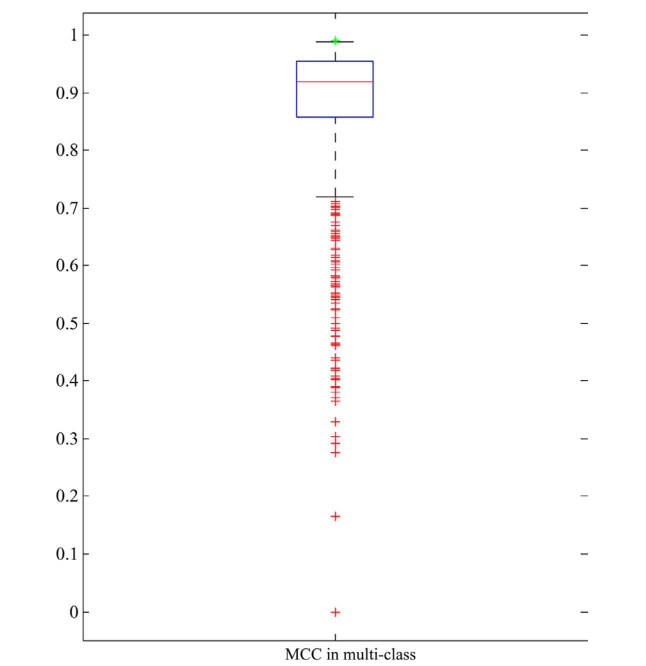


FIGURE 5 | Box plot to show the performance of RIPPER algorithm with 25 features that are randomly selected from all features. The green star strands for the MCC in multi-class yielded by RIPPER algorithm with top 25 features, which is higher than all other MCCs in multi-class on randomly selected 25 features.

TABLE 1 | Five classification rules produced by the RIPPER algorithm for Actinobacteria, Bacteroidetes, and Firmicutes.

Rules	Criteria	Bacteria group
Rule 1	Genetic Information Processing: Folding, sorting, and degradation: Proteasome > = 1	Actinobacteria
Rule 2	(Human Diseases: Drug resistance: Cationic antimicrobial peptide (CAMP) resistance < = 0) and (Genetic Information Processing: Folding, sorting, and degradation: Protein processing in endoplasmic reticulum > = 2)	Actinobacteria
Rule 3	(Cellular Processes: Transport and catabolism: Peroxisome < = 0) and (Genetic Information Processing: Folding, sorting, and degradation: Protein processing in endoplasmic reticulum > = 2) and (Human Diseases: Drug resistance: Cationic antimicrobial peptide (CAMP) resistance < = 1)	Actinobacteria
Rule 4	Organismal Systems: Digestive system: Protein digestion and absorption > = 1	Bacteroidetes
Rule 5	others	Firmicutes

DISCUSSION

In this study, we attempted to integrate different feature sets from ARGD (Liu and Pop, 2009), CARD (McArthur et al., 2013; Jia et al., 2017), VFDB (Liu et al., 2019), and KEGG (Kanehisa, 2002; Tanabe and Kanehisa, 2012) databases. Basing on these collective features and original datasets, we accurately distinguished the common gut bacteria into three major clusters: Actinobacteria, Bacteroidetes, and Firmicutes. We not only identified the crucial features from the four known datasets that contributed most to such clustering but also set up a novel quantitative rule set for the accurate clustering of gut bacteria. All the predicted results (i.e., features and rules) were supported by solid experimental evidence presented in literature. We screened the top features and rules in our optimal prediction list for further discussion and analyses below due to page limitation.

Analysis of Optimal Features for Subtyping of Gut Bacteria

Using machine learning models, we screened a group of proper features to distinguish three common gut bacterial subgroups. The first significant distinctive feature (F_740) is a metabolism describing feature: glycan biosynthesis and lipopolysaccharide biosynthesis associated metabolism. According to recent publications, bacteria from Actinobacteria (King et al., 2009; Alshalchi and Anderson, 2015), Bacteroidetes (Jacobson et al., 2018), and Firmicutes (d'Hennezel et al., 2017) participate in these biological processes. In contrast to Actinobacteria and Bacteroidetes, Firmicutes directly promotes the biosynthesis of lipids and contributes to the pathogenesis of obesity (d'Hennezel et al., 2017). The activation of such metabolic processes was finally decided by the relative abundance of Firmicutes compared with the other bacterial phyla. Therefore, F_740 could be a novel and effective feature for subtyping different bacterial subgroups.

The following feature marked as F_602 describes cell growth and death-associated processes, including apoptosis. In general, the balance between cell growth and death in the intestine is usually regulated and maintained by inflammatory reactions (Neurath et al., 1998; Pickard et al., 2017) and lipopolysaccharide production (Guo et al., 2013). The production of lipopolysaccharides is significant for the survival of gut cells. According to recent publications, lipopolysaccharide production is correlated with the relative abundance ratio between Bacteroidetes and Firmicutes (Jeong et al., 2015; Kim et al., 2016). Therefore, the stable status of cell growth and death-associated processes may be sufficiently effective and sensitive for evaluating the relative abundance of such two major bacterial subtypes, thereby validating the efficacy of our new method.

F_823 describes the general protein digestion and absorption processes of the digestive system, and different bacterial subgroups play different roles in the digestion and absorption of different nutrients (Flint et al., 2012; Valdes et al., 2018). For example, the digestion and absorption of lipids and proteins as a proper instance again; as such, different subgroups of bacteria contribute differently to such processes. In contrast to fat metabolism, a case of protein metabolism, the high abundance of bacterial subgroups, such as Bacteroidetes, indicates the high activation status of protein digestion and absorption (Turnbaugh et al., 2006). Therefore, F_823, as an indicator of the activity degree of protein metabolism, may contribute to the distinction of different bacterial subgroups.

F_608, as a complicated feature describing the formation of biofilm, was screened to distinguish different gut bacterial subgroups. In 2015, a systematic review on microbial biofilms and associated gut diseases confirmed that the abundances of Firmicutes and Bacteroidetes rather than that of Actinobacteria are functionally related to biofilm. The relative contributions of the three clusters of gut bacteria on biofilm regulation would be quite different (von Rosenvinge et al., 2013). Therefore, the biological characteristics of gut biofilm may also be a potential biomarker for the distinction of different bacteria subgroups.

The finally discussed high-ranked feature, named as F_756, describes the biosynthesis of steroid hormone. In 2013, a review on gut microbiome summarized the specific role of steroid hormones in the interactions between the gut bacteria and host humans (Garcia-Gomez et al., 2013). According to this review, only bacteria from clusters such as Actinobacteria, Proteobacteria, and Firmicutes were confirmed to participate in the biosynthesis and metabolism of steroid hormone to date. However, Bacteroidetes does not. In addition, the dominant phyla, such as Actinobacteria and Firmicutes, can express hydroxysteroid dehydrogenase; this phenomenon is essential for steroid hormone metabolism (Kisiela et al., 2012). Therefore, such feature has significant functional importance for bacterial subgrouping.

Analysis of the Optimal Rules for Gut Bacteria Subtyping

The use of our newly presented computational approaches to determine the optimal features has been validated by recent publications. Apart from such qualitative analysis results, quantitative analysis was performed to distinguish different bacterial subgroups. Based on Jrip algorithm, also known as the RIPPER algorithm, we identified five effective rules for explaining the distinction of bacterial subgroups.

The first rule contains one feature describing the biological processes of proteasomes involving folding, sorting, and degradation of functional proteins. According to recent publications, proteasomes are self-compartmentalized proteolytic organelles only identified in Archaea, Actinobacteria, and eukaryotes but not in Bacteroidetes or Firmicutes (Valas and Bourne, 2008; Ziemski et al., 2018). Therefore, regarding such feature as a quantitative parameter for the identification of Actinobacteria is quite reasonable.

The next rule indicates cationic antimicrobial peptide (CAMP) resistance (F12) and protein folding in the endoplasmic reticulum as another two quantitative parameters for the recognition of Actinobacteria subgroup. According to recent reports, cationic antimicrobial peptides mediate the bacterial resistance against most Actinobacteria and Firmicutes (Anaya-Lopez et al., 2013). Therefore, the first parameter may distinguish Actinobacteria and Firmicutes from other bacterial subgroups. As for the next parameter, Actinobacteria has a specific structure called peroxisomes, sharing similar biological functions with the endoplasmic reticulum (Duhita et al., 2010; Gabaldon and Capella-Gutierrez, 2010). Therefore, the combination of the two parameters refers to the accurate identification of Actinobacteria, thereby validating the efficacy and accuracy of our prediction.

Next, the third rule has three parameters involved in protein modification. Apart from parameters F24 and F12, the effective parameter F7 describes the transport and catabolism of peroxisomes, which were identified and discussed to be unique in Actinobacteria, thereby validating our prediction (Duhita et al., 2010; Gabaldon and Capella-Gutierrez, 2010).

The fourth rule is associated with the differential performance of the general protein digestion and absorption processes of the digestive system with different distribution patterns of bacteria. The high activation status of protein digestion and absorption pattern in the gut indicate the abundance of Bacteroidetes (Turnbaugh et al., 2006), corresponding with our rules.

Overall, all optimal features and rules for the distinction of different bacterial subgroups are accurate and efficient with solid publication supports. The accurate clustering of gut bacteria is the foundation for microbiome studies of the human intestine. For a long time, applying microbiome clustering based on sequencing data is difficult and time consuming due to the complicated described feature sets. Here, with the help of machine learning models, we identified the core features for microbiome distinction and set up a group of accurate distinctive rules for explaining such clustering problem. Therefore, using proper machine learning models, the present study reveals an accurate and elaborate panorama for gut microbe and provides a novel tool for further studies on the microbiome.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://db.cngb.org/search/project/CNP0000126/.

AUTHOR CONTRIBUTIONS

All authors contributed to the research and reviewed the manuscript. LC and YZ designed the study. LC and DL performed the experiments. YS, HW, and YL analyzed the results. LC wrote the manuscript.

FUNDING

This work was supported by grants from the Science and Technology Research Projects of Anhui Province (201904b11020043, 201904e01020014), Natural Science Foundation of Anhui Province (1808085MC60), and the National Key R&D Program of China (2018YFD1100104).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01146/full#supplementary-material

SUPPLEMENTARY TABLE S1 | Data matrix of analysis.

SUPPLEMENTARY TABLE S2 | Top features with their importance scores calculated by the MCFS.

SUPPLEMENTARY TABLE S3 | Ten-fold cross-validation performance of IFS with RIPPER algorithm.

REFERENCES

Alshalchi, S. A., and Anderson, G. G. (2015). Expression of the lipopolysaccharide biosynthesis gene lpxD affects biofilm formation of Pseudomonas aeruginosa. *Arch. Microbiol.* 197, 135–145. doi: 10.1007/s00203-014-1030-y

Anaya-Lopez, J. L., Lopez-Meza, J. E., and Ochoa-Zarzosa, A. (2013). Bacterial resistance to cationic antimicrobial peptides. *Crit. Rev. Microbiol.* 39, 180–195. doi: 10.3109/1040841X.2012.699025

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature* 473, 174–180. doi: 10.1038/nature09944

Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., et al. (2013). Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* 500, 232–236. doi: 10.1038/nature12331

Barcenilla, A., Pryde, S. E., Martin, J. C., Duncan, S. H., Stewart, C. S., Henderson, C., et al. (2000). Phylogenetic relationships of butyrate-producing bacteria from the human gut. *Appl. Environ. Microbiol.* 66, 1654–1661. doi: 10.1128/AEM.66.4.1654-1661.2000

Bunker, J. J., Flynn, T. M., Koval, J. C., Shaw, D. G., Meisel, M., Mcdonald, B. D., et al. (2015). Innate and Adaptive Humoral Responses Coat Distinct Commensal Bacteria with Immunoglobulin A. *Immunity* 43, 541–553. doi: 10.1016/j.immuni.2015.08.007

Cai, Y.-D., Zhang, S., Zhang, Y.-H., Pan, X., Feng, K., Chen, L., et al. (2018). Identification of the Gene Expression Rules That Define the Subtypes in Glioma. *J. Clin. Med.* 7, 350. doi: 10.3390/jcm7100350

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 42, 1387–1395. doi: 10.1007/s00726-011-0835-0

Chadchan, S. B., Cheng, M., Parnell, L. A., Yin, Y., Schriefer, A., Mysorekar, I. U., et al. (2019). Antibiotic therapy with metronidazole reduces endometriosis disease progression in mice: a potential role for gut microbiota. *Hum. Reprod.* 34, 1106–1116. doi: 10.1093/humrep/dez041

Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug-drug interactions using chemical interactions. *Curr. Bioinf.* 12, 526–534. doi: 10.2174/1574893611666160618094219

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018a). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell. Biochem.* 119, 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Pan, X., Zhang, Y.-H., Hu, X., Feng, K., Huang, T., et al. (2019a). Primary tumor site specificity is preserved in patient-derived tumor xenograft models. *Front. Genet.* 10, 738. doi: 10.3389/fgene.2019.00738

Chen, L., Pan, X., Zhang, Y.-H., Huang, T., and Cai, Y.-D. (2019b). Analysis of Gene Expression Differences between Different Pancreatic Cells. *ACS Omega* 4, 6421–6435. doi: 10.1021/acsomega.8b02171

Chen, L., Pan, X., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2019c). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* 120, 7068–7081. doi: 10.1002/jcb.27977

Chen, L., Pan, X., Zhang, Y.-H., Liu, M., Huang, T., and Cai, Y.-D. (2019d). Classification of widely and rarely expressed genes with recurrent neural network. *Comput. Struct. Biotechnol. J.* 17, 49–60. doi: 10.1016/j.csbj.2018.12.002

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017b). Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* 5, 26582–26590. doi: 10.1109/ACCESS.2017.2775703

Chen, L., Zeng, W.-M., Cai, Y.-D., and Huang, T. (2013). Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set. *Curr. Bioinf.* 8, 200–207. doi: 10.2174/1574893611308020008

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y. H., Yuan, F., et al. (2019e). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26, 29–39. doi: 10.1038/s41434-018-0051-6

Chen, L., Zhang, Y.-H., Lu, G., Huang, T., and Cai, Y.-D. (2017c). Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. *Artificial Intell. Med.* 76, 27–36. doi: 10.1016/j.artmed.2017.02.001

Chen, L., Zhang, Y. H., Huang, G., Pan, X., Wang, S., Huang, T., et al. (2018b). Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genomics* 293, 137–149. doi: 10.1007/s00438-017-1372-7

Cohen, W. W. (1995). "Fast effective rule induction," in *The twelfth international conference on machine learning*, 115–123. doi: 10.1016/B978-1-55860-377-6.50023-2

Cui, H., and Chen, L. (2019). A Binary Classifier for the Prediction of EC Numbers of Enzymes. *Curr. Proteomics* 16, 381–389. doi: 10.2174/1570164616666190126103036

d'Hennezel, E., Abubucker, S., Murphy, L. O., and Cullen, T. W. (2017). Total lipopolysaccharide from the human gut microbiome silences toll-like receptor signaling. *mSystems* 2, e00046–e00017. doi: 10.1128/mSystems.00046-17

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Duhita, N., Le, H. A., Satoshi, S., Kazuo, H., Daisuke, M., and Takao, S. (2010). The origin of peroxisomes: The possibility of an actinobacterial symbiosis. *Gene* 450, 18–24. doi: 10.1016/j.gene.2009.09.014

Flint, H. J., Scott, K. P., Louis, P., and Duncan, S. H. (2012). The role of the gut microbiota in nutrition and health. *Nat. Rev. Gastroenterol. Hepatol.* 9, 577–589. doi: 10.1038/nrgastro.2012.156

Foster, J. A., and McVey Neufeld, K. A. (2013). Gut-brain axis: how the microbiome influences anxiety and depression. *Trends Neurosci.* 36, 305–312. doi: 10.1016/j.tins.2013.01.005

Gabaldon, T., and Capella-Gutierrez, S. (2010). Lack of phylogenetic support for a supposed actinobacterial origin of peroxisomes. *Gene* 465, 61–65. doi: 10.1016/j.gene.2010.06.004

Garcia-Gomez, E., Gonzalez-Pedrajo, B., and Camacho-Arroyo, I. (2013). Role of sex steroid hormones in bacterial-host interactions. *Biomed. Res. Int.* 2013, 928290. doi: 10.1155/2013/928290

Ghaisas, S., Maher, J., and Kanthasamy, A. (2016). Gut microbiome in health and disease: Linking the microbiome-gut-brain axis and environmental factors in the pathogenesis of systemic and neurodegenerative diseases. *Pharmacol. Ther.* 158, 52–62. doi: 10.1016/j.pharmthera.2015.11.012

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* 28, 367–374. doi: 10.1016/j.compbiolchem.2004.09.006

Guo, S., Al-Sadi, R., Said, H. M., and Ma, T. Y. (2013). Lipopolysaccharide causes an increase in intestinal tight junction permeability in vitro and in vivo by inducing enterocyte membrane expression and localization of TLR-4 and CD14. *Am. J. Pathol.* 182, 375–387. doi: 10.1016/j.ajpath.2012.10.014

Huang, T., Cui, W., Hu, L., Feng, K., Li, Y. X., and Cai, Y. D. (2009). Prediction of pharmacological and xenobiotic responses to drugs based on time course gene expression profiles. *PloS One* 4, e8126. doi: 10.1371/journal.pone.0008126

Huang, T., Wang, P., Ye, Z. Q., Xu, H., He, Z., Feng, K. Y., et al. (2010). Prediction of Deleterious Non-Synonymous SNPs Based on Protein Interaction Network and Hybrid Properties. *PloS One* 5, e11900. doi: 10.1371/journal.pone.0011900

Jacobson, A. N., Choudhury, B. P., and Fischbach, M. A. (2018). The Biosynthesis of Lipooligosaccharide from Bacteroides thetaiotaomicron. *MBio* 9, e02289–e02217. doi: 10.1128/mBio.02289-17

Jeong, J. J., Kim, K. A., Jang, S. E., Woo, J. Y., Han, M. J., and Kim, D. H. (2015). Orally administrated Lactobacillus pentosus var. plantarum C29 ameliorates age-dependent colitis by inhibiting the nuclear factor-kappa B signaling pathway via the regulation of lipopolysaccharide production by gut microbiota. *PloS One* 10, e0116533. doi: 10.1371/journal.pone.0116533

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., et al. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 45, D566–D573. doi: 10.1093/nar/gkw1004

Kanehisa, M. (2002). The KEGG database. *Novartis Found Symp.* 247, 91–101; discussion 101-103, 119-128, 244-152. doi: 10.1002/0470857897.ch8

Kelly, D., Campbell, J. I., King, T. P., Grant, G., Jansson, E. A., Coutts, A. G., et al. (2004). Commensal anaerobic gut bacteria attenuate inflammation by regulating nuclear-cytoplasmic shuttling of PPAR-gamma and RelA. *Nat. Immunol.* 5, 104–112. doi: 10.1038/ni1018

Kelly, D., Conway, S., and Aminov, R. (2005). Commensal gut bacteria: mechanisms of immune modulation. *Trends Immunol.* 26, 326–333. doi: 10.1016/j.it.2005.04.008

Kim, K. A., Jeong, J. J., Yoo, S. Y., and Kim, D. H. (2016). Gut microbiota lipopolysaccharide accelerates inflamm-aging in mice. *BMC Microbiol.* 16, 9. doi: 10.1186/s12866-016-0625-7

King, J. D., Kocincova, D., Westman, E. L., and Lam, J. S. (2009). Review: Lipopolysaccharide biosynthesis in Pseudomonas aeruginosa. *Innate. Immun.* 15, 261–312. doi: 10.1177/1753425909106436

Kisiela, M., Skarka, A., Ebert, B., and Maser, E. (2012). Hydroxysteroid dehydrogenases (HSDs) in bacteria: a bioinformatic perspective. *J. Steroid Biochem. Mol. Biol.* 129, 31–46. doi: 10.1016/j.jsbmb.2011.08.002

Kohler, C. A., Maes, M., Slyepchenko, A., Berk, M., Solmi, M., Lanctot, K. L., et al. (2016). The gut-brain axis, including the microbiome, leaky gut and bacterial translocation: mechanisms and pathophysiological role in Alzheimer's disease. *Curr. Pharm. Des.* 22, 6152–6166. doi: 10.2174/1381612822666160907093807

Li, J., Chen, L., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2018). A computational method for classifying different human tissues with quantitatively tissue-specific expressed genes. *Genes* 9, 449. doi: 10.3390/genes9090449

Li, J., Lu, L., Zhang, Y. H., Xu, Y., Liu, M., Feng, K., et al. (2019). Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene. Ther*. doi: 10.1038/s41417-019-0105-y

Liu, B., and Pop, M. (2009). ARDB–Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 37, D443–D447. doi: 10.1093/nar/gkn656

Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* 47, D687–D692. doi: 10.1093/nar/gky1080

Liu, L., Chen, L., Zhang, Y. H., Wei, L., Cheng, S., Kong, X., et al. (2017). Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection. *J. Biomol. Struct. Dyn.* 35, 312–329. doi: 10.1080/07391102.2016.1138142

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA)-Protein Structure* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., et al. (2013). The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* 57, 3348–3357. doi: 10.1128/AAC.00419-13

Miller, T. L., and Wolin, M. J. (1979). Fermentations by saccharolytic intestinal bacteria. *Am. J. Clin. Nutr.* 32, 164–172. doi: 10.1093/ajcn/32.1.164

Neurath, M. F., Becker, C., and Barbulescu, K. (1998). Role of NF-kappaB in immune and inflammatory responses in the gut. *Gut* 43, 856–860. doi: 10.1136/gut.43.6.856

Pan, X., Chen, L., Feng, K.-Y., Hu, X.-H., Zhang, Y.-H., Kong, X.-Y., et al. (2019a). Analysis of Expression Pattern of snoRNAs in Different Cancer Types with Machine Learning Algorithms. *Int. J. Mol. Sci.* 20, 2185. doi: 10.3390/ijms20092185

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., et al. (2019b). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294, 95–110. doi: 10.1007/s00438-018-1488-4

Pan, X., Hu, X., Zhang, Y.-H., Feng, K., Wang, S. P., Chen, L., et al. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* 9, 208. doi: 10.3390/genes9040208

Pickard, J. M., Zeng, M. Y., Caruso, R., and Nunez, G. (2017). Gut microbiota: Role in pathogen colonization, immune responses, and inflammatory disease. *Immunol. Rev.* 279, 70–89. doi: 10.1111/imr.12567

Plummer, M. P., Meier, J. J., and Deane, A. M. (2013). The gut-brain axis in the critically ill: is glucagon-like peptide-1 protective in neurocritical care? *Crit. Care* 17, 163. doi: 10.1186/cc12758

Ramakrishna, B. S. (2013). Role of the gut microbiota in human nutrition and metabolism. *J. Gastroenterol. Hepatol.* 28 Suppl 4, 9–17. doi: 10.1111/jgh.12294

Reichardt, N., Duncan, S. H., Young, P., Belenguer, A., Mcwilliam Leitch, C., Scott, K. P., et al. (2014). Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *ISME J.* 8, 1323–1335. doi: 10.1038/ismej.2014.14

Riediger, T., Zuend, D., Becskei, C., and Lutz, T. A. (2004). The anorectic hormone amylin contributes to feeding-related changes of neuronal activity in key structures of the gut-brain axis. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 286, R114–R122. doi: 10.1152/ajpregu.00333.2003

Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., et al. (2016). Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* 1671125-1136, e1128. doi: 10.1016/j.cell.2016.10.020

Slack, E., Hapfelmeier, S., Stecher, B., Velykoredko, Y., Stoel, M., Lawson, M. A., et al. (2009). Innate and adaptive immunity cooperate flexibly to maintain host-microbiota mutualism. *Science* 325, 617–620. doi: 10.1126/science.1172747

Tanabe, M., and Kanehisa, M. (2012). Using the KEGG database resource. *Curr. Protoc. Bioinf.* 38, 1.12.1–1.12.43. doi: 10.1002/0471250953.bi0112s38

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444, 1027–1031. doi: 10.1038/nature05414

Valas, R. E., and Bourne, P. E. (2008). Rethinking proteasome evolution: two novel bacterial proteasomes. *J. Mol. Evol.* 66, 494–504. doi: 10.1007/s00239-008-9075-7

Valdes, A. M., Walter, J., Segal, E., and Spector, T. D. (2018). Role of the gut microbiota in nutrition and health. *BMJ* 361, k2179. doi: 10.1136/bmj.k2179

von Rosenvinge, E. C., O'may, G. A., Macfarlane, S., Macfarlane, G. T., and Shirtliff, M. E. (2013). Microbial biofilms and gastrointestinal diseases. *Pathog. Dis.* 67, 25–38. doi: 10.1111/2049-632X.12020

Wang, S.-B., and Huang, T. (2019). The early detection of asthma based on blood gene expression. *Mol. Biol. Rep*. 46, 217–223. doi: 10.1007/s11033-018-4463-6

Wells, C. L., Maddaus, M. A., Reynolds, C. M., Jechorek, R. P., and Simmons, R. L. (1987). Role of anaerobic flora in the translocation of aerobic and facultatively anaerobic intestinal bacteria. *Infect. Immun.* 55, 2689–2694.

Windey, K., De Preter, V., and Verbeke, K. (2012). Relevance of protein fermentation to gut health. *Mol. Nutr. Food Res.* 56, 184–196. doi: 10.1002/mnfr.201100542

Witten, IH, and Frank, E, editors. (2005). *Data Mining:Practical Machine Learning Tools and Techniques*. San Francisco, CA, USA: Morgan Kaufmann.

Wu, H. J., Ivanov, I., Darce, J., Hattori, K., Shima, T., Umesaki, Y., et al. (2010). Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells. *Immunity* 32, 815–827. doi: 10.1016/j.immuni.2010.06.001

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Zhang, N., Huang, T., and Cai, Y. D. (2015a). Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol. Genet. Genomics* 290, 343–352. doi: 10.1007/s00438-014-0922-5

Zhang, P. W., Chen, L., Huang, T., Zhang, N., Kong, X. Y., and Cai, Y. D. (2015b). Classifying ten types of major cancers based on reverse phase protein array profiles. *PloS One* 10, e0123147. doi: 10.1371/journal.pone.0123147

Zhang, T. M., Huang, T., and Wang, R. F. (2018). Cross talk of chromosome instability, CpG island methylator phenotype and mismatch repair in colorectal cancer. *Oncol. Lett.* 16, 1736–1746. doi: 10.3892/ol.2018.8860

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/ACCESS.2019.2944177

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinf*. doi: 10.2174/1574893614666190220114644

Zhao, X., Chen, L., and Lu, J. (2018). A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* 306, 136–144. doi: 10.1016/j.mbs.2018.09.010

Zhou, Y., Zhang, N., Li, B. Q., Huang, T., Cai, Y. D., and Kong, X. Y. (2015). A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. *J. Biomol. Struct. Dyn.* 33, 2479–2490. doi: 10.1080/07391102.2014.1001793

Ziemski, M., Jomaa, A., Mayer, D., Rutz, S., Giese, C., Veprintsev, D., et al. (2018). Cdc48-like protein of actinobacteria (Cpa) is a novel proteasome interactor in mycobacteria and related organisms. *Elife* 7, e34055. doi: 10.7554/eLife.34055

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., et al. (2019). 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nat. Biotechnol.* 37, 179–185. doi: 10.1038/s41587-018-0008-8

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2019
doi: 10.3389/fgene.2019.01154

The Gene Expression Biomarkers for Chronic Obstructive Pulmonary Disease and Interstitial Lung Disease

*Yangwei Yao, Yangyang Gu, Meng Yang, Dakui Cao and Fengjie Wu**

Department of Pulmonary and Critical Care Medicine, The Second Hospital of Jiaxing, Jiaxing, China

COPD (chronic obstructive pulmonary disease) and ILD (interstitial lung disease) are two common respiratory diseases. They share similar clinical traits but require different therapeutic treatments. Identifying the biomarkers that are differentially expressed between them will not only help the diagnosis of COPD and ILD, but also provide candidate drug targets that may facilitate the development of new treatment for COPD and ILD. Due to the irreversible complex pathological changes of COPD, there are very limited therapeutic options for COPD patients. In this study, we analyzed the gene expression profiles of two datasets: one training dataset that includes 144 COPD patients and 194 ILD patients, and one test dataset that includes 75 COPD patients and 61 ILD patients. Advanced feature selection methods, mRMR (minimal Redundancy Maximal Relevance) and incremental feature selection (IFS), were applied to identify the 38-gene biomarker. An SVM (support vector machine) classifier was built based on the 38-gene biomarker. Its accuracy, sensitivity, and specificity on training dataset evaluated by leave one out cross-validation were 0.905, 0.896, and 0.912, respectively. And on independent test dataset, the accuracy, sensitivity, and specificity on were as great as and were 0.904, 0.933, and 0.869, respectively. The biological function analysis of the 38 genes indicated that many of them can be potential treatment targets that may benefit COPD and ILD patients.

Keywords: chronic obstructive pulmonary disease, interstitial lung disease, biomarker, gene expression, treatment target

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS),
China

Reviewed by:
Ning Li,
Wuhan University,
China
Wen-Yu Chen,
First Hospital of Jiaxing,
China

****Correspondence:***
Fengjie Wu
Feng_wju@163.com

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

Received: *01 August 2019*
Accepted: *22 October 2019*
Published: *20 November 2019*

Citation:
Yao Y, Gu Y, Yang M, Cao D and Wu F (2019) The Gene Expression Biomarkers for Chronic Obstructive Pulmonary Disease and Interstitial Lung Disease. Front. Genet. 10:1154.
doi: 10.3389/fgene.2019.01154

INTRODUCTION

COPD (chronic obstructive pulmonary disease) and ILD (interstitial lung disease) are both common lung diseases (Andersen et al., 2013). And cigarette smoking is the biggest risk factor for COPD and ILD (Caminati et al., 2012). About 20% smokers will develop COPD (Bosse, 2012). COPD is also an independent risk factor of lung cancer. Both emphysema and non-emphysema COPD phenotypes significantly increased the risk of lung cancer (Wang et al., 2018). In addition, epidemiological studies have found that COPD increases the risk of lung cancer by two to six times, regardless of whether there is a history of smoking or not (Papi et al., 2004; Young et al., 2009). Since the complex pathological changes in COPD and most of ILD patients are not irreversible, the diseases cause extensive mortality and are great public health problems worldwide (Vogelmeier et al., 2017).

Although COPD and ILD share many common traits and have similar clinical phenotypes, their treatments and the therapeutic effects are different. The recommended treatments for COPD patients are smoking cessation and drugs that treat bronchoconstriction and inflammation, such as methylxanthines, β-adrenoceptor agonists, corticosteroids, phosphodiesterase type 4 (PDE-4) inhibitors, and anticholinergics (Andersen et al., 2013), while the ILD patients are treated with immunosuppressive agents, such as alkylating nitrogen mustard (du Bois, 2010). Inhaled corticosteroids (ICS) are important in managing exacerbations and symptoms in COPD (Lakshmi et al., 2017). However, a significant percentage of patients respond poorly or not at all to pharmacotherapies, especially for patients with severe disease (Nixon et al., 2017). In addition, poor adherence to medication is an essential factor in treatment failure. Therefore, new therapy strategies are needed urgently.

It is critical to classify COPD patients from ILD patients since it is the first step for choosing the right treatment. As we mentioned, COPD and ILD share similar pathogeny and have similar clinical phenotype; it is difficult to discriminate these two diseases and the underlying mechanisms of COPD and ILD are largely unknown. Identifying the biomarkers for COPD and ILD will not only provide a tool for disease diagnosis, but also reveal novel insights of the pathological mechanisms and help developing new treatment to benefit the survival of patients. Microarray is a reliable technology to measure the expression level of thousands of genes simultaneously and has been proven to be great data source for discovering biomarkers.

In this study, we analyzed two gene expression datasets of COPD and ILD: one training dataset of Agilent-028004 SurePrint G3 Human GE 8x60K Microarray including 144 COPD patients and 194 ILD patients, and one independent test data of Agilent-014850 Whole Human Genome Microarray 4x44K G4112F including 75 COPD patients and 61 ILD patients. Advanced feature selection methods, mRMR (minimal Redundancy Maximal Relevance) and IFS (incremental feature selection), were applied to get the optimal biomarkers on training dataset. The SVM (support vector machine) method was used to construct the classifier on training dataset and tested on independent test dataset. The 37-gene classifier achieved great performance on training and test datasets. The accuracies on training data and test data were 0.964 and 0.904, respectively. The 37 selected genes were involved in key biological pathways and functions of COPD and ILD. These results provided novel insight for understanding the mechanisms of COPD and ILD and shed light on new treatment development.

METHODS

The Gene Expression Profiles of COPD and ILD Patients

The gene expression profiles of COPD and ILD patients were downloaded from GEO (Gene Expression Omnibus) with accession number of GSE47460 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE47460). The original data were generated by Peng et al. (2016). They measured the gene expression levels of 144 COPD patients and 194 ILD patients with Agilent-028004 SurePrint G3 Human GE 8x60K Microarray and 75 COPD patients and 61 ILD patients with Agilent-014850 Whole Human Genome Microarray 4x44K G4112F. We extracted the common 15,180 genes between these two microarray platforms and quantile normalized the two datasets. Then the first dataset of 144 COPD patients and 194 ILD patients were considered as training dataset, while the second dataset of 75 COPD patients and 61 ILD patients were considered as independent test dataset.

Biomarker Selection Using mRMR and IFS Methods

We adopted the mRMR (minimal Redundancy Maximal Relevance) method (Peng et al., 2005) to rank the genes on the training dataset. The mutual information-based mRMR method is widely used and has been used in solving many biomedical problems (Niu et al., 2013; Zhao et al., 2013; Zhou et al., 2015). The C/C++ version mRMR program was downloaded from http://home.penglab.com/proj/mRMR/. Unlike the univariate method, such as t test and ANOVA (analysis of variance), mRMR considers not only the relevance between genes and disease types but also the redundancies between genes.

Ω, Ω_s, and Ω_t were used to represent the complete set of all 15,180 (N) candidate genes for biomarker ranking, the selected m genes, and the to-be-selected n genes, respectively. The relevance of gene g from Ω_t with disease type t can be measured with mutual information (*I*) (Sun et al., 2012; Huang and Cai, 2013):

$$\mathrm{D} = I(g, \mathrm{t}) \tag{1}$$

And the redundancy R of the gene g with the selected genes in Ω_s are

$$\mathrm{R} = \frac{1}{m}\left(\sum_{g_i \in \Omega_s} I\left(g, g_i\right)\right) \tag{2}$$

The goal of this algorithm is to get the gene g_j from Ω_s that has maximum relevance with disease type t and minimum redundancy with the selected genes in Ω_s, i.e. maximize the mRMR function

$$\max_{g_j \in \Omega_t}\left[I\left(g_j, t\right) - \frac{1}{m}\left(\sum_{g_i \in \Omega_s} I\left(g, g_i\right)\right)\right]\left(j = 1, 2, \ldots, n\right) \tag{3}$$

The evaluation procedure will be continued for N rounds, and all the genes will be ranked as a list

$$\mathrm{S} = \left\{g_1^{'}, g_2^{'}, \ldots, g_h^{'}, \ldots, g_N^{'}\right\} \tag{4}$$

The index h reflects the trade-off between relevance with disease type and redundancy with selected genes. The smaller the index h is, the better the discriminating power the gene has.

Based on the top 500 mRMR genes, we constructed 500 SVM classifiers and applied an IFS method (Jiang et al., 2013; Li et al., 2014; Shu et al., 2014; Zhang et al., 2014a; Zhang et al., 2015) to identify the optimal genes as biomarker. Each candidate gene set $S_k = \left\{g_1^{'}, g_2^{'}, \ldots, g_k^{'}\right\}\left(1 \le k \le 500\right)$ included the top k genes in the mRMR list.

Based on the leave-one-out cross-validation (LOOCV) accuracy of each candidate gene set on the training dataset, an IFS curve can be plotted. The x-axis denoted the number of top genes that were used to train the SVM classifier, and the y-axis denoted the LOOCV accuracies of trained classifiers. Based on the IFS curve, we can choose the right cutoff of gene numbers to achieve a good prediction performance.

Prediction Performance Evaluation of the Classifier

We used LOOCV (Cui et al., 2013; Yang et al., 2014) to evaluate the prediction performance of the SVM classifiers on the training dataset and then independently tested the final classifier that was trained using all training data on the independent test dataset. During LOOCV on training dataset, all of the N training samples were tested one by one. In each round, one sample was used for testing of the prediction model trained with all the other N-1 samples. After N rounds, all samples were tested one time, and the predicted disease types were compared with the actual disease types. The final classifier was trained using all the training samples and then tested on the independent test dataset. **Figure 1** showed the flowchart of biomarker selection, classifier construction, and prediction performance evaluation. The SVM method was applied using the svm function with default parameters in R package e10171 (https://cran.r-project.org/web/packages/e1071/).

Accuracy (ACC), Sensitivity (Sn), and Specificity (Sp) were calculated to evaluate the prediction performance

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$S_n = \frac{TP}{TP + FN} \tag{6}$$

$$S_p = \frac{TN}{TN + FP} \tag{7}$$

where TP, TN, FP, and FN stand for true positive (COPD), true negative (ILD), false positive (COPD), and false negative (ILD), respectively.

RESULTS AND DISCUSSION

The genes that showed different expression pattern between COPD and ILD patients

We obtained the top 500 most discriminative genes of COPD and ILD patient samples using the mRMR method on the training dataset. The mRMR ranked the genes based on their relevance with disease types, COPD or ILD, and their redundancy with selected genes. Both the relevance and redundancy were

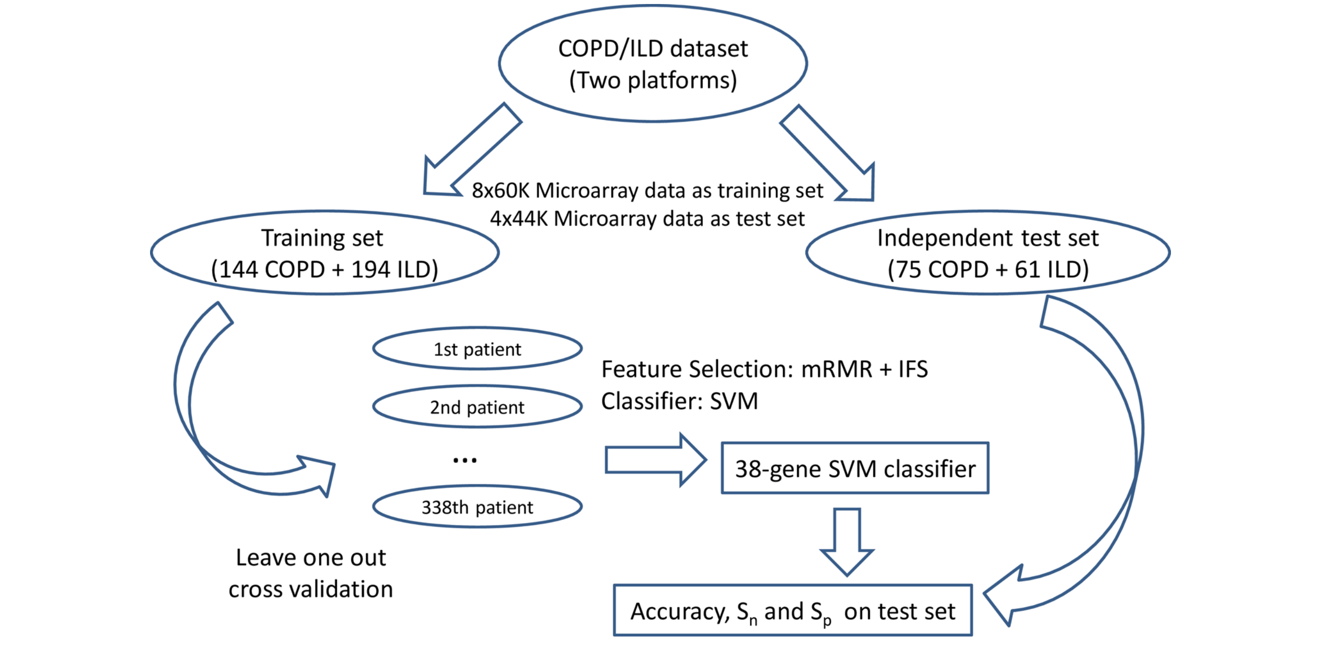


FIGURE 1 | The flowchart of biomarker selection, classifier construction, and prediction performance evaluation. First, the COPD/ILD samples were divided into training dataset and test dataset based on their platform: the 144 COPD samples and 194 ILD samples profiled with 8x60K Microarray was the training set; the 75 COPD samples and 61 ILD samples profiled with 4x44K Microarray were the test set. Then in the training set, we applied mRMR and IFS to select the optimal number of genes as biomarkers and evaluated its performance on the training dataset using leave-one-out cross-validation. At last, the final 38-gene SVM classifier was trained using all training dataset and tested on the independent test dataset. The accuracy, sensitivity, and specificity were calculated to objectively evaluate the prediction performance of the 38-gene classifier.

measured with mutual information. The mutual information has been proven to be a better statistic than correlation and was widely used. The top 500 mRMR genes were given in **Table S1**.

The Optimal Biomarkers Identified From the mRMR Gene List With IFS Methods

After mRMR analysis, the genes were ranked based on the gene expression profiles on training dataset. But we still did not know how many top genes should we choose. And the ideal biomarkers should use less genes and achieve great performance. Therefore, we applied the IFS procedure to select the optimal number of top mRMR genes to form the biomarker gene set. During each round of IFS, different numbers of top genes were used and the corresponding prediction performance, i.e., the LOOCV accuracy on training dataset, were calculated. The relationship between the number of genes and prediction accuracies was plotted as an IFS curve as shown in **Figure 2**. It can be seen that when 94 genes were used, the LOOCV accuracy on training dataset was the highest. But even early, when only 38 genes were used, the accuracy was over 0.90. To consider both using less genes and achieving higher prediction accuracy, we chose the 38 genes as the optimal biomarker gene set since increasing the number of genes will not significantly increase the accuracy any more after the 38 genes were used. The 38 genes were shown in **Table 1**.

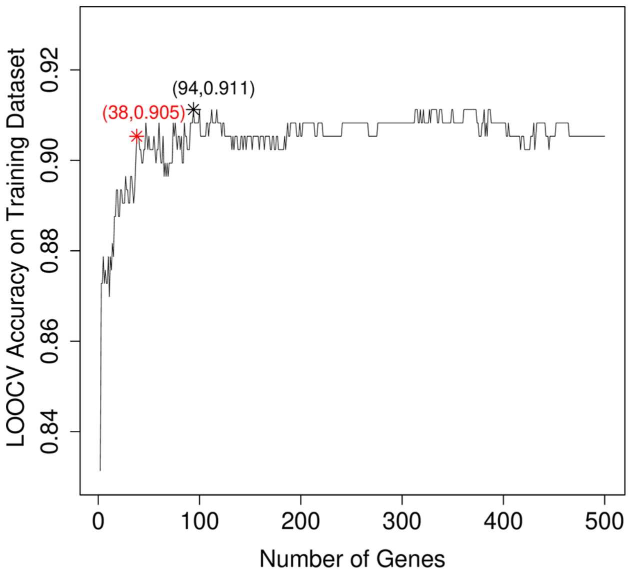


FIGURE 2 | The IFS curve that showed how the prediction performance improved when more and more genes were used to construct the classifier. The IFS curve explained the relationship between the number of genes and prediction accuracies. The x-axis denoted the number of top genes that were used to train the SVM classifier, and the y-axis denoted the LOOCV accuracies of trained classifiers. The highest accuracy was achieved when 94 genes were used. But after 38 genes were used, the IFS curve entered the plateau area and did not increase too much even when more and more genes were included. To consider both the model complexity and model performance, we chose the 38 genes as the optimal biomarker gene set.

TABLE 1 | The 38 genes selected by mRMR and IFS methods.

Order	Symbol	Name	Score
1	HBEGF	Heparin binding EGF like growth factor	0.288
2	DIO2	Iodothyronine deiodinase 2	0.187
3	CLCN3	Chloride voltage-gated channel 3	0.115
4	SEPT4	Septin 4	0.120
5	FAT1	FAT atypical cadherin 1	0.120
6	CTSE	Cathepsin E	0.116
7	CRIP1	Cysteine rich protein 1	0.108
8	ACADVL	Acyl-CoA dehydrogenase, very long chain	0.112
9	CNTN3	Contactin 3	0.118
10	UQCRQ	Ubiquinol-cytochrome c reductase complex III subunit VII	0.116
11	ASPN	Asporin	0.111
12	ZNF786	Zinc finger protein 786	0.110
13	RARRES2	Retinoic acid receptor responder 2	0.107
14	BTC	Betacellulin	0.111
15	FNDC1	Fibronectin type III domain containing 1	0.114
16	DUSP1	Dual specificity phosphatase 1	0.113
17	C6orf145	PX domain containing 1	0.104
18	NUTF2	Nuclear transport factor 2	0.105
19	TNN	Tenascin N	0.101
20	COQ9	Coenzyme Q9	0.103
21	SCG5	Secretogranin V	0.105
22	BCHE	Butyrylcholinesterase	0.099
23	NR4A2	Nuclear transport factor 2	0.100
24	HS6ST3	Heparan sulfate 6-O-sulfotransferase 3	0.103
25	SHE	Src homology 2 domain containing E	0.102
26	C20orf111	Oxidative stress responsive serine rich 1	0.098
27	REEP2	Receptor accessory protein 2	0.099
28	C19orf63	ER membrane protein complex subunit 10	0.097
29	IRS2	Nuclear receptor subfamily 4 group A member 2	0.098
30	FA2H	Fatty acid 2-hydroxylase	0.094
31	ACTL6A	Actin like 6A	0.094
32	NR4A3	Nuclear receptor subfamily 4 group A member 3	0.093
33	DAO	D-amino acid oxidase	0.095
34	VNN2	Vanin 2	0.093
35	IGFL2	IGF like family member 2	0.094
36	ZNF692	Zinc finger protein 692	0.093
37	CAMK1D	Calcium/calmodulin-dependent protein kinase ID	0.091
38	HCAR2	Hydroxycarboxylic acid receptor 2	0.092

The Prediction Performance of the 38-Gene Classifier

The 38 genes were chosen from the genome wide 15,180 genes based on mRMR and IFS methods. To objectively evaluate their prediction power, we calculated not only the LOOCV accuracy, sensitivity, and specificity on training dataset, but also the accuracy sensitivity and specificity on independent test dataset. The confusion matrix of predicted disease types and actual disease types on both training and test datasets were shown in **Table 2**. On training dataset, the LOOCV accuracy, sensitivity, and specificity were 0.905, 0.896, and 0.912, respectively. More importantly, the accuracy, sensitivity, and specificity on

TABLE 2 | The confusion matrix of predicted disease types and actual disease types on both training and test datasets.

Leave one out cross validation on Training set*			**Independent test on test set***		
	Actual COPD	**Actual ILD**		**Actual COPD**	**Actual ILD**
Predicted COPD	129	17	Predicted COPD	70	8
Predicted ILD	15	177	Predicted ILD	5	53
Accuracy: 0.905	Sensitivity: 0.896	Specificity: 0.912	Accuracy: 0.904	Sensitivity: 0.933	Specificity: 0.869

**COPD was considered as positive sample and ILD was considered as negative samples during sensitivity and specificity calculation.*

independent test dataset were as great as on the training dataset and were 0.904, 0.933, and 0.869, respectively.

To more intuitively demonstrate the discriminative power of these 38 genes for COPD and ILD samples, we combined the training dataset samples and test dataset samples and draw a heatmap using these 38 genes (**Figure 3**). It can be seen that even without advanced machine learning algorithm, such as SVM, the simple hierarchical clustering can group most COPD and ILD samples into the right clusters. And the upregulation and downregulation patterns of these 38 genes were very clear between COPD and ILD patients.

We also calculated the results of the 94 genes and plotted their heatmap as **Figure S1**. On training dataset, the LOOCV accuracy, sensitivity, and specificity of the 94-gene classifier were 0.911, 0.889, and 0.928, respectively. On independent test dataset, the accuracy, sensitivity, and specificity of the 94-gene classifier were 0.897, 0.933, and 0.852, respectively. The performance of the 94 genes was close to the 38 genes on both training and independent test datasets. The 38 genes were even slightly better than the 94 genes on independent test dataset.

The Biological Significance of the 38-Gene Biomarkers

As shown in **Table 1**, the first gene on the mRMR list was HBEGF (heparin binding EGF like growth factor). From **Figure 2**, it can be seen that HBEGF was highly expressed in COPD patients. HBEGF is a key member of the EGFR pathway. Its expression level has been reported to be increased in COPD samples and were significantly correlated with both diffusing capacity of the lung for carbon monoxide (DLCO) and Forced Expiratory Volume in 1 second (FEV1), two major clinical variables for COPD (Cockayne et al., 2012). We investigated the tissue specific expression pattern of HBEGF in ARCHS4 (Lachmann et al., 2018) and **Figure 4**, which were retrieved from ARCHS4, showed that HBEGF is very specifically highly expressed in lung.

The second gene was DIO2 (iodothyronine deiodinase 2). DIO2 plays an important role in biologically active triiodothyronine synthesis. Its expression level was consistent with the degree of lung injury: the more the lung injury, the higher the expression of DIO2 (Ma et al., 2011). Clearly, DIO2 is key for the inflammatory response (Ma et al., 2011). And COPD is a complex chronic inflammatory disease involving the dysfunction of a variety of inflammatory mediators (Thorley and Tetley, 2007). DIO2 could be a key factor in the inflammatory mechanism of COPD (Barnes, 2017).

CLCN3 (chloride voltage-gated channel 3) ranked third on the mRMR list. It has been reported that the CLCN3 mRNA was expressed in fetal airway epithelia and played important roles in pulmonary epithelium developing of human lung (Lamb et al., 2001). As we have known, COPD mainly affects pulmonary epithelium (Hiemstra et al., 1998). And it is believed that cigarette smoke triggers COPD through causing epithelial damage and interfering repair processes (Thorley and Tetley, 2007).

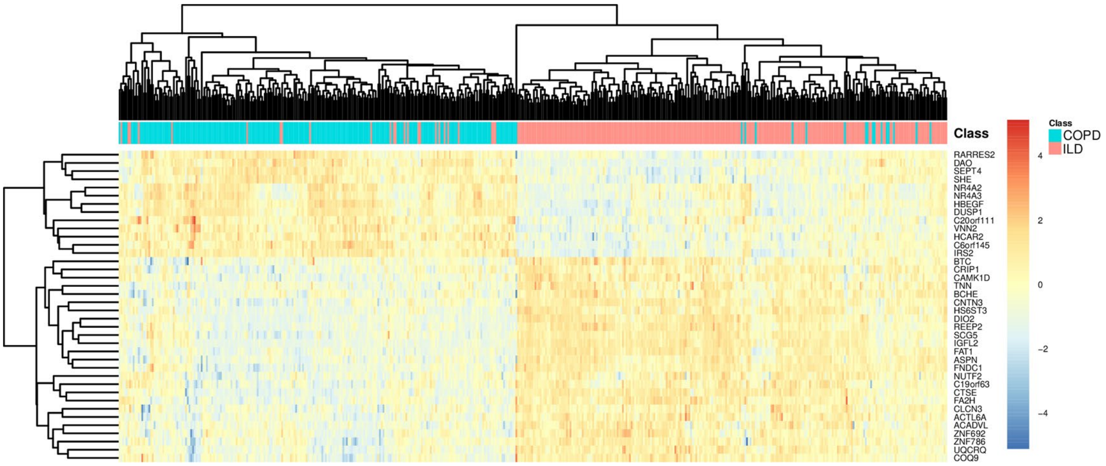


FIGURE 3 | The heatmap of COPD and ILD patients using the selected 38 genes. The COPD and ILD patients from training dataset and test dataset were hierarchically cluttered using the 38 selected genes. There were very clear clusters of COPD and cluster of ILD. Most samples were grouped into the right cluster.

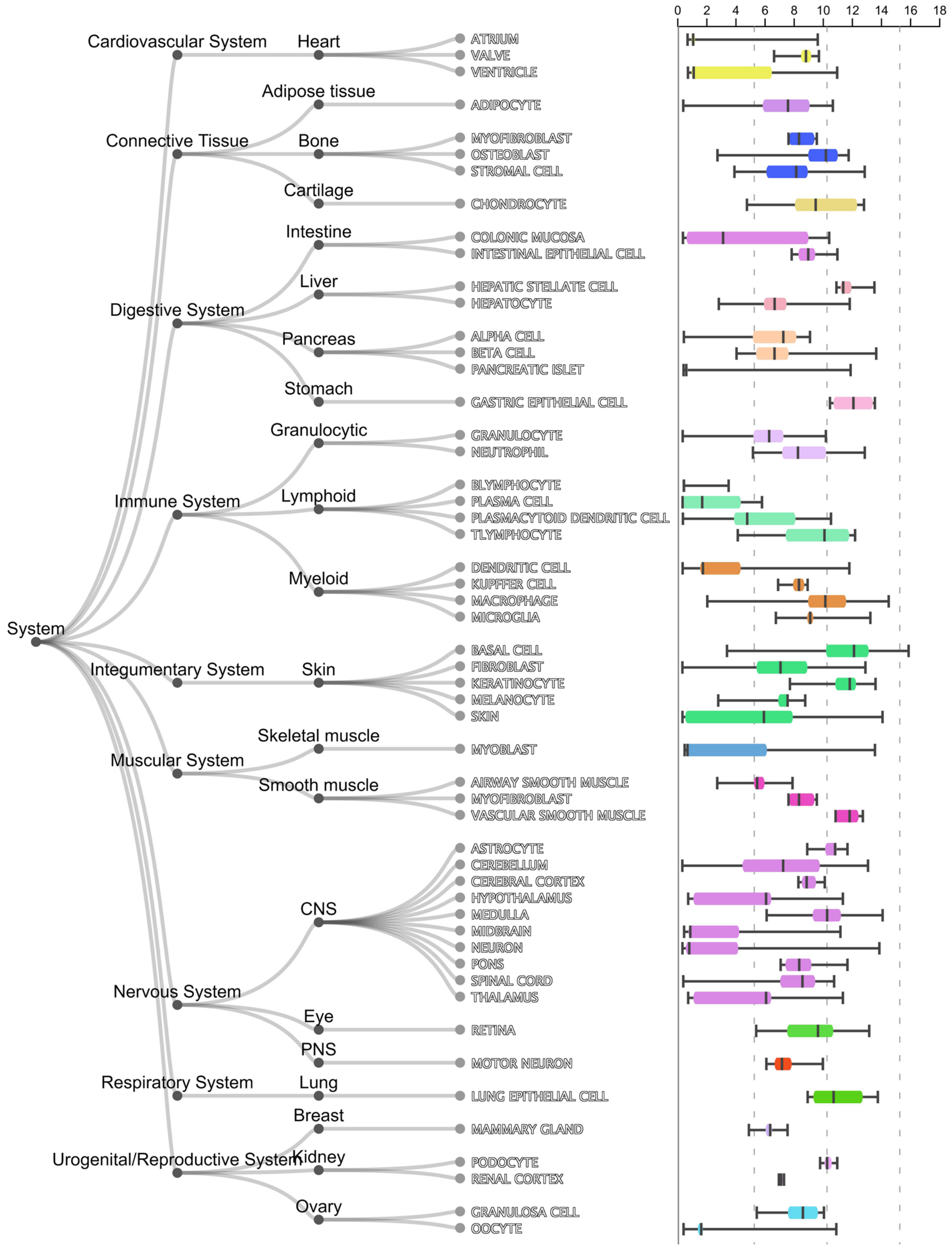


FIGURE 4 | The tissue specific expression pattern of HBEGF in ARCHS4. The tissue expression data from ARCHS4 showed that HBEGF is very specifically highly expressed in lung (https://amp.pharm.mssm.edu/archs4/gene/HBEGF#tissueexpression).

ILD and COPD are two kinds of chronic lung diseases with significant differences in etiology, incidence, pathology, and prognosis (McDonald, 2018). ILD is a heterogeneous group of diseases, characterized by chronic, progressive, mainly interstitial inflammation and is always accompanied by varying degrees of pulmonary parenchyma fibrosis (Doyle et al., 2012), while COPD is characterized by chronic airflow limitation caused by small airway disease and substantial destruction, which is not completely reversible and usually progressive (Song et al., 2012; Rabe and Watz, 2017). Generally, the diagnosis and classification of ILD and COPD severity depend on clinical evaluation, thoracic imaging, and pulmonary function testing (PFT) (Song et al., 2012; Du Plessis et al., 2018).

Among these identified genes, HBEGF has been found related with the invasion and progression of many malignant tumors including breast, pancreatic, and ovarian, and may be involved in macrophage-mediated cellular proliferation (Ray et al., 2014; He et al., 2015). He et al. (2019) conducted comprehensive bioinformatic analyses to predict target genes of ILD and identified HBEGF as one of the potential prognostic markers and therapeutic targets for ILD. Besides, SEPTIN4, a member of the septin family of nucleotide binding proteins, plays a role in apoptosis and cancer (Garcia et al., 2008), which may affect the occurrence and development of ILD.

We will not go through the mRMR table one by one. With only the top three genes, the LOOCV accuracy was 0.873 as shown in **Figure 2**. There are several genes in **Table 1** that are highly possible to play key roles in COPD. Notably, CTSE (cathepsin E) ranked sixth was reported to be able to promote pulmonary emphysema through causing mitochondrial fission and may be a candidate therapeutic target (Zhang et al., 2014b). BTC (betacellulin) ranked 14th was found to be higher expressed in COPD ex-smokers than ex-smokers without COPD (de Boer et al., 2006). DUSP1 (dual specificity phosphatase 1) ranked 16th was reported to have anti-inflammatory potential (Newton, 2014) and when COPD patients undertook fluticasone propionate, DUSP1 expression level was increased (Lee et al., 2016). BCHE (butyrylcholinesterase) ranked 22nd was associated with oxidative stress and inflammation, and its activity was found to be decreased in COPD patients (Sicinska et al., 2017). In **Figure 3**, we also observed the downregulation of BCHE in COPD cluster. SHE (Src homology 2 domain containing E) ranked 25th may play a critical role in promoting airway smooth muscle cell growth and migration during the airway remodeling of COPD patients (Krymskaya et al., 2005). DAO (D-amino acid oxidase) ranked 33rd was an enzyme for peroxisome, glyoxylate metabolism, and glycine degradation. The serum DAO activity was found to be higher in the intestinal tissue of COPD model rat than control (Xin et al., 2016). CAMK1D (calcium/calmodulin dependent protein kinase ID) ranked 37th was found to be a hub node on the protein–protein interaction network of differentially expressed gene (DEG) in COPD and was considered as candidate biomarker and potential target for clinical diagnosis and treatment of COPD (Yuan et al., 2014).

Since there are very few drugs for COPD, we searched DrugBank for possible COPD drugs and found that BCHE, DAO, UQCRQ, HCAR2, CAMK1D, and NR4A3 were drug targetable. The number of small molecule drugs that targeted BCHE, DAO, UQCRQ, HCAR2, CAMK1D, and NR4A3 were 31, 8, 8, 3, 2, and 1, respectively. These genes can be considered as therapeutic targets and may be helpful for development of COPD treatment.

The Associations Between the 38 Genes and Air Pollutants, Particulate Matter, and Tobacco Smoke Pollution

COPD has a close relationship with environmental factors. Pollution and smoking can trigger COPD. Some of the 38 genes have been reported to be associated with smoking by GWAS (genome-wide association study). For example, rs1374879 within CNTN3, which ranked 9th in **Table 1**, was found to be associated with smoking quantity (Argos et al., 2014). Therefore, we systematically studied the associations between signature genes and air pollutants, particulate matter, and tobacco smoke pollution in CTD (comparative toxicogenomics database) (Mattingly et al., 2006). **Table 3** listed how many manually curated literatures, the associations between the gene, and the environmental factor were reported.

It can be seen that 5 genes (HBEGF, DUSP1, NR4A2, NR4A3, and VNN2) were associated with all three environmental factors, 14 genes were associated with two environmental factors, and 4 genes were associated with one environmental factor. Column wise, there were 23 genes associated with particulate matter, 17 genes associated with tobacco smoke pollution, and 7 genes

TABLE 3 | The associations between the 38 genes and air pollutants, particulate matter, and tobacco smoke pollution.

Gene	Air pollutants*	Particulate matter*	Tobacco smoke pollution*
HBEGF	1	15	5
DIO2	0	5	1
CLCN3	0	0	0
SEPT4	0	1	1
FAT1	0	3	2
CTSE	0	4	4
CRIP1	1	1	0
ACADVL	0	4	0
CNTN3	0	0	0
UQCRQ	0	0	0
ASPN	0	0	0
ZNF786	0	0	0
RARRES2	0	3	1
BTC	0	0	0
FNDC1	0	2	1
DUSP1	1	12	3
C6orf145	0	0	0
NUTF2	0	1	1
TNN	0	2	1
COQ9	0	0	0
SCG5	0	1	0
BCHE	0	2	0
NR4A2	1	3	1
HS6ST3	0	0	0
SHE	0	0	0
C20orf111	0	0	0
REEP2	0	0	0
C19orf63	0	0	0
IRS2	0	2	1
FA2H	0	1	0
ACTL6A	1	1	0
NR4A3	1	2	1
DAO	0	1	1
VNN2	1	1	1
IGFL2	0	0	0
ZNF692	0	0	0
CAMK1D	0	3	1
HCAR2	0	2	1

**: The number literatures that suggested the association.*

associated with air pollutants. Particulate matter is a serious threat to health and can cause many lung diseases (Shu et al., 2016).

CONCLUSION

COPD and ILD are two common and similar lung diseases. Both diseases cause huge public health problems. The diagnosis of COPD and ILD is essential for early treatment. We analyzed the gene expression profiles of COPD and ILD patients from two batches that were measured with two microarray platforms. We chose one dataset as the training dataset and selected 38 genes that showed different expression pattern between COPD and ILD patients using advanced mRMR and IFS methods. Based on these 38 genes, a powerful COPD/ILD SVM classifier was built. The classifier had great performance both on training dataset evaluated by LOOCV and on independent test dataset. The 38-gene classifier demonstrated great robustness and excellent prediction accuracy. The biological function analysis of the 38 genes indicated that many of them can be potential treatment targets that may improve the current COPD and ILD therapeutic practice.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YY contributed to the study design. YG conducted the literature search. MY acquired the data. DC wrote the article. FW performed data analysis. YY revised the article and gave the final approval of the version to be submitted. All authors read and approved the final manuscript.

FUNDING

This study was supported by the funds from the Funds of Science and Technology Project of Jiaxing (2017AY33029).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01154/full#supplementary-material

TABLE S1 | The top 500 mRMR gene that showed different expression pattern between COPD and ILD patients. The first column is the rank and the second column is the gene symbol.

FIGURE S1 | The heatmap of COPD and ILD patients using the top 94 genes. The COPD and ILD patients from training dataset and test dataset were hierarchically cluttered using the top 94 genes. There were very clear cluster of COPD and cluster of ILD. Most samples were grouped into the right cluster.

REFERENCES

Andersen, C. U., Mellemkjær, S., Nielsen-Kudsk, J. E., Bendstrup, E., Hilberg, O., and Simonsen, U. (2013). Pulmonary hypertension in chronic obstructive and interstitial lung diseases. *Int. J. Cardiol.* 168 (3), 1795–1804. doi: 10.1016/j.ijcard.2013.06.033

Argos, M., Tong, L., Pierce, B. L., Rakibuz-Zaman, M., Ahmed, A., and Islam, T. (2014). Genome-wide association study of smoking behaviours among Bangladeshi adults. *J. Med. Genet.* 51 (5), 327–333. doi: 10.1136/jmedgenet-2013-102151

Barnes, P. J. (2017). Cellular and molecular mechanisms of asthma and COPD. *Clin. Sci. (Lond.)* 131 (13), 1541–1558. doi: 10.1042/cs20160487

Bosse, Y. (2012). Updates on the COPD gene list. *Int. J. Chron. Obstruct. Pulmon. Dis.* 7, 607–631. doi: 10.2147/COPD.S35294

Caminati, A., Cavazza, A., Sverzellati, N., and Harari, S. (2012). An integrated approach in the diagnosis of smoking-related interstitial lung diseases. *Eur. Respir. Rev.* 21 (125), 207–217. doi: 10.1183/09059180.00003112

Cockayne, D. A., Cheng, D. T., Waschki, B., Sridhar, S., Ravindran, P., and Hilton, H. (2012). Systemic biomarkers of neutrophilic inflammation, tissue injury and repair in COPD patients with differing levels of disease severity. *PloS One* 7 (6), e38629. doi: 10.1371/journal.pone.0038629

Cui, W., Chen, L., Huang, T., Gao, Q., Jiang, M., and Zhang, N. (2013). Computationally identifying virulence factors based on KEGG pathways. *Mol. Biosyst.* 9 (6), 1447–1452. doi: 10.1039/c3mb70024k

de Boer, W. I., Hau, C. M., van Schadewijk, A., Stolk, J., van Krieken, J. H., and Hiemstra, P. S. (2006). Expression of epidermal growth factors and their receptors in the bronchial epithelium of subjects with chronic obstructive pulmonary disease. *Am. J. Clin. Pathol.* 125 (2), 184–192. doi: 10.1309/w1ax-kgt7-ua37-x257

Doyle, T. J., Hunninghake, G. M., and Rosas, I. O. (2012). Subclinical interstitial lung disease: why you should care. *Am. J. Respir. Crit. Care Med.* 185 (11), 1147–1153. doi: 10.1164/rccm.201108-1420PP

du Bois, R. M. (2010). Strategies for treating idiopathic pulmonary fibrosis. *Nat. Rev. Drug Discovery* 9 (2), 129–140. doi: 10.1038/nrd2958

Du Plessis, J. P., Fernandes, S., Jamal, R., Camp, P., Johannson, K., and Schaeffer, M. (2018). Exertional hypoxemia is more severe in fibrotic interstitial lung disease than in COPD. *Respirol.* 23 (4), 392–398. doi: 10.1111/resp.13226

Garcia, W., Rodrigues, N. C., de Oliveira Neto, M., de Araujo, A. P., Polikarpov, I., and Tanaka, M. (2008). The stability and aggregation properties of the GTPase domain from human SEPT4. *Biochim. Biophys. Acta* 1784 (11), 1720–1727. doi: 10.1016/j.bbapap.2008.06.005

He, C., Lv, X., Hua, G., Lele, S. M., Remmenga, S., and Dong, J. (2015). YAP forms autocrine loops with the ERBB pathway to regulate ovarian cancer initiation and progression. *Oncogene* 34 (50), 6040–6054. doi: 10.1038/onc.2015.52

He, Y., Liu, H., Wang, S., and Chen, Y. (2019). In silico detection and characterization of micrornas and their target genes in microrna microarray datasets from patients with systemic sclerosis-interstitial lung disease. *DNA Cell Biol.* 38 (9), 933–944. doi: 10.1089/dna.2019.4780

Hiemstra, P. S., van Wetering, S., and Stolk, J. (1998). Neutrophil serine proteinases and defensins in chronic obstructive pulmonary disease: effects on pulmonary epithelium. *Eur. Respir. J.* 12 (5), 1200–1208. doi: 10.1183/09031936.98.12051200

Huang, T., and Cai, Y.-D. (2013). An information-theoretic machine learning approach to expression QTL analysis. *PloS One* 8 (6), e67899. doi: 10.1371/journal.pone.0067899

Jiang, Y., Huang, T., Chen, L., Gao, Y. F., Cai, Y., and Chou, K. C. (2013). Signal propagation in protein interaction network during colorectal cancer progression. *BioMed. Res. Int.* 2013, 287019. doi: 10.1155/2013/287019

Krymskaya, V. P., Goncharova, E. A., Ammit, A. J., Lim, P. N., Goncharov, D. A., and Eszterhas, A. (2005). Src is necessary and sufficient for human airway smooth muscle cell proliferation and migration. *FASEB J.* 19 (3), 428–430. doi: 10.1096/fj.04-2869fje

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., and Wang, L. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* 9 (1), 1366. doi: 10.1038/s41467-018-03751-6

Lakshmi, S. P., Reddy, A. T., and Reddy, R. C. (2017). Emerging pharmaceutical therapies for COPD. *Int. J. Chron. Obstruct. Pulmon. Dis.* 12, 2141–2156. doi: 10.2147/copd.s121416

Lamb, F. S., Graeff, R. W., Clayton, G. H., Smith, R. L., Schutte, B. C., and McCray, P. B. Jr. (2001). Ontogeny of CLCN3 chloride channel gene expression in human pulmonary epithelium. *Am. J. Respir. Cell Mol. Biol.* 24 (4), 376–381. doi: 10.1165/ajrcmb.24.4.4114

Lee, J., Machin, M., Russell, K. E., Pavlidis, S., Zhu, J., and Barnes, P. J. (2016). Corticosteroid modulation of immunoglobulin expression and B-cell function in COPD. *FASEB J.* 30 (5), 2014–2026. doi: 10.1096/fj.201500135

Li, B. Q., You, J., Huang, T., and Cai, Y. D. (2014). Classification of non-small cell lung cancer based on copy number alterations. *PloS One* 9 (2), e88300. doi: 10.1371/journal.pone.0088300

Ma, S. F., Xie, L., Pino-Yanes, M., Sammani, S., Wade, M. S., and Letsiou, E. (2011). Type 2 deiodinase and host responses of sepsis and acute lung injury. *Am. J. Respir. Cell Mol. Biol.* 45 (6), 1203–1211. doi: 10.1165/rcmb.2011-0179OC

Mattingly, C. J., Rosenstein, M. C., Davis, A. P., Colby, G. T., Forrest, J. N. Jr., and Boyer, J. L. (2006). The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks. *Toxicol. Sci.* 92 (2), 587–595. doi: 10.1093/toxsci/kfl008

McDonald, C. F. (2018). Exercise desaturation and oxygen therapy in ILD and COPD: Similarities, differences and therapeutic relevance. *Respirol.* 23 (4), 350–351. doi: 10.1111/resp.13242

Newton, R. (2014). Anti-inflammatory glucocorticoids: changing concepts. *Eur. J. Pharmacol.* 724, 231–236. doi: 10.1016/j.ejphar.2013.05.035

Niu, B., Huang, G., Zheng, L., Wang, X., Chen, F., and Zhang, Y. (2013). Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties. *BioMed. Res. Int.* 2013, 674215. doi: 10.1155/2013/674215

Nixon, J., Newbold, P., Mustelin, T., Anderson, G. P., and Kolbeck, R. (2017). Monoclonal antibody therapy for the treatment of asthma and chronic obstructive pulmonary disease with eosinophilic inflammation. *Pharmacol. Ther.* 169, 57–77. doi: 10.1016/j.pharmthera.2016.10.016

Papi, A., Casoni, G., Caramori, G., Guzzinati, I., Boschetto, P., and Ravenna, F. (2004). COPD increases the risk of squamous histological subtype in smokers who develop non-small cell lung carcinoma. *Thorax.* 59 (8), 679–681. doi: 10.1136/thx.2003.018291

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8), 1226–1238. doi: 10.1109/TPAMI.2005.159

Peng, X., Moore, M., Mathur, A., Zhou, Y., Sun, H., and Gan, Y. (2016). Plexin C1 deficiency permits synaptotagmin 7–mediated macrophage migration and enhances mammalian lung fibrosis. *FASEB J.* 30 (12), 4056–4070. doi: 10.1096/fj.201600373R

Rabe, K. F., and Watz, H. (2017). Chronic obstructive pulmonary disease. *Lancet* 389 (10082), 1931–1940. doi: 10.1016/s0140-6736(17)31222-9

Ray, K. C., Moss, M. E., Franklin, J. L., Weaver, C. J., Higginbotham, J., and Song, Y. (2014). Heparin-binding epidermal growth factor-like growth factor eliminates constraints on activated Kras to promote rapid onset of pancreatic neoplasia. *Oncogene* 33 (7), 823–831. doi: 10.1038/onc.2013.3

Shu, Y., Zhang, N., Kong, X., Huang, T., and Cai, Y. D. (2014). Predicting A-to-I RNA Editing by Feature Selection and Random Forest. *PloS One* 9 (10), e110607. doi: 10.1371/journal.pone.0110607

Shu, Y., Zhu, L., Yuan, F., Kong, X., Huang, T., and Cai, Y. D. (2016). Analysis of the relationship between PM2.5 and lung cancer based on protein-protein interactions. *Comb. Chem. High Throughput Screen* 19 (2), 100–108. doi: 10.2174/1386207319666151110123345.

Sicinska, P., Bukowska, B., Pajak, A., Koceva-Chyla, A., Pietras, T., and Nizinkowski, P. (2017). Decreased activity of butyrylcholinesterase in blood plasma of patients with chronic obstructive pulmonary disease. *Arch. Med. Sci.* 13 (3), 645–651. doi: 10.5114/aoms.2016.60760

Song, G., Mortani Barbosa, E. Jr., Tustison, N., Gefter, W. B., Kreider, M., and Gee, J. C. (2012). A comparative study of HRCT image metrics and PFT values for characterization of ILD and COPD. *Acad. Radiol.* 19 (7), 857–864. doi: 10.1016/j.acra.2012.03.007

Sun, L., Yu, Y., Huang, T., An, P., Yu, D., and Yu, Z. (2012). Associations between ionomic profile and metabolic abnormalities in human population. *PloS One* 7 (6), e38845. doi: 10.1371/journal.pone.0038845

Thorley, A. J., and Tetley, T. D. (2007). Pulmonary epithelium, cigarette smoke, and chronic obstructive pulmonary disease. *Int. J. Chron. Obstruct. Pulmon. Dis.* 2 (4), 409–428.

Vogelmeier, C. F., Criner, G. J., Martinez, F. J., Anzueto, A., Barnes, P. J., and Bourbeau, J. (2017). Global strategy for the diagnosis, management and prevention of chronic obstructive lung disease 2017 report: GOLD Executive Summary. *Respirol.* 22 (3), 575–601. doi: 10.1111/resp.13012

Wang, W., Xie, M., Dou, S., Cui, L., Zheng, C., and Xiao, W. (2018). The link between chronic obstructive pulmonary disease phenotypes and histological subtypes of lung cancer: a case-control study. *Int. J. Chron. Obstruct. Pulmon. Dis.* 13, 1167–1175. doi: 10.2147/copd.s158818

Xin, X., Dai, W., Wu, J., Fang, L., Zhao, M., and Zhang, P. (2016). Mechanism of intestinal mucosal barrier dysfunction in a rat model of chronic obstructive pulmonary disease: An observational study. *Exp. Ther. Med.* 12 (3), 1331–1336. doi: 10.3892/etm.2016.3493

Yang, J., Chen, L., Kong, X., Huang, T., and Cai, Y. D. (2014). Analysis of Tumor Suppressor Genes Based on Gene Ontology and the KEGG Pathway. *PloS One* 9 (9), e107202. doi: 10.1371/journal.pone.0107202

Young, R. P., Hopkins, R. J., Christmas, T., Black, P. N., Metcalf, P., and Gamble, G. D. (2009). COPD prevalence is increased in lung cancer, independent of age, sex and smoking history. *Eur. Respir. J.* 34 (2), 380–386. doi: 10.1183/09031936.00144208

Yuan, Y. P., Shi, Y. H., and Gu, W. C. (2014). Analysis of protein-protein interaction network in chronic obstructive pulmonary disease. *Genet. Mol. Res.* 13 (4), 8862–8869. doi: 10.4238/2014.October.31.1

Zhang, N., Huang, T., and Cai, Y. D. (2014a). Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol. Genet. Genomics*. 290 (1), 343–352. doi: 10.1007/s00438-014-0922-5

Zhang, X., Shan, P., Homer, R., Zhang, Y., Petrache, I., and Mannam, P. (2014b). Cathepsin E promotes pulmonary emphysema *via* mitochondrial fission. *Am. J. Pathol.* 184 (10), 2730–2741. doi: 10.1016/j.ajpath.2014.06.017

Zhang, P. W., Chen, L., Huang, T., Zhang, N., Kong, X. Y., and Cai, Y. D. (2015). Classifying ten types of major cancers based on reverse phase protein array profiles. *PloS One* 10 (3), e0123147. doi: 10.1371/journal.pone.0123147

Zhao, T. H., Jiang, M., Huang, T., Li, B. Q., Zhang, N., and Li, H. P. (2013). A novel method of predicting protein disordered regions based on sequence features. *BioMed. Res. Int.* 2013, 414327. doi: 10.1155/2013/414327

Zhou, Y., Zhang, N., Li, B. Q., Huang, T., Cai, Y. D., and Kong, X. Y. (2015). A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. *J. Biomol. Struct. Dyn.* 33 (11), 2479–2490. doi: 10.1080/07391102.2014.1001793

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.

METHODS
published: 11 December 2019
doi: 10.3389/fgene.2019.01234

Identifying Potential miRNAs–Disease Associations With Probability Matrix Factorization

Junlin Xu[1], Lijun Cai[1], Bo Liao[3]*, Wen Zhu[1], Peng Wang[1], Yajie Meng[1], Jidong Lang[2], Geng Tian[2] and Jialiang Yang[3]**

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China, [2] Department of Science, Geneis Beijing Co., Ltd., Beijing, China, [3] School of Mathematics and Statistics, Hainan Normal University, Haikou, China

In recent years, miRNAs have been verified to play an irreplaceable role in biological processes associated with human disease. Discovering potential disease-related miRNAs helps explain the underlying pathogenesis of the disease at the molecular level. Given the high cost and labor intensity of biological experiments, computational predictions will be an indispensable alternative. Therefore, we design a new model called probability matrix factorization (PMFMDA). Specifically, we first integrate miRNA and disease similarity. Next, the known association matrix and integrated similarity matrix are utilized to construct a probability matrix factorization algorithm to identify potentially relevant miRNAs for disease. We find that PMFMDA achieves reliable performance in the frameworks of global leave-one-out cross validation (LOOCV) and 5-fold cross validation (AUCs are 0.9237 and 0.9187, respectively) in the HMDD (V2.0) dataset, significantly outperforming a few state-of-the-art methods including CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA. In addition, case studies show that PMFMDA has good predictive performance for new associations, and the evidence can be identified by literature mining.

Keywords: diseases, miRNAs, probabilistic matrix factorization, association prediction, receiver operating characteristic curve (ROC)

OPEN ACCESS

Edited by:
Peilin Jia,
University of Texas Health Science Center, United States

Reviewed by:
Cheng Liang,
Shandong Normal University, China
Yin-Ying Wang,
University of Texas Health Science Center at Houston, United States

****Correspondence:***
Lijun Cai
ljcai@hnu.edu.cn
Bo Liao
dragonbw@163.com
Jialiang Yang
yangjl@geneis.cn

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

Received: *29 May 2019*
Accepted: *06 November 2019*
Published: *11 December 2019*

Citation:
Xu J, Cai L, Liao B, Zhu W, Wang P, Meng Y, Lang J, Tian G and Yang J (2019) Identifying Potential miRNAs–Disease Associations With Probability Matrix Factorization. Front. Genet. 10:1234. doi: 10.3389/fgene.2019.01234

INTRODUCTION

MicroRNAs are short non-coding RNAs. It plays a vital role in the regulation of many important biological processes (Bandyopadhyay et al., 2010; Hammond, 2015; Zhang et al., 2017). It has shown that human disease is associated with abnormal expression of miRNAs, whose analyses can guide the diagnosis, prognosis and treatment of certain diseases (Liang et al., 2019). However, identifying new miRNA–disease associations through bio-wet experiments not only has a high error rate, but also consumes huge financial resources (Feng et al., 2017). Therefore, *in-silicon* prediction of disease-associated miRNAs has become a critical step in prioritizing most confident targets for further experimental validation. Due to the growing power of sequencing technology, more and more omics data have been published (Yi et al., 2017), which provides a chance to reveal what role miRNAs play in physiology and pathology. Typical directions include miRNAs–disease interaction prediction, miRNA–miRNA regulatory module discovery, and so on (Chou et al., 2016). Undoubtedly, all these studies enrich our understanding of the functional regulation mechanisms of miRNA (Ha et al., 2019).

In recent years, in order to understand the pathogenesis of diseases, more and more computational models have been proposed by researchers to infer disease-related miRNAs, among which machine learning-based and network-based methods are most popular (Luo et al., 2017a). Network-based methods are based on a common assumption that miRNAs associated with diseases using similar phenotypes are similar in function, and vice versa. For example, Jiang et al. (2010) proposed the priority of disease-associated miRNAs through human peptide–microRNAome networks to identify potential associations. However, this method relies too much on known associations to make its prediction performance less effective. Subsequently, Chen et al. (2012) implemented a random walk with restart (RWRMDA) on its network to identify potentially associated miRNAs by building a network of similarities between miRNAs. Similarly, Shi et al. (2013) conducted random walks through functional linkages between miRNA targets and disease genes to explore the relationship between human miRNA diseases. Peng et al. (2017) constructed a multiple biological network by integrating the two-way relationship among microRNA, disease and environmental factors, and realized the unbalanced random walk algorithm on this network to achieve the purpose of prediction. However, these methods cannot predict miRNAs associated with isolated diseases. Later, Chen and Zhang (2013) used a network of consistent reasoning methods to infer unknown miRNAs associated with disease. Gu et al. (2016) created a network consistent projection algorithm to identify latent associations by integrating similarity networks and associated networks. The biggest advantage of these methods is that they can predict isolated disease-associated miRNAs, but the performance achieved is not very satisfactory.

More recently, machine learning-based models have been implemented to improve classification accuracy and prediction performance (Gu et al., 2016). For example, Xu et al. (2011) designed a support vector machine (SVM) classifier that combines four topological features extracted from a miRNA target disease network to distinguish between prostate cancer-associated miRNAs and non-prostate cancer-associated miRNAs. To construct a negative sample, they randomly paired the miRNA with the disease and then removed the pair present in the positive sample set. It is clear that negative samples constructed in this way are prone to false positives. Chen and Yan (2014) introduced a normalized least square method to identify the association between potential miRNAs–diseases (RLSMDA), which does not require negative samples. In addition, Luo et al. (2017b) developed a Kronecker regularized least squares method to predict the potential association of miRNAs–disease by combining multiple omics data. Liu et al. (2019) converted the miRNAs–disease association prediction problem into a complete bipartite graph model, and proposed a prediction algorithm based on a restricted Boltzmann machine to improve prediction performance. Shen et al. (2017) introduced the cooperative matrix decomposition (CMFMDA) algorithm in the recommendation system to infer potential associations. Finally, Chen et al. (2018) introduced an induction matrix-completed algorithm to identify unknown associations. However, these methods do not perform well in predicting associations related to new diseases or miRNAs, and the prediction accuracy is not as satisfactory as associations with known diseases or miRNAs.

In order to achieve better predictive performance, we construct a new model called probability matrix factorization (PMFMDA) to predict unknown miRNAs–disease associations in this study. PMFMDA makes full use of miRNA disease association, miRNA similarity and disease similarity. To evaluate the effectiveness of PMFMDA, we test it using frameworks of global 5-fold CV and global LOOCV. In addition, a validation method called CV_d is developed to estimate the performance in predicting novel diseases or miRNAs. Outperforming other state-of-the-arts methods, PMFMDA achieve reliable performance in the frameworks of global LOOCV and 5-fold CV (AUCs of 0.9237 and 0.9187, respectively) in the HMDD (V2.0) dataset (Li et al., 2014). To further demonstrate the superiority of PMFMDA, we conduct an analysis of three common diseases. According to the analysis of the test results, we can find that there are 20, 19 and 17 of 20 candidate miRNAs that are confirmed to be associated with esophageal neoplasms, breast neoplasms and lung neoplasms by dbDEMC and miRCancer, respectively.

MATERIALS AND METHODS

The general workflow of PMFMDA is shown in **Figure 1**. We first use matrix Y to represent 5,430 experimentally validated associations after preprocessing the HMDD V2.0 database (Li et al., 2014). Specifically, Y is a 495 × 383 matrix with row denoting miRNAs and column denoting diseases; $Y_{i,j} = 1$ if the i^{th} miRNA is associated with the j^{th} disease and 0 otherwise. We then calculate the disease similarity S_d and miRNA similarity S_m. Finally, a probability matrix factorization (PMF) model is proposed by integrating Y, S_d and S_m, the solution of which will recover unknown miRNAs–disease associations based on known ones.

Disease Semantic Similarity

The hierarchical directed acyclic graphs (DAGs), usually are obtained from the MeSH database, and are widely used to calculate the similarity between diseases (Gu et al., 2016). Specifically, for a disease d, let $DAGd = (d, T_d, E_d)$ represents its directed acyclic graph, where T_d denotes the set of the ancestors of d, and E_d represents the set of links in the MeSH tree. So, the semantic contribution of disease t to disease d is defined as:

$$D_d(t) = \begin{cases} 1 & if\ t = d \\ max\left\{\Delta \times D_d(t') \mid t' \in children\ of\ t\right\} & if\ t \neq d \end{cases} \tag{1}$$

Where Δ is a predefined sematic contribution factor, the value of Δ in this study is set to 0.5. Therefore, we can calculate the semantic similarity of between diseases by formula (2).

$$D(d_i, d_j) = \frac{\sum_{t \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(t) + D_{d_j}(t)\right)}{\sum_{t \in T_{d_i}} D_{d_i}(t) + \sum_{t \in T_{d_j}} D_{d_j}(t)} \tag{2}$$

miRNAs Functional Similarity

For the similarity between miRNAs, most studies use functional similarity measurements (Wang et al., 2010). Specifically, for any two miRNAs r_i and r_j, let $DT_i = \{d_{i1}, d_{i2}, \ldots, d_{ik}\}$ and $DT_j =$

$\{d_{j1},d_{j2},\ldots,d_{jl}\}$ be their associated disease sets, respectively. Similar to Wang et al. we first use $S(d, DT)=\max_{d_i \in DT} D(d,d_i)$ to represent the similarity between a disease d and DT. Then the similarity between r_i and r_j is defined as

$$R(r_i,r_j)=\frac{\sum_{m=1}^{k} S(d_{im},DT_j)+\sum_{n=1}^{l} S(d_{jn},DT_i)}{k+l} \tag{3}$$

The Gaussian Interaction Profile Kernel Similarity For Diseases and miRNAs

In the similarity measurement algorithm, Gaussian interaction profile kernel similarity is also a good measurement algorithm, which is widely used in various fields (Lu et al., 2019). Let $VP(d_i)$ be the vector associated with the disease d_i in Y, i.e. the i^{th} column of Y. Then, the Gaussian interaction kernel similarity between disease d_i and d_j is calculated as:

$$KD(d_i\ ,d_j)=exp(-\gamma_d \,||\, VP(d_i)-VP(d_j)||^2) \tag{4}$$

where γ_d is the adjustment parameter of the kernel bandwidth. The parameter γ_d update rule is as follows:

$$\gamma_d=\gamma_d^{'}/(\frac{1}{nd}\Sigma_{i=1}^{nd}\,||(d_i)||^2) \tag{5}$$

where $\gamma_d^{'}$ is usually set to 1.

Similarly, we can conclude that the Gaussian kernel similarity of miRNAs is as follows:

$$KM\left(r_i\ ,r_j\right)=exp(-\gamma_m \,||\, VP(r_i)-VP(r_j)||^2) \tag{6}$$

$$\gamma_m=\gamma_m^{'}/(\frac{1}{nm}\Sigma_{i=1}^{nm}\,||\,VP(r_i)||^2) \tag{7}$$

Where $\gamma_m^{'}$ is usually set to 1.

Integrated Similarity For Diseases and miRNAs

The similarity between disease d_i and disease d_j is constructed by combining the two similarities of the disease as follows:

$$S_d(d_i\ ,d_j)=\begin{cases} D(d_i\ ,d_j) & d_i \text{ and } d_j \text{ has semantic similarity} \\ KD(d_i\ ,d_j) & \text{otherwise} \end{cases} \tag{8}$$

Similarly, the similarity between miRNAs r_i and r_j can be redefined as:

$$S_m(r_i\ ,r_j)=\begin{cases} R(r_i\ ,r_j) & r_i \text{ and } r_j \text{ has functional similarity} \\ KM(r_i\ ,r_j) & \text{otherwise} \end{cases} \tag{9}$$

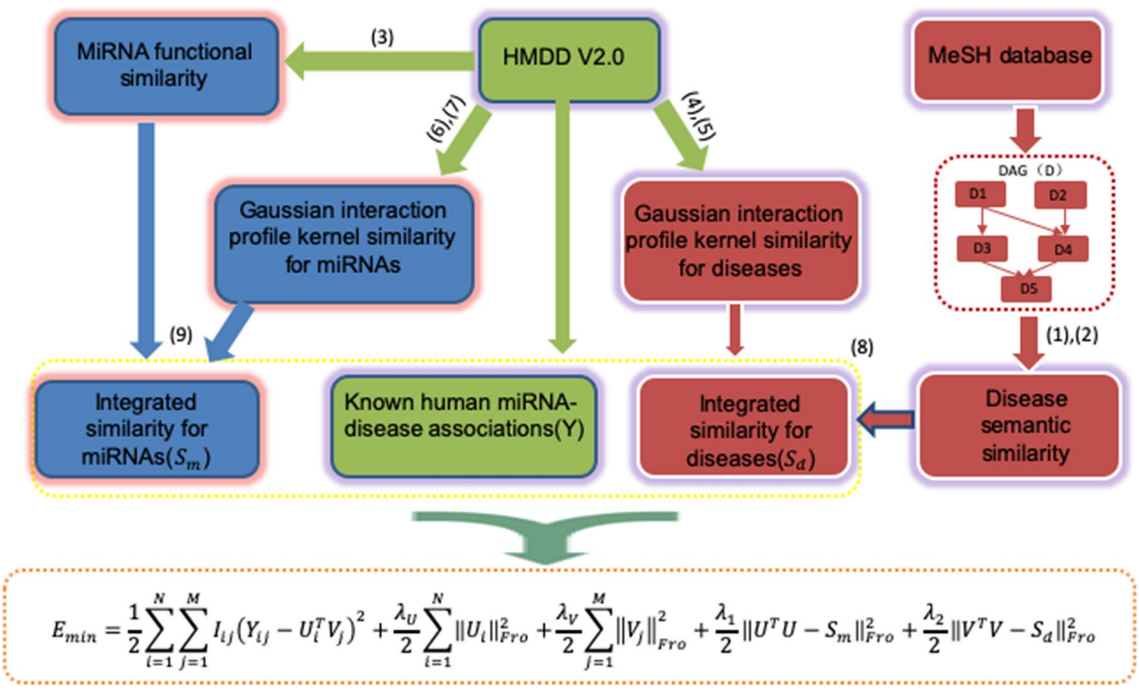

MiRNA functional similarity
(3)
HMDD V2.0
MeSH database
(6),(7)
(4),(5)
Gaussian interaction profile kernel similarity for miRNAs
Gaussian interaction profile kernel similarity for diseases
DAG（D）
D1
D2
D3
D4
D5
(9)
(8)
(1),(2)
Integrated similarity for miRNAs(S_m)
Known human miRNA-disease associations(Y)
Integrated similarity for diseases(S_d)
Disease semantic similarity
$E_{min}=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(Y_{ij}-U_i^T V_j)^2+\frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_{Fro}^2+\frac{\lambda_V}{2}\sum_{j=1}^{M}\|V_j\|_{Fro}^2+\frac{\lambda_1}{2}\|U^T U-S_m\|_{Fro}^2+\frac{\lambda_2}{2}\|V^T V-S_d\|_{Fro}^2$

FIGURE 1 | The workflow of PMFMDA is used to infer disease-associated unknown miRNAs.

PMFMDA

Probability Matrix Decomposition (PMF) is a probabilistic linear model of Gaussian observation noise and has been widely used in data representation (Salakhutdinov and Mnih, 2008). Let $Y \in R^{n \times m}$ be the known miRNAs–disease association matrix, U_i and V_i represent the D-dimensional miRNA-specific and disease-specific latent feature vectors, respectively. The conditional distribution of the observed associations $Y \in R^{n \times m}$ (likelihood term) and the prior distribution of $U \in R^{D \times n}$ and $V \in R^{D \times m}$ are given by:

$$P\left(Y \mid U, V, \alpha\right) = \prod_{i=1}^{N} \prod_{j=1}^{M} [N(Y_{ij} \mid U_i^T V_j, \alpha^{-1})]^{I_{ij}} \quad (10)$$

$$P(U|\alpha_U) = \prod_{i-1}^{N} N(U_i \mid 0, \alpha_U^{-1} I) \quad (11)$$

$$P(V \mid \alpha_V) = \prod_{j=1}^{M} N(V_j \mid 0, \alpha_V^{-1} I) \quad (12)$$

Where $N(x \mid \mu, \alpha^{-1})$ denotes the Gaussian distribution, $I_{ij} = 0$ if the entry(i,j) in Y is missing, and 1 otherwise.

The optimal model is obtained by maximizing the logarithmic a posterior of miRNAs and disease characteristics using fixed hyperparameters:

$$\ln P\left(U, V \mid Y, \alpha, \alpha_V, \alpha_U\right) = \ln P\left(Y \mid U, V, \alpha\right) + \ln P\left(U \mid \alpha_U\right) + \ln P\left(V \mid \alpha_V\right) + C \quad (13)$$

Where C is a constant. So, using a quadratic regularization term to minimize the sum of squares of the error functions instead of maximizing the posterior distribution relative to U and V:

$$E_{\min} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} (Y_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \| U_i \|_{Fro}^2 + \frac{\lambda_V}{2} \sum_{j=1}^{M} \| V_j \|_{Fro}^2 \quad (14)$$

Where $\lambda_U = \alpha_U / \alpha$ and $\lambda_v = \alpha_V / \alpha$ are regularization parameters, $\| \cdot \|_{Fro}^2$ denotes the Frobenius norm.

The standard PMF in Equation (10) does not consider the effect of similarity between miRNAs and the similarity between diseases. Since U_i represents the D-dimensional miRNA-specific latent feature vectors, $U^T U$ denotes the weighted similarity matrix of the miRNAs. Similarly, $V^T V$ denotes the weighted similarity matrix of the disease. Thus, we propose a new objective function by integrating miRNAs similarity and diseases similarity named PMFMDA as follows:

$$E_{\min} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} I_{ij} (Y_{ij} - U_i^T V_j)^2 + \frac{\lambda_U}{2} \sum_{i=1}^{N} \| U_i \|_{Fro}^2 + \frac{\lambda_V}{2} \sum_{j=1}^{M} \| V_j \|_{Fro}^2 + \frac{\lambda_1}{2} \| U^T U - S_m \|_{Fro}^2 + \frac{\lambda_2}{2} \| V^T V - S_d \|_{Fro}^2 \quad (15)$$

where S_m and S_d have been calculated before.

Optimization

In order to obtain the local optimal solution of Equation (15), we use the gradient descent algorithm to solve (Xiao et al., 2018). According to the nature of the Frobenius norm, the corresponding Lagrange function L_E of Equation (15) is defined as:

$$\begin{aligned} L_E = {} & \frac{1}{2} Tr\left(I \cdot (YY^T - YV^T U - U^T V Y^T + U^T V V^T U)\right) + \\ & \frac{\lambda_U}{2} Tr(UU^T) + \frac{\lambda_V}{2} Tr(VV^T) + \frac{\lambda_1}{2} Tr(S_m (S_m)^T) - \\ & S_m U^T U - U^T U (S_m) + U^T U U^T U) + \frac{\lambda_2}{2} Tr(S_d (S_d)^T - \\ & S_d V^T V - V^T V S_d + V V^T V) + Tr(\varnothing U^T) + Tr(\psi V^T) \end{aligned} \quad (16)$$

where $T_r()$ denotes the trace of a matrix, $\varnothing = [\varphi_{ik}]$ and $\Psi = [\omega_{jk}]$ are Lagrangian multipliers.

The partial derivatives of U and V are as follows:

$$\begin{aligned} \frac{\partial L_E}{\partial U} &= I \cdot \left(-VY^T + VV^T U\right) + \lambda_U U + 2\lambda_1 \left(-U\left(S_m\right) + UU^T U\right) + \varnothing, \\ \frac{\partial L_E}{\partial V} &= I \cdot \left(-UY + UU^T V\right) + \lambda_V V + 2\lambda_2 \left(-V\left(S_d\right) + VV^T V\right) + \Psi \end{aligned} \quad (17)$$

Finally, the Karush-Kuhn-Tucker (KKT) conditions $\varphi_{ikU_{ik}=0}$ and $\omega_{jkV_{Jk}=0}$ according to the gradient descent method. The following equations are obtained for U_{ik} and V_{jk}:

$$\begin{aligned} & \left(I \cdot \left(-VY^T + VV^T U\right)\right)_{ik} U_{ik} + (\lambda_U U)_{ik} U_{ik} \\ & + \left(2\lambda_1 \left(-U\left(S_m\right) + UU^T U\right)\right)_{ik} U_{ik} = 0, \\ & \left(I \cdot (-UY + UU^T V)\right)_{jk} V_{jk} + (\lambda_V V)_{jk} V_{jk} \\ & + \left(2\lambda_2 (-V(S_d) + VV^T V)\right)_{jk} V_{jk} = 0 \end{aligned} \quad (18)$$

Therefore, the updating rules for U and V as follows:

$$U_{ik}^{new} = U_{ik} \frac{\left(I \cdot (VY^T) + 2\lambda_1 (U(S_m))\right)_{ik}}{\left(I \cdot (VV^T U)\right)_{ik} + (\lambda_U U)_{ik} + \left(2\lambda_1 (UU^T U)\right)_{ik}} \quad (19)$$

$$V_{jk}^{new} = V_{jk} \frac{\left(I \cdot (UY) + 2\lambda_1 (U(S_m))\right)_{jk}}{\left(I \cdot (UU^T V)\right)_{jk} + (\lambda_V V)_{jk} + \left(2\lambda_2 (VV^T V)\right)_{jk}} \quad (20)$$

Update U and V according to Equation (19) and Equation (20) until the local minimum of the objective function. Finally, the predicted miRNAs–disease association matrix is $Y' = U^T V$. The ith column of Y' indicates the association score between disease d_i and miRNAs, and the larger the score, the more relevant it is.

Evaluation Methods

In order to test the performance of PMFMDA, we utilize a 5-fold CV experiment and global LOOCV on the HMDD database and compare it with a few recent methods including CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA. In the 5-fold CV experiment of a single disease *d*, known miRNAs associated with *d* (column vectors in matrix $A \in R^{m \times n}$) are randomly divided into five subsets of equal size. Associations related to all other diseases together with 4 subsets (with respect to *d*) are taken as training samples and the remaining subset is considered as testing samples. The process is performed for 5 times until all the associations associated with d have been predicted once. Global LOOCV was used to evaluate the model's global prediction ability for all miRNAs–disease association simultaneously. Specifically, we removed each known association in turn as a testing sample, with all remaining associations as training samples. We then predicted the removed entry and evaluated the performance. In addition, we perform CV_d experiment to test the performance of PMFMDA in predicting miRNAs associated to a novel disease *d*. In CV_d: CV on disease d_i, we remove all the known associations of the disease d_i (column vectors in matrix $Y \in R^{m \times n}$) and build prediction model (for inferring the deleted associations) using the remaining data.

Parameter Tuning

We cross-validate the training set to tune the parameters of PMFMDA. Specifically, the parameters λ_U,λ_V,λ_1, and λ_2 are increased from 0.001 to 1 with a step of 0.1 and the ones with the best AUC are selected. Since the other methods have also been tested on HMDD (V2.0) in published papers, we adopt the parameters provided by the authors. Specifically, *W*=0.9 for RLSMDA, $\lambda_U = \lambda_V = 1$,$\lambda_1 = \lambda_2 = 0.005$ for PMFMDA, $\lambda_1 = \lambda_2 = 1$ for IMCMDA, $\lambda_m = \lambda_d = 1$ for CMFMDA $r = 0.9$, for RWRMDA and NCPMDA is parameter free.

RESULTS

PMFMDA Outperforms Other Popular Methods In Predicting Potential Associations

We apply PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA into the HMDD database. Their receiver operating characteristic (ROC) curves and associated area under the curve (AUCs) of the global 5-fold CV and LOOCV are plotted in **Figure 2**. As can be seen, the AUCs of PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA are 0.9187, 0.8928, 0.8372, 0.8792, 0.8333, and 0.8168, respectively. Furthermore, PMFMDA also achieve the best AUC (0.9237) on global LOOCV, indicating that PMFMDA perform best in predicting miRNAs–disease associations. However, considering the limited number of known miRNAs–disease associations, it might be insufficient to evaluate the performance of the methods by AUC alone. Thus, we also plotted the precise recall (PR) curve and calculated the area under the PR curve (AUPR) based on the global 5-fold CV experiment in **Figure 3**. In a PR-curve, the precision refers to the ratio of correctly predicted associations to all associations with scores higher than a given threshold; by contrast, the recall refers to the ratio of correctly predicted associations to all known miRNAs–disease associations. In general, the ROC curve and the PR curve show similar trend. As shown in **Figure 3**, the AUPRs of PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA are 0.3535, 0.3428, 0.2509, 0.1176, 0.1234, and 0.1369 respectively, indicating that PMFMDA performed best in predicting miRNAs–disease associations. At the same time, in order to further prove the effectiveness of PMFMDA. We performed 10 times of global 5-fold CV and achieved an average AUC and AUPR of 0.9187

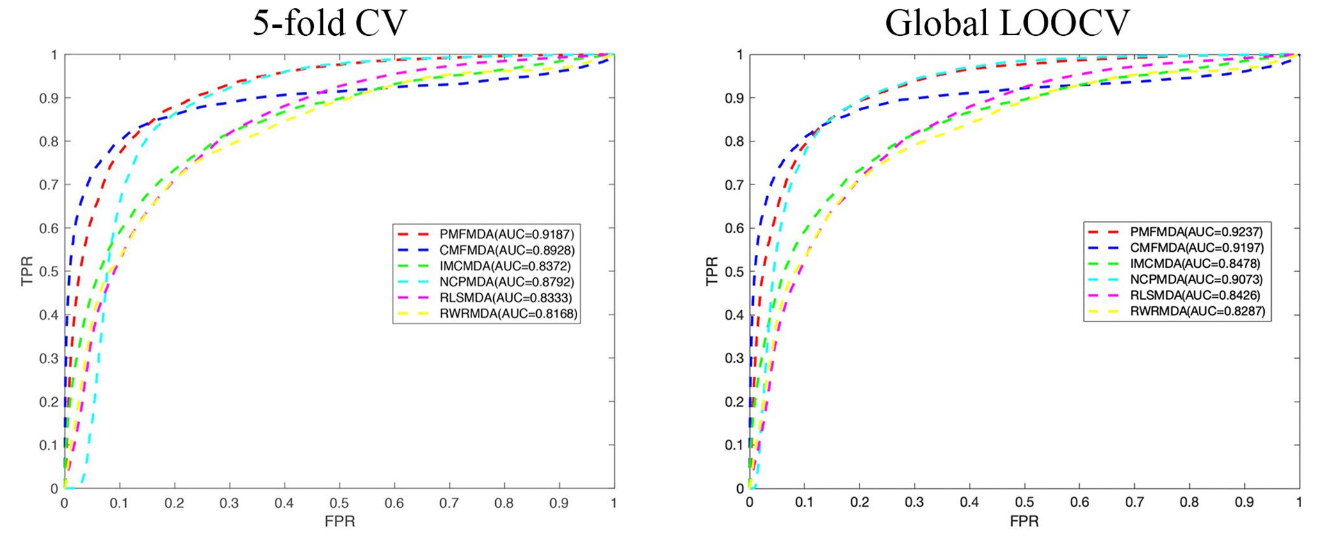


FIGURE 2 | The ROC curves for PMFMDA and benchmark algorithms for 5-fold CV and global LOOCV.

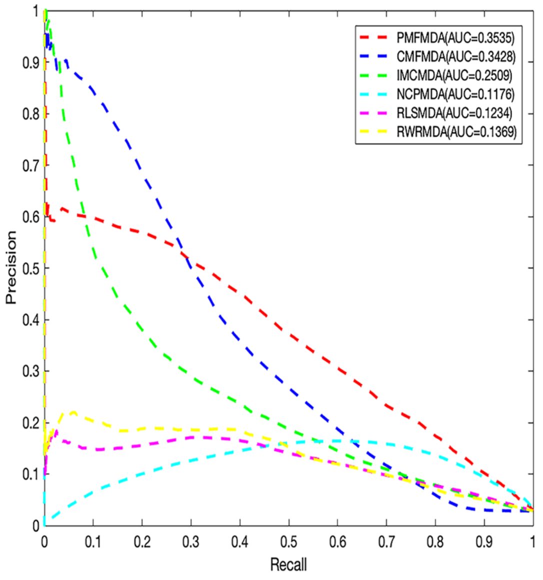


FIGURE 3 | The PR curves for PMFMDA and benchmark algorithms for 5-fold CV.

+/− 0.0013, 0.3535+/− 0.0015, respectively. This proves the reliability and stability of the PMFMDA algorithm.

PMFMDA Outperforms Other Popular Methods In Predicting miRNAs Associated With Novel Diseases

Besides global miRNAs–disease predictions, it is also critical to check the performance of the above methods on specific diseases. CV_d is used to measure the ability of an algorithm to predict a new disease-associated miRNA. In order to compare the fairness of the test, we conduct CV tests on 8 common diseases (Xuan et al., 2015) and use the area under the accurate recall curve (AUPR) as an indicator of predictive performance. The reason is that AUPR severely penalizes highly ranked non-interactions, which is desirable here because in practice we do not want to recommend incorrect predictions (i.e., AUPR metrics severely penalize highly ranked false positives). The results for CV_d are shown in **Table 1**. We can clearly see that the average AUPR of PMFMDA for the eight test diseases was 0.6687, which was significantly higher than IMCMDA (0.6377), CMFMDA (0.5091), NCPMDA (0.6121), and RLSMDA (0.5761). This also sufficient PMFMDA is also the best way to predict miRNAs associated with novel diseases.

Furthermore, in order to further evaluate our approach in predicting new diseases. We implement CV_d experiments on the above 8 diseases. We show the calculation of the number of disease-associated miRNAs identified at different ranking thresholds in **Table 2**. For example: We delete all miRNAs associated with breast tumors, and then use PMFMDA to predict its related miRNAs. we can find that 91 of the top 100 predictions are accurately predicted through the test results. This is ample indication that our approach can yield high quality predictions for isolated disease-associated miRNAs. In order to better understand the predicted eight disease-related miRNAs, we listed the names and predicted scores of the top 100 candidates related to the eight diseases in the **Supplementary Table S1**.

Evaluate Performance on Different Data Sources

To further test the versatility of PMFMDA. We obtain 60,576 experimental validation correlation data by preprocessing the MNDR (V2.0) dataset (Cui et al., 2018). The data contains 887 diseases and 3,954 miRNAs. We apply PMFMDA, CMFMDA, IMCMDA, NCPMDA, RLSMDA, and RWRMDA on the MNDR (V2.0) database. As shown in **Table 3**, the AUC of PMFMDA was 0.9885, significantly higher than those of CMFMDA (0.9799), IMCMDA (0.9171), NCPMDA (0.9480), RLSMDA (0.9358), and RWRMDA (0.9055) with increases of about 0.86, 7.14, 4.05, 5.27, and 8.3% respectively. The AUPR of PMFMDA was 0.5174, significantly higher than those of CMFMDA (0.5047), IMCMDA (0.3865), NCPMDA (0.2045), RLSMDA (0.2818), and RWRMDA (0.1907). In conclusion, PMFDA has been proven to be effective in inferring related miRNAs with diseases in terms of AUC values and AUPR values.

Parameter Sensitivity Analysis

In machine learning, parameter tuning is critical for the performance of a model. Thus, we presented in **Table 4** several sets of parameter settings based on the global 5-fold CV experiment on the HMDDV 2.0 dataset. We found that a better

TABLE 1 | Comparison of AUPR values predicted by PMFMDA and benchmark algorithms on novel diseases.

Disease name	AURP				
	PMFMDA	IMCMDA	CMFMDA	NCPMDA	RLSMDA
Melanoma	0.7149	0.6757	0.4574	0.6785	0.6940
Breast tumor	0.7895	0.7752	0.6135	0.7866	0.7749
Colorectal tumor	0.6585	0.6333	0.4725	0.5714	0.5315
Glioblastoma	0.5940	0.5076	0.4540	0.4779	0.4028
Heart failure	0.5956	0.6284	0.4510	0.6182	0.5510
Prostatic tumor	0.6578	0.5881	0.5963	0.5873	0.5208
Stomach tumor	0.6981	0.6438	0.5231	0.6269	0.6081
Bladder tumor	0.6409	0.5388	0.5051	0.5505	0.5255
Mean	0.6687	0.6237	0.5091	0.6121	0.5761

TABLE 2 | PMFMDA predicts the correct numbers of different ranking thresholds for 8 common diseases.

Cancer	No. of known associated miRNAs	Ranking threshold				
		20	40	60	80	100
Breast neoplasms	202	20	38	54	74	91
Colorectal neoplasms	147	17	30	45	58	70
Glioblastoma	96	17	30	36	43	53
Heart failure	120	17	28	39	51	58
Melanoma	141	19	35	51	63	77
Prostatic neoplasms	118	17	32	43	56	65
Stomach neoplasms	173	15	32	49	63	79
Urinary bladder neoplasms	92	18	31	42	51	55

TABLE 3 | The performance of PMFMDA and the baseline methods based on 5-fold CV on the MNDRV2.0 dataset.

	PMFMDA	CMFMDA	IMCMDA	NCPMDA	RLSMDA	RWRMDA
AUC	0.9885	0.9799	0.9171	0.9480	0.9358	0.9055
AUPR	0.5174	0.5047	0.3865	0.2045	0.2818	0.1907

TABLE 4 | Parameter tuning for PMFMDA based on 5-fold CV.

AUC	$\lambda_U = \lambda_V = 1$	$\lambda_U = \lambda_V = 0.1$	$\lambda_U = \lambda_V = 0.01$
$\lambda_1 = \lambda_2 = 1$	0.7905	0.7728	0.7588
$\lambda_1 = \lambda_2 = 0.1$	0.9040	0.8507	0.8381
$\lambda_1 = \lambda_2 = 0.01$	0.9185	0.9032	0.8692

prediction result will be achieved when the value of λ_1 and λ_2 are large and the value of λ_1 and λ_2 are small. This result further confirms the effectiveness of seeking an optimal combination of parameters in improving performance.

Finally, we explore the effect of the disease similarity and miRNA similarity on prediction performance. Specifically, we perform global 5-fold CV with parameters λ_1 and λ_2 setting to zero (**Figure 4**) in the HMDD (V2.0) dataset. We can see that the two similarities do contribute to prediction performance. In addition, PMFMDA achieve good results even in the model without integrating disease and miRNA similarity. However, this model is not good in predicting the association of new diseases or new miRNAs.

Case Studies

Another aspect of PMFMDA's strong predictive power is in case studies. Here, all the associations included in the HMDD (V2.0) database are used as training for the model, and the unincorporated associations are considered candidates for verification. In addition, miRCancer (Xie et al., 2013) and dbDEMC (Yang et al., 2010) were used to verify the correctness of the predictions. In this work, we mainly study three diseases including esophageal tumors, breast tumors, and lung tumors, and perform detailed analyses of the top 10 candidates predicted by PMFMDA in each disease (see **Table 5**).

Esophageal tumors are a disease with high morbidity and high mortality in the digestive system (Kano et al., 2010; He et al., 2012). Early diagnosis plays a crucial role in its treatment (Azmi, 2012). In this study, we use PMFMDA to identify potential miRNAs associated with esophageal tumors. The top 10 miRNAs to be all confirmed by the database were associated with esophageal tumors (see **Table 5**).

Breast neoplasm is the malignant tumor that is prone to occur in women, it is a systemic malignant disease, for which many related genes have been discovered (Venkatadri et al., 2016). MicroRNA (miRNA), as a kind of small RNA, can specifically bind to the 3′ untranslated region of its target mRNA, causing translational inhibition or degradation of target mRNA, and playing an oncogene in the process of cell growth and differentiation (Miller et al., 2008). Thus, MiRNAs present a new way for the study of pathogenic genes in breast neoplasms. As we can see from **Table 5**, 9 of the top 10 predictions have been confirmed by the relevant databases.

TABLE 5 | PMFMDA infers the top 10 miRNA candidates for the three selected diseases.

Cancer	Number of miRNAs identified by the literature	Top 10					
		Rank	miRNAs	Evidence	Rank	miRNAs	Evidence
Esophageal neoplasms	10	1	mir-17	dbDEMC	6	mir-1	dbDEMC
		2	mir-18a	dbDEMC	7	mir-200b	dbDEMC
		3	mir-221	dbDEMC	8	mir-222	dbDEMC
		4	mir-16	dbDEMC	9	mir-29a	dbDEMC
		5	mir-19b	dbDEMC	10	mir-133b	dbDEMC
Breast neoplasms	9	1	mir-142	miRCancer	6	mir-138	dbDEMC
		2	mir-150	dbDEMC, miRCancer	7	mir-15b	dbDEMC
		3	mir-106a	dbDEMC	8	mir-192	dbDEMC
		4	mir-99a	dbDEMC, miRCancer	9	mir-378a	Unconfirmed
		5	mir-130a	dbDEMC	10	mir-196b	dbDEMC
lung neoplasms	9	1	mir-16	dbDEMC	6	mir-99a	dbDEMC
		2	hsa-mir-15a	dbDEMC	7	mir-429	dbDEMC, miRCancer
		3	hsa-mir-106b	dbDEMC	8	mir-302b	dbDEMC, miRCancer
		4	mir-195	dbDEMC, miRCancer	9	mir-130a	dbDEMC
		5	mir-141	dbDEMC	10	mir-296	Unconfirmed

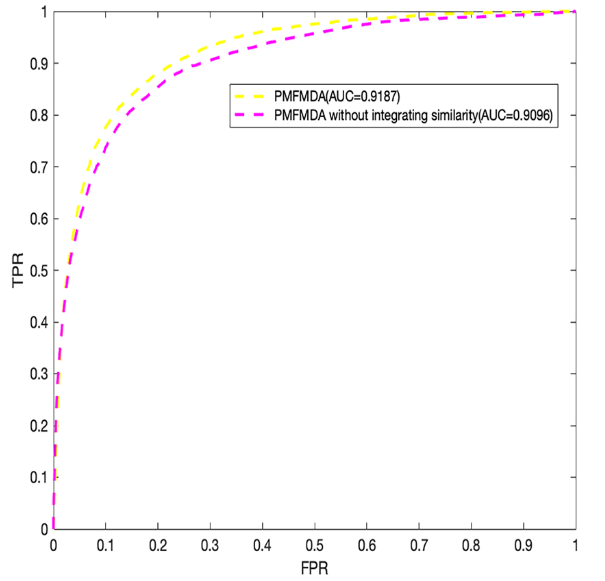


FIGURE 4 | Performance evaluation of PMFMDA in two situations for 5-fold cross validation. (1) PMFMDA with similarity information; (2) PMFMDA without similarity information.

The death rate from lung neoplasms is extremely high. About 1.3 million people die of lung neoplasms every year, accounting for about one-third of all neoplasms deaths worldwide (Yu et al., 2015; Sun et al., 2016). miRNAs have been found as a tumor suppressor gene and lung neoplasms. For example, Gu et al. found that miR-99a was significantly expressed in lung cancer tissues and lung neoplasm cells. In addition, the expression level of miR-99a is correlated with clinicopathological factors, the clinical stage and lymph node metastasis of lung cancer patients. We use PMFMDA to predict potential related miRNAs in lung tumors. As shown in **Table 5**, we can find that only one of the top 10 related miRNAs predicted is unconfirmed.

For a clear view, we show the top 20 miRNAs associated networks predicting three tumors in **Figure 5**. It is worth noting that some miRNA candidates are usually associated with several diseases. For example, mir-15b and mir-130a are associated with both Prostatic lung and Breast Neoplasms. Has-mir-16 is associated with both Esophageal Neoplasms and lung Neoplasms.

DISCUSSION

It is known that miRNAs often play an irreplaceable role in biological processes related to human diseases (Shen et al., 2017).

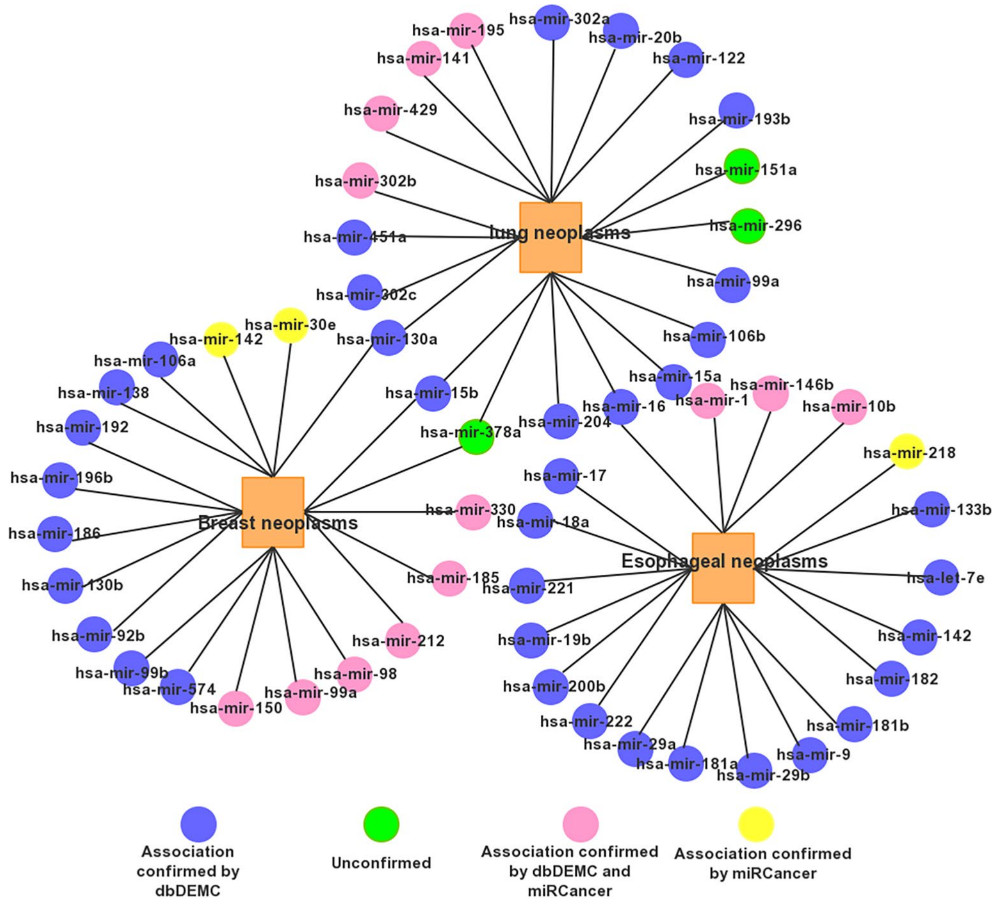


FIGURE 5 | The network of the top 20 predicted associations for the three selected diseases via PMFMDA.

Accurately inferring disease-related potential miRNAs is helpful for us to investigate the pathogenesis of the disease and find a more effective treatment. In this study, we construct a mathematical model based on probability matrix factorization (PMFMDA) to identifying potential miRNAs–disease associations. PMFMDA outperform a few state-of-the-art models in the HMDD V2.0 database due to a few factors. First, PMFMDA not only uses known correlation data, but also integrates the similarities between miRNAs and between diseases. This has enabled PMFMDA to achieve good results in predicting isolated disease-associated miRNAs since theoretically similar miRNAs may associate with similar diseases. Second, the model is a semi-supervised model, which does not rely on negative samples. Thus, it is better than most machine learning algorithms with strong requirement for good negative samples. Finally, in the model solving process, we use the alternating gradient descent algorithm to find the optimal solution to ensure the reliability of disease feature vectors and miRNA feature vectors. In terms of experiment, PMFMDA achieves the highest AUC (0.9187, 0.9237, respectively) in 5-fold CV and global LOOCV, demonstrates its most reliable prediction performances. At the same time, we also perform CV_d experiments to measure the ability of PMFMDA to predict miRNAs associated with novel diseases. We conduct CV testing on 8 common diseases, which have at least 80 associations are verified (Xuan et al., 2015). PMFMDA achieves the highest average AUPRs of 0.6687. Finally, to make the more comprehensive test of PMFMDA, we use the three most common diseases in humans for research. The number of other database validations in the top 20 predicted miRNAs for esophageal tumors, breast tumors, and lung tumors are found to be 20, 19, and 17, respectively. In conclusion, PMFMDA has achieved good results in predicting the potential association of miRNA disease and predicting new disease-associated miRNAs and can be used as a very useful supplement to existing prediction models.

Although quite satisfactory results have been achieved from PMFMDA, there are still some limitations to this approach. Firstly, we only use semantic similarity and the Gaussian kernel similarity to construct disease similarity network. It may be helpful to improve the predictive performance of PMFMDA by integrating disease or miRNA similarity from multiple data sources such sequence similarity. Secondly, the public data sets used in this study may have noise and outliers. A preprocessing step for de-noising and dimension reduction in raw input data might be useful. Thirdly, in the process of solving PMFMDA, the gradient descent method often obtains the local optimal solution, and how to further optimize its solution helps to improve the prediction performance of PMFMDA. Finally, as more and more miRNAs and disease associations are confirmed, collecting more validated data will help us to conduct more in-depth research.

DATA AVAILABILITY STATEMENT

The program and data used in this study are publicly available at: https://github.com/xujunlin123/PMFMDA.git.

AUTHOR CONTRIBUTIONS

JY, JX, LC, GT and BL conceived the concept of the work. JX, PW, WZ, YM, and JL performed the experiments. JX and JY wrote the paper. GT helped in revising the manuscript.

FUNDING

This study is supported by National Nature Science Foundation of China (Grant Nos. 11171369, 61272395, 61370171, 61300128, 61472127, 61572178, 61672214, and 61772192), and the Natural Science Foundation of Hunan, China (Grant Nos. 2018JJ2461, 2018JJ3570).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01234/full#supplementary-material

REFERENCES

Azmi, A. S. (2012). Systems biology in cancer research and drug discovery. Springer. doi: 10.1007/978-94-007-4819-4

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. Q. (2010). Development of the human cancer microRNA network. *Silence* 1, 6. doi: 10.1186/1758-907X-1-6

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4, 5501. doi: 10.1038/srep05501

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

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

Chen, X., 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

Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S., Lin, Y. L., Lee, W. H., et al. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* 44, D239–D247. doi: 10.1093/nar/gkv1258

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

Feng, P., Zhang, J., Tang, H., Chen, W., and Lin, H. (2017). Predicting the organelle location of noncoding RNAs using pseudo nucleotide compositions. *Interdiscip. Sci. Comput. Life Sci.* 9, 540–544. doi: 10.1007/s12539-016-0193-4

Gu, C., Bo, L., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6, 36054. doi: 10.1038/srep36054

Ha, J., Park, C., and Park, S. (2019). PMAMCA: prediction of microRNA-disease association utilizing a matrix completion approach. *BMC Syst. Biol.* 13 (1), 1–13. doi: 10.1186/s12918-019-0700-4

Hammond, S. M. (2015). An overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi: 10.1016/j.addr.2015.05.001

He, B., Yin, B., Wang, B., Xia, Z., Chen, C., and Tang, J. (2012). MicroRNAs in esophageal cancer (review). *Mol. Med. Rep.* 6, 459–465. doi: 10.3892/mmr.2012.975

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, S2–S2. doi: 10.1186/1752-0509-4-S1-S2

Kano, M., Seki, N., Kikkawa, N., Fujimura, L., Hoshino, I., Akutsu, Y., et al. (2010). MiR-145, miR-133a and miR-133b: tumor-suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma. *Int. J. Cancer* 127, 2804–2814. doi: 10.1002/ijc.25284

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, 1070–1074. doi: 10.1093/nar/gkt1023

Liang, C., Yu, S., and Luo, J. (2019). Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs. *PloS Comput. Biol.* 15, e1006931. doi: 10.1371/journal.pcbi.1006931

Liu, Y., Luo, J., and Ding, P. (2019). Inferring MicroRNA targets based on restricted boltzmann machines. *IEEE J. Biomed. Heal. Inf.* 23, 427–436. doi: 10.1109/JBHI.2018.2814609

Lu, X., Qian, X., Li, X., Miao, Q., and Peng, S. (2019). DMCM: a data-adaptive mutation clustering method to identify cancer-related mutation clusters. *Bioinformatics* 35, 389–397. doi: 10.1093/bioinformatics/bty624

Luo, J., Ding, P., Liang, C., Cao, B., and Chen, X. (2017a). Collective prediction of disease-associated miRNAs based on transduction learning. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14, 1468–1475. doi: 10.1109/TCBB.2016.2599866

Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017b). Predicting microRNA-disease associations using Kronecker Regularized Least Squares based on heterogeneous omics data. *IEEE Access* 5, 2503–2513. doi: 10.1109/ACCESS.2017.2672600

Miller, T. E., Ghoshal, K., Ramaswamy, B., Roy, S., Datta, J., Shapiro, C. L., et al. (2008). MicroRNA-221/222 confers tamoxifen resistance in breast cancer by targeting p27Kip1. *J. Biol. Chem.* 283, 29897–29903. doi: 10.1074/jbc.M804612200

Peng, W., Lan, W., Yu, Z., Wang, J., and Pan, Y. (2017). A framework for integrating multiple biological networks to predict MicroRNA-disease associations. *IEEE Trans. Nanobiosci.* 16, 100–107. doi: 10.1109/TNB.2016.2633276

Salakhutdinov, R., and Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. in, 880–887. doi: 10.1145/13901561390267

Shen, Z., Zhang, Y.-H., Han, K., Nandi, A. K., Honig, B., and Huang, D.-S. (2017). miRNA-disease association prediction with collaborative matrix factorization. *Complexity* 2017, 1–9. doi: 10.1155/2017/2498957

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., et al. (2013). Walking the interactome to identify human miRNA-disease associations;through the functional link between miRNA targets and disease genes. *BMC Syst. Biol.* 7, 101. doi: 10.1186/1752-0509-7-101

Sun, M., Hong, S., Li, W., Wang, P., You, J., Zhang, X., et al. (2016). MIR-99a regulates ROS-mediated invasion and migration of lung adenocarcinoma cells by targeting NOX4. *Oncol. Rep.* 35, 2755–2766. doi: 10.3892/or.2016.4672

Venkatadri, R., Muni, T., Iyer, A. K. V., Yakisich, J. S., and Azad, N. (2016). Role of apoptosis-related miRNAs in resveratrol-induced breast cancer cell death. *Cell Death Dis.* 7, e2104. doi: 10.1038/cddis.2016.6

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 34, 239–248. doi: 10.1093/bioinformatics/btx545

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). MiRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 29, 638–644. doi: 10.1093/bioinformatics/btt014

Xu, J., Li, C.-X., Lv, J.-Y., Li, Y.-S., Xiao, Y., Shao, T.-T., et al. (2011). Prioritizing candidate disease mirnas by topological features in the mirna target-dysregulated network: case study of prostate cancer. *Mol. Cancer Ther.* 10, 1857–1866. doi: 10.1158/1535-7163.MCT-11-0055

Xuan, P., Han, K., Guo, Y., Li, J., Li, X., Zhong, Y., et al. (2015). Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics* 31, 1805–1815. doi: 10.1093/bioinformatics/btv039

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 11, 1–8. doi: 10.1186/1471-2164-11-S3-I1

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

Yu, S. H., Zhang, C. L., Dong, F. S., and Zhang, Y. M. (2015). miR-99a suppresses the metastasis of human non-small cell lung cancer cells by targeting AKT1 signaling pathway. *J. Cell. Biochem.* 116, 268–276. doi: 10.1002/jcb.24965

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, 135–138. doi: 10.1093/nar/gkw728

Conflict of Interest: The authors JL and GT were employed by Geneis Beijing Co., Ltd.

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

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fgene.2020.00017

The Functional Effects of Key Driver KRAS Mutations on Gene Expression in Lung Cancer

Jisong Zhang[1], Huihui Hu[1], Shan Xu[1], Hanliang Jiang[1], Jihong Zhu[2], E. Qin[3], Zhengfu He[4] and Enguo Chen[1*]*

[1] Department of Pulmonary and Critical Care Medicine, Sir Run Run Shaw Hospital of Zhejiang University, Hangzhou, China, [2] Department of Anesthesiology, Sir Run Run Shaw Hospital of Zhejiang University, Hangzhou, China, [3] Department of Respiratory Medicine, Shaoxing People's Hospital (Shaoxing Hospital, Zhejiang University School of Medicine), Shaoxing, China, [4] Department of Thoracic Surgery, Sir Run Run Shaw Hospital of Zhejiang University, Hangzhou, China

Lung cancer is a common malignant cancer. Kirsten rat sarcoma oncogene (KRAS) mutations have been considered as a key driver for lung cancers. KRAS p.G12C mutations were most predominant in NSCLC which was comprised about 11–16% of lung adenocarcinomas (p.G12C accounts for 45–50% of mutant KRAS). But it is still not clear how the KRAS mutation triggers lung cancers. To study the molecular mechanisms of KRAS mutation in lung cancer. We analyzed the gene expression profiles of 156 KRAS mutation samples and other negative samples with two stage feature selection approach: (1) minimal Redundancy Maximal Relevance (mRMR) and (2) Incremental Feature Selection (IFS). At last, 41 predictive genes for KRAS mutation were identified and a KRAS mutation predictor was constructed. Its leave one out cross validation MCC was 0.879. Our results were helpful for understanding the roles of KRAS mutation in lung cancer.

Keywords: Kirsten rat sarcoma oncogene (KRAS), mutation, lung cancer, predictor, gene expression

OPEN ACCESS

Edited by:
*Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China*

Reviewed by:
*Jing Feng,
Tianjin Medical University General Hospital, China
Xiaoying Huang,
Wenzhou Medical University, China*

***Correspondence:**
*Zhengfu He
hezhengfu@zju.edu.cn
Enguo Chen
3195024@zju.edu.cn*

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

Received: *11 October 2019*
Accepted: *07 January 2020*
Published: *04 February 2020*

Citation:
Zhang J, Hu H, Xu S, Jiang H, Zhu J, Qin E, He Z and Chen E (2020) The Functional Effects of Key Driver KRAS Mutations on Gene Expression in Lung Cancer. Front. Genet. 11:17. doi: 10.3389/fgene.2020.00017

INTRODUCTION

Lung cancer, known as a malignant cancer which defined as the overgrowth of uncontrolled cell in lung tissues, has proved be a key cause of cancer death. Each year, 1.3 million people die of lung cancer (Jemal et al., 2006; Jemal et al., 2011). Non-small-cell lung cancer (NSCLC) accounts for more than 85% of diagnosed lung cancer patients (Morgensztern et al., 2010). NSCLC can be further divided into adenocarcinoma, squamous cell carcinoma (SCC), and large cell carcinoma (Sandler et al., 2006; Morgensztern et al., 2010).

At present, the pathogenesis of lung cancer is not very clear, but is generally believed that one of the most important reason is the accumulation of mutations including single nucleotide transformation, small fragments of insertions and deletions, the changes of copy number, and chromosome rearrangement. Moreover, these mutations are closed with cell proliferation, invasion, metastasis, and apoptosis (Scagliotti et al., 2008; Liu et al., 2012). So, studying mutations in living systems will be helpful to understand how mutations are associated with lung-cancer biological processes.

In the last decade, researchers have uncovered the source of one of the important mutations is called as Kirsten rat sarcoma oncogene (KRAS) mutations in lung cancers using molecular studies (Gautschi et al., 2007). KRAS is the principal isoform of RAS. KRAS p.G12C mutations were most predominant in NSCLC which was comprised about 11–16% of lung adenocarcinomas (p.G12C accounts for 45–50% of mutant KRAS) (Cox et al., 2014). Other common KRAS mutations in lung cancer are G12V and G12D. In other cancers, such as pancreatic cancer and colorectal cancer, KRAS mutations are also frequent. Based on the TCGA data in cBioPortal (Gao et al., 2013), the most frequent KRAS mutations in pancreatic cancer are G12D, G12V, and G12R; the most frequent KRAS mutations in colorectal cancer are G12D, G12V, and G13D. KRAS may be a good lung cancer therapeutic target for searching potential drugs.

As above mentioned, mutations in KRAS is the most usual mutations that occur in lung cancer, especially in NSCLC (Mao et al., 1994; Mills et al., 1995; Nakamoto et al., 2001). KRAS mutation is more frequent in Caucasians than in Asians. Moreover, smokers may have more KRAS mutations than nonsmokers (Westcott and To, 2013; Ferrer et al., 2018). Single amino acid substitutions in codon 12 were most common KRAS mutations in NSCLC (Graziano et al., 1999). Therefore, the search for how the KRAS mutations affected the gene in lung cancer has been a long-standing goal in cancer biology.

In this study, to study the functional effects of key driver KRAS mutations on gene expression in lung cancer, we analyzed the gene expression profiles of 156 lung cancer cell lines with KRAS mutations and other 3,582 lung cancer cell lines without KRAS mutations. Forty-one discriminative genes for KRAS mutations were identified using two stage feature selection approach: (1) minimal Redundancy Maximal Relevance (mRMR) and (2) Incremental Feature Selection (IFS).

METHODS

The Gene Expression Profiles of Cell Lines With and Without KRAS Mutations

To identify the key genes that distinguishes key driver KRAS mutations from other mutations, we downloaded the gene expression profiles of 156 lung cancer cell lines with KRAS mutations as positive samples and other 3,582 lung cancer cell lines without KRAS mutations as negative samples from publicly available Gene Expression Omnibus (GEO) database under accession number of GSE83744 (Berger et al., 2016). The expression levels of 978 representative genes from Broad Institute Human L1000 landmark were measured. The L1000 landmark was derived from the Connectivity Map (CMap) project (Subramanian et al., 2017). CMap is a large gene-expression dataset of human cells perturbed with many chemicals and genetic reagents (Lamb et al., 2006). These 1,000 genes were sensitive to perturbations and can reflect 81% of non-measured transcripts (Subramanian et al., 2017).

Two Stage Feature Selection Approach

We applied two stage feature selection approach to select the biomarker genes. First, the genes were ranked based on not only their relevance with mutation samples, but also their redundancy among genes using the mRMR algorithm (Peng et al., 2005). It had a wide range of applications in bioinformatics for feature selection (Chen et al., 2018c; Chen et al., 2019e; Li and Huang, 2018; Li et al., 2019b; Wang and Huang, 2019a). As the equation shown below, Ω_s, Ω_t and Ω were the set of m selected genes, n to-be-selected genes, and all m+n genes, respectively. We use mutual information (I) to measure the relevance of the expression levels of gene g from Ω_t with KRAS mutation status t (Huang and Cai, 2013):/>

$$\mathrm{D} = I(\mathrm{g}, \mathrm{t}) \tag{1}$$

Meanwhile, the redundancy R of the gene g with the selected genes in Ω_s can be calculated as below:

$$\mathrm{R} = \frac{1}{m}\left(\Sigma_{g_i \in \Omega_s} I(g, g_i)\right) \tag{2}$$

The optimal gene g_j from Ω_t with max relevance with KRAS mutation status t and min redundancy with the selected genes in Ω_s can be selected by maximizing mRMR function listed below

$$\max_{g_j \in \Omega_t}\left[I(g_j, \mathrm{t}) - \frac{1}{m}\left(\Sigma_{g_i \in \Omega_s} I(g_j, g_i)\right)\right] (j = 1, 2, \ldots, n) \tag{3}$$

With N round evaluations, genes can be ranked as

$$\mathrm{S} = \left\{g_1', g_2', \ldots, g_h', \ldots, g_N',\right\} \tag{4}$$

The top ranked genes were associated with KRAS mutation status, and had little redundancy with other genes. Such genes were suitable for biomarkers. The top 200 genes were further analyzed at the second stage.

The second stage was to determine the number of selected genes using the IFS method (Chen et al., 2018b; Chen et al., 2019b; Chen et al., 2019c; Chen et al., 2019d; Chen et al., 2019f; Li et al., 2019a; Pan et al., 2019a; Pan et al., 2019b;). To do so, 200 classifiers were constructed using top 1, top 2, top 200 genes. The LOOCV (leave-one-out cross validation) MCC (Mathew's correlation coefficient) of the top k-gene classifier was calculated each time.

We tried several different classifiers: (1) SVM (Support Vector Machine) (Jiang et al., 2019; Yan et al., 2019; Chen et al., 2019a; Li et al., 2019a; Pan et al., 2019a; Wang and Huang, 2019b; Chen et al., 2019d), (2) 1NN (1 Nearest Neighbor) (Lei et al., 2013; Chen et al., 2016; Wang et al., 2017a), (3) 3NN (3 Nearest Neighbors), (4) 5NN (5 Nearest Neighbors), (5) Decision Tree (DT) (Huang et al., 2008; Huang et al., 2011; Chen et al., 2015), (6) Neural Network (NN) (Liu et al., 2017; Pan et al., 2018; Chen et al., 2019e). The function svm from R package e1071, function knn from R package class, function rpart from R package rpart, function nnet from R package nnet were used to apply these classification algorithms.

Based on the IFS curve in which x-axis was the number of genes and y-axis was the corresponding LOOCV MCC, we can decide the best gene combinations we should select. The peak of the curve was the optimal selection.

Prediction Performance Evaluation of the Classifier

As we mentioned before, the prediction performance of each classifier was evaluated with leave-one-out cross validation (LOOCV) (Cui et al., 2013; Yang et al., 2014). It will go through N rounds and each sample will be tested during the N rounds. In each round, one sample will be tested using the model trained with the other N-1 samples. It can objectively evaluate all samples (Chou, 2011).

The performance metrics, including Sensitivity (Sn), Specificity (Sp), Accuracy (ACC), and Mathew's correlation coefficient (MCC) were all calculated:

$$S_n = \frac{TP}{TP + FN} \tag{5}$$

$$S_p = \frac{TN}{TN + FP} \tag{6}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

where TP, TN, FP, and FN stand for the number of true positive samples, true negative samples, false positive samples, and false negative samples, respectively. Since the sizes of KRAS mutation + samples and KRAS mutation - samples were imbalance and MCC can trade-off sensitivity and specificity (Chen et al., 2018a; Li et al., 2018; Pan et al., 2018; Pan et al., 2019a; Pan et al., 2019b), MCC was used as the main performance metric.

RESULTS AND DISCUSSION

The Genes That Showed Different Expression Pattern Between KRAS Mutations From Other Mutations Samples

The top 200 most informative genes for KRAS mutations were identified using the mRMR method which has been widely used in bioinformatics filed (Zhao et al., 2013; Zhang et al., 2016). The C/C++ version software written by Peng et al. (Peng et al., 2005; Best et al., 2017) (http://home.penglab.com/proj/mRMR/) was used to apply the mRMR algorithm. Unlike the traditional statistical test based univariate feature selection methods, mRMR considers the relevance between gene expression and KRAS mutation status, and the redundancy among genes.

The Optimal Biomarkers Identified From the mRMR Gene List With IFS Methods

After genes were ranked by mRMR, the IFS procedure was applied to find the optimal number of genes to be selected. The IFS curve in **Figure 1** showed the relationship between the number of genes and their MCCs. The peak LOOCV MCCs of SVM, 1NN, 3NN, 5NN, DT, and NN were 0.858 with 8 genes, 0.853 with 48 genes, 0.879 with 41 genes, 0.878 with 59 genes, 0.871 with 69 genes, 0.842 with 174 genes. 3NN performed best. The corresponding 41 genes were shown in **Table 1**.

The Prediction Metrics of the 41 Genes

The 41 genes were chosen with two stage feature selection methods: mRMR and IFS. To more carefully evaluate their prediction power, we checked their confusion matrix which showed the overlaps between actual KRAS mutation status and predicted KRAS mutation status using 3NN (**Table 2**). The LOOCV sensitivity, specificity, accuracy, and MCC were 0.840, 0.997, 0.991, and 0.879, respectively.

The Network Associations Between KRAS and the 41 Genes

We searched KRAS and the eight genes in STRING database Version: 11.0 (https://string-db.org) and **Figure 2** showed their functional association networks. It can be seen that 20 out of 41 genes (CCND3, CDK19, CEBPA, CEBPD, CSNK1E, CTSL, DUSP6, GRB10, HMGA2, MMP1, MTHFD2, NR3C1, PAK4, PMAIP1, RAP1GAP, SDHB, STX1A, TP53, TRIB3, UBE2L6)

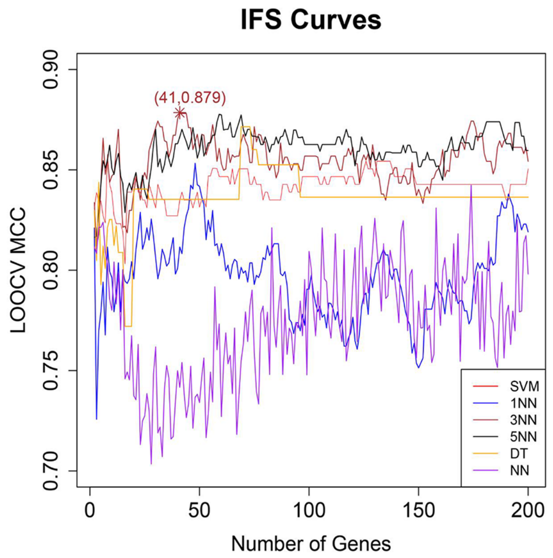


FIGURE 1 | The IFS curves of six different classifiers. The x-axis was the number of genes and the y-axis was the then leave one out cross validation (LOOCV) MCC. The red, blue, brown, black, orange, and purple curves were the IFS results of SVM, 1NN, 3NN, 5NN, DT, and NN, respectively. Peak LOOCV MCCs of SVM, 1NN, 3NN, 5NN, DT, and NN were 0.858 with 8 genes, 0.853 with 48 genes, 0.879 with 41 genes, 0.878 with 59 genes, 0.871 with 69 genes, 0.842 with 174 genes. 3NN performed best. Therefore, the corresponding 41 genes were finally selected.

TABLE 1 | The 41 genes selected by mRMR and IFS.

Rank	Gene	Rank	Gene
1	CTSL1	22	CCDC92
2	GNPDA1	23	BRP44
3	TRIB3	24	CDK19
4	STX1A	25	CD320
5	PHKA1	26	ATP1B1
6	CSNK1E	27	DRAP1
7	COL4A1	28	DUSP6
8	CEBPA	29	RAP1GAP
9	CEBPD	30	GALE
10	NSDHL	31	SSBP2
11	TP53	32	UBE2L6
12	MTHFD2	33	CCND3
13	RGS2	34	PAFAH1B1
14	NR3C1	35	RBM6
15	PPIC	36	C5
16	BAMBI	37	SDHB
17	PAK4	38	GRB10
18	FEZ2	39	UFM1
19	KTN1	40	ARL4C
20	HMGA2	41	PMAIP1
21	MMP1		

TABLE 2 | The confusion matrix of actual sample classes and predicted sample classes using 3NN.

	Predicted KRAS mutation +	Predicted KRAS mutation –
Actual KRAS mutation +	131	25
Actual KRAS mutation –	10	3572
MCC = 0.879	Sensitivity = 0.840	Specificity = 0.997

had direct interactions with KRAS. The STRING network results supported that most of the 41 genes had direct interactions with KRAS.

The Biological Significance of the Selected Genes in Lung Cancer

As mentioned earlier, we used mRMR algorithm and IFS program to screen out 41 genes which may be molecular markers for identifying KARS mutations. Subsequently, we reviewed studies of these genes in lung cancer and other cancers with high frequency of KARS mutations such as colorectal and pancreatic cancer. In the study of Zhang X et al., Tribbles-3 (TRIB3) pseudokinase can activate the β-catenin signal pathway, which in turn promotes the proliferation and migration of NSCLC cells (Zhang et al., 2019). In addition, blocking the activity of TRIB3 may be one of the mechanisms for the treatment of lung cancer (Ding et al., 2018). Wang X et al. have found that PAK4 is significantly associated with poor prognosis of NSCLC (Wang et al., 2016b), and LIMK1 phosphorylation mediated by it regulates the migration and invasion of NSCLC. Therefore, PAK4 may be an important prognostic indicator and a potential molecular target for treatment of NSCLC (Cai et al., 2015). HMGA2 affects apoptosis and is highly expressed in metastatic LUAD through Caspase 3/9 and Bcl-2. It is also considered to be a biomarker and potential therapeutic target for lung cancer therapy (Kumar et al., 2014; Gao et al., 2017b). A meta-analysis of lung cancer showed that metalloproteinase 1 (MMP1)-16071G/2G polymorphism was a risk factor for lung cancer in Asians (Li et al., 2015). In addition, DUSP6 rs2279574 gene polymorphism is thought to predict the survival time of NSCLC patients after chemotherapy (Wang et al., 2016a). Cyclin D3 gene (CCND3) is a key cell cycle gene of NSCLC, which can promote the growth of LUAD (Zhang et al., 2017). Casein kinase I epsilon (CSNK1E), a circadian rhythm gene, whose genetic variation has a very significant correlation with the risk of lung cancer (Ortega and Mas-Oliva, 1986). CEPBA, can be used as a new tumor suppressor factor, Lu H et al. through clinical experiments, it was found that up-regulation of CEBPA is an effective method for the treatment of human NSCLC (Halmos et al., 2002; Lu et al., 2015). In addition, a comprehensive analysis of lung cancer genes by, Lv M shows that CEPBD may be involved in the development of lung cancer (Lv and Wang, 2015). TP53 mutation is very common in NSCLC and is considered to be a marker of poor prognosis and a prognostic indicator of lung cancer (Gao et al., 2017a; Labbe et al., 2017). Methylenetetrahydrofolate dehydrogenase 2 (MTHFD2) has redox homeostasis and can be used in the treatment of lung cancer (Nishimura et al., 2019). NR3C1 is reported to be involved in the pathways related to the biological process of lung cancer, and as a gene marker has a significant correlation with the survival of LUAD (Zhao et al., 2015; Luo et al., 2018). Cathepsin L1, as a protein was encoded by the CTSL1 gene, could reduce the cellular matrix and proteolytic cascades which resulting to promote invasion or metastatic activity (Duffy, 1996; Turk et al., 2012). Elevated expression of extracellular Cathepsin L was related with cancer progression of lung cancer cells (Okudela et al., 2016). Moreover, Cathepsin L is viewed as a downstream target of oncogenic KRAS mutations.

The above genes have not only been proved to be closely related to the prognosis, diagnosis, and treatment of lung cancer, but also have a direct interaction with KRAS. Some of the 41 selected genes have no direct interaction with KRAS, but are considered to be involved in the occurrence and development of lung cancer. RBM6 protein is located at 3p21.3, and its expression changes regulate many of the most common abnormal splicing events in lung cancer (Sutherland et al., 2010; Coomer et al., 2019). The double up-regulation of RGS2 gene is related to the poor overall survival rate of patients with lung adenocarcinoma (Yin et al., 2016). Epigenetic silencing of BAMBI has been identified as a marker of NSCLC, and overexpression of BAMBI may become a new target for the treatment of this cancer (Marwitz et al., 2016; Wang et al., 2017b). Overexpression of PAFA-H1B1 can lead to the occurrence and poor prognosis of lung cancer (Lo et al., 2012). Collagen alpha-1(IV) chain (COL4A1), encoded by the COL4A1 gene, was found previously to play a crucial role in the coordinating alveolar morphogenesis and formatting the epithelium vasculature lung tissue (Abe et al., 2017).

The Potential Roles of the Selected Genes in Other Cancers

KRAS related genes are likely to be diagnostic, prognostic markers and therapeutic targets of lung cancer. We also

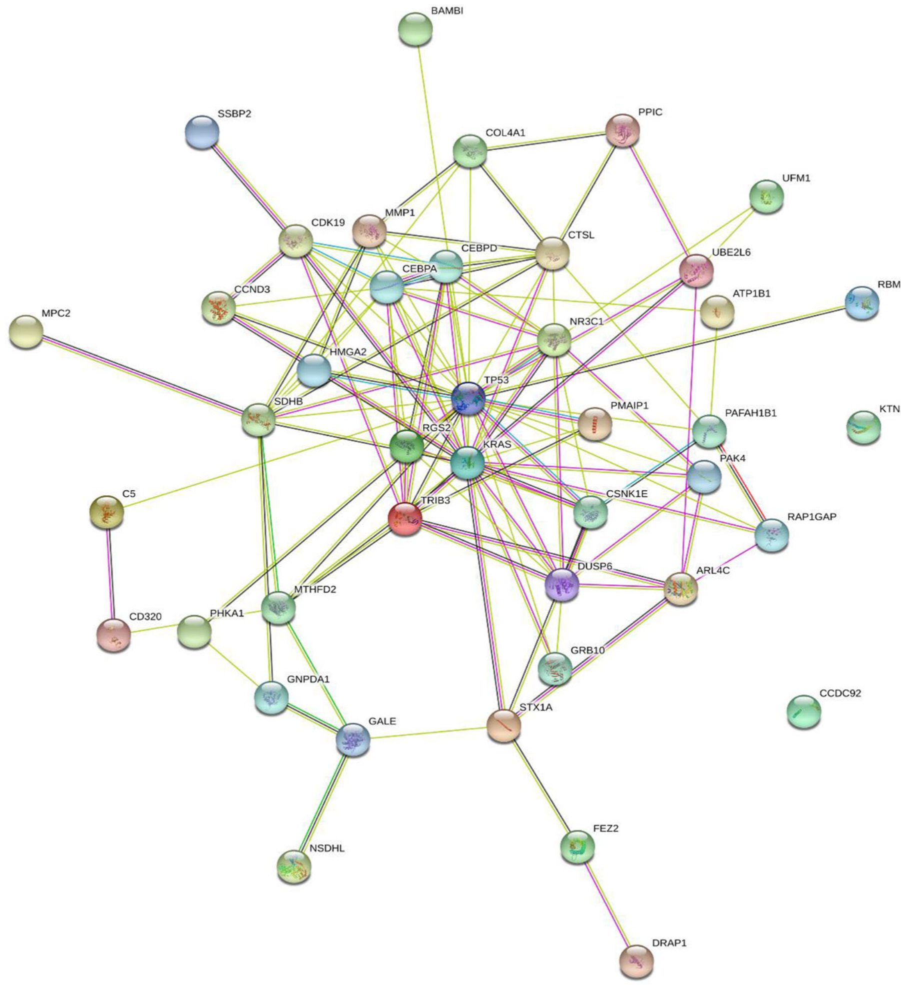


FIGURE 2 | The functional association network of KRAS and the selected genes based on STRING database. Twenty out of 41 genes (CCND3, CDK19, CEBPA, CEBPD, CSNK1E, CTSL, DUSP6, GRB10, HMGA2, MMP1, MTHFD2, NR3C1, PAK4, PMAIP1, RAP1GAP, SDHB, STX1A, TP53, TRIB3, UBE2L6) had direct interactions with KRAS. Each line represented an interaction supported by different evidences. The skype-blue, purple, green, red, blue, grass green, black, and navy-blue edges were interactions from curated databases, experiment, gene neighborhood, gene fusions, gene co-occurrence, text mining, co-expression, and protein homology, respectively. For more detailed explanations, please refer to STRING database (https://string-db.org).

looked for studies of these genes and KRAS high-frequency mutations in other cancers, mainly in colorectal and pancreatic cancer. According to Hua F et al., TRIB 3 gene knockout can reduce the occurrence of colon tumors in mice, reduce the migration of colorectal cancer cells, and reduce their growth in mouse transplanted tumors. The strategy of blocking the activity of TRIB3 can be used to treat colorectal cancer (Hua et al., 2019). Tyagi N et al. have found that PAK4 can maintain the stem cell phenotype of pancreatic cancer cells by activating STAT3 signal, which can be used as a new therapeutic target (Tyagi et al., 2016). TP53 mutation is associated with early stage of colorectal cancer (Laurent et al., 2011). There was a significant correlation between MMP1 and colon cancer mortality (Slattery and Lundgreen, 2014).

DATA AVAILABILITY STATEMENT

We downloaded the blood gene expression profiles of 156 KRAS mutations as positive samples and other 3582 mutations as negative samples from publicly available GEO (Gene Expression Omnibus) under accession number of GSE83744.

AUTHOR CONTRIBUTIONS

JZha conceived and designed the study. HH and SX performed data analysis. HJ wrote the paper. JZhu, EC and ZH reviewed and edited the manuscript. JZha approved final version of the manuscript. All authors read and approved the manuscript.

FUNDING

This study was supported by the Funds from Science Technology Department of Zhejiang Province (LGF19H010010), Medical and Health Research Foundation of Zhejiang Province (2016ZDB005, 2017ZD020), China, WU JIEPING MEDICAL foundation (320.6750.19092-12), Beijing Xisike Clinical Oncology Research Foundation (Y-HS2017-037) and Medical Health and Scientific Technology Project of Zhejiang Province (2019RC182).

REFERENCES

Abe, Y., Matsuduka, A., Okanari, K., Miyahara, H., Kato, M., Miyatake, S., et al. (2017). A severe pulmonary complication in a patient with COL4A1-related disorder: a case report. *Eur. J. Med. Genet.* 60 (3), 169–171. doi: 10.1016/j.ejmg.2016.12.008

Berger, A. H., Brooks, A. N., Wu, X., Shrestha, Y., Chouinard, C., Piccioni, F., et al. (2016). High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* 30 (2), 214–228. doi: 10.1016/j.ccell.2016.06.022

Best, M. G., Sol, N., In 't Veld, S., Vancura, A., Muller, M., Niemeijer, A. N., et al. (2017). Swarm intelligence-enhanced detection of non-small-cell lung cancer using tumor-educated platelets. *Cancer Cell* 32 (2), 238–252.e239. doi: 10.1016/j.ccell.2017.07.004

Cai, S., Ye, Z., Wang, X., Pan, Y., Weng, Y., Lao, S., et al. (2015). Overexpression of P21-activated kinase 4 is associated with poor prognosis in non-small cell lung cancer and promotes migration and invasion. *J. Exp. Clin. Cancer Res.* 34, 48. doi: 10.1186/s13046-015-0165-2

Chen, L., Chu, C., Huang, T., Kong, X., and Cai, Y. D. (2015). Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino Acids* 47 (7), 1485–1493. doi: 10.1007/s00726-015-1974-5

Chen, L., Zhang, Y. H., Huang, T., and Cai, Y. D. (2016). Gene expression profiling gut microbiota in different races of humans. *Sci. Rep.* 6, 23075. doi: 10.1038/srep23075

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018a). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell Biochem.* 119 (4), 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Zhang, Y.-H., Pan, X., Liu, M., Wang, S., Huang, T., et al. (2018b). Tissue Expression difference between mRNAs and lncRNAs. *Int. J. Mol. Sci.* 19 (11), 3416. doi: 10.3390/ijms19113416

Chen, L., Zhang, Y. H., Huang, G., Pan, X., Wang, S., Huang, T., et al. (2018c). Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genomics* 293 (1), 137–149. doi: 10.1007/s00438-017-1372-7

Chen, L., Pan, X., Zeng, T., Zhang, Y., Huang, T., and Cai, Y. (2019a). Identifying essential signature genes and expression rules associated with distinctive development stages of early embryonic cells. *IEEE Access* 7, 128570–128578. doi: 10.1109/ACCESS.2019.2939556

Chen, L., Pan, X., Zhang, Y.-h., Hu, X., Feng, K., Huang, T., et al. (2019b). Primary tumor site specificity is preserved in patient-derived tumor xenograft models. *Front. In Genet*. doi: 10.3389/fgene.2019.00738

Chen, L., Pan, X., Zhang, Y.-H., Huang, T., and Cai, Y.-D. (2019c). Analysis of gene expression differences between different pancreatic cells. *ACS Omega* 4 (4), 6421–6435. doi: 10.1021/acsomega.8b02171

Chen, L., Pan, X., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2019d). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* 120 (5), 7068–7081. doi: 10.1002/jcb.27977

Chen, L., Pan, X., Zhang, Y.-H., Liu, M., Huang, T., and Cai, Y.-D. (2019e). Classification of widely and rarely expressed genes with recurrent neural network. *Comput. Struct. Biotechnol. J.* 17, 49–60. doi: 10.1016/j.csbj.2018.12.002

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y. H., Yuan, F., et al. (2019f). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26 (1-2), 29–39. doi: 10.1038/s41434-018-0051-6

Chou, K. C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* 273 (1), 236–247. doi: 10.1016/j.jtbi.2010.12.024

Coomer, A. O., Black, F., Greystoke, A., Munkley, J., and Elliott, D. J. (2019). Alternative splicing in lung cancer. *Biochim. Biophys. Acta Gene Regul. Mech.* 1862 (11-12), 194388. doi: 10.1016/j.bbagrm.2019.05.006

Cox, A. D., Fesik, S. W., Kimmelman, A. C., Luo, J., and Der, C. J. (2014). Drugging the undruggable RAS: mission possible? *Nat. Rev. Drug Discovery* 13 (11), 828–851. doi: 10.1038/nrd4389

Cui, W., Chen, L., Huang, T., Gao, Q., Jiang, M., Zhang, N., et al. (2013). Computationally identifying virulence factors based on KEGG pathways. *Mol. Biosyst.* 9 (6), 1447–1452. doi: 10.1039/c3mb70024k

Ding, C. Z., Guo, X. F., Wang, G. L., Wang, H. T., Xu, G. H., Liu, Y. Y., et al. (2018). High glucose contributes to the proliferation and migration of non-small cell lung cancer cells via GAS5-TRIB3 axis. *Biosci. Rep.* 38 (2), BSR20171014. doi: 10.1042/BSR20171014

Duffy, M. J. (1996). PSA as a marker for prostate cancer: a critical review. *Ann. Clin. Biochem.* 33 (Pt 6), 511–519. doi: 10.1177/000456329603300604

Ferrer, I., Zugazagoitia, J., Herbertz, S., John, W., Paz-Ares, L., and Schmid-Bindert, G. (2018). KRAS-Mutant non-small cell lung cancer: From biology to therapy. *Lung Cancer* 124, 53–64. doi: 10.1016/j.lungcan.2018.07.013

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* 6 (269), pl1. doi: 10.1126/scisignal.2004088

Gao, W., Jin, J., Yin, J., Land, S., Gaither-Davis, A., Christie, N., et al. (2017a). KRAS and TP53 mutations in bronchoscopy samples from former lung cancer patients. *Mol. Carcinog.* 56 (2), 381–388. doi: 10.1002/mc.22501

Gao, X., Dai, M., Li, Q., Wang, Z., Lu, Y., and Song, Z. (2017b). HMGA2 regulates lung cancer proliferation and metastasis. *Thorac. Cancer* 8 (5), 501–510. doi: 10.1111/1759-7714.12476

Gautschi, O., Huegli, B., Ziegler, A., Gugger, M., Heighway, J., Ratschiller, D., et al. (2007). Origin and prognostic value of circulating KRAS mutations in lung cancer patients. *Cancer Lett.* 254 (2), 265–273. doi: 10.1016/j.canlet.2007.03.008

Graziano, S. L., Gamble, G. P., Newman, N. B., Abbott, L. Z., Rooney, M., Mookherjee, S., et al. (1999). Prognostic significance of K-ras codon 12 mutations in patients with resected stage I and II non-small-cell lung cancer. *J. Clin. Oncol.* 17 (2), 668–675. doi: 10.1200/JCO.1999.17.2.668

Halmos, B., Huettner, C. S., Kocher, O., Ferenczi, K., Karp, D. D., and Tenen, D. G. (2002). Down-regulation and antiproliferative role of C/EBPalpha in lung cancer. *Cancer Res.* 62 (2), 528–534.

Hua, F., Shang, S., Yang, Y. W., Zhang, H. Z., Xu, T. L., Yu, J. J., et al. (2019). TRIB3 Interacts with beta-Catenin and TCF4 to increase stem cell features of colorectal cancer stem cells and tumorigenesis. *Gastroenterology* 156 (3), 708–721.e715. doi: 10.1053/j.gastro.2018.10.031

Huang, T., and Cai, Y. D. (2013). An information-theoretic machine learning approach to expression QTL analysis. *PloS One* 8 (6), e67899. doi: 10.1371/journal.pone.0067899

Huang, T., Tu, K., Shyr, Y., Wei, C. C., Xie, L., and Li, Y. X. (2008). The prediction of interferon treatment effects based on time series microarray gene expression profiles. *J. Trans. Med.* 6 (1), 44. doi: 10.1186/1479-5876-6-44

Huang, T., Chen, L., Liu, X. J., and Cai, Y. D. (2011). Predicting triplet of transcription factor - mediating enzyme - target gene by functional profiles. *Neurocomputing* 74 (17), 3677–3681. doi: 10.1016/j.neucom.2011.07.019

Jemal, A., Siegel, R., Ward, E., Murray, T., Xu, J., Smigal, C., et al. (2006). Cancer Statistics, 2006. *CA: A Cancer J. Clin.* 56 (2), 106–130. doi: 10.3322/canjclin.56.2.106

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61 (2), 69–90. doi: 10.3322/caac.20107

Jiang, Y., Pan, X., Zhang, Y., Huang, T., and Gao, Y. (2019). Gene expression difference between primary and metastatic renal cell carcinoma using patient-derived xenografts. *IEEE Access* 7, 142586–142594. doi: 10.1109/ACCESS.2019.2944132

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

Labbe, C., Cabanero, M., Korpanty, G. J., Tomasini, P., Doherty, M. K., Mascaux, C., et al. (2017). Prognostic and predictive effects of TP53 co-mutation in patients with EGFR-mutated non-small cell lung cancer (NSCLC). *Lung Cancer* 111, 23–29. doi: 10.1016/j.lungcan.2017.06.014

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 (5795), 1929–1935. doi: 10.1126/science.1132939

Laurent, C., Svrcek, M., Flejou, J. F., Chenard, M. P., Duclos, B., Freund, J. N., et al. (2011). Immunohistochemical expression of CDX2, beta-catenin, and TP53 in inflammatory bowel disease-associated colorectal cancer. *Inflammation Bowel Dis.* 17 (1), 232–240. doi: 10.1002/ibd.21451

Lei, C., Wei-Ming, Z., Yu-Dong, C., and Tao, H. (2013). Prediction of metabolic pathway using graph property, chemical functional group and chemical structural set. *Curr. Bioinf.* 8 (2), 200–207. doi: 10.2174/15748936 11308020008

Li, J., and Huang, T. (2018). Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies. *Biochim. Biophys. Acta* 1864 (6 Pt B), 2241–2246. doi: 10.1016/j.bbadis.2017.10.036

Li, H., Liang, X., Qin, X., Cai, S., and Yu, S. (2015). Association of matrix metalloproteinase family gene polymorphisms with lung cancer risk: logistic regression and generalized odds of published data. *Sci. Rep.* 5, 10056. doi: 10.1038/srep10056

Li, J., Lan, C.-N., Kong, Y., Feng, S.-S., and Huang, T. (2018). Identification and analysis of blood gene expression signature for osteoarthritis with advanced feature selection methods. *Front. Genet.* 9, 246. doi: 10.3389/fgene.2018.00246

Li, J., Lu, L., Zhang, Y.-H., Xu, Y., Liu, M., Feng, K., et al. (2019a). Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther*. doi: 10.1038/s41417-019-0105-y

Li, J., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019b). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell Biochem.* 120 (1), 405–416. doi: 10.1002/jcb.27395

Liu, P., Morrison, C., Wang, L., Xiong, D., Vedell, P., Cui, P., et al. (2012). Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* 33 (7), 1270–1276. doi: 10.1093/carcin/bgs148

Liu, C., Cui, P., and Huang, T. (2017). Identification of cell cycle-regulated genes by convolutional neural network. *Comb. Chem. High Throughput Screen* 20 (7), 603–611. doi: 10.2174/1386207320666170417144937

Lo, F. Y., Chen, H. T., Cheng, H. C., Hsu, H. S., and Wang, Y. C. (2012). Overexpression of PAFAH1B1 is associated with tumor metastasis and poor survival in non-small cell lung cancer. *Lung Cancer* 77 (3), 585–592. doi: 10.1016/j.lungcan.2012.05.105

Lu, H., Yu, Z., Liu, S., Cui, L., Chen, X., and Yao, R. (2015). CUGBP1 promotes cell proliferation and suppresses apoptosis via down-regulating C/EBPalpha in human non-small cell lung cancers. *Med. Oncol.* 32 (3), 82. doi: 10.1007/s12032-015-0544-8

Luo, J., Shi, K., Yin, S. Y., Tang, R. X., Chen, W. J., Huang, L. Z., et al. (2018). Clinical value of miR-182-5p in lung squamous cell carcinoma: a study combining data from TCGA, GEO, and RT-qPCR validation. *World J. Surg. Oncol.* 16 (1), 76. doi: 10.1186/s12957-018-1378-6

Lv, M., and Wang, L. (2015). Comprehensive analysis of genes, pathways, and TFs in nonsmoking Taiwan females with lung cancer. *Exp. Lung Res.* 41 (2), 74–83. doi: 10.3109/01902148.2014.971472

Mao, L., Hruban, H. R., Boyle, J. O., Ms, T., and Sidransky, D. (1994). Detection of oncogene mutations in sputum precedes diagnosis of lung cancer. 54 (7), 1634–1637.

Marwitz, S., Depner, S., Dvornikov, D., Merkle, R., Szczygiel, M., Muller-Decker, K., et al. (2016). Downregulation of the TGFbeta pseudoreceptor bambi in non-small cell lung cancer enhances TGFbeta signaling and invasion. *Cancer Res.* 76 (13), 3785–3801. doi: 10.1158/0008-5472.CAN-15-1326

Mills, N. E., Fishman, C. L., Scholes, J., Anderson, S. E., Rom, W. N., and Jacobson, D. R. (1995). Detection of K-ras oncogene mutations in bronchoalveolar lavage fluid for lung cancer diagnosis. *JNCI: J. Natl. Cancer Institute* 87 (14), 1056–1060. doi: 10.1093/jnci/87.14.1056

Morgensztern, D., Ng, S. H., Gao, F., and Govindan, R. (2010). Trends in stage distribution for patients with non-small cell lung cancer: a national cancer database survey. *J. Thoracic Oncol.* 5 (1), 29–33. doi: 10.1097/JTO.0b013e3181c5920c

Nakamoto, M., Teramoto, H., Matsumoto, S., Igishi, T., and Shimizu, E. (2001). K-ras and rho A mutations in malignant pleural effusion. *Int. J. Oncol.* 19 (5), 971–976. doi: 10.3892/ijo.19.5.971

Nishimura, T., Nakata, A., Chen, X., Nishi, K., Meguro-Horike, M., Sasaki, S., et al. (2019). Cancer stem-like properties and gefitinib resistance are dependent on purine synthetic metabolism mediated by the mitochondrial enzyme MTHFD2. *Oncogene* 38 (14), 2464–2481. doi: 10.1038/s41388-018-0589-1

Okudela, K., Mitsui, H., Woo, T., Arai, H., Suzuki, T., Matsumura, M., et al. (2016). Alterations in cathepsin L expression in lung cancers. *Pathol. Int.* 66 (7), 386–392. doi: 10.1111/pin.12424

Ortega, A., and Mas-Oliva, J. (1986). Direct regulatory effect of cholesterol on the calmodulin stimulated calcium pump of cardiac sarcolemma. *Biochem. Biophys. Res. Commun.* 139 (3), 868–874. doi: 10.1016/S0006-291X(86)80258-3

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., Chen, L., et al. (2018). Identifying Patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes (Basel)* 9 (4). doi: 10.3390/genes9040208

Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019a). Analysis of Expression Pattern of snoRNAs in Different Cancer Types with Machine Learning Algorithms. *Int. J. Mol. Sci.* 20 (9). doi: 10.3390/ijms20092185

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., et al. (2019b). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294 (1), 95–110. doi: 10.1007/s00438-018-1488-4

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27 (8), 1226–1238. doi: 10.1109/TPAMI.2005.159

Sandler, A., Gray, R., Perry, M. C., Brahmer, J., Schiller, J. H., Dowlati, A., et al. (2006). Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N. Engl. J. Med.* 355 (24), 2542–2550. doi: 10.1056/NEJMoa061884

Scagliotti, G. V., Parikh, P., von Pawel, J., Biesma, B., Vansteenkiste, J., Manegold, C., et al. (2008). Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J. Clin. Oncol.* 26 (21), 3543–3551. doi: 10.1200/JCO.2007.15.0375

Slattery, M. L., and Lundgreen, A. (2014). The influence of the CHIEF pathway on colorectal cancer-specific mortality. *PloS One* 9 (12), e116169. doi: 10.1371/journal.pone.0116169

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 171 (6), 1437–1452.e1417. doi: 10.1016/j.cell.2017.10.049

Sutherland, L. C., Wang, K., and Robinson, A. G. (2010). RBM5 as a putative tumor suppressor gene for lung cancer. *J. Thorac. Oncol.* 5 (3), 294–298. doi: 10.1097/JTO.0b013e3181c6e330

Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., et al. (2012). Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim. Biophys. Acta* 1824 (1), 68–88. doi: 10.1016/j.bbapap.2011.10.002

Tyagi, N., Marimuthu, S., Bhardwaj, A., Deshmukh, S. K., Srivastava, S. K., Singh, A. P., et al. (2016). p-21 activated kinase 4 (PAK4) maintains stem cell-like phenotypes in pancreatic cancer cells through activation of STAT3 signaling. *Cancer Lett.* 370 (2), 260–267. doi: 10.1016/j.canlet.2015.10.028

Wang, S.-B., and Huang, T.J.M.B.R. (2019a). The early detection of asthma based on blood gene expression 46, 1, 217–223. doi: 10.1007/s11033-018-4463-6

Wang, S. B., and Huang, T. (2019b). The early detection of asthma based on blood gene expression. *Mol. Biol. Rep.* 46 (1), 217–223. doi: 10.1007/s11033-018-4463-6

Wang, T. L., Song, Y. Q., Ren, Y. W., Zhou, B. S., Wang, H. T., Bai, L., et al. (2016a). Dual Specificity Phosphatase 6 (DUSP6) Polymorphism Predicts Prognosis of

Inoperable Non-Small Cell Lung Cancer after Chemoradiotherapy. *Clin. Lab.* 62 (3), 301–310. doi: 10.7754/Clin.Lab.2015.150432

Wang, X., Lu, Y., Feng, W., Chen, Q., Guo, H., Sun, X., et al. (2016b). A two kinase-gene signature model using CDK2 and PAK4 expression predicts poor outcome in non-small cell lung cancers. *Neoplasma* 63 (2), 322–329. doi: 10.4149/220_150817N448

Wang, S., Zhang, Y. H., Zhang, N., Chen, L., Huang, T., and Cai, Y. D. (2017a). Recognizing and predicting thioether bridges formed by lanthionine and beta-methyllanthionine in lantibiotics using a random forest approach with feature selection. *Comb. Chem. High Throughput Screen* 20 (7), 582–593. doi: 10.2174/1386207320666170310115754

Wang, X., Li, M., Hu, M., Wei, P., and Zhu, W. (2017b). BAMBI overexpression together with beta-sitosterol ameliorates NSCLC via inhibiting autophagy and inactivating TGF-beta/Smad2/3 pathway. *Oncol. Rep.* 37 (5), 3046–3054. doi: 10.3892/or.2017.5508

Westcott, P. M., and To, M. D. (2013). The genetics and biology of KRAS in lung cancer. *Chin. J. Cancer* 32 (2), 63–70. doi: 10.5732/cjc.012.10098

Yan, X., Yu-Hang, Z., JiaRui, L., Xiaoyong, P., Tao, H., and Yu-Dong, C. (2019). New computational tool based on machine-learning algorithms for the identification of rhinovirus infection-related genes. *Combinatorial Chem. High Throughput Screening* 22, 1–1. doi: 10.2174/1386207322666191129114741

Yang, J., Chen, L., Kong, X., Huang, T., and Cai, Y. D. (2014). Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. *PloS One* 9 (9), e107202. doi: 10.1371/journal.pone.0107202

Yin, H., Wang, Y., Chen, W., Zhong, S., Liu, Z., and Zhao, J. (2016). Drug-resistant CXCR4-positive cells have the molecular characteristics of EMT in NSCLC. *Gene* 594 (1), 23–29. doi: 10.1016/j.gene.2016.08.043

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860 (11 Pt B), 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zhang, K., Wang, J., Tong, T. R., Wu, X., Nelson, R., Yuan, Y. C., et al. (2017). Loss of H2B monoubiquitination is associated with poor-differentiation and enhanced malignancy of lung adenocarcinoma. *Int. J. Cancer* 141 (4), 766–777. doi: 10.1002/ijc.30769

Zhang, X., Zhong, N., Li, X., and Chen, M. B. (2019). TRIB3 promotes lung cancer progression by activating beta-catenin signaling. *Eur. J. Pharmacol.* 863, 172697. doi: 10.1016/j.ejphar.2019.172697

Zhao, T. H., Jiang, M., Huang, T., Li, B. Q., Zhang, N., Li, H. P., et al. (2013). A novel method of predicting protein disordered regions based on sequence features. *BioMed. Res. Int.* 2013, 414327. doi: 10.1155/2013/414327

Zhao, N., Liu, Y., Chang, Z., Li, K., Zhang, R., Zhou, Y., et al. (2015). Identification of biomarker and co-regulatory motifs in lung adenocarcinoma based on differential interactions. *PloS One* 10 (9), e0139165. doi: 10.1371/journal.pone.0139165

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: 04 February 2020
doi: 10.3389/fgene.2020.00005

Pathogenic Gene Prediction Algorithm Based on Heterogeneous Information Fusion

Chunyu Wang[1]*, Jie Zhang[1], Xueping Wang[1], Ke Han[2] and Maozu Guo[3,4]*

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] School of Computer and Information Engineering, Harbin University of Commerce, Harbin, China, [3] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China, [4] Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing, China

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Xiucai Ye,
University of Tsukuba, Japan
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

***Correspondence:**
Chunyu Wang
chunyu@hit.edu.cn
Maozu Guo
guomaozu@bucea.edu.cn

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

Received: *26 November 2019*
Accepted: *06 January 2020*
Published: *04 February 2020*

Citation:
Wang C, Zhang J, Wang X, Han K and Guo M (2020) Pathogenic Gene Prediction Algorithm Based on Heterogeneous Information Fusion. Front. Genet. 11:5. doi: 10.3389/fgene.2020.00005

Complex diseases seriously affect people's physical and mental health. The discovery of disease-causing genes has become a target of research. With the emergence of bioinformatics and the rapid development of biotechnology, to overcome the inherent difficulties of the long experimental period and high cost of traditional biomedical methods, researchers have proposed many gene prioritization algorithms that use a large amount of biological data to mine pathogenic genes. However, because the currently known gene–disease association matrix is still very sparse and lacks evidence that genes and diseases are unrelated, there are limits to the predictive performance of gene prioritization algorithms. Based on the hypothesis that functionally related gene mutations may lead to similar disease phenotypes, this paper proposes a PU induction matrix completion algorithm based on heterogeneous information fusion (PUIMCHIF) to predict candidate genes involved in the pathogenicity of human diseases. On the one hand, PUIMCHIF uses different compact feature learning methods to extract features of genes and diseases from multiple data sources, making up for the lack of sparse data. On the other hand, based on the prior knowledge that most of the unknown gene–disease associations are unrelated, we use the PU-Learning strategy to treat the unknown unlabeled data as negative examples for biased learning. The experimental results of the PUIMCHIF algorithm regarding the three indexes of precision, recall, and mean percentile ranking (MPR) were significantly better than those of other algorithms. In the top 100 global prediction analysis of multiple genes and multiple diseases, the probability of recovering true gene associations using PUIMCHIF reached 50% and the MPR value was 10.94%. The PUIMCHIF algorithm has higher priority than those from other methods, such as IMC and CATAPULT.

Keywords: pathogenic gene prediction, induction matrix completion, compact feature learning, PU-Learning, mean percentile ranking

INTRODUCTION

The discovery of disease-causing genes plays an important role in understanding the causes of diseases, clinically diagnosing diseases, and achieving early prevention and treatment (Cheng et al., 2016; Zeng et al., 2017; Cheng et al., 2019). It is also an important goal of human genome research, with great scientific and social significance. Prioritization of potentially pathogenic genes is an important step in the discovery of disease-causing genes and obtaining an understanding of genetic diseases.

Early studies of gene–disease associations were based on clinical and biological experiments, which are expensive and time-consuming. Owing to the inherent difficulties and delays in the study of human genetic diseases, there are very few known identified gene–disease links in public databases, such as the widely used Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2015) and Genetic Association Database (Becker et al., 2004). Because of the specificity of the study of disease-causing genes, we do not know the genes that are not related to a particular disease. We only know the few genes that have been proven to be related to it. Against this background, with the emergence of bioinformatics, researchers have begun to focus on and study genetic disease prioritization algorithms, and use computer technology to mine candidate pathogenic genes from massive data (Liu et al., 2020; Wang et al., 2018; Zeng et al., 2018; Zhang et al., 2019; Zeng et al., 2019; Pan et al., 2019). The selected genes are more likely to be related to diseases, and gene sorting algorithms with better predictive performance would be more helpful to conduct targeted biological experiments and understand pathogenic mechanisms.

Early gene sorting algorithms based on network similarity focused on local information in the gene–disease network, namely, nodes adjacent to gene or disease nodes; an example of these is the molecular triangulation method (Krauthammer et al., 2004). It has been found that the global topology of a network can improve the performance in predicting disease-causing genes (Pan et al., 2019; Chen et al., 2019). Kohler et al. (Kohler et al., 2008) used the random walk (RWR) algorithm to analyze candidate disease-causing genes, which further improved the predictive performance.

Complex biological systems cannot always meet the needs of analysis with single network data (Chen et al., 2019). The continuous growth of biological data, such as high-throughput sequencing, also brings opportunities to study new predictive methods. The more commonly used databases include the gene expression database GEO (Barrett et al., 2007), the cancer gene information TCGA database (Cancer Genome Atlas Research et al., 2013), the protein interaction network database STRING (Szklarczyk et al., 2017), the Gene Ontology (GO) database (Ashburner et al., 2000), and Disease Ontology (DO) (Schriml et al., 2012). Recently, there has been increasing interest in studying gene sorting algorithms and starting to integrate a large amount of biological data and analyze heterogeneous networks (Gomez-Cabrero et al., 2014; Jiang, 2015; Zhang et al., 2019; Deng et al., 2019). In 2008, the CIPHER algorithm (Wu et al., 2008) was proposed by Wu et al., which combines protein interaction and disease-like networks but only considers local information in the network and lacks global topology. In 2010, Vanunu et al. (Vanunu et al., 2010) proposed the PRINCE algorithm, based on the idea of global network information and network dissemination. In the same year, Yongjin Li et al. (Li and Patra, 2010) proposed the restarted random walk algorithm (RWRH) that fused a gene similarity network, a disease phenotypic similarity network, and a large heterogeneous network composed of a disease phenotype–gene relationship network. In addition, Singh-Blom et al. (Singh-Blom et al., 2013) further improved the predictive performance in 2013 using the Katz method commonly used in the field of social networks for the task of predicting gene–disease relationships.

With the rapid development of machine learning and artificial intelligence in recent years, new algorithms based on machine learning have been applied to predict candidate pathogenic genes; they have shown good predictive performance (Zou et al., 2018; Peng et al., 2018; Liao et al., 2018; Zhang et al., 2018; Xiong et al., 2018; He et al., 2018; Cheng et al., 2018; Cheng et al., 2018; Zeng et al., 2019; Ding et al., 2019; Liu, 2019; Liu et al., 2019a; Zhu et al., 2019). In 2011, Mordelet et al. (Mordelet and Vert, 2011) considered the problem of genetic prediction as a supervised machine learning problem and proposed the ProDiGe method. Moreover, in 2013, Singh-Blom et al. (Kohler et al., 2008) proposed the supervised machine-learning method CATAPULT using a variety of data sources. Then, Natarajan et al. (Natarajan and Dhillon, 2014) applied the inductive matrix completion algorithm (IMC) in the recommendation system to predict pathogenic genes. This algorithm can not only predict existing genes and diseases but also predict new genes and diseases that have not previously been shown to be related. To compensate for the impact of a data sparseness and the PU problem, the PUIMCHIF algorithm is proposed in this paper. Specifically, on the basis of the original IMC algorithm, the main innovations and contributions of this paper can be summarized as follows: (1) owing to the sparsity of gene–disease association data, we used a variety of data sources to construct the characteristics of genes and diseases, and added a STRING data set for the compact feature learning of genes, which contained the physical relationships and other interactions that were not in the original data set. (2) For the gene–gene network and the disease–disease network (Li et al., 2019), we used the RWR method to obtain the diffusion state of each node in the network under a steady state in accordance with the network topology, used diffusion component analysis (DCA) to reduce the dimensions of the data, and finally obtained the network characteristics of genes or diseases. One advantage of this approach is the ability to analyze both HumanNet and STRING networks. (3) Self-encoders in machine learning can learn efficient representations of data for dimensionality reduction. Combined with the characteristics of biological data, the work described in this paper used denoising self-encoding to reduce the dimensionality of high-dimensional data features of genes and diseases. (4) Considering the sparse disease–gene association data and the prior knowledge that most unknown associations are negative cases, we adopted the PU-Learning strategy to treat unlabeled data as negative cases for biased learning, so as to replace the IMC method involving learning for only positive cases. (5) To verify the effectiveness of the PUIMCHIF method proposed in this paper, we used two commonly used evaluation indexes, Precision and Recall. On this basis, we added the MPR index of mean percentile ranking to further analyze the experimental results comprehensively.

INTRODUCTION TO METHODS

We are interest in kinds of associations between the genes and diseases, but only part of them are known. So we want to make a prediction about the unknown pare from the known ones. As shown in **Figure 1**, our goal was to predict these unknown associations based on the constructed low-dimensional characteristics of the genes and diseases, and some known items in the gene–disease association matrix *P*, that is, to predict candidate genes potentially involved in the pathogenicity of the disease.

First, we constructed a low-dimensional eigenvector of genes and diseases from different biological sources (compact feature learning). We proposed different methods for learning compact features based on different forms of data. For the network data of genes and diseases, the random walk with restart algorithm (RWR) was first used to extract the diffusion state of each node in the network, and then DCA was used for dimensionality reduction to obtain the similarity of each gene (or disease) node in the heterogeneous network encoded by low-dimensional feature vectors. This is because genes (or diseases) with similar topological properties in the network are more likely to be functionally related.

Second, for common feature matrix data, to reduce the influence of high noise and data loss of biological data, we used denoising autoencoder (DAE) to reduce the dimensions of features.

Next, we applied the partial inductive matrix completion algorithm to predict the relationship between genes and diseases by combining the characteristics of multiple diseases and genes. One of the main advantages of this method is that it is generalized and can be applied to diseases that are not present during training, which cannot be predicted by traditional matrix completion methods. This allows us to take advantage of previous knowledge of known gene–disease interactions to predict unknown gene–disease interactions. Because we added an unbiased learning scheme for the unknown association relationship as a negative example, we finally adopted the PUIMC method for disease-causing gene prediction. The details of the PUIMCHIF algorithm are described below.

Compact Feature Learning

In machine learning, the data are more important than the algorithm because the generalization of machine learning algorithm is about the ability from known data to the unknown data. Therefore, when we choose the prediction method based on machine learning to predict the disease-causing gene. First, we need to use high-quality data. Second, we need to conduct feature processing on the data to obtain more favorable data features for the prediction task.

We integrated a variety of biological data to extract characteristics of genes or diseases. Moreover, our goal was to obtain a low-dimensional effective data feature matrix, where one row of the feature matrix refers to a gene or disease, and the columns of the matrix represent different characteristics. The different compact feature learning methods that we used are described below.

RWR

Closely linked or functionally similar genes are more likely to cause the same or similar diseases. Random walk provides an effective framework for exploring relationships in networks.

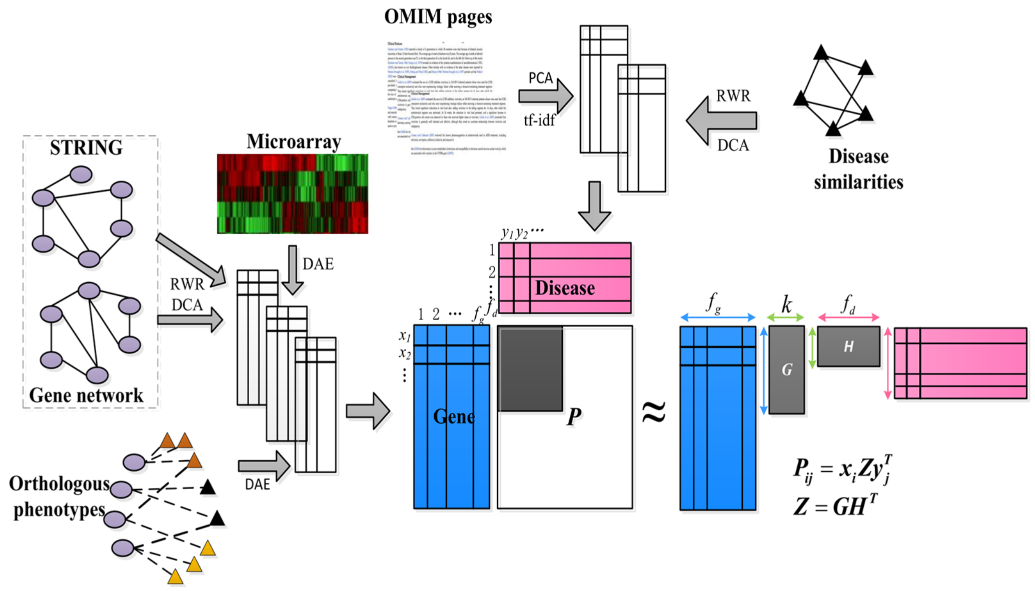


FIGURE 1 | Schematic diagram of PUIMCHIF model framework.

Random walk with restart is referred to as RWR, which is a network diffusion algorithm widely used in the analysis of complex biological network data (Navlakha and Kingsford, 2010; Cao et al., 2014). Different from the traditional random walk method, each iteration of RWR introduces a predefined restart probability at the initial node, which can consider both local and global topological connection patterns within the network and take full advantage of direct or indirect relationships between nodes.

Here, matrix A and B are defined. Matrix A represents the weighted adjacency matrix of the interaction network of genes (or diseases). And in matrix B as shown in equation (1), each element B_{ij} describes the probability of transition from node i to node j. s_i^t represents an n-dimensional distribution vector, and each element stores the probability that a node is accessed after iterating t times from node i during the random walk. The formula for calculating RWR is shown in equation (2).

$$B_{ij} = \frac{A_{ij}}{\Sigma_{j'} A_{ij'}} \tag{1}$$

$$s_i^{t+1} = (1 - p_r) s_i^t B + p_r \delta_i \tag{2}$$

In equation (2), δ_i represents an n-dimensional standard basis vector and $\delta_i(i) = 1$, $\delta_i\ (j) = 0$, for $\forall j \neq i$. And p_r is a predefined restart probability that controls the relative influence of local structure and global structure in the diffusion process. With a higher value, more attention is paid to the local structure in the network.

For a node in the iterative process, we can obtain a stable distribution s_i^∞, so we define s_i as the "diffusion state" of node i, that is $s_i = s_i^\infty$. The jth element s_{ij} of s_i represents the probability that the RWR starts from node i and ends at node j in equilibrium. When two nodes have similar diffusion states, it generally means that they are more similar than other nodes in the network and may have similar functions. This discovery provides a basis for predicting unknown gene–disease associations.

Diffusion Component Analysis

Although the diffuse states generated by the above RWR process represent the underlying topological environment and intrinsic connectivity spectrum of each gene or disease node in the network, they may not be completely accurate due to the low-quality and high-dimensional nature of biological data. For example, a small number of missing or false interactions in the network can significantly affect the outcome of the diffusion process (Kim and Leskovec, 2011). It is often inconvenient to directly use high-dimensional diffusion states as topological features in prediction tasks.

To solve this problem, we used a dimensionality reduction method called DCA to reduce the dimensions of the feature space and obtain important topological features from the diffusion state. In addition, for multi-omics networks, DCA also performs very well. The key idea of DCA is to obtain an informative but low-dimensional vector representation. Similar to principal component analysis (PCA), which seeks the inherent low-dimensional linear structure of data to best interpret variances, DCA learns the low-dimensional vector representation of all nodes to best interpret their patterns of connection in heterogeneous networks. We will briefly describe the DCA framework below.

To achieve the purposes of noise reduction and dimensionality reduction, DCA uses the polynomial logic model represented by a low-dimensional vector to approximate the obtained diffusion state distribution, and it has far fewer dimensions than the original n-dimensional vector representing the diffusion state. Specifically, the probability of assigning node i to node j in the diffusion state is modeled as:

$$\hat{s}_{ij} = \frac{\exp\{x_i^T w_j\}}{\sum_{j'} \exp\{x_i^T w_{j'}\}} \tag{3}$$

In equation (3), x_i, $w_i \in \mathbb{R}^d$, $d \ll n$. We take w_i as the context feature and x_i as the node feature of node i, both of which describe the topological properties of the network. If x_i and w_i point in similar directions, we obtain a larger inner product. This means that node j may be frequently visited in a random walk starting from node i. DCA uses the obtained diffusion state S={ $s_1,\cdots,\ s_n$ } as input to optimize w and x of all nodes. The optimization method uses KL divergence, as shown in equation (4).

$$\min_{w,\ x} C(s,\ \hat{s}) = \min_{w,x} \frac{1}{n} \sum_{i=1}^{n} D_{KL}(s_i || \hat{s}_i) \tag{4}$$

$D_{KL}(\cdot||\cdot)$ is the KL divergence between the two distributions. We use w and x to represent this formula according to the definition of KL invergence and $\hat{s}$.

$$C(s, \hat{s}) = \tfrac{1}{n} \sum_{i=1}^{n} \left[H(s_i) - \sum_{j=1}^{n} s_{ij} \left(w_i^T x_j - \log\left(\sum_{j'=1}^{n} \exp\{w_i^T, x_{j'}\} \right) \right) \right] \tag{5}$$

In equation (5), $H(\cdot)$ represents entropy. The objective function can find the low-dimensional vector representation of w and x using the standard quasi-Newton L-BFGS method. Although the obtained low-dimensional vector can effectively capture the network structure, we found that this optimization method is time-consuming.

To make the DCA framework more suitable for large biological networks, we use a more efficient method, clusDCA (Wang et al., 2015), which is based on matrix factorization, to decompose the diffusion states and obtain their low-dimensional vector representations. According to the definition, the following formula can be obtained:

$$\log \hat{s}_{ij} = x_i^T w_j - \log \sum_{j'} \exp\{x_i^T w_{j'}\} \tag{6}$$

The first term corresponds to the low-dimensional approximation of $\hat{s}_{ij}$. The second term is a normalization factor, ensuring that $\hat{s}_i$ is a well-defined distribution. By removing the second term, we relax the constraint that the elements in $\hat{s}_{ij}$ must add up to 1. Although the obtained low-dimensional approximation of the diffusion state is no longer a strictly valid probability distribution, it is found that these approximations are very close to the true distribution, and the effects of relaxation are negligible. Therefore, it can be simplified as:

$$\log \hat{s}_{ij} = x_i^T w_j \,. \tag{7}$$

In addition, we use the sum of squared errors as the objective function, instead of minimizing the relative entropy between the original diffusion state and the approximate diffusion state.

$$\min_{w,\, x} C(s,\ \hat{s}) = \min_{w,x} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(w_i^T x_j - \log s_{ij}\right)^2 \tag{8}$$

The obtained objective function can be optimized by singular value decomposition (SVD). To avoid taking the logarithm of 0, we add a small positive number $\frac{1}{n}$ to s_{ij}. The calculation formula of the logarithm diffusion state matrix L is as follows:

$$L = \log(S + Q) - \log(Q)\,. \tag{9}$$

In equation (9), $S \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n}$ and $Q_{ij} = \frac{1}{n}$, for $\forall i, j$. Using the singular value decomposition method, we decompose L into three matrices:

$$L = U\Sigma V^T \tag{10}$$

In equation (10), $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{n\times n}$, $\Sigma \in \mathbb{R}^{n\times n}$ and Σ is a diagonal singular value matrix. To obtain the low-dimensional vectors w_j and x_i in d dimensions, we simply select the first d singular vectors U_d, V_d, and Σ_d. Each row of matrix $X = [x_1,\ \ldots,\ x_n]^T$ represents the low-dimensional eigenvector corresponding to each node in the network. In matrix $W = [w_1,\ \ldots,\ w_n]^T$, each row represents the corresponding vector of the context feature. The formulas for calculating X and W are as follows:

$$X = U_d \Sigma_d^{1/2}, \quad W = V_d \Sigma_d^{1/2}\,. \tag{11}$$

Denoising Autoencoder

Autoencoder is an unsupervised neural network model. It learns the implicit features of input data, which is called "coding." At the same time, the original input data can be reconstructed with the learned new features, which is called "decoding." Intuitively, autoencoder can be used for reducing feature dimensionality, like principal components analysis (PCA), but with stronger performance than PCA because the neural network model can extract more effective new features.

The denoising autoencoder adds noise to the input x to obtain $\tilde{x}$, and after training, it obtains a noiseless output z, as shown in **Figure 2**.

This prevents the autoencoder from simply copying the input to the output, so as to extract useful patterns in the data and improve the weight robustness. Noise can be either pure gaussian noise added to the input or randomly discarding a feature at input layer, similar to dropout. The specific equation for calculating z is as follows:

$$\begin{aligned} y &= f(\tilde{x}W_1 + b_1) \\ z &= g(yW_2 + b_2) \end{aligned} \tag{12}$$

In addition, network parameters are trained to minimize reconstruction errors, namely:

$$\min L_H(x, z) = \min ||x - z_p\,. \tag{13}$$

Pathogenic Gene Prediction Method

Standard Inductive Matrix Completion

In the gene–disease association matrix $P \in \mathbb{R}^{N_g \times N_d}$, each row represents a gene ID and the number of genes is N_g. Each column represents a disease phenotype and the number of diseases is N_d. If $P_{ij} = 1$, this means that gene i is related to disease j, and $P_{ij} = 0$ means that the relationship between gene i and disease j is uncertain. Based on the most successful and deeply studied matrix completion method in the recommender systems, the IMC algorithm was used to complete the task of learning gene–disease associations. The advantage of this is that this method is inductive, and it can achieve the prediction of new genes or diseases that have rarely been studied.

IMC assumes that the association matrix has a low rank, with the goal of recovering Z using the observed values of P and the eigenvectors of genetic diseases, as shown in **Figure 3**.

The eigenvector matrix of N_g genes is represented by $X \in \mathbb{R}^{N_g \times f_g}$, and the eigenvector of gene i is represented by $x_i \in \mathbb{R}^{f_g}$.

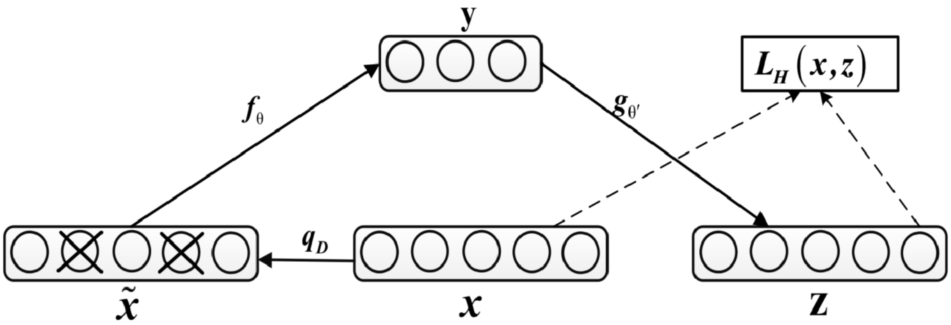


FIGURE 2 | Diagram of Denoising Autoencoder.

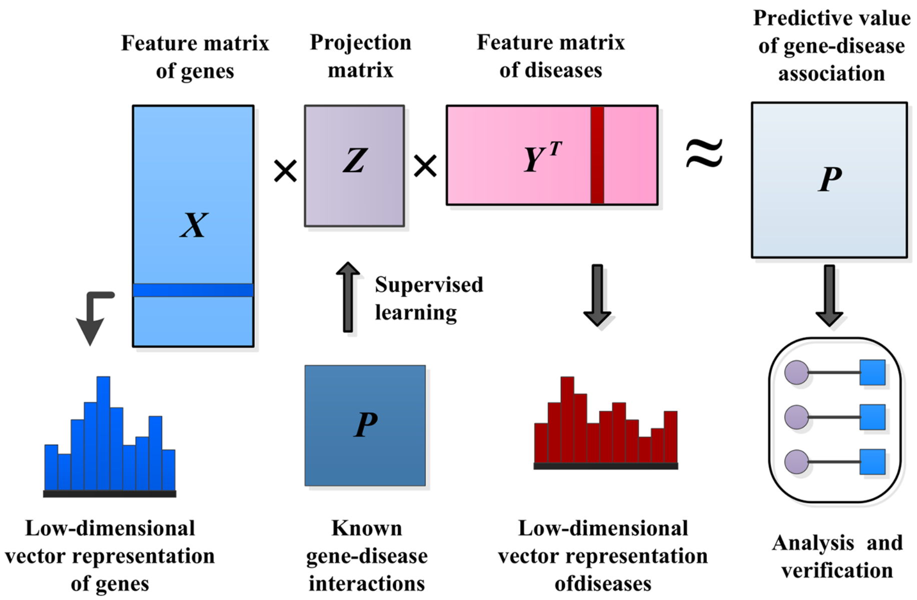

Feature matrix of genes
Projection matrix
Feature matrix of diseases
Predictive value of gene-disease association
X
×
Z
×
Y^T
≈
P
Supervised learning
P
Low-dimensional vector representation of genes
Known gene-disease interactions
Low-dimensional vector representation ofdiseases
Analysis and verification

FIGURE 3 | Methods of predicting pathogenic genes.

Similarly, $Y \in \mathbb{R}^{N_d \times f_d}$ is used to represent the eigenvector matrix of N_d diseases, and $y_i \in \mathbb{R}^{f_d}$ is used to represent the eigenvector of disease j. The inductive matrix completion problem is to recover a low-rank matrix Z by using the known association Ω^+ from the gene–disease association matrix P. We established a bilinear function to learn the projection matrix Z between the gene space and the disease space to predict the interaction between unknown genes and diseases. We modeled the matrix P as $XZY^T \approx P$. Then, we used the following formula to measure the probability of pairwise interaction score between gene i and disease j, and the higher the score(i, j) value, the more likely gene i and disease j interact.

$$score(i,j) = x_i Z y_j^T \tag{14}$$

There is usually a significant correlation between spatially close eigenvectors of genes or diseases, which can greatly reduce the number of effective parameters needed to model gene–disease interactions in Z. To consider this problem, we applied a low-rank constraint on Z and learned only a few potential factors. Let $Z = GH^T$, where $G \in \mathbb{R}^{f_g \times k}$, $H \in \mathbb{R}^{f_d \times k}$, and k is small. This low-rank constraint not only alleviates the overfitting problem, but also facilitates the process of optimizing the calculation (Wang et al., 2015). The optimization problem of low-rank constraint is NP-hard on the original matrix Z. One standard method of relaxing the low-rank constraint is to minimize the trace norm, that is, the sum of the singular values. Minimizing the trace norm of $Z = GH^T$ is equivalent to minimizing $\frac{1}{2}(\|G\|_F^2 + \|H\|_F^2)$. The decomposition of Z into G and H solves the following optimization problems by alternating minimization. A common choice for the loss function ℓ is the square loss function. λ is the regularization parameter.

$$\min \sum_{(i,j)\in\Omega^+} \ell\left(P_{ij}, x_i^T GH^T y_i\right) + \frac{\lambda}{2}\left(\|G\|_F^2 + \|H\|_F^2\right) \tag{15}$$

Improved Inductive Matrix Completion

To optimize the objective function, we introduce the idea of PU-Learning. Although we predicted positive examples from unknown relationships, that is, candidate disease-causing genes, it was undeniable that these unknown genes-disease pairs may be unrelated. Therefore, unknown association relationship information was added to the learning process as a negative example, and the objective function was as follows:

$$\min \sum_{(i,j)\in\Omega^+} \ell\left(P_{ij}, x_i^T GH^T y_i\right) + \alpha \sum_{(i,j)\in\Omega^-} \ell\left(P_{ij}, x_i^T GH^T y_i\right) + \frac{\lambda}{2}\left(\|G\|_F^2 + \|H\|_F^2\right) \tag{16}$$

We represent the unknown association in the gene–disease association matrix P as Ω^-. The key parameter $\alpha < 1$ because the penalty weight of the known relationship must be greater than the unknown relationship. Finally, equation (14) was still used to calculate the interaction score between gene i and disease j. The scores are sorted in descending order, and the first k genes were selected as candidate pathogenic genes for the corresponding disease.

DATA SETS AND FEATURES

The data sets used in this paper can be divided into three categories: gene–disease association data, gene characteristic data, and disease characteristic data.

Gene–Disease Associations

The known gene–disease association data that we used were from the OMIM database, which contained 12,331 genes, 3,209 diseases, and 3,954 known gene–disease associations (the total number of nonzero elements in the gene–disease association matrix). It can be seen that the data in the incidence matrix are very sparse, with more than 90% of the columns having only one nonzero item and 70% of the rows having no nonzero elements.

Gene Characteristics

Gene characteristics were obtained by processing four different data sources through compact feature learning (*Compact Feature Learning*). The first source of gene characteristics was gene microarray data, which contained 8,755 genes and 4,536 characteristics. First, we linearly transformed the expression range of each gene to between 0 and 1. Because these characteristics are highly correlated, we used four layers of denoising autoencoder to reduce the dimensionality of the data, and the number of cells in each hidden layer was 3,000-800-300-100, respectively. Moreover, gaussian noise with a noise factor of 0.2 was added to the input data, and sigmoid was used to activate each layer. The model was optimized with Adam, and epoch was 100.

The second source of gene characteristics was from homologous gene phenotypic associations in eight other species, which were more abundant than in studies of human genetic diseases. The data used in the experiment are shown in **Table 1**. The features were extracted by two-layer denoising autoencoder with the following specific parameters: the number of nodes in each layer is "200–100," the corruption level of data is 0.2, the activation function is sigmoid function, the batch size is set as 150, and the model is optimized by Adam.

In addition, the data on interactions between genes can also be used as a part of the characteristics of genes. We integrated two networks, HumanNet (Lee et al., 2011) and STRING (Szklarczyk et al., 2017), for unified analysis. These two sets of data represent gene–gene interaction networks, but there are differences between them (Kuang et al., 2018). The integrated analysis of different sets of data can verify each set, and they can help to validate each other and expand understanding the potential rules. We used the RWR and DCA methods to fuse two networks to extract gene features. We set the restart probability to 0.05 and extracted the 600-dimensional gene characteristics. Finally, the gene characteristics used in the model were 800 dimensions.

TABLE 1 | Species Details.

Number	Species name	Number of disease phenotypes	Number of associations
1	Human	3209	3954
2	Arabidopsis thaliana	1137	12010
3	Worm	744	30519
4	Drosophila	2503	68525
5	Zebrafish	1143	4500
6	Escherichiacoli	324	72846
7	Gallus	1188	22150
8	Mouse	4662	75199
9	Saccharomyce	1243	73284

Disease Characteristics

The disease characteristics are mainly derived from two data sources: the disease similarity network MimMiner and clinical manifestation data of the disease, as well as a large amount of data from analysis of the medical literature.

MimMiner data are processed by literature (van Driel et al., 2006) and are freely available online. This data set has been applied in gene prioritization methods (Vanunu et al., 2010; Singh-Blom et al., 2013; Natarajan and Dhillon, 2014). RWR and DCA were used to extract 100-dimension disease features in the disease similarity network, and the restart probability was set as 0.05.

Another disease feature that we incorporated was from the OMIM disease webpage. We paid special attention to the clinical features and clinical management of webpages. We obtained disease features through text mining. We used PCA to reduce the dimensions of feature space and retained the first 100 principal components. Finally, we obtained 200-dimension disease characteristics.

EXPERIMENT

Evaluation Indexes and Methods introduces the evaluation indexes and methods of the experiment. *Parameter Settings* describes the influence of important parameters in the experiment. In *Global Performance*, the global performance of the experiment is compared. *Prediction of New Genes and New Diseases* compares the ability to predict new genes and new diseases. *Newly Discovered Genes* compares the ability to predict newly discovered associations.

Evaluation Indexes and Methods

In the experiment, to quantitatively evaluate our method and compare it with the most advanced disease-causing gene prioritization methods, we used a cross-validation strategy to measure gene recovery. We divided the known gene–disease pairs into three groups of the same size. The associations in one group were hidden, and the associations in the remaining two groups were used as training data, repeated three times to ensure that each group was hidden only once. For each disease in our data set, we ranked all of the genes according to the degree to which they were associated with the disease. The first r genes were taken as candidate pathogenic genes for corresponding diseases; namely, the top-r ranking method was used. The performance of the algorithm was analyzed by comparing the recall and precision of each method under different thresholds r, usually $r \le 100$. The formula for calculating this was as follows:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{17}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{18}$$

Recall rate refers to the proportion of positive cases correctly judged by the model relative to all positive cases (TP+FN) in the data set. FN represents the data that are mistaken as negative cases by the model but are actually positive cases. The precision rate is the proportion of true positive cases (TP) relative to all positive cases (TP+FP) judged by the model (Xiong et al., 2012; Xu et al., 2017; Cheng et al., 2019; Cheng et al., 2019).

To further confirm the value of our approach, we also used the mean percentile ranking (MPR), an evaluation index based on recall, to evaluate the performance of the algorithm. This evaluation index has been applied in recommendation algorithm and analyses of the performance for predicting drug-targets (Hu et al., 2008; Johnson, 2014; Li et al., 2015; Ding et al., 2017; Hao et al., 2019; Liu et al., 2019b; Liu et al., 2019c; Zeng et al., 2019) and disease biomarkers (Chen et al., 2016; Zeng et al., 2016; Hong et al., 2019; Xu et al., 2019). For each disease, the genes were ranked in descending order according to the calculated gene–disease predictive value. The average ranking of the true and established associations among them is the final result. Here, rank_{ji} can be used to represent the percentile ranking (PR) of gene j and disease i. rank $_{ji}$ = 0% indicates that disease i is most likely to interact with gene j. Similarly, rank $_{ji}$ = 100% indicates that disease i has the lowest probability of interacting with gene j. Therefore, the definition of MPR is as follows:

$$MPR = \frac{\sum_{i=1}^{N_D^t} R_i}{N_D^t} \tag{19}$$

N_T^t represents the number of diseases in the test set, and the formula for calculating R_i is as follows:

$$R_i = \frac{\sum_{j=1}^{N_T^t} \text{rank}_{ji}}{N_T^t} \tag{20}$$

N_T^t represents the number of genes in the test set for current disease i. It is important to emphasize that lower MPR values are preferable because they indicate that our approach has a higher probability, which means that the model works better. Conversely, a higher MPR indicates a lower likelihood of gene interactions with disease. Clearly, the randomly generated list is expected to have an MPR of 50%. Using this measure, we can obtain a list of recommended candidate pathogenic genes, where the recommended optimal prediction is used for higher priority experimental validation.

Parameter Settings

The key parameters of PUIMCHIF are the rank k of matrix $Z\in\mathbb{R}^{800\times200}$, the regularized parameter λ, and the penalty weight α for the unknown relation. As can be seen from **Figure 4**, the performance of the PUIMCHIF method increases with the increase of k. When k = 100, 150, and 200, the three curves are very close. In the following experiment, the PUIMCHIF method uniformly set parameters as follows: k = 200, λ = 0.02, and α=0.0035.

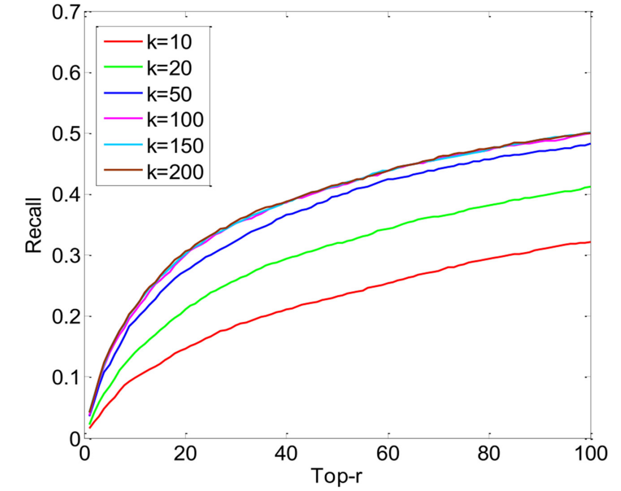


FIGURE 4 | Performance Comparison of PUIMCHIF with Different k Values.

As mentioned earlier, our approach features four improvements over the original IMC approach. For **Figure 5**, recall, precision, and MPR were used to analyze the effect of our improved method. The four experimental results in the figure represent (a) the initial experimental results of the original IMC method, (b) the results of extracting features by using RWR and DCA, instead of PCA, for the network data of diseases and genes, (c) the prediction results of adding STRING data to the gene interaction network, and (d) the experimental results of each index of the PUIMCHIF method.

We found that using RWR and DCA can better extract the gene–gene and disease–disease relationships, and helps to improve the prediction of candidate pathogenic genes. Meanwhile, it was also found that the protein interaction network STRING improved the prediction recall rate to 47.45%, and the MPR value also decreased significantly. Using denoising autoencoder to represent the characteristics of genes and diseases, and introducing the idea of PU-Learning into the inductive matrix completion can further improve the predictive performance.

Global Performance

In this experiment, the threefold cross-validation method was used to compare the overall performance of the proposed method with CATAPULT, Katz, and IMC. As shown in **Figure 6A**, the vertical axis gives the probability of recovering the true gene association in the top-r prediction of different r values on the horizontal axis. The experimental results show that the PUIMCHIF algorithm proposed in this paper has a much higher probability of recovering true gene associations under different thresholds than the other methods. **Figure 6B** presents the precision–recall curve.

In addition, **Table 2** shows the results of three evaluation indexes for each method when the threshold r=100. It is worth mentioning that a smaller value of MPR is associated with a higher probability and a better effect. It can be seen that the MPR value of PUIMCHIF is the lowest and the recall rate reaches 50%, while the best method among other methods, IMC, is only 25%, that is, the recall is doubled. The precision rate was also twice that of Katz which is the best method of other methods, reaching 4.87%. The overall performance of PUIMCHIF has been further improved, confirming the superiority of our method.

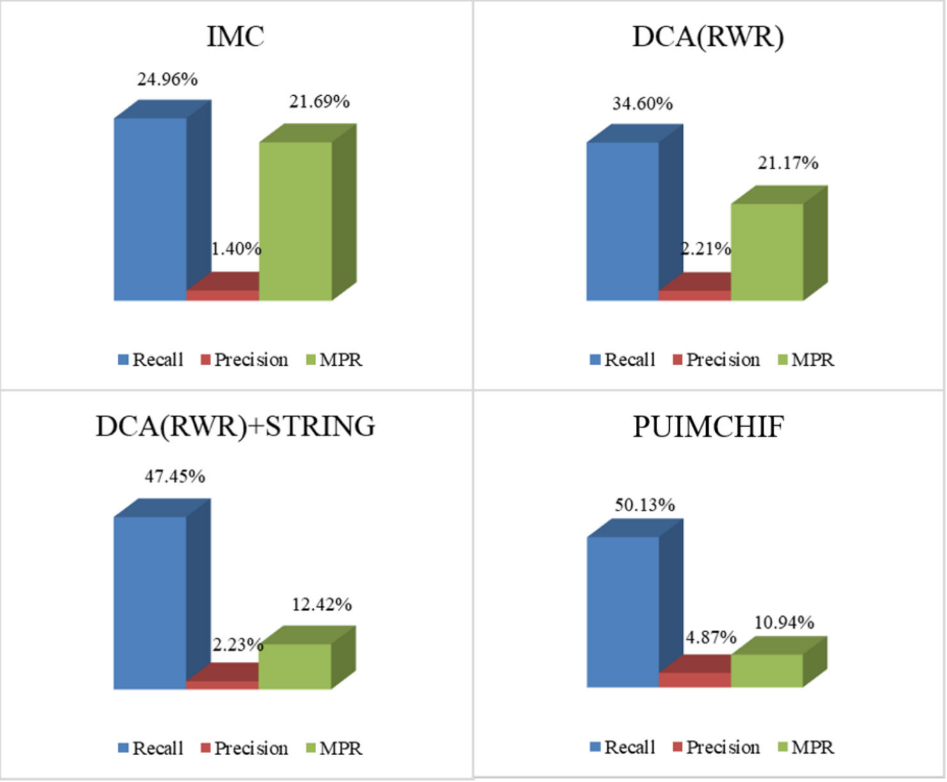


FIGURE 5 | Model Optimization Results.

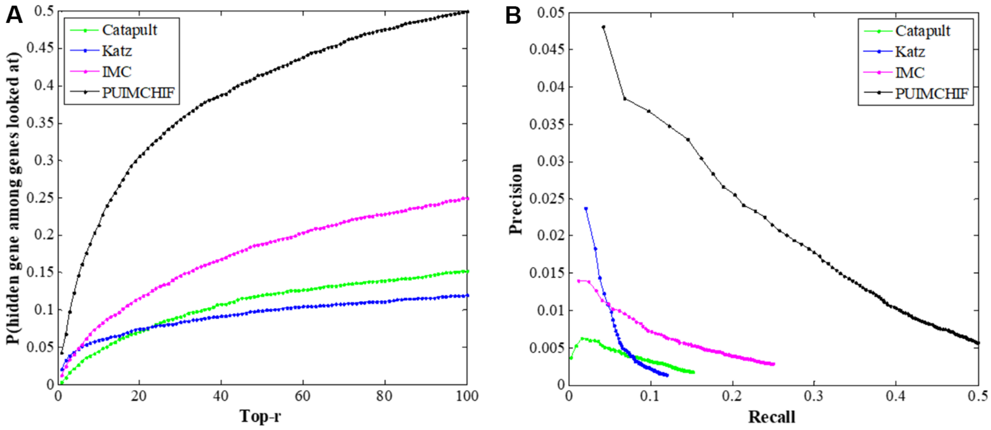


FIGURE 6 | Global Performance with Different Thresholds *r*. **(A)** Recall rate at different threshold *r*. **(B)** Precision-recall curve.

Prediction of New Genes and New Diseases

Prediction of New Genes

One problem affecting prioritization assessments is that well-related genes and diseases tend to be more predictable and therefore tend to generate inflated recall rates. Here, we focued only on genes that are known to have a single association in the gene-disease association data set. In other words, we selected the gene corresponding to the row with only 1 non-zero element in the gene-disease association matrix as the validation set, and hided these known associations in the training process. After repeated three-fold cross validations, **Figure 7A** shows the predictive power of different methods within the threshold $r < 100$. The Y-axis represents the probability of a true known single gene association hidden during recovery training.

TABLE 2 | Experimental Results with Threshold *r* =100.

Methods	Recall	Precision	MPR
CATAPULT	0.152251	0.006289	0.319410
Katz	0.120132	0.023752	0.335564
IMC	0.249621	0.014036	0.216856
PUIMCHIF	0.501265	0.048681	0.109412

Table 3 shows the specific experimental results of each method when r = 100. For the prediction of new genes, although the precision rate was slightly lower than Katz, the recall rate of our PUIMCHIF was significantly higher than other methods, reaching 40.7% when the recall rate of IMC method was only 13.7%. At the same time, we found that using the MPR index to evaluate the results, the PUIMCHIF method was only 13.5%, much lower than Katz and CATAPULT. This also shows that our method is more reliable.

Prediction of New Diseases

Similar to the prediction of new genes, we only considered diseases with a single known association in the gene-disease association data set as the validation set, that is, diseases corresponding to the columns with only 1 non-zero element in the gene-disease association matrix, and hided these known associations during training. Similarly, a three-fold cross-validation analysis was used, and the results are shown in **Figure 7B**. The probability that the proposed method could recover the true association of new diseases reached 48%, which was a significant improvement compared with other methods. Moreover, the MPR value of our method was lower than that of other methods, and the precision rate was nearly 2.7 percentage points higher than that of IMC method. As can be seen from **Table 4**, PUIMCHIF method is superior to other methods in three evaluation indexes.

Newly Discovered Genes

Cross-validation of retrospective data can lead to overly optimistic performance estimates. For example, certain gene interactions may be found because of associations with specific diseases being evaluated. Although the association itself is hidden, other features are contaminated by this information. Therefore, the use of recently reported associations to assess gene prioritization tools is unbiased in this assessment.

We trained all methods using all the gene associations of the 3,209 OMIM diseases collected. We found 162 newly discovered associations, of which 83 genes had no known associations previously. Thus, the assessment of new associations also helps determine the ability of methods to recommend new genes. The ranking performance of each method in 162 new associations is shown in **Figure 8**. We can see that the IMC method is superior to other methods in the range of threshold 6 ≤ r ≤ 100.

CONCLUSION

In this paper, a PU induction matrix completion algorithm based on heterogeneous information fusion, PUIMCHIF, was proposed to predict gene–disease associations. Based on the specific advantages of IMC method, PUIMCHIF can predict new genes and diseases, and has good predictive performance. In addition, because closely connected or functionally similar genes are more likely to cause the same or similar diseases, we

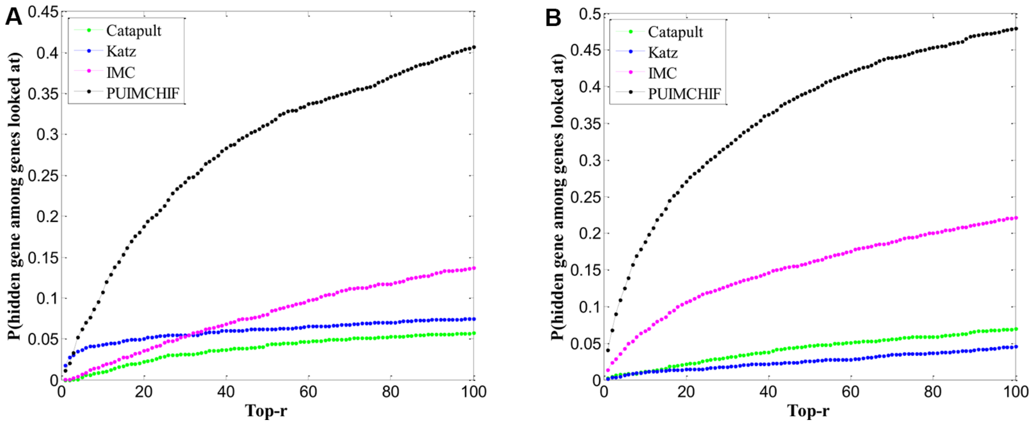


FIGURE 7 | Prediction of New Genes and New Diseases. **(A)** Prediction of New Genes. **(B)** Prediction of New Diseases.

TABLE 3 | Prediction of New Genes with Threshold *r* =100.

Methods	Recall	Precision	MPR
CATAPULT	0.056943	0.001227	0.497410
Katz	0.074838	0.018446	0.466105
IMC	0.137195	0.001935	0.284610
PUIMCHIF	0.407281	0.013840	0.135043

TABLE 4 | Prediction of New Disease with Threshold *r* =100.

Methods	Recall	Precision	MPR
CATAPULT	0.070060	0.002392	0.346974
Katz	0.045454	0.001709	0.363452
IMC	0.221804	0.014012	0.226905
PUIMCHIF	0.479836	0.040671	0.112801

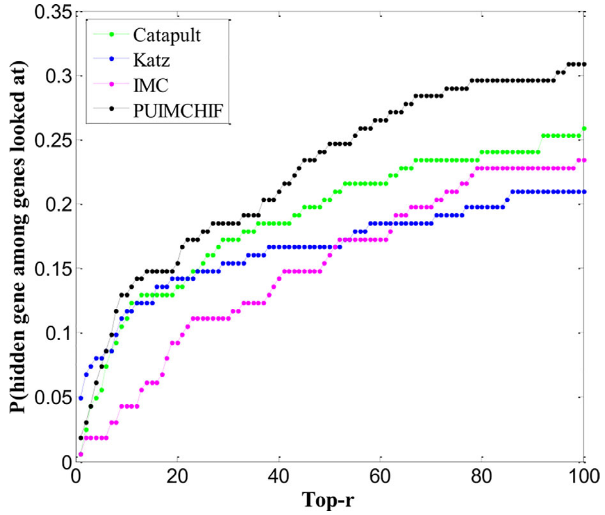


FIGURE 8 | Newly Discovered Prediction of Association.

constructed low-dimensional feature representations of genes and diseases from various data sources such as STRING using the compact feature learning method, which effectively alleviated the impact of data sparsity. Although there is no evidence that genes are unrelated to diseases in the data set, it is clear that most of the unknown associations are negative. PUIMCHIF conducts biased learning by treating unlabeled data as negative cases and constraining the penalty weight of known relationships to be greater than that of unknown relationships. Compared with the existing prediction methods, the PUIMCHIF method can significantly improve the prediction results regarding recall rate, precision rate, and MPR. According to the evaluation index of MPR, the experimental results of the PUIMCHIF method that we proposed are the lowest; that is to say, the candidate genes given by our algorithm have a higher priority for validation by biological experiments.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this article are not publicly available. Requests to access the datasets should be directed to chunyu@hit.edu.cn.

AUTHOR CONTRIBUTIONS

CW initiated the idea, conceived the whole process and drafted the manuscript. JZ and XW implemented the experiments and designed the figures. KH helped with data analysis and revised the manuscript. MG and finalized the paper. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

The work was supported Natural Science Foundation of China (61872114, 61571163, 61532014, and 61871020), and the National Key Research and Development Plan Task of China (No. 2016YFC0901902). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

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 (Database issue), D789–D798. doi: 10.1093/nar/gku1205

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Eppig JT *et al*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res.* 35 (Database issue), D760–D765. doi: 10.1093/nar/gkl887

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36 (5), 431–432. doi: 10.1038/ng0504-431

Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45 (10), 1113–1120. doi: 10.1038/ng.2764

Cao, M., Pietras, C. M., Feng, X., Doroschak, K. J., Schaffner, T., Park, J., et al. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. *Bioinformatics* 30 (12), i219–i227. doi: 10.1093/bioinformatics/btu263

Chen, X., Pérez-Jiménez, M. J., Valencia-Cabrera, L., Wang, B., and Zeng, X. (2016). Computing with viruses. *Theor. Comput. Sci.* 623, 146–159. doi: 10.1016/j.tcs.2015.12.006

Chen, L., Zeng, T., Pan, X. Y., Zhang, Y. H., Huang, T., and Cai, Y. D. (2019). Identifying Methylation Pattern and Genes Associated with Breast Cancer Subtypes. *Int. J. Mol. Sci.* 20 (17), 20. doi: 10.3390/ijms20174269

Chen, L., Pan, X. Y., Zhang, Y. H., Kong, X. Y., Huang, T., and Cai, Y. D. (2019). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell Biochem.* 120 (5), 7068–7081. doi: 10.1002/jcb.27977

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., 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 (Suppl 1), 919. doi: 10.1186/s12864-017-4338-6

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 (11), 1953–1956. doi: 10.1093/bioinformatics/bty002

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

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

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 (1), 203–209. doi: 10.1093/bib/bbx103

Deng, L., Li, W., and Zhang, J. (2019). LDAH2V: Exploring meta-paths across multiple networks for lncRNA-disease association prediction. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2019.2946257

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions *via* multiple information integration. *Inf. Sci.* 418-419, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8 (Suppl 2), I1. doi: 10.1186/1752-0509-8-S2-I1

Hao, M., Bryant, S. H., and Wang, Y. (2019). Open-source chemogenomic data-driven algorithms for predicting drug-target interactions. *Brief Bioinform.* 20 (4), 1465–1474. doi: 10.1093/bib/bby010

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinf.* 19 (1), 306. doi: 10.1186/s12859-018-2321-0

Hong, Z., Zeng, X., Wei, L., and Liu, X. J. B. (2019). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*. (Pisa, Italy: IEEE). doi: 10.1093/bioinformatics/btz694

Hu, Y., Koren, Y., and Volinsky, C. (2008). "Collaborative Filtering for Implicit Feedback Datasets," in *2008 Eighth IEEE International Conference on Data Mining: 15-19 Dec. 2008*, (Pisa, Italy: IEEE), 263–272. doi: 10.1109/ICDM.2008.22

Jiang, R. (2015). Walking on multiple disease-gene networks to prioritize candidate genes. *J. Mol. Cell Biol.* 7 (3), 214–230. doi: 10.1093/jmcb/mjv008

Johnson, C. (2014). "Logistic matrix factorization for implicit feedback data," in *Advances in Neural Information Processing Systems*. Montréal, Canada. vol. 27

Kim, M., and Leskovec, J. (2011). "The Network Completion Problem: Inferring Missing Nodes and Edges in Networks," in *Proceedings of the 2011 SIAM International Conference on Data Mining*. Mesa, Arizona, U. S. A. 47–58. doi: 10.1137/1.9781611972818.5

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82 (4), 949–958. doi: 10.1016/j.ajhg.2008.02.013

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A* 101 (42), 15148–15153. doi: 10.1073/pnas.0404315101

Kuang, L., Yu, L., Huang, L., Wang, Y., Ma, P., Li, C., et al. (2018). A Personalized QoS Prediction Approach for CPS Service Recommendation Based on Reputation and Location-Aware Collaborative Filtering. *Sensors* 18 (5), 1556. doi: 10.3390/s18051556

Lee, I., Blom, U. M., Wang, P. I., Shim, J. E., and Marcotte, E. M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* 21 (7), 1109–1121. doi: 10.1101/gr.118992.110

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26 (9), 1219–1224. doi: 10.1093/bioinformatics/btq108

Li, W., Yu, J., Lian, B., Sun, H., Li, J., Zhang, M., et al. (2015). Identifying prognostic features by bottom-up approach and correlating to drug repositioning. *PloS One* 10 (3), e0118672. doi: 10.1371/journal.pone.0118672

Li, J. R., Lu, L., Zhang, Y. H., Liu, M., Chen, L., Huang, T., et al. (2019). Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell Biochem.* 120 (1), 405–416. doi: 10.1002/jcb.27395

Liao, Z. J., Li, D. P., Wang, X. R., Li, L. S., and Zou, Q. (2018). Cancer diagnosis through isomiR expression with machine learning method. *Curr. Bioinform.* 13 (1), 57–63. doi: 10.2174/1574893611666160609081155

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(D1):D871–D881. doi: 10.1093/nar/gkz1007

Liu, B. (2019) BioSeq-Analysis: a platform for DNA, RNA, and protein sequence analysis based on machine learning approaches. *Briefings In Bioinf.* 20 (4),1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019a). 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.* 47 (20), e127. doi: 10.1093/analys/anz032

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

Liu, B., Zhu, Y., and Yan, K. (2019c). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Briefings Bioinf.* doi: 10.1093/bib/bbz139

Mordelet, F., and Vert, J. P. (2011). ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinf.* 12, 389. doi: 10.1186/1471-2105-12-389

Natarajan, N., and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene-disease associations. *Bioinformatics* 30 (12), i60–i68. doi: 10.1093/bioinformatics/btu269

Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26 (8), 1057–1063. doi: 10.1093/bioinformatics/btq076

Pan, X. Y., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019). Analysis of Expression Pattern of snoRNAs in Different Cancer Types with Machine Learning Algorithms. *Int. J. Mol. Sci.* 20 (9), 16. doi: 10.3390/ijms20092185

Pan, X. Y., Hu, X. H., Zhang, Y. H., Chen, L., Zhu, L. C., Wan, S. B., et al. (2019). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294 (1), 95–110. doi: 10.1007/s00438-018-1488-4

Peng, L., Peng, M. M., Liao, B., Huang, G. H., Li, W. B., and Xie, D. F. (2018). The advances and challenges of deep learning application in biological big data processing. *Curr. Bioinform.* 13 (4), 352–359. doi: 10.2174/1574893612666170707095707

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W., Mazaitis, M., Felix, V., et al. (2012). Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res.* 40 (Database issue), D940–D946. doi: 10.1093/nar/gkr972

Singh-Blom, U. M., Natarajan, N., Tewari, A., Woods, J. O., Dhillon, I. S., and Marcotte, E. M. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PloS One* 8 (5), e58977. doi: 10.1371/journal.pone.0058977

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). Bork P *et al*: The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45 (D1), D362–D368. doi: 10.1093/nar/gkw937

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G., and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* 14 (5), 535–542. doi: 10.1038/sj.ejhg.5201585

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

Wang, S., Cho, H., Zhai, C., Berger, B., and Peng, J. (2015). Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics* 31 (12), i357–i364. doi: 10.1093/bioinformatics/btv260

Wang, L., Ping, P. Y., Kuang, L. N., Ye, S. T., Lqbal, F. M. B., and Pei, T. R. (2018). A novel approach based on bipartite network to predict human microbe-disease associations. *Curr. Bioinform.* 13 (2), 141–148. doi: 10.2174/1574893612666170911143601

Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Mol. Syst. Biol.* 4, 189. doi: 10.1038/msb.2008.27

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Sci.* 10 (Suppl 1), S20. doi: 10.1186/1477-5956-10-S1-S20

Xiong, Y., Wang, Q., Yang, J., Zhu, X., and Wei, D. Q. (2018). PredT4SE-Stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method. *Front. Microbiol.* 9, 2571. doi: 10.3389/fmicb.2018.02571

Xu, Q., Xiong, Y., Dai, H., Kumari, K. M., Xu, Q., Ou, H. Y., et al. (2017). PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *J. Theor. Biol.* 417, 1–7. doi: 10.1016/j.jtbi.2017.01.019

Xu, H., Zeng, W., and Zhang, D. (2019). Zeng XJIToC: MOEA/HD: A Multiobjective Evolutionary Algorithm Based on Hierarchical Decomposition. *IEEE Trans. Cybernetics* 49 (2), 517–526. doi: 10.1109/TCYB.2017.2779450

Zeng, X., Zhang, X., and Liao, Y. (2016). Pan LJBeBA-GS: Prediction and validation of association between microRNAs and diseases by multipath methods. *Biochim. Biophys. Acta -General Subj.* 1860 (11), 2735–2739. doi: 10.1016/j.bbagen.2016.03.016

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and Validation of Disease Genes Using HeteSim Scores. *Ieee-Acm Trans. Comput. Biol. And Bioinf.* 14 (3), 687–695. doi: 10.1109/TCBB.2016.2520947

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

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

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2019). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Briefings In Bioinf.* doi: 10.1093/bib/bbz080

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*. 35 (24), 5191–5198. doi: 10.1093/bioinformatics/btz418

Zhang, J., Zhang, Z., Wang, Z., Liu, Y., and Deng, L. (2018). Ontological function annotation of long non-coding RNAs through hierarchical multi-label classification. *Bioinformatics* 34 (10), 1750–1757. doi: 10.1093/bioinformatics/btx833

Zhang, X., Zou, Q., Rodriguez-Paton, A., and Zeng, X. X. (2019). Meta-path methods for prioritizing candidate disease miRNAs. *Ieee-Acm Trans. Comput. Biol. Bioinf.* 16 (1), 283–291. doi: 10.1109/TCBB.2017.2776280

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. Bioinf.* 16 (2), 407–416. doi: 10.1109/TCBB.2017.2704587

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Brief Funct. Genomics*. 18 (6), 367–376. doi: 10.1093/bfgp/elz018

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.

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fbioe.2020.00045

Gene-Focused Networks Underlying Phenotypic Convergence in a Systematically Phenotyped Cohort With Heterogeneous Intellectual Disability

Yan Wang[1,2,3†], *Li-Na Zhu*[1,2,3†], *Xiu-Wei Ma*[1,2,3†], *Fang Yang*[4], *Xi-Lin Xu*[4], *Yao Yang*[1,2,3], *Xiao Yang*[1,2,3], *Wei Peng*[1,2,3], *Wan-Qiao Zhang*[1,2,3], *Jin-Yu Liang*[5], *Wei-Dong Zhu*[5], *Tai-Jiao Jiang*[4,6], *Xin-Lei Zhang*[7] *and Zhi-Chun Feng*[1,2,3*]

[1] *BaYi Children's Hospital, The Seventh Medical Center of PLA General Hospital, Beijing, China,* [2] *National Engineering Laboratory for Birth Defects Prevention and Control of Key Technology, Beijing, China,* [3] *Beijing Key Laboratory of Pediatric Organ Failure, Beijing, China,* [4] *Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences, Suzhou, China,* [5] *The Second People's Hospital of Aohanqi, Inner Mongolia, China,* [6] *Center of Systems Medicine, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China,* [7] *Suzhou Geneworks Technology Co., Ltd., Suzhou, China*

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Kuo Yang,
Beijing Jiaotong University, China
Shur-Jen Wang,
Medical College of Wisconsin, United States

***Correspondence:**
Zhi-Chun Feng
fengzc81@sina.com;
zhjfengzc@126.com

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

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

Received: *18 November 2019*
Accepted: *21 January 2020*
Published: *07 February 2020*

Citation:
Wang Y, Zhu L-N, Ma X-W, Yang F, Xu X-L, Yang Y, Yang X, Peng W, Zhang W-Q, Liang J-Y, Zhu W-D, Jiang T-J, Zhang X-L and Feng Z-C (2020) Gene-Focused Networks Underlying Phenotypic Convergence in a Systematically Phenotyped Cohort With Heterogeneous Intellectual Disability. Front. Bioeng. Biotechnol. 8:45.
doi: 10.3389/fbioe.2020.00045

The broad spectrum of intellectual disability (ID) patients' clinical manifestations, the heterogeneity of ID genetic variation, and the diversity of the phenotypic variation represent major challenges for ID diagnosis. By exploiting a manually curated systematic phenotyping cohort of 3803 patients harboring ID, we identified 704 pathogenic genes, 3848 pathogenic sites, and 2075 standard phenotypes for underlying molecular perturbations and their phenotypic impact. We found the positive correlation between the number of phenotypes and that of patients that revealed their extreme heterogeneities, and the relative contribution of multiple determinants to the heterogeneity of ID phenotypes. Nevertheless, despite the extreme heterogeneity in phenotypes, the ID genes had a specific bias of mutation types, and the top 44 genes that ranked by the number of patients accounted for 39.9% of total patients. More interesting, enriched co-occurrent phenotypes and co-occurrent phenotype networks for each gene had the potential for prioritizing ID genes, further exhibited the convergences of ID phenotypes. Then we established a predictor called IDpred using machine learning methods for ID pathogenic genes prediction. Using10-fold cross-validation, our evaluation shows remarkable AUC values for IDpred (auc = 0.978), demonstrating the robustness and reliability of our tool. Besides, we built the most comprehensive database of ID phenotyped cohort to date: IDminer http://218.4.234.74:3100/IDminer/, which included the curated ID data and integrated IDpred tool for both clinical and experimental researchers. The IDminer serves as an important resource and user-friendly interface to help researchers investigate ID data, and provide important implications for the diagnosis and pathogenesis of developmental disorders of cognition.

Keywords: intellectual disability, phenotypic convergence, gene-focused networks, co-occurrent phenotype, machine learning, pathogenic genes prediction

INTRODUCTION

Intellectual disability (ID), also known as mental retardation, is characterized by significant impairment in cognition. The patients with ID usually have the obvious shortcomings of adaptive behavior before the age of 18, and a high incidence rate, 1–3%, making it a worldwide social problem (Maulik et al., 2011; Mefford et al., 2012). It can occur in isolation or in combination with congenital malformations or other neurological features such as epilepsy, congenital malformations, sensory impairment, and autism spectrum disorders (ASD), and its severity (mild, moderate, severe, and profound) is highly variable (Vissers et al., 2016). The heterogeneity of phenotypes poses additional challenges for understanding the complex etiology, with contributions by environmental factors, perinatal hypoxia, and genetic factors. In recent years, genetic factors including chromosomal abnormalities, single and multiple gene mutations have found to become increasingly prominent for the disease (Gilissen et al., 2014; Lelieveld et al., 2016; Reichenberg et al., 2016). With the increasing number of ID cases identified in clinics, its phenotypes have found to be extremely heterogeneous. Previous studies found that patients with identical mutations in a single gene could give rise to different phenotypes (Hoischen et al., 2014). As the limitations of detection technologies and the heterogeneity of ID genes and phenotypes, many patients still lack appropriate diagnosis.

In the past 10 years, a large number of studies have been carried out in order to explore the genetic mechanism of ID (Gécz et al., 2009; Ellison et al., 2013). In particular, the development of second-generation sequencing technology facilitates the rapid investigation of more DNA samples from ID cases (Rauch et al., 2012; Gilissen et al., 2014). This led to an expansion in the number of genes associated with ID. Having mass data about ID genes, clinical phenotypes, and pedigrees available in the public domain could shed insights into ID mechanisms. A previous report suggests that ID genes are substantially enriched with co-expression, protein-protein interactions, and specific biological functions. Furthermore, they also revealed combinations of typical phenotypes within process-defined groups of ID disorders by clusters of ID genes with significantly elevated biological coherence (Kochinke et al., 2016). This suggests that ID genes and phenotypes have their own characteristics, and these data can be used to define mechanisms of ID and may improve the diagnosis of patients.

In this study, the ID genes, phenotypes, and pedigrees were extracted manually and analyzed and then integrated to build a standard ID database IDminer, which analyzed the phenotypes, genes, families and their relationships based on the individual patient. Furthermore, the candidate pathogenic genes for ID patients could be prioritized based on the molecular feature of ID genes and the genes specific phenotypes and phenotypic pairs. Furthermore, the similarity between patients was also evaluated via clinical features and could help patients with effective intervention. Importantly, the curated data including ID phenotypes, genes and pedigrees, their integrated analysis and their applications are accessible online via http://218.4.234.74:3100/IDminer/.

MATERIALS AND METHODS

Analysis of Specific Phenotypes and Phenotypic Pairs

Each pathogenic gene could be associated with multiple patient samples, and each patient may have different phenotypes. For each gene, the specific phenotypes were obtained with the enrichment analysis using the hypergeometric distribution. A gene could correspond to multiple patients. For each patient, any two of their phenotypes formed a phenotype pair, referred to as co-occurrence. A phenotype pair could appear in N patients (N represents the frequency of phenotypic pairs). In situations with a single gene affecting multiple cases, multi-phenotypic pairs and their frequencies were obtained. For each phenotypic pair, we analyzed whether the co-occurrence was enriched in the affected patients or not.

Construction of Co-occurrence Network

For constructing a co-occurrence network, all phenotypic pairs with a P-value = 0.05 for at least one gene were built as a non-directional network. In this network, each node represents a phenotype and the node size indicates the frequency of the phenotype in the database, while the edges denote significantly enrichment between phenotype pairs. Then the modules were extracted with the R igraph package.

Phenotype-Based Samples Similarity Analysis

The same phenotype may appear in different patient samples. Based on the number of the same phenotypes between these samples, similarity scores between pairs of patient samples was calculated.

The Phenotype Converting Tool

The tool was used to calculate the similarity between the users' input phenotypes and the 2,075 standard phenotypes in this website. The python module named FuzzyWuzzy was used to calculate the similarity score [0,100]. The higher the score, the more similar the two phenotypes.

Supervised Machine Learning Prediction

In this study, the supervised machine learning method, Support Vector Machine (SVM), was employed for ID pathogenic genes prediction. The R language interface of LIBSVM was used to construct the SVM-based pathogenic predictors. The radial basis function was chosen as the kernel function, and the other parameters were set at the default. A prediction model was trained using repeated 10-fold cross-validation of the training dataset, and their predictive performance was evaluated in the independent test dataset.

Web Interface Configuration

The interface has two main parts: one part displayed the ID knowledge base data and the search results, while the other displayed the input and results of the analysis tool. Through the search box on the main page, users could search for a gene or a phenotype. Through the tools button in the main menu, users could enter the analysis interface, and according to the given phenotypes and genes, the ID genes were identified, and the association between the genes and their phenotypes were visualized. The web service was mainly based on java server pages, JavaScript, R, Python, Ajax, Apache, and MySQL.

RESULTS

Data Curation

We first employed the keywords, such as ID, mental retardation, developmental delay, cognitive impairment, developmental disability, and learning disability to accomplish the literature searches by using PubMed. Then the literature was filtered through the artificial proofing method, and the ID-related papers and genes were retained. The text mining method was used to mark phenotypes in the literature using the HPO[1] database phenotypic information as a reference. Then the gene name, mutation site, and phenotypes were curated manually (**Figure 1**). Based on the sample description in the literature, the family information of the samples were also collated from the HGNC (HUGO Gene Nomenclature Committee) database according to the acquired ID-related gene name information, such as gene alias, chromosome localization, corresponding OMIM ID, and Ensembl ID, and the biological function and pathway information for these genes were marked simultaneously through GO[2] and KEGG[3] databases.

The Landscape and Convergence of ID Genes

Through 1174 ID papers, we obtained a total of 3803 samples with 2075 phenotypic descriptions, that were caused by 704 ID genes. Among these genes, there are 3848 mutations, containing 1793 missense/non-sense mutations, 182 splicings, and 610 indels. We found that the majority of the genes were identified in less than 10 patients, and 305 genes (43.3%) found in only one patient and 103 genes (14.8%) in two patients (**Figure 2A**). Also, a small set of genes caused more patients than other genes, as shown in **Figure 2B**, the top nine genes ranked by the number of patients accounted for 14.9% of the total patient group, and the top 44 genes included 39.9% of patients. Moreover, our analysis also showed some ID genes had the dominant mutation types (**Figure 2C**). For the top 57 genes ranked by the number of ID patients, the majority of mutations of patients harboring mutated MECP2, HUWE1, and CREBBP are gross insertions. In addition, the predominant mutation type of patients with mutated THOC2, KIF1A, KDM5C, IQSEC2, SLC6A8, TBC1D24, MAN1B1, YAP1, GRIN2B, PAK3, NALCN, CLPB, and GRIN1 genes are missense/non-sense mutations, while deletions are mainly found in patients harboring SOX4, NRXN1, FMR1, MEF2C, OPHN1, PQBP1, AUTS1, MYT1L, CNTNAP2, MAPT, and TUSC3 genes. Importantly, the mutation types of 47 of the top 57 genes contained gross insertions (most duplications) and missense/non-sense, suggesting that both deletion and overexpression of these genes were likely to cause ID disease. These findings suggested that despite the diversity of ID genetic variation, most ID patients are caused by a small number of genes based on its genetic bias and convergence.

The Heterogeneities of ID Phenotypes

Among the patient cohort, 637 (16.6%) patients have a unique phenotype, while 901 (23.7%) patients have more than ten phenotypes (**Figure 3A**). Also, our data showed that the number of phenotypes for each patient had a positive correlation with the number of the patients, which showed a significant linear relationship (Spearman *P*-value < 0.001, **Figure 3B**) and indicated the heterogeneity of the ID phenotypes. Additionally, HPO structure analysis found the accompanying phenotypes of ID were also widely distributed, including symptoms in many parts of the body (**Figure 3C**). For these phenotypes, as shown in **Figure 3D**, the top 50 phenotypes ranked by the number of patients exhibited that the ID was usually accompanied by other mental diseases, such as seizure, epilepsy, microcephaly, ataxia, microcephaly and autism, abnormal behaviors containing hypotonia, strabismus, sleep disturbance, constipation, delayed or absent speech, motor delay, hyperactivity, feeding difficulties and inability to walk, and dysmorphism about spine, face, stature, and cryptorchidism. These results showed that the phenotypes of ID patients had extreme heterogeneity.

The Convergences of ID Phenotypes

The phenotypes that were converged for each gene based on the fact that intra-similarity between patients caused by one gene were more than inter-similarity between different genes' patients (**Figure 4A**) and the phenotypes in patients caused by the identified mutations in the same family had more similarity than other families (**Figure 4B**). To better understand the convergence of the ID clinic features, we first obtained the specific phenotypes for each ID gene with enrichment analysis. A total of 143 phenotypes, appearing in at least five patients caused by the same gene, were enriched in some genes' patients (**Figure 4C**). Importantly, among the phenotypes, 47 appeared in only single gene's patients and accounted for 30 genes, which could help to diagnosis the patients caused by the genes (**Figure 4C**). To illustrate the relationships between phenotypes, we also investigated the situation of two phenotypes could be co-occurred in one patient, and the co-occurrence phenotypes were recorded as "phenotypic pairs." We analyzed these phenotypic pairs presented in patients with an enrichment analysis. Interestingly, we found that most enriched phenotypic pairs were specific for a single gene. Like single phenotype analysis, phenotypic pairs made it easy to diagnosis patients with 82 ID genes (**Figure 4D**).

Gene-Focused Network for Phenotype Enrichment

Then we analyzed the network diagram of the phenotypic pairs for each gene, which revealed the gene-focused network (**Figure 5A**) and three typical sub-networks (**Figure 5B**). The

[1] http://human-phenotype-ontology.github.io

[2] http://geneontology.org/

[3] http://www.genome.jp/kegg/

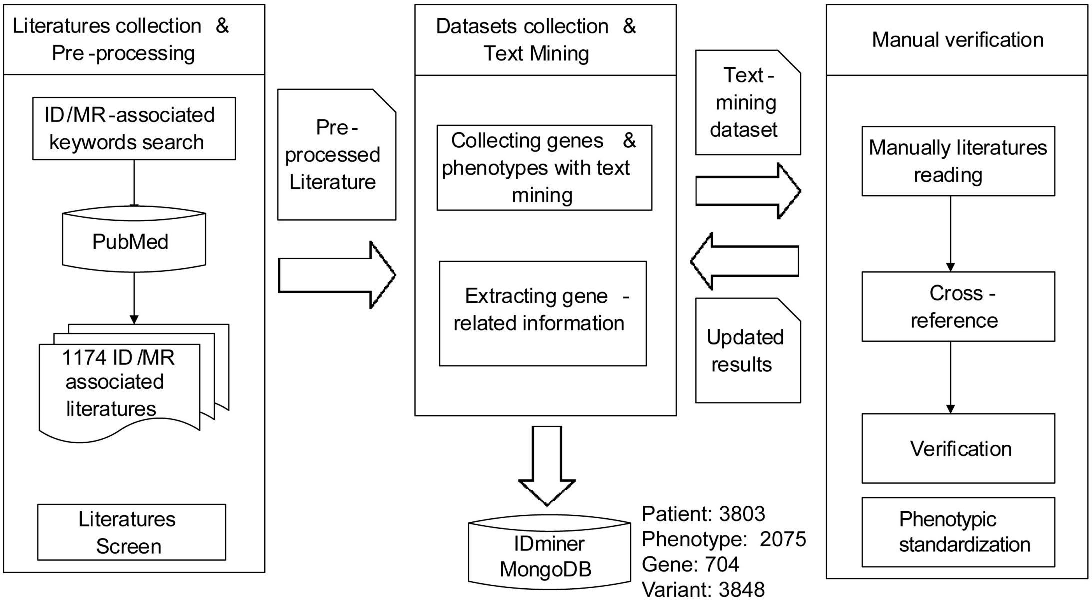

Literatures collection & Pre-processing
ID/MR-associated keywords search
PubMed
1174 ID/MR associated literatures
Literatures Screen
Pre-processed Literature
Datasets collection & Text Mining
Collecting genes & phenotypes with text mining
Extracting gene-related information
Text-mining dataset
Updated results
IDminer MongoDB
Patient: 3803
Phenotype: 2075
Gene: 704
Variant: 3848
Manual verification
Manually literatures reading
Cross-reference
Verification
Phenotypic standardization

FIGURE 1 | The flowchart of data collection and curation. The framework for genes extracting, paper downloading, phenotypes, and pedigrees obtaining and data curating of this project.

first type of sub-network was radial, indicating that most of the phenotypes co-occurred with another one phenotype (like gene ZNF711). The pathogenic genes with the first type of sub-network may have a core phenotype, or an important phenotype that appeared more frequently, and it illustrated that there are strong association between the core phenotype and the biological function. The second type of sub-network was dense, and the phenotypes co-occurred with each other (like gene PIGO). The pathogenic genes with the second type of sub-network often result in a set of concurrent phenotypes. In this case, the prediction of pathogenic genes by phenotype may be more accurate. The third type of sub-network was the mixed state of the above two types (like gene MECP2). With the third type of sub-network of pathogenic genes, the mutations are usually more extensive, the phenotypes are complex, and one independent group phenotypes is often insufficient to reveal the pathogenic genes information. Our analysis showed that the co-occurrence network of each gene had its own characteristics, and the phenotypes in the co-occurrence network of each gene are different. And the co-occurrence networks of different genes had commonality in their structural similarity. Analysis of co-occurring networks further illustrated the phenotypic conservation relative to genes, despite the heterogeneity of phenotypes. Based on the above discoveries, we inferred that the pathogenic genes for patients could be achieved by analyzing specific phenotypes and phenotypic pairs. Our analyses indeed showed that the more the patients' phenotypes, the more accurate the prediction of pathogenic genes (**Figure 5C**). Furthermore, given more phenotypes, the predicted pathogenic genes incline to have a more significant *P*-values (**Figure 5D**). These results showed that phenotypic analysis could reveal the convergences of ID phenotypes and be used for clinical pathogenic gene analysis.

Pathogenic Gene Prediction

Support Vector Machine is one of the most widely used machine learning algorithms in computational biology. It was previously used for predicting virulent proteins in bacterial pathogens (Garg and Gupta, 2008), the clinical outcome from cancer patients (Yeoh et al., 2002) and gene interactions in genetic diseases (Upstill-Goddard et al., 2013). As shown in **Supplementary Figure S1**, developed SVM-based predictor, a 10-fold cross-validation was employed on the training datasets for model selection purpose (**Figure 6A**), and the final performance of the predictor was measured on the independent testing dataset (Ortiz-Gonzalez et al., 2018) compared with other ID pathogenic gene prediction models (Yang et al., 2015; Stelzer et al., 2016; **Figures 6B,C**). The receiver operating characteristic curve (sensitivity against 1-specificity) was used to measure the prediction performance under different decision thresholds, and the area under the curve (AUC) was calculated as the main performance evaluation metric. For calculating variable importance for prediction, 100 sets of independent training were performed using different random seed. The median of variable importance obtained in each training was used as a representative value (**Supplementary Figure S2**).

Database and Tool for ID Research and Diagnosis

In order to represent the ID data and the analysis tools for ID research and diagnosis, the IDminer system was designed. The database included a number of components, including a knowledge base for intellectual disabilities, specific phenotypes and phenotypic pairs for genes, co-occurrence networks, and analysis tools for converting phenotypes to standard phenotypes

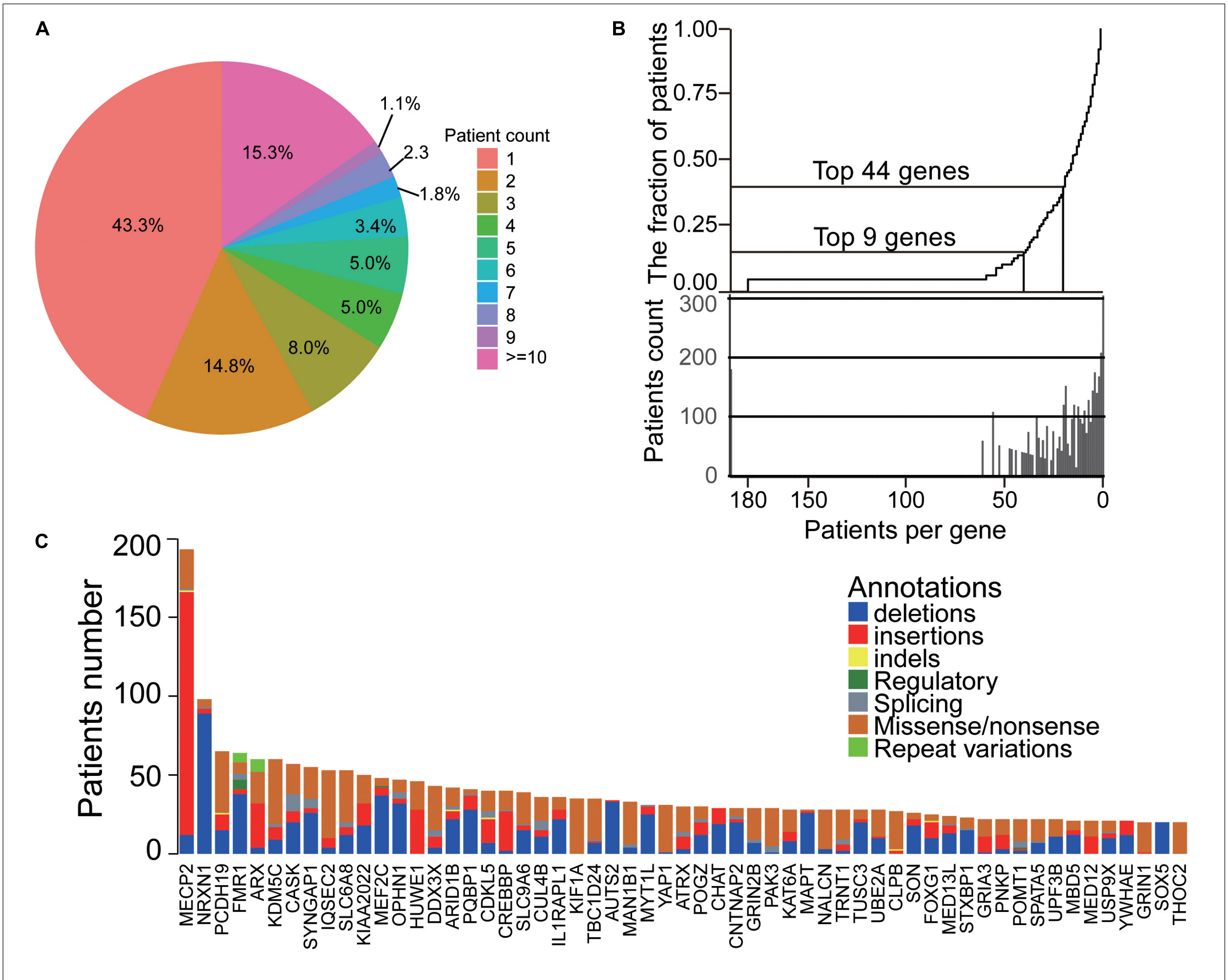


FIGURE 2 | The landscape and convergence of ID genes. **(A)** The distribution of patient number for each gene. Most genes had less than three patients. **(B)** The top genes accounted for most patients. **(C)** The heatmap of genes and their mutations/indels in ID patients.

and exploring the expressions of interesting genes in the brain (**Figure 7**). IDminer was built on open sources software systems, such as MongoDB database, Express web development framework, Nginx web server, and Ubuntu operating system. Python and R were used for data collection processing and analysis. A user-friendly web interface was provided to help users search and analyze the data online at http://218.4.234.74:3100/IDminer/. The interface consists of seven parts: Home, Browser, Tools, Statistics, Download, Help, and Q&A. On the Home page, an introduction to the IDminer outlines a description statistic about all the data integrated into the database and the search module for gene and phenotype. There are two analysis tools for converting phenotypes and prioritizing candidate genes, respectively. Converting phenotypes is to help user mapping their clinical descriptions to our standard ID phenotypes, while co-expression analysis can be based on the brain gene expression data to study the expression profile of the interesting genes and its related genes. In the Document and Q&A pages, the guidelines for the database, and frequently asked questions and answers were showed. Furthermore, our database could be easily updated with the latest published information. For gene query, we provided basic gene information and linked it to multiple external databases, such as containing Ensemble, UniProtKB, GO, KEGG, and OMIM. Reported mutations, ID phenotypes, and patient information were also represented. Additionally, the gene's phenotypic pairs were also interactively visualized. When users entered a phenotypic item in the input box, we listed its basic information such as HPO ID, synonyms and phenotype definitions, reported patients with this phenotype, reported causative genes causing the patients, and its co-occurrence network. For reported genes, in addition to displaying detailed mutation information of these genes, we also annotated the genes' functions and performed PPI network analysis. Importantly, the query clinic feature could be enriched for some genes, and the

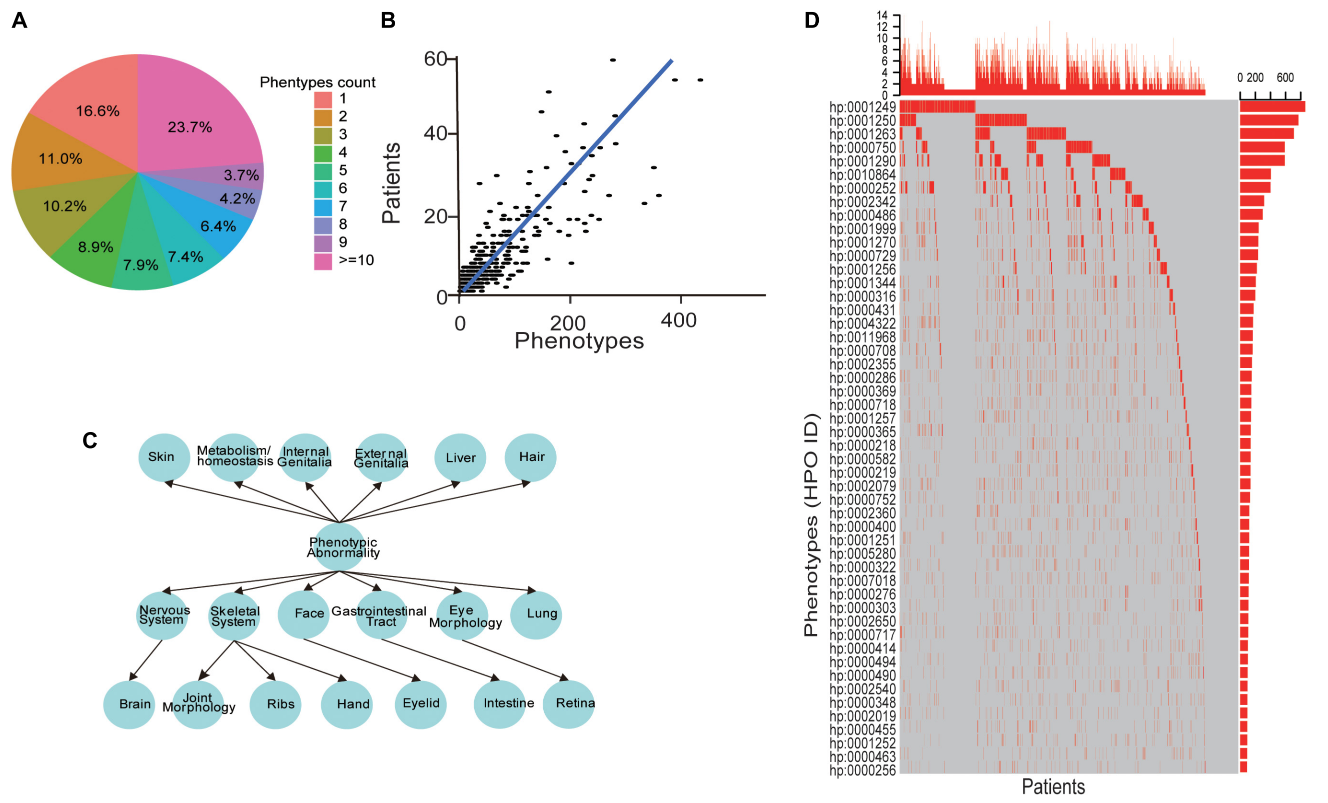


FIGURE 3 | The heterogeneities of ID phenotypes. **(A)** The distribution of the phenotypes number for each patient. **(B)** A scatter point and line fitting showing the correlation between the patient number and phenotypes count. The patient number and phenotype count were derived from each gene in the database. **(C)** The phenotypes structure of ID patients. **(D)** The oncoprint-like representation of phenotypes in ID patients.

genes were also listed. Finally, the top co-occurred phenotypic pairs ranked by their frequencies were shown as a network and the enriched genes for each pair were shown by clicking the edge.

A Use Case for the IDpred

The case of a real patient with the pathogenic gene AAR2 and the standardized phenotypes [Microcephaly (HP:0000252), Cochlear malformation (HP:0008554), Hypoplasia of the corpus callosum (HP:0002079), Ventricular septal defect (HP:0001629), Global developmental delay (HP:0001263), Anteriorly placed anus (HP:0001545), Macule (HP:0012733), Patent foramen ovale (HP:0001655)] was selected based on the previous studies (Charng et al., 2016). The other input candidate genes were randomly selected from the gene list in our database. Then, the query genes list consisted of MXRA8, DMBX1, AAR2, CLIC2, PLA2G6, and phenotypes list consisted of all the standardized phenotypes of this patient (genes and phenotypes are separated by semicolons) were entered into the corresponding box on the page of the website. Then the selection of the models (for example, SVM) with the appropriate parameters should be submitted (**Supplementary Figure S3A**). The result page contains seven columns (GeneSymbol, PathogenicGeneRank, PathogenicScore, Pathogenicity, SimilarRank, SimilarScore, and Phenotypes) would be displayed. On the result table, PathogenicGeneRank is the rank of the input pathogenic genes compared to all deposited genes in our database, PathogenicScore is the score of the pathogenic genes, and Pathogenicity is defined as "Probably" (PathogenicScore > 0.5) or "Less likely" (PathogenicScore = 0.5). SimilarRank refers to the rank of similarity between gene and phenotypes, and SimilarScore refers to the calculated score of similarity between gene and phenotypes. Phenotypes listed the phenotypes related to the GeneSymbol. As shown in the result of this case, AAR2 was predicted as the pathogenic gene with the highest pathogenic score of 0.788 (**Supplementary Figure S3B**).

DISCUSSION

Our work manually extracted a large number of genes, clinic phenotypes and basic information of the patients from published ID literature. By integrating these data for comprehensive analysis, we have provided a holistic view of the current genetic research of ID and made the correlation of various clinic factors of ID patients, prompting researchers to further explore the mechanisms causing ID. The mutation spectrum delineated in our datasets provided essential information for molecular diagnosis in ID patients. Though most genes had its major mutation types, the spectrum showed that all mutation types were identified in ID cases. This combination of mutation types

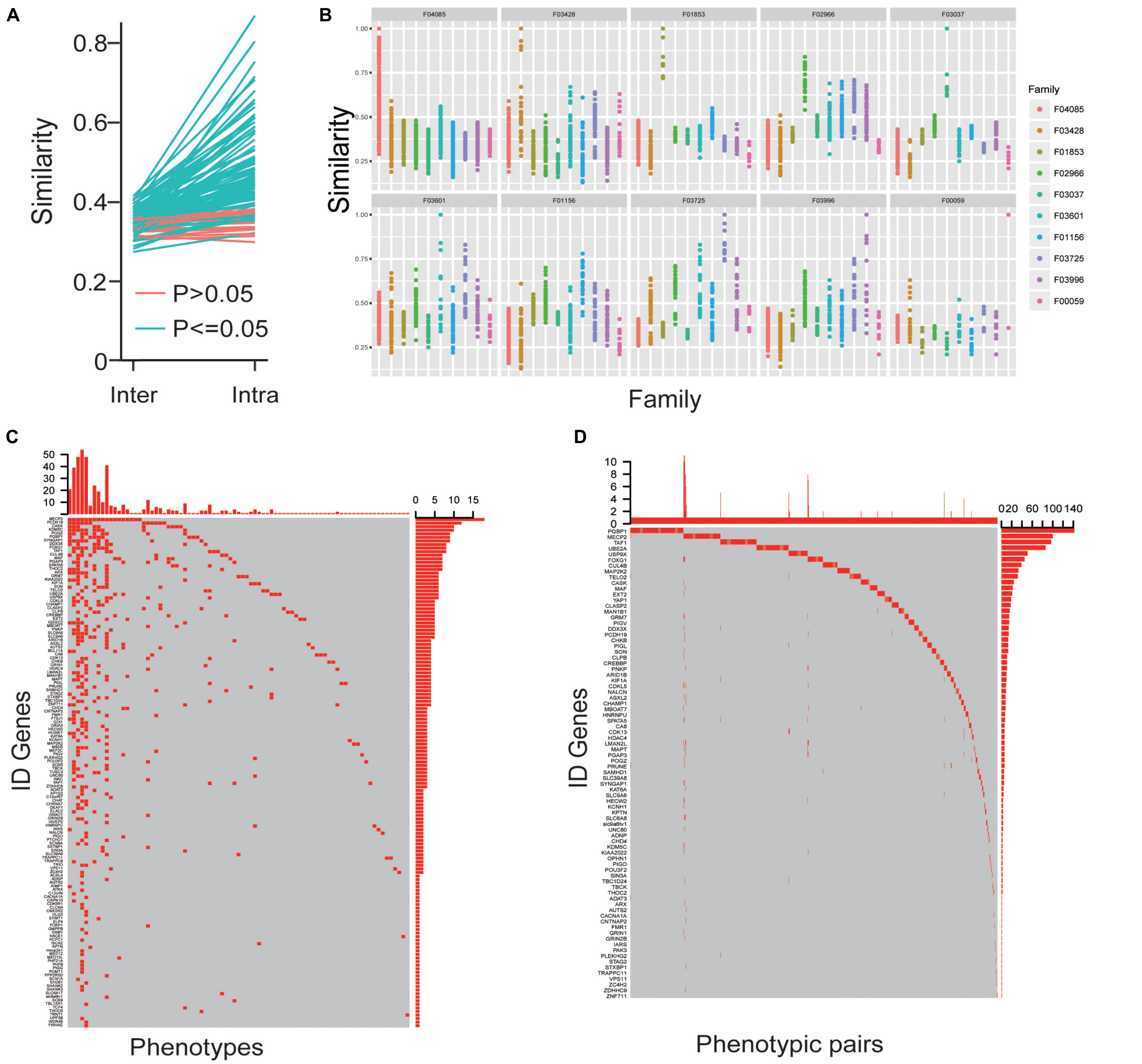


FIGURE 4 | The convergences of ID phenotypes. **(A)** The mean value of intra-similarity between patients caused by one gene was higher than the mean value of inter-similarity between the gene's patients and other genes' patients. **(B)** The similarity of phenotypes in patients caused by identical mutations among the same and other families. **(C)** The oncoprint-like representation of specific phenotypes for genes. **(D)** The oncoprint-like representation of specific phenotypic pairs for genes.

raises the need of using several clinical detection methods for ID diagnoses such as Array Comparative Genomic Hybridization (aCGH), target panel sequencing, whole exon sequencing, and even whole genome sequencing (De Ligt et al., 2013; Redin et al., 2014). Notably, because a small number set of genes accounted for most ID patients, targeted panel sequencing may be favorable than other methods in consideration of cost, time and the difficulty of the data analysis.

The phenotypes of ID patients were extremely diverse and heterogeneous. Unlike the previous study of phenotype-based clustering (Kochinke et al., 2016), we mapped the phenotypes of ID patients to HPO items and found the 2075 phenotypes in total 3803 patients. We confirmed not only mutations in different genes could lead to various phenotypes, but defects in a single gene had been implicated in different phenotypes. Interestingly, there was also considerable phenotypic heterogeneity even among individuals who have identical mutations in the disease gene. We speculated that, besides various genes, the heterogeneity of phenotypes could be affected by other factors, such as mutation types, genetic background, and environment. Though the phenotypes of ID patients were heterogeneous, the specific phenotypes for genes could be analyzed and used for prioritizing caused genes. A previous report suggests that, for tubulinopathies, each mutated gene has

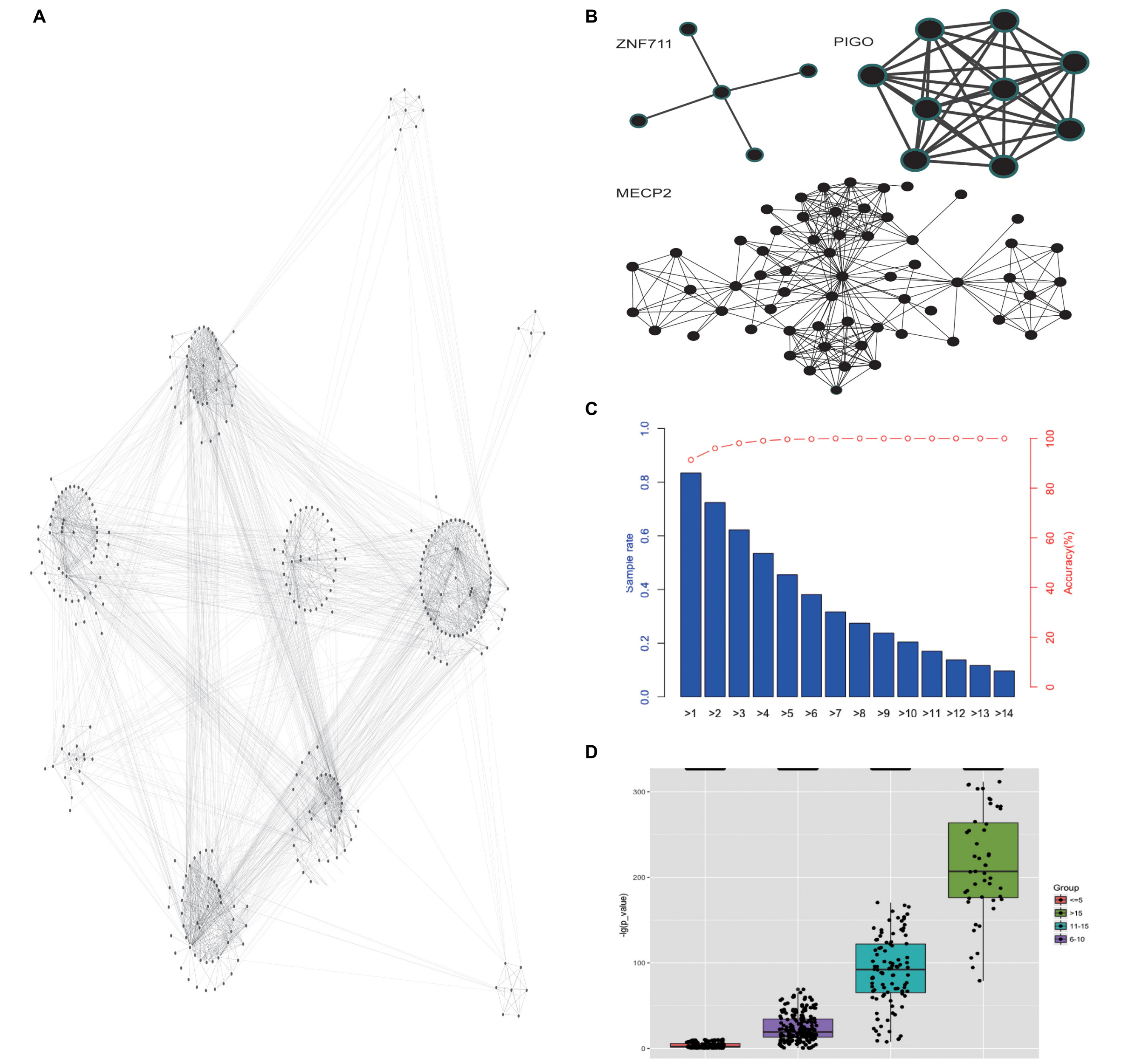


FIGURE 5 | Gene-focused network for phenotype enrichment. **(A)** The network of phenotypes that were enriched in genes. **(B)** The three types of co-occurrence sub-networks. **(C)** The accuracy of the predicted pathogenic gene with phenotypes. **(D)** The *P*-value distribution of predicted pathogenic gene based on different number of given phenotypes.

an associated predominant pattern of cortical dysgenesis (Bahi-Buisson et al., 2014). Additionally, the previous studies in ID found that convergent molecular pathways result in common phenotypes (Kochinke et al., 2016), allowing some phenotype-genotype correlation. However, the common phenotypes for each gene could be achieved until recently the applications of NGS, aCGH, target sequencing, WES, and WGS to ID patients, which lead to an increase of diagnosis. This larger sample size could raise the power of the statistical significance test. Then, for some genes, a large number of patients are sufficient to statistically to find the specific phenotypes, phenotypic pairs and co-occurrence networks for the genes. These features were extracted with enrichment in patients subgroup caused by each gene, confirming the phenotype-genotype correlations and the convergence of ID phenotypes among their extreme heterogeneities.

With the deepening of ID research and the increase of reported patients, it also requires the development of analytical tools for ID researchers to understand the data. Therefore, providing online friendly and easy-to-use analysis tools will also greatly assist in the research of the entire ID field. So, our website not only provides

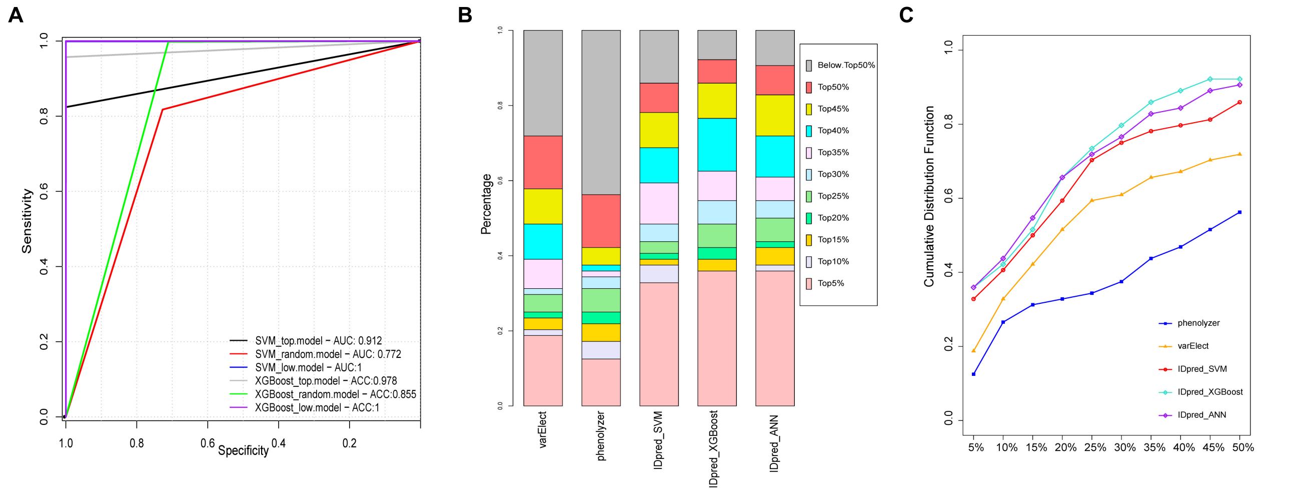
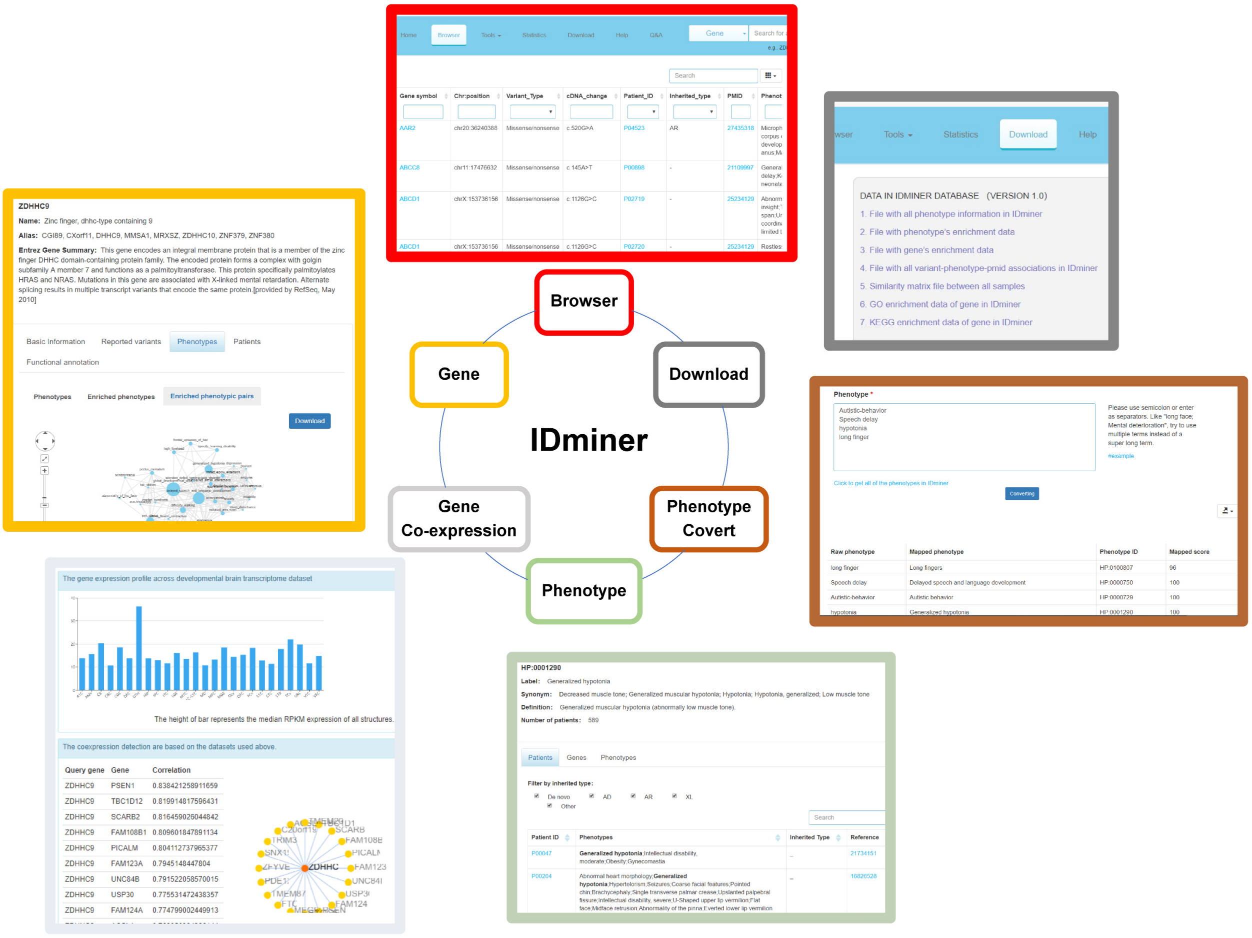


FIGURE 6 | Performance comparison of pathological gene prediction between IDpred and other algorithms. **(A)** ROC curve derived from IDpred model based on 10 fold cross validation. **(B)** the percentage of predicted pathogenic gene derived from IDpred, phenolyzer, and varElect. **(C)** cumulative distribution of TopN rate base on the rank of the pathogenic gene derived from IDpred, phenolyzer, and varElect.



Gene symbol	Chr:position	Variant_Type	cDNA_change	Patient_ID	Inherited_type	PMID
AAR2	chr20:36240388	Missense/nonsense	c.520G>A	P04523	AR	27435318
ABCC8	chr11:17476632	Missense/nonsense	c.145A>T	P00888	-	21109997
ABCD1	chrX:153736156	Missense/nonsense	c.1126G>C	P02719	-	25234129
ABCD1	chrX:153736156	Missense/nonsense	c.1126G>C	P02720	-	25234129

ZDHHC9

Name: Zinc finger, dhhc-type containing 9

Alias: CGI89, CXorf11, DHHC9, MMSA1, MRXSZ, ZDHHC10, ZNF379, ZNF380

Entrez Gene Summary: This gene encodes an integral membrane protein that is a member of the zinc finger DHHC domain-containing protein family. The encoded protein forms a complex with golgin subfamily A member 7 and functions as a palmitoyltransferase. This protein specifically palmitoylates HRAS and NRAS. Mutations in this gene are associated with X-linked mental retardation. Alternate splicing results in multiple transcript variants that encode the same protein.[provided by RefSeq, May 2010]

DATA IN IDMINER DATABASE (VERSION 1.0)

1. File with all phenotype information in IDminer
2. File with phenotype's enrichment data
3. File with gene's enrichment data
4. File with all variant-phenotype-pmid associations in IDminer
5. Similarity matrix file between all samples
6. GO enrichment data of gene in IDminer
7. KEGG enrichment data of gene in IDminer

Phenotype *

Autistic-behavior
Speech delay
hypotonia
long finger

Please use semicolon or enter as separators. Like "long face; Mental deterioration", try to use multiple terms instead of a super long term.

Raw phenotype	Mapped phenotype	Phenotype ID	Mapped score
long finger	Long fingers	HP:0100807	96
Speech delay	Delayed speech and language development	HP:0000750	100
Autistic-behavior	Autistic behavior	HP:0000729	100
hypotonia	Generalized hypotonia	HP:0001290	100

The gene expression profile across developmental brain transcriptome dataset

The height of bar represents the median RPKM expression of all structures.

The coexpression detection are based on the datasets used above.

Query gene	Gene	Correlation
ZDHHC9	PSEN1	0.838421258911659
ZDHHC9	TBC1D12	0.819914817596431
ZDHHC9	SCARB2	0.816459026044842
ZDHHC9	FAM108B1	0.809601847891134
ZDHHC9	PICALM	0.804112737965377
ZDHHC9	FAM123A	0.7945148447804
ZDHHC9	UNC84B	0.791522058570015
ZDHHC9	USP30	0.775531472438357
ZDHHC9	FAM124A	0.774799002449913

HP:0001290

Label: Generalized hypotonia

Synonym: Decreased muscle tone; Generalized muscular hypotonia; Hypotonia; Hypotonia, generalized; Low muscle tone

Definition: Generalized muscular hypotonia (abnormally low muscle tone).

Number of patients: 589

Patients Genes Phenotypes

Filter by inherited type: ☑ De novo ☑ AD ☑ AR ☑ XL ☑ Other

Patient ID	Phenotypes	Inherited Type	Reference
P00047	**Generalized hypotonia**;Intellectual disability, moderate;Obesity;Gynecomastia	-	21734151
P00204	Abnormal heart morphology;**Generalized hypotonia**;Hypertelorism;Seizures;Coarse facial features;Pointed chin;Brachycephaly;Single transverse palmar crease;Upslanted palpebral fissure;Intellectual disability, severe;U-Shaped upper lip vermilion;Flat face;Midface retrusion;Abnormality of the pinna;Everted lower lip vermilion	-	16826528

FIGURE 7 | The illustration of functional modules of IDminer database. The six functional modules of IDminer: Brower, Genes, Gene co-expression, Phenotype, Phenotype convert, and Download.

a knowledge base of ID but also aggregates tools commonly used in ID analysis. And more analysis tools for ID will be added in the future to promote ID research as much as possible.

Overall, our data and analysis showed the convergences of ID genes and phenotypes among their extreme heterogeneities. For genes, the convergence was characterized by the fact that a small percentage of genes could explain the majority of ID phenotypes. And for phenotypes, it was represented as genes' specific phenotype and phenotypic pairs. Importantly, we provided analysis tools based on ID genes and phenotypes in hopes of establishing the standard ID gene and phenotype libraries and, in turn, aiding in clinical diagnosis. Overall, the findings and tools could contribute to the understanding of the genetic basis of ID disease and ultimately improve the diagnosis and treatment of the disease.

CONCLUSION

Our analysis provided evidence to support, though the ID genes and phenotypes were extremely heterogeneous, the genetic bias and phenotypic convergence deserved our more attention, which may help to help us to quickly diagnose ID patients and further promote the studies of disease mechanisms. Moreover, our curated data, analysis, and developed tools were integrated to build a standard ID database IDminer, which could be accessed through http://218.4.234.74:3100/IDminer/. The database and interface are user-friendly for geneticists and clinicians, and a very wide range of ID researchers.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://218.4.234.74:3100/IDminer/.

AUTHOR CONTRIBUTIONS

Z-CF, YW, L-NZ, and X-WM conceived the project, analyzed and interpreted the data, and wrote the manuscript. FY, X-LX, YY, XY, WP, W-QZ, J-YL, W-DZ, and X-LZ acquired data and performed bioinformatics analyses. T-JJ edited the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (Project Nos. 2016YFC1000707 and 2017YFC1001701), and the Sanming Project of Medicine in Shenzhen (SZSM201606088).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00045/full#supplementary-material

FIGURE S1 | The flowchart of pathogenic gene prediction model specification. The framework for training and test data generation, model construction, and validation.

FIGURE S2 | Feature importance scores derived from IDpred. Feature importance is defined as the average gain of the feature in trees from XGBoost in IDpred_XGBoost.top, IDpred_XGBoost.random, and IDpred_XGBoost.low model.

FIGURE S3 | Dpred interface and direct mode example. **(A)** The user enters three types of input: gene symbol list, phenotype expression and modeling type. **(B)** Output results presented in a tab with seven columns which were defined as GeneSymbol, PathogenicGeneRank, PathogenicScore, pathogenicity, SimilarRank, SimilarScore, and Phenotypes.

REFERENCES

Bahi-Buisson, N., Poirier, K., Fourniol, F., Saillour, Y., Valence, S., Lebrun, N., et al. (2014). The wide spectrum of tubulinopathies: what are the key features for the diagnosis? *Brain* 137, 1676–1700. doi: 10.1093/brain/awu082

Charng, W. L., Karaca, E., Coban Akdemir, Z., Gambin, T., Atik, M. M., Gu, S., et al. (2016). Exome sequencing in mostly consanguineous Arab families with neurologic disease provides a high potential molecular diagnosis rate. *BMC Med. Genomics* 9:42. doi: 10.1186/s12920-016-0208-3

De Ligt, J., Willemsen, M. H., Van Bon, B. W. M., Kleefstra, T., Yntema, H. G., Kroes, T., et al. (2013). Diagnostic exome sequencing in persons with severe intellectual disability. *Obstet. Gynecol. Surv.* 68, 191–193. doi: 10.1097/01.ogx.0000428160.59063.a6

Ellison, J. W., Rosenfeld, J. A., and Shaffer, L. G. (2013). Genetic Basis of Intellectual Disability. *Annu. Rev. Med.* 64, 441–450. doi: 10.1146/annurev-med-042711-140053

Garg, A., and Gupta, D. (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics* 9:62. doi: 10.1186/1471-2105-9-62

Gécz, J., Shoubridge, C., and Corbett, M. (2009). The genetic landscape of intellectual disability arising from chromosome X. *Trends Genet.* 25, 308–316. doi: 10.1016/j.tig.2009.05.002

Gilissen, C., Hehir-Kwa, J. Y., Thung, D. T., Van De Vorst, M., Van Bon, B. W. M., Willemsen, M. H., et al. (2014). Genome sequencing identifies major causes of severe intellectual disability. *Nature* 511, 344–347. doi: 10.1038/nature13394

Hoischen, A., Krumm, N., and Eichler, E. E. (2014). Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat. Neurosci.* 17, 764–772. doi: 10.1038/nn.3703

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., et al. (2016). Systematic phenomics analysis deconvolutes genes mutated in intellectual disability into biologically coherent modules. *Am. J. Hum. Genet.* 98, 149–164. doi: 10.1016/j.ajhg.2015.11.024

Lelieveld, S. H., Reijnders, M. R. F., Pfundt, R., Yntema, H. G., Kamsteeg, E. J., De Vries, P., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* 19, 1194–1196. doi: 10.1038/nn.4352

Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. *Res. Dev. Disabil.* 32, 419–436. doi: 10.1016/j.ridd.2010.12.018

Mefford, H. C., Batshaw, M. L., and Hoffman, E. P. (2012). Genomics, intellectual disability, and autism. *N. Engl. J. Med.* 366, 733–743. doi: 10.1056/NEJMra1114194

Ortiz-Gonzalez, X. R., Tintos-Hernandez, J. A., Keller, K., Li, X., Foley, A. R., Bharucha-Goebel, D. X., et al. (2018). Homozygous boricua TBCK mutation causes neurodegeneration and aberrant autophagy. *Ann. Neurol.* 83, 153–165. doi: 10.1002/ana.25130

Rauch, A., Wieczorek, D., Graf, E., Wieland, T., Endele, S., Schwarzmayr, T., et al. (2012). Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *Lancet* 380, 1674–1682. doi: 10.1016/S0140-6736(12)61480-9

Redin, C., Gérard, B., Lauer, J., Herenger, Y., Muller, J., Quartier, A., et al. (2014). Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. *J. Med. Genet.* 51, 724–736. doi: 10.1136/jmedgenet-2014-102554

Reichenberg, A., Cederlöf, M., McMillan, A., Trzaskowski, M., Kapara, O., Fruchter, E., et al. (2016). Discontinuity in the genetic and environmental causes of the intellectual disability spectrum. *Proc. Natl. Acad. Sci. U.S.A.* 113, 1098–1103. doi: 10.1073/pnas.1508093112

Stelzer, G., Plaschkes, I., Oz-Levi, D., Alkelai, A., Olender, T., Zimmerman, S., et al. (2016). VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. *BMC Genomics* 17(Suppl. 2):444. doi: 10.1186/s12864-016-2722-2

Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A. (2013). Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief. Bioinform* 14, 251–260. doi: 10.1093/bib/bbs024

Vissers, L. E. L. M., Gilissen, C., and Veltman, J. A. (2016). Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* 17, 9–18. doi: 10.1038/nrg3999

Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* 12, 841–843. doi: 10.1038/nmeth.3484

Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., et al. (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1, 133–143. doi: 10.1016/S1535-6108(02)00032-6

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 07 February 2020
doi: 10.3389/fbioe.2020.00003

Integrative Analysis of Methylation and Gene Expression in Lung Adenocarcinoma and Squamous Cell Lung Carcinoma

Hao Zhang[1†], *Zhou Jin*[1,2†], *Ling Cheng*[3] *and Bin Zhang*[1*]

[1] *Department of Respiratory and Critical Care Medicine, Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, China,* [2] *Department of Respiration, Hospital of Traditional Chinese Medicine of Zhenhai, Ningbo, China,* [3] *Shanghai Engineering Research Center of Pharmaceutical Translation, Shanghai, China*

Lung cancer is a highly prevalent type of cancer with a poor 5-year survival rate of about 4–17%. Eighty percent lung cancer belongs to non-small-cell lung cancer (NSCLC). For a long time, the treatment of NSCLC has been mostly guided by tumor stage, and there has been no significant difference between the therapy strategy of lung adenocarcinoma (LUAD) and squamous cell lung carcinoma (SCLC), the two major subtypes of NSCLC. In recent years, important molecular differences between LUAD and SCLC are increasingly identified, indicating that targeted therapy will be more and more histologically specific in the future. To investigate the LUAD and SCLC difference on multi-omics scale, we analyzed the methylation and gene expression data together. With the Boruta method to remove irrelevant features and the MCFS (Monte Carlo Feature Selection) method to identify the significantly important features, we identified 113 key methylation features and 23 key gene expression features. HNF1B and TP63 were found to be dysfunctional on both methylation and gene expression levels. The experimentally determined interaction network suggested that TP63 may play an important role in connecting methylation genes and expression genes. Many of the discovered signature genes have been supported by literature. Our results may provide directions of precision diagnosis and therapy of LUAD and SCLC.

Keywords: lung adenocarcinoma, squamous cell lung carcinoma, methylation, gene expression, Boruta, Monte Carlo Feature Selection

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Yunguang Tong,
University of California, Los Angeles, United States
Hong Liu,
Temple University, United States

Correspondence:
Bin Zhang
beanzhang0821@zju.edu.cn;
binzhang0821@zju.edu.cn

† *These authors have contributed equally to this work*

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

Received: *31 October 2019*
Accepted: *03 January 2020*
Published: *07 February 2020*

Citation:
Zhang H, Jin Z, Cheng L and Zhang B (2020) Integrative Analysis of Methylation and Gene Expression in Lung Adenocarcinoma and Squamous Cell Lung Carcinoma. Front. Bioeng. Biotechnol. 8:3. doi: 10.3389/fbioe.2020.00003

INTRODUCTION

Lung cancer, considered to be a highly prevalent type of cancer, is a leading cause of cancer-related mortality worldwide, resulting in 1.6 million deaths each year with poor 5-year survival rate of about 4–17% (Hirsch et al., 2017; Altorki et al., 2019). Lung cancer is classified as follows: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC), accounting for approximately 20 and 80% of all lung cancer cases, respectively (Oser et al., 2015). NSCLC is a complex systems disease with dysfunctions on multiple pathways and multiple molecular levels (Huang et al., 2012, 2015; Li et al., 2013; Zhou et al., 2015; Chen et al., 2016; Liu et al., 2017). It can also be typically divided into three main subtypes, lung adenocarcinoma (LUAD), squamous cell lung carcinoma (SCLC), and large cell cancer (LCC), according to standard pathology methods

(Socinski et al., 2016; Swanton and Govindan, 2016; Herbst et al., 2018). Compared with squamous lung cancer, adenocarcinoma was associated with better prognosis. Despite the advances in diagnostic and therapeutic technology, lung cancer remains a serious global public health concern.

For a long time, the treatment of NSCLC has been mostly guided by tumor stage, and there has been no significant difference between the therapy strategy of LUAD and SCLC. Most lung cancers are usually diagnosed at an advanced stage and are treated primarily with systemic chemotherapy, typically with platinum-based regimens (Bishop et al., 2010). Recent progress in characterization of NSCLC by molecular typing, especially in adenocarcinomas of the lung, have brought new investigation of therapeutic agents that target dominant oncogenic mutations, such as epidermal growth factor receptor (EGFR)-targeted therapies, which have showed improved response rates in patients with NSCLC (Shigematsu et al., 2005).

Currently, progress in molecular biology of lung cancer has resulted in the identification of multiple potential biomarkers that may be related to the clinical management of NSCLC patients. In recent years, with the emergence of next-generation sequencing technologies, important molecular differences between LUAD and SCLC are increasingly identified, indicating that targeted therapy will be more and more histologically specific in the future (Kim et al., 2005; Sun et al., 2007; Li et al., 2014). Several studies have identified multiple gene expression subtypes that differ in prognosis, genomic alterations, clinical characteristics, including tumor differentiation, stage-specific survival, underlying drivers, and potential responses to treatment within LUAD and SCLC (Wilkerson et al., 2010; Thomas et al., 2014; Lu et al., 2016). For example, LUAD patients that harbor EGFR, ALK, ROS1, or BRAF mutations were discovered to benefit the most (Villalobos and Wistuba, 2017; Herbst et al., 2018). Targeted therapies for gene abnormalities of HER2, MET, RET, and NTRK1 appear to be an effective approach to treat LUAD (Dearden et al., 2013; Mazieres et al., 2013). SCLC shows different mutation spectrum from that of adenocarcinoma, and the mutation targeted therapy for SCLC has not been thoroughly studied to obtain approved treatment (Bunn et al., 2016; Soldera and Leighl, 2017).

A series of imaging studies suggested that NSCLC may progress rapidly between occurrence and primary treatment (Koh et al., 2017). Therefore, it is necessary for clinicians to identify between these two subtypes of NSCLC in a convenient and rapid way. With the improvement of the above clinical and molecular levels, growing evidences have shown that immunohistochemistry (IHC) is an effective tool for differentiating adenocarcinoma from squamous cell carcinoma (Bass et al., 2009; Weiss et al., 2010).

It is reported that the formation and development of lung cancer are related to the accumulation of permanent genetic changes and dynamic epigenetic changes. Therefore, enhancing our understanding of tumor biology and gene expression profiles will be critical for cancer treatment and diagnosis. In this study, an integrative analysis of lung cancer methylation data and gene expression data was performed, and mixed features were also screened out for analysis.

MATERIALS AND METHODS

The Joint Methylation and Expression Profiles of Lung Cancer Patients

The methylation and gene expression profiles of lung cancer patients were obtained from GEO (Gene Expression Omnibus)[1]. The data were originally generated by Karlsson et al. (2014). They used the data to cluster the patients into five groups, and these groups showed different overall survival (Karlsson et al., 2014). We were more interested in how the methylation and expression differ from well-known subtypes, especially LUAD and SCLC. Therefore, we analyzed the 77 LUAD and 22 SCLC patients who had both methylation and expression data.

The methylation profiles were measured with Illumina HumanMethylation450 BeadChip while the gene expression profiles were measured with Illumina HumanHT-12 V4.0 expression BeadChip. The probe expression levels were averaged onto 20,178 genes. The 354,251 methylation sites within genes were analyzed. Therefore, each patient was represented with 20,178 genes and 354,251 methylation sites.

Screen for the Relevant Methylation and Expression Features

Since the number of methylation and expression features was very large, it was difficult to analyze directly. We applied the Boruta method (Kursa and Rudnicki, 2010) to screen the combined data and identify the relevant methylation and expression features. The Boruta method was based on random forest classification, and the relevance of features to sample classes was measured by the ensemble of the random forest classifier's stochasticity.

Evaluate the Importance of Relevant Methylation and Expression Features

After the irrelevant features were removed, the relevant methylation and expression features were ranked based on their importance evaluated with MCFS (Monte Carlo Feature Selection) (Draminski et al., 2008). The MCFS was a widely used method to rank features based on classification trees (Chen et al., 2018, 2019; Pan et al., 2018, 2019a,b; Li et al., 2019). First, for the d features, we selected s subsets and each subset included m features (m was much smaller than d). Then, for each subset, t trees were constructed. Based on the s × t trees, we can estimate a feature's importance by considering how many times it appeared in these trees and how well it performed in these trees as a node. By comparing the permutation results, the significance of features was evaluated.

Perdition Performance of the Mixed Methylation and Expression Signature

The MCFS can find the significant top-ranking features by comparing with permutations. To objectively evaluate the significant top-ranking features' prediction performance, we performed LOOCV (Leave One Out Cross Validation) using

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60645

SVM (Support Vector Machine) classifier (Li et al., 2018; Sun et al., 2018; Pan et al., 2019a). Each time, one sample was chosen as test samples and all other samples were used to train the SVM predictor. After all samples were tested once, we compared the actual sample classes with predicted sample classes and calculated the sensitivity, specificity, accuracy, and Mathew's correlation coefficient (MCC) based on the confusion matrix (Huang et al., 2011, 2013; Cai et al., 2012).

RESULTS AND DISCUSSION

Rank the Methylation and Expression Features

The methylation and gene expression data were combined and, therefore, each lung cancer patient was represented with mixed methylation and gene expression features. The number of mixed features (20,178 gene expression features and 354,251 methylation features) was too large to conduct sophisticated statistical analysis. So, we removed irrelevant features using the Boruta method (Kursa and Rudnicki, 2010). At last, 711 relevant features were remained.

Then, these 711 Boruta selected features were further ranked with the MCFS method (Draminski et al., 2008). As a classification tree-based ensemble learning algorithm, MCFS can rank the features based on how many times and how much it contributed to the sample classification in s × t trees. By comparing with permutation results, it can evaluate the significance of features.

Identify the Methylation and Expression Signature

The 136 significant top-ranking features were identified using the latest dmLab version 2.3.0 software downloaded from[2]

TABLE 1 | The 136 methylation and gene expression signature identified with the MCFS method.

Rank	Feature	Rank	Feature	Rank	Feature	Rank	Feature
1	DSC3	35	cg08796240	69	cg14487292	103	cg08621277
2	KRT5	36	cg08198430	70	cg03545620	104	cg13387113
3	cg02194717	37	cg10969178	71	DSG3	105	S1PR5
4	cg17814481	38	cg07838427	72	cg10991454	106	cg14769121
5	cg00415665	39	cg15958289	73	ANXA8L1	107	cg25634000
6	cg04432660	40	cg19445207	74	cg18736431	108	cg07417666
7	cg12932675	41	DLX5	75	cg14108894	109	cg18383680
8	cg13715502	42	cg26117023	76	cg17775621	110	cg11640015
9	cg08436756	43	cg16148454	77	cg15221831	111	cg02328660
10	cg02771299	44	cg13089599	78	cg26150462	112	cg08379517
11	cg06555468	45	cg00180559	79	cg11288202	113	cg04778236
12	cg13626676	46	cg21845794	80	cg27623451	114	cg11416243
13	KRT6C	47	cg26819757	81	cg02459569	115	cg18368125
14	cg01397507	48	cg03782130	82	cg24228306	116	cg09853371
15	SPRR2A	49	cg17005319	83	RORC	117	cg16260888
16	cg23613253	50	cg26795540	84	cg07538160	118	cg10842126
17	cg24235613	51	cg17957094	85	cg12448539	119	cg17094593
18	cg16969274	52	cg17543218	86	cg08774902	120	cg15335334
19	FAT2	53	cg13522118	87	cg04488647	121	KRT17
20	cg02579706	54	cg26431815	88	cg08190615	122	RFC4
21	TMEM63A	55	cg06332339	89	cg09470758	123	cg27009392
22	cg07568117	56	cg19883066	90	cg21922731	124	TP63
23	KRT6A	57	cg21013395	91	cg20197694	125	cg08327518
24	cg25922471	58	cg19526267	92	ACSL5	126	cg05800082
25	cg23628350	59	cg02634861	93	KRT6B	127	cg05128003
26	cg19032799	60	cg20803931	94	RAE1	128	cg04926361
27	cg04703476	61	cg05351785	95	cg24083274	129	cg01943337
28	cg01176141	62	cg21936454	96	cg23037777	130	cg06520450
29	cg12788467	63	cg03361585	97	cg07112556	131	cg15441535
30	cg24211826	64	cg20637223	98	cg26807301	132	cg25521254
31	MUC1	65	ANXA8	99	HNF1B	133	cg21176488
32	FMO5	66	cg15247247	100	cg18771553	134	cg05267427
33	cg06200607	67	cg06411879	101	cg18720506	135	cg05575304
34	VSNL1	68	cg10720966	102	cg04345366	136	cg20544605

[2] https://home.ipipan.waw.pl/m.draminski/mcfs.html

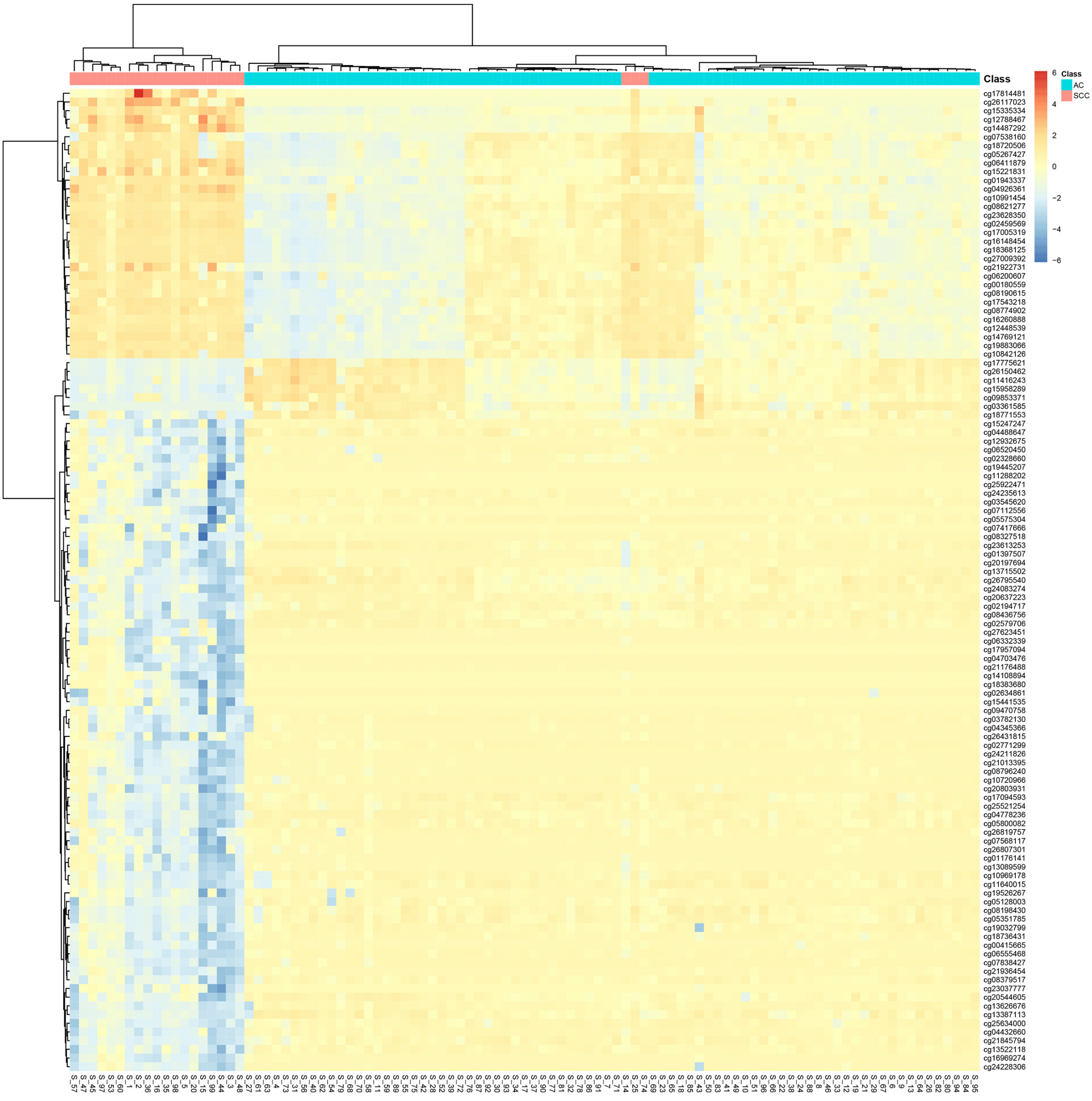


FIGURE 1 | The heatmap of LUAD and SCLC lung cancer patients with 113 methylation features. Almost all samples were correctly clustered using the 113 methylation features and only three SCLC samples were misclassified.

with default parameters. These 136 methylation and expression signatures are given in **Table 1**.

It can be seen that within these 136 signature features, there were 113 methylation features and 23 gene expression features. The annotations of the 113 methylation features based on GPL13534[3] are provided in **Supplementary Table S1**. We plotted the heatmaps of LUAD and SCLC lung cancer patients with 113 methylation features and 23 gene expression features in **Figures 1**, **2**, respectively. Both the 113 methylation features and 23 gene expression features can successfully group almost all samples with only three misclassified SCLC samples. They did not show difference on cluster results.

To more objectively and carefully compare the performance of the 113 methylation features and 23 gene expression features, we conducted LOOCV with SVM classifier. The LOOCV prediction performances of the 136 mixed features, 113 methylation features and 23 gene expression features are listed in **Tables 2–4**. It can be seen that the prediction results of 113 methylation features

[3] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534

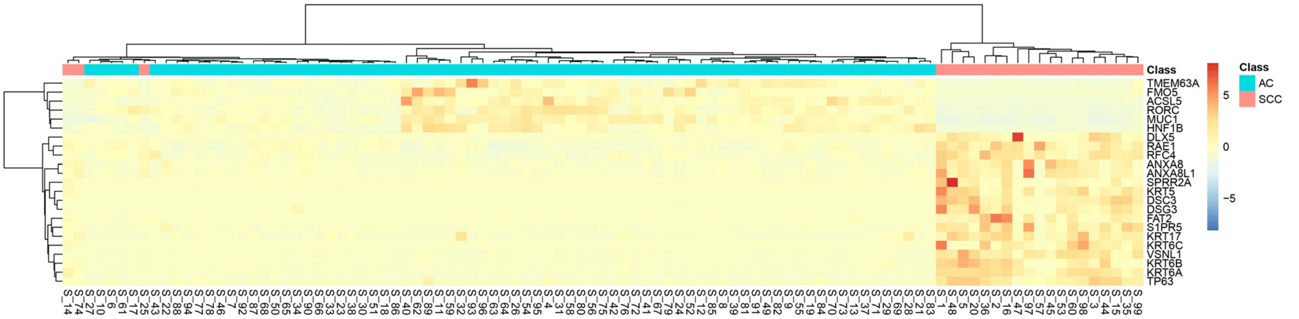


FIGURE 2 | The heatmap of LUAD and SCLC lung cancer patients with 23 gene expression features. Almost all samples were correctly clustered using the 23 gene expression features and only three SCLC samples were misclassified.

TABLE 2 | The confusion matrix using 136 mixed methylation and gene expression features.

	Actual LUAD	Actual SCLC
Predicted LUAD	77	2
Predicted SCLC	0	20
Performance Measurements	Sensitivity: 1.000, specificity: 0.909, accuracy: 0.980, MCC: 0.941	

TABLE 3 | The confusion matrix using 113 methylation features.

	Actual LUAD	Actual SCLC
Predicted LUAD	77	2
Predicted SCLC	0	20
Performance Measurements	Sensitivity: 1.000, specificity: 0.909, accuracy: 0.980, MCC: 0.941	

TABLE 4 | The confusion matrix using 23 gene expression features.

	Actual LUAD	Actual SCLC
Predicted LUAD	77	3
Predicted SCLC	0	19
Performance Measurements	Sensitivity: 1.000, specificity: 0.864, accuracy: 0.970, MCC: 0.912	

were the same as the 136 mixed features and better than the 23 gene expression features. The 23 gene expression features had one more misclassified SCLC samples. It seemed that methylation had better performance.

Comparison With CNV Signature

Comparing with the 136 LUAD and SQCLC CNV signatures identified by Li et al. (2014), we found that the methylated genes HORMAD2, KLHL3, LPP, and PTPN3 are also CNAs genes. HORMAD2 is expressed in nearly 10% of Chinese Han lung cancer tissues, which is a new target for lung cancer research (Liu et al., 2012). Lipoma preferred partner (LPP) may be an important candidate molecular marker for the classification of NSCLC tissue subtypes. PTPN3 can inhibit lung cancer by regulating EGFR signal (Li et al., 2015). However, there are no reports of KLHL3 in lung cancer, which also suggests a new idea of candidate molecular markers for the identification of lung cancer subtypes.

The Relationship Between Methylation and Expression Signature Genes

The 113 methylation features can be mapped onto 93 genes. We overlapped the selected methylation feature genes and expression feature genes and found that HNF1B and TP63 were dysfunctional on both methylation and gene expression levels. HNF1B was one of the DNA methylated markers of the same subtype (Matsuo et al., 2014; Shi et al., 2017). TP63, also known as P63, was considered to be the most common marker for SCLC (Bishop et al., 2012; Van de Laar et al., 2014).

We downloaded the 66 lung cancer genes from KEGG hsa05223 NSCLC[4] and mapped them and the overlapped two genes: HNF1B and TP63, onto STRING network (Szklarczyk et al., 2018). TP63 interacted with 39 KEGG lung cancer genes: AKT1, AKT3, ALK, BAK1, BAX, CASP9, CCND1, CDK4, CDK6, CDKN1A, CDKN2A, DDB2, E2F1, E2F2, E2F3, EGF, EGFR, EML4, ERBB2, FHIT, FOXO3, GADD45A, GRB2, HRAS, KRAS, MAP2K1, MAPK1, MAPK3, NRAS, PIK3CA, PIK3CB, PIK3R1, RB1, STAT3, STAT5A, STAT5B, STK4, TGFA, and TP53. HNF1B interacted with 14 KEGG lung cancer genes: AKT1, AKT2, CCND1, CDKN1A, CDKN2A, EGF, HRAS, KRAS, MAPK1, MAPK3, PIK3CA, RXRA, STAT3, and TP53.

What's more, we searched the methylation genes and expression genes in STRING database (Szklarczyk et al., 2018) and extracted the experimentally determined interaction and plotted the network in **Figure 3**. The light-yellow nodes were methylation genes, the light-blue nodes were expression genes. The overlapped methylation and expression genes were marked in red, the overlapped methylation and CNV genes from Li et al. (2014) were marked in pink. It can be seen that TP63 played an important role in connecting methylation genes and expression genes. The methylation genes and expression genes were closely connected to form a dense functional module on the network.

[4] https://www.genome.jp/dbget-bin/www_bget?pathway+hsa05223

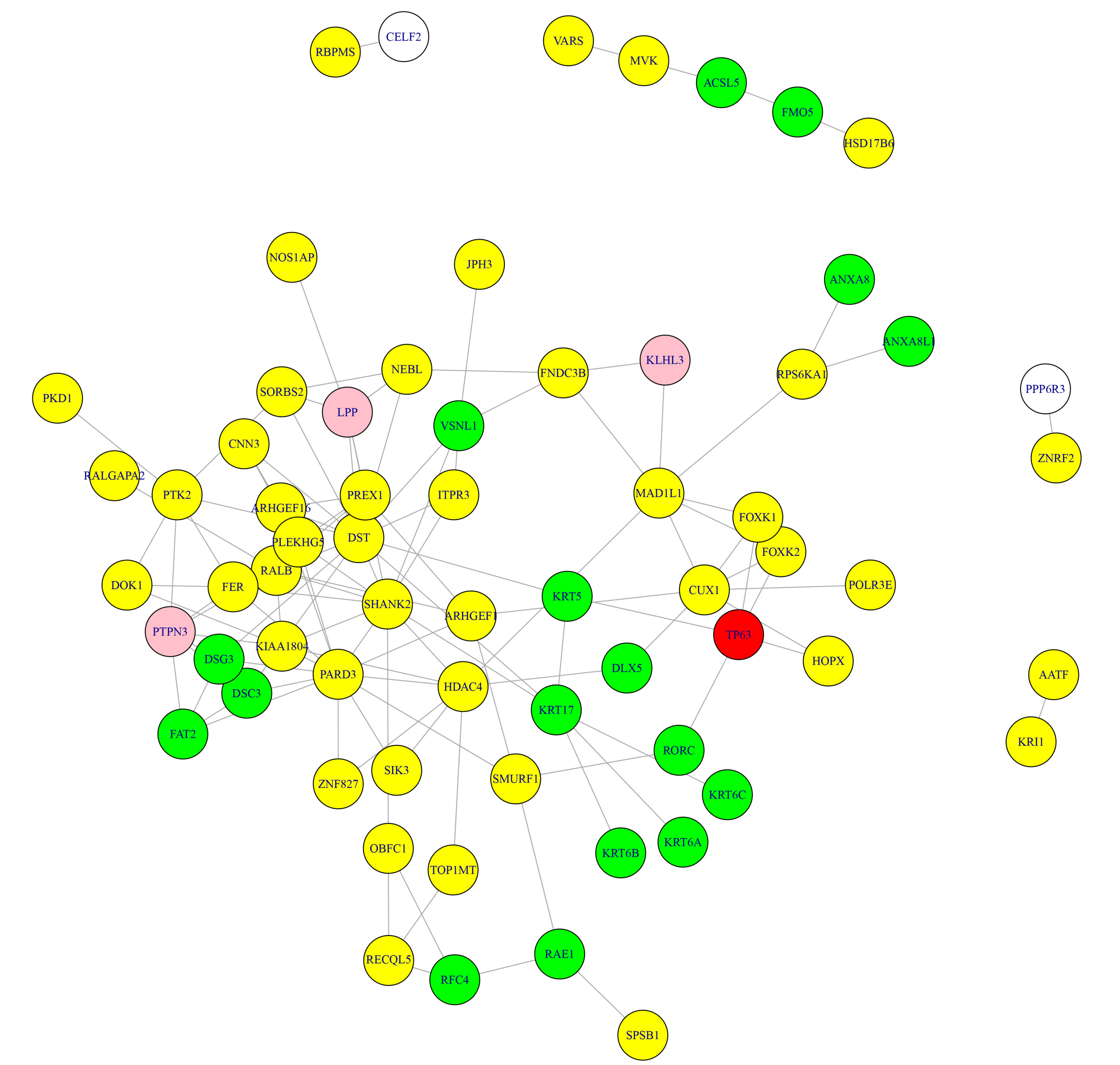


FIGURE 3 | The methylation genes and expression genes with experimentally determined interactions on STRING network. The light-yellow nodes were methylation genes, and the light-blue nodes were expression genes. The overlapped methylation and expression genes were marked in red, and the overlapped methylation and CNV genes were marked in pink. TP63 played an important role in connecting methylation genes and expression genes.

The Biological Significance of the Identified Signature

To develop more specific and individualized targeted therapy, there is an urgent need to improve our knowledge on the molecular basis, in addition to different phenotypes. It is noteworthy that adenocarcinoma and squamous cell carcinoma show marked differences in expression profiles, DNA methylation, and lesion location. In this study, the features containing methylation and expression data were screened by Boruta and then further sorted by MCFS. After comparing the selected features with related literatures, a certain correlation was found between these features and lung cancer subtypes.

In this study, 113 methylation features were screened and mapped to 93 genes. We inquired about the functions of these genes and their relationship with lung cancer to discuss whether they have the potential as molecular markers to recognize LUAD and SQCLC. Many genes have been proved to promote or inhibit the progression of lung cancer. For instance, FOXK1 was expressed in many malignant tissues (Huang and Lee, 2004) and Ma et al. (2018) also found that FOXK1 plays a carcinogenic role in lung cancer. MAD1L1 is a checkpoint gene, with its mutation been proved to play a pathogenic role in lung cancer (Tsukasaki et al., 2001). Some genes have been reported to be related with the prognosis of NSCLC, such as HORMAD2 and ANO1. The overexpression of ANO1 is related to the high expression of EGFR, which can be used as a predictor of recurrence after NSCLC (He et al., 2017). In addition, according to Zhang et al. (2014) HORMAD2 gene polymorphism has great potential prognostic value in Chinese patients with NSCLC. Other genes are associated with NSCLC subtypes, such as another member of the FOX family, FOXK2, which was reported to be closely related to the overall survival of LUAD (Chen et al., 2017). DOK1 and HOPX were found to serve as lung tumor suppressors for LUAD (Berger et al., 2010; Chen et al., 2015). In the study of Zhou et al. (2017) the methylation locus of PARD3 gene was positively correlated with the expression of PARD3 and suppression of PARD3 intensified chemoresistance in LUAD cells. SFTA3 was found obviously overexpressed in LUAD, and its expression in LUAD and SQCLC was quite different. Therefore, the sensitivity and specificity of using SFTA3 to distinguish the two subtypes will be relatively high (Zhan et al., 2015). ARHGEF1 aliased p114RhoGEF and its expression might help to predict progression and survival of SQCLC patients (Song et al., 2013). Notably, LPP has multiple functions of actin binding protein and transcriptional coactivator (Kuriyama et al., 2016). Ngan et al. (2017) proved that the expression of LPP reduces the number of circulating tumor cells and inhibits lung cancer metastasis. Kang et al. (2009) used high-resolution array-CGH to find that the difference in genomic imbalance patterns between SQCLC and LUAD was most significant in 3q26.2-q29, while LPP (3q28) was significantly targeted in SQCLC, suggesting that LPP may be an attractive candidate molecular marker for histological subtype classification of NSCLC and may be involved in the pathogenesis of SQCLC.

We also investigated 23 expressed genes in lung cancer, and found that many studies clearly indicated that some genes were associated with LUAD or SQCLC. DSC3 (Han et al., 2014; Lv et al., 2015) and KRT5 (Xu et al., 2014; Travis et al., 2015) have been proved to be an effective marker of SQCLC. ANXA8 (Chao et al., 2006) and DSG3 (Savci-Heijink et al., 2009) were significantly over-expressed in SQCLC, and DSG3 could be an effective ancillary marker to identify SQCLC (Sanchez-Palencia et al., 2011; Gómez-Morales et al., 2013). VSNL1, also known as VILIP-1, was a tumor suppressor gene specific to SQCLC (Fu et al., 2008). KRT6A, KRT6B, and KRT6C, members of the keratin protein family, are specific to squamous cells and associated with epidermis of squamous epithelium (Fujii et al., 2002; Hawthorn et al., 2006; Chang et al., 2011). In addition, we also identified several genes primarily associated with LUAD. According to Balabko et al. (2014) RORC is a specific transcription factor in the tumor area of lung tissue in patients with LUAD. DLX5 (Kato et al., 2008; Balabko et al., 2014), MUC1 (Mashima et al., 2005; Molina-Pinelo et al., 2014), and KRT17 (Erdogan et al., 2009; Liu et al., 2018) were found to be overexpressed in LUAD.

TABLE 5 | The GO enrichment results of the identified signature.

GO Term	FDR	*P* value	Number of overlapped genes
GO:0070268 cornification	8.58E-05	5.39E-09	9
GO:0009913 epidermal cell differentiation	0.0109	1.42E-06	11
GO:0031424 keratinization	0.0109	2.05E-06	9
GO:0030216 keratinocyte differentiation	0.0109	2.73E-06	10
GO:0060429 epithelium development	0.0115	3.59E-06	20
GO:0030855 epithelial cell differentiation	0.0130	4.91E-06	15
GO:0043588 skin development	0.0172	7.57E-06	11
GO:0009888 tissue development	0.0202	1.01E-05	25
GO:0008544 epidermis development	0.0319	1.80E-05	11
GO:0005737 cytoplasm	0.0045	2.34E-06	79
GO:0005829 cytosol	0.0083	8.55E-06	46

The GO Enrichment Analysis of the Identified Signature

In order to further analyze the relationship between mixed characteristics and lung cancer, we carried out GO enrichment analysis. The results suggest that characteristic genes are mainly related to keratinization, epidermal cell differentiation, tissue development, and cytoplasm. The GO enriched results with FDR (False Discovery Rate) smaller than 0.05 are listed in **Table 5**. P63 appears to be useful in differentiating SQCLC from LUAD in small biopsies with no keratosis or glandular differentiation, helping to establish different treatments (Camilo et al., 2006). The expression of keratinocyte transglutaminase and cytokeratin 10 was measured as markers of squamous differentiation (Lokshin et al., 1999). Epidermal cell differentiation is related to EGFR signal pathway, which can inhibit the proliferation and metastasis of cancer cells, while EGFR mutation is largely limited to LUAD (Ladanyi and Pao, 2008). The expression of Promyelocytic leukemia zinc finger (PLZF) in SQCLC was weak or absent, which was significantly lower than that in LUAD (Xiao et al., 2015).

To sum up, most of the 113 methylated genes and 23 expressed genes we found are closely related to lung cancer, and some of them have the possibility of distinguishing SQCLC from LUAD, which is helpful for the targeted selection of lung cancer treatment and provide more research support for lung cancer molecular markers.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

HZ, ZJ, LC, and BZ contributed to the study design. HZ, ZJ, and LC conducted the literature search. HZ, ZJ, and BZ acquired the data. ZJ and LC wrote the manuscript. HZ and BZ performed the data analysis. All authors gave the final approval of the version to be submitted, read, and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00003/full#supplementary-material

TABLE S1 | The annotations of the 113 methylation features.

REFERENCES

Altorki, N. K., Markowitz, G. J., Gao, D., Port, J. L., Saxena, A., Stiles, B., et al. (2019). The lung microenvironment: an important regulator of tumour growth and metastasis. *Nat. Rev. Cancer* 19, 9–31. doi: 10.1038/s41568-018-0081-9

Balabko, L., Andreev, K., Burmann, N., Schubert, M., Mathews, M., Trufa, D. I., et al. (2014). Increased expression of the Th17-IL-6R/pSTAT3/BATF/RorγT-axis in the tumoural region of adenocarcinoma as compared to squamous cell carcinoma of the lung. *Sci. Rep.* 4:7396. doi: 10.1038/srep07396

Bass, A. J., Watanabe, H., Mermel, C. H., Yu, S., Perner, S., Verhaak, R. G., et al. (2009). SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. *Nat. Genet.* 41, 1238–1242. doi: 10.1038/ng.465

Berger, A. H., Niki, M., Morotti, A., Taylor, B. S., Socci, N. D., Viale, A., et al. (2010). Identification of DOK genes as lung tumor suppressors. *Nat. Genet.* 42, 216–223. doi: 10.1038/ng.527

Bishop, J. A., Benjamin, H., Cholakh, H., Chajut, A., Clark, D. P., and Westra, W. H. (2010). Accurate classification of non-small cell lung carcinoma using a novel microRNA-based approach. *Clin. Cancer Res.* 16, 610–619. doi: 10.1158/1078-0432.Ccr-09-2638

Bishop, J. A., Teruya-Feldstein, J., Westra, W. H., Pelosi, G., Travis, W. D., and Rekhtman, N. (2012). p40 (ΔNp63) is superior to p63 for the diagnosis of pulmonary squamous cell carcinoma. *Mod. Pathol.* 25, 405–415. doi: 10.1038/modpathol.2011.173

Bunn, P. A. Jr., Minna, J. D., Augustyn, A., Gazdar, A. F., Ouadah, Y., et al. (2016). Small cell lung cancer: can recent advances in biology and molecular biology be translated into improved outcomes? *J. Thorac. Oncol.* 11, 453–474. doi: 10.1016/j.jtho.2016.01.012

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 42, 1387–1395. doi: 10.1007/s00726-011-0835-0

Camilo, R., Capelozzi, V. L., Siqueira, S. A., and Del Carlo Bernardi, F. (2006). Expression of p63, keratin 5/6, keratin 7, and surfactant-A in non-small cell lung carcinomas. *Hum. Pathol.* 37, 542–546. doi: 10.1016/j.humpath.2005.12.019

Chang, H. H., Dreyfuss, J. M., and Ramoni, M. F. (2011). A transcriptional network signature characterizes lung cancer subtypes. *Cancer* 117, 353–360. doi: 10.1002/cncr.25592

Chao, A., Wang, T. H., Lee, Y. S., Hsueh, S., Chao, A. S., Chang, T. C., et al. (2006). Molecular characterization of adenocarcinoma and squamous carcinoma of the uterine cervix using microarray analysis of gene expression. *Int. J. Cancer* 119, 91–98. doi: 10.1002/ijc.21813

Chen, L., Huang, T., Zhang, Y. H., Jiang, Y., Zheng, M., and Cai, Y. D. (2016). Identification of novel candidate drivers connecting different dysfunctional levels for lung adenocarcinoma using protein-protein interactions and a shortest path approach. *Sci. Rep.* 6:29849. doi: 10.1038/srep29849

Chen, L., Li, J., Zhang, Y. H., Feng, K., Wang, S., Zhang, Y., et al. (2018). Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell. Biochem.* 119, 3394–3403. doi: 10.1002/jcb.26507

Chen, L., Pan, X., Zhang, Y.-H., Kong, X., Huang, T., and Cai, Y.-D. (2019). Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* 120, 7068–7081. doi: 10.1002/jcb.27977

Chen, S., Jiang, S., Hu, F., Xu, Y., Wang, T., and Mei, Q. (2017). Foxk2 inhibits non-small cell lung cancer epithelial-mesenchymal transition and proliferation through the repression of different key target genes. *Oncol. Rep.* 37, 2335–2347. doi: 10.3892/or.2017.5461

Chen, Y., Yang, L., Cui, T., Pacyna-Gengelbach, M., and Petersen, I. (2015). HOPX is methylated and exerts tumour-suppressive function through Ras-induced senescence in human lung cancer. *J. Pathol.* 235, 397–407. doi: 10.1002/path.4469

Dearden, S., Stevens, J., Wu, Y. L., and Blowers, D. (2013). Mutation incidence and coincidence in non small-cell lung cancer: meta-analyses by ethnicity and histology (mutMap). *Ann. Oncol.* 24, 2371–2376. doi: 10.1093/annonc/mdt205

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2008). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Erdogan, E., Klee, E. W., Thompson, E. A., and Fields, A. P. (2009). Meta-analysis of oncogenic protein kinase Ciota signaling in lung adenocarcinoma. *Clin. Cancer Res.* 15, 1527–1533. doi: 10.1158/1078-0432.Ccr-08-2459

Fu, J., Fong, K., Bellacosa, A., Ross, E., Apostolou, S., Bassi, D. E., et al. (2008). VILIP-1 downregulation in non-small cell lung carcinomas: mechanisms and prediction of survival. *PLoS One* 3:e1698. doi: 10.1371/journal.pone.0001698

Fujii, T., Dracheva, T., Player, A., Chacko, S., Clifford, R., Strausberg, R. L., et al. (2002). A preliminary transcriptome map of non-small cell lung cancer. *Cancer Res.* 62, 3340–3346.

Gómez-Morales, M., Cámara-Pulido, M., Miranda-León, M. T., Sánchez-Palencia, A., Boyero, L., Gómez-Capilla, J. A., et al. (2013). Differential immunohistochemical localization of desmosomal plaque-related proteins in non-small-cell lung cancer. *Histopathology* 63, 103–113. doi: 10.1111/his.12126

Han, F., Dong, Y., Liu, W., Ma, X., Shi, R., Chen, H., et al. (2014). Epigenetic regulation of sox30 is associated with testis development in mice. *PLoS One* 9:e97203. doi: 10.1371/journal.pone.0097203

Hawthorn, L., Stein, L., Panzarella, J., Loewen, G. M., and Baumann, H. (2006). Characterization of cell-type specific profiles in tissues and isolated cells from squamous cell carcinomas of the lung. *Lung Cancer* 53, 129–142. doi: 10.1016/j.lungcan.2006.04.015

He, Y., Li, H., Chen, Y., Li, P., Gao, L., Zheng, Y., et al. (2017). Expression of anoctamin 1 is associated with advanced tumor stage in patients with non-small cell lung cancer and predicts recurrence after surgery. *Clin. Transl. Oncol.* 19, 1091–1098. doi: 10.1007/s12094-017-1643-0

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454. doi: 10.1038/nature25183

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Wu, Y. L., et al. (2017). Lung cancer: current therapies and new targeted treatments. *Lancet* 389, 299–311. doi: 10.1016/S0140-6736(16)30958-8

Huang, J. T., and Lee, V. (2004). Identification and characterization of a novel human FOXK1 gene in silico. *Int. J. Oncol.* 25, 751–757. doi: 10.3892/ijo.25.3.751

Huang, T., He, Z. S., Cui, W. R., Cai, Y. D., Shi, X. H., Hu, L. L., et al. (2013). A sequence-based approach for predicting protein disordered regions. *Protein Pept. Lett.* 20, 243–248. doi: 10.2174/0929866511320030002

Huang, T., Jiang, M., Kong, X., and Cai, Y. D. (2012). Dysfunctions associated with methylation, MicroRNA expression and gene expression in lung cancer. *PLoS One* 7:e43441. doi: 10.1371/journal.pone.0043441

Huang, T., Niu, S., Xu, Z., Huang, Y., Kong, X., Cai, Y. D., et al. (2011). Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. *PLoS One* 6:e22940. doi: 10.1371/journal.pone.0022940

Huang, T., Yang, J., and Cai, Y.-D. (2015). Novel candidate key drivers in the integrative network of genes, MicroRNAs, methylations, and copy number variations in squamous cell lung carcinoma. *BioMed Res. Int.* 2015:358125. doi: 10.1155/2015/358125

Kang, J. U., Koo, S. H., Kwon, K. C., Park, J. W., and Kim, J. M. (2009). Identification of novel candidate target genes, including EPHB3, MASP1 and SST at 3q26.2-q29 in squamous cell carcinoma of the lung. *BMC Cancer* 9:237. doi: 10.1186/1471-2407-9-237

Karlsson, A., Jonsson, M., Lauss, M., Brunnstrom, H., Jonsson, P., Borg, A., et al. (2014). Genome-wide DNA methylation analysis of lung carcinoma reveals one neuroendocrine and four adenocarcinoma epitypes associated with patient outcome. *Clin. Cancer Res.* 20, 6127–6140. doi: 10.1158/1078-0432.Ccr-14-1087

Kato, T., Sato, N., Takano, A., Miyamoto, M., Nishimura, H., Tsuchiya, E., et al. (2008). Activation of placenta-specific transcription factor distal-less homeobox 5 predicts clinical outcome in primary lung cancer patients. *Clin. Cancer Res.* 14, 2363–2370. doi: 10.1158/1078-0432.Ccr-07-1523

Kim, C. F., Jackson, E. L., Woolfenden, A. E., Lawrence, S., Babar, I., Vogel, S., et al. (2005). Identification of bronchioalveolar stem cells in normal lung and lung cancer. *Cell* 121, 823–835. doi: 10.1016/j.cell.2005.03.032

Koh, W. J., Greer, B. E., Abu-Rustum, N. R., Campos, S. M., Cho, K. R., Chon, H. S., et al. (2017). Vulvar cancer, version 1.2017, NCCN clinical practice guidelines in oncology. *J. Natl. Compr. Canc. Netw.* 15, 92–120.

Kuriyama, S., Yoshida, M., Yano, S., Aiba, N., Kohno, T., Minamiya, Y., et al. (2016). LPP inhibits collective cell migration during lung cancer dissemination. *Oncogene* 35, 952–964. doi: 10.1038/onc.2015.155

Kursa, M., and Rudnicki, W. (2010). Feature selection with the Boruta Package. *J. Stat. Softw. Artic.* 36, 1–13. doi: 10.18637/jss.v036.i11

Ladanyi, M., and Pao, W. (2008). Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. *Mod. Pathol.* 21(Suppl. 2), S16–S22. doi: 10.1038/modpathol.3801018

Li, B. Q., You, J., Chen, L., Zhang, J., Zhang, N., Li, H. P., et al. (2013). Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network. *BioMed Res. Int.* 2013:267375. doi: 10.1155/2013/267375

Li, B. Q., You, J., Huang, T., and Cai, Y. D. (2014). Classification of non-small cell lung cancer based on copy number alterations. *PLoS One* 9:e88300. doi: 10.1371/journal.pone.0088300

Li, J., Lan, C.-N., Kong, Y., Feng, S.-S., and Huang, T. (2018). Identification and analysis of blood gene expression signature for osteoarthritis with Advanced feature selection methods. *Front. Genet.* 9:246. doi: 10.3389/fgene.2018.00246

Li, J., Lu, L., Zhang, Y. H., Xu, Y., Liu, M., Feng, K., et al. (2019). Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther.* doi: 10.1038/s41417-019-0105-y

Li, M. Y., Lai, P. L., Chou, Y. T., Chi, A. P., Mi, Y. Z., Khoo, K. H., et al. (2015). Protein tyrosine phosphatase PTPN3 inhibits lung cancer cell proliferation and migration by promoting EGFR endocytic degradation. *Oncogene* 34, 3791–3803. doi: 10.1038/onc.2014.312

Liu, C., Zhang, Y. H., Huang, T., and Cai, Y. (2017). Identification of transcription factors that may reprogram lung adenocarcinoma. *Artif. Intell. Med.* 83, 52–57. doi: 10.1016/j.artmed.2017.03.010

Liu, J., Liu, L., Cao, L., and Wen, Q. (2018). Keratin 17 promotes lung adenocarcinoma progression by enhancing cell proliferation and invasion. *Med. Sci. Monit.* 24, 4782–4790. doi: 10.12659/msm.909350

Liu, M., Chen, J., Hu, L., Shi, X., Zhou, Z., Hu, Z., et al. (2012). HORMAD2/CT46.2, a novel cancer/testis gene, is ectopically expressed in lung cancer tissues. *Mol. Hum. Reprod.* 18, 599–604. doi: 10.1093/molehr/gas033

Lokshin, A., Zhang, H., Mayotte, J., Lokshin, M., and Levitt, M. L. (1999). Early effects of retinoic acid on proliferation, differentiation and apoptosis in non-small cell lung cancer cell lines. *Anticancer Res.* 19, 5251–5254.

Lu, C., Chen, H., Shan, Z., and Yang, L. (2016). Identification of differentially expressed genes between lung adenocarcinoma and lung squamous cell carcinoma by gene expression profiling. *Mol. Med. Rep.* 14, 1483–1490. doi: 10.3892/mmr.2016.5420

Lv, J., Zhu, P., Yang, Z., Li, M., Zhang, X., Cheng, J., et al. (2015). PCDH20 functions as a tumour-suppressor gene through antagonizing the Wnt/β-catenin signalling pathway in hepatocellular carcinoma. *J. Viral Hepat.* 22, 201–211. doi: 10.1111/jvh.12265

Ma, X., Yang, X., Bao, W., Li, S., Liang, S., and Sun, Y. (2018). Circular RNA circMAN2B2 facilitates lung cancer cell proliferation and invasion via miR-1275/FOXK1 axis. *Biochem. Biophys. Res. Commun.* 498, 1009–1015. doi: 10.1016/j.bbrc.2018.03.105

Mashima, T., Oh-hara, T., Sato, S., Mochizuki, M., Sugimoto, Y., Yamazaki, K., et al. (2005). p53-defective tumors with a functional apoptosome-mediated pathway: a new therapeutic target. *J. Natl. Cancer Inst.* 97, 765–777. doi: 10.1093/jnci/dji133

Matsuo, T., Dat le, T., Komatsu, M., Yoshimaru, T., Daizumoto, K., and Sone, S. (2014). Early growth response 4 is involved in cell proliferation of small cell lung cancer through transcriptional activation of its downstream genes. *PLoS One* 9:e113606. doi: 10.1371/journal.pone.0113606

Mazieres, J., Peters, S., Lepage, B., Cortot, A. B., Barlesi, F., Beau-Faller, M., et al. (2013). Lung cancer that harbors an HER2 mutation: epidemiologic characteristics and therapeutic perspectives. *J. Clin. Oncol.* 31, 1997–2003. doi: 10.1200/jco.2012.45.6095

Molina-Pinelo, S., Gutiérrez, G., Pastor, M. D., Hergueta, M., Moreno-Bueno, G., García-Carbonero, R., et al. (2014). MicroRNA-dependent regulation of transcription in non-small cell lung cancer. *PLoS One* 9:e90524. doi: 10.1371/journal.pone.0090524

Ngan, E., Stoletov, K., Smith, H. W., Common, J., Muller, W. J., Lewis, J. D., et al. (2017). LPP is a Src substrate required for invadopodia formation and efficient breast cancer lung metastasis. *Nat. Commun.* 8:15059. doi: 10.1038/ncomms15059

Oser, M. G., Niederst, M. J., Sequist, L. V., and Engelman, J. A. (2015). Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. *Lancet Oncol.* 16, e165–e172. doi: 10.1016/s1470-2045(14)71180-5

Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., and Kong, X. Y. (2019a). Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms. *Int. J. Mol. Sci.* 20:2185. doi: 10.3390/ijms20092185

Pan, X., Hu, X., Zhang, Y.-H., Chen, L., Zhu, L., Wan, S., et al. (2019b). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294, 95–110. doi: 10.1007/s00438-018-1488-4

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., and Chen, L. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes (Basel)* 9:208. doi: 10.3390/genes9040208

Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R., et al. (2011). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer* 129, 355–364. doi: 10.1002/ijc.25704

Savci-Heijink, C. D., Kosari, F., Aubry, M. C., Caron, B. L., Sun, Z., Yang, P., et al. (2009). The role of desmoglein-3 in the diagnosis of squamous cell carcinoma of the lung. *Am. J. Pathol.* 174, 1629–1637. doi: 10.2353/ajpath.2009.080778

Shi, Y. X., Wang, Y., Li, X., Zhang, W., Zhou, H. H., Yin, J. Y., et al. (2017). Genome-wide DNA methylation profiling reveals novel epigenetic signatures in squamous cell lung cancer. *BMC Genomics* 18:901. doi: 10.1186/s12864-017-4223-3

Shigematsu, H., Lin, L., Takahashi, T., Nomura, M., Suzuki, M., Wistuba, I. I., et al. (2005). Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J. Natl. Cancer Inst.* 97, 339–346. doi: 10.1093/jnci/dji055

Socinski, M. A., Obasaju, C., Gandara, D., Hirsch, F. R., Bonomi, P., Bunn, P., et al. (2016). Clinicopathologic features of advanced squamous NSCLC. *J. Thorac. Oncol.* 11, 1411–1422. doi: 10.1016/j.jtho.2016.05.024

Soldera, S. V., and Leighl, N. B. (2017). Update on the treatment of metastatic squamous non-small cell lung cancer in new era of personalized medicine. *Front. Oncol.* 7:50. doi: 10.3389/fonc.2017.00050

Song, C., Gao, Y., Tian, Y., Han, X., Chen, Y., and Tian, D. L. (2013). Expression of p114RhoGEF predicts lymph node metastasis and poor survival of squamous-cell lung carcinoma patients. *Tumour. Biol.* 34, 1925–1933. doi: 10.1007/s13277-013-0737-8

Sun, S., Schiller, J. H., and Gazdar, A. F. (2007). Lung cancer in never smokers–a different disease. *Nat. Rev. Cancer* 7, 778–790. doi: 10.1038/nrc2190

Sun, X., Li, J., Gu, L., Wang, S., Zhang, Y., Huang, T., et al. (2018). Identifying the characteristics of the hypusination sites using SMOTE and SVM algorithm with feature selection. *Curr. Proteom.* 15, 111–118. doi: 10.2174/1570164614666171109120615

Swanton, C., and Govindan, R. (2016). Clinical implications of genomic discoveries in lung cancer. *N. Engl. J. Med.* 374, 1864–1873. doi: 10.1056/NEJMra1504688

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., and Huerta-Cepas, J. (2018). 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

Thomas, J. K., Kim, M. S., Balakrishnan, L., Nanjappa, V., Raju, R., Marimuthu, A., et al. (2014). Pancreatic cancer database: an integrative resource for pancreatic cancer. *Cancer Biol. Ther.* 15, 963–967. doi: 10.4161/cbt.29188

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., et al. (2015). The 2015 world health organization classification of lung tumors: impact of genetic. clinical and radiologic advances since the 2004 classification. *J. Thorac. Oncol.* 10, 1243–1260.

Tsukasaki, K., Miller, C. W., Greenspun, E., Eshaghian, S., Kawabata, H., and Fujimoto, T. (2001). Mutations in the mitotic check point gene, MAD1L1, in human cancers. *Oncogene* 20, 3301–3305. doi: 10.1038/sj.onc.1204421

Van de Laar, E., Clifford, M., Hasenoeder, S., Kim, B. R., Wang, D., Lee, S., et al. (2014). Cell surface marker profiling of human tracheal basal cells reveals distinct subpopulations, identifies MST1/MSP as a mitogenic signal, and identifies new biomarkers for lung squamous cell carcinomas. *Respir. Res.* 15:160. doi: 10.1186/s12931-014-0160-8

Villalobos, P., and Wistuba, I. I. (2017). Lung cancer biomarkers. *Hematol. Oncol. Clin. North Am.* 31, 13–29. doi: 10.1016/j.hoc.2016.08.006

Weiss, J., Sos, M. L., Seidel, D., Peifer, M., Zander, T., and Heuckmann, J. M. (2010). Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer. *Sci. Transl. Med.* 2:62ra93. doi: 10.1126/scitranslmed.3001451

Wilkerson, M. D., Yin, X., Hoadley, K. A., Liu, Y., Hayward, M. C., Cabanski, C. R., et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* 16, 4864–4875. doi: 10.1158/1078-0432.ccr-10-0199

Xiao, G. Q., Li, F., Findeis-Hosey, J., Hyrien, O., Unger, P. D., and Xiao, L. (2015). Down-regulation of cytoplasmic PLZF correlates with high tumor grade and tumor aggression in non-small cell lung carcinoma. *Hum. Pathol.* 46, 1607–1615. doi: 10.1016/j.humpath.2015.06.021

Xu, C., Fillmore, C. M., Koyama, S., Wu, H., Zhao, Y., Chen, Z., et al. (2014). Loss of Lkb1 and Pten leads to lung squamous cell carcinoma with elevated PD-L1 expression. *Cancer Cell* 25, 590–604. doi: 10.1016/j.ccr.2014.03.033

Zhan, C., Yan, L., Wang, L., Sun, Y., Wang, X., Lin, Z., et al. (2015). Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. *J. Thorac. Dis.* 7, 1398–1405. doi: 10.3978/j.issn.2072-1439.2015.07.25

Zhang, K., Tang, S., Cao, S., Hu, L., Pan, Y., and Ma, H. (2014). Association of polymorphisms at HORMAD2 and prognosis in advanced non-small-cell lung cancer patients. *Cancer Epidemiol.* 38, 414–418. doi: 10.1016/j.canep.2014.03.013

Zhou, Q., Dai, J., Chen, T., Dada, L. A., Zhang, X., Zhang, W., et al. (2017). Downregulation of PKCζ/Pard3/Pard6b is responsible for lung adenocarcinoma cell EMT and invasion. *Cell. Signal.* 38, 49–59. doi: 10.1016/j.cellsig.2017.06.016

Zhou, Y., Wu, K., Jiang, J., Huang, J., Zhang, P., and Zhu, Y. (2015). Integrative analysis reveals enhanced regulatory effects of human long intergenic non-coding RNAs in lung adenocarcinoma. *J. Genet. Genomics* 42, 423–436. doi: 10.1016/j.jgg.2015.07.001

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: 07 February 2020
doi: 10.3389/fgene.2019.01367

Analysis of the Clinicopathologic Characteristics of Lung Adenocarcinoma With *CTNNB1* Mutation

*Chao Zhou[1,2], Wentao Li[2], Jinchen Shao[3], Jikai Zhao[3] and Chang Chen[1]**

[1] Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China, [2] Department of Thoracic Surgery, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China, [3] Department of Pathology, Shanghai Chest Hospital, Shanghai Jiao Tong University, Shanghai, China

Introduction: Lung adenocarcinoma with *CTNNB1* mutation is relatively uncommon, and its clinicopathologic characteristics, disease course, and prognosis have not been well-studied.

Methods: A total of 564 lung adenocarcinoma patients were enrolled in this study. The relationship between *CTTNB1* mutational status and clinicopathologic parameters, the rates of relapse-free survival (RFS) and overall survival (OS), and the mutational status of other genes commonly mutated in lung adenocarcinoma were analyzed.

Results: Of 564 lung adenocarcinoma patients, 30 (5.3%) harbored *CTNNB1* mutations. Univariate analyses revealed that gender, smoking history, pleural invasion, and histological subtype were all significant predictors of RFS and OS. Pleural invasion and histological subtype remained significant predictors of RFS and OS in a multivariate analysis. There were no significant differences in RFS (p = 0.504) or OS (p = 0.054) between lung adenocarcinoma patients with *CTNNB1* mutation and those without *CTNNB1* mutation. However, patients with *CTNNB1* mutation tended to have a worse OS.

Conclusions: Female patients and nonsmokers are likely to harbor *CTNNB1* mutation and primary lung adenocarcinoma with mutated *CTNNB1* has a poor prognosis.

Keywords: adenocarcinoma, lung cancer, *CTNNB1*, mutation, prognosis

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Liang Liu,
Fudan University Shanghai Cancer Center, China
Deli Liu,
Cornell University, United States

***Correspondence:**
Chang Chen
chenthoracic@163.com

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

Received: *17 October 2019*
Accepted: *12 December 2019*
Published: *07 February 2020*

Citation:
Zhou C, Li W, Shao J, Zhao J and Chen C (2020) Analysis of the Clinicopathologic Characteristics of Lung Adenocarcinoma With CTNNB1 Mutation. Front. Genet. 10:1367. doi: 10.3389/fgene.2019.01367

INTRODUCTION

Lung cancer, which has the highest incidence of all cancers and the highest rate of disease-related fatalities, is the main cause of cancer-related death worldwide (Torre et al., 2015; Gu et al., 2017a; Gu et al., 2018). Lung adenocarcinoma is the most common pathological subtype, accounting for nearly 70% of all lung tumors (Sun et al., 2010). With the introduction of low-dose computed tomography, which enables earlier detection, the incidence of lung cancer, especially early-stage lung cancer, has risen sharply in recent years (Field et al., 2012).

With the advent of genomics, molecular or genetic variants affecting disease risk can be identified (Field, 2008). Mutations in the gene encoding β-catenin (*CTNNB1*) have been detected in numerous human malignancies, including lung cancer (Woenckhaus et al., 2008), malignant mesothelioma (Shigemitsu et al., 2001), desmoid tumors (Colombo et al., 2013), colon cancer (Akyol et al., 2019), and others. Woenckhaus et al. (2008) identified a number of differentially expressed genes in smoke-exposed bronchial epithelium and nonsmall cell lung cancers (NSCLCs), they found in adenocarcinomas, the cytoplasmic expression of beta-catenin was associated with shorter survival (p = 0.012). Shigemitsu et al. (2001) found *CTNNB1* is infrequently mutated in lung cancer. Akyol et al. (2019) defined an immunohistochemical algorithm to dissect Wnt pathway alterations in formalin-fixed and paraffin-embedded neoplastic tissues and found all six colon adenomas of the 126 total adenomas studied for the altered/mutant β-catenin staining pattern had presumptively pathogenic point mutations or deletions in CTNNB1. The N-terminus of β-catenin, with contains conserved phosphorylated threonine/serine amino acid residues, is the most frequent location of cancer-related *CTNNB1* mutations (Dar et al., 2017). The level of free β-catenin in the cytoplasmic pool is regulated by ubiquitination and proteasomal degradation (Akyol et al., 2019). β-catenin is a member of the Wnt signaling cascade and is associated with cadherin-mediated cell–cell adhesion systems (Woenckhaus et al., 2008). In lung tumors, the immunohistologic loss of β-catenin membrane staining along with a corresponding increase cytoplasmic or nuclear staining has been reported (Nozawa et al., 2006).

Although *CTNNB1* mutation occurs in many tumors types, it has not been well-studied in the context of lung adenocarcinoma, and the clinicopathologic characteristics and prognosis of lung adenocarcinoma with mutated *CTNNB1* has not been described. Therefore, we compared the clinicopathologic characteristics of 30 lung adenocarcinomas with *CTNNB1* mutations with those of 534 lung adenocarcinomas with wild-type *CTNNB1*.

MATERIALS AND METHODS

From July 2008 to April 2013, resected primary lung adenocarcinomas were collected at the Department of Thoracic Surgery of Shanghai Chest Hospital, Shanghai Jiaotong University. To confirm the diagnosis of primary lung cancer, all the patients received thorough preoperative testing at our hospital, including physical exams, serological tests, pulmonary function tests, chest/brain computed tomography (CT), technetium bone scanning, and abdominal ultrasound. Biopsies were done by bronchoscopy or endobronchial ultrasound-guided transbronchial needle aspiration, and in some cases, positron emission tomography CT was used to exclude mediastinal lymph node metastases (Gu et al., 2017b). The lung adenocarcinoma subtype was determined by light microscopy intraoperatively, using frozen sections, and confirmed postoperatively, using paraffin-embedded sections. All surgical samples had at least 5% tumor content. Each case was reviewed by at least two junior pathologists and a senior pathologist to confirm the histologic subtype of resected lung neoplasms. The combination of routine preoperative examination and intra-/postoperative pathological diagnosis is recommended to make an exact lung cancer diagnosis.

In total, 601 patients with primary lung adenocarcinoma were identified. Of these, 17 and 20 patients were excluded because they received neoadjuvant chemotherapy or were lost to follow-up, respectively. The remaining 564 patients were enrolled in this study.

Informed consent was given by all patients or their legal representatives. The study was initiated after obtaining Institutional Review Board approval. The medical records for all patients were reviewed to collect corresponding clinicopathologic data, including sex, age, smoking status, pathologic tumor, node, and metastasis (TNM) stage [according to the staging system of the 7th edition of the American Joint Committee on Cancer (Edge et al., 2010)], thyroid transcription factor-1 status, and treatment information. Data on disease recurrence and survival were obtained from follow-up clinic visits or by telephone.

Bioinformatics Analysis

Data of The Cancer Genome Atlas (TCGA) were analyzed by Gene Expression Profiling interactive Analysis (http://gepia.cancer-pku.cn/) and Kaplan–Meier Plotter (http://kmplot.com/analysis/index.php?p=service&cancer=lung). Gene *CTNNB1* were further analyzed by Gene Expression Profiling interactive Analysis and the survival curves were draw and compared by Kaplan–Meier Plotter.

Mutational Analysis

The mutational status of *EGFR*, *KRAS*, and *CTNNB1* was determined by targeted sequencing and verified by DNA sequencing analysis. Relevant primers were designed to amplify all known *ALK* fusion variants by quantitative real-time reverse transcriptase PCR of cDNA. *ALK* fluorescent *in situ* hybridization was used to confirm the presence of *ALK* gene fusions (Wang et al., 2012).

Statistical Analysis

Clinicopathologic data was analyzed using the SPSS 22.0 software package (SPSS Inc, Chicago, IL). Relapse-free survival (RFS) and overall survival (OS) were estimated by the Kaplan–Meier method, and differences were compared by log-rank testing using Prism 6 (GraphPad Software, La Jolla, CA). A p value of <0.05 was considered statistically significant.

RESULTS

Mutational Status of Lung Adenocarcinomas

Of the 564 lung adenocarcinoma patients examined, 30 (5.3%) harbored *CTNNB1* mutations (**Table 1**). The distributions of specific mutation types are shown in **Figure 1**.

TABLE 1 | Characteristics of lung adenocarcinoma with CTNNB1 mutation.

Cases	Gender	Age	Smoking	Subtype	Tumor size (cm)	Stage	CTNNB1 mutation	RFS (months)	OS (months)
1	F	57	Never smoker	A + P	3	2a	S45F	35.4	46.8
2	F	52	Never smoker	S + P	3	3a	S45F	6.3	26.1
3	F	60	Never smoker	A + P	4.1	3a	D32Y	46.5	82+
4	F	59	Never smoker	P	2.8	3a	D32Y	3.6	22.3
5	F	44	Never smoker	S + A	3	3a	D32Y	22	56+
6	F	49	Never smoker	P + S + L	8.4	3a	S33C	3.2	16.8
7	M	59	Never smoker	L + A	1.9	1a	S37A	25.4	47+
8	M	65	Smoker	P	4.6	3a	S33C	25	68+
9	M	62	Smoker	P	2.4	1a	S37F	45+	45+
10	F	55	Never smoker	A + P	5	1b	S45P	12	43+
11	M	59	Smoker	IMA	5	3a	G34V	2.4	19
12	F	75	Never smoker	P + M	2.1	1a	S33Y	45+	45+
13	F	60	Never smoker	A + P	1.6	1b	S33C	63+	63+
14	F	74	Never smoker	P + M	4.3	1b	S37C	63+	63+
15	M	67	Smoker	A + P + M	2.1	2b	S37C	3.2	44
16	F	69	Never smoker	A + P + M	2.9	3a	S37F	16.8	29
17	F	70	Never smoker	A + M	1.7	1a	D32H	56+	56+
18	F	62	Never smoker	S + P	2.1	1b	S33F	6.4	10
19	F	55	Never smoker	A + L	2.8	3a	S37F	58+	58+
20	M	41	Never smoker	A + P	2.6	1a	S33C	62+	62+
21	F	59	Never smoker	P + A	4.5	1b	S37F	60+	60+
22	F	68	Never smoker	P + A + M	4.3	1b	G34R	4.8	13.8
23	F	72	Never smoker	S + P	2.1	1a	S45F	54+	54+
24	F	68	Never smoker	A + P	2.9	1b	S33C	19	29.4
25	F	59	Never smoker	A + P + M	2.4	1a	S33C	16	34.2
26	M	46	Never smoker	A + P	2.6	1a	G34R	15	35
27	F	70	Never smoker	P + A	2.1	1b	S37C	19.6	35.3
28	M	74	Never smoker	M	5.6	2a	S45P	56+	56+
29	F	60	Never smoker	A + S	3.8	1b	S33C	48+	48+
30	M	61	Smoker	P + M	4.6	1b	G34V	61+	61+

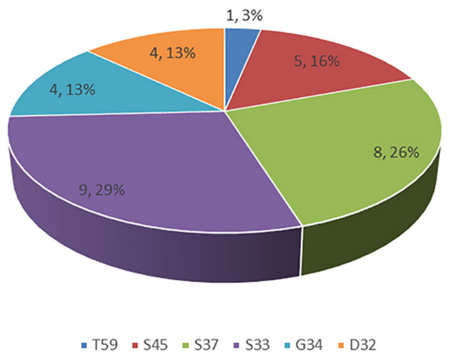


FIGURE 1 | Spectrum of CTNNB1 mutations in lung adenocarcinoma.

Relationship Between Clinicopathologic Factors and *CNTTB1* Mutational Status

Of the 30 patients with *CTNNB1* mutations, there were 21 (70%) female patients and 9 (30%) male patients, ranging in age from 22 to 81 years (median, 59 years). Histologically, 263 of the tumors were acinar-predominant (47%), 115 were micropapillary-predominant (20%), 94 were papillary-predominant (17%), 49 were lepidic-predominant (9%), 32 were mucinous adenocarcinoma-predominant (5%), and 11 were solid-predominant (2%). Most of the patients had early-stage lung cancer in both the *CNTTB1* mutation group (stage I: 18/30, 60%) and the *CNTTB1* wild-type group (stage I: 241/534, 45%). Age ($p = 0.851$), tumor size ($p = 0.256$), lymph node status ($p = 0.184$), pathologic stage ($p = 0.322$), and the presence of pleural invasion ($p = 0.459$) were similar between lung adenocarcinomas with *CTNNB1* mutation and lung adenocarcinomas without *CTNNB1* mutation, but the former group tended to have more female patients ($p < 0.001$) and more smokers ($p = 0.019$) (**Table 2**).

Relationship Between *CNTTB1* Mutational Status and Survival

Univariate analysis revealed that gender, smoking history, pleural invasion, and histological subtype were all significant predictors of RFS and OS (**Table 3**). Pleural invasion and histological subtype were still significant predictors of RFS and OS in a multivariate analysis (**Table 4**).

During follow-up, 19 (63.3%) patients with lung adenocarcinomas with mutated *CTNNB1* and 259 (48.5%) patients with lung adenocarcinomas with wild-type *CTNNB1* experienced a relapse, and 10 (33.3%) and 111 (20.8%) patients died, respectively. There were no statistically significant differences in RFS ($p = 0.504$) or OS ($p = 0.054$) between patients with *CTNNB1* mutation and patients without *CTNNB1* mutation (**Figure 2**). However, patients with *CTNNB1* mutation tended to have a worse OS.

TABLE 2 | Features of patients with lung adenocarcinoma harboring *CTNNB1* mutations.

	CTNNB1 mutation		**CTNNB1 wild type**		
	No.	**Percent**	**No.**	**Percent**	***p* value**
Total	30	5.3%	534	94.7%	
Sex					
Male	9	30%	259	48.5%	
Female	21	70%	275	51.5%	<0.001
Age					
≥60 years	14	47%	234	44%	
<60 years	16	53%	300	56%	0.851
Smoking status					
Smoker	5	17%	332	62%	
Never-smoker	25	83%	202	38%	0.019
Tumor size					
≤3c m	21	70%	314	59%	
>3 cm	9	30%	220	41%	0.256
Lymph Node status					
N0	21	70%	300	56%	
N1/2	9	30%	234	44%	0.184
Pathologic stage					
I	18	60%	241	45%	
II	3	10%	81	15%	
III	9	30%	187	35%	
IV	0	/	25	5%	0.322
Pleural invasion					
0	15	50%	299	56%	
1/2	15	50%	235	44%	0.459
Pathological subtype					
Lepidic	1	3%	48	8%	
Acinar	13	43%	250	47%	
Papillary	10	33%	84	16%	
Micropapillary	4	13%	111	21%	
Solid	1	3%	10	2%	
IMA	1	3%	31	6%	0.168
TTF1					
Positive	16	53%	339	63%	
Negative	14	47%	195	37%	0.331
EGFR					
Present	21	70%	314	59%	
Absent	9	30%	220	41%	0.256
KRAS					
Present	1	3.3%	56	10.5%	0.347
Absent	29	96.7%	478	89.5%	
ALK					
Present	2	7%	28	5%	
Absent	28	93%	506	95%	0.669

As for lung adenocarcinomas from TCGA, there was no significant differences in the distributions of CTNNB1 mRNA expression among different lung adenocarcinoma stages (**Supplementary Figure 1**). Besides, between lung adenocarcinoma patients with and without CTNNB1 mutation, there was no significant differences in RFS (p = 0.49), while significant differences were found in OS (p = 8.9e−05). (**Supplementary Figures 2** and **3**).

DISCUSSION

Lung cancer remains the leading cause of cancer-related death worldwide (Gu et al., 2017). Low-dose computed tomography

TABLE 3 | Independent predictors of overall survival.

Univariate analysis	**HR**	**95% CI**	***p* value**
Gender, male vs. female	1.706	1.194–2.438	0.003
Age	0.988	0.971–1,006	0.18
Smoke, never vs. ever	1.464	1.025–2.09	0.036
Pleural invasion, yes vs.no	0.671	0.4–0.814	0.002
Subtypes			
Lepidic	0.041	0.004–0.479	0.011
Acinar	0.801	0.561–1.142	0.801
Papillary	0.927	0.574–1.497	0.757
Micropapillary	0.438	0.061–3.134	0.411
Solid	2.918	2.021–4.213	0.0001
Invasive mucinous	0.726	0.267–1.97	0.529
EGFR mutation, no vs. yes	0.746	0.523–1.065	0.106
ALK, negative vs. positive	1.411	0.689–2.89	0.347
CTNNB1 mutation, yes vs.no	1.746	0.982–3.103	0.058
Multivariate analysis	**HR**	**95% CI**	***p* value**
Gender, male vs. female	1.995	1.183–3.367	0.01
Age	0.991	0.974–1.009	0.341
Smoke, never vs. ever	0.769	0.449–1.318	0.339
Pleural invasion, yes vs.no	0.8	0.668–0.957	0.015
Subtypes			
Lepidic	0.001	/	0.949
Acinar	1.321	0.456–3.826	0.608
Papillary	1.344	0.431–4.188	0.611
Micropapillary	0.641	0.067–6.13	0.7
Solid	3.247	1.117–9.439	0.031
Invasive mucinous	/	/	/
EGFR mutation, no vs. yes	1.14	0.745–1.744	0.547
ALK, negative vs. positive	1.494	0.665–3.358	0.331
CTNNB1 mutation, yes vs.no	1.784	0.981–3.247	0.058

TABLE 4 | Independent predictors of relapse-free survival.

Univariate analysis	**HR**	**95% CI**	***p* value**
Gender, male vs. female	1.706	1.194–2.438	0.003
Age	0.988	0.971–1,006	0.18
Smoke, never vs. ever	1.464	1.025–2.09	0.036
Pleural invasion, yes vs.no	0.671	0.4–0.814	0.002
Subtypes			
Lepidic	0.041	0.004–0.479	0.011
Acinar	0.801	0.561–1.142	0.801
Papillary	0.927	0.574–1.497	0.757
Micropapillary	0.438	0.061–3.134	0.411
Solid	2.918	2.021–4.213	0.0001
Invasive mucinous	0.726	0.267–1.97	0.529
EGFR mutation, no vs. yes	0.746	0.523–1.065	0.106
ALK, negative vs. positive	1.411	0.689–2.89	0.347
CTNNB1 mutation, yes vs.no	1.746	0.982–3.103	0.058
Multivariate analysis	**HR**	**95% CI**	***p* value**
Gender, male vs. female	1.127	0.76–1.673	0.552
Age	0.995	0.988–1.007	0.435
Smoke, never vs. ever	1.435	0.957–2.15	0.081
Pleural invasion, yes vs.no	0.78	0.692–0.88	< 0.001
Subtypes			
Lepidic	0.345	0.145–0.822	0.016
Acinar	0.997	0.54–1.839	0.992
Papillary	0.967	0.497–1.881	0.92
Micropapillary	1.45	0.516–4.077	0.481
Solid	1.731	0.929–3.224	0.084
Invasive mucinous	/	/	/
EGFR mutation, no vs. yes	1.19	0.889–1.592	0.243
ALK, negative vs. positive	1.159	0.641–2.095	0.626
CTNNB1 mutation, yes vs.no	1.206	0.737–1.974	0.457

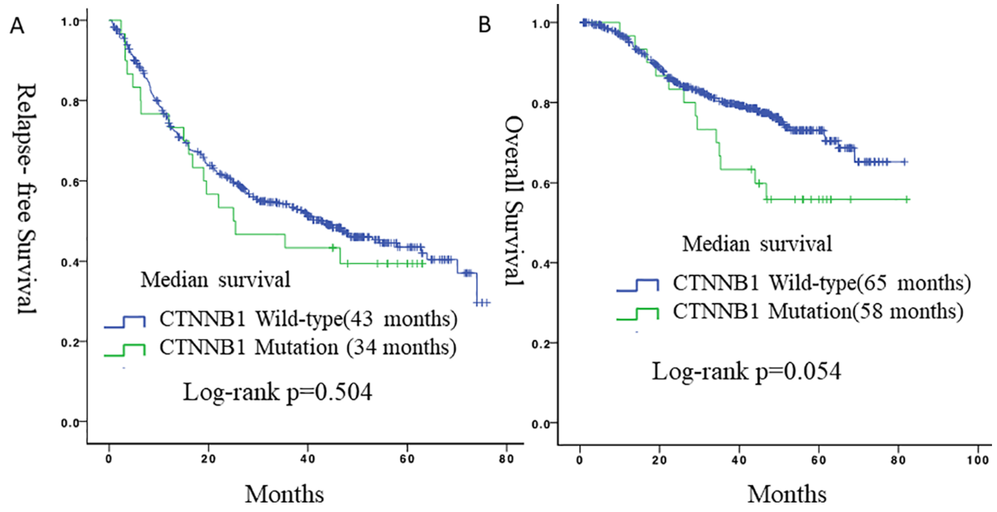


FIGURE 2 | Survival curves for relapse-free survival and overall survival according to CTNNB1 status. **(A)** Relapse-free survival between the two groups. **(B)** Overall survival between the two groups.

screening reduces the mortality of lung cancer by as much as 20% in high-risk patients (National Lung Screening Trial Research Team, 2011). Early detection and diagnosis increases the number of patients who are newly diagnosed with lung cancer while it is still early-stage, improving the prognosis of lung cancer patients as a whole. For individuals at a high risk of developing lung cancer, periodic screening could have a survival benefit. Recently, some lung cancer risk prediction models have been constructed to make lung cancer screening more efficient (Spitz et al., 2007; Raji et al., 2012). With the development of gene mutation testing, targeted therapy has changed the treatment strategy for lung cancer. In this study, we describe the clinicopathological characteristics of lung adenocarcinoma with *CTNNB1* mutation.

β-catenin is important for the establishment and maintenance of the epithelial layer and is a key downstream component of the canonical Wnt signaling pathway. The WNT/β-catenin pathway is involved in cancer and pluripotent stem cell signaling, which may suggest the mechanism underlying cancer stem cells. In this study, out of 564 patients, 30 (5.3%) patients with *CTNNB1* mutations were identified. Kase et al. (2000) conducted an immunohistochemical analysis of 331 lung cancer specimens and reported that β-catenin expression was reduced in 122 (37%) of the samples, which was associated with significantly worse patient survival. Similarly, Woenckhaus et al. (2008) reported that reduced membrane staining of β-catenin and its abnormal accumulation in the cytoplasm and/or nuclei of lung adenocarcinoma cells was associated with shorter survival (p = 0.012). Another study also suggested that reduced β-catenin expression in surgically resected non-small cell lung cancer specimens was associated with lymph node metastasis and a poor prognosis (Retera et al., 1998). These studies suggest that decreased expression of β-catenin is associated with an unfavorable prognosis in lung cancer.

In our study, during follow-up, 19 patients (63.3%) with lung adenocarcinomas with *CTNNB1* mutations and 259 patients (48.5%) with lung adenocarcinomas with wild-type *CTNNB1* relapsed, and 10 (33.3%) and 111 (20.8%) patients died, respectively. Patients with *CTNNB1* mutations therefore tended to have a worse prognosis, although this did not reach statistical significance. When compare with data from TCGA, patients with CTNNB1 mutation in TCGA also had worse OS. Our findings therefore correspond well to the results of previous studies and common directory (Sunaga et al., 2001).

In Cox proportional hazards models, univariate analyses revealed that gender, smoking history, the presence of pleural invasion, and histological subtype were all significant predictors of RFS and OS. Pleural invasion and histological subtype remained significant predictors of RFS and OS in a multivariate analysis. With respect to histological subtype, adenocarcinoma patients with micropapillary or solid subtypes, which are defined as high-risk subtypes in the 2011 classification proposed by the International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society (Travis et al., 2011), had significantly worse prognosis. As for pleural invasion, pleural invasion, as well as visceral invasion, is considered an aggressive and invasive factor in NSCLC and has been included in the TNM staging system as a factor that should upstage the T factor (Rami-Porta et al., 2007; Travis et al., 2008; Butnor and Travis, 2012). Shimizu et al. (2005) demonstrated that velopharyngeal insufficiency (VPI) is a significant and independent predictor of a poor prognosis regardless of tumor size or N status, and as a result, VPI is a good indicator of the degree of invasion and aggressiveness of NSCLC. As more early-staged lung neoplasms are detected, whether VPI has impact on survival of patients with early-staged lung cancer is unknown. Therefore, Jiang et al. (2015) published a meta-analysis and found VPI together with tumor size has a synergistic effect on survival in patients with N0 disease. Patients with stage IB NSCLC and larger tumor size with VPI might be considered for adjuvant chemotherapy after surgical resection and need careful preoperative evaluation and postoperative follow-up.

There are several limitations to this study. First, the sample size was relatively small. Contributing to the small sample size, there were several patients with *CTNNB1* gene mutations who could not be included in the data analysis because of incomplete clinicopathological records. Finally, the patients' outcomes could have been influenced by the use of different treatment strategies, which may confound the survival analysis.

In summary, our results suggest that female patients and nonsmokers are likely to harbor *CTNNB1* mutation and primary lung adenocarcinoma with mutated *CTNNB1* has a poor prognosis. Further research is needed to verify our results. However, these data suggest that β-catenin could be a potential therapeutic target for advanced-stage lung cancer.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics committee of Shanghai chest hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conception and design: CZ and CC. Development of methodology: CZ, WL, and CC. Acquisition of data (provided surgical samples, gene detection, pathological diagnosis, acquired and managed patients, provided facilities, etc.): CZ, WL, JS, JZ, and CC. Analysis and interpretation of data: CZ. Writing, review, and/or revision of the manuscript: CZ and CC.

FUNDING

This work was supported by Shanghai Pulmonary Hospital Innovation Team Program (fkcx1906).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01367/full#supplementary-material

REFERENCES

Akyol, A., Güner, G., Özşeker, H. S., Işık, A., Atcı, O., Fearon, E. R., et al. (2019). An immunohistochemical approach to detect oncogenic CTNNB1 mutations in primary neoplastic tissues. *Lab. Invest.* 99 (1), 128. doi: 10.1038/s41374-018-0121-9

Butnor, K. J., and Travis, W. D. (2012). Editorial comment: recent advances in our understanding of lung cancer visceral pleural invasion and other forms of minimal invasion: implications for the next TNM classification.

Colombo, C., Miceli, R., Lazar, A. J., Perrone, F., Pollock, R. E., Gronchi, A., et al. (2013). CTNNB1 45F mutation is a molecular prognosticator of increased postoperative primary desmoid tumor recurrence: an independent, multicenter validation study. *Cancer* 119 (20), 3696–3702. doi: 10.1002/cncr.28271

Dar, M. S., Singh, P., Mir, R. A., et al. (2017). Beta-catenin N-terminal domain: an enigmatic region prone to cancer causing mutations. *Mutat. Res. Rev. Mutat. Res.* 773, 122–133. doi: 10.1016/j.mrrev.2017.06.001

Edge, S. B., Byrd, D. R., and Compton, C. C. (2010). *Cancer staging manual. American Joint Committee on Cancer (AJCC). 7th ed.* (New York: Springer).

Field, J. K., Smith, R. A., Aberle, D. R., Oudkerk, M., Baldwin, D. R., Yankelevitz, D., et al. (2012). International association for the study of lung cancer computed tomography screening workshop 2011 report. *J. Thoracic Oncol.* 7 (1), 10–19. doi: 10.1097/JTO.0b013e31823c58ab

Field, J. K. (2008). Lung cancer risk models come of age. *Cancer Prev. Res.* 1 (4), 226–228. doi: 10.1158/1940-6207.CAPR-08-0144

Gu, C., Wang, R., Pan, X., Huang, Q., Zhang, Y., Yang, J., et al. (2017a). Sublobar resection versus lobectomy in patients aged ≤ 35 years with stage IA non-small cell lung cancer: a SEER database analysis. *J. Cancer Res. Clin. Oncol.* 143 (11), 2375–2382. doi: 10.1007/s00432-017-2499-y

Gu, C., Pan, X., Chen, Y., Yang, J., Zhao, H., Shi, J., et al. (2017b). Short-term and mid-term survival in bronchial sleeve resection by robotic system versus thoracotomy for centrally located lung cancer. *Eur. J. Cardio-Thoracic Surg.* 53 (3), 648–655. doi: 10.1093/ejcts/ezx355

Gu, C., Wang, R., Pan, X., Huang, Q., Luo, J., Zheng, J., et al. (2017c). Comprehensive study of prognostic risk factors of patients underwent pneumonectomy. *J. Cancer* 8 (11), 2097. doi: 10.7150/jca.19454

Gu, C., Huang, Z., Dai, C., Wang, Y., Ren, Y., Chen, C., et al. (2018). Prognostic analysis of limited resection versus lobectomy in stage IA small cell lung cancer patients based on the surveillance, epidemiology, and end results registry database. *Front. Genet.* 9, 1–6. doi: 10.3389/fgene.2018.00568

Jiang, L., Liang, W., Shen, J., Chen, X., Shi, X., He, J., et al. (2015). The impact of visceral pleural invasion in node-negative non-small cell lung cancer: a systematic review and meta-analysis. *Chest* 148 (4), 903–911. doi: 10.1378/chest.14-2765

Kase, S., Sugio, K., Yamazaki, K., Okamoto, T., Yano, T., Sugimachi, K., et al. (2000). Expression of E-cadherin and β-catenin in human non-small cell lung cancer and the clinical significance. *Clin. Cancer Res.* 6 (12), 4789–4796.

National Lung Screening Trial Research Team (2011). The national lung screening trial: overview and study design. *Radiology* 258 (1), 243–253. doi: 10.1148/radiol.10091808

Nozawa, N., Hashimoto, S., Nakashima, Y., Matsuo, Y., Koga, T., Sugio, K., et al. (2006). Immunohistochemical α-and β-catenin and E-cadherin expression and their clinicopathological significance in human lung adenocarcinoma. *Pathol. Res. Pract.* 202 (9), 639–650. doi: 10.1016/j.prp.2006.03.007

Raji, O. Y., Duffy, S. W., Agbaje, O. F., Baker, S. G., Christiani, D. C., Field, J. K., et al. (2012). Predictive accuracy of the Liverpool Lung Project risk model for stratifying patients for computed tomography screening for lung cancer: a case–control and cohort validation study. *Ann. Internal Med.* 157 (4), 242–250. doi: 10.7326/0003-4819-157-4-201208210-00004

Rami-Porta, R., Ball, D., Crowley, J., Giroux, D. J., Jett, J., Travis, W. D., et al. (2007). The IASLC Lung Cancer Staging Project: proposals for the revision of the T descriptors in the forthcoming (seventh) edition of the TNM classification for lung cancer. *J. Thoracic Oncol.* 2 (7), 593–602. doi: 10.1097/JTO.0b013e31807a2f81

Retera, J. M., Leers, M. P., Sulzer, M. A., and Dar, M. J. (1998). The expression of beta-catenin in non-small-cell lung cancer: a clinicopathological study. *J. Clin. Pathol.* 51 (12), 891–894. doi: 10.1136/jcp.51.12.891

Shigemitsu, K., Sekido, Y., Usami, N., Mori, S., Sato, M., Horio, Y., et al. (2001). Genetic alteration of the β-catenin gene (CTNNB1) in human lung cancer and malignant mesothelioma and identification of a new 3p21. 3 homozygous deletion. *Oncogene* 20 (31), 4249. doi: 10.1038/sj.onc.1204557

Shimizu, K., Yoshida, J., Nagai, K., Nishimura, M., Ishii, G., Morishita, Y., et al. (2005). Visceral pleural invasion is an invasive and aggressive indicator of non-small cell lung cancer. *J. Thoracic Cardiovasc. Surg.* 130 (1), 160–165. doi: 10.1016/j.jtcvs.2004.11.021

Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., et al. (2007). A risk model for prediction of lung cancer. *J. Natl. Cancer Institute* 99 (9), 715–726. doi: 10.1093/jnci/djk153

Sun, Y., Ren, Y., Fang, Z., Li, C., Fang, R., Gao, B., et al. (2010). Lung adenocarcinoma from East Asian never-smokers is a disease largely defined by targetable oncogenic mutant kinases. *J. Clin. Oncol.* 28 (30), 4616. doi: 10.1200/JCO.2010.29.6038

Sunaga, N., Kohno, T., Kolligs, F. T., Fearon, E. R., Saito, R., and Yokota, J. (2001). Constitutive activation of the Wnt signaling pathway by CTNNB1 (β-catenin) mutations in a subset of human lung adenocarcinoma. *Genes Chromosomes Cancer* 30 (3), 316–321. doi: 10.1002/1098-2264(2000)9999:9999<::aid-gcc1097>3.0.co;2-9

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA: Cancer J. Clinicians* 65 (2), 87–108. doi: 10.3322/caac.21262

Travis, W. D., Brambilla, E., Rami-Porta, R., Vallières, E., Tsuboi, M., Rusch, V., et al. (2008). Visceral pleural invasion: pathologic criteria and use of elastic stains: proposal for the 7th edition of the TNM classification for lung cancer. *J. Thoracic Oncol.* 3 (12), 1384–1390. doi: 10.1097/JTO.0b013e31818e0d9f

Travis, W. D., Brambilla, E., Noguchi, M., Nicholson, A. G., Geisinger, K. R., Yatabe, Y., et al. (2011). IASLC/ATS/ERS international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* 6 (2), 244–285. doi: 10.1097/JTO.0b013e318206a221

Wang, R., Pan, Y., Li, C., Hu, H., Zhang, Y., Li, H., et al. (2012). The use of quantitative real-time reverse transcriptase PCR for 5′ and 3′ portions of ALK transcripts to detect ALK rearrangements in lung cancers. *Clin. Cancer Res.* 18 (17), 4725–4732. doi: 10.1158/1078-0432.CCR-12-0677

Woenckhaus, M., Merk, J., Stoehr, R., Schaeper, F., Gaumann, A., Wiebe, K., et al. (2008). Prognostic value of FHIT, CTNNB1, and MUC1 expression in non–small cell lung cancer. *Hum. Pathol.* 39 (1), 126–136. doi: 10.1016/j.humpath.2007.05.027

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: 07 February 2020
doi: 10.3389/fgene.2019.01359

Transcriptome Changes of *Mycobacterium marinum* in the Process of Resuscitation From Hypoxia-Induced Dormancy

Jun Jiang [1†], *Chen Lin* [1†], *Junli Zhang* [1], *Yuchen Wang* [1], *Lifang Shen* [1], *Kunpeng Yang* [1], *Wenxuan Xiao* [1], *Yao Li* [1*], *Lu Zhang* [1,2*] and *Jun Liu* [1,3*]

[1] State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai, China, [2] State Key Laboratory of Genetic Engineering, MOE Engineering Research Center of Gene Technology, School of Life Science, Fudan University, Shanghai, China, [3] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS),
China

Reviewed by:
Li Zhang,
East China Normal University,
China
Yongyong Yang,
Northwestern University,
United States

***Correspondence:**
Yao Li
yaoli@fudan.edu.cn
Lu Zhang
zhanglu407@fudan.edu.cn
Jun Liu
jun.liu@utoronto.ca

†These authors have contributed equally to this work

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

Received: 14 October 2019
Accepted: 11 December 2019
Published: 07 February 2020

Citation:
Jiang J, Lin C, Zhang J, Wang Y, Shen L, Yang K, Xiao W, Li Y, Zhang L and Liu J (2020) Transcriptome Changes of Mycobacterium marinum in the Process of Resuscitation From Hypoxia-Induced Dormancy. Front. Genet. 10:1359. doi: 10.3389/fgene.2019.01359

Nearly one-third of the world's population is latently infected with *Mycobacterium tuberculosis* (*M. tb*), which represents a huge disease reservoir for reactivation and a major obstacle for effective control of tuberculosis. During latent infection, *M. tb* is thought to enter nonreplicative dormant states by virtue of its response to hypoxia and nutrient-deprived conditions. Knowledge of the genetic programs used to facilitate entry into and *exit* from the nonreplicative dormant states remains incomplete. In this study, we examined the transcriptional changes of *Mycobacterium marinum* (*M. marinum*), a pathogenic mycobacterial species closely related to *M. tb*, at different stages of resuscitation from hypoxia-induced dormancy. RNA-seq analyses were performed on *M. marinum* cultures recovered at multiple time points after resuscitation. Differentially expressed genes (DEGs) at each time period were identified and analyzed. Co-expression networks of transcription factors and DEGs in each period were constructed. In addition, we performed a weighted gene co-expression network analysis (WGCNA) on all genes and obtained 12 distinct gene modules. Collectively, these data provided valuable insight into the transcriptome changes of *M. marinum* upon resuscitation as well as gene module function of the bacteria during active metabolism and growth.

Keywords: transcriptional regulation, resuscitation, *M. marinum*, hypoxia, latency

INTRODUCTION

Mycobacterium tuberculosis (*M. tb*), the causative agent of tuberculosis (TB), is the leading cause of death due to an infectious disease globally, with an estimated 10 million new cases and 1.3 million deaths in 2017. There were an additional 300,000 deaths from TB among HIV-positive people. The success of *M. tb* as a leading pathogen is associated with its ability to infect and persist in the host. About 1.7 billion people, 23% of the world's population, are estimated to have a latent TB infection (LTBI), which is asymptomatic but can persist for decades (Stewart et al., 2003; North and Jung., 2004). About 5–10% of LTBI will eventually develop active disease, and host immunosuppression (e.g., HIV coinfection) markedly increases the risk of reactivation (Corbett et al., 2007). LTBI poses a major challenge to the effective control of TB because of the difficulty in treatment and the fact that LTBI represents a huge disease reservoir.

During LTBI, *M. tb* is thought to enter nonreplicative 'dormant' states by virtue of its lowered or altered metabolism in response to hypoxia, nitrosative stress, and/or nutrient deprivation (Boshoff and Barry, 2005). Accordingly, much research has been focused on environmental conditions and genetic programs that induce bacteriostasis, and the most extensively studied culture condition is hypoxia (Wayne and Sohaskey, 2001; Rustad et al., 2009). It was shown that an immediate bacterial response (2 hr) was the coordinated upregulation of 47 *M. tb* genes under the control of the response regulator (DosR) and two sensor kinases (DosS and DosT), known as the DosR regulon (Sherman et al., 2001; Boon and Dick, 2002; Park et al., 2003). A second set of 230 genes, induced by longer hypoxia exposure (7 days), was also identified (Rustad et al., 2008). These genes, collectively known as the enduring hypoxic response (EHR), were DosR-independent genes (Rustad et al., 2008).

During the reactivation of LTBI, the dormant bacteria are believed to resuscitate and resume active growth and metabolism. A few recent studies have used reaeration of hypoxic cultures for *in vitro* modeling of reactivation or resuscitation (Veatch and Kaushal, 2018). Several regulatory proteins, such as transcription factor ClgR and sigma factors SigH and SigE, were found to play a role in *M. tb* resuscitation from hypoxia (Mcgillivray et al., 2015; Iona et al., 2016; Veatch et al., 2016).

Despite the progress, knowledge of the genetic programs used to facilitate entry into and *exit* from the nonreplicative dormant states remains incomplete. In this study, we examined the transcriptional changes of *Mycobacterium marinum* (*M. marinum*) at different stages of resuscitation from hypoxia-induced dormancy. *M. marinum* is a pathogenic *Mycobacterium* and the closest genetic relative of the *M. tb* complex. *M. marinum* is an excellent model through which to understand various aspects of host–pathogen interactions in *M. tb* pathogenesis. For example, *M. marinum* and *M. tb* share many virulence determinants, such as the ESX-1 secretion system (Tobin and Ramakrishnan, 2008) and lipid virulence factors phthiocerol dimycocerosates and phenolic glycolipids (Yu et al., 2012). As such, findings from the current study of *M. marinum* may be applicable to *M. tb*.

RESULT

RNA-Seq Analysis of *M. marinum* Recovered From Hypoxia

Larry Wayne and co-workers were the first to develop an *in vitro* model to mimic the hypoxic environment of the human granuloma (Wayne, 1977; Wayne and Hayes., 1996; Wayne and Sohaskey., 2001). In the Wayne model, a sealed, standing culture is incubated over an extended period while the bacteria deplete the available oxygen. The gradual depletion of oxygen leads to nonreplicating persistence states with a concomitant shift in gene expression and metabolism.

To gain insight into the genetic mechanisms that facilitate the exit of mycobacteria from the nonreplicative state, we grew *M. marinum* under hypoxia for 7 days using the Wayne model and then reaerated the cultures. At different time points thereafter (0, 0.5, 4, 12, 24, and 48 hr), *M. marinum* cultures were collected and subjected to RNA-seq analysis. The growth curve of the bacteria is shown in **Supplementary Figure 1**. A total of 18 samples were collected (three biological replicates at each time point) and analyzed.

The RNA-seq reads showed a high mapping ratio for all samples (>96%) (**Table 1**), supporting the overall sequencing accuracy. Transcripts of more than 4,900 genes were detected in each sample. We compared the RNA-seq data of cultures recovered at different time points under aerobic conditions. As expected, results showed that the correlation coefficient decreased as the interval between two samples increased (**Figure 1**). This result also suggested that the recovery from hypoxia is a gradual but dynamic process.

Dynamic Changes of Gene Expression at Different Stage of Resuscitation

To analyze transcriptome changes of *M. marinum*, we focused on genes with RPKM ≥ 10 and compared samples from adjacent intervals: between 0.5 and 0 hr, 4 and 0.5 hr, 12 and 4 hr, 24 and 12 hr, as well as 48 and 24 hr. Differentially expressed genes (DEGs) were identified, and this was defined as a fold change greater than 2 and false discovery rate *P* value less than 0.05.

At the earliest time point after resuscitation (0.5 hr), 136 DEGs were detected, of which 71 were upregulated and 65 were downregulated. Between 4 and 0.5 hr, most of the DGEs were downregulated (81 out of 88 total DEGs). The numbers of DGEs found between 12 and 4 hr, 24 and 12 hr, as well as 48 and 24 hr

TABLE 1 | RNA-seq reads of 18 samples (three biological replicates at each time point).

Sample	Total reads	Mapped reads	Pair mapped reads	Single mapped reads	Mapped ratio(%)
0hr_rep1	25667618	25206648	25071544	135104	98.2
0hr_rep2	24461374	24025951	23875630	150321	98.22
0hr_rep3	24356981	23991626	23851008	140618	98.5
0.5hr_rep1	27750348	27226045	27037288	188757	98.11
0.5hr_rep2	29227286	28693463	28508636	184827	98.17
0.5hr_rep3	30215819	29671934	29509385	162549	98.2
4hr_rep1	29552072	28996782	28795942	200840	98.12
4hr_rep2	25707932	25227133	25058798	168335	98.13
4hr_rep3	26459312	25956585	25781988	174597	98.1
12hr_rep1	20056376	19435701	19294860	140841	96.91
12hr_rep2	25849366	25305478	25143598	161880	97.9
12hr_rep3	24639687	24048335	23903266	145069	97.6
24hr_rep1	28399172	27857378	27662324	195054	98.09
24hr_rep2	31131638	30539305	30334952	204353	98.1
24hr_rep3	29686534	29152176	28950812	201364	98.2
48hr_rep1	28813928	28353514	28145540	207974	98.4
48hr_rep2	27427442	26995636	26807948	187688	98.43
48hr_rep3	28062635	27641695	27444061	197634	98.5

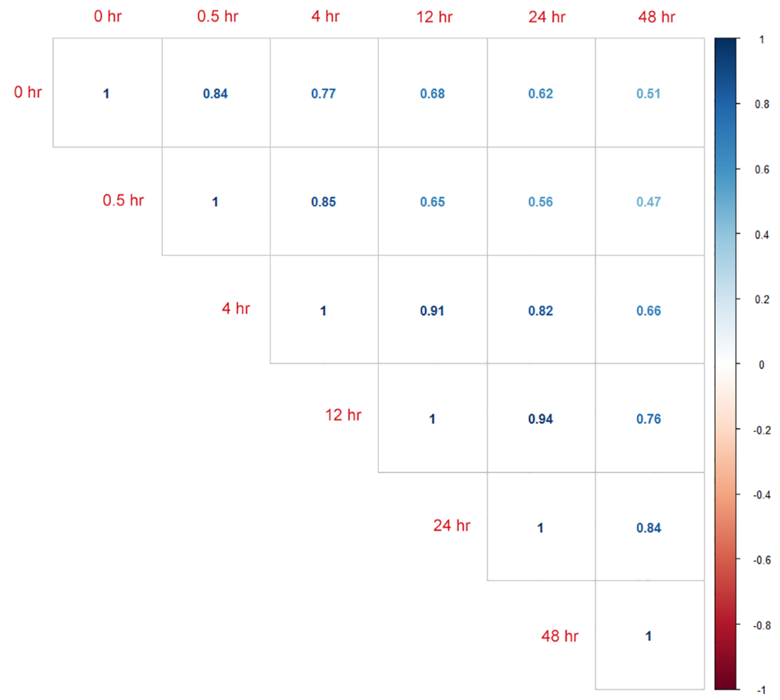


FIGURE 1 | Calculated correlation coefficients between RNA-seq data from samples of different time points (0_hour, 0.5_hour, 4_hour, 12_hour, 24_hour, and 48_hour) during resuscitation.

were 72, 85, and 172, respectively. The heat map of the five groups of DEGs is shown in **Figure 2**.

We performed a Venn analysis of the five DEG groups (**Figure 3A**). The proportion of unique genes in each group was high: 68.4% (93/136, between 0.5 and 0 hr), 60.2% (53/88, between 4 and 0.5 hr), 46.7% (35/72, between 12 and 4 hr), 47.1% (40/85, between 24 and 12 hr), and 74.4% (128/172, between 48 and 24 hr). This suggests that a variety of genes were involved at different stages of resuscitation.

Combing the DEGs of different stages, a total of 440 genes were identified (**Supplementary Table 1**), and their expression underwent dynamic changes during resuscitation (**Figure 3B**). For example, the expression of *MMAR_0922*, *MMAR_3562*, and *MMAR_1654* were significantly changed at the early stage of resuscitation, suggesting that they may play an important role in this period. Some genes had changed in multiple time periods. For example, the expression of *MMAR_3403* was changed in last three periods, suggesting that this gene may be associated with the late stage of resuscitation.

Validation of RNA-seq Results by RT-qPCR

To validate the RNA-seq results, a real-time quantitative (RT-qPCR) analysis was performed. Three biologically independent samples at each time point were used for this experiment. For each DEG group, we selected the top 10 upregulated genes and 10 downregulated genes for this analysis (**Figure 4**). Between 0 and 0.5 hr, *MMAR_4852* and *MMAR_5170* (*whiB4*) were significantly downregulated, and six genes, *MMAR_5122* (*lipX*), *MMAR_0548* (*espG3*), *MMAR_0547* (*esxR*), *MMAR_0551* (*eccE3*), *MMAR_0546* (*esxG*), and *MMAR_0550* (*mycP3*), were significantly upregulated (**Figure 4A**). Between 0.5 to 4 hr, six genes, including *MMAR_1656*, *MMAR_1658* (*hycQ*), *MMAR_1653* (*Rv0081*), *MMAR_1655*, *MMAR_5122* (*lipX*), and *MMAR_5170* (*whiB4*), were significantly downregulated and four genes *MMAR_0845* (*hemB*), *MMAR_5484*, *MMAR_1908* (*ATC1*), and *MMAR_3776* (*rpfE*) were significantly upregulated (**Figure 4B**). Between 4 and 12 hr, *MMAR_2343* (*papA1*), *MMAR_3555*, and *MMAR_2320* (*wecE*) were significantly upregulated and *MMAR_4903* were significantly downregulated (**Figure 4C**). Between 12 and 24 hr, six genes, *MMAR_0335, MMAR_0602, MMAR_4903, MMAR_4899, MMAR_2009*, and *MMAR_0615 (iniA)*, were significantly downregulated (**Figure 4D**). Between 24 and 48 hr, six genes, *MMAR_3465* (*PPE51*), *MMAR_4824, MMAR_4482* (*cypM*), *MMAR_4750, MMAR_2944*, and *MMAR_1790* (*PPE2*), were significantly downregulated, and seven genes, *MMAR_2651, MMAR_2914* (*katG*), *MMAR_2649, MMAR_5315* (*lpqH*), *MMAR_5319, MMAR_2839* (*mpt63*), and *MMAR_0656*, were significantly upregulated (**Figure 4E**).

There is a good agreement between the RNA-seq and qPCR data, evident from the scatter plot using the expression levels of all 97 genes that were analyzed by both RNA-seq and qPCR ($R^2 = 0.784$) (**Figure 4F**). Based on this result, we consider that the RNA-seq data is reliable.

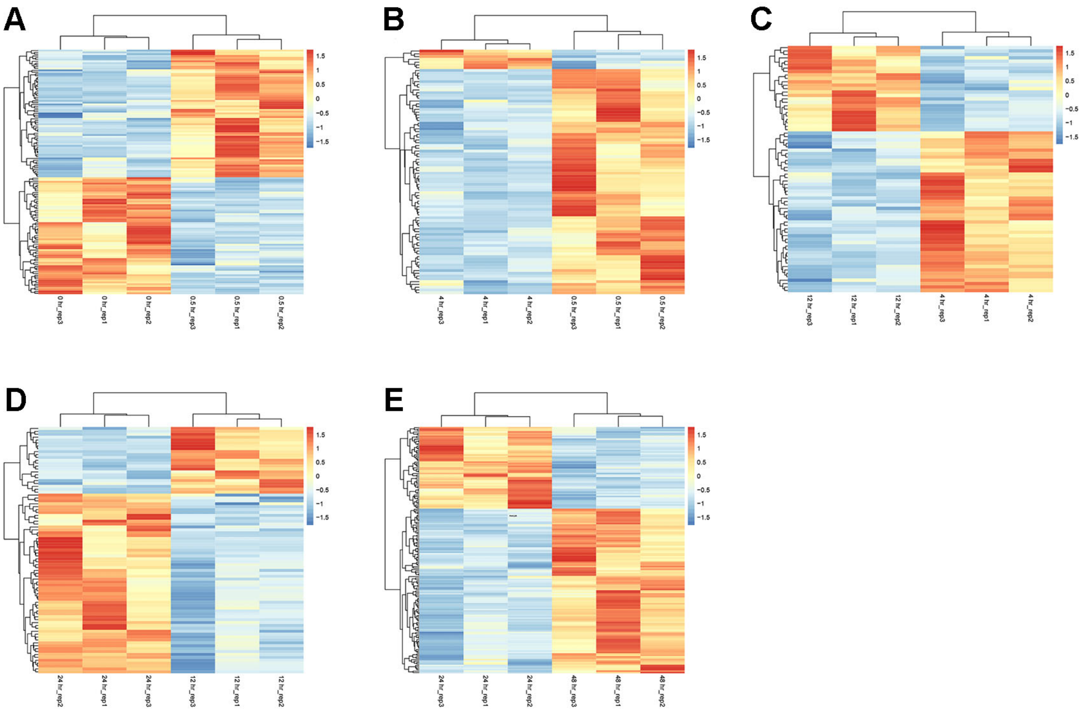


FIGURE 2 | Heatmaps of DEGs between adjacent time points. At each time point, data from three biologically independent samples were included. **(A)** 0.5 vs. 0 hr; **(B)** 4 vs. 0.5 hr; **(C)** 12 vs. 4 hr; **(D)** 24 vs. 12 hr; and **(E)** 48 vs. 24 hr. The red color indicates upregulation. The blue color indicates downregulation.

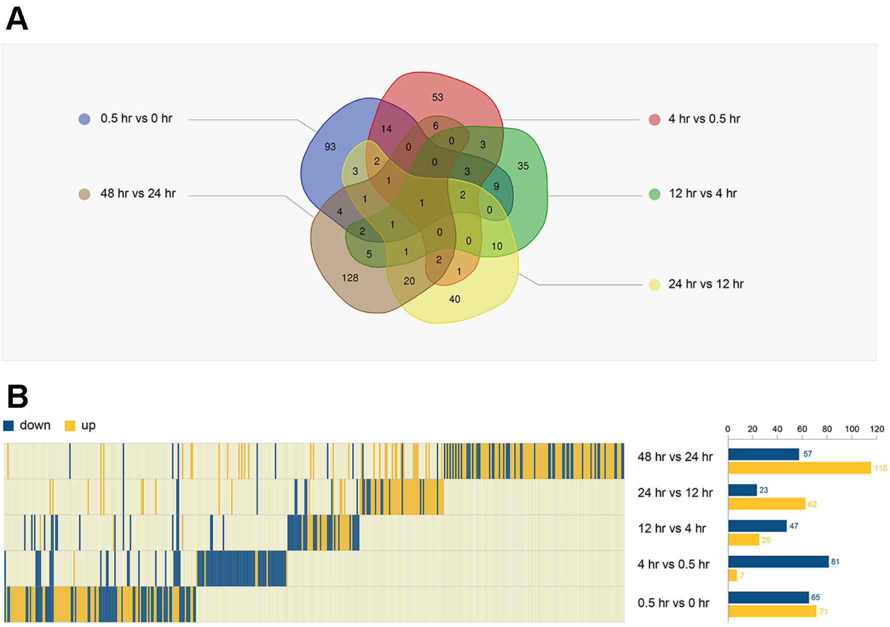


FIGURE 3 | Analysis of DEGs in five groups. **(A)** The Venn diagram of five DEGs. **(B)** The expression trend of all DEGs (440 genes) in the five periods (left). The number of DEGs in each period is shown on the right.

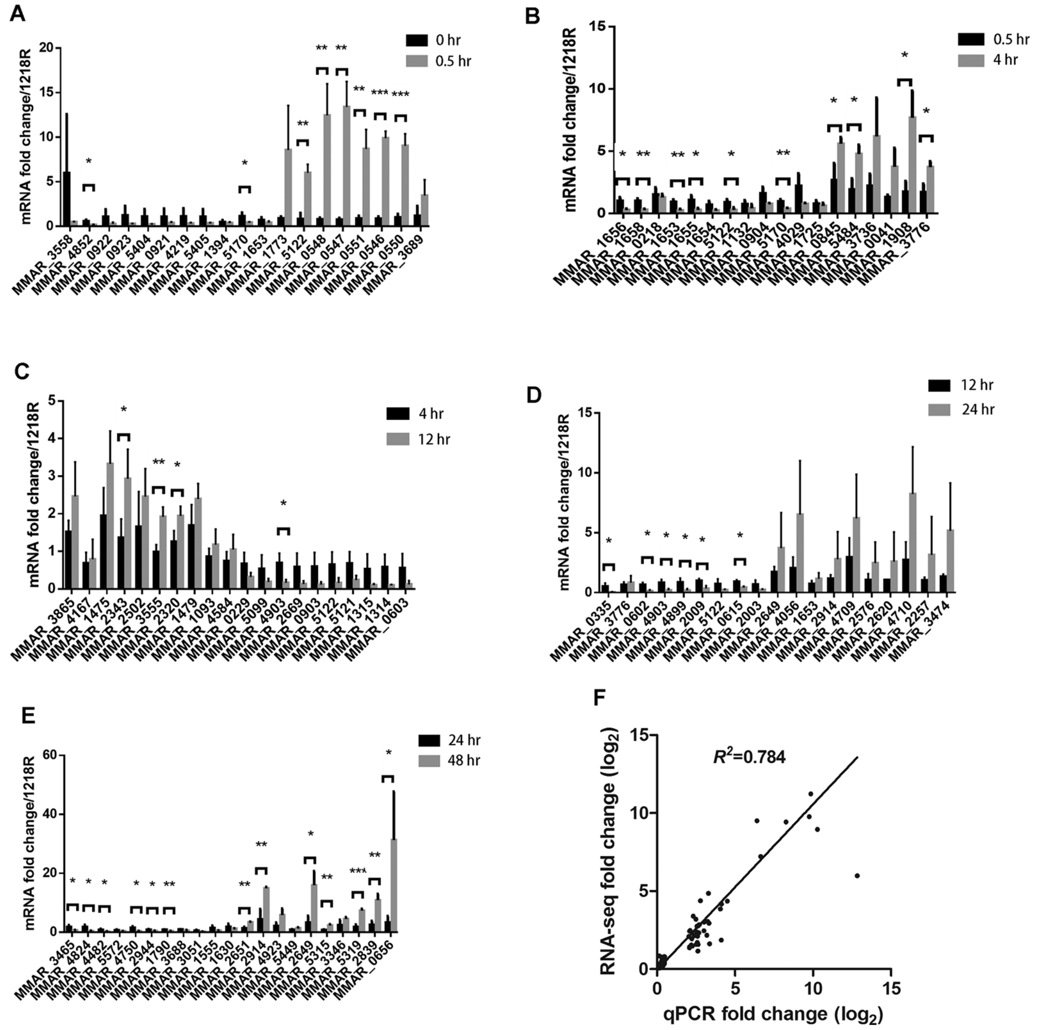


FIGURE 4 | qRT-PCR results. **(A–E)** Approximately 20 genes from each period were selected and analyzed by qRT-PCR. *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$. **(F)** Scatter plot of RT-PCR data of all genes analyzed in **(A–E)**, comparing them to RNA-seq data of the same genes.

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Enrichment Analysis

To gain insight into the biological consequence of the observed transcriptome changes, we performed GO and KEGG pathway analyses of the five DEG groups (**Figure 5**). A GO analysis was applied to identify the functional categories of DEGs. Between 0.5 and 0 hr, more than half of DEGs were involved in membrane and cell wall processes and were significantly enriched (**Figure 5**). This is consistent with the notion that, upon reaeration, the bacteria resumed cell division, which involved cell wall and membrane biogenesis. A KEGG analysis consistently revealed that DEGs involved in metabolic pathways and biosynthesis were significantly enriched and accounted for the largest number. A similar trend was observed for later periods, in which genes involved in cell wall and membrane biogenesis were highly enriched and accounted for the majority of DEGs at these stages. These results provide snapshots into the recovery of the bacteria from hypoxia and active growth under aerobic conditions.

Co-Expression Analysis Between Transcriptional Regulators and mRNAs

Co-expression networks can show relationships between genes. To explore the regulatory mechanisms at different stages of resuscitation, we constructed a co-expression network between transcriptional regulators and mRNAs. For this, we selected known transcriptional regulators, including transcription factors and sigma factors from the five DEGs, and calculated the correlation coefficients between these transcription factors and the remaining DEGs in the same group. We considered that a relationship existed between a given transcriptional regulator and other genes if the absolute value of the correlation coefficient was greater than 0.9, which included both positive and negative correlation. Based on these results, we constructed five

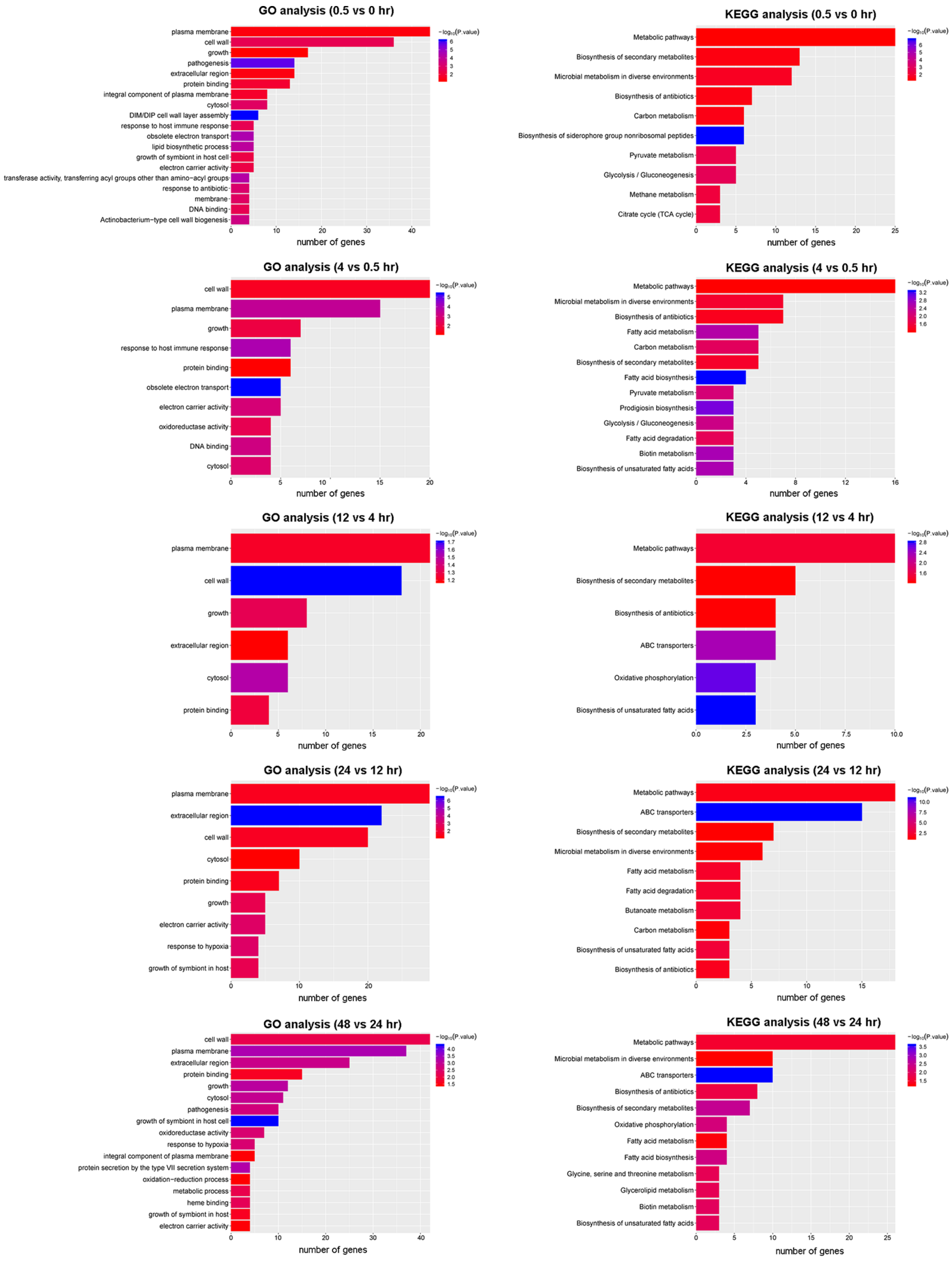


FIGURE 5 | GO and KEGG analysis of DEGs in the five periods.

co-expression modules and integrated them into a large network (**Figure 6**). The dark blue nodes in the figure represent transcription factors, and the green nodes denote DEGs in the same period. The size of the node is determined by the degree of connectivity. Greater degrees of connectivity are indicated by larger points. If there is a line between two nodes, then there is a relationship between them.

In the first co-expression module (between 0 and 0.5 hr), three transcription factors, MMAR_4874 (CosR), MMAR_1653 (Rv0081), and MMAR_4852 (KmtR), formed the major regulatory hubs, and MMAR_4874 (CosR) was the largest hub and interacted with other hubs in the network (**Figure 6**). The MMAR_4874 (CosR) and MMAR_1653 (Rv0081) hubs remained in the second co-expression module (between 0.5

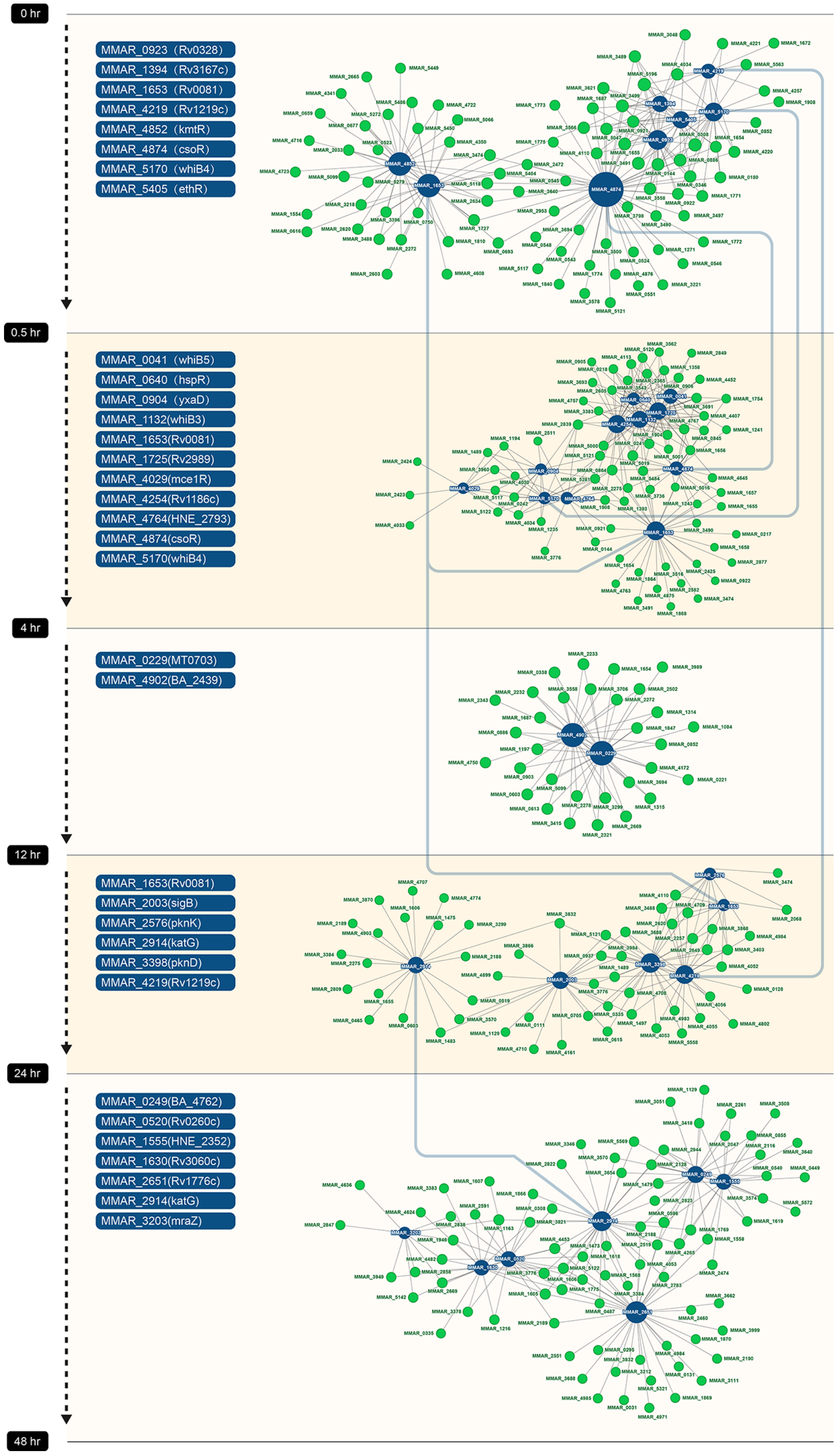


FIGURE 6 | Co-expression networks in five periods. Each dot represents a gene. Transcription factors are labeled in blue, and other genes are labeled in green. Line between dots represents co-expression relationship between genes. The size of dot is proportional to the level of connectivity.

and 4 hr), in addition to three new hubs formed by MMAR_4254, MMAR_1725 and MMAR_1132. In the third period (between 4 and 12 hr), MMAR_0229 and MMAR_4902 formed the hubs. In the fourth period (between 12 and 24 hr), MMAR_2003 (SigB) and MMAR_4219 formed the hubs. In the final period (24 to 48 hr), MMAR_2651, MMAR_1555, and MMAR_0249 formed the hubs.

Weighted Gene Co-Expression Network Analysis (WGCNA)

The DEG analysis focused on partial dynamic changes during resuscitation. While the co-expression network of transcription factors provides an overview of the regulatory programs enabling the resuscitation of *M. marinum*, our knowledge on the overall genetic changes is still missing. Therefore, in this section, we analyzed the expression of all genes from cultures at different stages of resuscitation.

Weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath, 2008) is a method for analyzing the gene expression patterns of multiple samples. It clusters genes into modules by similar expression trends and reveals the relationship between gene modules and specific traits or phenotypes. We applied this method to analyze the RNA-seq data of *M. marinum* at different stages of resuscitation.

The gene modules were identified by the WGCNA package in R software. We first determined the appropriate "soft-thresholding" value, which emphasizes strong gene–gene correlations at the expense of weak correlations. An optimal parameter (power = 20) was determined by plotting the strength of correlation against a series (range 2 to 20) of soft threshold powers (**Figure 7A**).

An unsigned pairwise correlation matrix was calculated, and the WGCNA algorithm was used to transfer the correlation coefficient between genes into the adjacent coefficient. Then, the dissimilarity of the topological overlap matrix was calculated based on the adjacent coefficient. Using the calculated dissimilarity, we carried out a hierarchical analysis by using agglomerative hierarchical clustering, also known as the bottom-up method. Other assumptions were made: (i) distances between

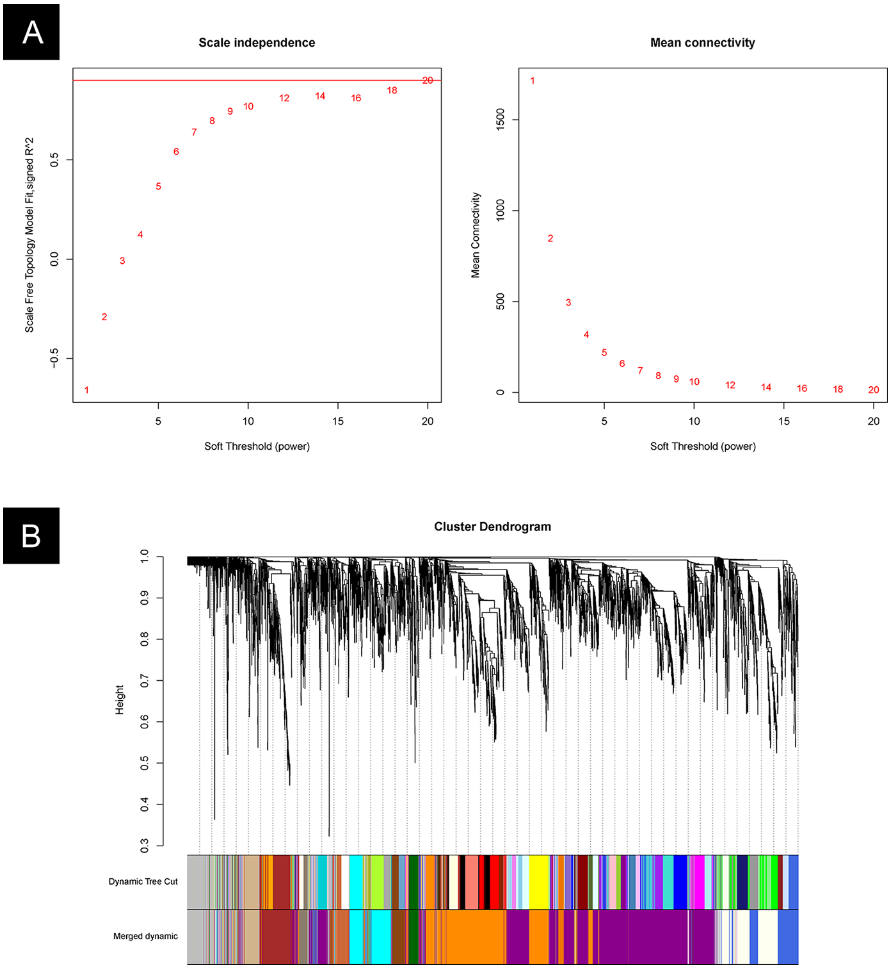


FIGURE 7 | WGCNA cluster analysis. **(A)** Plot of the strength of correlation against a series (range 2 to 20) of soft threshold powers. **(B)** Gene clusters and gene module fusion.

different classes were measured by the average connectivity, and (ii) there should be at least 30 genes in each gene module.

Based on these analyses, we initially obtained 48 gene modules. The hierarchical cluster tree was then treated using the dynamic tree cut algorithm in the WGCNA package. A total of 13 gene modules were obtained. The "gray" module was the default module, which included discarded genes that could not be clustered. Thus, we focused on the analysis in the remaining 12 gene modules. The process of fusion is shown in **Figure 7B**. The number of genes varied among these 12 modules, and the detailed information is listed in **Table 2** and **Supplementary Table 2**.

The first principal component analysis (PCA) was performed on the 12 gene modules (**Figure 8**). The PCA results reflected the main trend of gene expression in the modules. Module 20 played an important role in the early stage (0 to 0.5 hr) of recovery, module 35 played a role mostly in the middle stage (4 to 12 hr), and module 12 was only involved in the last stage (24 to 48 hr). Other modules appeared to play roles in more extended periods.

Identification of Key Gene Modules Associated With Different Stages of Resuscitation

Thus far, we have identified 5 DEGs and 12 gene modules. We then performed an enrichment analysis between them. When the *P* value of Fisher's exact test was less than 0.001, we considered that these gene modules were significantly enriched in the DEG sets. The results are shown in **Figure 9**. Interestingly, module 20 was significantly enriched in DEGs of the early stage (0 to 0.5 hr) and module 12 was significantly enriched in DEGs of the last stage (24 to 48 hr) of the recovery. This is consistent with the result that these two modules were only involved in the early and last stages of resuscitation, respectively (**Figure 8**).

DISCUSSION

In this study, we examined the transcriptome changes of *M. marinum* recovered from hypoxia-induced dormancy. To gain a comprehensive view, multiple time points, including shortly after resuscitation (0.5 hr) to more extended periods up to 48 hr, were included. For each time point, three biologically independent samples were analyzed. Transcripts of the whole genome were analyzed by RNA-seq, and the quality of the RNA-seq data was reflected by the high genome mapping ratio and further validated by qPCR analysis of close to 100 genes. With these high-quality sequence data, we performed in-depth analyses, which included the identification of DEGs and the construction of co-expression network of transcription factors in each period. The availability of transcriptomes of independent samples at multiple time points also allowed us to employ a weighted gene co-expression network analysis to identify gene modules of *M. marinum*. Collectively, these data provide valuable insight into not only the genetic changes of the bacteria upon resuscitation but also the gene module function of *M. marinum* during active metabolism and growth.

TABLE 2 | Information of gene modules identified by WGCNA.

Module_ID	Number of genes
module_34	1711
module_26	919
module_43	424
module_14	320
module_20	298
module_3	268
module_35	142
module_29	136
module_12	127
module_22	89
module_47	89
module_45	46

A total of 136 DEGs were identified in *M. marinum* upon resuscitation from dormancy (0 to 0.5 hr), including eight transcription factors (**Figure 6**). Interestingly, all of these transcription factors were significantly downregulated. Among them, MMAR_1653 is a homolog of Rv0081 in *M. tb*. Previous studies have shown that Rv0081 is a member of the DosR regulon and is induced at the early stage of hypoxia (Sherman et al., 2001). Rv0081 is a major regulator of *M. tb* response to hypoxia and forms a large regulatory hub (Galagan et al., 2013; Chawla et al., 2018; Sun et al., 2018). Rv0081 plays an important role connecting the early and enduring hypoxic responses (Sun et al., 2018). WhiB4 (MMAR_5170) is an oxygen-sensitive transcription factor and has been shown to regulate PE/PPE family proteins, and it plays a role in *M. marinum* virulence (Chawla et al., 2012; Wu et al., 2017). CosR (MMAR_4874) is a copper-inducible transcriptional regulator, and the loss of *cosR* resulted in a hypoxia-type response with the induction of the DosR regulon (Talaat et al., 2004; Ward et al., 2008; Marcus et al., 2016). Given their roles in the hypoxic response, it is not surprising that Rv0081, WhiB4, and CosR underwent dynamic changes in expression upon resuscitation by reaeration. Two other transcription factors that were downregulated at this stage, MMAR_5405 (EthR) and MMAR_1394 (Rv3176c), belong to the TetR family transcription factors (Leiba et al., 2014; Sharma et al., 2017).

We also found that multiple members of the ESAT-6 family proteins were upregulated upon resuscitation (Harboe et al., 1996; Priscille et al., 2004), including EsxA (Sandra et al., 2010; Zhang et al., 2016), EsxB (Sandra et al., 2010), EsxG (Sweeney et al., 2011), EsxH (Alka et al., 2013; Portal-Celhay et al., 2016), EsxK, and EsxN (Zhigang et al., 2017). These proteins are components of the Type VII secretion systems, and many of them are important T cell antigens and play a critical role modulating the host–pathogen interactions (Abdallah et al., 2007).

From 0.5 to 4 hr after reaeration, 11 transcription factors were downregulated (**Figure 6**), including Rv0081, CsoR, and WhiB4 as mentioned above. Notably, the expression of two other WhiB family proteins (Averina et al., 2012), WhiB3 and WhiB5, were also significantly altered. WhiB3 responds to dormancy signals, including hypoxia and NO, and controls redox homeostasis of the bacteria (Priscille et al., 2004). WhiB5 responds to oxygen and controls the expression type VII secretion systems (Priscille et al., 2004). Consistently, we found that *whiB3* was downregulated while *whiB5* was upregulated at this stage of resuscitation.

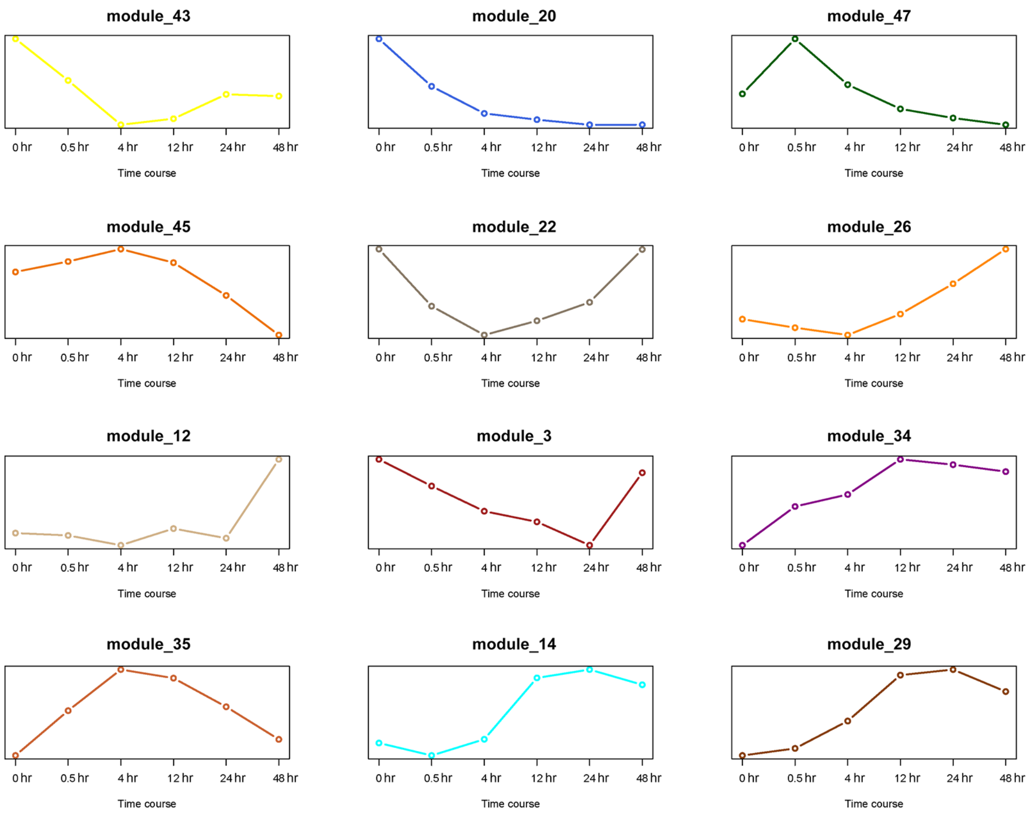


FIGURE 8 | PCA analysis of 12 gene modules. The X axis represents time period, and the Y axis represents expression of first principal component.

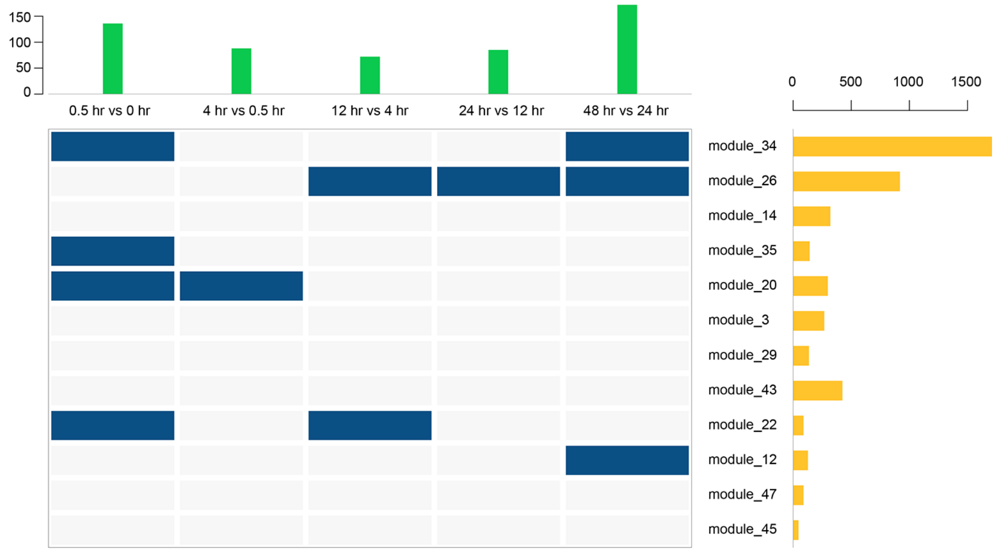


FIGURE 9 | Enrichment analysis between 12 gene modules and 5 DEGs. The above bar chart represents the number of DEGs in each period, and the bar chart on the right represents the number of genes in each gene module. Blue squares represent a significant enrichment between the row set and the column set.

During the period of 4 to 12 hr, we identified 72 DEGs, including 7 *PPE* family genes (*MMAR_5121, MMAR_1095, MMAR_1235, MMAR_1847, MMAR_1905, MMAR_2669,* and *MMAR_3989*). The PE/PPE family proteins play a critical role in mycobacterial pathogenesis (Fishbein et al., 2015). In addition, several genes of the Mce family were upregulated, including *MMAR_3865 (mce2B), MMAR_3868 (mce5E), MMAR_3867 (mce5D), MMAR_3866 (mce5C)*. The Mce family genes comprise four mammalian cell invasion factor (*mce*) operons (*mce1-4*), and some of these are involved in the invasion of host cells (Zhang and Xie, 2011).

A picture appears to have emerged from these analyses; in the early stage of resuscitation from the hypoxia-induced dormancy, transcription factors critical for a hypoxia-induced response are downregulated, and, as the recovery continues, genes important for virulence and host interactions are upregulated.

A WGCNA analysis revealed 12 distinct gene modules. Of particular interest is gene module 20, which was involved in the early stage of resuscitation only (**Figure 8**). This module comprises of ~300 genes, many of which have unknown functions or annotations. Future studies focusing on genes in this module may help to understand the molecular machinery enabling the exit of the bacteria from dormancy.

METHODS AND MATERIALS

Bacterial Strain, Media, and Growth Conditions

M. marinum 1218R (ATCC 927) was grown in Middlebrook 7H9 broth to OD600~0.5, at which point they were aliquoted and cultured in screw-capped conical flasks at 30°C without additional oxygen. The hypoxic culture conditions were described previously by Wayne and Hayes (Wayne and Hayes, 1996). After 7 days in hypoxic conditions, the screw cap was replaced with a permeable membrane, and the rest of the conditions were unchanged. After aeration, samples were taken at 0 h, 0.5 h, 4 h, 12 h, 24 h, and 48 h, and an aliquot was used to measure the OD value (**Supplementary Figure 1**). The remaining samples were collected and snap frozen in liquid nitrogen for RNA sequencing.

RNA Extraction, Illumina Sequencing, and RT-qPCR

M. marinum cultures were centrifuged at 4,500 × g for 5 min at room temperature and frozen on dry ice. The frozen cell pellets were resuspended in 1 mL TRIzol reagent (CW Bio). RNA extraction and illumina sequencing were performed as previously described (Wu et al., 2017). Raw data of RNA sequencing have been uploaded to the GEO database (BioProject ID : PRJNA588556).

For RT-qPCR validation of RNA-seq data, 1 μg RNA was reversed-transcribed to cDNA, which was then used as the template for RT-PCR analysis. The primers for analyzing the selected genes were listed in **Supplementary Table 3**.

Transcriptome and Bioinformatics Analysis

The RNA-seq analysis and identification of differentially expressed genes (DEGs) were performed as previously described (Lee et al., 2019).

GO and KEGG Analysis

We used the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Ogata et al., 2000) (www.genome.jp/kegg/) and the Gene Ontology (GO) database (Ashburner et al., 2000) (www.geneontology.org/) for our data analysis.

Construction of the Gene Co-Expression Network

We calculated the correlation coefficient between identified transcription factors and other DEGs in each time period. The correlation coefficient value ranges from -1 to 1, representing a negative and positive correlation, respectively. We considered the expression of two genes as correlated if the absolute value of the correlation coefficient was larger than 0.8. The results were imported into Cytoscape 3.0 to generate the network map (Kohl et al., 2011).

Weighted Gene Co-Expression Network Analysis (WGCNA)

We used the RNA-seq data from multiple time points (three biological independent samples at each time point) for WGCNA analysis. We used the WGCNA package to cluster gene modules as follows.

(a) Define gene co-expression similarity: calculate the similarity between any two genes using Pearson's correlation coefficient (Sij = |cor(i,j)|, the correlation coefficient of gene i and gene j), which then forms the correlation matrix ($S = [S_{ij}]$).

(b) Define the exponential weighted value β: for any gene pair (i and j), apply the exponential adjacency function in the WGCNA algorithm to measure their relation index, namely, the exponential weighted β square of the correlation coefficient ($a_{ij} = \text{power}(S_{ij}, \beta) = |S_{ij}|^{\beta}$). Exponential weighted β is the power of the correlation coefficient. We selected β = 5 after the analysis (fit value R^2 to approximately 0.9).

(c) Define a measure of node dissimilarity: after determining the adjacency function parameter β, the correlation matrix $S = [S_{ij}]$ is switched into the adjacency matrix $A = [a_{ij}]$ and converted into the topological overlap matrix $\Omega = [\omega_{ij}]$. k_i or k_j indicate the sum of one node's adjacency coefficients. The node is a gene (i or j).

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - \ a_{ij}}$$

$$l_{ij} = \sum\nolimits_{\mu} a_{i\mu} a_{\mu j} \ k_i = \sum\nolimits_{\mu} a_{i\mu} \ k_i = \sum\nolimits_{\mu} a_{i\mu}$$

(d) Build hierarchical clustering tree to identify gene modules: the hierarchical clustering tree built using the dissimilarity coefficient d_{ij}^{ω} ($d_{ij}^{\omega} = 1 - \omega_{ij}$), and the different branches represent the gene modules.

Enrichment Analysis and PCA Analysis

To determine whether one set of genes were more enriched in another set of genes, we used the Chi-square test or Fisher's exact test (Upton, 1992). First, the two sets of genes were taken and used to form a 2*2 contingency table. If there was a value less than or equal to five in the table, the Fisher's exact test was applied; otherwise, the Chi-square test was applied. When the p value was less than 0.05, the two sets were considered significantly enriched to each other. PCA analyses were performed by the princomp function in R software (version 3.5.1)

DATA AVAILABILITY STATEMENT

Raw data of RNA sequencing have been uploaded to the GEO database (BioProject ID:PRJNA588556).

AUTHOR CONTRIBUTIONS

JJ and CL performed the experiment. JJ, CL, and JL wrote the manuscript. YW and JZ prepared **Figures 1–9**. LS, KY, and WX prepared **Tables 1** and **2**. LZ, YL, and JL provided guidance for experiments. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

This study was supported by grants from China's 13th Five Year Programs for the prevention and cure of great infectious diseases (2017ZX10201301-005), the National Key R&D Program of China (No.2018YFD0500900), and the Open Fund of Shanghai Key Laboratory of Tuberculosis (2018kf03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01359/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | Growth curve of M. marinum after reaeration from hypoxic cultures.

SUPPLEMENTARY TABLE 1 | Detailed information on the differentially expressed gene list at different time periods.

SUPPLEMENTARY TABLE 2 | List of genes identified in 12 gene modules by WGCNA analysis.

SUPPLEMENTARY TABLE 3 | Primer sequences of genes analyzed by RT-PCR.

REFERENCES

Abdallah, A. M., Gey van Pittius, N. C., Champion, P. A., Cox, J., Luirink, J., Vandenbroucke-Grauls, C. M., et al. (2007). Type VII secretion–mycobacteria show the way. *Nat. Rev. Microbiol.* 5 (11), 883–891. doi: 10.1038/nrmicro1773

Alka, M., Aleena, Z., Victor, T., Natalie, S., Ashley, W., Maura, P., et al. (2013). Mycobacterium tuberculosis Type VII secreted effector EsxH targets host ESCRT to impair trafficking. *PLoS Pathogens* 9 (10), e1003734. doi: 10.1371/journal.ppat.1003734

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Gene Ontol. Consortium Nat. Genet.* 25 (1), 25–29. doi: 10.1038/75556

Averina, O. V., Zakharevich, N. V., and Danilenko., V. N. (2012). Identification and characterization of WhiB-like family proteins of the Bifidobacterium genus. *Anaerobe* 18 (4), 421–429. doi: 10.1016/j.anaerobe.2012.04.011

Boon, C., and Dick., T. (2002). Mycobacterium bovis BCG response regulator essential for hypoxic dormancy. *J. Bacteriol.* 184 (24), 6760–6767. doi: 10.1128/JB.184.24.6760-6767.2002

Boshoff, H. I., and Barry, C. E.3rd (2005). Tuberculosis - metabolism and respiration in the absence of growth. *Nat. Rev. Microbiol.* 3 (1), 70–80. doi: 10.1038/nrmicro1065

Chawla, M., Parikh, P., Saxena, A., Munshi, M., Mehta, M., Mai, D., et al. (2012). Mycobacterium tuberculosis WhiB4 regulates oxidative stress response to modulate survival and dissemination *in vivo*. *Mol. Microbiol.* 85 (6), 1148–1165. doi: 10.1111/j.1365-2958.2012.08165.x

Chawla, M., Mishra, S., Anand, K., Parikh, P., Mehta, M., Vij, M., et al. (2018). Redox-dependent condensation of the mycobacterial nucleoid by WhiB4. *Redox Biol.* 19, 116–133. doi: 10.1016/j.redox.2018.08.006

Corbett, E. L., Bandason, T., Cheung, Y. B., Munyati, S., Godfrey-Faussett, P., Hayes, R., et al. (2007). Epidemiology of tuberculosis in a high HIV prevalence population provided with enhanced diagnosis of symptomatic disease. *PLoS Med.* 4 (1), e22. doi: 10.1371/journal.pmed.0040022

Fishbein, S., van Wyk, N., Warren, R. M., and Sampson., S. L. (2015). Phylogeny to function: PE/PPE protein evolution and impact on Mycobacterium tuberculosis pathogenicity. *Mol. Microbiol.* 96 (5), 901–916. doi: 10.1111/mmi.12981

Galagan, J. E., Minch, K. J., Peterson, M. W., Lyubetskaya, A., Azizi, E., Sweet, L., et al. (2013). The Mycobacterium tuberculosis regulatory network and hypoxia. *Nature* 499 (7457), 178–183. doi: 10.1038/nature12337

Harboe, M., Oettinger, T., Wiker, H. G., Rosenkrands, I., and Andersen., P. (1996). Evidence for occurrence of the ESAT-6 protein in Mycobacterium tuberculosis and virulent Mycobacterium bovis and for its absence in Mycobacterium bovis BCG. *Infection Immunity* 64 (1), 16–22. doi: 10.1128/IAI.64.1.16-22

Iona, E., Pardini, M., Mustazzolu, A., Piccaro, G., Nisini, R., Fattorini, L., et al. (2016). Mycobacterium tuberculosis gene expression at different stages of hypoxia-induced dormancy and upon resuscitation. *J. Microbiol.* 54 (8), 565–572. doi: 10.1007/s12275-016-6150-4

Kohl, M., Wiese, S., and Warscheid., B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* 696 (696), 291–303. doi: 10.1007/978-1-60761-987-1_18

Langfelder, P., and Horvath., S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform.* 9 (1), 559–559. doi: 10.1186/1471-2105-9-559

Lee, J., Lee, S.-G., Kim, K. K., Lim, Y.-J., Choi, J.-A., Cho, S.-N., et al. (2019). Characterisation of genes differentially expressed in macrophages by virulent and attenuated Mycobacterium tuberculosis through RNA-Seq analysis. *Sci. Rep.* 9 (1), 4027. doi: 10.1038/s41598-019-40814-0

Leiba, J., Carrère-Kremer, S., Blondiaux, N., Dimala, M. M., Wohlkönig, A., Baulard, A., et al. (2014). The mycobacterium tuberculosis transcriptional repressor EthR is negatively regulated by Serine/Threonine phosphorylation. *Biochem. Biophys. Res. Commun.* 446 (4), 1132–1138. doi: 10.1016/j.bbrc.2014.03.074

Marcus, S. A., Sidiropoulos, S. W., Steinberg, H., and Talaat., A. M. (2016). csor is essential for maintaining copper homeostasis in mycobacterium tuberculosis. *PLoS One* 11 (3), e0151816. doi: 10.1371/journal.pone.0151816

Mcgillivray, A., Golden, N. A., and Kaushal., D. (2015). The mycobacterium tuberculosis Clp gene regulator is required for *in vitro* reactivation from hypoxia-induced dormancy. *J. Biol. Chem.* 290 (4), 2351–2367. doi: 10.1074/jbc.M114.615534

North, R. J., and Jung., Y. J. (2004). Immunity to tuberculosis. *Annu. Rev. Immunol.* 22, 599–623. doi: 10.1146/annurev.immunol.22.012703.104635

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa., M. (2000). (KEGG). Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27 (1), 29–34. doi: 10.1093/nar/28.1.27

Park, H. D., Guinn, K. M., Harrell, M. I., Liao, R., Voskuil, M. I., Tompa, M., et al. (2003). Rv3133c/dosR is a transcription factor that mediates the hypoxic response of Mycobacterium tuberculosis. *Mol. Microbiol.* 48 (3), 833–843. doi: 10.1046/j.1365-2958.2003.03474.x

Portal-Celhay, C., Tufariello, J. M., Srivastava, S., Zahra, A., Klevorn, T., Grace, P. S., et al. (2016). Mycobacterium tuberculosis EsxH inhibits ESCRT-dependent CD4+ T-cell activation. *Nat. Microbiol.* 2, 16232. doi: 10.1038/nmicrobiol.2016.232

Priscille, B., Ida, R., Peter, A., Cole, S. T., and Roland., B. (2004). ESAT-6 proteins: protective antigens and virulence factors? *Trends Microbiol.* 12 (11), 500–508. doi: 10.1016/j.tim.2004.09.007

Rustad, T. R., Harrell, M. I., Liao, R., and Sherman., D. R. (2008). The enduring hypoxic response of Mycobacterium tuberculosis. *PLoS One* 3 (1), e1502. doi: 10.1371/journal.pone.0001502

Rustad, T. R., Sherrid, A. M., Minch, K. J., and Sherman., D. R. (2009). Hypoxia: a window into Mycobacterium tuberculosis latency. *Cell Microbiol.* 11 (8), 1151–1159. doi: 10.1111/j.1462-5822.2009.01325.x

Sandra, A. S. R., Facey, P. D., Lorena, F. M., Caridad, R., Carlos, V., Ricardo, D. S., et al. (2010). A heterodimer of EsxA and EsxB is involved in sporulation and is secreted by a type VII secretion system in Streptomyces coelicolor. *Microbiology* 156 (Pt 6), 1719. doi: 10.1099/mic.0.037069-0

Sharma, A., Singh, K., and Kaur., J. (2017). MesT, a unique epoxide hydrolase, is essential for optimal growth of Mycobacterium tuberculosis in the presence of styrene oxide. *Future Microbiol.* 00 (00), 527–546. doi: 10.2217/fmb-2016-0206

Sherman, D. R., Voskuil, M., Schnappinger, D., Liao, R., Harrell, M. I., and Schoolnik., G. K. (2001). Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. *Proc. Natl. Acad. Sci. U. S. A.* 98 (13), 7534–7539. doi: 10.1073/pnas.121172498

Stewart, G. R., Robertson, B. D., and Young., D. B. (2003). Tuberculosis: a problem with persistence. *Nat. Rev. Microbiol.* 1 (2), 97–105. doi: 10.1038/nrmicro749

Sun, X., Zhang, L., Jiang, J., Ng, M., Cui, Z., Mai, J., et al. (2018). Transcription factors Rv0081 and Rv3334 connect the early and the enduring hypoxic response of Mycobacterium tuberculosis. *Virulence* 9 (1), 1468. doi: 10.1080/21505594.2018.1514237

Sweeney, K. A., Dao, D. N., Goldberg, M. F., Tsungda, H., Venkataswamy, M. M., Marcela, H. T., et al. (2011). A recombinant Mycobacterium smegmatis induces potent bactericidal immunity against Mycobacterium tuberculosis. *Nat. Med.* 17 (10), 1261. doi: 10.1038/nm.2420

Talaat, A. M., Rick, L., Howard, S. T., and Stephen Albert., J. (2004). The temporal expression profile of Mycobacterium tuberculosis infection in mice. *Proc. Natl. Acad. Sci. U. S. A.* 101 (13), 4602–4607. doi: 10.1073/pnas.0306023101

Tobin, D. M., and Ramakrishnan., L. (2008). Comparative pathogenesis of Mycobacterium marinum and Mycobacterium tuberculosis. *Cell Microbiol.* 10 (5), 1027–1039. doi: 10.1111/j.1462-5822.2008.01133.x

Upton, G. J. G. (1992). Fisher's exact test. *J. Royal Stat. Soc.* 155 (3), 395–402.

Veatch, A. V., and Kaushal., D. (2018). Opening Pandora's box: mechanisms of mycobacterium tuberculosis resuscitation. *Trends Microbiol.* 26 (2), 145. doi: 10.2307/2982890

Veatch, A. V., Niu, T., Caskey, J., Mcgillivray, A., Gautam, U. S., Subramanian, R., et al. (2016). Sequencing-relative to hybridization-based transcriptomics approaches better define Mycobacterium tuberculosis stress-response regulons. *Tuberculosis* 101S, S9–S17.

Ward, S. K., Hoye, E. A., and Talaat., A. M. (2008). The Global responses of mycobacterium tuberculosis to physiological levels of copper. *J. Bacteriol.* 190 (8), 2939–2946. doi: 10.1128/JB.01847-07

Wayne, L. G., and Hayes., L. G. (1996). An *in vitro* model for sequential study of shiftdown of Mycobacterium tuberculosis through two stages of nonreplicating persistence. *Infect. Immun.* 64 (6), 2062–2069. doi: 10.1016/1380-2933(96)00040-1

Wayne, L. G., and Sohaskey., C. D. (2001). Nonreplicating persistence of Mycobacterium tuberculosis. *Annu. Rev. Microbiol.* 55, 139–163. doi: 10.1146/annurev.micro.55.1.139

Wayne, L. G. (1977). Synchronized replication of Mycobacterium tuberculosis. *Infect. Immun.* 17 (3), 528–530. doi: 10.1128/IAI.17.3.528-530.1977

Wu, J., Ru, H., Xiang, Z., Jiang, J., Wang, Y., Zhang, L., et al. (2017). WhiB4 regulates the PE/PPE gene family and is essential for virulence of mycobacterium marinum. *Sci. Rep.* 7 (1), 3007. doi: 10.1038/s41598-017-03020-4

Yu, J., Tran, V., Li, M., Huang, X., Niu, C., Wang, D., et al. (2012). Both phthiocerol dimycocerosates and phenolic glycolipids are required for virulence of Mycobacterium marinum. *Infect. Immun.* 80 (4), 1381–1389. doi: 10.1128/IAI.06370-11

Zhang, F., and Xie., J. P. (2011). Mammalian cell entry gene family of Mycobacterium tuberculosis. *Mol. Cell. Biochem.* 352 (1-2), 1–10. doi: 10.1007/s11010-011-0733-5

Zhang, Q., Wang, D., Jiang, G., Liu, W., Deng, Q., Li, X., et al. (2016). EsxA membrane-permeabilizing activity plays a key role in mycobacterial cytosolic translocation and virulence: effects of single-residue mutations at glutamine 5. *Sci. Rep.* 6 (1), 32618. doi: 10.1038/srep32618

Zhigang, X. U., Dezhou, L. I., Chen, X., Duan, Z., Mao, J., and Wen., J. (2017). Identification and application of Mycobacterium tuberculosis esxN-specific cell epitopes in the diagnosis of pulmonary tuberculosis. *Chin. J. Cell. Mol. Immunol.* 33 (12), 1686–1691.

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: 11 February 2020
doi: 10.3389/fgene.2020.00027

Identification of the Prognosis-Related lncRNAs and Genes in Gastric Cancer

*Xiaohui Su**, *Jianjun Zhang*, *Wei Yang*, *Yanqing Liu*, *Yang Liu*, *Zexing Shan* **and** *Wentao Wang*

Department of Gastric Surgery, Cancer Hospital of China Medical University, Liaoning, China

Gastric cancer is a common malignant tumor with high occurrence and recurrence and is the leading cause of death worldwide. However, the prognostic value of protein-coding and non-coding RNAs in stage III gastric cancer has not been systematically analyzed. In this study, using TCGA data, we identified 585 long noncoding RNAs (lncRNAs) and 927 protein-coding genes (PCGs) correlated with the overall survival rate of gastric cancer. Functional enrichment analysis revealed that the prognostic genes positively correlated with death rates were enriched in pathways, including gap junction, focal adhesion, cell adhesion molecules (CAMs), and neuroactive ligand-receptor interaction, that are involved in the tumor microenvironment and cell-cell communications, suggesting that their dysregulation may promote the tumor progression. To evaluate the performance of the prognostic genes in risk prediction, we built three multivariable Cox models based on prognostic genes selected from the prognostic PCGs and lncRNAs. The performance of the three models based on features from only PCGs or lncRNAs or from all prognostic genes were systematically compared, which revealed that the features selected from all the prognostic genes showed higher performance than the features selected only from lncRNAs or PCGs. Furthermore, the multivariable Cox regression analysis revealed that the stratification with the highest performance was an independent prognostic factor in stage III gastric cancer. In addition, we explored the underlying mechanism of the prognostic lncRNAs in the Cox model by predicting the lncRNA and protein interaction. Specifically, *CTD-2218G20.2* was predicted to interact with *PSG4, PSG5,* and *PSG7*, which could also interact with cancer-related proteins, including *KISS1, TIMP2, MMP11, IGFBP1, EGFR,* and *CDKN1C*, suggesting that CTD-2218G20.2 might participate in the cancer progression *via* these cancer-related proteins. In summary, the systematic analysis of the prognostic lncRNAs and PCGs was of great importance to the understanding of the progression of stage III gastric cancer.

Keywords: gastric cancer, long noncoding RNA, prognostic, TCGA, model

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Xialin Yan,
Tongji University, China
Yingbin Liu,
Xinhua Hospital, China
Yang Yang
Shanghai Jiaotong University, China,
in collaboration with reviewer YL

****Correspondence:***
Xiaohui Su
yazhan775169617@163.com

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

Received: *07 November 2019*
Accepted: *08 January 2020*
Published: *11 February 2020*

Citation:
Su X, Zhang J, Yang W, Liu Y, Liu Y, Shan Z and Wang W (2020) Identification of the Prognosis-Related lncRNAs and Genes in Gastric Cancer. Front. Genet. 11:27. doi: 10.3389/fgene.2020.00027

INTRODUCTION

Gastric cancer is one of the most commonly diagnosed cancers worldwide, with an incidence of 1,031,700 new cases in 2018 and poor survival rates, causing approximately 787,200 deaths that year (Bray et al., 2018). Incidence rates of gastric cancer exhibit significant differences among regions, as its rates in Eastern Asia are markedly higher than those in Northern America and Northern Europe, and about 70% of gastric cancer is reported in developing countries with a higher mortality ratio, reflecting the importance of modern surgical and medical technology in gastric cancer treatment (Guggenheim and Shah, 2013). Environmental risk factors for gastric cancer include Helicobacter pylori infection, tobacco and alcohol use, and dietary salt intake (Zhang and Zhang, 2017), while genetic studies have revealed several key genetic factors in gastric cancer, including chromosomal instability, changes in microRNA profile, and somatic gene mutations (McLean and El-Omar, 2014).

According to the TNM system, most of GC patients are suffering from stage III or stage IV disease (Washington, 2010; Coburn et al., 2018). Surgery may seem to be the only approach to ensure long-time survival; however, for patients who have undergone surgical resection, the recurrence-free survival time remains poor, with a median length shorter than two years (Spolverato et al., 2014; Chan et al., 2016). Although adjuvant radiotherapy and chemotherapy are utilized to reduce its recurrence after surgery, the five-year survival rate for all stages is still unsatisfying, as it merely becomes 65% for patients with stage I disease, and the situation is much worse for patients with more advanced stages (Spolverato et al., 2014).

The discovery of biomarkers will greatly help deliver personalized treatment, with the goal of reducing gastric cancer recurrence and mortality rates. Currently, most studies investigating biomarkers in gastric cancer focus on protein-coding genes (PCGs), but noncoding RNAs are less addressed (Nagarajan et al., 2012). Though a growing number of long noncoding RNAs (lncRNAs), including *HOTAIR, MEG3, MALAT1, H19, GAPLINC,* and *GClnc1*, have been reported to be associated with gastric cancer tumorigenesis, the role of lncRNAs in human gastric cancer and their prognostic value are still inadequately explored (Kogo et al., 2011; Gu et al., 2015; Sun et al., 2016). Furthermore, the performance of the protein-coding genes and lncRNAs in risk prediction has not been systematically compared in gastric cancer. In the present study, we collected gene expression data for stage III gastric cancer and aimed to identify key prognostic lncRNAs in gastric cancer. Moreover, we built a Cox model based on features from both protein-coding genes and lncRNAs and compared the performance of Cox models based on features from prognostic lncRNAs, from pPCGs, and from all prognostic genes. The systematic analysis of the prognostic lncRNAs and PCGs is of great importance for the understanding of the progression of stage III gastric cancer.

MATERIALS AND METHODS

TCGA Gene Expression Data Collection and Processing

The gene expression data of TCGA stomach adenocarcinoma (TCGA-STAD) and the associated clinical data were downloaded from the UCSC Xena database (Goldman et al., 2018) (https://xenabrowser.net/datapages/). Samples diagnosed with TNM stage III in TCGA-STAD were selected for the downstream data analysis. Each gene was discretized as of high and low expression status if its expression level was higher or lower than the median, respectively. The survival analysis was conducted based on the discretized expression status.

Overrepresentation Enrichment Analysis (ORA)

To characterize the prognostic genes, we employed overrepresentation enrichment analysis (ORA), which was implemented by R package *clusterProfiler* with the *enrichKEGG* and *enricher* functions (Yu et al., 2012). The gene sets used for the enrichment analysis of the lncRNA interacting proteins were collected from MSigDB gene sets (Liberzon et al., 2011) (http://software.broadinstitute.org/gsea/index.jsp). The significant pathways were selected based on a threshold of 0.05 for adjusted *P*-value.

Cox Proportional Hazards Regression Analysis

The two-sample comparisons of overall survival were performed by Cox proportional hazards regression analysis and the differences tested by log-rank test, implemented using the R package *survival* with the *coxph* function. The predicted risk score for the patients was calculated based on the expression status of the prognostic genes, implemented in R with the *predict.coxph* function. Particularly, the features (prognostic genes) were selected by the Maximum Minimum Parents and Children (MMPC) algorithm (Lagani et al., 2016) and implemented by the *MXM* package in R.

lncRNA–Protein Interaction Analysis

To predict the potential lncRNA–protein interactions, we used the pre-trained LncADeep (Yang et al., 2018) model, a deep learning model, and utilized the sequences of differentially expressed lncRNAs and proteins to predict their interactions. In addition, we also conducted Pearson correlation analysis between the lncRNAs and proteins, with a threshold of 0.3 for Pearson correlation coefficients (PCC).

Statistical Analysis

The statistical analyses were conducted in R programming software, version 3.6.0. The two-sample or multiple-sample comparisons were performed using the Wilcoxon rank-sum test or analysis of variance (ANOVA). *P*-value < 0.05 was considered statistically significant difference.

RESULTS

Identification of Prognostic Genes by a Univariable Cox Model

To identify the prognostic genes, including the long noncoding RNAs (lncRNAs) and protein-coding genes (PCGs), we collected 152 stage III gastric cancer samples from the TCGA gastric cancer cohort. The univariable survival analysis was then conducted on all the genes with stable expression (FPKM > 1 in 10% of samples). In total, we identified 585 lncRNAs and 927 PCGs correlated with overall gastric cancer survival (**Supplementary Table S1**). Notably, 57.95% of PCGs and 68.72% of lncRNAs positively correlated with death rates in the Cox models were identified (**Figure 1A**), suggesting that these genes might drive the cancer progression. The two proportions showed significant difference (two-sample proportion test, $P < 0.05$), which might be caused by the relatively lower expression of lncRNAs. Moreover, we also investigated the distribution of the prognostic gene expression levels. The prognostic PCGs had significantly higher expression than the prognostic lncRNAs (**Figure 1B**). As shown in **Figure 1C**, the top five genes positively and negatively correlated with death rates included *RP13-577H12.2, AJ239318.1, CLDN9, OLFML2A, RP11-1021N1.1, CTD-2218G20.2, LMNB2, RP11-291L22.4, SRSF7,* and *PPP1R15B*. Notably, *RP13-577H12.2, CTD-2218G20.2,* and *RP11-291L22.4* were prognostic lncRNAs.

Functional Characterization of the Prognostic Genes

To characterize the functions of the prognostic genes, the prognostic genes positively or negatively correlated with death rates were subjected to KEGG enrichment analysis. The genes promoting the progression of gastric cancer were enriched in pathways such as adrenergic signaling in cardiomyocytes, axon guidance, gap junction, insulin secretion, the cAMP signaling pathway, bladder cancer, focal adhesion, cell adhesion molecules (CAMs), the PI3K-

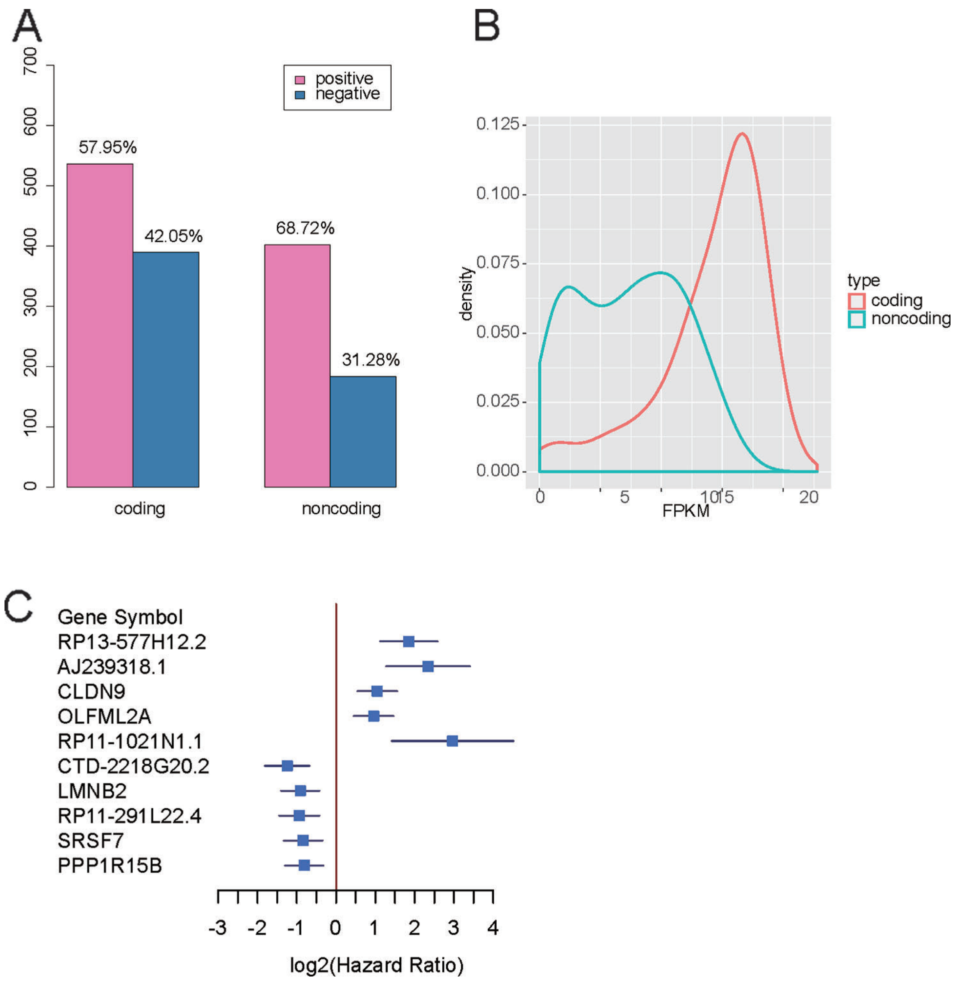


FIGURE 1 | Overview of the prognostic protein-coding genes (PCGs) and long noncoding RNAs (lncRNAs). **(A)** Proportions of prognostic PCGs and lncRNAs positively and negatively correlated with death rates. **(B)** Distribution of expression levels for the PCGs and lncRNAs. **(C)** Forest plot for the top five genes positively and negatively correlated with death rates.

Akt signaling pathway, and neuroactive ligand-receptor interaction (**Figure 2**). In contrast, the genes with higher expression in samples with better prognosis were enriched in base excision repair, transcriptional misregulation in cancer, breast cancer, Fanconi anemia pathway, pancreatic cancer, platinum drug resistance, RNA transport, hepatitis C, homologous recombination, and spliceosome (**Figure 2**). It should be noted that the gap junction, focal adhesion, cell adhesion molecules (CAMs), and neuroactive ligand-receptor interaction pathways were involved in tumor microenvironments and cell-cell communications, suggesting that their dysregulation may promote the tumor progression.

The Performance of the Prognostic lncRNAs and PCGs in Risk Prediction

To evaluate the performance of the prognostic genes in risk prediction, we first selected features from the lncRNAs, PCGs, and all genes, respectively, with a significance level of 0.01 using the MMPC algorithm. Specifically, 10 lncRNAs and seven PCGs were selected for the construction of Cox models based on only lncRNAs or PCGs (**Figures 3A, B**). Additionally, another nine genes including five lncRNAs and four PCGs were selected to build the model under both lncRNAs and PCGs (**Figure 3C**). As shown in **Figure 3**, the risk groups stratified by the three Cox models showed significantly different overall survival ($P <$ 0.0001), and the selected features were highly correlated with the risk. Furthermore, we also compared the performance of the three models based on the criteria of log-rank test, Wald test, and C-index (**Table 1**). Consistently, the features selected from all the prognostic genes showed higher performance than the features selected only from lncRNAs or PCGs (**Table 1**), suggesting that stratification by feature by integrating PCGs and lncRNAs was superior to using either of the two alone.

TABLE 1 | Performance of three Cox models based on features selected from all genes, PCGs, and lncRNAs.

Features	logtest.pvalue	waldtest.pvalue	C-index	sd(C-index)
All genes	3.56E-20	1.48E-16	0.84	0.04
PCGs	2.73E-16	8.20E-14	0.80	0.04
LncRNAs	2.99E-18	8.59E-14	0.83	0.04

The Stratification Based on the Features From All Genes Is an Independent Prognostic Factor in Stage III Gastric Cancer

As the prognostic model based on the features from all genes exhibited satisfying performance on all stage III gastric cancer patients, it was also necessary to investigate whether this stratification was a prognostic factor independent of clinical indicators such as age, gender, race, and histology grade. The multivariable Cox regression model was then constructed by group and these co-factors. We observed that both age and group were significantly associated with stage III gastric cancer survival (P < 0.05). Remarkably, the group had the highest statistical significance (P = 1.54E-14), suggesting that the stratification based on the features from all genes was an independent prognostic factor in stage III gastric cancer (**Table 2**).

Prediction of the Underlying Mechanism of the lncRNAs in the Cox Model

As the lncRNAs could perform their function by interacting with proteins, we then predicted the interactions between the

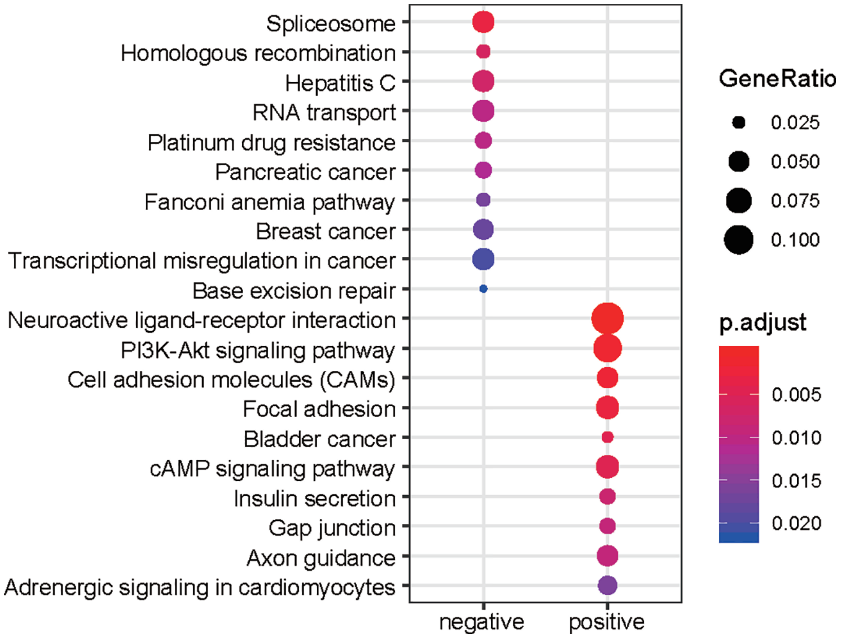


FIGURE 2 | KEGG enrichment of the prognostic genes. The node size represents the ratio of the genes in the pathway. The colors represent the statistical significance of the pathways.

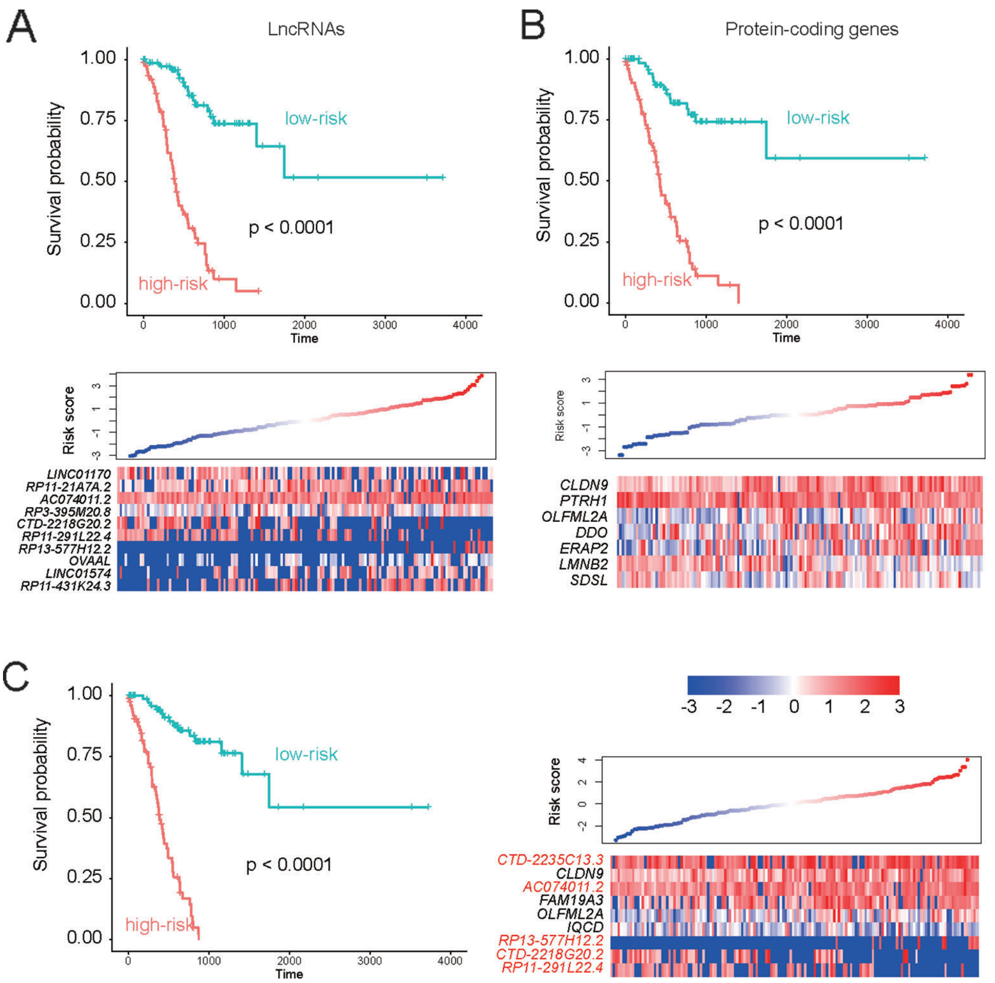


FIGURE 3 | Performance of the three Cox models in risk prediction. Performance of three Cox models based on the features selected from only the prognostic lncRNAs, only the prognostic PCGs, and all the prognostic genes are displayed in **(A–C)**, respectively.

TABLE 2 | Multivariable Cox model with age, gender, race, and histology grade as co-factors.

Variables		Coef	exp(Coef)	se(Coef)	Z	P
Age		1.12E-04	1.00E+00	3.86E-05	2.896	0.00378
Gender	Male	4.12E-01	1.51E+00	2.69E-01	1.532	0.12547
Race	Black or African American	-1.13E-01	8.93E-01	7.00E-01	-0.162	0.87143
	White	4.04E-01	1.50E+00	3.83E-01	1.055	0.29138
Histology grade	G2	-1.82E-01	8.34E-01	1.05E+00	-0.173	0.86275
	G3	-8.96E-02	9.14E-01	1.03E+00	-0.087	0.93095
	GX	-1.54E+01	2.10E-07	3.40E+03	-0.005	0.99639
Group	Low-risk	-2.88E+00	5.59E-02	3.75E-01	-7.684	1.54E-14

prognostic lncRNAs in the Cox model and proteins using a deep learning method, LncADeep. Moreover, we also conducted a correlation analysis between the proteins and lncRNAs. However, only one of the five lncRNAs in the Cox model, *CTD-2218G20.2*, was predicted to interact with 86 proteins (Pearson correlation coefficient, PCC > 0.3). The gene set enrichment analysis revealed that these interacting proteins also interacted with cancer-related proteins, including *KISS1, TIMP2, MMP11, IGFBP1, EGFR*, and *CDKN1C* (**Figure 4A**). Specifically, pregnancy-specific glycoproteins, including *PSG4, PSG5*, and *PSG7*, were those proteins jointly interacting with *CTD-2218G20.2* and cancer-related proteins, which were highly correlated with *CTD-2218G20.2* (**Figure 4B**, PCC > 0.3). These results suggested that *CTD-2218G20.2* might participate in the cancer progression *via* these cancer-related proteins.

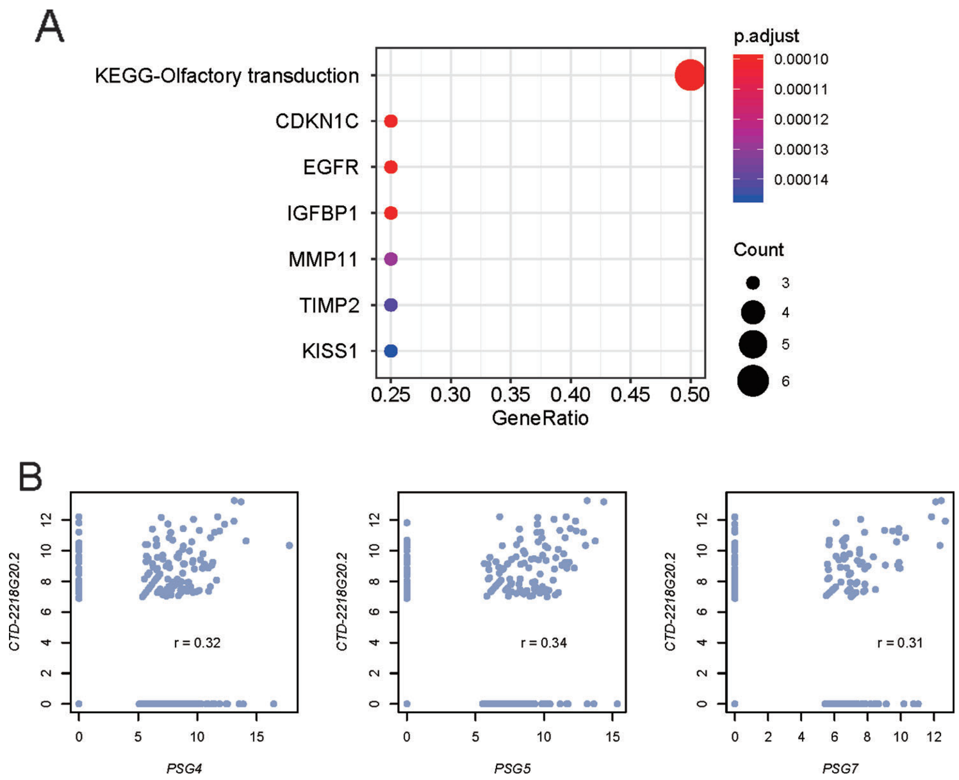


FIGURE 4 | Predicted function of the prognostic lncRNA CTD-2218G20.2. **(A)** Gene sets enriched by the proteins predicted to interact with CTD-2218G20.2. **(B)** Correlation analysis between CTD-2218G20.2 and three pregnancy-specific glycoproteins. The x- and y-axes represent the expression levels (log2 (FPKM+1)) of the PSG genes and CTD-2218G20.2, respectively.

DISCUSSION

Gastric cancer is a common malignant tumor with high occurrence and recurrence and is the leading cause of death worldwide (Bray et al., 2018). However, the prognostic value of protein-coding and non-coding RNAs in stage III gastric cancer has not been systematically analyzed. In this study, we identified 585 lncRNAs and 927 PCGs correlated with the overall survival rate of gastric cancer. Notably, 57.95% of PCGs and 68.72% of lncRNAs were positively correlated with the death rate in the Cox models (**Figure 1A**). To characterize the function of the prognostic genes, the prognostic genes positively or negatively correlated with death rates were subjected to KEGG enrichment analysis. Notably, the pathways of gap junction, focal adhesion, cell adhesion molecules (CAMs), and neuroactive ligand-receptor interaction were involved in the tumor microenvironments and cell-cell communications, suggesting that their dysregulation may promote the tumor progression. In accordance with previous studies (Wang et al., 2013; Yan et al., 2018; Zhao et al., 2019), the genes in gap junction, focal adhesion, and CAMs were significantly associated with gastric cancer prognosis. In addition, PI3K/Akt signaling pathway has been widely reported to regulate the tumorigenesis and progression (Tapia et al., 2014; Matsuoka and Yashiro, 2014) and act as a potential therapeutic target in gastric cancer (Ye et al., 2012; Singh et al., 2015).

To evaluate the performance of the prognostic genes in risk prediction, we built three Cox models based on prognostic lncRNAs, PCGs, and both (**Figure 3C**). The performances of the three models were systematically compared based on the criteria of log-rank test, Wald test, and C-index, which revealed that the features selected from all the prognostic genes showed higher performance than the features selected only from lncRNAs or PCGs. Furthermore, we investigated whether the stratification with the highest performance was a prognostic factor independent of clinical indicators, such as age, gender, race, and histology grade. The multivariable Cox regression analysis revealed that the stratification had the highest statistical significance ($P = 1.54E-14$), suggesting that the stratification based on the features from all genes was an independent prognostic factor in stage III gastric cancer. In addition, as the CEA and CA19-9 are commonly used biomarkers for gastric cancer risk prediction, we compared their prognostic values with those of the genes included in the multivariable Cox model. The hazard ratios (HR) of CEA and CA19-9 were estimated as 1.681 and 1.83 by meta-analysis (Song et al., 2015; Deng et al., 2015). However, the HR of CTD-2218G20.2 in the multivariable Cox model reached 3.48, suggesting that the lncRNA CTD-2218G20.2 was superior to the common clinical biomarkers like serum CEA and CA19-9.

Furthermore, we explored the underlying mechanism of the prognostic lncRNAs in the Cox model by predicting the lncRNA and protein interaction. Specifically, CTD-2218G20.2 was predicted to interact with 86 proteins (Pearson correlation coefficient, PCC > 0.3), some of which, including *PSG4, PSG5,* and *PSG7*, could also interact with cancer-related proteins, including *KISS1, TIMP2, MMP11, IGFBP1, EGFR*, and *CDKN1C* (**Figure 4A**). Notably, *KISS1, TIMP2, MMP11, IGFBP1,* and *EGFR* have been reported to be involved in the metastasis of gastric cancer (Guan-Zhen et al., 2007; Kou et al., 2013; Wang et al., 2017; Wang et al., 2018; Sato et al., 2019). These results suggested that *CTD-2218G20.2* might participate in the cancer progression *via* these cancer-related proteins.

The present study still had some limitations, such as lack of experimental validation or large sample size. However, we aimed to discover some key prognostic PCGs and lncRNAs in stage III gastric cancer that could not be extrapolated to early stage GC patients. In summary, this systematic analysis of the prognostic lncRNAs and PCGs was of great importance to the understanding of the progression of stage III gastric cancer.

DATA AVAILABILITY STATEMENT

All datasets generated and analyzed for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

No identifiable data is present in this paper.

AUTHOR CONTRIBUTIONS

Conception and design: XS. Administrative support: XS. Provision of study materials or patients: XS, JZ, and WY. Collection and assembly of data: all authors. Data analysis and interpretation: all authors. Manuscript writing: all authors. Final approval of manuscript: all authors.

FUNDING

This work was funded by a Key Project of Natural Science Foundation of Liaoning Province (NO. 20170540566).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00027/full#supplementary-material

REFERENCES

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68 (6), 394–424. doi: 10.3322/caac.21492

Chan, B. A., Jang, R. W., Wong, R. K., Swallow, C. J., Darling, G. E., and Elimova, E. (2016). Improving outcomes in resectable gastric cancer: a review of current and future strategies. *Oncol. (Williston Park)* 30 (7), 635–645.

Coburn, N., Cosby, R., Klein, L., Knight, G., Malthaner, R., Mamazza, J., et al. (2018). Staging and surgical approaches in gastric cancer: a systematic review. *Cancer Treat Rev.* 63, 104–115. doi: 10.1016/j.ctrv.2017.12.006

Deng, K., Yang, L., Hu, B., Wu, H., Zhu, H., and Tang, C. (2015). The prognostic significance of pretreatment serum CEA levels in gastric cancer: a meta-analysis including 14651 patients. *PloS One* 10 (4), e0124151. doi: 10.1371/journal.pone.0124151

Goldman, M., Craft, B., Kamath, A., Brooks, A., Zhu, J., and Haussler, D. (2018). The UCSC xena platform for cancer genomics data visualization and interpretation 326470. doi: 10.1101/326470

Gu, Y., Chen, T., Li, G., Yu, X., Lu, Y., Wang, H., et al. (2015). LncRNAs: emerging biomarkers in gastric cancer. *Future Oncol.* 11 (17), 2427–2441. doi: 10.2217/fon.15.175

Guan-Zhen, Y., Ying, C., Can-Rong, N., Guo-Dong, W., Jian-Xin, Q., and Jie-Jun, W. (2007). Reduced protein expression of metastasis-related genes (nm23, KISS1, KAI1 and p53) in lymph node and liver metastases of gastric cancer. *Int. J. Exp. Pathol.* 88 (3), 175–183. doi: 10.1111/j.1365-2613.2006.00510.x

Guggenheim, D. E., and Shah, M. A. (2013). Gastric cancer epidemiology and risk factors. *J. Surg. Oncol.* 107 (3), 230–236. doi: 10.1002/jso.23262

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71 (20), 6320–6326. doi: 10.1158/0008-5472.CAN-11-1021

Kou, Y. B., Zhang, S. Y., Zhao, B. L., Ding, R., Liu, H., and Li, S. (2013). Knockdown of MMP11 inhibits proliferation and invasion of gastric cancer cells. *Int. J. Immunopathol. Pharmacol.* 26 (2), 361–370. doi: 10.1177/039463201302600209

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2016). Feature selection with the R package MXM. *F1000Res.* 2018 (7), 1505. doi: 10.12688/f1000research.16216.2

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27 (12), 1739–1740. doi: 10.1093/bioinformatics/btr260

Matsuoka, T., and Yashiro, M. (2014). The role of PI3K/Akt/mTOR signaling in gastric carcinoma. *Cancers (Basel)* 6 (3), 1441–1463. doi: 10.3390/cancers6031441

McLean, M. H., and El-Omar, E. M. (2014). Genetics of gastric cancer. *Nat. Rev. Gastroenterol. Hepatol.* 11 (11), 664–674. doi: 10.1038/nrgastro.2014.143

Nagarajan, N., Bertrand, D., Hillmer, A. M., Zang, Z. J., Yao, F., Jacques, P. E., et al. (2012). Whole-genome reconstruction and mutational signatures in gastric cancer. *Genome Biol.* 13 (12), R115. doi: 10.1186/gb-2012-13-12-r115

Sato, Y., Inokuchi, M., Takagi, Y., and Kojima, K. (2019). IGFBP1 Is a predictive factor for haematogenous metastasis in patients with gastric cancer. *Anticancer Res.* 39 (6), 2829–2837. doi: 10.21873/anticanres.13411

Singh, S. S., Yap, W. N., Arfuso, F., Kar, S., Wang, C., Cai, W., et al. (2015). Targeting the PI3K/Akt signaling pathway in gastric carcinoma: a reality for personalized medicine? *World J. Gastroenterol.* 21 (43), 12261–12273. doi: 10.3748/wjg.v21.i43.12261

Song, Y. X., Huang, X. Z., Gao, P., Sun, J. X., Chen, X. W., Yang, Y. C., et al. (2015). Clinicopathologic and prognostic value of Serum Carbohydrate Antigen 19-9 in gastric cancer: a meta-analysis. *Dis. Markers* 2015, 549843. doi: 10.1155/2015/549843

Spolverato, G., Ejaz, A., Kim, Y., Squires, M. H., Poultsides, G. A., Fields, R. C., et al. (2014). Rates and patterns of recurrence after curative intent resection for gastric cancer: a United States multi-institutional analysis. *J. Am. Coll. Surg.* 219 (4), 664–675. doi: 10.1016/j.jamcollsurg.2014.03.062

Sun, T. T., He, J., Liang, Q., Ren, L. L., Yan, T. T., Yu, T. C., et al. (2016). LncRNA GClnc1 promotes gastric carcinogenesis and may act as a modular scaffold of WDR5 and KAT2A complexes to specify the histone modification pattern. *Cancer Discovery* 6 (7), 784–801. doi: 10.1158/2159-8290.CD-15-0921

Tapia, O., Riquelme, I., Leal, P., Sandoval, A., Aedo, S., Weber, H., et al. (2014). The PI3K/AKT/mTOR pathway is activated in gastric cancer with potential prognostic and predictive significance. *Virchows Arch.* 465 (1), 25–33. doi: 10.1007/s00428-014-1588-4

Wang, Y. Y., Li, L., Zhao, Z. S., Wang, Y. X., Ye, Z. Y., and Tao, H. Q. (2013). L1 and epithelial cell adhesion molecules associated with gastric cancer progression and prognosis in examination of specimens from 601 patients. *J. Exp. Clin. Cancer Res.* 32, 66. doi: 10.1186/1756-9966-32-66

Wang, D., Wang, B., Wang, R., Zhang, Z., Lin, Y., Huang, G., et al. (2017). High expression of EGFR predicts poor survival in patients with resected T3 stage gastric adenocarcinoma and promotes cancer cell survival. *Oncol. Lett.* 13 (5), 3003–3013. doi: 10.3892/ol.2017.5827

Wang, W., Zhang, Y., Liu, M., Wang, Y., Yang, T., Li, D., et al. (2018). TIMP2 is a poor prognostic factor and predicts metastatic biological behavior in gastric cancer. *Sci. Rep.* 8 (1), 9629. doi: 10.1038/s41598-018-27897-x

Washington, K. (2010). 7th edition of the AJCC cancer staging manual: stomach. *Ann. Surg. Oncol.* 17 (12), 3077–3079. doi: 10.1245/s10434-010-1362-z

Yan, P., He, Y., Xie, K., Kong, S., and Zhao, W. (2018). In silico analyses for potential key genes associated with gastric cancer. *PeerJ* 6, e6092. doi: 10.7717/peerj.6092

Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., et al. (2018). LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* 34 (22), 3825–3834. doi: 10.1093/bioinformatics/bty428

Ye, B., Jiang, L. L., Xu, H. T., Zhou, D. W., and Li, Z. S. (2012). Expression of PI3K/AKT pathway in gastric cancer and its blockade suppresses tumor growth and metastasis. *Int. J. Immunopathol. Pharmacol.* 25 (3), 627–636. doi: 10.1177/039463201202500309

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 (5), 284–287. doi: 10.1089/omi.2011.0118

Zhang, X. Y., and Zhang, P. Y. (2017). Gastric cancer: somatic genetics as a guide to therapy. *J. Med. Genet.* 54 (5), 305–312. doi: 10.1136/jmedgenet-2016-104171

Zhao, X., Yu, C., Zheng, M., and Sun, J. (2019). Prognostic value of the mRNA expression of gap junction alpha members in patients with gastric cancer. *Oncol. Lett.* 18 (2), 1669–1678. doi: 10.3892/ol.2019.10516

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fbioe.2020.00081

miR-221-3p Delivered by BMMSC-Derived Microvesicles Promotes the Development of Acute Myelocytic Leukemia

Xuewu Zhang[†], *Yu Xu*[†], *Jinghan Wang, Shuqi Zhao, Jianhu Li, Xin Huang, Huan Xu, Xiang Zhang, Shanshan Suo, Yunfei Lv, Yi Zhang and Wenjuan Yu*[*]

Department of Hematology, Zhejiang University School of Medicine First Affiliated Hospital, Hangzhou, China

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Jinhang Zhu,
Shanghai Jiao Tong University, China
Minzhi Zhao,
National Center for Nanoscience and Technology (CAS), China

Correspondence:
Wenjuan Yu
1306023@zju.edu.cn

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

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

Received: *20 December 2019*
Accepted: *29 January 2020*
Published: *14 February 2020*

Citation:
Zhang X, Xu Y, Wang J, Zhao S, Li J, Huang X, Xu H, Zhang X, Suo S, Lv Y, Zhang Y and Yu W (2020) miR-221-3p Delivered by BMMSC-Derived Microvesicles Promotes the Development of Acute Myelocytic Leukemia. Front. Bioeng. Biotechnol. 8:81. doi: 10.3389/fbioe.2020.00081

Objective: The study aims to investigate the effects of miR-221-3p in bone marrow mesenchymal stem cell (BMMSC)-derived microvesicles (MVs) on cell cycle, proliferation and invasion of acute myelocytic leukemia (AML).

Methods: Bioinformatics was used to predict differentially expressed miRNAs (DEmiRNAs) in AML. The morphology of BMMSC-derived MVs was observed under an electron microscope, and the positional relation of MVs and OCI-AML2 cells was observed by a fluorescence microscope. MTT, Transwell, and flow cytometry assays were used to analyze the effects of MVs on OCI-AML2 cells. The targeted relationship between miR-221-3p and CDKN1C was detected by dual luciferase assay.

Results: It was verified that miR-221-3p promoted the proliferation, invasion and migration of OCI-AML2 cells, and induced the cell cycle arrest in G1/S phase as well as inhibited cell apoptosis. Further studies showed that MVs promoted the proliferation, migration and invasion of AML, and induced the cell cycle arrest in G1/S phase through miR-221-3p. It was confirmed that miR-221-3p can directly target CDKN1C to regulate cell cycle, proliferation and invasion of AML.

Conclusion: miR-221-3p in BMMSC-derived MVs regulated AML cell cycle, cell proliferation and invasion through targeting CDKN1C. miR-221-3p and CDKN1C were considered to be potential targets and biomarkers for the treatment of AML in clinic.

Keywords: BMMSC, microvesicles, miR-221-3p, AML, cell proliferation and invasion, cell cycle

INTRODUCTION

Acute myelocytic leukemia (AML) is a malignant tumor of abnormal clonal in immature myeloid hematopoietic cells with high heterogeneity, which is characterized by differentiation and maturation disorders along with block of apoptosis in clonal hematopoietic stem cells or progenitor cells, leading to malignant proliferation and accumulation of cells in the bone marrow, thus affecting normal hematopoiesis (Coombs et al., 2016; Khwaja et al., 2016). AML is the most-common acute leukemia in adults, but it predominantly occurs in older people (>60 years of age), with a median age at diagnosis of 67 (Coombs et al., 2016). It typically presents with a rapid onset of symptoms that are attributable to bone marrow failure and may be fatal within weeks or months when left untreated. Currently, chemotherapy and hematopoietic stem cell transplantation are the main treatments, but the success rate of AML cure remains low (Cornelissen and Blaise, 2016; Stein and Tallman, 2016). Therefore, it is particularly important to study the pathogenesis of AML and explore the possible therapeutic approaches.

Microvesicles (MVs) are extracellular vesicles between 100 nm and 1 μm that derived from normal cells and cancer cells. MVs can transfer proteins, glycoproteins, lipids, nucleic acids, and cytokines from maternal cells to recipient cells, promoting phenotype changes of recipient cells and playing an important role in intercellular communication (Hansen et al., 2014; Gopal et al., 2017; Abbasian et al., 2018). Studies have found that tumor-derived MVs (TMV) can interact directly with tumor cells and play a macro-messenger role to promote the transfer of molecular substances between tumor cells to facilitate tumor growth (Stec et al., 2015a,b). MVs derived from bone marrow mesenchymal stem cell (BMMSC) can promote tumorigenesis and development (Crompot et al., 2017; Boyiadzis and Whiteside, 2018). miRNAs in MVs, as post-transcriptional regulatory elements, directly regulate gene expression, target mRNA expression and translation or induce mRNA degradation to reduce protein synthesis by directly binding with the 3′-untranslation region (3′-UTR) of specific mRNA targets (Braicu et al., 2015; Jerez et al., 2019), ultimately induce multiple pathophysiological processes, such as leukemia stem cell formation, regulation of tumor cell proliferation, angiogenesis, invasion, metastasis, and immune escape to modulate leukemia development (Braicu et al., 2015; Del Principe et al., 2017).

microRNAs (miRNA) are a class of evolutionarily conserved 22 to 24-nucleotide small RNAs in length, which are widely found in eukaryotic cells with molecular functions to regulate cell differentiation, proliferation and apoptosis (Summerer et al., 2013; Zheng et al., 2013). miR-221-3p has important regulatory effects on a variety of cancers as an important miRNA. Studies have reported that in cervical squamous cell carcinoma, miR-221-3p in MVs promotes lymph angiogenesis and lymphatic metastasis by targeting VASH1 (Zhou et al., 2019), and promotes angiogenesis by targeting THBS2 (Wu et al., 2019). We previously found that miR-221-3p was significantly highly expressed in AML patients through bioinformatics, and miR-221-3p mainly existed in BMMSC-derived MVs. These results suggest that miR-221-3p in BMMSC-derived MVs has certain regulatory effects on AML cells.

Therefore, in this paper, we explored the regulatory effects and mechanism of miR-221-3p in BMMSC-derived MVs on cell cycle, cell proliferation and invasion of AML through *in vitro* experiments, so as to further understand the pathogenesis of AML and provide new ideas for future clinical diagnosis and treatment.

MATERIALS AND METHODS

Cell Lines and Patients

Normal human BMMSCs were purchased from Kunming cell bank, Chinese Academy of Sciences (No. 3153C0001000000244). BMMSCs were isolated from AML patients and human AML cells OCI-AML2 (BNCC341618) were purchased from BeNa Culture Collection (China).

Fifteen AML patients and 18 control samples (peripheral blood or bone marrow) were obtained with the informed consent of the patient or healthy subject and were collected at the First Affiliated Hospital of Zhejiang University through the protocol approved by the review committee.

Bioinformatics Analysis

AML-related miRNA expression dataset GSE49665 was obtained from GEO database (https://www.ncbi.nlm.nih.gov/geoprofiles/) to screen differentially expressed miRNAs (DEmiRNAs) and determine target miRNAs. Target miRNAs were found to be highly expressed in the MVs of fiber cells and mesenchymal stem cells (MSCs) via searching expression location In the EV miRNA database (http://bioinfo.life.hust.edu.cn/EVmiRNA). The downstream target genes of the target miRNAs were predicted by TargetScan database (http://www.targetscan.org/vert_72/), miRSearch database (https://www.exiqon.com/miRSearch), and mirDIP database (http://ophid.utoronto.ca/mirDIP/index.jsp), and differential analysis was conducted on AML gene expression in TCGA. The down-regulated genes in AML were selected to intersect with the predicted downstream target genes. Finally, the target genes with the most significant expression changes were detected by signaling pathway enrichment analysis.

Isolation, Culture and Analysis of BMMSC

BMMSCs were obtained by density gradient centrifugation. The bone marrow fluids were centrifuged at 1,000 rpm for 10 min, while the lipids and supernatant were absorbed and discarded. The remaining marrow fluids were added with equal quantity of PBS buffer and mixtured, centrifuged at 1,000 rpm for 10 min, and the supernatant was discarded. Then cell suspensions were prepared with 2 mL PBS buffer at a density of 4×10^7 cells, carefully superimposed on 5 mL Percoll separation solution (at a density of 1.077 g/mL), and centrifuged at 2,300 rpm for 30 min. After centrifugation, the liquids from top to bottom are: platelet and plasma diluent layer, yellow-brown annular cloud-like mononuclear cell layer, lymphocyte separation liquid layer, red blood cells and granulocyte layer. The mononuclear cell layer was absorbed and mixed with PBS buffer at a ratio of 1:2, and then centrifuged at 1,500 rpm for 10 min. All centrifugations were carried out at room temperature. The supernatant was discarded and cells were washed twice. 1×10^6 cells/mL were inoculated in a 25 cm^2 culture bottle with 5 mL BMMSCs medium (containing 10% fetal bovine serum, FBS). After 2–3 days, non-adhesive cells were removed, and monolayer adherent cells were spread to 70–80% of the bottom of the culture bottle. Cells were then isolated in a trypsin solution (0.25% trypsin/0.1% EDTA PBS solution, free of magnesium/magnesium and phenolic red) (Aurogene, Rome, Italy) and re-inoculated at a density of 3.5×10^3 cells/cm^2. The 3–5 generation cells were used for the experiment. Cell growth was analyzed by direct cell count at every passage.

Isolation and Identification of MVs

BMMSC-derived MVs were isolated using the exoEasy Maxi Kit (qiagen, Germany) according to the manufacturer's instructions.

MVs were observed by Philips CM120 BioTwin transmission electron microscope (FEI, USA).

Inhibition/Overexpression of miRNA and mRNA

miR-221-3p inhibitor, 100 nmol/L miR-221-3p mimic, 100 nmol/L overexpression of CDKN1C and the corresponding negative control (NC) were purchased from GenePharma (Shanghai, China). Approximately 1 × 10^5 cells were inoculated into 12-well plates during transfection. CDKN1C, miR-221-3p and negative control were transfected into the cells using LipoFiter kit (Hanbio, Shanghai, China) according to the kit instructions. RNA and proteins were extracted 48 h after transfection. The sequences of synthesized primers were shown in **Supplement Table 1**.

Construction of Lentivirus Expression Vector and Cell Transfection

Human miR-221-3p sequences were amplified and then bound to pcDNA3.1 (+) to form miR-221-3p expression vector (GenePharma, ShangHai, China). PcDNA3.1 carrier was used as blank control. Lentivirus coated miR-221-3p or blank lentivirus was transfected into OCI-AML2 cells and cultured for 96 h and then treated with puromycin for 4 weeks to screen cells (Santa Cruz organisms).

qRT-PCR

Total RNA was extracted from tissues and cells using Trizol (Invitrogen) according to the manufacturer's protocol. cDNA was synthesized using reverse transcription system kit (Invitrogen). qRT-PCR was performed on ABI 7900HT instrument (Applied Biosystems, USA). Quantitative PCR was performed using the miScript SYBR Green PCR Kit (Qiagen, Germany) under the following thermal cycling conditions: pre-denaturation at 95°C for 10 min, followed by 40 cycles of denaturation at 95°C for 2 min, annealing at 95°C for 5 s and extending at 60°C for 30 s. CDKN1C was normalized with β-Actin as an internal reference, and miR-221-3p was normalized with U6 as an internal reference. The relative expression of the target gene mRNAs in the control group and the experimental group were analyzed by $2^{-\Delta\Delta Ct}$ method. The primers used in the experiment were shown in **Supplement Table 1**.

Western Blot

Forty-eight hours after transfection of cells from different treatment groups, the cells were washed three times with cold PBS (Thermo fisher, USA), and lysed on ice using whole protein lysate for 10 min. BCA quantitative kit (Thermo fisher, USA) was used for protein quantification, then 10 µl loading buffer was added and proteins were boiled at 95°C for 10 min. the proteins were loaded onto SDS-PAGE at 100 V and transferred to the NC membrane blocked with 5% BSA/TBST for 60 min. The membrane was incubated with primary antibodies at 4°C overnight and then washed with 1 × TBST solution (Solarbio, Beijing, China) at room temperature for 5 min × 3 times. the membrane was probed with HRP labeled goat-anti-rabbit IgG at room temperature for 120 min, and washed by TBST for three times. After each 20 min, the ECL kit (Solarbio, Beijing, China) was used for detecting luminescence reaction, and the protein blot was photographed and observed. The antibodies used in experiment were listed in **Supplement Table 2**.

MTT Assay

OCI-AML2 cells (5 × 10^3 cells/100 ul) were seeded into 96-well plates. Each group was made in triplicate. Proliferation of cells were evaluated by sterile MTT solution (Beyptime) according to the instructions after culture for 12, 24, 48, and 72 h, respectively. Absorbance at 490 nm was measured using a spectrophotometer (Molecular Devices, Sunnyvale, CA, USA).

Transwell Assays

Transwell migration assay was used to evaluate the migration ability of OCI-AML2 cells. 24-well Transwell Chambers (8 µm aperture, BD Biosciences) were used. For migration assay, cells at a density of 1 × 10^5 cells/chamber were seeded into the upper chamber and the 600 µL of medium containing 10% FBS (Thermo fisher, USA) was placed in the lower chamber. For invasion assay, ∼2 × 10^4 cells/chamber were seeded in the upper chamber, which was coated with Matrigel. Dulbecco's modified Eagle culture medium (DMEM) containing 10% FBS (Thermo fisher, USA) was filled into the lower chamber. After incubation at 37°C for 48 h, the cells that were not migrated/invaded were cleared away with a cotton swab and the migrated/invaded cells on the lower side were stained with 0.5% crystal violet. Cells were observed under a microscope, and photographed.

Flow Cytometry (FCM)

Cell cycle detection: OCI-AML2 cells in growth phase were added with 3 mL PBS and harvested with 1 mL trypsin for 1–5 min after removing the liquid. The cell suspension was prepared by adding 5 mL PBS, and then transferred to a 15 mL centrifuge tube for centrifugation at 1,500 rpm for 5 min to discard the supernatant. Five hundred microliters PBS was added for cell suspension, and 2 mL of cold ethanol of 95% at 20°C was added to the suspension. After mixing, the suspension was fixed for 30 min. Five milliliters PBS was added and centrifuged at 1,500 rpm for 5 min to remove the supernatant and then added with 5 mL PBS and centrifuged at 1,500 rpm for 5 min to discard the supernatant. The cells were stained with 800 µL PI at room temperature for 30 min in darkness. Cell cycle was detected by FCM.

Dual Luciferase Assay

In order to determine the binding probability of miR-221-3p and 3′UTR of CDKN1C, a psiCHECK luciferase reporter vector (Sangon Co., LTD, ShangHai, China) was inserted into 3′UTR of CDKN1C wild type (WT) and mutated type (MUT). HEK293T cells (Thermo fisher, USA) were then inoculated in a 48-well plate and cultured for 24 h. miR-221-3p/NC and psiCHECK WT/MUT plasmids were co-transfected into cells. Finally, luciferase activity was measured by luciferase assay reagent (Promega, Fitchburg, WI, USA).

Statistical Analysis

All data were processed by SPSS 22.0 statistical software. The measurement data were expressed as mean ± standard deviation.

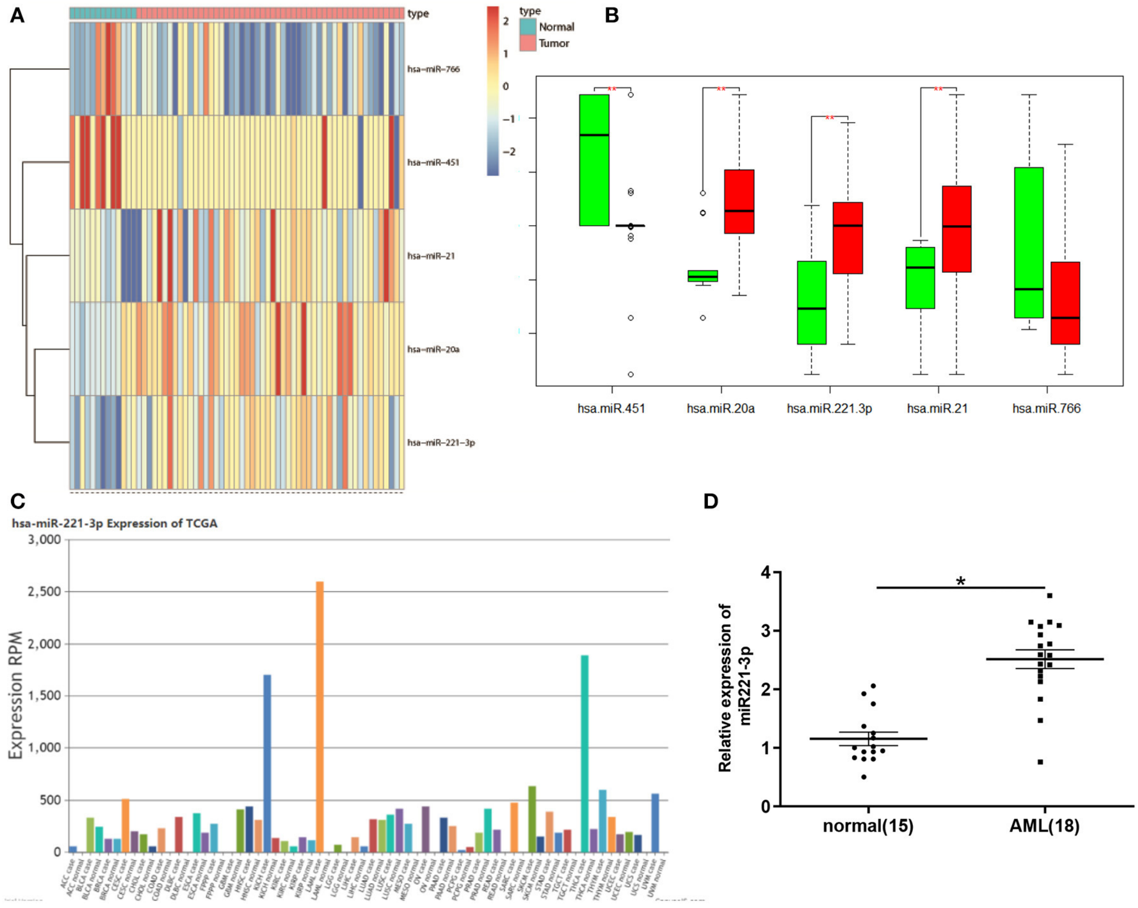


FIGURE 1 | miR-221-3p is highly expressed in peripheral blood of patients with leukemia. **(A)** Heat map of DEmiRNAs in GSE49665 dataset. **(B)** Boxplot of DEmiRNA. **(C)** Expression levels of miR-221-3p in different cancer species in TCGA database. **(D)** miR-221-3p was significantly overexpressed in AML patients.

The comparison between the two groups was analyzed by *t*-test, in which * stood for $P < 0.05$.

RESULTS

miR-221-3p Is Highly Expressed in Peripheral Blood of Patients With AML

Bioinformatics analysis found that in the miRNA expression dataset GSE49665 of AML patients in the GEO database, 5 DEmiRNAs were obtained and the expression of miR-221-3p changed most significantly (**Figures 1A,B**). At the same time, we detected the expression level of miR-221-3p in various cancers in TCGA database and found that its expression was most significant in AML (**Figure 1C**), so we chose miR-221-3p for follow-up study. In order to further confirm the high expression of miR-221-3p in the peripheral blood of AML patients, we used qRT-PCR to detect the expression of miR-221-3p in the peripheral blood of 15 normal people and 18 AML patients, and discovered that miR-221-3p was highly expressed in the peripheral blood of AML patients (**Figure 1D**), which was consistent with the bioinformatics results.

miR-221-3p Regulates AML Cell Cycle, Proliferation, and Invasion

miR-221-3p was overexpressed in OCI-AML2 cells (**Figure 2A**) to further explore its role in AML. Analysis of MTT (**Figure 2B**) and Transwell (**Figure 2C**) assays revealed that overexpression of miR-221-3p could significantly improve the viability, migration and invasion abilities of OCI-AML2 cells. FCM assay were performed on NC-mimic and oe-miR-221-3p OCI-AML2 cells. The results indicated that overexpression of miR-221-3p reduced the number of cells in the G0/G1 phase with the number of cells in the divisions increased in OCI-AML2 cells (**Figure 2D**). The expressions of PARP, caspase 8, cleave caspase 8, caspase 9, and other apoptosis-related proteins detected by western blot were decreased after overexpression of miR-221-3p (**Figure 2E**), indicating that overexpression of miR-221-3p weakened the apoptosis of OCI-AML2 cells, and the results were consistent with FCM.

BMMSC-Derived MVs Regulate the Function of AML Cells

At present, studies have found that miRNAs can be produced by other cells and transported to target cells through MVs for

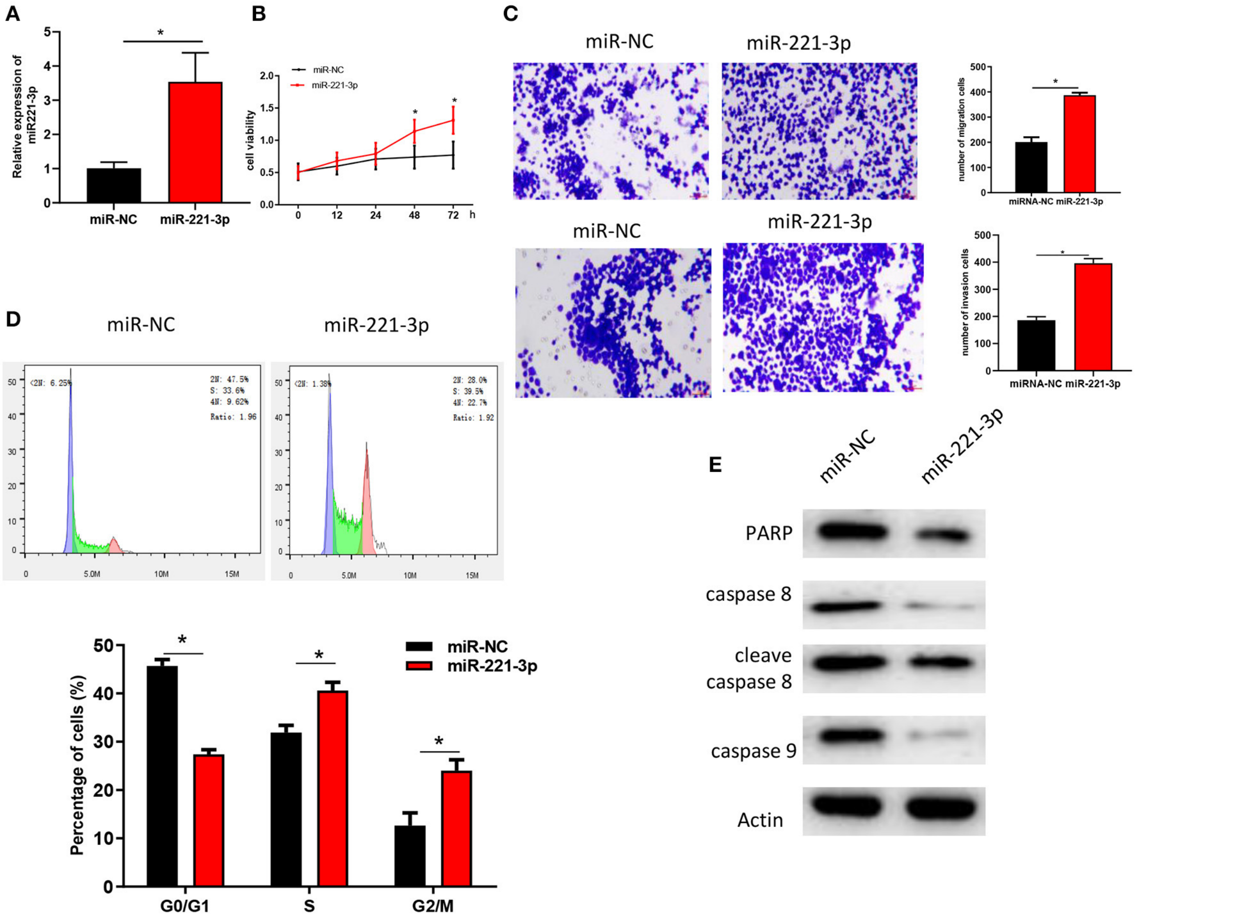


FIGURE 2 | miR-221-3p regulates AML cell cycle, proliferation, and invasion. **(A)** miR-221-3p expression in each group. **(B)** The effect of miR-221-3p overexpression on the activity of OCI-AML2 cells was tested by MTT. **(C)** The effects of miR-221-3p overexpression on invasion and migration of OCI-AML2 cells were detected by Transwell assay (100×). **(D)** The effects of overexpression of miR-221-3p overexpression on the cell cycle of OCI-AML2 was detected by FCM. **(E)** The effect of miR-221-3p overexpression on expressions of apoptosis-related proteins in OCI-AML2 cells.

further function (Momen-Heravi et al., 2015; Hornick et al., 2016; Lu, 2017), so we speculated that miR-221-3p may be carried by MVs to act on AML cells and thus exerting its regulatory role. We searched the expression location of miR-221-3p in the EVmiRNA database and found that its content in MVs of fibroblast and MSCs was significantly higher than that in MVs of other cells (**Figure 3A**). Meanwhile, studies have reported that BMMSC can affect the morphology, adhesion and microenvironment of leukemia stem cells (Roversi et al., 2019). Then we hypothesized that miR-221-3p was contained in the BMMSC-derived MVs and entered the blood to affect the morphological function of AML cells. MVs of BMMSC from normal subjects and AML patients were extracted, and the morphology of MVs was observed under electron microscopy to verify the hypothesis. The MVs showed double-concave disk-like particles with a diameter of about 100 nm−1 μm (**Figure 3B**). The contents of MV marker proteins CD63, TSG101, HSP70, CD9, and CD81 were detected by western blot (**Figure 3C**) to verify the successful extraction of MVs. Further detection revealed that the expression of miR-221-3p in the BMMSC-derived MVs of AML patients was significantly higher than that of normal subjects (**Figure 3D**). To identify the delivery of MVs, we labeled BMMSC-derived MVs and OCI-AML2 cells with Dil (red) or Dio (green), respectively. After co-culture, it was observed that Dil spots presented in the OCI-AML2 cells under laser scanning confocal microscope, indicating that the MVs released by the BMMSC were delivered to the OCI-AML2 cells (**Figure 3E**). Finally, the results of Transwell (**Figure 3F**) and MTT (**Figure 3G**) assays showed that the BMMSC-derived MVs significantly improved the migration and invasion abilities as well as cellular activity of OCI-AML2 cells. The results of FCM on co-cultured cells showed that the number of cells was reduced in G0/G1 phase and increased in division stage of the OCI-AML2 cell cycle induced by BMMSC-derived MCs (**Figure 3H**). The expressions of apoptosis-related proteins including PARP, caspase 8, cleave caspase 8, and caspase 9 determined by western blot were decreased after the co-culture of BMMSC-derived MVs with OCI-AML2 cells (**Figure 3I**), indicating that the MVs could

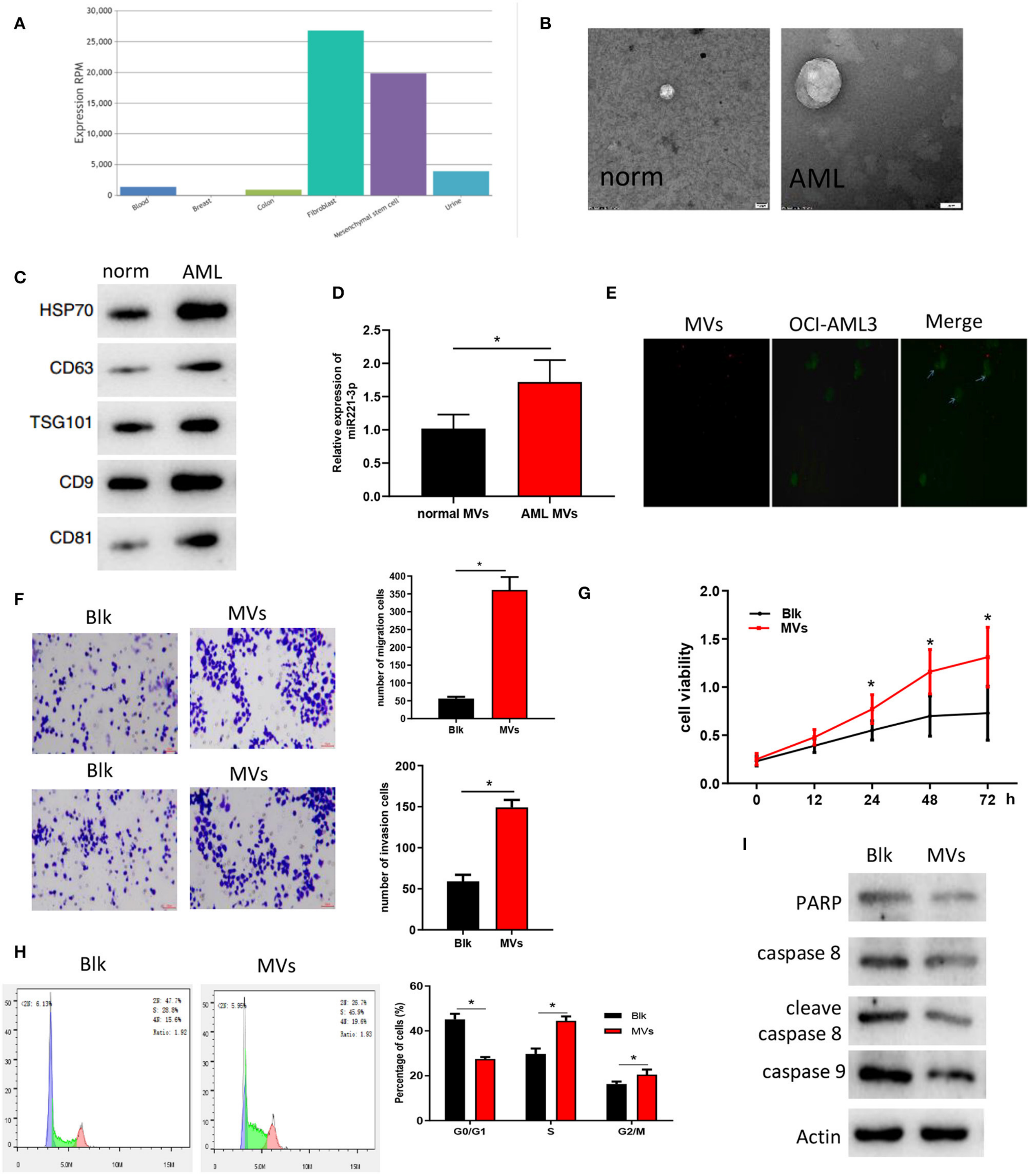


FIGURE 3 | BMMSC-derived MVs regulate AML cell function. **(A)** The expression of miR-221-3p in the MVs of each cell, the abscissa represents the cell name and the ordinate represents the miRNA expression value. **(B)** The morphology and size of BMMSC-derived MVs from normal subjects and AML patients were observed under electron microscopy. **(C)** The changes of MV marker proteins from different sources were detected by western blot. **(D)** miR-221-3p expression in BMMSC-derived MVs in normal subjects and AML patients. **(E)** Fluorescence microscopy showed that Dio-labeled OCI-AML2 cells (green) were transferred to Dil-labeled MVs (red). **(F)** Transwell assay was conducted to detect the effects of MVs on the invasion and migration ability of OCI-AML2 cells (100×). **(G)** The effect of MVs on the activity of OCI-AML2 cells was determined by MTT assay. **(H)** The effects of MVs from different sources on the cell cycle of OCI-AML2 was detected by FCM. **(I)** The expression changes of apoptosis-related proteins in OCI-AML2 cells affected by MVs from different sources were determined by western blot.

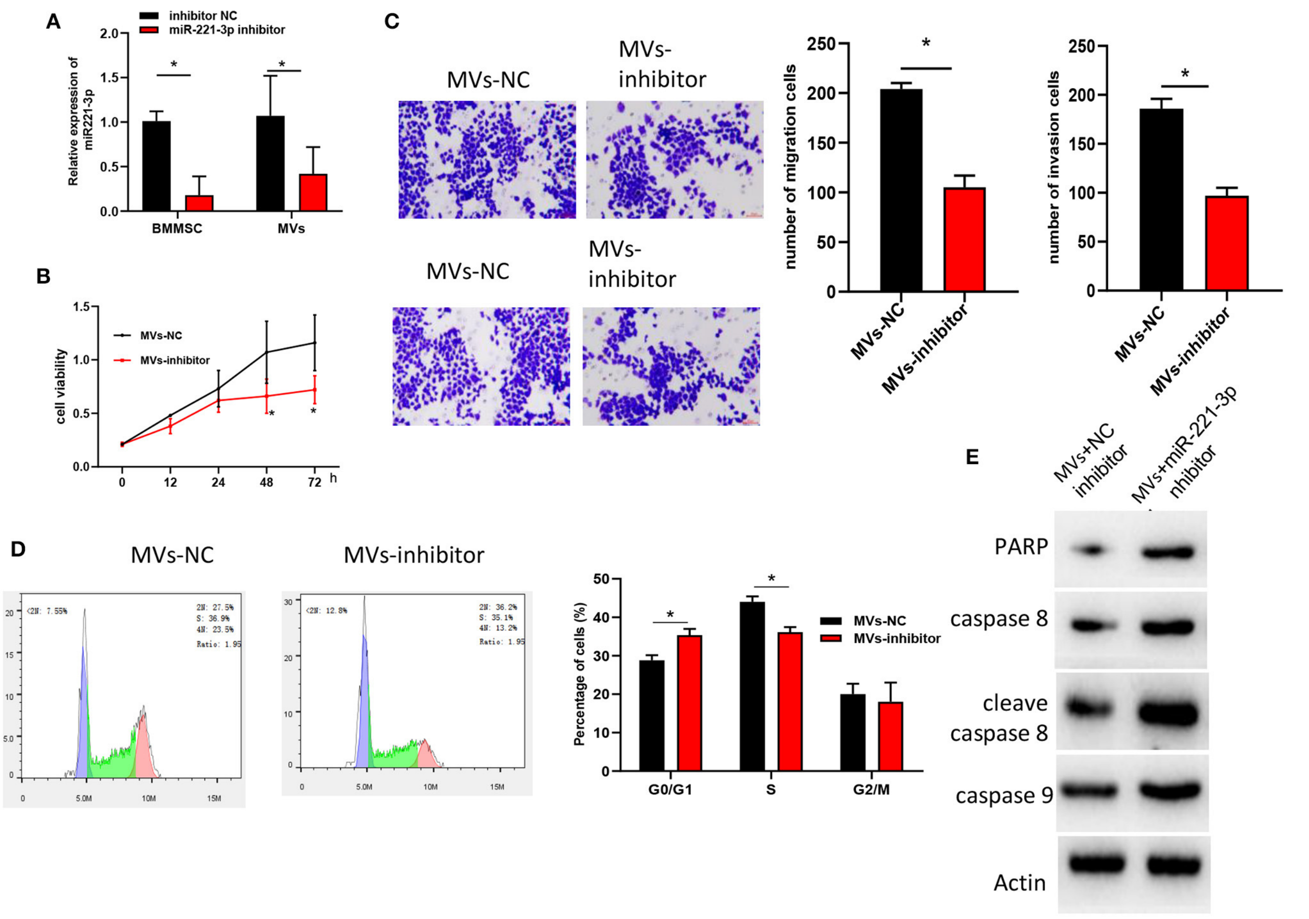


FIGURE 4 | BMMSC-derived MVs regulates cell biological behaviors in AML via miR-221-3p. **(A)** miR-221-3p expression in BMMSC and MVs. **(B)** OCI-AML2 cell viability detected by MTT. **(C)** Cell migration and invasion assayed by Transwell (100×). **(D)** Cell cycle and cell apoptosis test by FCM. **(E)** Levels of apoptosis-related proteins determined by western blot.

reduce the apoptosis of OCI-AML2 cells, which was in keeping with the results of FCM. These results demonstrated that the BMMSC-derived MVs could enter OCI-AML2 cells, promote the proliferation, migration and invasion, weaken the apoptosis and regulate the cell cycle of OCI-AML2 cells.

BMMSC-Derived MVs Regulates Cell Biological Behaviors in AML via miR-221-3p

To further investigate the regulatory mechanism of BMMSC-derived MVs on OCI-AML2 cell proliferation, invasion and cell cycle via miR-221-3p, inhibitor NC and miR-221-3p inhibitor were transfected into BMMSC, respectively. MVs in two groups were extracted and we found that miR-221-3p was significantly decreased in MVs with miR-221-3p inhibitor relative to that in MVs with NC inhibitor (**Figure 4A**). Then, the MVs were co-cultured with OCI-AML2 cells, showing that MVs in miR-221-3p inhibitor group suppressed the promotive effect of BMMSC-derived MVs on cell proliferation, migration and invasion of OCI-AML cells (**Figures 4B,C**). Meanwhile, FCM revealed that miR-221-3p inhibitor induced BMMSC cell cycle arrested in G0/G1 phase, indicating that miR-221-3p inhibitor could reverse the effect of BMMSC-derived MVs on cell cycle (**Figure 4D**). Besides, high expressions of apoptosis-related proteins PARP, caspase 8, cleave caspase 8, and caspase 9 were detected by western blot (**Figure 4E**), and the results suggested that miR-221-3p was capable of abrogating the inhibitory effect of BMMSC-derived MVs on OCI-AML2 cells, which was consistent with the FCM results. In all, these findings shed light on that BMMCS-derived MVs regulated OCI-AML2 cell biological behaviors via miR-221-3p.

miR-221-3p Regulates Cell Proliferation, Invasion and Cell Cycle in AML via Targeting CDKN1C

The downstream targets of miR-221-3p were predicted by TargetScan, miRSearch and mirDIP databases. Differential analysis was performed on the mRNAs procured from TCGA-AML dataset, and eventually 12 potential targets were obtained after the intersection between the identified down-regulated

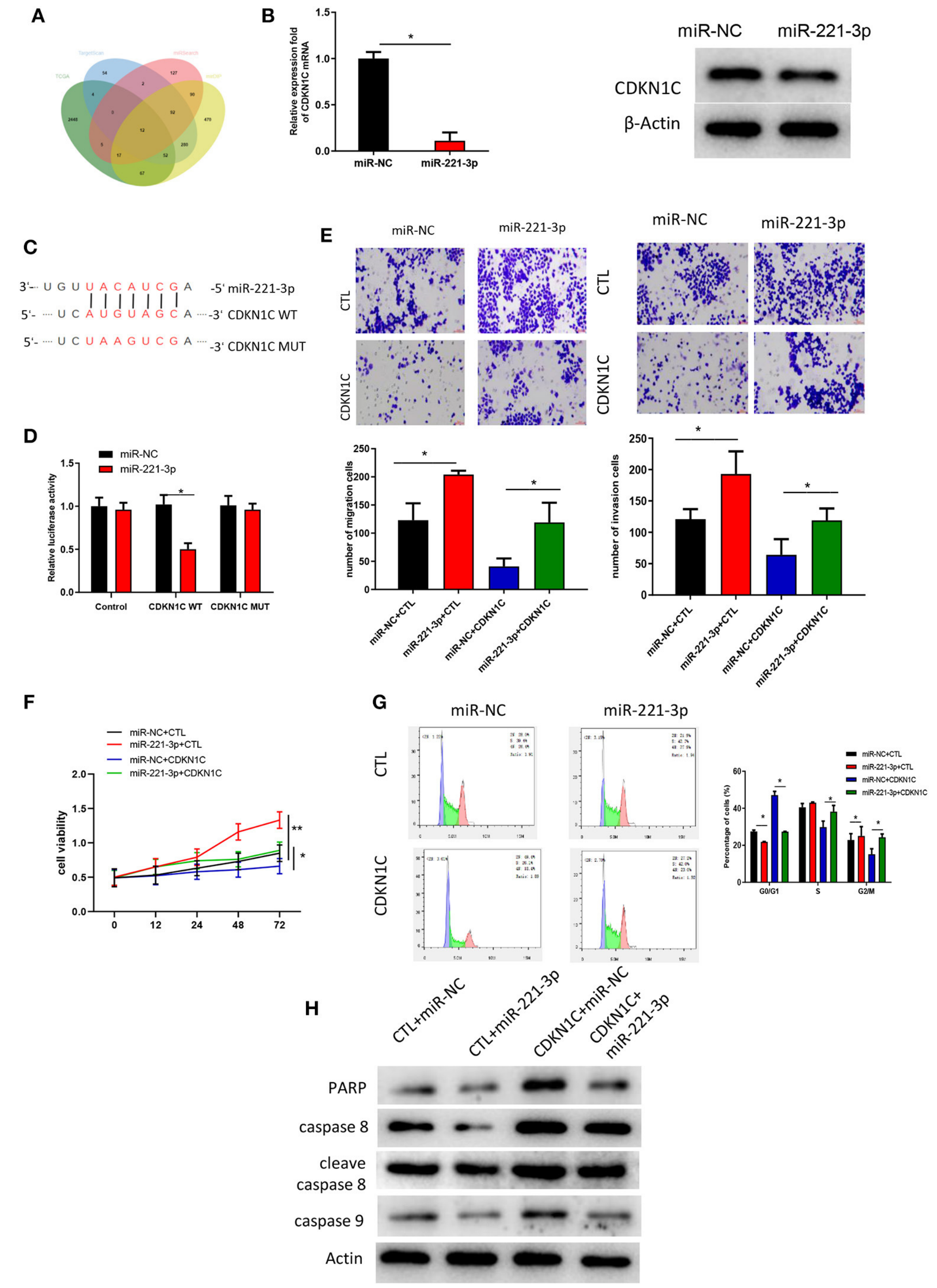


FIGURE 5 | miR-221-3p regulates cell proliferation, invasion and cell cycle in AML via targeting CDKN1C. **(A)** Venn daigram was plotted to find the potential target genes of miR-221-3p. **(B)** CDKN1C expression in mRNA and protein levels in the presence of miR-221-3p overexpression. **(C)** The binding sites of miR-221-3p on CDKN1C 3′-UTR-WT and CDKN1C 3′-UTR-MUT. **(D)** Relative luciferase activity in each group. **(E)** Cell migration and invasion detected by Transwell (100×). **(F)** Cell viability test by MTT. **(G)** Cell cycle determined by FCM. **(H)** Protein levels of apoptosis-related proteins measured by western blot.

DEmRNAs and the predicted targets (**Figure 5A**). Among the 12 target genes, CDKN1C alteration in AML was shown to be the most significant (**Table 1**). Thereafter, to further validate the relationship between miR-221-3p and CDKN1C, miR-221-3p mimic and NC-mimic were, respectively, transfected into OCI-AML2 cells. Western blot and qRT-PCR suggested

TABLE 1 | Differential expression of the identified 12 potential target genes in the TCGA-AML dataset.

Gene symbol	Gene ID	Median (Tumor)	Median (Normal)	Log2 (Fold Change)	Adjp
CDKN1C	ENSG00000129757.12	13.44	118.61	−3.05	4.88E-22
NRK	ENSG00000123572.16	0.93	9.153	−2.395	1.77E-10
MYLIP	ENSG00000007944.14	2.83	13.58	−1.929	8.84E-26
POGZ	ENSG00000143442.21	11.19	31.24	−1.403	1.20E-27
FOS	ENSG00000170345.9	28.099	75.439	−1.393	3.10E-09
ARHGEF7	ENSG00000102606.17	9.05	23.855	−1.306	4.83E-36
CREBZF	ENSG00000137504.13	13.06	32.565	−1.255	2.01E-27
FAM214A	ENSG00000047346.12	12.69	30.505	−1.202	6.79E-18
ADAM22	ENSG00000008277.14	1.37	4.23	−1.142	2.88E-14
CD4	ENSG00000010610.9	3.35	8.185	−1.078	2.74E-07
HMBOX1	ENSG00000147421.17	14.57	31.545	−1.064	1.92E-21
RFX7	ENSG00000181827.14	1.66	4.535	−1.057	3.91E-26

that the mRNA and protein expressions CDKN1C were reduced in the cells transfected with miR-221-3p mimic (**Figure 5B**). Then, online miRNA data analysis software (starBase) was applied, finding that miR-221-3p was targeted binding with the CDKN1C 3′-UTR (**Figure 5C**). Meanwhile, dual-luciferase assay demonstrated that miR-221-3p inhibitory functioned on the luciferase activity in cells transfected with CDKN1C-WT, whereas there was no difference observed in cells transfected with CDKN1C-MUT (**Figure 5D**). Taken together, we could conclude that CDKN1C was a direct target of miR-221-3p. Subsequently, a series of *in vitro* experiments were conducted to explore the miR-221-3p-dependent mechanism on cell biological behaviors via CDKN1C. Transwell and MTT assays showed that overexpressing of CDKN1C could reverse the promotive role of miR-221-3p overexpression in cell migration, invasion, proliferation and colony forming of OCI-AML cells (**Figures 5E,F**). Besides, the effects of miR-221-3p overexpression on cell cycle could also be reversed when CDKN1C was simultaneously increased (**Figure 5G**). Moreover, apoptosis-related proteins were all observed to be elevated after CDKN1C being overexpressed (**Figure 5H**), elucidating that CDKN1C overexpression was capable of rescuing the decrease of cell apoptosis induced by miR-221-3p overexpression.

DISCUSSION

MVs were primarily regarded as unfunctional cellular components to be discarded, yet it has been increasingly suggested that MVs are important tools for the exchange of cellular information and materials, and closely correlated with tumor distant metastasis and immune inhibition (Steinbichler et al., 2017; Fan et al., 2018; Seo et al., 2018; Jerez et al., 2019). MVs are capable of inducing various biological processes after being transferred into recipient cells, such as angiogenesis, metastasis formation, therapeutic resistance, epithelial-mesenchymal transition (EMT) and epigenetic programming (Kreimer et al., 2015; Milane et al., 2015; Gopal et al., 2017). In the present study, we found that miR-221-3p was highly expressed in BMMSC-derived MVs. Besides, it has been reported that bone marrow stromal cell-derived MVs can attenuate the B cell apoptosis in chronic lymphocytic leukemia, also promote cell migration and induce gene expression and modification (Crompot et al., 2017). Hence, this study focused attention on the BMMSC MVs-derived miR-221-3p. Enormous studies have revealed that miR-221-3p is aberrantly expressed in various cancers and participate in the regulation of tumorigenesis and development, like cervical squamous carcinoma (Wu et al., 2019), hepatocellular carcinoma (Li et al., 2019), medulloblastoma (Yang et al., 2019), and breast cancer (Ergun et al., 2015). However, the role of miR-221-3p in AML has not been reported. Therefore, the purpose of this study is to explore the mechanism of miR-221-3p in AML. In our study, we discovered that miR-221-3p was mainly present in BMMSC-derived MVs, and found to be overexpressed in AML patients. Then we constructed miR-221-3p overexpression and found that elevated miR-221-3p was responsible for the promotion of OCI-AML2 cell proliferation, migration and invasion. Moreover, miR-221-3p has been reported to play an important role in other cancers. Wu et al. have found that miR-221-3p from tumor cell-derived MVs targets THBS2 to facilitate the angiogenesis in cervical squamous carcinoma (Wu et al., 2019). Wei et al. have reported that miR-221-3p can potentiate metastasis in cervical cancer via directly targeting THBS2 (Wei et al., 2017). In addition, Shi et al. have revealed that miR-221-3p serving as an oncogene promotively functions on cell proliferation, migration and invasion in gastric cancer through inhibiting PTEN (Shi et al., 2017). Collectively, we believed that miR-221-3p from BMMSC-derived MVs could act as an oncogene beneficial for the cell proliferation, migration and invasion in AML.

In order to further understand the molecular mechanism of miR-221-3p regulating the function of AML cells in BMMSC-derived MVs, we proved that miR-221-3p can directly target CDKN1C through bioinformatics analysis and dual-luciferase assay. Besides, there was a negative correlation showed in miR-221-3p and CDKN1C expressions both in tissues and cells. CDKN1C is a cyclin-dependent kinase inhibitor 1C, which can inhibit cell proliferation (Adkins and Lumb, 2002; Qiu et al., 2018, 2019). Abnormal expression of CDKN1C plays a role in breast cancer (Qiu et al., 2018), gastric cancer (Sun et al., 2017), glioma (Zhang et al., 2015) and other cancers. And some studies have found that CDKN1C is often methylated in acute lymphoblastic leukemia, and methylation is associated with poor prognosis (Shen et al., 2003). It is found that the expression of CDKN1C is related to the prognosis of patients with AML (Radujkovic et al., 2016), but the biological function of CDKN1C in AML is unclear. This study found that overexpressing CDKN1C could suppress cell proliferation, migration and invasion in AML. Moreover, CDKN1C was able to reverse the regulation of miR-221-3p overexpression on AML cell biological behaviors when it was concurrently elevated. Taken together, these results suggest that miR-221-3p in BMMSC-derived MVs in AML patients regulates the proliferation, invasion, migration and cell cycle by targeting CDKN1C.

In conclusion, our study confirmed that miR-221-3p from BMMSC-derived MVs had the functions of promoting cell proliferation, migration, invasion and regulating cell cycle in AML via targeting CDKN1C. This finding extends our knowledge on the role of miR-221-3p in AML, and helps to further explore the novel approaches for AML targeted therapy.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are included within the article. The data and materials in the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

XuZ and WY contributed to the study design. YX, JW, SZ, and JL conducted the literature search. XH and HX acquired the data. XiZ and SS wrote the article. YL performed data analysis and drafted. YZ and YX revised the article.

FUNDING

This study was supported by the funds from National Natural Science Foundation of China (No. 81670124).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00081/full#supplementary-material

REFERENCES

Abbasian, N., Herbert, K. E., Pawluczyk, I., Burton, J. O., and Bevington, A. (2018). Vesicles bearing gifts: the functional importance of micro-RNA transfer in extracellular vesicles in chronic kidney disease. *Am. J. Physiol. Renal Physiol.* 315, F1430–F1443. doi: 10.1152/ajprenal.00318.2018

Adkins, J. N., and Lumb, K. J. (2002). Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57Kip2. *Proteins* 46, 1–7. doi: 10.1002/prot.10018

Boyiadzis, M., and Whiteside, T. L. (2018). Exosomes in acute myeloid leukemia inhibit hematopoiesis. *Curr. Opin. Hematol.* 25, 279–284. doi: 10.1097/MOH.0000000000000439

Braicu, C., Tomuleasa, C., Monroig, P., Cucuianu, A., Berindan-Neagoe, I., and Calin, G. A. (2015). Exosomes as divine messengers: are they the hermes of modern molecular oncology? *Cell Death Differ.* 22, 34–45. doi: 10.1038/cdd.2014.130

Coombs, C. C., Tallman, M. S., and Levine, R. L. (2016). Molecular therapy for acute myeloid leukaemia. *Nat. Rev. Clin. Oncol.* 13, 305–318. doi: 10.1038/nrclinonc.2015.210

Cornelissen, J. J., and Blaise, D. (2016). Hematopoietic stem cell transplantation for patients with AML in first complete remission. *Blood* 127, 62–70. doi: 10.1182/blood-2015-07-604546

Crompot, E., Van Damme, M., Pieters, K., Vermeersch, M., Perez-Morga, D., Mineur, P., et al. (2017). Extracellular vesicles of bone marrow stromal cells rescue chronic lymphocytic leukemia B cells from apoptosis, enhance their migration and induce gene expression modifications. *Haematologica* 102, 1594–1604. doi: 10.3324/haematol.2016.163337

Del Principe, M. I., Del Principe, D., and Venditti, A. (2017). Thrombosis in adult patients with acute leukemia. *Curr. Opin. Oncol.* 29, 448–454. doi: 10.1097/CCO.0000000000000402

Ergun, S., Tayeb, T. S., Arslan, A., Temiz, E., Arman, K., Safdar, M., et al. (2015). The investigation of miR-221-3p and PAK1 gene expressions in breast cancer cell lines. *Gene* 555, 377–381. doi: 10.1016/j.gene.2014.11.036

Fan, Q., Yang, L., Zhang, X., Peng, X., Wei, S., Su, D., et al. (2018). The emerging role of exosome-derived non-coding RNAs in cancer biology. *Cancer Lett.* 414, 107–115. doi: 10.1016/j.canlet.2017.10.040

Gopal, S. K., Greening, D. W., Rai, A., Chen, M., Xu, R., Shafiq, A., et al. (2017). Extracellular vesicles: their role in cancer biology and epithelial-mesenchymal transition. *Biochem. J.* 474, 21–45. doi: 10.1042/BCJ20160006

Hansen, H. P., Engels, H. M., Dams, M., Paes Leme, A. F., Pauletti, B. A., Simhadri, V. L., et al. (2014). Protrusion-guided extracellular vesicles mediate CD30 trans-signalling in the microenvironment of Hodgkin's lymphoma. *J. Pathol.* 232, 405–414. doi: 10.1002/path.4306

Hornick, N. I., Doron, B., Abdelhamed, S., Huan, J., Harrington, C. A., Shen, R., et al. (2016). AML suppresses hematopoiesis by releasing exosomes that contain microRNAs targeting c-MYB. *Sci. Signal* 9:ra88. doi: 10.1126/scisignal.aaf2797

Jerez, S., Araya, H., Hevia, D., Irarrázaval, C. E., Thaler, R., van Wijnen, A. J., et al. (2019). Extracellular vesicles from osteosarcoma cell lines contain miRNAs associated with cell adhesion and apoptosis. *Gene* 710, 246–257. doi: 10.1016/j.gene.2019.06.005

Khwaja, A., Bjorkholm, M., Gale, R. E., Levine, R. L., Jordan, C. T., Ehninger, G., et al. (2016). Acute myeloid leukaemia. *Nat. Rev. Dis. Primers* 2:16010. doi: 10.1038/nrdp.2016.10

Kreimer, S., Belov, A. M., Ghiran, I., Murthy, S. K., Frank, D. A., and Ivanov, A. R. (2015). Mass-spectrometry-based molecular characterization of extracellular vesicles: lipidomics and proteomics. *J. Proteome Res.* 14, 2367–2384. doi: 10.1021/pr501279t

Li, H., Zhang, B., Ding, M., Lu, S., Zhou, H., Sun, D., et al. (2019). C1QTNF1-AS1 regulates the occurrence and development of hepatocellular carcinoma by regulating miR-221-3p/SOCS3. *Hepatol. Int.* 13, 277–292. doi: 10.1007/s12072-019-09944-5

Lu, X. (2017). The role of exosomes and exosome-derived microRNAs in atherosclerosis. *Curr. Pharm. Des.* 23, 6182–6193. doi: 10.2174/1381612823666170413125507

Milane, L., Singh, A., Mattheolabakis, G., Suresh, M., and Amiji, M. M. (2015). Exosome mediated communication within the tumor microenvironment. *J. Control Release* 219, 278–294. doi: 10.1016/j.jconrel.2015.06.029

Momen-Heravi, F., Bala, S., Kodys, K., Szabo, G., Momen-Heravi, F., Bala, S., et al. (2015). Exosomes derived from alcohol-treated hepatocytes horizontally transfer liver specific miRNA-122 and sensitize monocytes to LPS. *Sci. Rep.* 5:9991. doi: 10.1038/srep09991

Qiu, Z., Li, Y., Zeng, B., Guan, X., and Li, H. (2018). Downregulated CDKN1C/p57(kip2) drives tumorigenesis and associates with poor overall survival in breast cancer. *Biochem. Biophys. Res. Commun.* 497, 187–193. doi: 10.1016/j.bbrc.2018.02.052

Qiu, Z., Zhu, W., Meng, H., Tong, L., Li, X., Luo, P., et al. (2019). CDYL promotes the chemoresistance of small cell lung cancer by regulating H3K27 trimethylation at the CDKN1C promoter. *Theranostics* 9, 4717–4729. doi: 10.7150/thno.33680

Radujkovic, A., Dietrich, S., Andrulis, M., Benner, A., Longerich, T., Pellagatti, A., et al. (2016). Expression of CDKN1C in the bone marrow of patients with myelodysplastic syndrome and secondary acute myeloid leukemia is associated with poor survival after conventional chemotherapy. *Int. J. Cancer* 139, 1402–1413. doi: 10.1002/ijc.30181

Roversi, F. M., Cury, N. M., Lopes, M. R., Ferro, K. P., Machado-Neto, J. A., Alvarez, M. C., et al. (2019). Up-regulation of SPINT2/HAI-2 by azacytidine in bone marrow mesenchymal stromal cells affects leukemic stem cell survival and adhesion. *J. Cell. Mol. Med.* 23, 1562–1571. doi: 10.1111/jcmm.14066

Seo, N., Akiyoshi, K., and Shiku, H. (2018). Exosome-mediated regulation of tumor immunology. *Cancer Sci.* 109, 2998–3004. doi: 10.1111/cas.13735

Shen, L., Toyota, M., Kondo, Y., Obata, T., Daniel, S., Pierce, S., et al. (2003). Aberrant DNA methylation of p57KIP2 identifies a cell-cycle regulatory pathway with prognostic impact in adult acute lymphocytic leukemia. *Blood* 101, 4131–4136. doi: 10.1182/blood-2002-08-2466

Shi, J., Zhang, Y., Jin, N., Li, Y., Wu, S., and Xu, L. (2017). MicroRNA-221-3p plays an oncogenic role in gastric carcinoma by inhibiting PTEN expression. *Oncol. Res*. 25, 523–536. doi: 10.3727/096504016X14756282819385

Stec, M., Baj-Krzyworzeka, M., Baran, J., Weglarczyk, K., Zembala, M., Barbasz, J., et al. (2015b). Isolation and characterization of circulating micro(nano)vesicles in the plasma of colorectal cancer patients and their interactions with tumor cells. *Oncol. Rep*. 34, 2768–2775. doi: 10.3892/or.2015.4228

Stec, M., Szatanek, R., Baj-Krzyworzeka, M., Baran, J., Zembala, M., Barbasz, J., et al. (2015a). Interactions of tumour-derived micro(nano)vesicles with human gastric cancer cells. *J. Transl. Med.* 13:376. doi: 10.1186/s12967-015-0737-0

Stein, E. M., and Tallman, M. S. (2016). Emerging therapeutic drugs for AML. *Blood* 127, 71–78. doi: 10.1182/blood-2015-07-604538

Steinbichler, T. B., Dudás, J., Riechelmann, H., and Skvortsova, I. I. (2017). The role of exosomes in cancer metastasis. *Semin. Cancer Biol.* 44, 170–181. doi: 10.1016/j.semcancer.2017.02.006

Summerer, I., Niyazi, M., Unger, K., Pitea, A., Zangen, V., Hess, J., et al. (2013). Changes in circulating microRNAs after radiochemotherapy in head and neck cancer patients. *Radiat. Oncol*. 8:296. doi: 10.1186/1748-717X-8-296

Sun, C., Ma, P., Wang, Y., Liu, W., Chen, Q., Pan, Y., et al. (2017). KLF15 inhibits cell proliferation in gastric cancer cells via up-regulating CDKN1A/p21 and CDKN1C/p57 expression. *Dig. Dis. Sci*. 62, 1518–1526. doi: 10.1007/s10620-017-4558-2

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

Wu, X. G., Zhou, C. F., Zhang, Y. M., Yan, R. M., Wei, W. F., Chen, X. J., et al. (2019). Cancer-derived exosomal miR-221-3p promotes angiogenesis by targeting THBS2 in cervical squamous cell carcinoma. *Angiogenesis* 22, 397–410. doi: 10.1007/s10456-019-09665-1

Yang, Y., Cui, H., and Wang, X. (2019). Downregulation of EIF5A2 by miR-221-3p inhibits cell proliferation, promotes cell cycle arrest and apoptosis in medulloblastoma cells. *Biosci. Biotechnol. Biochem.* 83, 400–408. doi: 10.1080/09168451.2018.1553604

Zhang, J., Gong, X., Tian, K., Chen, D., Sun, J., Wang, G., et al. (2015). miR-25 promotes glioma cell proliferation by targeting CDKN1C. *Biomed. Pharmacother.* 71, 7–14. doi: 10.1016/j.biopha.2015.02.005

Zheng, H., Liu, J. Y., Song, F. J., and Chen, K. X. (2013). Advances in circulating microRNAs as diagnostic and prognostic markers for ovarian cancer. *Cancer Biol. Med.* 10, 123–130. doi: 10.7497/j.issn.2095-3941.2013.03.001

Zhou, C. F., Ma, J., Huang, L., Yi, H. Y., Zhang, Y. M., Wu, X. G., et al. (2019). Cervical squamous cell carcinoma-secreted exosomal miR-221-3p promotes lymphangiogenesis and lymphatic metastasis by targeting VASH1. *Oncogene* 38, 1256–1268. doi: 10.1038/s41388-018-0511-x

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 February 2020
doi: 10.3389/fgene.2019.01325

An Information Entropy-Based Approach for Computationally Identifying Histone Lysine Butyrylation

Guohua Huang[1]*, Yang Zheng[1], Yao-Qun Wu[1,2], Guo-Sheng Han[2] and Zu-Guo Yu[2,3]

[1] Provincial Key Laboratory of Informational Service for Rural Area of Southwestern Hunan, Shaoyang University, Shaoyang, China, [2] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan, China, [3] School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia

Butyrylation plays a crucial role in the cellular processes. Due to limit of techniques, it is a challenging task to identify histone butyrylation sites on a large scale. To fill the gap, we propose an approach based on information entropy and machine learning for computationally identifying histone butyrylation sites. The proposed method achieves 0.92 of area under the receiver operating characteristic (ROC) curve over the training set by 3-fold cross validation and 0.80 over the testing set by independent test. Feature analysis implies that amino acid residues in the down/upstream of butyrylation sites would exhibit specific sequence motif to a certain extent. Functional analysis suggests that histone butyrylation was most possibly associated with four pathways (systemic lupus erythematosus, alcoholism, viral carcinogenesis and transcriptional misregulation in cancer), was involved in binding with other molecules, processes of biosynthesis, assembly, arrangement or disassembly and was located in such complex as consists of DNA, RNA, protein, *etc*. The proposed method is useful to predict histone butyrylation sites. Analysis of feature and function improves understanding of histone butyrylation and increases knowledge of functions of butyrylated histones.

Keywords: butyrylation, random forest, histone, post-translational modification, information entropy

OPEN ACCESS

Edited by:
Peilin Jia,
University of Texas Health Science Center, United States

Reviewed by:
Xiong Li,
East China Jiaotong University, China
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

****Correspondence:***
Guohua Huang
guohuahhn@163.com

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

Received: *12 September 2019*
Accepted: *05 December 2019*
Published: *14 February 2020*

Citation:
Huang G, Zheng Y, Wu Y-Q, Han G-S and Yu Z-G (2020) An Information Entropy-Based Approach for Computationally Identifying Histone Lysine Butyrylation. Front. Genet. 10:1325. doi: 10.3389/fgene.2019.01325

INTRODUCTION

Butyrylation, a type of post-translation modification (PTM), refers to a biochemical interaction process where the butyryl functional group covalently modifies the lysine amino acid (Chen et al., 2007; Lu et al., 2018). Protein butyrylation is a newly discovered PTM (Chen et al., 2007). In the past 5 years, butyrylation's roles in the cellular process have been gradually uncovered. For example, Goudarzi et al. (2016) confirmed that histone butyrylation directly stimulates gene expression and inhibits Brdt Binding, Xu et al. (2018) found that butyrylation and acetylation are responsible for the phenotype and metabolic shifts of the endospore-forming *Clostridium acetobutylicum*, and Lu et al. (2018) revealed that butyrylation prefers poising gene activation by external stresses in the rise of submergence and starvation. Nevertheless, compared to such extensively-studied PTMs as acetylation (Kiemer et al., 2005; Basu et al., 2009; Gnad et al., 2010; Choudhary et al., 2014) and methylation (Chen et al., 2006; Shi et al., 2012b; Hamamoto et al., 2015; Shi et al., 2015; Wei et al., 2018), few functions of butyrylation are known. With in-depth exploration of butyrylation, more biological functions of butyrylation will undoubtedly be found.

Identifying butyrylation sites is an important foundation to further explore its functions. Biotechnologies whose representative is mass spectrometry provide a necessary approach to identify PTMs including butyrylation. Zhang et al. (2008) found four lysine butyrylation sites in histone yeast, Xu et al. (2014) 11 histone butyrylation sites in human cells, and Lu et al. (2018) identified four histone butyrylation sites in rice using mass spectrometry. Obviously, this strategy is not only labor-intensive and time-consuming, but also generally low-throughput. On the contrary, bioinformatics approaches provide an alternative to explore PTM sites, with characteristic being high-throughput. Since Hansen et al. (1995; 1998) proposed a method for computationally predicting mucin type O-glycosylation sites in the 1990s, dozens of computational approaches have been developed for identifying PTM sites (Blom et al., 2004; Xue et al., 2006; Zhou et al., 2006; Xu et al., 2008; Xu et al., 2010; Liu et al., 2011; Cai et al., 2012; Shi et al., 2012b; Zhang et al., 2012; Zhao et al., 2012; Xu et al., 2013; Zhang et al., 2013; Zhao et al., 2013; Huang et al., 2014; Shi et al., 2015; Xu et al., 2015a; Zhou et al., 2016). For instances, glycosylation identification includes the neural network-based method (Hansen et al., 1998), the support vector machine-based method (Li et al., 2006; Chen et al., 2008; Sasaki et al., 2009), the random forest-based method (Hamby and Hirst, 2008; Chuang et al., 2012), and ensemble learning algorithms (Caragea et al., 2007). Features used for predicting methylation sites are from protein sequences (Shao et al., 2009; Zhang et al., 2013; Qiu et al., 2014; Zhang et al., 2015; Wei et al., 2018), structure (Shien et al., 2009) or amino acid properties (Shi et al., 2012a). Xu et al. (2015b) proposed a pseudo amino acid composition-based method for predicting lysine succinylation. Zhou et al. (2004) proposed the GPS method for phosphorylation prediction, and Xu et al. (2008) proposed the method SUMOpre for sumoylation prediction. These computational methods are capable of screening potential modified sites on a large scale in a little time and help the former methods narrow the scope of verification of it. Here, we didn't plan to comprehensively review and discuss them, but propose a novel method based on information entropy and random forest for predicting histone butyryllysine. To the best of my knowledge, this is the first computational method for predicting butyrylation.

METHOD AND MATERIALS

Materials

One hundred butyrylated proteins were retrieved by searching both the Uniprot database (UniProt Consortium, 2018): https://www.uniprot.org/ and the Protein Lysine Modifications Database (PLMD): http://plmd.biocuckoo.org/ (Xu et al., 2017). The Uniprot database is a comprehensive repository of function annotation and sequences of proteins, which is updated every 2 months. The PLMD is dedicated to specifically collect lysine-modified proteins, and the current version 3.0 contains 284,780 modification events of 20 types of lysine-modified PTMs from 53501 proteins, including butyrylation, crotonylation and propionylation. Searching the Uniprot database with the keyword “butyryllysine”, we retrieved 91 butyrylated histones containing 317 butyrylation sites with the manual assertion. We downloaded the butyrylation data from the PLMD. Merging these two datasets and then removing abnormal proteins, we got 100 unique histones. To eliminate dependency of the computational method on homology, it is a general step to remove homology among protein sequences. The computational clustering tool (Huang et al., 2010) was used to cluster these 100 protein sequences with the sequence identity cut-off 0.7. Thirteen representative protein sequences were obtained among which sequence identity of any two is no more than 0.7. We selected six proteins from the Uniprot database as the training set which contained 17 butyrylation sites and the remaining seven from the PLMD as the testing set which contained nine butyrylation sites.

Method

As shown in **Figure 1**, the overall workflow of the proposed method consists mainly of four steps: cutting sequence, sequence encoding, training and predicting. The training and the predictive butyrylation histone sequences were cut into fragments which centered lysine with respectively N amino acid residues in the upstream and the downstream of it. That is, the window of (2N+1) residues centering lysine were separated out. For the windows containing lysine but less than 2N+1 residues, we prefixed or suffixed the character “X” to it for complement. The fragments undergoing butyrylation event were viewed as positive samples. We randomly selected 18 non-butyrylation fragments from the training set as training negative samples, and 18 non-butyrylation ones from the testing set as the testing negative samples. The **Supplementary Table 1** listed all the training and the testing butyrylation as well as the non-butyrylation sites. For each fragment with (2N+1) resides, the information entropy-based encoding (IEE) and the composition of k-space amino acid pair (CKSAAP) transformed it into numerical feature. After the random forest algorithm trained a classifier using the training set with the numerical features, the unknown protein sequences were input into the trained classifier for final prediction.

IEE

Histone butyrylation is assumed as a stochastic system described as $P^i(\alpha)$ which stands for probability of the amino acid α occurring at the i-th position. Obviously, $P^i(\alpha)$ is an *m*-by-*n* matrix where *m* is the number of characters of amino acid (here *m* is 21) and *n* the length of the sequence (here *n*=2*N+1). This stochastic system is measured by the information entropy of

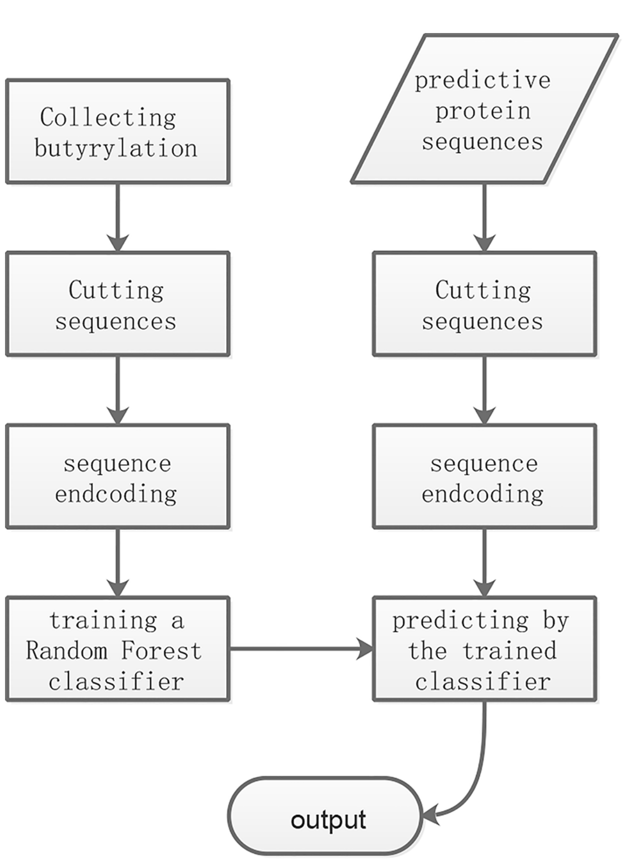

Collecting butyrylation
Cutting sequences
sequence endcoding
training a Random Forest classifier
predictive protein sequences
Cutting sequences
sequence endcoding
predicting by the trained classifier
output

FIGURE 1 | The flowchart of the proposed method.

amino acid (PIEA) and the information entropy of position (PIEP), which are denoted respectively by

$$\mathrm{PIEA}(\alpha) = \sum_{i=1}^{2N+1} -P^i(\alpha)\log P^i(\alpha) \tag{1}$$

and

$$\mathrm{PIEP}(i) = \sum_{\alpha\in\Phi} -P^i(\alpha)\log P^i(\alpha) \tag{2}$$

where Φ represents the set of characters of amino acid. $\mathrm{P}^i(\alpha)$ can be estimated by calculating frequencies of amino acid over all the positive samples in the training set, respectively. The PIEA and the PIEP represent uncertainty of the butyrylation system. The more the PIEA and the PIEP are, the more uncertainty the system is. After a new sample *s* was added to the system, its information entropies of amino acid and position are denoted by PIEPs and PIEAs. The variation of information entropies after addition of the new sample to the system is defined by

$$\mathrm{PVIEA} = \mathrm{PIEA}(\alpha) - \mathrm{PIEAs}(\alpha) \tag{3}$$

and

$$\mathrm{PVIEP} = \mathrm{PIEP(i)} - \mathrm{PIEPs(i)}\,. \tag{4}$$

Similarly, the non-butyrylation system is also assumed as a distinct stochastic system $\mathrm{N}^i(\alpha)$ which is estimated by calculating frequencies of amino acid over all the negative samples in the training set, respectively. The information entropies of amino acid (NIEA) and the information entropies of position (NIEP) for the non-butyrylation system are defined by

$$\mathrm{NIEA}(\alpha) = \sum_{i=1}^{2N+1} -\mathrm{N}^i(\alpha)\log \mathrm{N}^i(\alpha) \tag{5}$$

and

$$\mathrm{NIEP}(i) = \sum_{\alpha\in\Phi} -\mathrm{N}^i(\alpha)\log N^i(\alpha) \tag{6}$$

The variation of information entropies after addition of the new sample *s* to the non-butyrylation system is defined by

$$\mathrm{NVIEA} = \mathrm{NIEA}(\alpha) - \mathrm{NIEAs}(\alpha) \tag{7}$$

and

$$\mathrm{NVIEP} = \mathrm{NIEP(i)} - \mathrm{NIEPs(i)}, \tag{8}$$

where NIEAs and NIEPs denote respectively information entropies of amino acid and position after addition of the new sample to the non-butyrylation system. The new sample is encoded by PVIEA-NVIEA and PVIEP-NVIEP. Therefore, for each sample, we obtain (21 + 2N+1) feature to represent it.

CKSAAP

The CKSAAP is occurrence frequency of *k*-spaced amino acid pair which is spaced by up to *k* residues. *k* is equal to or more than 0. For example, AA, AC, ..., YX and XX belong to 0-spaced amino acid pair, while AA, AC, ..., XX, ABA, ABC, ..., and XBX to 1-spaced amino acid pair. Generally, there are (K+1)*21*21 features for k-spaced amino acid pair. The CKSAAP were widely applied to prediction of phosphorylation, methylation, palmitoylation, pupylation, ubiquitination and O-glycosylation (Chen et al., 2008; Wang et al., 2009; Chen et al., 2011; Zhao et al., 2012; Tung, 2013; Zhang et al., 2013).

Feature Normalization

All the features are normalized by the following formula

$$\mathrm{X}_k^n = \frac{x_k^n - \min\limits_m\{x_k^m\}}{\max\limits_m\{x_k^m\} - \min\limits_m\{x_k^m\}}, \tag{9}$$

where $x_k{}^n$ denotes the k-th non-normalized feature of the sample *n*. The normalized feature lies between 0 and 1.

Random Forest

Random forest by Breiman (2001) is an ensemble learning algorithm which combines decision trees for vote. The random forest is composed mainly of constructing of decision trees and voting over all the decision trees for the given sample. Each decision tree grow out of the new training set drawn with replacement from the training set and with *m* << *M* randomly selected features (*M* is the total number of sample features). The majority of vote for a sample is the output class for classification. The advantage of Random forests is that it overcome overfitting which occurred in decision trees, and meanwhile produce a limiting value of the generalization error. For more details of random forest, readers can refer to relevant references. Here, we use Weka software package (Hall et al., 2009) which realized a wide range of machine learning algorithms using the Java programming language.

CROSS VALIDATION AND METRICS

We used 3-fold cross validation to examine performance of the proposed method. For 3-fold cross validation, *n* training samples are divided into three parts in approximate or equal size. Each part is in turn used as the testing set which is predicted by the trained classifier over the other two parts. Independent test was used to examine generalization ability of the proposed method.

The receiver operating characteristic (ROC) curve was used to assess the predictive performances, which is plotting true positive rate against false positive rate under various threshold. Area under the ROC curve (AUC) was used to compare it, ranging from 0 to 1. The AUC was 1, meaning the perfect prediction, while the AUC was 0.5, indicating the uninformative classifier.

RESULTS AND DISCUSSION

To investigate effects of the parameter *N* (length of amino acid residues in the upstream or the downstream of the butyrylation sites) on the predictive performances, we conducted 3-fold cross validation over the training set. Most approaches for predicting PTM sites generally set *N* to the interval of 10 to 15 (Hou et al., 2014; Huang et al., 2014; Xu et al., 2015a; Hasan et al., 2016; Jia et al., 2016a; Jia et al., 2016b; Xu et al., 2016; Wang et al., 2017). For example, the iSulf-Cys for predicting s-sulfenylation sites (Xu et al., 2016) adopted a window of 21 residues (i.e., *N*=10), while the iSuc-PseOpt (Jia et al., 2016a), a tool for predicting lysine succinylation sites, used *N*=15 amino acid residues of the upstream/downstream of the modified site. Therefore, we tested *N* only between 10 to 15. As shown in **Figure 2**, the ROC curves of 3-fold cross validation under various *N* were plotted. The best AUC (N=13) is 0.92, while the worst (*N*=15) is 0.73. Therefore, we set N to 13.

ROC curves of 3-fold cross validation over the training set for single type of IEE and for single type of CKSAAP features were shown in **Figure 3A**. The IEE outperformed the CKSAAP and the combination of two. ROC curves of independent test were plotted in **Figure 3B**. Obviously, the combination performs best, followed by the CKSAAP and then by the IEE feature. The single performance of the IEE feature is best over the training set, but worst over the testing set. The single performance of the CKSAAP is worst over the training set. The combination of IEE and CKSAAP features performs most stable, with 0.92 of AUC over 3-fold cross validation and 0.80 of AUC over independent test respectively.

Analysis Of Sequence Pattern

We used the WebLogo program (Crooks et al., 2004) to draw a sequence logo of all the 26 positive samples both from the training and the testing sets, as shown in **Figure 4A**. The stacks at the positions 13, 25 and 26 is higher, followed by the positions 22, 18 and 11, indicating that these positions would be more evolutionarily conservative. On the contrary, the stacks at the positions 1, 7, 8 and 19 is lower, implying these positions would be less conservative. The symbols A (alanine) at the positions 3, 6, 12, 13, and 26, K (lysine) at the positions 5, 10, 18, 21 and 24, G (glycine) at the positions 9, 11 and 22, and R (arginine) at the position 25 are higher at respective stack, indicating that these amino acids alanine, lysine, glycine and arginine would appear more frequently at these corresponding positions. The two-sample sequence logo was plotted using a web-based software (Vacic et al., 2006) http://www.twosamplelogo.org/index.html. The positive samples were 26 non-redundant fragments containing butyrylation sites, while the negative ones were 36 fragments, 62 in total. In comparison to previous single-sample sequence logo, the two-sample logo more intuitively exhibited statistically significant differential residues between two classes. As shown in **Figure 4B** , the symbols K at these positions 21 and 22, A at these positions 3, 13,19 and 20, P (proline) at the position 2, M (methionine) at the position 9, Q (glutamine) at the position 10, S (serine) at the position 12, G and R at the position 25, were enriched in the butyrylation fragments, while G at the position 1, A at the position 9, K at the position 13, S at the position 22, V (valine) at the positions 15 and 25, and T (threonine) at the position 25 were depleted. Combining the information from **Figures 4A, B**, we speculated that alanine at the position 3 and 13, lysine at the position 21 and arginine at the position 25 would be associated with histone butyrylation.

Analysis of Information Entropy Feature

As shown in **Figure 5**, we calculated information entropies of all the used positive and the negative samples in the experiment using the equations (1) and (2). Regardless of amino acid or position, information entropies of butyrylation wholly are less than those of non-butyrylation, indicating that the distribution of amino acid followed more a rule in the butyrylation than at random. The information entropies of C (cysteine) and W (tryptophan) are near or equal to zero (**Figure 5A**), implying that two types of amino acid would occur in a fixed way not at random. The information entropies of F (phenylalanine) and N (asparagine) are much less in the butyrylation than in the non-butyrylation, indicating that phenylalanine and asparagine would play a role in the butyrylation. Information entropies of G, P, M and R in the positive sample is approximately equal or more than those in the negative samples, respectively. This indicated non-difference of these amino acids between butyrylation and non-butyrylation. The information entropies of position in the butyrylation is less than those in the non-butyrylation exception the position 14 (**Figure 5B**), indicating that amino acid distribution in the butyrylation would follow more rules than at random.

Analysis of CKSAAP Feature

We calculated pairs of amino acid separated by up to one residue. Namely, amino acid pair might be of such form as $\alpha\beta$ and $\alpha\Delta\beta$, where Δ represent an amino acid. **Figure 6** shows frequency of pair of amino acid. Obviously, distribution of amino acid pairs in

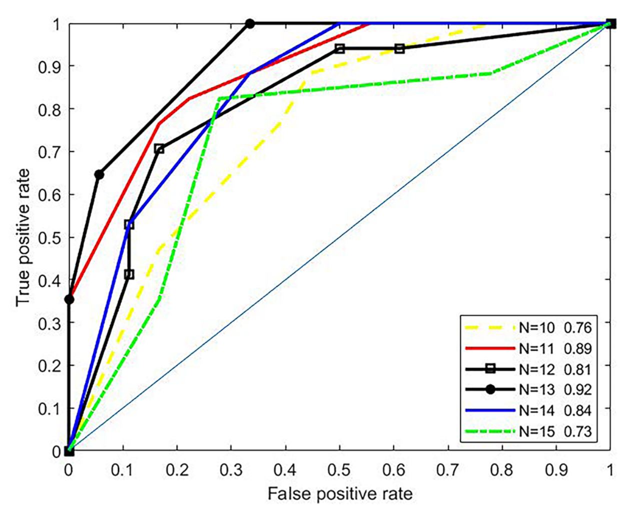


FIGURE 2 | ROC curves under various parameter *N*.

the butyrylation differs largely from that in the non-butyrylation. The butyrylation focuses mainly on these amino acid pairs of DN, GG, GK, KA, KD, KL, KP, KS, KV, PE, RH, RN, VY, XM and XX, while the non-butyrylation on GK, KA, KK and XX.

Analysis of Function for Histone Butyrylation

We used the PANTHER classification system (Mi et al., 2013) (http://www.pantherdb.org/) for functional analysis of histone butyrylation. Both statistical over-representation tests of *Homo sapiens* butyrylation histones against the whole *H. sapiens* genes and of *Mus musculus* butyrylation histones against the whole *M. musculus* genes were performed. The significantly over-represented GO terms ($P < 0.05$) for biological process, molecular function and cellular composition are listed in **Supplementary Table 2–7**. It is obviously observed that all GO terms of *M. musculus* butyrylated histones appeared in the *H. sapiens* histones, except cytosol (GO:0005829) which is defined as the part of the cytoplasm which does not contain organelles but contain such particulate matter as protein complexes. However, some GO terms of *H. sapiens* butyrylated histones fail to fall into the set of GO terms of *M. musculus* histones. For example, in terms of molecular function, STAT family protein binding (GO:0097677), RNA polymerase II core promoter sequence-specific DNA binding (GO:0000979), core promoter sequence-specific DNA binding (GO:0001046), core promoter binding (GO:0001047), chromatin binding (GO:0003682), protein-containing complex binding (GO:0044877), protein binding (GO:0005515) and binding (GO:0005488) are significant over-represented GO terms in *H. sapiens* butyrylation histone, not in *M. musculus* histones. The difference of the first four molecular functions between two species would be caused by the small-sample question. The number of studied *M. musculus* butyrylated histones is 17, less than the number of *H. sapiens* histones. The term GO:0097677 appeared two times, and these three terms GO:0000979, GO:0001046 and GO:0001047 appeared three times in these 30 butyrylated *H. sapiens* histones, while they would likely appear less than two times in these 17 butyrylated *M. musculus* histones. Only functions appearing two times or more would be statistically analyzed. Therefore, these four molecular functions could not separate *H. sapiens* from *M. musculus* histones. GO:0044877 appeared 10 times, GO:0003682 11 times, GO:0005515 29 times, while GO:0005488 appeared 30 times in the *H. sapiens* histone. It is rational to infer occurring more than two times in 17 *M. musculus* butyrylation histones, but they were not significant over-represented GO terms. This indicated that

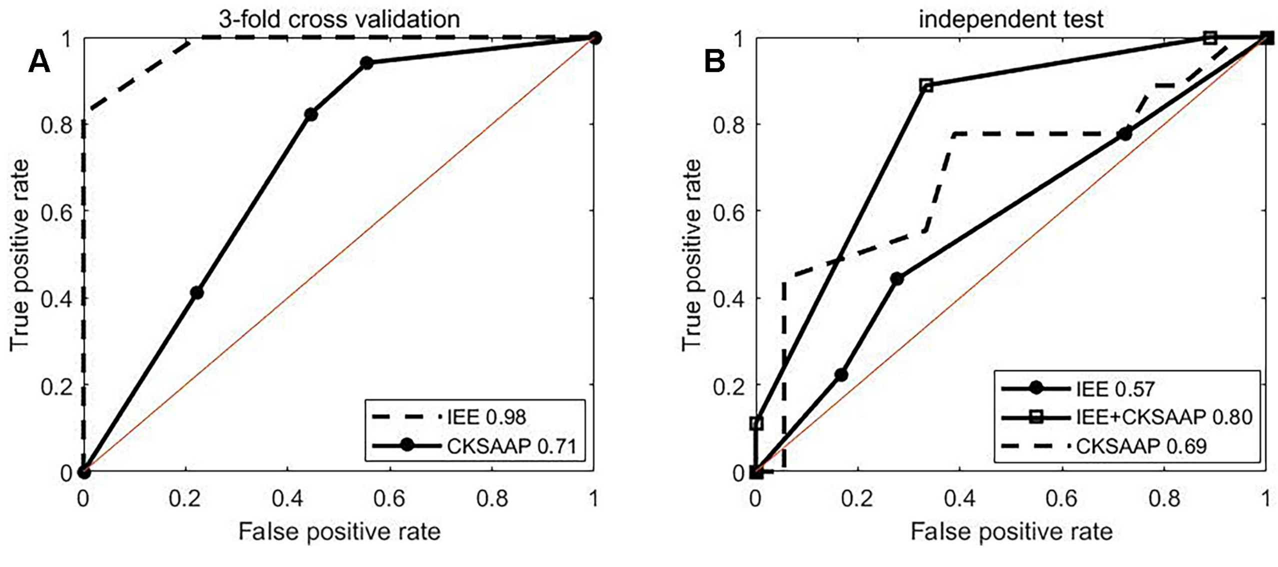


FIGURE 3 | ROC curves. **(A, B)** depict ROC curves of 3-fold cross validation and independent test, respectively.

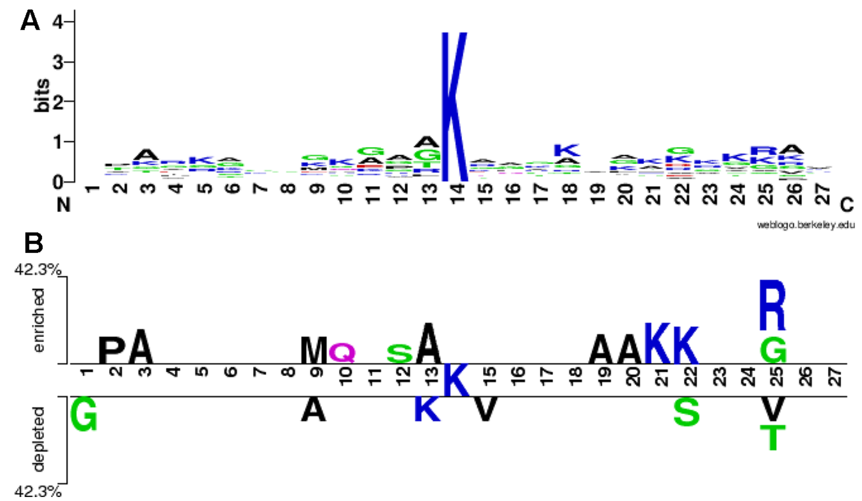


FIGURE 4 | Sequence logo. **(A)** is sequence logo of all the positive samples and **(B)** is sequence logo of all the positive and the negative samples.

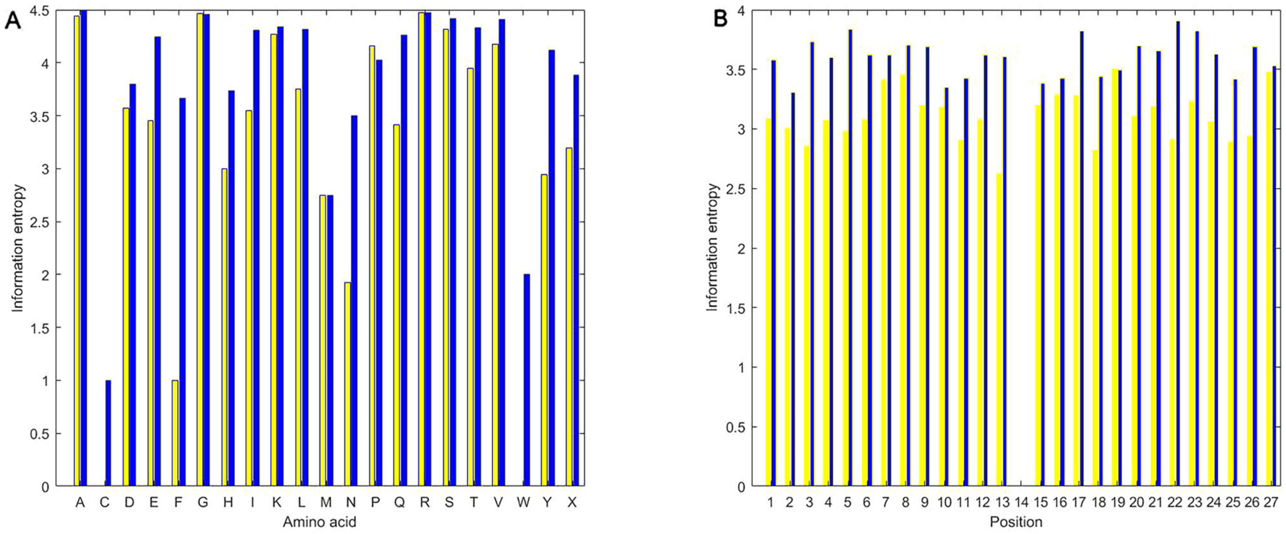


FIGURE 5 | Information entropies. **(A)** represents information entropies of PIEA and NIEA. **(B)** represents information entropies of PIEP and NIEP.

these later four molecular functions were enriched only in the *H. sapiens*, not in all the species.

Table 1 listed the most significant five GO terms of molecular function, biological process and cellular component in the butyrylated *H. sapiens* histone which all belonged to the set of the over-represent GO terms in the *M. musculus* histones respectively. These three terms GO:0003677 (DNA binding), GO:0003676 (nucleic acid binding) and GO:0031492 (nucleosomal DNA binding) are defined as interacting selectively and non-covalently with DNA, with any nucleic acid and with the DNA portion of a nucleosome, respectively. GO:0046982 (protein heterodimerization activity) is defined as interacting selectively and non-covalently with a non-identical protein to form a heterodimer, whose relationship with GO:0046983 (protein dimerization activity) is "is a". All the five terms belongs to the ancestor GO:0005488 (binding) via the "is a" relationship, implying that butyrylation histones could bind other molecules such as DNA, nucleic acid or protein. GO:0006334 (nucleosome assembly) is defined as the aggregation, arrangement and bonding together of a nucleosome, the beadlike structural units of eukaryotic chromatin composed of histones and DNA, which is of "is a" relationship with GO:0034728 (nucleosome organization) and of "part of " relationship with GO:0031497 (chromatin assembly). The term GO:0031497 is of "part of" relationship with GO:0006323 (DNA packaging) and of "is a" relationship with

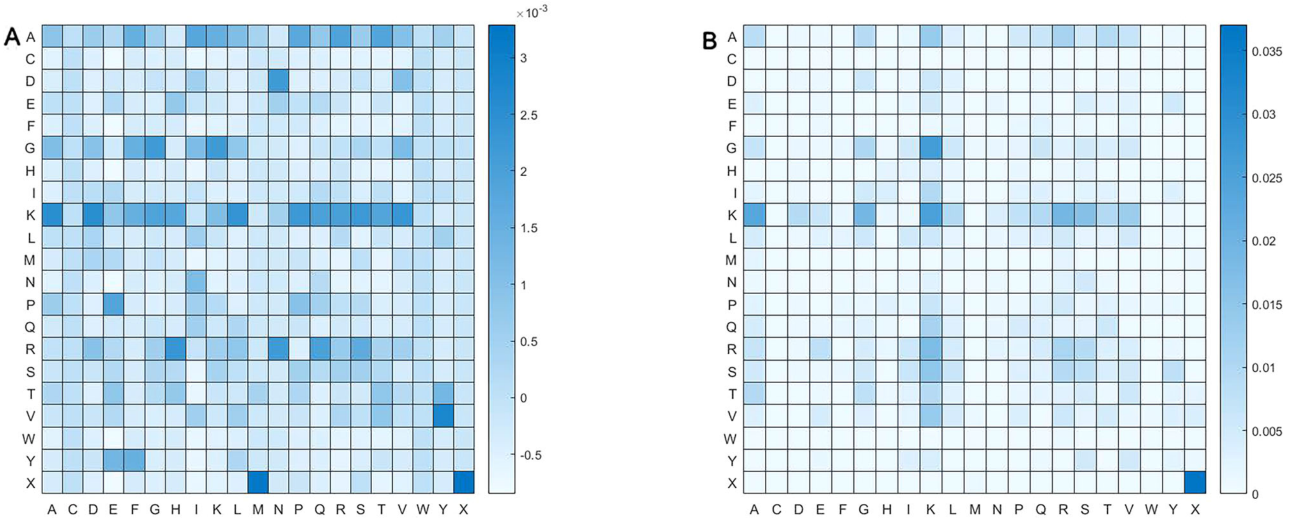


FIGURE 6 | Heatmap of amino acid pair. **(A)** represents heatmap of all the positive samples and **(B)** heatmap of all the negative samples.

TABLE 1 | Most significant five GO terms of molecular function, biological process and cellular component for *Homo sapiens*.

Molecular function	Biological process	Cellular component
Protein heterodimerization activity (GO:0046982)	Nucleosome assembly (GO:0006334)	Nucleosome (GO:0000786)
DNA binding (GO:0003677)	Chromatin assembly (GO:0031497)	DNA packaging complex (GO:0044815)
Protein dimerization activity (GO:0046983)	Chromatin assembly or disassembly (GO:0006333)	Protein-DNA complex (GO:0032993)
Nucleic acid binding (GO:0003676)	Nucleosome organization (GO:0034728)	Chromatin (GO:0000785)
Nucleosomal DNA binding (GO:0031492)	DNA packaging (GO:0006323)	Chromosome (GO:0005694)

GO:0006333 (chromatin assembly or disassembly). These five terms finally are traced up to two terms: GO:0016043 (cellular component organization) and GO:0044085 (cellular component biogenesis), indicating that butyrylation histones might be associated with these processes of biosynthesis, assembly, arrangement or disassembly. The term GO:0000786 (nucleosome) refers to a complex consisting of DNA wound around a multi-subunit core and associated proteins, which forms the primary packing unit of DNA into higher order structures. The term GO:0000786 is of "is a" relationships both with the term GO:0044815 (DNA packaging complex) and with GO:0032993 (protein-DNA complex) and is of "part of" relationship with the term GO:0000785 (chromatin) which is of part of relationship with the term GO:0005694 (chromosome). These results indicate that butyrylation histone might be located in a complex composed of DNA, proteins, *etc.*

We used the *David* (Database for Annotation, Visualization and Integrated Discovery) (Huang da et al., 2009a; Huang da et al., 2009b) to explore biological pathways in which the butyrylated histones are potential to be involved. The *David* is one of most popular tool for enrichment analysis of gene function, currently including over 40 annotation categories, such as ordinary GO terms, protein functional domains, bio-pathways, *etc.* The backgrounds for *H. sapiens* and *M. musculus* butyrylation histones were respectively the whole *H. sapiens* and the whole *M. musculus* genes. The statistically significant Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (*P*-value < 0.01) are systemic lupus erythematosus, alcoholism, viral carcinogenesis and transcriptional misregulation in cancer, whether for *H. sapiens* or for *M. musculus* genes, indicating that histone butyrylation is involved in similar bio-pathway.

CONCLUSION

Histone butyrylation is a newly discovered PTM, whose mechanism remains unknown. In this paper, we presented an approach based on information entropy and machine learning for identifying histone butyrylation sites. To the best of our knowledge, this is the first computational method for identifying histone butyrylation sites. By comparing sequences, IEE and CKSAAP between butyrylation and non-butyrylation, we found some specific characteristics implying potential and hidden pattern of histone butyrylation. The statistical test suggests that the butyrylation histone might be of binding with other molecules, be associated with the processes of biosynthesis, assembly, arrangement or disassembly, be located in the complex of DNA, protein, *etc*, and be involved in the such pathway as systemic lupus erythematosus, alcoholism, viral carcinogenesis and transcriptional misregulation in cancer.

DATA AVAILABILITY STATEMENT

Butyrylated proteins were retrieved by searching both the Uniprot database(UniProt Consortium, 2018): https://www.uniprot.org/ and the Protein Lysine Modifications Database: http://plmd.biocuckoo.org/_ (P68432; P62804; P62807;P02294; Q6DN03; Q3SZB8; O75367; P0C0S5; P16104; Q02539; Q75WM6; Q8IZA3; Q92522).

AUTHOR CONTRIBUTIONS

GH, YZ and Z-GY conceived the method. GH and YZ collected data. GH and YZ performed the experiment. GH, YZ, Y-QW, G-SH and Z-GY analyzed the results and wrote the manuscript. All the authors read the manuscript and approved the final manuscript.

FUNDING

This work is supported by National Natural Science Foundation of China (61672356, 11871061), by Scientific Research Fund of Hunan Provincial Education Department (18A394), and by the open project of Hunan Key Laboratory for Computation and Simulation in Science and Engineering (2019LCESE03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01325/full#supplementary-material

REFERENCES

Basu, A., Rose, K. L., Zhang, J., Beavis, R. C., Ueberheide, B., Garcia, B. A., et al. (2009). Proteome-wide prediction of acetylation substrates. *Proc. Natl. Acad. Sci. U. S. A.* 106 (33), 13785–13790. doi: 10.1073/pnas.0906801106

Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4 (6), 1633–1649. doi: 10.1002/pmic.200300771

Breiman, L. (2001). Random forests. *Mach. Learn.* 45 (1), 5–32. doi: 10.1023/A:1010933404324

Cai, Y., Huang, T., Hu, L., Shi, X., Xie, L., and Li, Y. (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids* 42 (4), 1387–1395. doi: 10.1007/s00726-011-0835-0

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., and Honavar, V. (2007). Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinf.* 8, 438. doi: 10.1186/1471-2105-8-438

Chen, H., Xue, Y., Huang, N., Yao, X., and Sun, Z. (2006). MeMo: a web tool for prediction of protein methylation modifications. *Nucleic Acids Res.* 34 (Web Server issue), W249–W253. doi: 10.1093/nar/gkl233

Chen, Y., Sprung, R., Tang, Y., Ball, H., Sangras, B., Kim, S. C., et al. (2007). Lysine propionylation and butyrylation are novel post-translational modifications in histones. *Mol. Cell Proteomics* 6 (5), 812–819. doi: 10.1074/mcp.M700021-MCP200

Chen, Y. Z., Tang, Y. R., Sheng, Z. Y., and Zhang, Z. (2008). Prediction of mucin-type O-glycosylation sites in mammalian proteins using the composition of k-spaced amino acid pairs. *BMC Bioinf.* 9, 101. doi: 10.1186/1471-2105-9-101

Chen, Z., Chen, Y. Z., Wang, X. F., Wang, C., Yan, R. X., and Zhang, Z. (2011). Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs. *PloS One* 6 (7), e22930. doi: 10.1371/journal.pone.0022930

Choudhary, C., Weinert, B. T., Nishida, Y., Verdin, E., and Mann, M. (2014). The growing landscape of lysine acetylation links metabolism and cell signalling. *Nat. Rev. Mol. Cell Biol.* 15 (8), 536–550. doi: 10.1038/nrm3841

Chuang, G.-Y., Boyington, J. C., Joyce, M. G., Zhu, J., Nabel, G. J., Kwong, P. D., et al. (2012). Computational prediction of N-linked glycosylation incorporating structural properties and patterns. *Bioinformatics* 28 (17), 2249–2255. doi: 10.1093/bioinformatics/bts426

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14 (6), 1188–1190. doi: 10.1101/gr.849004

Gnad, F., Ren, S., Choudhary, C., Cox, J., and Mann, M. (2010). Predicting post-translational lysine acetylation using support vector machines. *Bioinformatics* 26 (13), 1666–1668. doi: 10.1093/bioinformatics/btq260

Goudarzi, A., Zhang, D., Huang, H., Barral, S., Kwon, O. K., Qi, S., et al. (2016). Dynamic competing histone H4 K5K8 acetylation and butyrylation are hallmarks of highly active gene promoters. *Mol. Cell* 62 (2), 169–180. doi: 10.1016/j.molcel.2016.03.014

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* 11 (1), 10–18. doi: 10.1145/1656274.1656278

Hamamoto, R., Saloura, V., and Nakamura, Y. (2015). Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nat. Rev. Cancer* 15 (2), 110–124. doi: 10.1038/nrc3884

Hamby, S. E., and Hirst, J. D. (2008). Prediction of glycosylation sites using random forests. *BMC Bioinf.* 9, 500. doi: 10.1186/1471-2105-9-500

Hansen, J. E., Lund, O., Engelbrecht, J., Bohr, H., and Nielsen, J. O. (1995). Prediction of O-glycosylation of mammalian proteins: specificity patterns of UDP-GalNAc: polypeptide N-acetylgalactosaminyltransferase. *Biochem. J.* 308, 801–813. doi: 10.1042/bj3080801

Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L., and Brunak, S. (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj. J.* 15 (2), 115–130. doi: 10.1023/a:1006960004440

Hasan, M. M., Yang, S., Zhou, Y., and Mollah, M. N. (2016). SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Mol. Biosyst.* 12 (3), 786–795. doi: 10.1039/c5mb00853k

Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., et al. (2014). LAceP: lysine acetylation site prediction using logistic regression classifiers. *PloS One* 9 (2), e89575. doi: 10.1371/journal.pone.0089575

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37 (1), 1–13. doi: 10.1093/nar/gkn923

Huang da, W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4 (1), 44–57. doi: 10.1038/nprot.2008.211

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26 (5), 680–682. doi: 10.1093/bioinformatics/btq003

Huang, G., Lu, L., Feng, K., Zhao, J., Zhang, Y., Xu, Y., et al. (2014). Prediction of S-nitrosylation modification sites based on kernel sparse representation classification and mRMR algorithm. *BioMed. Res. Int.* 2014, 438341. doi: 10.1155/2014/438341

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K. C. (2016a). iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Anal. Biochem.* 497, 48–56. doi: 10.1016/j.ab.2015.12.009

Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K. C. (2016b). pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *J. Theor. Biol.* 394, 223–230. doi: 10.1016/j.jtbi.2016.01.020

Kiemer, L., Bendtsen, J. D., and Blom, N. (2005). NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics* 21 (7), 1269–1270. doi: 10.1093/bioinformatics/bti130

Li, S., Liu, B., Zeng, R., Cai, Y., and Li, Y. (2006). Predicting O-glycosylation sites in mammalian proteins by using SVMs. *Comput. Biol. Chem.* 30 (3), 203–208. doi: 10.1016/j.compbiolchem.2006.02.002

Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J., and Xue, Y. (2011). GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Mol. Biosyst.* 7 (4), 1197–1204. doi: 10.1039/c0mb00279h

Lu, Y., Xu, Q., Liu, Y., Yu, Y., Cheng, Z. Y., Zhao, Y., et al. (2018). Dynamics and functional interplay of histone lysine butyrylation, crotonylation, and acetylation in rice under starvation and submergence. *Genome Biol.* 19 (1), 144. doi: 10.1186/s13059-018-1533-y

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nat. Protoc.* 8 (8), 1551–1566. doi: 10.1038/nprot.2013.092

Qiu, W. R., Xiao, X., Lin, W. Z., and Chou, K. C. (2014). iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. *BioMed. Res. Int.* 2014, 947416. doi: 10.1155/2014/947416

Sasaki, K., Nagamine, N., and Sakakibara, Y. (2009). Support vector machine prediction of N-and O-glycosylation sites using whole sequence information and subcellular localization. *IPSJ Trans. Bioinf.* 2, 25–35. doi: 10.2197/ipsjtbio.2.25

Shao, J., Xu, D., Tsai, S. N., Wang, Y., and Ngai, S. M. (2009). Computational identification of protein methylation sites through bi-profile Bayes feature extraction. *PloS One* 4 (3), e4920. doi: 10.1371/journal.pone.0004920

Shi, S. P., Qiu, J. D., Sun, X. Y., Suo, S. B., Huang, S. Y., and Liang, R. P. (2012a). PLMLA: prediction of lysine methylation and lysine acetylation by combining multiple features. *Mol. Biosyst.* 8 (5), 1520–1527. doi: 10.1039/c2mb05502c

Shi, S. P., Qiu, J. D., Sun, X. Y., Suo, S. B., Huang, S. Y., and Liang, R. P. (2012b). PMeS: prediction of methylation sites based on enhanced feature encoding scheme. *PloS One* 7 (6), e38772. doi: 10.1371/journal.pone.0038772

Shi, Y., Guo, Y., Hu, Y., and Li, M. (2015). Position-specific prediction of methylation sites from sequence conservation based on information theory. *Sci. Rep.* 5, 12403. doi: 10.1038/srep12403

Shien, D. M., Lee, T. Y., Chang, W. C., Hsu, J. B., Horng, J. T., Hsu, P. C., et al. (2009). Incorporating structural characteristics for identification of protein methylation sites. *J. Comput. Chem.* 30 (9), 1532–1543. doi: 10.1002/jcc.21232

Tung, C. W. (2013). Prediction of pupylation sites using the composition of k-spaced amino acid pairs. *J. Theor. Biol.* 336, 11–17. doi: 10.1016/j.jtbi.2013.07.009

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 46 (5), 2699. doi: 10.1093/nar/gky092

Vacic, V., Iakoucheva, L. M., and Radivojac, P. (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22 (12), 1536–1537. doi: 10.1093/bioinformatics/btl151

Wang, X. B., Wu, L. Y., Wang, Y. C., and Deng, N. Y. (2009). Prediction of palmitoylation sites using the composition of k-spaced amino acid pairs. *Protein Eng. Des. Sel.* 22 (11), 707–712. doi: 10.1093/protein/gzp055

Wang, L. N., Shi, S. P., Xu, H. D., Wen, P. P., and Qiu, J. D. (2017). Computational prediction of species-specific malonylation sites via enhanced characteristic strategy. *Bioinformatics* 33 (10), 1457–1463. doi: 10.1093/bioinformatics/btw755

Wei, L., Xing, P., Shi, G., Ji, Z. L., and Zou, Q. (2018). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Trans. Comput. Biol. Bioinf.* PP (99), 1–1. doi: 10.1109/TCBB.2017.2670558

Xu, J., He, Y., Qiang, B., Yuan, J., Peng, X., and Pan, X. M. (2008). A novel method for high accuracy sumoylation site prediction from protein sequences. *BMC Bioinf.* 9, 8. doi: 10.1186/1471-2105-9-8

Xu, Y., Wang, X.-B., Ding, J., Wu, L.-Y., and Deng, N.-Y. (2010). Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *J. Theor. Biol.* 264 (1), 130–135. doi: 10.1016/j.jtbi.2010.01.013

Xu, Y., Ding, J., Huang, Q., and Deng, N.-Y. (2013). Prediction of protein methylation sites using conditional random field. *Protein Pept. Lett.* 20 (1), 71–77. doi: 10.2174/092986613804096865

Xu, G., Wang, J., Wu, Z., Qian, L., Dai, L., Wan, X., et al. (2014). SAHA regulates histone acetylation, Butyrylation, and protein expression in neuroblastoma. *J. Proteome Res.* 13 (10), 4211–4219. doi: 10.1021/pr500497e

Xu, H. D., Shi, S. P., Wen, P. P., and Qiu, J. D. (2015a). SuccFind: a novel succinylation sites online prediction tool *via* enhanced characteristic strategy. *Bioinformatics* 31 (23), 3748–3750. doi: 10.1093/bioinformatics/btv439

Xu, Y., Ding, Y. X., Ding, J., Lei, Y. H., Wu, L. Y., and Deng, N. Y. (2015b). iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Sci. Rep.* 5, 10184. doi: 10.1038/srep10184

Xu, Y., Ding, J., and Wu, L. Y. (2016). iSulf-Cys: prediction of S-sulfenylation sites in proteins with physicochemical properties of amino acids. *PloS One* 11 (4), e0154237. doi: 10.1371/journal.pone.0154237

Xu, H., Zhou, J., Lin, S., Deng, W., Zhang, Y., and Xue, Y. (2017). PLMD: an updated data resource of protein lysine modifications. *J. Genet. Genomics* 44 (5), 243–250. doi: 10.1016/j.jgg.2017.03.007

Xu, J. Y., Xu, Z., Liu, X., Tan, M., and Ye, B. C. (2018). Protein acetylation and butyrylation regulate the phenotype and metabolic shifts of the endospore-forming *Clostridium acetobutylicum*. *Mol. Cell Proteomics* 17 (6), 1156–1169. doi: 10.1074/mcp.RA117.000372

Xue, Y., Chen, H., Jin, C., Sun, Z., and Yao, X. (2006). NBA-Palm: prediction of palmitoylation site implemented in naive Bayes algorithm. *BMC Bioinf.* 7, 458. doi: 10.1186/1471-2105-7-458

Zhang, K., Chen, Y., Zhang, Z., and Zhao, Y. (2008). Identification and verification of lysine propionylation and butyrylation in yeast core histones using PTMap Software. *J. Proteome Res.* 8 (2), 900–906. doi: 10.1021/pr8005155

Zhang, N., Li, B.-Q., Gao, S., Ruan, J.-S., and Cai, Y.-D. (2012). Computational prediction and analysis of protein γ-carboxylation sites based on a random forest method. *Mol. Biosyst.* 8 (11), 2946–2955. doi: 10.1039/c2mb25185j

Zhang, W., Xu, X., Yin, M., Luo, N., Zhang, J., and Wang, J. (2013). Prediction of methylation sites using the composition of k-spaced amino acid pairs. *Protein Pept. Lett.* 20 (8), 911–917. doi: 10.2174/0929866511320080008

Zhang, Y., Tang, L., Zou, H., Yang, Q., Yu, X., Jiang, J., et al. (2015). Identifying protein arginine methylation sites using global features of protein sequence coupled with support vector machine optimized by particle swarm optimization algorithm. *Chemom. Intell. Lab. Syst.* 146, 102–107. doi: 10.1016/j.chemolab.2015.05.011

Zhao, X., Zhang, W., Xu, X., Ma, Z., and Yin, M. (2012). Prediction of protein phosphorylation sites by using the composition of k-spaced amino acid pairs. *PloS One* 7 (10), e46302. doi: 10.1371/journal.pone.0046302

Zhao, X., Dai, J., Ning, Q., Ma, Z., Yin, M., and Sun, P. (2013). Position-specific analysis and prediction of protein pupylation sites based on multiple features. *BioMed. Res. Int.* 2013, 109549. doi: 10.1155/2013/109549

Zhou, F. F., Xue, Y., Chen, G. L., and Yao, X. (2004). GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun.* 325 (4), 1443–1448. doi: 10.1016/j.bbrc.2004.11.001

Zhou, F., Xue, Y., Yao, X., and Xu, Y. (2006). CSS-Palm: palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics* 22 (7), 894–896. doi: 10.1093/bioinformatics/btl013

Zhou, Y., Huang, T., Huang, G., Zhang, N., Kong, X., and Cai, Y.-D. (2016). Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method. *Neurocomputing* 217, 53–62. doi: 10.1016/j.neucom.2015.10.148

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

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 18 February 2020
doi: 10.3389/fbioe.2020.00054

LINC00346 Acts as a Competing Endogenous RNA Regulating Development of Hepatocellular Carcinoma via Modulating CDK1/CCNB1 Axis

Jinglan Jin[1], Hongqin Xu[1], Wanyu Li[1], Xiaotong Xu[1], Huan Liu[2] and Feng Wei[2]*

[1] Department of Hepatology, The First Hospital of Jilin University, Changchun, China, [2] Department of Hepatobiliary and Pancreatic Surgery, The First Hospital of Jilin University, Changchun, China

Hepatocellular carcinoma (HCC) is one of the important types of liver cancer. LncRNA is an important regulatory factor that regulates many biological processes such as tumor cells during tumorigenesis and metastasis. LINC00346 has been associated with various types of liver cancer, but its role and regulatory mechanism in HCC remain unclear. In our study, we found the LINC00356-miR-199a-3p-CDK1/CCNB1 axis through bioinformatics analysis. The expressions of LINC00356, miR-199a-3p, CDK1, and CCNB1 in HCC and normal hepatocytes were determined by qRT-PCR and WB. The results showed that LINC00356, CDK1 and CCNB1 were highly expressed in HCC, while miR-199a-3p was lowly expressed. Dual luciferase reporter gene assay, RIP and RNA-pull down assays demonstrated the targeted binding relationship of LINC00346-miR-199a-3p-CDK1/CCNB1. Overexpressing LINC00460 and silencing miR-199a-3p promoted cell invasion, inhibited apoptosis of HCC, and arrested the cell cycle in S phase while opposite results were obtained when silencing LINC00346, CDK1, and CCNB1. LINC00346 indirectly affects liver cancer by promoting the expression of CDK1/CCNB1 through competitive adsorption of miR-199a-3p. In addition, the study also demonstrated that overexpression of LINC00346 indirectly inhibited the expression of p53 and p21 proteins by promoting CDK1/CCNB1 expressions, thereby blocking the p53 signaling pathway. These results proved that LINC00346 could regulate the expression of CDK1/CCNB1 through the competitive adsorption of miR-199a-3p, thereby affecting the p53 signaling pathway and finally regulating the apoptosis, invasion and cell cycle of HCC cells. In conclusion, LINC00346 can be used as a tumor promoter and potential therapeutic target for HCC metastasis and prognosis.

Keywords: LINC00346, miR-1991-3p, CDK1, CCNB1, p53 signaling pathway, hepatocellular carcinoma

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Da Li,
Sir Run Run Shaw Hospital, China
Jiangshan Lian,
Zhejiang University, China

***Correspondence:**
Feng Wei
w_f_eng@163.com

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

Received: 19 December 2019
Accepted: 22 January 2020
Published: 18 February 2020

Citation:
Jin J, Xu H, Li W, Xu X, Liu H and Wei F (2020) LINC00346 Acts as a Competing Endogenous RNA Regulating Development of Hepatocellular Carcinoma via Modulating CDK1/CCNB1 Axis. Front. Bioeng. Biotechnol. 8:54. doi: 10.3389/fbioe.2020.00054

INTRODUCTION

Primary liver cancer is one of the second leading causes of death worldwide, and hepatocellular carcinoma (HCC) is one of the major types of it (Sia et al., 2017). In recent years, incidence rates continue to increase rapidly for liver cancer, by about 3% per year in women and 4% per year in men (Siegel et al., 2017). Although the treatment of HCC has improved significantly in recent decades, including surgical resection, liver transplantation, radiotherapy and chemotherapy, the overall 5-year survival rate is not improved (Ulahannan et al., 2014). Over the years, the occurrence of liver cancer is considered to be a complex multi-step process involving multiple molecules and multiple signaling pathways (Setshedi et al., 2018). A better understanding of the occurrence and development of HCC will help to further improve treatment strategies. Therefore, it is of great significance to explore the mechanism of the occurrence and development of HCC, determine the effective therapeutic targets of HCC, and find a new way for HCC treatment.

Recently, increasing evidence confirms that long non-coding RNAs (lncRNAs) and microRNAs (miRNAs) have been identified as important regulators in a variety of cancers including HCC (Bartel, 2004; Li and Chang, 2014; Wang et al., 2017). Abnormal expression of lncRNA plays a key role in cancer progression and carcinogenesis through a variety of mechanisms (Chen et al., 2016). LINC00346 is located on chromosome 13q34, with a total length of 6322bp, and is up-regulated and has oncogenic effects in non-small cell lung cancer and bladder cancer (Ye et al., 2017; Zhang B. et al., 2018). Overexpression of LINC00346 was positively correlated with poor prognosis of pancreatic cancer (Brown et al., 2018). Zhang et al. (2015) have found that the up-regulated expression of LINC00346 in HCC is significantly negatively correlated with the survival of HCC patients. These results indicate that LINC00346 plays a significant oncogenic role in a variety of cancers, but the specific biological function and mechanism of LINC00346 in HCC have not been studied.

Cyclin-dependent kinase (CDK) is an important cell cycle-regulating protein, belonging to the serine/threonine kinase family, which includes catalytic kinase subunits and cell cycle protein conjugates. Only CDK1 in the CDK family can promote cell cycle independently (Santamaría et al., 2007; Malumbres and Barbacid, 2009). CDK1 has been reported to be highly expressed in human colorectal cancer (Sung et al., 2014), prostate cancer (Willder et al., 2013). Therefore, CDK1 is closely related to cancer progression. In the study of the specific role of CDK1 in cancer. Danhier et al. (2010) have reported that CDK1/CyclinB1 inhibitor JNJ-7706621 and aurora kinase combined with paclitaxel can effectively treat transplantable liver cancer and inhibit tumor growth. In addition, studies have shown that CDK1 and CCNB1 have inhibitory effects on p53 signaling pathway as regulatory factors in HCC (Qin et al., 2019). Cell cyclin B1 (CCNB1) is an important cell cycle protein whose abnormal expression plays an important role in regulating cell cycle (Jin et al., 1998). Recent studies have demonstrated that CCNB1 is highly expressed in various human cancers, including breast cancer (Niméus-Malmström et al., 2010), cervical cancer (Kreis et al., 2010) and lung cancer (Yoshida et al., 2004). Moreover, Porter et al. (2003) found that inhibition of CCNB1 nuclear export and CCNB1 accumulation in the nucleus induced apoptosis. CCNB1 has also been proved to significantly correlate with overall survival of HBV-related HCC recurrence. These results indicate that CDK1/CCNB1 are significantly positively correlated with the development of various cancers including HCC, but the specific role of these two genes in HCC has not been investigated.

In this study, we found that LINC00346 was highly expressed in HCC cells, and it regulated the expressions of CDK1/CCNB1 through competitive adsorption of miR-199-3p as a ceRNA, thereby promoting the proliferation and metastasis of HCC cells. LINC00346 also regulated the p53 signaling pathway by regulating the miR-199-3p/CDK1/CCNB1 signaling axis. These results demonstrated that LINC00346 played a significant oncogenic role in the development of HCC and LINC00346 can be used as a prognostic target and potential biomarker in the diagnosis and prognosis of HCC.

MATERIALS AND METHODS

Bioinformatics Analysis

Hepatocellular carcinoma expression datasets GSE62232 (including 10 normal samples and 81 HCC samples) and GSE74618 (including 10 normal samples and 218 HCC samples) were obtained through GEO database[1]. Normal samples were set as control, and “limma” package was used for differential analysis with the threshold of $|\text{logFC}| > 2$ and $P\text{-value} < 0.05$. KEGG pathway enrichment analysis of the differential genes was conducted by “clusterprofiler” package, and the pathview diagram was plotted by “pathview” package. Gene interaction analysis was conducted through STRING database[2], and gene interaction network map was drawn by cytocape v3.7.1. The expression levels of CDK1 and CCNB1 in TCGA[3] database were retrieved through GEPIA database[4]. The upstream regulatory miRNAs of CDK1 and CCNB1 were predicted and the binding site information of miRNA-mRNA was obtained by TargetScan database[5]. The upstream lncRNAs of miR-199a-3p were predicted by RAID database[6], and lncRNA-miRNA binding site information was obtained through RNA22 database[7].

Cell Lines and Transfection

Human normal liver cells L-02 (BNCC351907), HCC cell lines HepG2 (BNCC338070), Huh-7 (BNCC337690), Hep3b (BNCC337952), and SMMC-7721 (BNCC352197) were all purchased from BeNa Culture Collection. All cell lines were incubated with 90% Dulbecco’s Modified Eagle Medium (DMEM)-H containing 10% fetal bovine serum (FBS, Gibco, Grand Island, NY) and maintained in an incubator with 5% CO_2 at 37°C.

LINC00346, oe-LINC00346, CDK1 and CCNB1 siRNA, miR-199a-3p inhibitors, miR-199a-3p mimic or corresponding controls were obtained from GENECHEM (Shanghai, China) and transfected into cells in 6-well plates by Lipofectamine 3000 (Invitrogen, United States) according to the manufacturer’s instructions. In addition, PFT β [Pifithrin-β, p53 protein

[1] https://www.ncbi.nlm.nih.gov/geoprofiles/
[2] https://string-db.org
[3] https://www.cancer.gov/
[4] http://gepia.cancer-pku.cn/
[5] http://www.targetscan.org/vert_71/
[6] http://www.rna-society.org/raid2/index.html
[7] https://cm.jefferson.edu/rna22/

inhibitor, HY-16702, MedChemExpress (Da Pozzo et al., 2014)] was added to the culture medium at 10 μM per well plate. All cells were transfected for 48 h and collected for subsequent experiments.

qRT-PCR

Total RNA from transfected cells was extracted by TriZol reagent (Qiagen) according to the instructions. Complementary DNA (cDNA) synthesis and quantitative polymerase chain reaction (qPCR) procedures were performed for mRNA and lncRNA using PrimeScript RT Master Mix and TB Green Premix Ex Taq II (TaKaRa, Dalian, China). RNA of miRNA was isolated using the miRNeasy Mini Kit (Qiagen, Shenzhen, China). Mir-x miRNA first strand synthesis kit and Mir-x miRNA quantitative real-time polymerase chain reaction (qRT-PCR) SYBR kit (TaKaRa) were used for reverse transcription and qPCR. The primers were listed in **Table 1**. U6 snRNA was used as an internal reference for miR-199a-3p, and GAPDH was the internal reference of LINC00346, CDK1 and CCNB1. The data were analyzed by $2^{-\Delta\ \Delta\ Ct}$ method.

Western Blot (WB)

Total protein was extracted from transfected cells by RIPA lysis buffer (Beyotime, Shanghai, China). The purity of the protein in the whole extract was determined by bicinchoninic acid (BCA) protein assay kit (Pierce, Appleton, WI). The proteins were separated by SDS-PAGE and transferred to PVDF membrane (Millipore). After incubation with bovine serum albumin (BSA) in Tris-HCl buffered saline containing 0.1% TBST, the membrane was added with corresponding primary antibodies CDK1 (ab32094, 1:2000, Abcam, Cambridge, MA), CCNB1 (ab32053, 1:50000, Abcam, Cambridge, MA), p53 (ab32389, 1:1000, Abcam, Cambridge, MA), p21 (ab109520, 1:1000, Abcam, Cambridge, MA), and GAPDH (ab181602, 1: 10, 000, Abcam, Cambridge, MA). Second antibody Goat anti-rabbit IgG H&L (horseradish peroxidase, HRP) (ab6721, 1:2000, Abcam, Cambridge, MA, United States) was then used to incubate the membrane. The positive bands were detected by Immobilon Western Chemiluminescent HRP Substrate (Millipore) and the strength of the target strip was quantified by Image Lab Software (Bio-Rad).

TABLE 1 | Primer sequences.

Genes	Primer sequences
LINC00346	F: 5′-TCTCACCAGCATTTGACGCT-3′
	R: 5′-ACGTGCGCAAGTAAGTCTCA-3′
CDK1	F:5′-AAACTACAGGTCAAGTGGTAGCC-3′
	R:5′-TCCTGCATAAGCACATCCTGA-3′
CCNB1	F:5′-GACCTGTGTCAGGCTTTCTCTG-3′
	R:5′-GGTATTTTGGTCTGACTGCTTGC-3′
GAPDH	F:5′-CAGGAGGCATTGCTGATGAT-3′
	R:5′-GAAGGCTGGGGCTCATTT-3′
miR-199a-3p	F:5′-CTCACAGTAGTCTGCACA-3′
	R:5′-GACTGTTCCTCTCTTCCTC-3′
U6	F:5′-CTCGCTTCGGCAGCACA-3′
	R:5′-AACGCTTCACGAATTTGCGT-3′

Flow Cytometry (FCM)

Apoptosis assay: transfected cells for 48 h were collected and the apoptosis rate was measured by Annexin V-FITC/PI apoptosis assay kit (Beijing Biosea Biotechnology, Beijing, China). In brief, the cells were stained with 10 μL Annexin V-FITC and 5 μL propidium iodide (PI). After incubation in darkness at room temperature for 30 min, the samples were analyzed by FCM (Beckman Coulter, Fullerton, CA, United States). Annexin V-PE (+)/PI (-) represents apoptotic cells, while Annexin V-PE (+)/PI (+) represents early apoptotic or dead cells.

Cell cycle determination: after transfection for 48 h, cells were harvested and stained with PI using the CycleTest Plus DNA Reagent kit (BD) in accordance with the manufacturer's guidelines. Finally, the percentage of cells in G0/G1, S and G2/M phases was counted.

Transwell

Invasion measurement was performed using Transwell Chambers consisting of an 8 μm membrane filter (Corning Incorporated, Corning, NY, United States) coated with Matrigel (BD Biosciences, San Jose, CA, United States). Cells were trypsinized and suspended in serum-free medium. Next, 2 × 10 (Setshedi et al., 2018) cells were plated in the upper chamber, and the lower chamber was filled with a medium containing 10% FBS. After incubation for 36 h, the cells in the lower chamber were fixed with 4% paraformaldehyde and stained with crystal violet. Five fields were randomly selected and cells were counted under a microscope.

RIP Assay

The EZ-magna RIP kit (Millipore, United States) was applied to carry out the RIP assay according to the product specifications. First, the HepG2 cells were collected and lysed in a full RIP lysis buffer. Cell extracts were then incubated with RIP buffer containing magnetic beads conjugated to human AGO2 antibodies (ab32381, abcam, Cambridge, United Kingdom), and IgG antibody (ab6702, abcam, Cambridge, United Kingdom) was used as controls. The samples were incubated with protease K and oscillated to digest the protein and isolate the immunoprecipitated RNA. The concentration of RNA was then measured using a NanoDrop spectrophotometer and real-time PCR analysis of the purified RNA was performed.

Dual Luciferase Reporter Gene Assay

CDK1 and CCNB1 fragments containing miR-199a-3p binding sites were amplified by PCR and cloned into the downstream of luciferase reporter gene in pmirGLO vector, which were named CDK1-WT and CCNB1-WT. CDK1-MUT and CCNB1-MUT (mutations within the binding sites) were generated using the Quickchange XL Site-Directed Mutagenesis Kit (Stratagene) according to the manufacturer's protocol. Mimic NC and miR-199a-3p mimic were co-transfected with CDK1-WT or CDK1-MUT and CCNB1-WT or CCNB1-MUT, respectively, into HepG2 cells. After 48 h of transfection, cells were harvested and luciferase assay was performed using the dual luciferase reporter system (Promega).

RNA Pull-Down

RNA pull-down measurements were performed using the Pierce TM Magnetic RNA-Protein Pull-Down Kit (Millipore) according to the manufacturer's instructions. In brief, HepG2 cells were transfected with 3′-terminal biotin-labeled LNC00346 probe and its control probe. 24 h after transfection, the cells were incubated with streptomycin-coated magnetic beads. The expressions of LNC00346 and miR-199a-3p in the binding portion were determined by qRT-PCR.

Statistical Analysis

SPSS 21 (IBM Corp., Armonk, NY, United States) was used for statistical analysis of data between different groups. All data were expressed as Mean ±SD. The comparison between two groups was analyzed by *t*-test. $P < 0.05$ was statistically significant.

RESULTS

CDK1 and CCNB1 Are Possible Targets for HCC

In order to find the genes associated with HCC development, we firstly conducted differential analysis on GES62232 dataset in GEO database to analyze the mRNAs with significantly different expressions, and 230 DEGs in HCC were obtained (**Supplementary Table S1**). **Figure 1A** showed the expressions of the first 100 DEGs. Further analysis of KEGG function enrichment of these DEGs revealed that six genes were enriched in p53 signaling pathway (**Figure 1B**). P53 signaling pathway was believed to be involved in cell apoptosis, cell cycle and other activities, and many studies indicated that it was involved in tumor development (Smal et al., 1989; Meng et al., 2014). Protein interaction analysis was performed on these six genes (**Figure 1C**) and it was found that CDK1 and other three genes were at the core position. CDK1 and CCNB1 were chosen for the follow-up studies. The expression levels of CDK1 and CCNB1 in HCC tumor samples and normal samples from TCGA database were analyzed (**Figures 1D,E**) and we found that CDK1 and CCNB 1 were highly expressed in HCC, which suggested that CDK1 and CCNB1 may play important regulatory roles in HCC.

CDK1/CCNB1 Regulate Cell Cycle, Apoptosis and Invasion in HCC

We first detected the expression of CDK1 and CCNB1 in human normal liver cells L-02 and HCC cell lines HepG2, Huh-7, Hep3b, and SMMC-7721 by qRT-PCR. The results exhibited that both CDK1 and CCNB1were highly expressed in HCC cells (**Figure 2A**) ($P < 0.05$), and then the two HCC cell lines HepG2 and Huh-7 with higher expression of CDK1 and CCNB1 were selected for subsequent experiments. WB was used to detect the protein expressions of CDK1 and CCNB1 in L-02, HepG2 and Huh-7 cell lines (**Figure 2B**). Compared with L-02, the protein expressions of CDK1 and CCNB1 were significantly increased in HepG2 and Huh-7 cell lines, which was consistent with the expression trend of mRNA. In order to further study the functional role of CDK1 and CCNB1 in HCC, we established CDK1 and CCNB1 silencing cell lines and the silencing efficiency was detected by qRT-PCR (**Figures 2C,D**). Compared with si-NC group, CDK1 and CCNB1 were effectively silenced ($P < 0.05$), and si-CDK1-2 and si-CCNB1-2 sequences with better silencing efficiency were selected for subsequent experiments. Then, FCM was performed to detect apoptosis in si-NC group, si-CDK1 group and si-CCNB1 group (**Figure 2E**). The results showed that silencing CDK1/CCNB1 promoted apoptosis of HepG2 and Huh-7 cells in HCC ($P < 0.05$). After studying the effect of CDK1/CCNB1 on apoptosis of HCC cells, its effects on cell invasion and cell cycle were also studied. Transwell assay was used to determine cell invasion (**Figure 2F**). The results exhibited that the invasion ability of HCC cells decreased significantly after CDK1/CCNB1 was silenced ($P < 0.05$). Then FCM was used to detect the cell cycle (**Figure 2G**). The proportion of cells in G0/G1 phase in the si-CDK1 group and si-CCNB1 group increased significantly, and the proportion of cells in S phase decreased greatly ($P < 0.05$). In conclusion, silencing CDK1 or CCNB1 can promote the apoptosis of HCC cell lines HepG2 and Huh-7, inhibit cell invasion, and block cells in G0/G1 phase.

CDK1/CCNB1 Affect the Apoptosis, Invasion and Cell Cycle of HCC by Regulating p53 Pathway

After determining the biological function of CDK1/CCNB1 on HCC cells, we investigated the effect of CDK1 and CCNB1 on HCC cells by regulating p53 pathway. The transfected HepG2 cells with the highest CDK1 expression were divided into si-NC + DMSO group, si-NC + PFT β group, si-CDK1 + DMSO group, si-CDK1 + PFT β group, and transfected Huh-7 cells with the highest CCNB1 expression were divided into si-NC + DMSO group, si-NC + PFT β group, si-CCNB1 + DMSO group, si-CCNB1 + PFT β group. First, WB was used to detect the protein expressions of CDK1, CCNB1, p53 and p21 in cells in each group. As displayed in **Figures 3A,B**, silencing CDK1/CCNB1 can promote the protein expressions of p53 and p21 in cancer cells ($P < 0.05$), while the result was reverse with the addition of PFT β, a p53 pathway inhibitor. This result indicated that CDK1/CCNB1 could negatively regulate p53 signaling pathway.

In order to study the function of this regulation in HCC, we used FCM to detect the apoptosis of each group (**Figures 3C,D**). The results displayed that compared with the si-NC + DMSO group, the apoptosis rate of the si-NC + PFT β group was significantly reduced. CDK1 or CCNB1 silencing group can eliminate the inhibitory effect on apoptosis of pathway inhibitor group. Then Transwell was used to detect the cell invasion of each group (**Figures 3E,F**). The results showed that PFT β promoted invasive ability of cancer cells while CDK1 or CCNB1 silencing group canceled out the promoting effect of PFT β on invasion ability of HepG2 or Huh-7 cells ($P < 0.05$). Finally, FCM was used to detect the cell cycle (**Figures 3G,H**). We observed that the proportion of cells in the G0/G1 phase in the si-NC + PFT β group decreased significantly, while that in the S phase increased

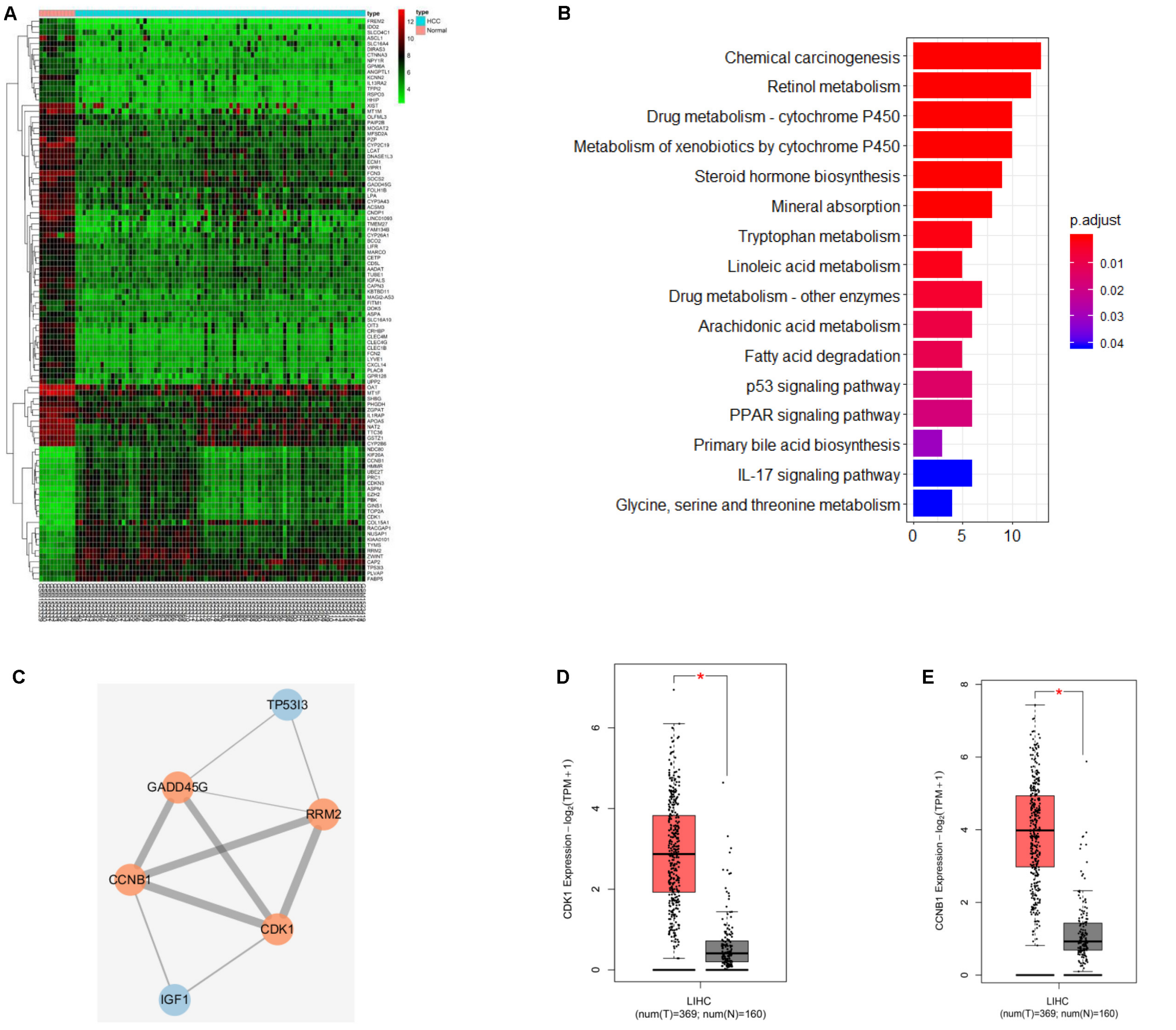


FIGURE 1 | CDK1 and CCNB1 are possible targets for HCC. **(A)** Heat map of the first 100 DEGs expression in GSE62232 dataset from GEO database; **(B)** KEGG pathway enrichment analysis of DEGs in GSE62232 dataset; **(C)** Interaction analysis of DEGs in p53 signaling pathway, orange represents high expression and blue represents low expression; **(D,E)** CDK1 and CCNB1 gene expression in HCC samples and normal samples from TCGA database, red is the tumor sample and black is the normal sample. *represents $P < 0.05$.

significantly. Silencing CDK1 or CCNB1 blocked cell cycle and eliminates the effect of PFT β. These results suggested that inhibiting CDK1/CCNB1 could promote invasion, inhibit apoptosis and regulate cell cycle of HCC cells by blocking the p53 signaling pathway.

miR-199a-3p Targeted Inhibits Both CDK1 and CCNB1 Expression

In order to investigate the upstream miRNAs that regulate CDK1/CCNB1, the TargetScan database was further used to predict the upstream regulatory miRNAs of CDK1 and CCNB1. At the same time, a miRNA expression dataset GSE74618 of HCC was obtained through the GEO database and analyzed. Finally, two miRNAs that were significantly down-regulated in HCC were obtained. The microarray analysis results and TargetScan prediction results were intersected (**Figure 4A**) and it was found that there was only one miRNA that was miR-199a-3p in the intersection. The expression level of miR-199a-3p in GSE74618 dataset was significantly lower in HCC samples (**Figure 4B**). QRT-PCR was used to detect the expression of miR-199a-3p in human normal liver cells L-02 and HCC cell lines HepG2 and Huh-7 (**Figure 4C**). The result showed that miR-199a-3p was significantly lowly expressed in HCC cells ($P < 0.05$). Next, the binding sites of miR-199a-3p and CDK1 or CCNB1 were predicted by bioinformatics website,

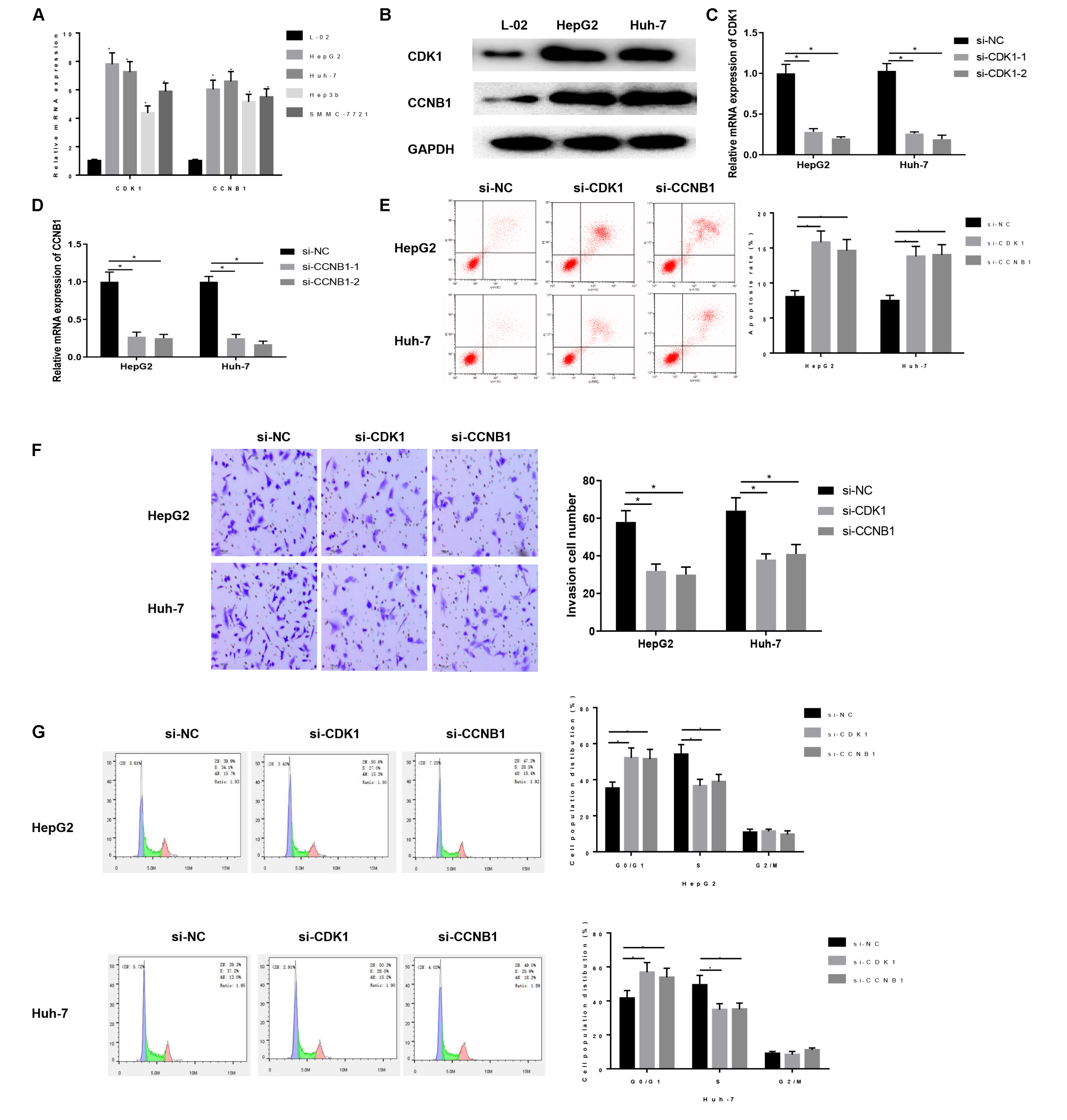


FIGURE 2 | Expression and regulation of CDK1 and CCNB1 in HCC cells. **(A)** The expressions of CDK1 and CCNB1 in normal human cells and four HCC cell lines were detected by qRT-PCR; **(B)** The protein expressions of CDK1 and CCNB1 in HepG2 and Huh-7 were detected by WB; **(C,D)** The silencing efficiency of CDK1 and CCNB1 was measured by qRT-PCR; Cells development in each group was detected by FCM and Transwell assays (100×). **(E)** Cell apoptosis; **(F)** Cell invasion; **(G)** Cell cycle. *means $P < 0.05$. Representative of three independent experiments.

respectively, to explore the targeting relationship between mir-199a-3p and CDK1 or CCNB1 (**Figure 4D**). miR-199a-3p can bind to the 3′UTR of CDK1 or CCNB1, respectively. Then RIP assay was performed on HepG2 cells to detect whether miR-199a-3p could bind with CDK1 or CCNB1, as shown in **Figure 4E**. Compared with IgG antibodies, CDK1 and CCNB1 enriched in AGO2 antibody group were significantly increased ($P < 0.05$). Moreover, dual luciferase reporter gene assay was used to verify the targeted binding relationship (**Figure 4F**). Compared with the mimic NC group, the relative fluorescence

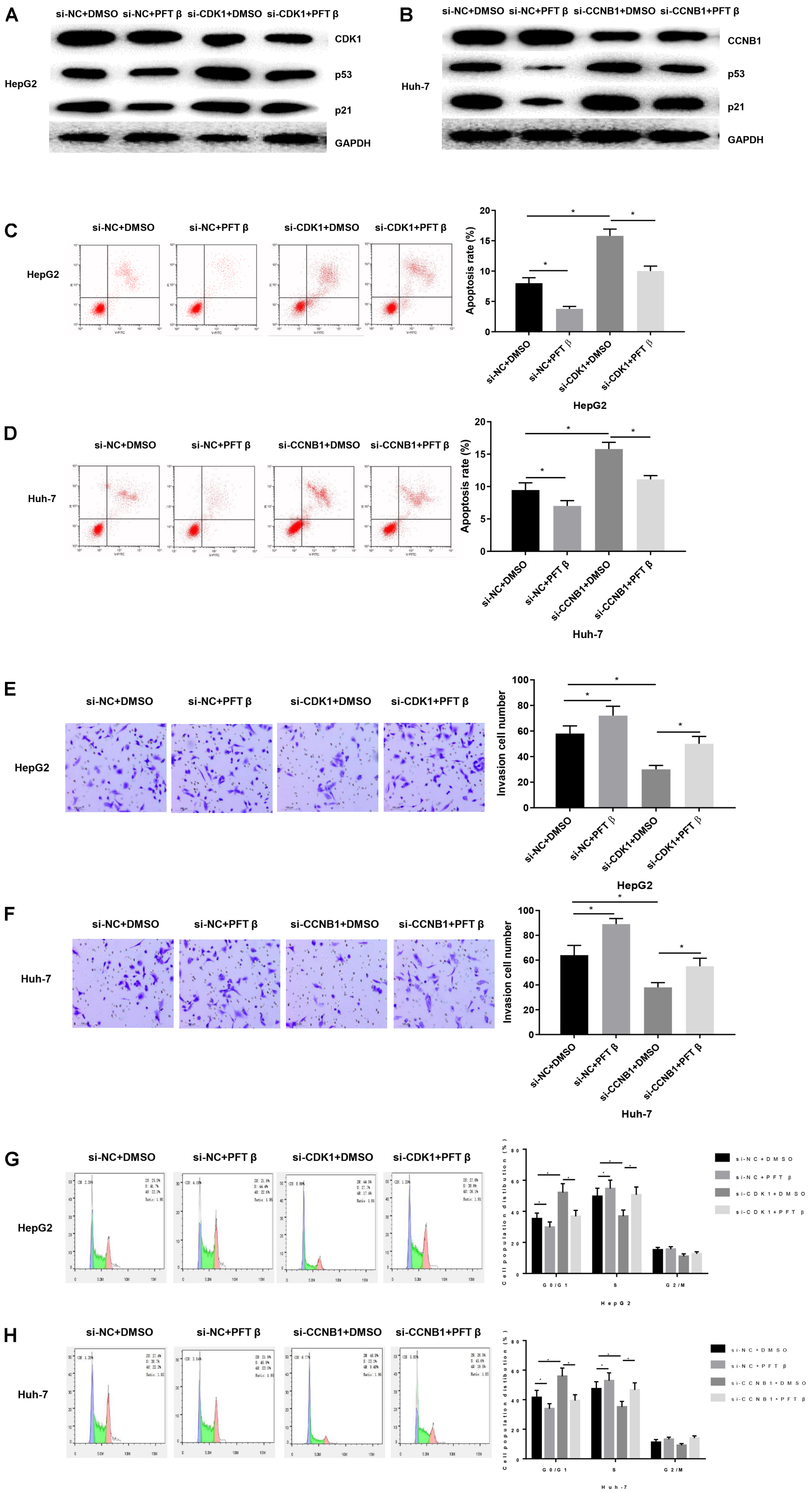


FIGURE 3 | CDK1 and CCNB1 affect the apoptosis, invasion and cell cycle of HCC by regulating p53 pathway. **(A)** The protein expressions of CDK1, p53 and p21 in HepG2 cells were detected by WB; **(B)** The protein expressions of CCNB1, p53 and p21 in Huh-7 cells were detected by WB; Cells development in each group was detected by FCM and Transwell assays (100×). **(C,D)** Cell apoptosis; **(E,F)** Cell invasion; **(G,H)** Cell cycle. *means $P < 0.05$.

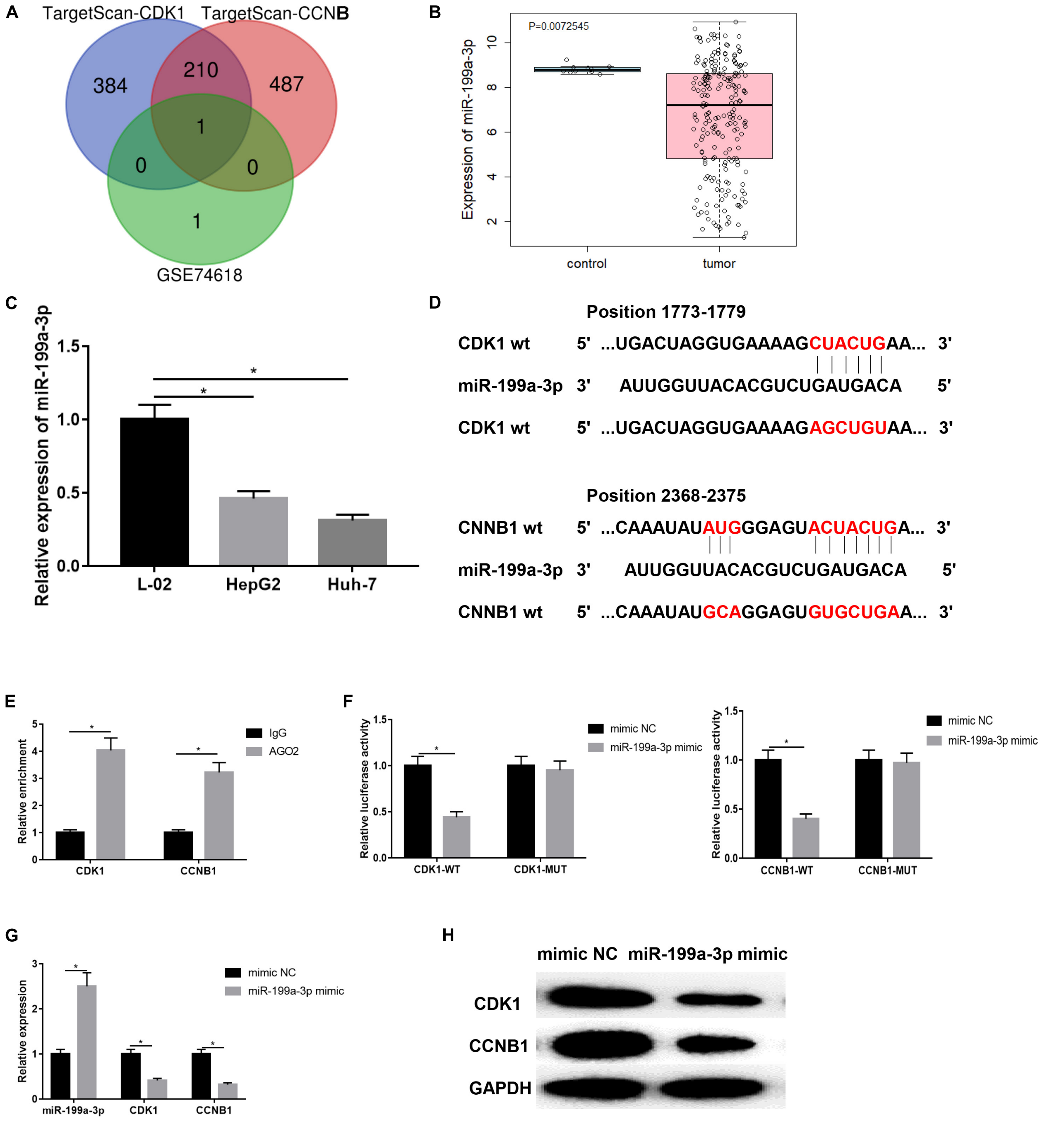


FIGURE 4 | miR-199a-3p targeted inhibited CDK1 and CCNB1 expression. **(A)** The miRNAs that regulate CDK1 and CCN1 were predicted by TargetScan and intersections with significantly down-regulated expression in GSE74618 of HCC; **(B)** miR-199a-3p expression in GSE74618 dataset, black represents the normal sample and red represents the tumor sample; **(C)** The expression of miR-199a-3p in L-02, HepG2 and Huh-7 cell lines was detected by qRT-PCR; **(D)** The binding sites of miR-199a-3p and CDK1 or CCNB1; **(E)** RIP assay was used to detect whether miR-199a-3p could bind with CDK1 and CCNB1; **(F)** Dual luciferase reporter gene assay was used to verify the targeted binding relationship between miR-199a-3p and CDK1/CCNB1; **(G)** The expressions of miR-199a-3p, CDK1 and CCNB1 in HepG2 cells were detected by qRT-PCR; **(H)** Protein expressions of CDK1 and CCNB1 in HepG2 cells was detected by WB. *represents $P < 0.05$.

activity of miR-199a-3p mimic was significantly decreased in the co-transfection group with CDK1-WT or CCNB1-WT, indicating that miR-199a-3p could target CDK1 and CCNB1, respectively. In addition, HepG2 cell line was transfected into NC mimic group and miR-199a-3p mimic group. Expressions of miR-199a-3p, CDK1 and CCNB1 were detected by qRT-PCR (**Figure 4G**), and protein expressions of CDK1 and CCNB1 were detected by WB (**Figure 4H**). Overexpression

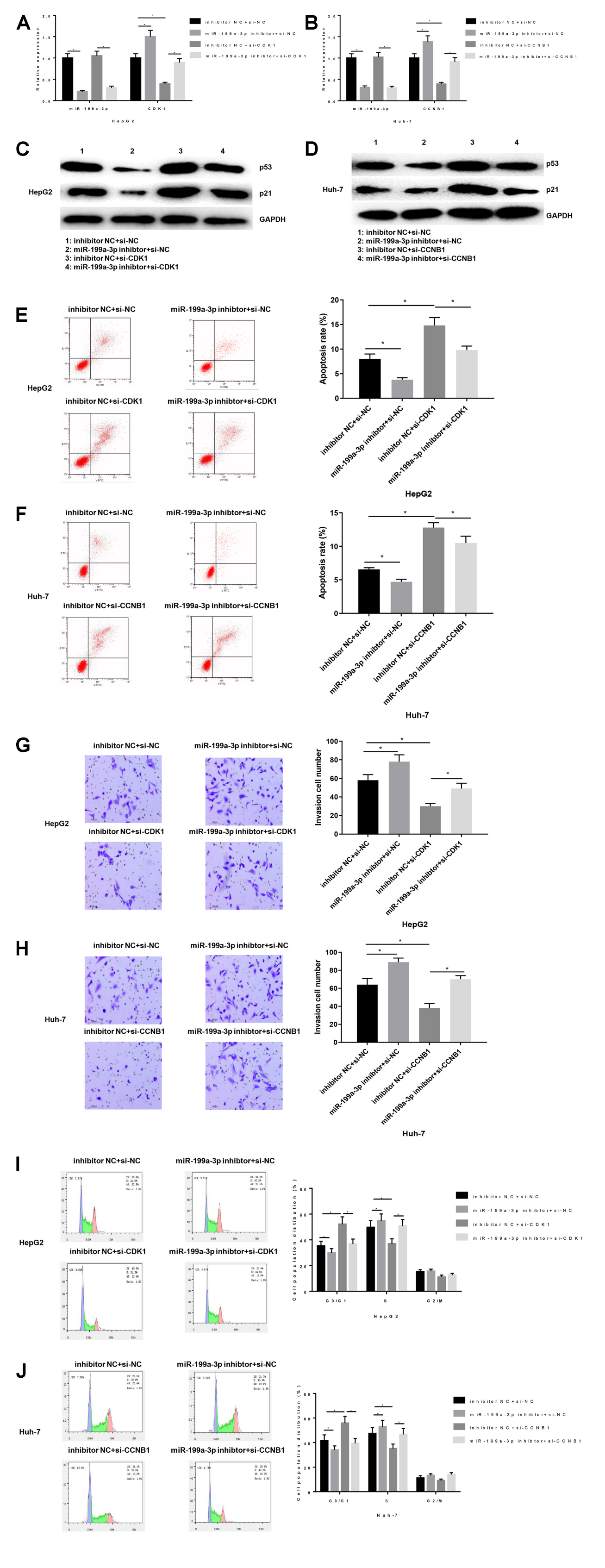


FIGURE 5 | Continued

FIGURE 5 | miR-199a-3p affects the apoptosis, invasion and cell cycle of HCC through CDK1 and CCNB1. The mRNA and protein expressions of miR-199a-3p, CNK1, CCNB1, p53 and p21 in each group were detected by qRT-PCR and WB, respectively. Cells development in each group was detected by FCM and Transwell assays (100×). **(A)** MRNA expressions of miR-199a-3p and CDK1 in HepG2 cells; **(B)** MRNA expressions of miR-199a-3p and CCNB1 in Huh-7 cells; **(C)** The protein expressions of p53 and p21 in HepG2 cells; **(D)** The protein expressions of p53 and p21 in Huh-7 cells; **(E,F)** Cell apoptosis; **(G,H)** Cell invasion; **(I,J)** Cell cycle. *represents $P < 0.05$. Representative of three independent experiments.

of miR-199a-3p resulted in down-regulation of mRNA and protein expression levels of CDK1 and CCNB1 ($P < 0.05$). These results proved that miR-199a-3p inhibit CCNB1 and CDK1 expressions.

miR-199a-3p Affects the Apoptosis, Invasion and Cell Cycle of HCC Through Targeting CDK1 and CCNB1

Next, in order to explore the effect of miR-199a-3p on HCC by targeting CDK1 and CCNB1, HepG2 cells were transfected into inhibitor NC + si-NC, miR-199a-3p inhibitor + si-NC, inhibitor NC + si-CDK1, miR-199a-3p inhibitor + si-CDK1 groups, and Huh-7 cells were transfected into inhibitors NC + si-NC, miR-199a-3p inhibitor + si-NC, inhibitor NC + si- CCNB1, miR-199a-3p inhibitor + si-CCNB1 groups. QRT-PCR was used to detect the expressions of miR-199a-3p and CDK1 in HepG2 cells, as well as expressions of miR-199a-3p and CCNB1 in Huh-7 cells. As exhibited in **Figures 5A,B**. Silencing miR-199a-3p significantly up-regulated the expressions of CDK1 and CCNB1 ($P < 0.05$). WB was performed to examine the protein expressions of p53 and p21 in cells of each group (**Figures 5C,D**). The results revealed that silencing miR-199a-3p inhibited the protein expressions of p53 and p21, while simultaneously silencing CDK1/CCNB1 offset the inhibitory effect of silencing miR-199a-3p ($P < 0.05$).

Functionally, FCM was used to detect the apoptosis rate (**Figures 5E,F**), compared with the inhibitor NC + si-NC group, the miR-199a-3p inhibitor + si-NC group had a significantly lower apoptosis rate, while inhibitor NC + si-CDK1 and the inhibitor NC + si-CCNB1 group had a significantly increased apoptosis rate. The co-transfection groups of miR-199a-3p inhibitor with si-CDK1 or si-CCNB1 offset the effect of both on apoptosis. Similarly, Transwell was used to measure cell invasion (**Figures 5G,H**). The results showed that silencing miR-199-3p could promote the invasion of HCC cells while silencing CDK1/CCNB1 would counteract the promoting effect of miR-199-3p ($P < 0.05$). Finally, FCM was used to detect the cell cycle (**Figures 5I,J**). We found that the cell ratio in G0/G1 phase in the miR-199a-3p inhibitor group was significantly reduced, and the cell ratio in S phase was significantly increased. While silencing CDK1 or CCNB1 improved miR-199a-3p inhibitory effect on cell cycle arrest.

In conclusion, miR-199a-3p activated the p53 signaling pathway by targeted inhibiting the expressions of CDK1 and

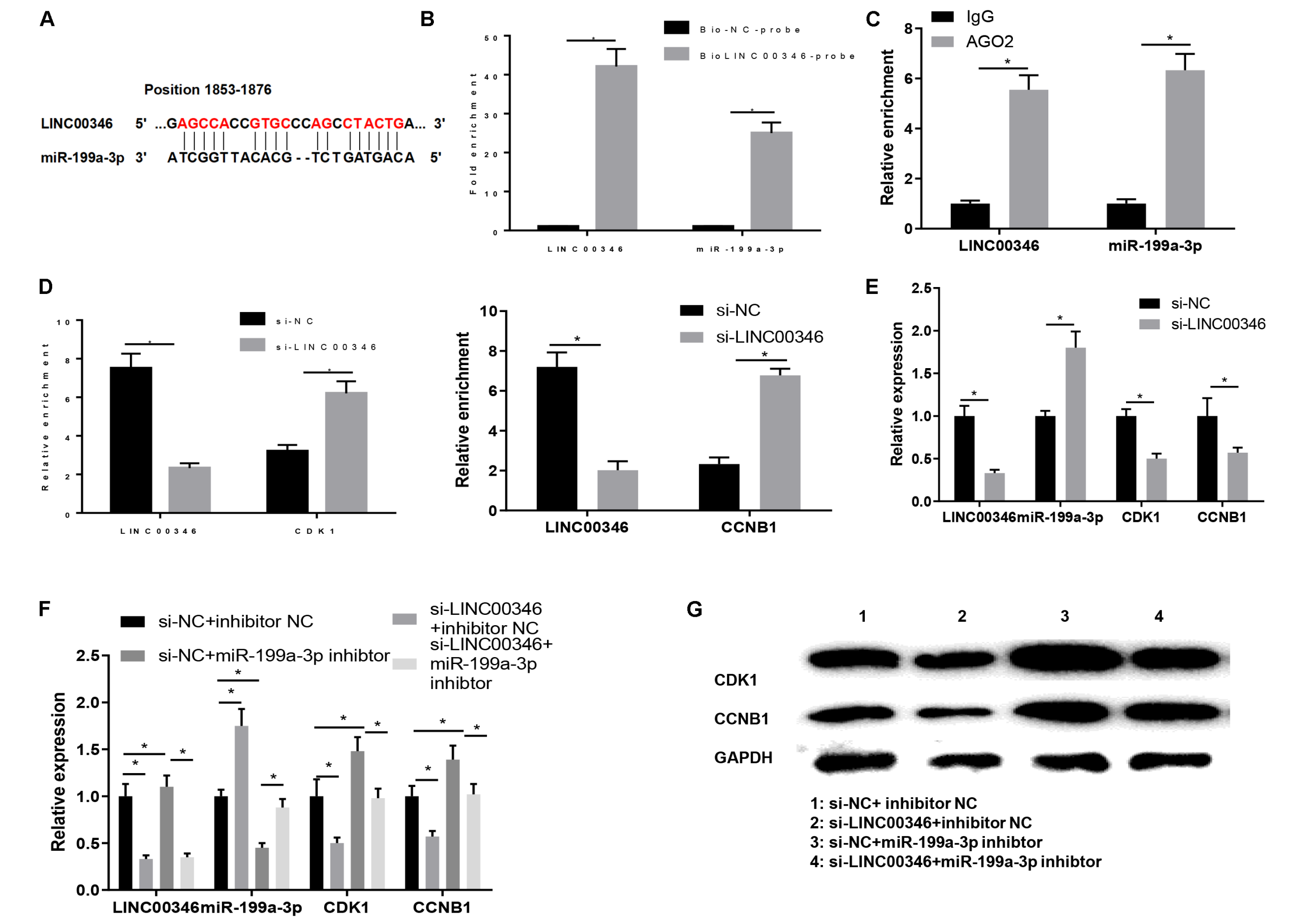


FIGURE 6 | LINC00346 regulates the expression of CDK1 and CCNB1 by sponging miR-199a-3p. **(A)** The binding sites of LINC00346 and miR-199a-3p; **(B)** RNA pull-down assay was used to verify the interaction between LINC00346 and miR-199a-3p, red letters indicate predicted binding sites of LINC00346 and mRNA-199a-3p **(C)** RIP was used to detect the binding of LINC00346 and miR-199a-3p in each group; **(D)** RIP was used to detect changes in LINC00346, CDK1 and CCNB1 binding to miR-199a-3p in each group; **(E)** Expression changes of LINC00346, miR-199a-3p, CDK1 and CCNB1 after LINC00346 silencing were detected by qRT-PCR; **(F)** The expressions of LINC00346, miR-199A-3p, CDK1 and CCNB1F in each group were detected by qRT-PCR and WB, respectively: The expressions of LINC00346, miR-199a-3p, CDK1 and CCNB1; **(G)** Protein expressions of CDK1 and CCNB1. *represents $P < 0.05$.

CCNB1, thus promoting the apoptosis of HepG2 or Huh-7 cells, inhibiting the invasion and regulating cell cycle.

LINC00346 Regulates the Expression of CDK1 and CCNB1 by Sponging miR-199a-3p

After confirming that miR-199a-3p regulated HCC cells by targeting CDK1/CCNB1, we then used the RNA22 database to find the corresponding lncRNA to miR-199a-3p and found that it was regulated by LINC00346. Moreover, studies have reported that LINC00346 was significantly negatively correlated with the survival of HCC patients, and could play a regulatory role through the mechanism of ceRNA (Zhang et al., 2015). These results suggested that LINC00346 may target and regulate the expression of CDK1 and CCNB1 by sponging miR-199a-3p, thereby affecting the p53 signaling pathway and ultimately participating in the development of HCC. For verification, the binding sites of LINC00346 and miR-199a-3p (**Figure 6A**) were predicted by bioinformatics website, and RNA pull-down assay was performed to verify the interaction between LINC00346 and miR-199a-3p in HepG2 cells. As shown in **Figure 6B**, both LINC00346 and miR-199a-3p were significantly enriched in the biotin-labeled LINC00346 drop-down conjugate ($P < 0.05$), indicating that LINC00346 could directly bind to miR-199a-3p. Then RIP was conducted to detect the binding of LINC00346 and miR-199a-3p. As displayed in **Figure 6C**, compared with IgG antibody, LINC00346 and miR-199a-3p enriched in AGO2 antibody group were significantly increased ($P < 0.05$). After silencing LINC00346, the relative enrichment of LINC00346 was significantly decreased, while that of CDK1 and CCNB1 was significantly increased (**Figure 6D**).

We then determined whether LINC00346 could modulate the expression of miR-199a-3p in HCC cells. HepG2 cells were transfected into si-NC and si-LINC00346 groups. QRT-PCR was used to detect the changes of LINC00346, miR-199a-3p, CDK1

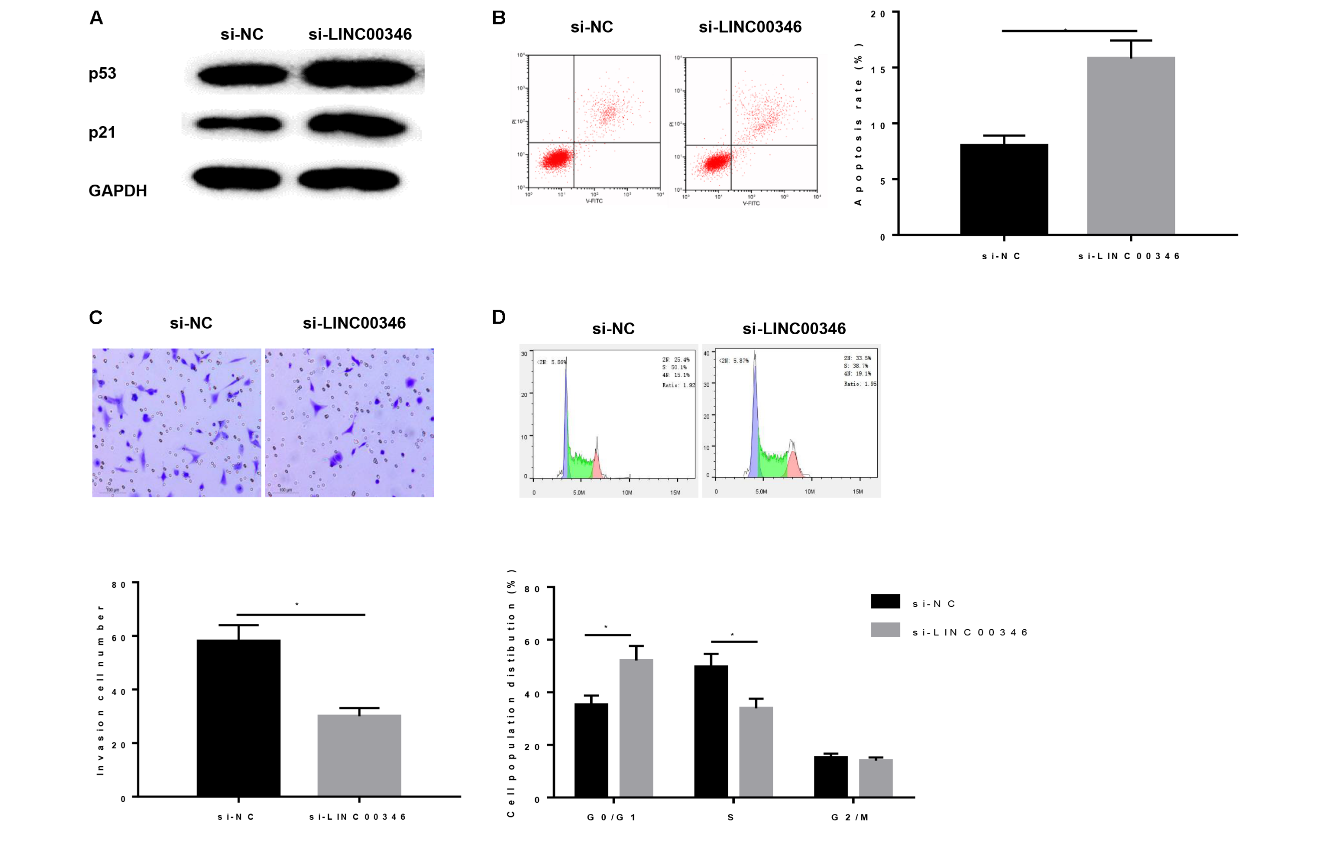


FIGURE 7 | Silencing LINC00346 affects the apoptosis, invasion and cell cycle of HCC. **(A)** Protein expressions of p53 and p21 in each group were detected by WB; Cells development in each group was detected by FCM and Transwell assays (100×). **(B)** Cell apoptosis; **(C)** Cell invasion; **(D)** Cell cycle. *represents $P < 0.05$.

and CCNB1 after silencing LINC00346. As shown in **Figure 6E**, silencing LINC00346 promoted the expression of miR-199a-3p and decreased the expression of CDK1 and CCNB1 ($P < 0.05$). Finally, HepG2 cells were transfected into si-NC + inhibitor NC, si-LINC00346 + inhibitor NC, si-NC + miR-199a-3p inhibitor, si-LINC00346 + miR-199a-3p inhibitor groups and the expressions of LINC00346, miR-199a-3p, CDK1, and CCNB1 in each group were detected (**Figure 6F**). WB was used to detect the protein expressions of CDK1 and CCNB1 in each group (**Figure 6G**). The results revealed that the mRNA and protein expressions of CDK1 and CCNB1 were significantly down-regulated after silencing LINC00346, and those were significantly up-regulated after silencing miR-199a-3p. Silencing LINC00346 and miR-199a-3p simultaneously canceled out the effects of silencing CDK1 and CCNB1 expression ($P < 0.05$). In conclusion, LINC00346 promoted the expression of CDK1 and CCNB1 by sponging miR-199a-3p.

LINC00346 Affects the Apoptosis, Invasion and Cell Cycle of HCC by Regulating the Expression of CDK1 and CCNB1

Finally, we explored the effect of LINC00346 on CDK1 and CCNB1 regulation on HCC. HepG2 cells were transfected and divided into si-NC group and si-LINC00346 group. Protein expressions of p53 and p21 were detected by WB (**Figure 7A**). The results showed that silencing LINC00346 promoted the protein expressions of p53 and p21 ($P < 0.05$). Functionally, the apoptotic rate was detected by FCM (**Figure 7B**) and observed that the apoptosis rate in si-LINC00346 group was significantly increased ($P < 0.05$). Cell invasion was detected by Transwell (**Figure 7C**). The results exhibited that silencing LINC00346 decreased the invasion ability of HCC cells ($P < 0.05$). FCM was used to detect the cell cycle (**Figure 7D**), we found that the cell proportion of G0/G1 phases in the si-LINC00346 group increased significantly, and the cell proportion of S phase decreased significantly. The results showed that silencing LINC00346 promoted p53 signaling pathway, promoted apoptosis, inhibited invasion, and blocked cells in G0/G1 phase.

HepG2 cells after transfection were then grouped into oe-NC + si-NC, oe-LINC00346 + si-NC, oe-LINC00346 + si-CDK1, and oe-LINC00346 + si-CCNB1. The expressions of LINC00346, CDK1 and CCNB1 in each group were detected by qRT-PCR (**Figure 8A**). The results indicated that the expressions of LINC00346, CDK1 and CCNB1 in the oe-LINC00346 + si-NC group were significantly up-regulated compared with those in the oe-NC + si-NC group ($P < 0.05$). Compared with oe-LINC00346 + si-NC group, the expression of CDK1 in oe-LINC00346 + si-CDK1 group and CCNB1 in

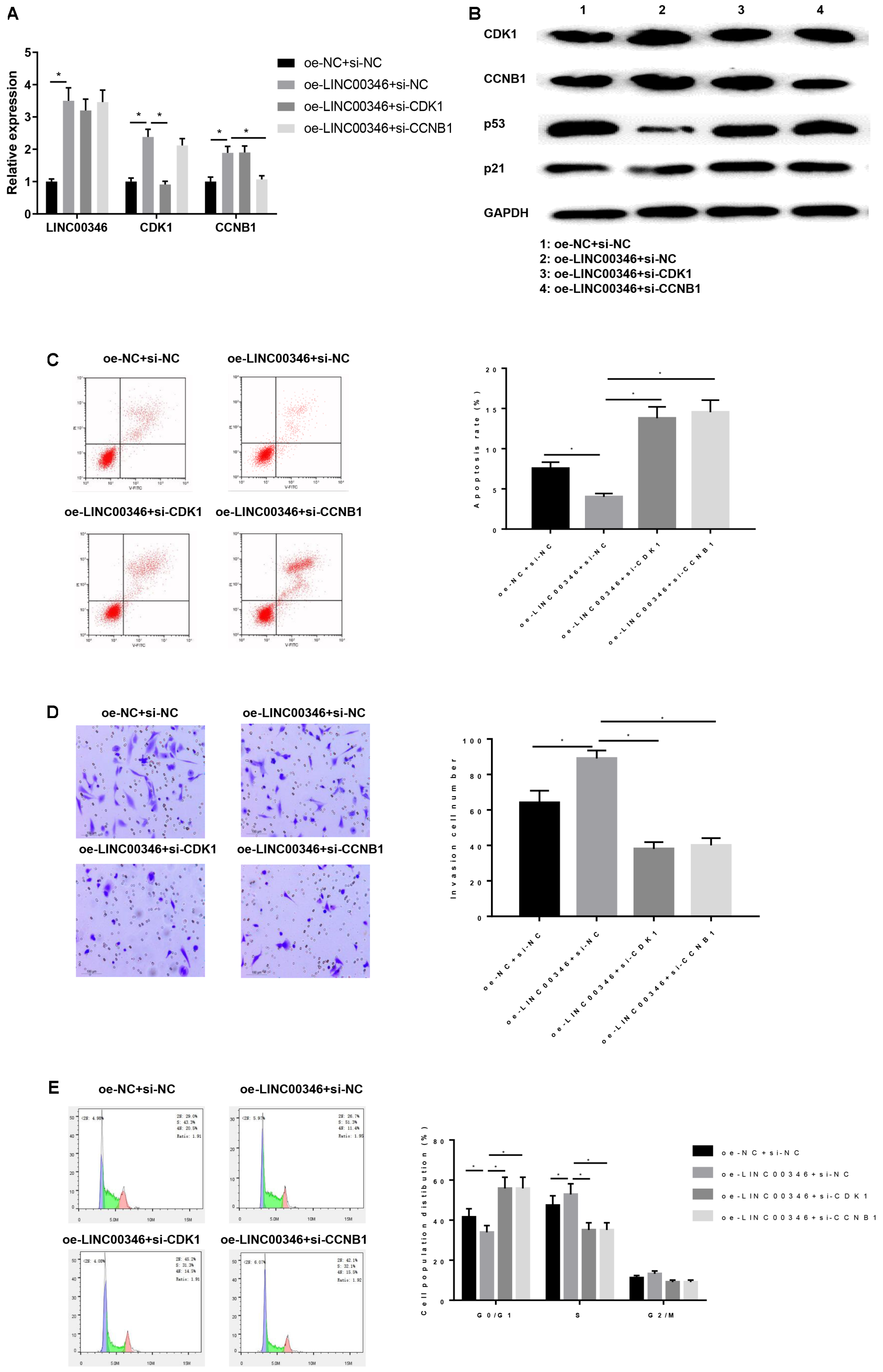


FIGURE 8 | LINC00346 affects the apoptosis, invasion and cell cycle of HCC by regulating the expression of CDK1 and CCNB1. The mRNA and protein expressions of LINC00346, CNK1, CCNB1, p53 and p21 in each group were detected by qRT-PCR and WB, respectively. Cells development in each group was detected by FCM and Transwell assays (100×). **(A)** The expressions of LINC00346, CDK1 and CCNB1; **(B)** Protein expressions of CDK1, CCNB1, p53 and p21; **(C)** Cell apoptosis; **(D)** Cell invasion; **(E)** Cell cycle. *represents $P < 0.05$.

oe-LINC00346 + si-CCNB1 group were significantly down-regulated ($P < 0.05$). WB was performed to detect the protein expressions of CDK1, CCNB1, p53 and p21 in each group (**Figure 8B**). The results revealed that overexpression of LINC00346 promoted the expressions of CDK1 and CCNB1 and inhibited the expressions of p53 and p21 ($P < 0.05$). FCM was also used to detect the apoptosis rate (**Figure 8C**). Overexpression of LINC00346 inhibited cell apoptosis, while silencing CDK1 and CCNB1 reversed the inhibitory effect of LINC00346 overexpression on apoptosis. Transwell was conducted to detect cell invasion (**Figure 8D**). Overexpression of LINC00346 promoted cell invasion, while the co-transfection group of over-expressing LINC00346 and silenced CDK1 or CCNB1 reversed the promotion effect of over-expressing LINC00346. FCM was used to measure cell cycle (**Figure 8E**). We found that compared with the oe-NC + si-NC group, the cell ratio of G0/G1 phase in the oe-LINC00346 group was significantly reduced while the cell ratio in S phase was significantly increased, and silencing CDK1 or CCNB1 improved the inhibitory effect of overexpression of LINC00346 on cell cycle arrest. The results above showed that overexpression of LINC00346 promoted the expressions of LINC00346, CDK1, and CCNB1 and invasion, inhibited the p53 pathway and apoptosis, and blocked cells in the S phase. These results indicated that LINC00346 could block the p53 signaling pathway by promoting the expressions of CDK1/CCNB1, thereby promoting the invasion of cancer cells, inhibiting apoptosis and regulating cell cycle.

DISCUSSION

Because of the high recurrence and metastasis rates, the overall survival of HCC patients remains low, and the study of molecular therapies for HCC has been a hot topic (Setshedi et al., 2018). This study demonstrated that CDK1 and CCNB1 were highly expressed in HCC tissues and cells through bioinformatics analysis combined with cell experiments, which was consistent with previous results (Wu et al., 2018; Gu et al., 2019). Then, we analyzed the effects of CDK1 and CCNB1 on the biological behavior of HCC, and found that CDK1 and CCNB1, as two oncogenes, could inhibit the apoptosis of HCC cells and promote cell invasion. These two genes also play a carcinogenic role in other cancers. Yang et al. put forward that the expression of CDK1 as an oncogene would increase with the progressive deterioration of epithelial ovarian cancer (Yang et al., 2016). Ding et al. (2014) demonstrated that CCNB1 can act as a biomarker of ER + breast cancer and play an oncogenic role in the occurrence of ER + breast cancer (Ding et al., 2014). These results indicated that CDK1/CCNB1 played an important role as an oncogene in a variety of cancers including HCC.

In exploring the specific molecular mechanism of CDK1/CCNB1 in regulating cancer, Zhang et al. revealed that CCNB1 could affect the cell cycle and apoptosis of pancreatic cancer cells by regulating p53 signaling pathway (Zhang H. et al., 2018). Qin G et al. reported that the p53 signaling pathway may be regulated by multiple genes to affect the development of liver cells (Qin et al., 2019). In addition, KEGG pathway enrichment analysis (**Figure 1B**) showed that multiple DEGs, including CDK1/CCNB1, were enriched in the p53 signaling pathway. Therefore, we speculated that CNK1/CCNB1 might regulate the occurrence of HCC by affecting p53 signaling pathway. We had conducted several experiments to verify this hypothesis, and the results showed that silencing CDK1/CCNB1 could promote the protein expressions of p53 and p21, thus promoting the apoptosis and inhibiting the invasion of HCC cells. These results all suggested that CDK1 and CCNB1 affected the apoptosis, invasion and cell cycle of HCC by regulating p53 signaling pathway.

In order to further explore the genes related to CDK1/CCNB1, we proved the signal axis of LINC00346-miR-199a-3p-CDK1/CCNB1 through bioinformatics analysis and molecular experiments and found that miR-199a-3p could bind to the 3′UTR of CDK1 and CCNB1. Meanwhile, the targeted relationship was verified by the dual luciferase reporter gene assay, and the results of WB confirmed that miR-199a-3p inhibited the expressions of CDK1 and CCNB1. The results are consistent with the results of previous studies. Ma et al. (2019) reported that miR-199a-3p is poorly expressed in HCC cells and HEIH silence suppressed the activation of mTOR signaling via upregulating miR-199a-3p (Ma et al., 2019). Ren et al. (2016) reported that miR-199a-3p could inhibit the proliferation of HCC cells by targeted down-regulating YAP1 expression (Ren et al., 2016). Our study on the regulatory effect of miR-199a-3p on HCC also found that miR-199a-3p could activate the p53 signaling pathway by targeting the expressions of CDK1/CCNB1, thereby inhibiting the development of HCC.

In the study of LINC00346 on HCC, we found that LINC00346 significantly promoted the invasion and inhibited cell apoptosis of HCC through the competitive adsorption of miR-199a-3p to promote the expressions of CDK1/CCNB1. In recent years, lncRNAs analysis and functional assays for various types of cancer have provided increasing evidence supporting the critical role of lncRNAs in HCC tumor growth and progression, such as HOTAIR (Hu et al., 2018), MALAT1 (Tao et al., 2018) and TUG1 (Sun et al., 2018). Previous studies have found that LINC00346 is up-regulated in HCC tissues (Zhang et al., 2015). As a gene regulator, lncRNA regulates gene expression through a variety of mechanisms, one of which is sponging miRNA to up-regulate or down-regulate miRNA expression. Previous studies have reported that the LINC00346-miR-10a-5p-CDK1 axis may be an important mechanism for HBV-related HCC, and genes in this ceRNA axis may be potential prognostic biomarkers and therapeutic targets (Li et al., 2019). The effects of the ceRNA axis of LINC00346-miR-199a-3p-CDK1/CCNB1 on cell invasion, apoptosis and cell cycle of HCC was demonstrated in this study, which further proved the role of LINC00346 as an oncogene in HCC, and the mechanism of HCC was further investigated.

CONCLUSION

In conclusion, LINC00346 has a positive regulatory effect on HCC. LINC00346 can regulate CDK1/CCNB1 to inhibit apoptosis, promote cell invasion and regulate cell cycle of HCC by targeting miR-199a-3p, while LINC00346-miR-199a-3p-CDK1/CCNB1 signal axis can regulate p53 signaling pathway. This result provides a deeper understanding of LINC00346's role in HCC, and lays the foundation for searching new targeted therapies for HCC.

DATA AVAILABILITY STATEMENT

The data used to support the findings of this study are included within the article. The data and materials in the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

JJ contributed to the study design. HX conducted the literature search. WL and XX wrote the article. HL and FW revised the article.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00054/full#supplementary-material

TABLE S1 | 230 differential genes screened by differential analysis of GES62232 dataset in GEO database.

REFERENCES

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Brown, J. D., Feldman, Z. B., Doherty, S. P., Reyes, J. M., Rahl, P. B., Lin, C. Y., et al. (2018). BET bromodomain proteins regulate enhancer function during adipogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2144–2149. doi: 10.1073/pnas.1711155115

Chen, X., Fan, S., and Song, E. (2016). Noncoding RNAs: new players in cancers. *Adv. Exp. Med. Biol.* 927, 1–47. doi: 10.1007/978-981-10-1498-7_1

Da Pozzo, E., La Pietra, V., Cosimelli, B., Da Settimo, F., Giacomelli, C., Marinelli, L., et al. (2014). p53 functional inhibitors behaving like pifithrin-β counteract the Alzheimer peptide non-β-amyloid component effects in human SH-SY5Y cells. *ACS Chem. Neurosci.* 5, 390–399. doi: 10.1021/cn4002208

Danhier, F., Ucakar, B., Magotteaux, N., Brewster, M. E., and Préat, V. (2010). Active and passive tumor targeting of a novel poorly soluble cyclin dependent kinase inhibitor, JNJ-7706621. *Int. J. Pharm.* 392, 20–28. doi: 10.1016/j.ijpharm.2010.03.018

Ding, K., Li, W., Zou, Z., Zou, X., and Wang, C. (2014). CCNB1 is a prognostic biomarker for ER+ breast cancer. *Med. Hypotheses* 83, 359–364. doi: 10.1016/j.mehy.2014.06.013

Gu, J., Liu, X., Li, J., and He, Y. (2019). MicroRNA-144 inhibits cell proliferation, migration and invasion in human hepatocellular carcinoma by targeting CCNB1. *Cancer Cell Int.* 19:15. doi: 10.1186/s12935-019-0729-x

Hu, J., Wang, Z., Shan, Y., Pan, Y., Ma, J., and Jia, L. (2018). Long non-coding RNA HOTAIR promotes osteoarthritis progression via miR-17-5p/FUT2/β-catenin axis. *Cell Death Disease* 9:711. doi: 10.1038/s41419-018-0746-z

Jin, P., Hardy, S., and Morgan, D. O. (1998). Nuclear localization of cyclin B1 controls mitotic entry after DNA damage. *J. Cell Biol.* 141, 875–885. doi: 10.1083/jcb.141.4.875

Kreis, N. N., Sanhaji, M., Krämer, A., Sommer, K., Rödel, F., Strebhardt, K., et al. (2010). Restoration of the tumor suppressor p53 by downregulating cyclin B1 in human papillomavirus 16/18-infected cancer cells. *Oncogene* 29, 5591–5603. doi: 10.1038/onc.2010.290

Li, H., Zhao, X., Li, C., Sheng, C., and Bai, Z. (2019). Integrated analysis of lncRNA-associated ceRNA network reveals potential biomarkers for the prognosis of hepatitis B virus-related hepatocellular carcinoma. *Cancer Manag. Res.* 11, 877–897. doi: 10.2147/CMAR.S186561

Li, L., and Chang, H. Y. (2014). Physiological roles of long noncoding RNAs: insight from knockout mice. *Trends Cell Biol.* 24, 594–602. doi: 10.1016/j.tcb.2014.06.003

Ma, Y., Cao, D., Li, G., Hu, J., Liu, X., and Liu, J. (2019). Silence of lncRNA HEIH suppressed liver cancer cell growth and metastasis through miR-199a-3p/mTOR axis. *J. Cell. Biochem.* 120, 17757–17766. doi: 10.1002/jcb.29041

Malumbres, M., and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nat. Rev. Cancer* 9, 153–166. doi: 10.1038/nrc2602

Meng, X., Franklin, D. A., Dong, J., and Zhang, Y. (2014). MDM2-p53 pathway in hepatocellular carcinoma. *Cancer Res.* 74, 7161–7167. doi: 10.1158/0008-5472.CAN-14-1446

Niméus-Malmström, E., Koliadi, A., Ahlin, C., Holmqvist, M., Holmberg, L., Amini, R. M., et al. (2010). Cyclin B1 is a prognostic proliferation marker with a high reproducibility in a population-based lymph node negative breast cancer cohort. *Int. J. Cancer* 127, 961–967. doi: 10.1002/ijc.25091

Porter, L. A., Cukier, I. H., and Lee, J. M. (2003). Nuclear localization of cyclin B1 regulates DNA damage-induced apoptosis. *Blood* 101, 1928–1933. doi: 10.1182/blood-2002-04-1103

Qin, G., Tu, X., Li, H., Cao, P., Chen, X., Song, J., et al. (2019). Long noncoding RNA p53-Stabilizing and activating RNA promotes p53 signaling by inhibiting heterogeneous nuclear ribonucleoprotein K deSUMOylation and suppresses Hepatocellular Carcinoma. *Hepatology* 71, 112–129. doi: 10.1002/hep.30793

Ren, K., Li, T., Zhang, W., Ren, J., Li, Z., and Wu, G. (2016). miR-199a-3p inhibits cell proliferation and induces apoptosis by targeting YAP1, suppressing Jagged1-Notch signaling in human hepatocellular carcinoma. *J. Biomed. Sci.* 23:79.

Santamaría, D., Barrière, C., Cerqueira, A., Hunt, S., Tardy, C., Newton, K., et al. (2007). Cdk1 is sufficient to drive the mammalian cell cycle. *Nature* 448, 811–815. doi: 10.1038/nature06046

Setshedi, M., Andersson, M., Kgatle, M. M., and Roberts, L. (2018). Molecular and cellular oncogenic mechanisms in hepatocellular carcinoma. *South Afr. Med. J.* 108, 41–46. doi: 10.7196/SAMJ.2018.v108i8b.13500

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

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer statistics, 2017. *CA* 67, 7–30. doi: 10.3322/caac.21387

Smal, M. A., Baldo, B. A., and Redmond, J. W. (1989). Production of antibodies to platelet activating factor. *Mol. Immunol.* 26, 711–719. doi: 10.1016/0161-5890(89)90030-8

Sun, J., Hu, J., Wang, G., Yang, Z., Zhao, C., Zhang, X., et al. (2018). LncRNA TUG1 promoted KIAA1199 expression via miR-600 to accelerate cell metastasis and epithelial-mesenchymal transition in colorectal cancer. *J. Exp. Clin. Cancer Res.* 37, 106–106. doi: 10.1186/s13046-018-0771-x

Sung, W.-W., Lin, Y.-M., Wu, P.-R., Yen, H.-H., Lai, H.-W., Su, T.-C., et al. (2014). High nuclear/cytoplasmic ratio of Cdk1 expression predicts poor prognosis in colorectal cancer patients. *BMC Cancer* 14:951. doi: 10.1186/1471-2407-14-951

Tao, F., Tian, X., Ruan, S., Shen, M., and Zhang, Z. (2018). miR-211 sponges lncRNA MALAT1 to suppress tumor growth and progression through inhibiting PHF19 in ovarian carcinoma. *FASEB J.* doi: 10.1096/fj.201800495RR [Epub ahead of print],

Ulahannan, S. V., Duffy, A. G., McNeel, T. S., Kish, J. K., Dickie, L. A., Rahma, O. E., et al. (2014). Earlier presentation and application of curative treatments in hepatocellular carcinoma. *Hepatology* 60, 1637–1644. doi: 10.1002/hep.27288

Wang, Y., Liu, Z., Yao, B., Li, Q., Wang, L., Wang, C., et al. (2017). Long non-coding RNA CASC2 suppresses epithelial-mesenchymal transition of hepatocellular carcinoma cells through CASC2/miR-367/FBXW7 axis. *Mol. Cancer* 16:123. doi: 10.1186/s12943-017-0702-z

Willder, J. M., Heng, S. J., McCall, P., Adams, C. E., Tannahill, C., Fyffe, G., et al. (2013). Androgen receptor phosphorylation at serine 515 by Cdk1 predicts biochemical relapse in prostate cancer patients. *Br. J. Cancer* 108, 139–148. doi: 10.1038/bjc.2012.480

Wu, C. X., Wang, X. Q., Chok, S. H., Man, K., Tsang, S. H. Y., Chan, A. C. Y., et al. (2018). Blocking CDK1/PDK1/β-Catenin signaling by CDK1 inhibitor RO3306 increased the efficacy of sorafenib treatment by targeting cancer stem cells in a preclinical model of hepatocellular carcinoma. *Theranostics* 8, 3737–3750. doi: 10.7150/thno.25487

Yang, W., Cho, H., Shin, H.-Y., Chung, J.-Y., Kang, E. S., Lee, E.-J., et al. (2016). Accumulation of cytoplasmic Cdk1 is associated with cancer growth and survival rate in epithelial ovarian cancer. *Oncotarget* 7, 49481–49497. doi: 10.18632/oncotarget.10373

Ye, T., Ding, W., Wang, N., Huang, H., Pan, Y., and Wei, A. (2017). Long noncoding RNA linc00346 promotes the malignant phenotypes of bladder cancer. *Biochem. Biophys. Res. Commun.* 491, 79–84. doi: 10.1016/j.bbrc.2017.07.045

Yoshida, T., Tanaka, S., Mogi, A., Shitara, Y., and Kuwano, H. (2004). The clinical significance of Cyclin B1 and Wee1 expression in non-small-cell lung cancer. *Ann. Oncol.* 15, 252–256. doi: 10.1093/annonc/mdh073

Zhang, B., Li, C., and Sun, Z. (2018). Long non-coding RNA LINC00346, LINC00578, LINC00673, LINC00671, LINC00261, and SNHG9 are novel prognostic markers for pancreatic cancer. *Am. J. Transl. Res.* 10, 2648–2658.

Zhang, H., Zhang, X., Li, X., Meng, W.-B., Bai, Z.-T., Rui, S.-Z., et al. (2018). Effect of CCNB1 silencing on cell cycle, senescence, and apoptosis through the p53 signaling pathway in pancreatic cancer. *J. Cell. Physiol.* 234, 619–631. doi: 10.1002/jcp.26816

Zhang, J., Fan, D., Jian, Z., Chen, G. G., and Lai, P. B. S. (2015). Cancer specific long noncoding RNAs show differential expression patterns and competing endogenous RNA potential in hepatocellular carcinoma. *PLoS One* 10:e0141042. doi: 10.1371/journal.pone.0141042

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

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 20 February 2020
doi: 10.3389/fbioe.2020.00075

MiR-29c-3p Suppresses the Migration, Invasion and Cell Cycle in Esophageal Carcinoma via CCNA2/p53 Axis

*Haiyong Wang, Linhai Fu, Desheng Wei, Bin Wang, Chu Zhang, Ting Zhu, Zhifeng Ma, Zhupeng Li, Yuanlin Wu and Guangmao Yu**

Department of Thoracic and Cardiovascular Surgery, Shaoxing People's Hospital (Shaoxing Hospital, Zhejiang University School of Medicine), Shaoxing, China

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Zhenda Xie,
Taizhou University, China
Xiaodong Liang,
Zhejiang Provincial People's Hospital, China

****Correspondence:***
Guangmao Yu
yu_guangmao@163.com

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

***Received:** 02 December 2019*
***Accepted:** 28 January 2020*
***Published:** 20 February 2020*

Citation:
Wang H, Fu L, Wei D, Wang B, Zhang C, Zhu T, Ma Z, Li Z, Wu Y and Yu G (2020) MiR-29c-3p Suppresses the Migration, Invasion and Cell Cycle in Esophageal Carcinoma via CCNA2/p53 Axis. Front. Bioeng. Biotechnol. 8:75. doi: 10.3389/fbioe.2020.00075

Objective: In the present study, we tried to describe the role of miR-29c-3p in esophageal carcinoma (EC) and the relationship of miR-29c-3p with CCNA2 as well as cell cycle, accordingly revealing the potential molecular mechanism across cell proliferation, migration and invasion.

Methods: Expression profiles of EC miRNAs and matched clinical data were accessed from TCGA database for differential and survival analyses. Bioinformatics databases were employed to predict the downstream targets of the potential miRNA, and enrichment analysis was performed on the miRNA and corresponding target gene using GSEA software. qRT-PCR was conducted to detect the expression levels of miR-29c-3p and CCNA2 mRNA in EC tissues and cells, and Western blot was performed for the examination of CCNA2, CDK1 and p53 protein levels. Subsequently, cells were harvested for MTT, Transwell as well as flow cytometry assays to examine cell viability, migration, invasion and cell cycle. Dual-luciferase reporter gene assay and RIP were carried out to further investigate and verify the targeted relationship between miR-29c-3p and CCNA2.

Results: MiR-29c-3p was shown to be significantly down-regulated in EC tissues and able to predict poor prognosis. CCNA2 was found to be a downstream target of miR-29c-3p and mainly enriched in cell cycle and p53 signaling pathway, whereas miR-29c-3p was remarkably activated in cell cycle. MiR-29c-3p overexpression inhibited cell proliferation, migration and invasion, as well as arrested cells in G0/G1 phase. As suggested by dual-luciferase reporter gene assay and RIP, CCNA2 was under the regulation of miR-29c-3p, and the negative correlation between the two genes was verified. Silencing CCNA2 could suppress cell proliferation, migration and invasion, as well as activate p53 pathway, even was seen to reverse the inhibitory effect of PFTβ on p53. Besides, in the presence of low miR-29c-3p, CCNA2 was up-regulated while p53 was simultaneously inhibited, resulting in the promotion of cell migration, invasion and cell cycle arrest.

Conclusion: MiR-29c-3p plays a regulatory role in EC tumorigenesis and development. MiR-29c-3p can target CCNA2 to mediate p53 signaling pathway, finally attributing to the inhibition of cell proliferation, migration and invasion, and making cells arrest in G0/G1 phase.

Keywords: miR-29c-3p, CCNA2, p53, esophageal carcinoma, migration, invasion, cell cycle

INTRODUCTION

Esophageal carcinoma (EC), a common gastrointestinal neoplasm, is the fifth cause of cancer-related death in China (Wang et al., 2012). EC can be classified as esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC) (Zaidi and Kelly, 2017), among which ESCC is the most common histopathological type with relative high morbidity in China (Kamangar et al., 2006; Bohanes et al., 2012). At present, the main curative option of EC is surgery with adjuvant radiotherapy and chemotherapy, but the overall prognosis remains poor (Gertler et al., 2011). Therefore, studying potential molecular mechanism is of great importance for the exploration of novel therapies.

MiRNAs is a large family, and some of the miRNAs show targeted relationship with miRNAs. Genome analysis suggests that miRNA-mediated genes account for nearly 30% of the total human genome, and their expressions are firmly associated with cancers (Calin et al., 2004; Lewis et al., 2005). Studies have found that the alteration of miRNAs expression can lead to the changes of oncogenes and tumor suppressor genes, thus affecting cell proliferation, migration, invasion and apoptosis in gastrointestinal neoplasms including ESCC (Harada et al., 2016). In addition, miRNAs have been observed to be differentially expressed in ESCC as reported by multiple microarray studies. Ogawa et al. found that 22 miRNAs were up-regulated in ESCC tissues relative to that in adjacent normal tissues, whereas 4 miRNAs were down-regulated (Fu et al., 2013). Fu et al. revealed that among the 43 differentially expressed miRNAs (DEmiRNAs) found in ESCC, 27 miRNAs were decreased and the rest were increased, of which miRNA-1 was significantly reduced and attributed to the inhibition of cell proliferation, clone, migration and invasion (Yao et al., 2015). Moreover, miR-34a was found to suppress cell migration and invasion in ESCC via targeting Yin Yang-1 (Nie et al., 2015), and miR-29b was shown to function on ESCC progression through targeting MMP-2 (Qi et al., 2015).

CCNA2 (cyclin A2) is a cyclin accumulated in G1 phase and plays a regulatory role in the transitional period of G1/S and G2/M (Krasnov et al., 2017). Published literature has reported that CCNA2 functions on various cancers, like colorectal cancer (Huang et al., 2017), liver cancer (Yang et al., 2016), breast cancer (Gao et al., 2014), cervical cancer (Wu et al., 2019), and EC (Ma, 2019). In the present study, we found that miR-29c-3p was remarkably down-regulated in EC cells. Then bioinformatics methods were performed to predict the targets of miR-29c-3p, and CCNA2 was selected for further investigation, in turn evaluating the potential of miR-29c-3p/CCNA2 axis as an effective therapy for EC.

MATERIALS AND METHODS

Bioinformatics Analysis

The miRNA and mRNA expression profiles of ESCA were downloaded from the TCGA database[1]. "edgeR" package was used to perform differential analysis, and $|\text{logFC}| > 1.5$ and $P\text{-adj} < 0.01$ were set as the threshold to screen out DEGs. Survival analysis was performed on DEmiRNAs to confirm the potential target miRNA. Four databases miRDB[2], mirDIP[3], starBase[4], and miRTarBase[5] were utilized to predict the targets of the miRNA, and Venn diagram was plotted to find the potential target genes. GSEA 4.0.1 software was applied to carry out enrichment analysis on the miR-29c-3p and its target gene CCNA2. According to the median expression level of CCNA2 and miR-29c-3p, EC tissue samples were divided into high ($n = 80$) and low ($n = 80$) expression groups. MSigDB[6] was applied to access "c2.cp.kegg.v7.0.symbols.gmt" data as reference.

Cell Culture

Human normal esophageal epithelial cell HET-1A (BNCC342346) and EC cell lines Eca-109 (BNCC337687), EC9706 (BNCC339892), KYSE150 (BNCC342590), and KYSE180 (BNCC351871) were purchased from BeNa Culture Collection (Beijing, China). All cells were grown in the Dulbecco's Modified Eagle Medium (DMEM; Gibco, United States) supplemented with 10% fetal bovine serum (FBS; Gibco, United States), streptomycin (100 mg/mL; Gibco, United States) and penicillin (100 units/mL; Gibco, United States), and maintained in 5% CO_2 at 37°C.

Sample Collection

A total of 30 cases of EC tissues and matched adjacent normal tissues (2 cm in margin) were collected in the Shaoxing People's Hospital from January 2018 to May 2019. All samples were obtained during the intraoperative period as well as firmly

[1] https://portal.gdc.cancer.gov/

[2] http://mirdb.org/miRDB/index.html

[3] http://ophid.utoronto.ca/mirDIP/index.jsp#r

[4] http://starbase.sysu.edu.cn/

[5] http://mirtarbase.mbc.nctu.edu.tw/php/index.php

[6] http://software.broadinstitute.org/gsea/msigdb/index.jsp

diagnosed by experienced pathologists, and none of the patients had received preoperative chemotherapy or radiotherapy. EC tissues separated were rapidly stored in RNA preservation solution. All procedures were performed with the approval of the Ethics Committee in the Shaoxing People's Hospital and informed consent was obtained from all patients before this study. Patients' clinicopathological characteristics like gender, age, histology identification results and tumor location were detailed in **Table 1**.

Cell Transfection

For preparation, cells were grown in complete medium for at least 24 h, and washed by phosphate buffered saline (PBS; pH 7.4) before transfection. Plasmids were all purchased from GenePharma (Shanghai, China), and transiently transfected into EC cells using Lipofectamine2000 (Thermo Fisher Scientific, Inc.), consequently forming six groups of NC mimic, miR-29c-3p mimic, NC inhibitor, miR-29c-3p inhibitor, si-NC, and si-CCNA2. Transfected cells were cultured in DMEM containing 5% CO_2 at 37°C for subsequent experiments.

RNA Extraction and qRT-PCR

TRIzol Reagent (Invitrogen) was utilized to isolate the total RNA and Superscript II reverse transcriptase (Invitrogen) was applied for cDNA synthesis via reverse transcription using 2 μg of total samples. qRT-PCR was conducted for the detection of miR-29c-3p and CCNA2 mRNA using the Applied Biosystems 7300 Real-Time PCR System (Applied Biosystems, United States), with U6 and GAPDH as internal control. All steps were followed the manufacturer's instructions. Primers used were designed by Sangon Biotech (Shanghai, China) as listed in **Table 2**. $2^{-\Delta\Delta Ct}$ method was used for the normalization of miR-29c-3p and CCNA2 mRNA expression levels.

Western Blot

After 48 h of transfection, cells were washed three times with cold PBS. Then proteins were extracted from cells on ice by whole cell lysate, and the concentration was assayed using BCA kit (Thermo Fisher Scientific, Waltham, MA, United States). 30 μg of the total extraction was separated through polyacrylamide gel electrophoresis (PAGE), and transferred onto the PVDF membranes (Amersham, United States) that were sequentially blocked in 5% skim milk for 1 h. Afterward, the membrane was incubated with primary rabbit polyclonal antibodies overnight at 4°C, followed by horseradish peroxidase (HRP)-labeled secondary antibody goat anti-rabbit IgG H&L (ab6721, 1:2000, Abcam, Cambridge, United Kingdom) at room temperature for 1 h. Primary antibodies contained CCNA2 (ab181591, 1:2000, Abcam, Cambridge, United Kingdom), p53 (ab32389, 1:1000, Abcam, Cambridge, United Kingdom) and GAPDH (ab9485, 1:2500, Abcam, Cambridge, United Kingdom). PBST (PBS containing 0.1% Tween-20) was utilized to wash the membranes three times following each step. Protein bands were visualized by chemiluminescence apparatus (GE, United States) and then photographed.

TABLE 1 | Basic information of patients and correlation with miR-29c-3p expression.

Characteristic	Total	miR-29c-3p expression	
		Low	High
Gender			
Male	13	4	9
Female	17	6	11
Age			
<50	12	3	9
≥50	18	7	11
Histology			
Adenocarcinoma	5	2	3
Squamous cell carcinoma	25	8	17
Tumor location			
Upper esophagus	2	0	2
Middle esophagus	23	8	15
Lower esophagus	5	2	3

TABLE 2 | Primer sequence.

Target gene	Primer (5′–3′)
CCNA2	F: CAGAAAACCATTGGTCCCTC
	R: CACTCACTGGCTTTTCATCTTC
miR-29c-3p	F: TAGCACCATTTGAAATCGGTTA
U6	F: CTCGCTTCGGCAGCACA
	R: AACGCTTCACGAATTTGCGT
GAPDH	F: GCACCGTCAAGGCTGAGAAC
	R: TGGTGAAGACGCCAGTGGA

MTT

Transfected cells were digested, resuspended and then plated into 96-well plates at a density of 5 × 10^3 cells/well. At 24, 48, 72, and 96 h, 10 μL of MTT regent (5 mg/mL) was added into per well and cells were continuously cultured at 37°C for 4 h. Thereafter, the supernatant was removed and the precipitate was solubilized in 200 μL of dimethyl sulfoxide (DMSO). Neo multimode reader (Thermo Fisher Scientific) was applied to measure the absorbance at 595 nm.

Transwell

Migration assay: Cells in logarithmic phase were placed in serum-free medium for 24 h. On the following day, cell suspension at a concentration of 2 × 10^4 cells/mL was prepared after digestion and centrifugation. 0.2 mL of suspension was added into the Transwell inserts, and 700 μL of pre-cooled DMEM containing 10% FBS was placed out of the inserts. After 24 h of incubation in 5% CO_2 at 37°C, unmigrated cells were wiped off with a cotton swab, while cells migrated out of the inserts were fixed by methanol for 30 min and stained in 0.1% crystal violet for 20 min. Images were captured under an inverted microscope, and five fields were randomly selected for cell count.

Invasion assay: Around 2 × 10^4 cells were added into the upper chambers pre-coated with Matrigel matrix (Corning, NY, United States), and DMEM supplemented with 10% FBS was placed in the lower chambers. Follow-up steps were similar with migration assay as detailed above.

Flow Cytometry

Petri dishes (6 cm) were utilized to culture transfected cells (2×10^5 cells/dish) until 80% in confluence. Subsequently, cells were digested with trypsin, washed with ice-cold PBS and collected. After being fixed by 75% methanol, cells were centrifuged and suspended in RNase A (Sigma-Aldrich) followed by stained using 500 μL PI solution (Sigma-Aldrich). Flow cytometer (Beckman-Coulter) was employed to analyze the cell cycle. The percentage of cells in G0/G1, S and G2/M was calculated, respectively, and compared in each group.

Dual-Luciferase Reporter Gene Assay

CCNA2 vectors bearing mutant and wild type 3′UTR (MUT- and WT-3′UTR) were cloned into pmiRGLO (Promega, Madison, WI, United States), forming the luciferase reporter plasmids WT-CCNA2 and MUT-CCNA2. Then the two plasmids were co-transfected with miR-29c-3p mimic or NC mimic into EC cell lines, respectively, with the Renilla luciferase expression vector pRL-TK (TaKaRa, Dalian, China) as the internal control. Cells were grown in DMEM containing 10% FBS. After 48 h, dual-luciferase detection kit (Promega, Madison, WI, United States) was utilized to examine the luciferase activity.

RNA-Binding Protein Immunoprecipitation Assay (RIP)

Magna RIP RNA-Binding Protein Immunoprecipitation Kit (Millipore, United States) was applied in this experiment following the manufacture's protocols. Cells were firstly lysed in RIP lysate buffer for 30 min, then the proteins obtained were incubated in RIP buffer containing magnetic beads. Ago2 antibody (ab32381, 1:50, Abcam, Cambridge, United Kingdom) was used to immune-precipitate miR-29c-3p protein complex taking normal rabbit antibody IgG (ab6712, 1:1000, Abcam, Cambridge, United Kingdom) as negative control. Consequently, protease K was utilized to purify the immunoprecipitation, and qRT-PCR was performed to detect the expression of CCNA2 mRNA.

Statistical Analysis

SPSS 22.0 statistical software was utilized to process all data. Measurement data were expressed as mean ± standard deviation

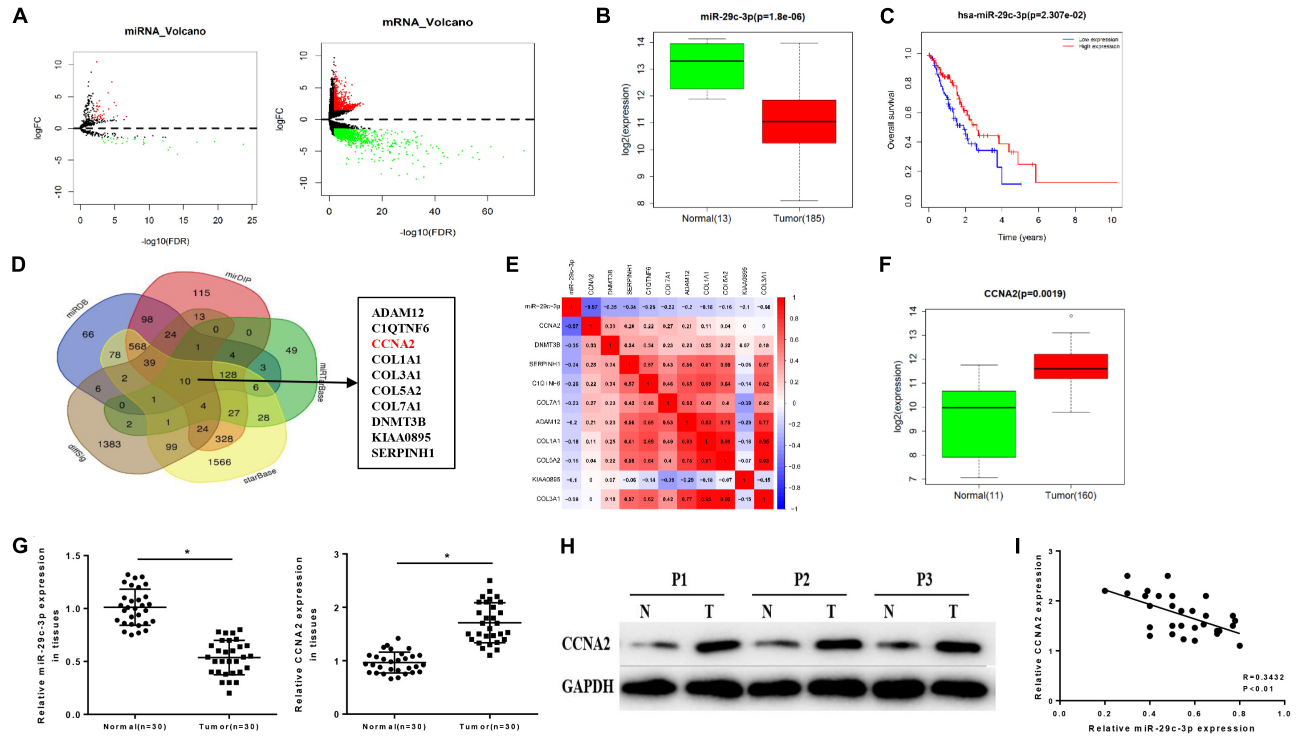


FIGURE 1 | MiR-29c-3p is decreased in EC tissues accompanied by low survival rate and associated with the increase of CCNA2. TCGA database was utilized to access expression data of miRNAs and mRNAs of ESCA, and **(A)** the results of differential analysis were plotted in Volcano plots, with red representing high expression and green representing low expression. In panel **(B)**, miR-29c-3p level in EC tissues were determined as shown in a box plot. **(C)** Survival analysis of miR-29c-3p in TCGA-ESCA dataset was performed, with the red line as high expression and blue line as low expression. In panel **(D)**, Venn diagram was made to find the candidate targets of miR-29c-3p, acquiring 10 DEmRNAs. In panel **(E)**, correlation analysis was conducted between miR-29c-3p and CCNA2 (–0.57) as plotted in a heat map. In panel **(F)**, CCNA2 expression in EC cells was examined. Clinical tissue samples were used to further explore the **(G)** expression of miR-29c-3p and CCNA2 mRNA in EC tissues by qRT-PCR, **(H)** the protein level of CCNA2 (P1, P2, P3 referred to three EC samples) via Western blot and **(I)** the correlation between miR-29c-3p and CCNA2. **P* < 0.05.

(SD). *t*-test and one-way ANOVA were employed to perform comparisons between two groups and among multiple groups, respectively. Kaplan–Meier survival analysis was conducted using the log-rank method. Pearson analysis was used to analyze the correlation between miR-29c-3p and CCNA2. Statistical significance was considered when $P < 0.05$.

RESULTS

MiR-29c-3p Is Decreased in EC Tissues Accompanied by Low Survival Rate and Associated With the Increase of CCNA2

Differential analysis was conducted on the gene expression profiles in TCGA-ESCA dataset using "edgeR" package, acquiring 62 DEmiRNAs and 1609 DEmRNAs (**Figure 1A**). Among them, miR-29c-3p showed significantly low expression in EC tissues (**Figure 1B**). Meanwhile, survival analysis suggested that low miR-29c-3p predicted poor prognosis, showing the survival time of patients with low miR-29c-3p shorter than those with high expression (**Figure 1C**). In addition, miRDB, mirDIP, starBase, and miRTarBase four databases were applied to predict candidate targets of miR-29c-3p and Venn diagram was plotted to find the potential target genes. As revealed in **Figure 1D**, 10 DEmRNAs were obtained, among which CCNA2 presented relative high correlation with miR-29c-3p (−0.57), as well as significantly increased expression in cancer cells relative to the normal control (**Figures 1E,F**). GSEA suggested that miR-29c-3p was highly enriched in cell cycle, and CCNA2 was mainly activated in cell cycle and p53 signaling pathway (**Supplementary Figures S1A–C**).

Moreover, we detected the expression levels of miR-29c-3p and CCNA2 mRNA in clinical samples of cancer tissues by qRT-PCR, and the results showed that miR-29c-3p in EC epithelial tissues was significantly lower than that in normal tissues, while CCNA2 was highly expressed (**Figure 1G**). Western blot indicated that CCNA2 showed remarkably up-regulated expression in the protein level as well (**Figure 1H**). Besides, correlation analysis was carried out, and there was a reverse association between miR-29c-3p and CCNA2 (**Figure 1I**). All above results elucidated that miR-29c-3p and CCNA2 could be served as potential biomarkers for EC diagnosis.

MiR-29c-3p Functions on the Migration, Invasion and Cell Cycle of EC Epithelial Cells

To further verify the role of miR-29c-3p in EC, qRT-PCR was firstly performed in 5 cell lines (including one normal

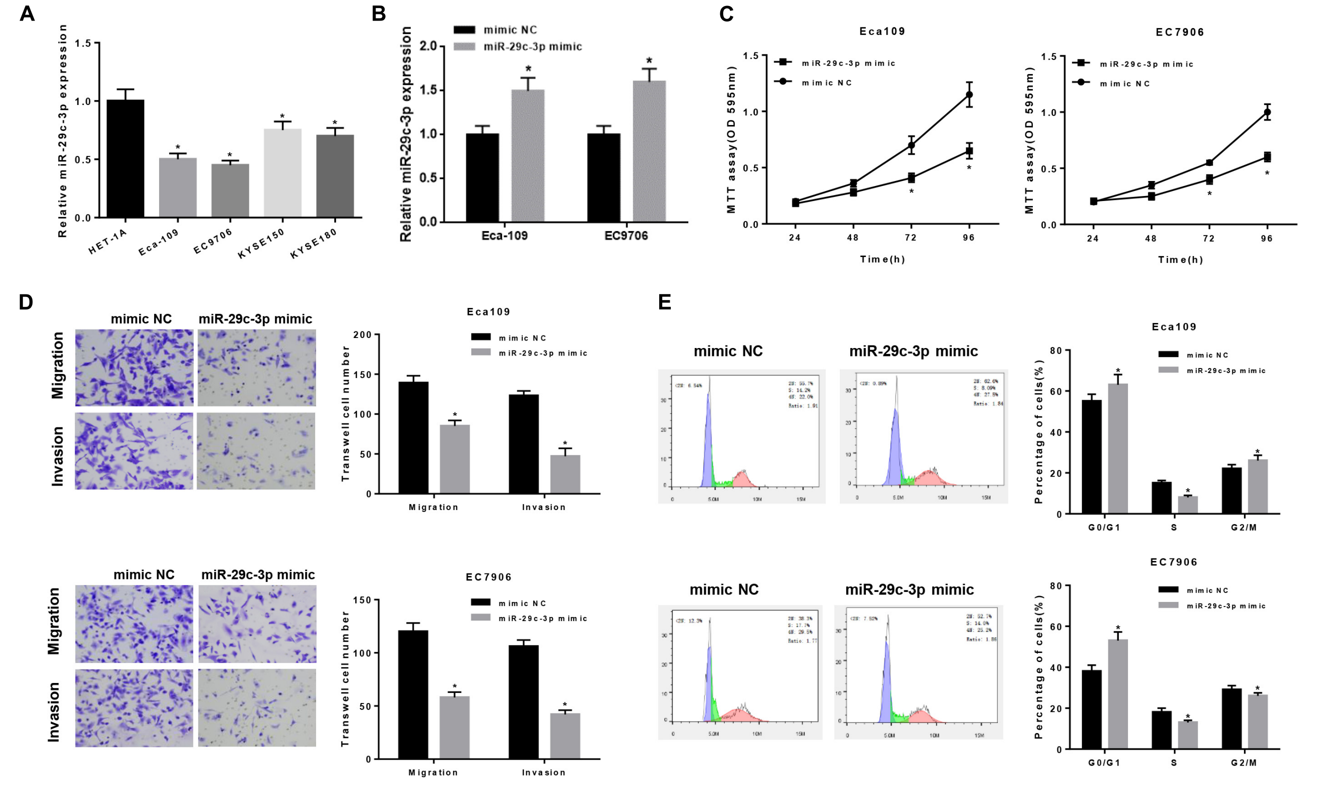


FIGURE 2 | MiR-29c-3p functions on the migration, invasion and cell cycle of EC epithelial cells. qRT-PCR was performed to test **(A)** the expression of miR-29c-3p in each cell line and **(B)** that in Eca109 and EC7906 cells after transfected with miR-29c-3p mimic. To further evaluate the role of miR-29c-3p in EC cells, with the presence of miR-29c-3p mimic and normal cells as control, **(C)** cell viability was detected by MTT, **(D)** migration and invasion ability was examined by Transwell and **(E)** cell cycle was assayed by flow cytometry. *$P < 0.05$.

cell line and four EC cell lines), suggesting the decreased expression of miR-29c-3p in EC cells (**Figure 2A**). For better observation, Eca109 and EC7906 were selected for subsequent analysis. As shown in **Figure 2B**, miR-29c-3p was significantly up-regulated in the two cells lines after transfection with miR-29c-3p mimic. It was also determined that cell viability was suppressed in the cells transfected with miR-29c-3p mimic as detected by MTT, indicting the weakening of cell proliferation (**Figure 2C**). Transwell assay revealed that EC cells had reduced migration and invasion abilities when miR-29c-3p was overexpressed (**Figure 2D**). Furthermore, flow cytometry showed the accumulation of miR-29c-3p mimic transfected cells in G0/G1phase (**Figure 2E**), indicating that miR-29c-3p could arrest cells from entering S phase, thus inhibiting cell proliferation.

MiR-29c-3p Targets CCNA2 and Inhibits Its Expression

As aforementioned, CCNA2 was a downstream target of miR-29c-3p (**Figure 3A**) and highly expressed in EC cells. To make a better understanding of their targeted relationship, dual-luciferase reporter gene assay was carried out via the construction of WT-CCNA2 and MUT-CCNA2. As shown in **Figure 3B**, luciferase activity was significantly decreased in the cells transfected with miR-29c-3p mimic and WT-CCNA2. Meanwhile, findings concluded by RIP suggested the remarkably up-regulated CCNA2 in miR-29c-3p mimic transfected cells (**Figure 3C**). Together, it could be elucidated that miR-29c-3p could specifically bind with CCNA2, leading to its reduced expression.

Besides, qRT-PCR and Western blot were conducted to further verify such relationship. As revealed by qRT-PCR plotted in **Figure 3D**, significantly reduced CCNA2 mRNA was observed when miR-29c-3p was overexpressed. Similar trend could be seen in Western blot, showing the down-regulation of CCNA2 protein level in cells transfected with miR-29c-3p mimic (**Figure 3E**). Collectively, there was a targeted relationship between miR-29c-3p and CCNA2.

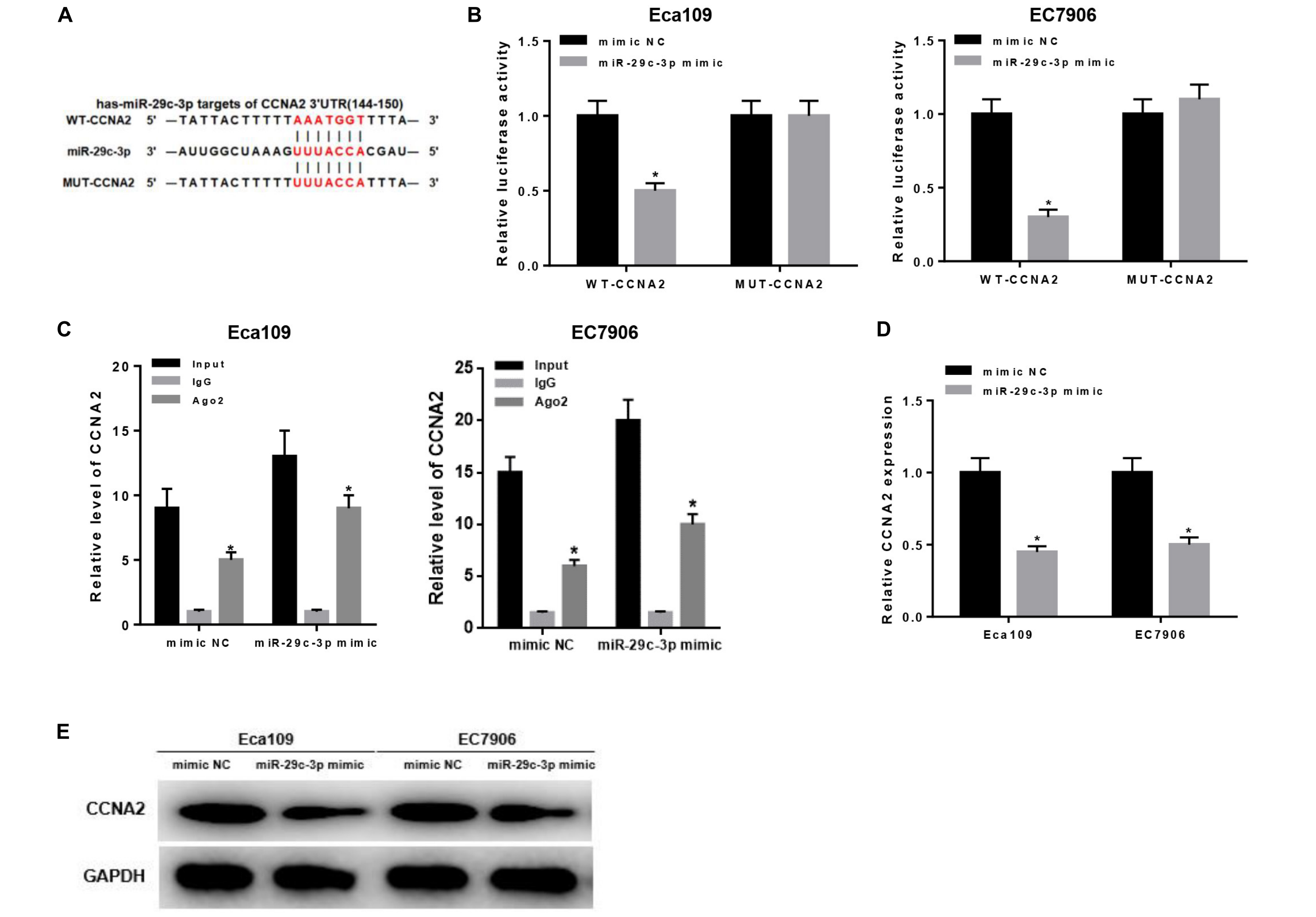


FIGURE 3 | MiR-29c-3p targets CCNA2 and inhibits its expression. Targeted binding sites of miR-29c-3p and CCNA2 were predicted before as shown in panel **(A)**. To investigate their targeted relationship, **(B)** dual-luciferase assay was performed to confirm their targeted binding, and **(C)** RIP was conducted to describe the effect of miR-29c-3p on CCNA2. Moreover, **(D,E)** qRT-PCR and Western blot were carried out to determine CCNA2 expression in mRNA and protein levels in miR-29c-3p mimic transfected cells, so as to further verify such relationship. *$P < 0.05$.

CCNA2 Silencing Regulates the Migration, Invasion and Cell Cycle in EC by Promoting p53 Signaling Pathway

Prior studies have found that CCNA2, a type of cyclin, is able to promote cell proliferation, migration and invasion (Ma, 2019; Shekhar et al., 2019; Tu et al., 2019), as well as associated with various biological pathways like p53 signaling pathway (Zhang et al., 2018; Doan et al., 2019). In order to investigate the underlying mechanism of CCNA2 on EC cells, cells were classified into four groups: si-NC + DMSO, si-CCNA2 + DMSO, si-NC + PFTβand si-CCNA2 + PFTβ groups [PFTβ, p53 inhibitor, HY-16702, MedChemExpress, 10 μM (Da Pozzo et al., 2014)]. qRT-PCR was firstly performed to test CCNA2 level in each group, finding that CCNA2 was markedly decreased in cells transfected with si-CCNA2 + DMSO (**Figure 4A**). Western blot suggested that CCNA2 silencing was with a concomitant increase in p53 level. si-NC + PFTβ group presented the lowest p53 expression, while in si-CCNA2 + PFTβ group, p53 level was restored, indicating that CCNA2 silencing could abrogate the inhibitory effect of PFTβ on p53 (**Figure 4B**). Subsequently, transfected cells were harvested for MTT and Transwell assays, revealing that the decrease of cell viability was more prominent with the reduction of CCNA2 (**Figure 4C**), as well as cell migration and invasion (**Figure 4D**). Furthermore, flow cytometry results plotted in **Figure 4E** showed the accumulation of cells in G0/G1 phase with the reduction of CCNA2. Taken together, CCNA2 silencing could repress EC epithelial cell activities.

MiR-29c-3p Mediates the Migration, Invasion and Cell Cycle in EC via CCNA2/p53 Axis

In order to uncover the miR-29c-3p-dependent mechanism in EC, inhibitor NC + si-NC, inhibitor NC + si-CCNA2, miR-29c-3p inhibitor + si-NC and miR-29c-3p inhibitor + si-CCNA2 were designed to transfect cells. As shown in **Figure 5A**, p53 expression was up-regulated when CCNA2 was silenced, whereas when miR-29c-3p was inhibited, CCNA2 was increased with a concomitant of p53 decrease. Besides, when miR-29c-3p and CCNA2 were repressed simultaneously, p53 expression

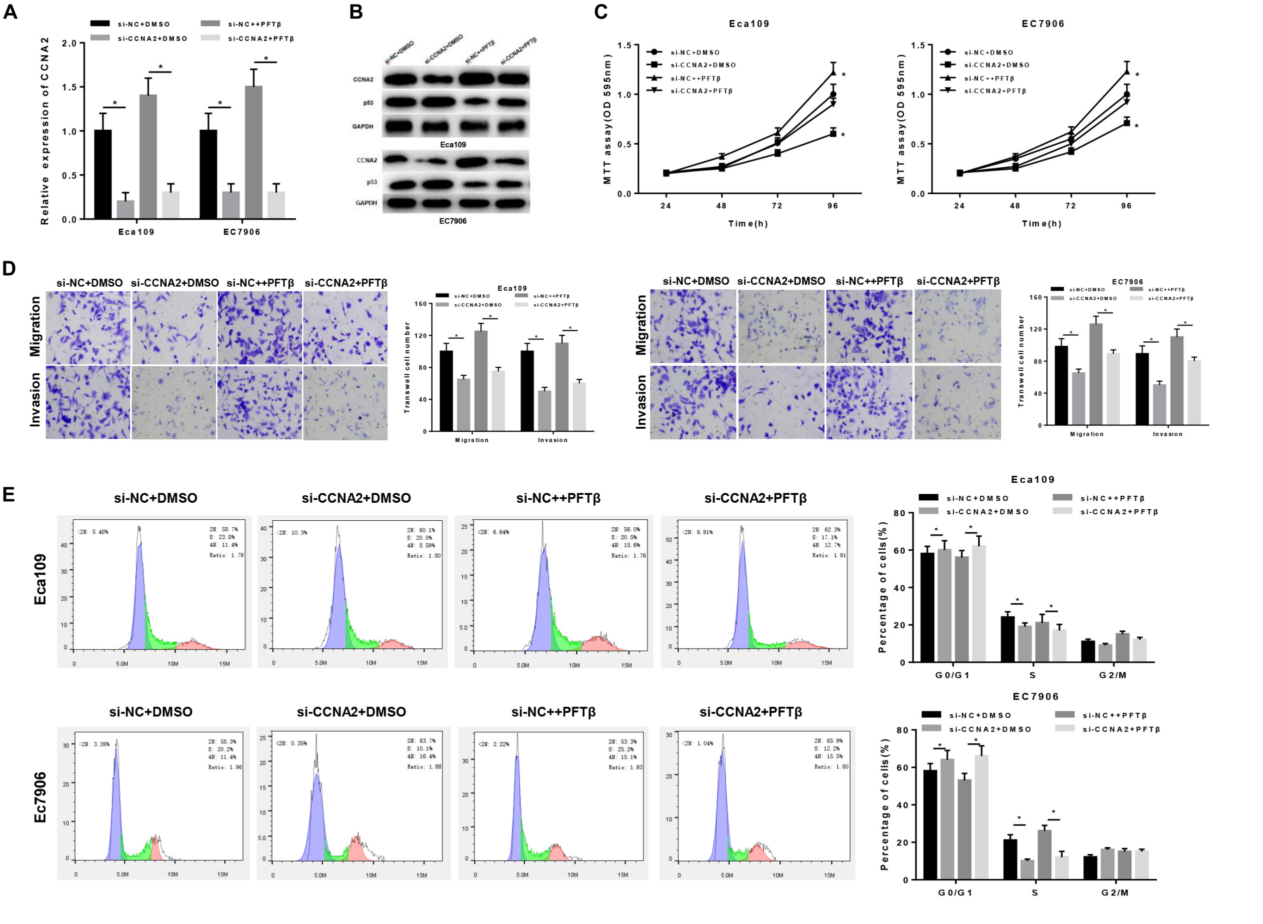


FIGURE 4 | CCNA2 silencing regulates the migration, invasion and cell cycle in EC by promoting p53 signaling pathway. si-NC + DMSO, si-CCNA2 + DMSO, si-NC + PFTβ and si-CCNA2 + PFTβ were transfected into cells. qRT-PCR and Western blot were conducted to determine **(A)** the CCNA2 mRNA and **(B)** protein levels of CCNA2 as well as p53. MTT, Transwell, and flow cytometry were performed to investigate the effects of silencing CCNA2 on EC cell activities, including **(C)** cell viability, **(D)** migration and invasion, **(E)** cell cycle. *$P < 0.05$.

was restored. Moreover, MTT assay showed that miR-29c-3p inhibitor could promote cell proliferation, while such promotive effect was inhibited when CCNA2 was silenced (**Figure 5B**). Similar results could be concluded as suggested by Transwell migration and invasion assays (**Figure 5C**). Meanwhile, with the results of flow cytometry plotted in **Figure 5D**, it could be indicated that the accumulation of cells in G0/G1 phase was positively associated with miR-29c-3p expression. Collectively, we speculated that miR-29c-3p targeted CCNA2 to regulate p53 signaling pathway, thereby repressing cell migration, invasion and resulting in cell cycle arrest, culminating in the inhibition of tumorigenesis.

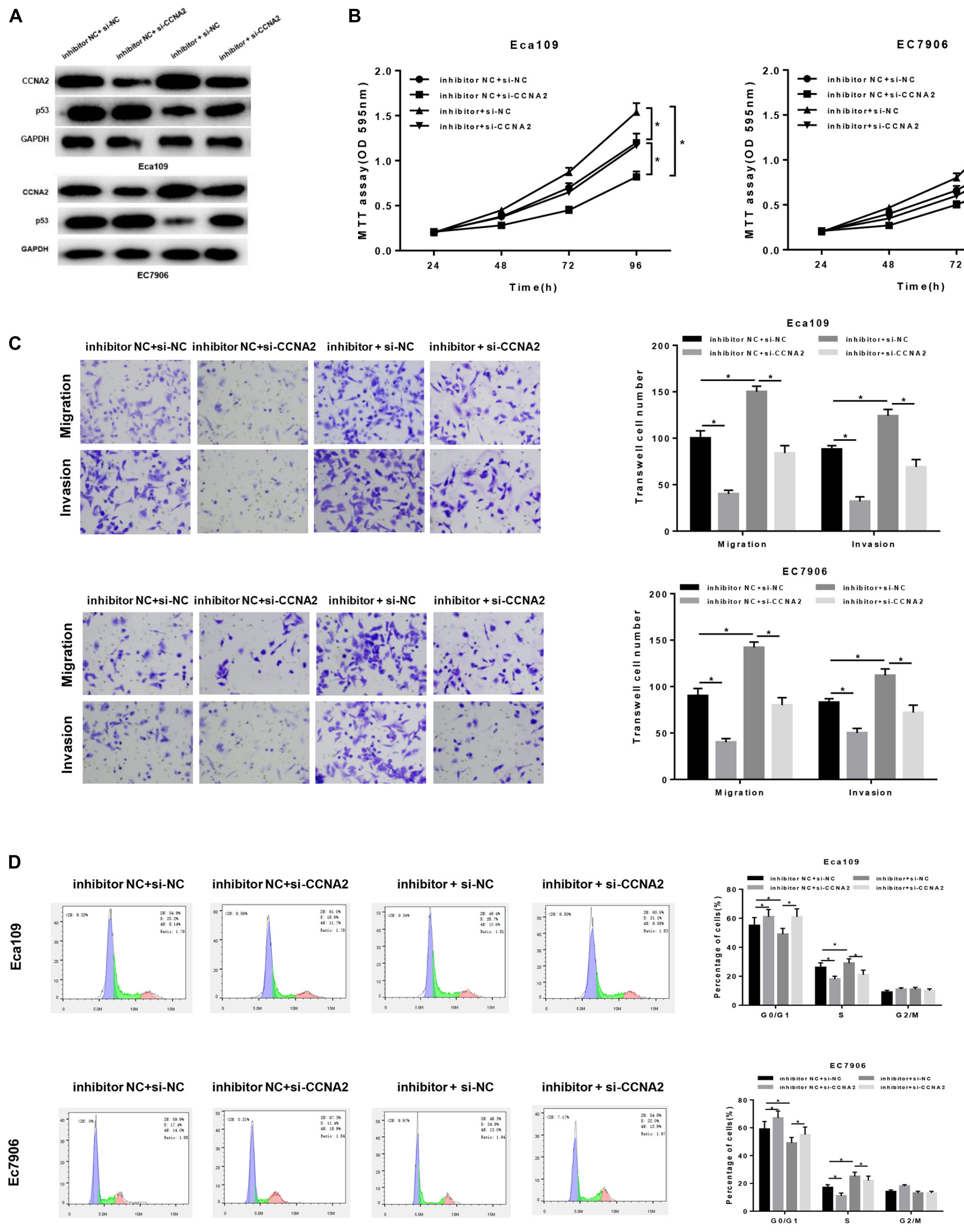


FIGURE 5 | MiR-29c-3p mediates the migration, invasion and cell cycle in EC via CCNA2/p53 axis. Cells were treated with inhibitor NC + si-NC, inhibitor NC + si-CCNA2, miR-29c-3p inhibitor + si-NC and miR-29c-3p inhibitor + si-CCNA2, and then harvested for **(A)** Western blot to detect the protein levels of CCNA2 and p53. **(B)** MTT was performed to test cell viability, **(C)** Transwell was conducted to assay the ability of cell migration and invasion, and **(D)** flow cytometry was carried out to determine the effect of miR-29c-3p on cell cycle. *$P < 0.05$.

DISCUSSION

Enormous evidence has revealed that miRNAs possess the ability of mediating cancer cell proliferation, migration and invasion, exerting their important regulatory roles in tumorigenesis and development (Lu et al., 2005; Deng et al., 2016). In ESCC, miR-124-3p could targeted bind with 3′UTR of BCAT1, and was found to be firmly correlated with cell proliferation and migration (Zeng et al., 2019). Renata Hezova et al. (2015) reported that miR-21, miR-29c, miR-148, and miR-203 could serve as potential diagnostic and prognostic biomarkers in EAC and ESCC. Therein, miR-29c has been reported to function on cell proliferation, migration, invasion and cell cycle, supporting its potential as a therapeutic target (Rao and Pattabiraman, 1989; Fan et al., 2014; Zhao et al., 2015; Li et al., 2018). For example, miR-29c could induce cell cycle arrest in ESCC through mediating the expression of cyclin E (Ding et al., 2011).

In the present study, we verified the low expression of miR-29c-3p in EC epithelial tissues and cells, and found that cell viability, migration and invasion were significantly decreased in cells treated with miR-29c-3p mimic. In addition, miR-29c-3p was seen to be associated with cell cycle as indicated by GSEA. Compared with the NC group, overexpression of miR-29c-3p attributed to the cell cycle arrest in G0/G1 phase, thus inhibiting the cell proliferation.

CCNA2 was found to be a direct target of miR-29c-3p, which was predicted by bioinformatics methods and verified by dual-luciferase reporter gene and RIP assays. Silencing CCNA2 could decrease the inhibitory effect of miR-29c-3p on EC epithelial cell proliferation, migration and invasion. CCNA2, a member of cyclin family, is a core regulatory factor during cell cycle progression, and participates in the regulation of S phase and mitosis. In addition, another study has showed that CCNA2 is involved in cytoskeleton dynamics behaviors and cell activities (Bendris et al., 2012), and the dysregulation of CCNA2 expression can be used as a marker of metastasis (Loukil et al., 2015). Notably, CCNA2 is commonly related to cell proliferation and highly expressed in many cancers. For instance, Cyclin A2 and Cyclin E2 can be mediated by SOSTDC1 and potentiate cell proliferation in thyroid cancer (Liang et al., 2015). FH535 can suppress cell proliferation and migration in colorectal cancer through regulating cyclin A2 and Claudin1 (Tu et al., 2019). Besides, Xu et al. (2019) found that miRNAs-mediated CCNA2 targeted p53 to inhibit cell senescence, in other words, p53/miRNAs/CCNA2 axis could be used as a novel regulator for cell senescence.

In this research, CCNA2 was found to be mainly activated in p53 signaling pathway and cell cycle detected by GSEA, and significantly up-regulated in EC tissues and cells relative to the normal controls. Silencing CCNA2 could remarkably repress cell proliferation, migration and invasion. In addition, Western blot showed that the decrease of CCNA2 attributed to the increase of p53, which played a crucial role in the regulation of cell cycle and apoptosis, as well as in the response of cell to DNA damage. When CCNA2 and p53 were simultaneously silenced, the protein level of p53 was seen to be up-regulated relative to that with p53 inhibitor alone, elucidating that silencing CCNA2 could promote p53 expression to some extent, thus activating p53 signaling pathway, consequently inhibiting cell proliferation, migration and invasion, and inducing cell arrest in G0/G1 phase.

In summary, our study confirmed that high miR-29c-3p expression can inhibit cell proliferation, migration and invasion in EC via CCNA2/p53 axis, which helps us to explore a novel approach on EC diagnosis and treatment.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

All procedures were performed with the approval of the Ethics Committee in the Shaoxing People's Hospital and informed consent was obtained from all patients before this study.

AUTHOR CONTRIBUTIONS

HW and LF conceived and designed the study. DW, BW, CZ, and TZ performed the experiments. ZM and ZL wrote the manuscript. YW, HW, LF, and GY reviewed and edited the manuscript. All authors read and approved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00075/full#supplementary-material

FIGURE S1 | GSEA analysis of miR-29c-3p and CCNA2. **(A)** GSEA suggested that miR-29c-3p was highly activated in cell cycle. **(B,C)** GSEA revealed the correlation of high CCNA2 expression with cell cycle and p53 signaling pathway. FDR < 0.05.

REFERENCES

Bendris, N., Loukil, A., Cheung, C., Arsic, N., Rebouissou, C., Hipskind, R., et al. (2012). Cyclin A2: a genuine cell cycle regulator? *Biomol. Concepts* 3, 535–543. doi: 10.1515/bmc-2012-0027

Bohanes, P., Yang, D., Chhibar, R. S., Labonte, M. J., Winder, T., Ning, Y., et al. (2012). Influence of sex on the survival of patients with esophageal cancer. *J. Clin. Oncol.* 30, 2265–2272. doi: 10.1200/JCO.2011.38.8751

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., et al. (2004). Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2999–3004. doi: 10.1073/pnas.0307323101

Da Pozzo, E., La Pietra, V., Cosimelli, B., Da Settimo, F., Giacomelli, C., Marinelli, L., et al. (2014). p53 functional inhibitors behaving like pifithrin-beta counteract the Alzheimer peptide non-beta-amyloid component effects in human SH-SY5Y cells. *ACS Chem. Neurosci.* 5, 390–399. doi: 10.1021/cn4002208

Deng, D., Xue, L., Shao, N., Qu, H., Wang, Q., Wang, S., et al. (2016). miR-137 acts as a tumor suppressor in astrocytoma by targeting RASGRF1. *Tumour. Biol.* 37, 3331–3340. doi: 10.1007/s13277-015-4110-y

Ding, D. P., Chen, Z. L., Zhao, X. H., Wang, J. W., Sun, J., Wang, Z., et al. (2011). miR-29c induces cell cycle arrest in esophageal squamous cell carcinoma by modulating cyclin E expression. *Carcinogenesis* 32, 1025–1032. doi: 10.1093/carcin/bgr078

Doan, P., Musa, A., Candeias, N. R., Emmert-Streib, F., Yli-Harja, O., and Kandhavelu, M. (2019). Alkylaminophenol induces G1/S phase cell cycle arrest in glioblastoma cells through p53 and cyclin-dependent kinase signaling pathway. *Front. Pharmacol.* 10:330. doi: 10.3389/fphar.2019.00330

Fan, Y., Song, X., Du, H., Luo, C., Wang, X., Yang, X., et al. (2014). Down-regulation of miR-29c in human bladder cancer and the inhibition of proliferation in T24 cell via PI3K-AKT pathway. *Med. Oncol.* 31:65. doi: 10.1007/s12032-014-0065-x

Fu, H. L., Wu, D. P., Wang, X. F., Wang, J. G., Jiao, F., Song, L. L., et al. (2013). Altered miRNA expression is associated with differentiation, invasion, and metastasis of esophageal squamous cell carcinoma (ESCC) in patients from Huaian. China. *Cell Biochem. Biophys.* 67, 657–668. doi: 10.1007/s12013-013-9554-3

Gao, T., Han, Y., Yu, L., Ao, S., Li, Z., and Ji, J. (2014). CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance. *PLoS One* 9:e91771. doi: 10.1371/journal.pone.0091771

Gertler, R., Stein, H. J., Langer, R., Nettelmann, M., Schuster, T., Hoefler, H., et al. (2011). Long-term outcome of 2920 patients with cancers of the esophagus and esophagogastric junction: evaluation of the New Union Internationale Contre le Cancer/American Joint cancer committee staging system. *Ann. Surg.* 253, 689–698. doi: 10.1097/SLA.0b013e31821111b5

Harada, K., Baba, Y., Ishimoto, T., Shigaki, H., Kosumi, K., Yoshida, N., et al. (2016). The role of microRNA in esophageal squamous cell carcinoma. *J. Gastroenterol.* 51, 520–530. doi: 10.1007/s00535-016-1161-9

Hezova, R., Kovarikova, A., Srovnal, J., Zemanova, M., Harustiak, T., Ehrmann, J., et al. (2015). Diagnostic and prognostic potential of miR-21, miR-29c, miR-148 and miR-203 in adenocarcinoma and squamous cell carcinoma of esophagus. *Diagn Pathol.* 10:42. doi: 10.1186/s13000-015-0280-6

Huang, H. L., Chen, W. C., Hsu, H. P., Cho, C. Y., Hung, Y. H., Wang, C. Y., et al. (2017). Silencing of argininosuccinate lyase inhibits colorectal cancer formation. *Oncol. Rep.* 37, 163–170. doi: 10.3892/or.2016.5221

Kamangar, F., Dores, G. M., and Anderson, W. F. (2006). Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J. Clin. Oncol.* 24, 2137–2150. doi: 10.1200/JCO.2005.05.2308

Krasnov, G. S., Puzanov, G. A., Kudryavtseva, A. V., Dmitriev, A. A., Beniaminov, A. D., Kondratieva, T. T., et al. (2017). [Differential expression of an ensemble of the key genes involved in cell-cycle regulation in lung cancer]. *Mol. Biol.* 51, 849–856. doi: 10.7868/S0026898417050135

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi: 10.1016/j.cell.2004.12.035

Li, W., Yi, J., Zheng, X., Liu, S., Fu, W., Ren, L., et al. (2018). miR-29c plays a suppressive role in breast cancer by targeting the TIMP3/STAT1/FOXO1 pathway. *Clin. Epigenetics* 10:64. doi: 10.1186/s13148-018-0495-y

Liang, W., Guan, H., He, X., Ke, W., Xu, L., Liu, L., et al. (2015). Down-regulation of SOSTDC1 promotes thyroid cancer cell proliferation via regulating cyclin A2 and cyclin E2. *Oncotarget* 6, 31780–31791. doi: 10.18632/oncotarget.5566

Loukil, A., Cheung, C. T., Bendris, N., Lemmers, B., Peter, M., and Blanchard, J. M. (2015). Cyclin A2: at the crossroads of cell cycle and cell invasion. *World J. Biol. Chem.* 6, 346–350. doi: 10.4331/wjbc.v6.i4.346

Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., et al. (2005). MicroRNA expression profiles classify human cancers. *Nature* 435, 834–838. doi: 10.1038/nature03702

Ma, Q. (2019). MiR-219-5p suppresses cell proliferation and cell cycle progression in esophageal squamous cell carcinoma by targeting CCNA2. *Cell Mol. Biol. Lett.* 24:4. doi: 10.1186/s11658-018-0129-6

Nie, J., Ge, X., Geng, Y., Cao, H., Zhu, W., Jiao, Y., et al. (2015). miR-34a inhibits the migration and invasion of esophageal squamous cell carcinoma by targeting Yin Yang-1. *Oncol. Rep.* 34, 311–317. doi: 10.3892/or.2015.3962

Qi, Y., Li, X., and Zhao, S. (2015). miR-29b inhibits the progression of esophageal squamous cell carcinoma by targeting MMP-2. *Neoplasma* 62, 384–390. doi: 10.4149/neo_2015_046

Rao, P., and Pattabiraman, T. N. (1989). Reevaluation of the phenol-sulfuric acid reaction for the estimation of hexoses and pentoses. *Anal. Biochem.* 181, 18–22. doi: 10.1016/0003-2697(89)90387-4

Shekhar, R., Priyanka, P., Kumar, P., Ghosh, T., Khan, M. M., Nagarajan, P., et al. (2019). The microRNAs miR-449a and miR-424 suppress osteosarcoma by targeting cyclin A2 expression. *J. Biol. Chem.* 294, 4381–4400. doi: 10.1074/jbc.RA118.005778

Tu, X., Hong, D., Jiang, Y., Lou, Z., Wang, K., Jiang, Y., et al. (2019). FH535 inhibits proliferation and migration of colorectal cancer cells by regulating CyclinA2 and Claudin1 gene expression. *Gene* 690, 48–56. doi: 10.1016/j.gene.2018.12.008

Wang, B., Yin, B. L., He, B., Chen, C., Zhao, M., Zhang, W., et al. (2012). Overexpression of DNA damage-induced 45 alpha gene contributes to esophageal squamous cell cancer by promoter hypomethylation. *J. Exp. Clin. Cancer Res.* 31:11. doi: 10.1186/1756-9966-31-11

Wu, Y., Li, H., Wang, H., Zhang, F., Cao, H., and Xu, S. (2019).). MSK2 promotes proliferation and tumor formation in squamous cervical cancer via PAX8/RB-E2F1/cyclin A2 axis. *J. Cell Biochem.* doi: 10.1002/jcb.28421 [Epub ahead of print].

Xu, S., Wu, W., Huang, H., Huang, R., Xie, L., Su, A., et al. (2019). The p53/miRNAs/Ccna2 pathway serves as a novel regulator of cellular senescence: complement of the canonical p53/p21 pathway. *Aging Cell* 18:e12918. doi: 10.1111/acel.12918

Yang, F., Gong, J., Wang, G., Chen, P., Yang, L., and Wang, Z. (2016). Waltonitone inhibits proliferation of hepatoma cells and tumorigenesis via FXR-miR-22-CCNA2 signaling pathway. *Oncotarget* 7, 75165–75175. doi: 10.18632/oncotarget.12614

Yao, L., Zhang, Y., Zhu, Q., Li, X., Zhu, S., Gong, L., et al. (2015). Downregulation of microRNA-1 in esophageal squamous cell carcinoma correlates with an advanced clinical stage and its overexpression inhibits cell migration and invasion. *Int. J. Mol. Med.* 35, 1033–1041. doi: 10.3892/ijmm.2015.2094

Zaidi, N., and Kelly, R. J. (2017). The management of localized esophageal squamous cell carcinoma: western approach. *Chin. Clin. Oncol.* 6:46. doi: 10.21037/cco.2017.07.07

Zeng, B., Zhang, X., Zhao, J., Wei, Z., Zhu, H., Fu, M., et al. (2019). The role of DNMT1/hsa-miR-124-3p/BCAT1 pathway in regulating growth and invasion of esophageal squamous cell carcinoma. *BMC Cancer* 19:609. doi: 10.1186/s12885-019-5815-x

Zhang, H. P., Li, S. Y., Wang, J. P., and Lin, J. (2018). Clinical significance and biological roles of cyclins in gastric cancer. *Onco. Targets Ther.* 11, 6673–6685. doi: 10.2147/OTT.S171716

Zhao, X., Li, J., Huang, S., Wan, X., Luo, H., and Wu, D. (2015). MiRNA-29c regulates cell growth and invasion by targeting CDK6 in bladder cancer. *Am. J. Transl. Res.* 7, 1382–1389.

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: 28 February 2020
doi: 10.3389/fgene.2020.00147

Fragment Enrichment of Circulating Tumor DNA With Low-Frequency Mutations

Xiaojun Liu[1], Jidong Lang[2†], Shijun Li[3†], Yuehua Wang[3], Lihong Peng[1], Weitao Wang[2], Yingmin Han[2], Cuixiao Qi[2], Lei Song[2], Shuangshuang Yang[2], Kaixin Zhang[2], Guoliang Zang[2], Hong Pei[2], Qingqing Lu[2], Yonggang Peng[2], Shuxue Xi[2], Weiwei Wang[2], Dawei Yuan[2], Pingping Bing[4], Liqian Zhou[1*] and Geng Tian[1,2*]*

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] Bioinformatics Department, Geneis (Beijing) Co. Ltd., Beijing, China, [3] Department of Pathology, Chifeng Municipal Hospital, Chifeng, China, [4] Academics Working Station, Changsha Medical University, Changsha, China

Human blood contains cell-free DNA (cfDNA), with circulating tumor-derived DNAs (ctDNAs) widely used in cancer diagnosis and treatment. However, it is still difficult to efficiently and accurately identify and distinguish specific ctDNAs from normal cfDNA in cancer patient blood samples. In this study, ctDNA fragment length distribution analysis showed that ctDNA fragments are frequently shorter than the normal cfDNAs, which is consistent with previous findings. Interestingly, the ctDNA fragment length was found to be partially associated with the mutant allele frequency, with a low mutant allele frequency (< ~0.6%) associated with a longer ctDNA fragment length when compared to normal cfDNAs. The findings of this study contribute to improving the detection of low-frequency tumor mutations.

Keywords: low-frequency tumor mutation, cell-free DNA, circulating tumor-derived DNA, fragment length enrichment, mutant allele frequency, next generation sequencing

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS),
China

Reviewed by:
Lu Zhang,
Hong Kong Baptist University,
Hong Kong
Bing Wang,
Anhui University of Technology,
China

***Correspondence:**
Jidong Lang
langjd@geneis.cn
Liqian Zhou
zhoulq11@163.com
Geng Tian
tiang@geneis.cn

†These authors have contributed equally to this work

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

Received: *27 November 2019*
Accepted: *07 February 2020*
Published: *28 February 2020*

Citation:
Liu X, Lang J, Li S, Wang Y, Peng L, Wang W, Han Y, Qi C, Song L, Yang S, Zhang K, Zang G, Pei H, Lu Q, Peng Y, Xi S, Wang W, Yuan D, Bing P, Zhou L and Tian G (2020) Fragment Enrichment of Circulating Tumor DNA With Low-Frequency Mutations. Front. Genet. 11:147. doi: 10.3389/fgene.2020.00147

INTRODUCTION

In modern medicine, liquid biopsies are widely used in prenatal diagnoses and cancer treatment. When utilizing a liquid biopsy, circulating cell-free DNA (cfDNA), circulating tumor cells (CTCs), or exosomes are isolated for evaluation (Bardelli and Pantel, 2017; Wan et al., 2017; Siravegna et al., 2017). Of these, circulating tumor-derived DNA (ctDNA) is widely utilized as a tumor biomarker in translational and clinical research (Diaz and Bardelli, 2014; Donaldson and Park, 2018), while fetal cfDNA obtained from maternal blood is widely used as a noninvasive method for prenatal diagnoses (Lun et al., 2008; Lo et al., 2010; Yu et al., 2014; Sun et al., 2018).

About 30 years ago, Stroun et al. first discovered that cancer patient blood samples contain cfDNA of cancer origin (Stroun et al., 1989; Thierry et al., 2016). In the following decades, ctDNA has been gradually developed as a clinical tool for cancer diagnosis and treatment, and has even been used as a prognostic or predictive factor (Mao et al., 1994; Lecomte et al., 2002; Kimura, 2006; Diehl et al., 2008). Currently, the use of ctDNA detection in cancer therapy has been approved by the US Food and Drug Administration as a treatment determinant (osimertinib or erlotinib) in non-small-cell lung carcinoma (NSCLC) patients with an *EGFR* mutation in the event that a tumor biopsy cannot be performed (US Food & Drug Administration, 2016). The application of ctDNA in cancer therapy is reliant on precise polymerase chain reaction (PCR)-based technologies, such as droplet digital PCR (ddPCR) or amplification refractory mutation system (ARMS)-PCR, and deep-sequencing technologies; these techniques aid in distinguishing ctDNAs from other normal cfDNAs within the plasma and enable hotspot mutation detection within cancer driver genes (Taly et al., 2013; Newman et al., 2014; Frenel et al., 2015; Azizi et al., 2018). However, ctDNAs are usually present in low abundance relative to the normally occurring cfDNAs derived from normal cells, particularly in non-metastatic solid tumors (Tug et al., 2014; Siravegna et al., 2017). Consequently, there is an urgent need to reliably distinguish ctDNAs from normal cfDNA to improve the accuracy of identifying driver gene mutations.

Recently, tumor-derived ctDNAs have been shown to vary in size and are shorter than normal cfDNAs in healthy people (Umetani et al., 2006; Thierry et al., 2010; Mouliere et al., 2011; Mouliere et al., 2013). This trend was also observed during pregnancy, with fetal cfDNA usually of a different fragment size than the maternal cfDNA (Lun et al., 2008; Lo et al., 2010). Furthermore, in one study examining ctDNA length distributions in hepatocellular carcinoma (HCC) patients, copy number aberrations were leveraged and showed that high-concentration ctDNA fractions were more fragmented, while low-concentration fractions were paradoxically longer (Jiang et al., 2015; Mouliere and Rosenfeld, 2015; Jiang and Lo, 2016). In another study, ctDNAs were found to be consistently shorter than normal cfDNA, in both animal xenograft models and clinical plasma samples (Underhill et al., 2016). Additionally, mutant ctDNA fragments from tumor patients were always shorter than wild-type cfDNA fragments from healthy donors, with mutant alleles more commonly having shorter fragment lengths, something that could potentially be exploited to improve ctDNA detection (Underhill et al., 2016; Hellwig et al., 2018). Moreover, a later study confirmed that this size difference could be exploited to enhance sensitivity when monitoring ctDNAs and for noninvasive genomic analysis of various cancers (Mouliere et al., 2018). However, few studies have examined the impact of mutant allele frequency on the size distribution of ctDNA fragments, and most studies were conducted in cancer patients with relatively high mutant allele frequencies.

Thus, the aim of this study was to examine ctDNA fragment distributions in patients with low mutant allele frequencies and determine whether the ctDNA fragment length is affected by the mutant allele frequency. This was accomplished by utilizing blood samples from cancer patients with a variety of different histological types and stages. Key driver gene mutation frequencies were determined using deep-sequencing technologies and ddPCR, and fragment length differences between mutant ctDNAs and normal cfDNAs obtained from the cancer patient samples were examined.

MATERIALS AND METHODS

Sample Collection

All 105 samples (male: 49.52%, female: 50.48%) were obtained from lung cancer patients from Chifeng Municipal Hospital. All patients provided informed written consent before de-identification. The median age of the patients was 63.5 years old (range from 36 to 85 years old). Our research was approved by the Medical Research Ethics Committee of Chifeng Municipal Hospital (Ethics [2018] No. 017).

Next-Generation Sequencing (NGS) Library Preparation, Sequencing, and Bioinformatics

Cell-free DNA was extracted using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The extracted DNA (20 ng/sample) was then used to build libraries using Accel-NGS® 2S Plus DNA Library Kits (96 reactions; Swift BioSciences, Ann Arbor, MI, USA). Customized probes were obtained from Integrated DNA Technologies (IDT, Skokie, IL, USA) and were used for hybridization capture. All cfDNA libraries utilized a 38-hotspot gene panel (**Supplementary File**) and were quantified using a Universal Library Quantification Kit (Kapa Biosystems, Wilmington, MA, USA) on an ABI 7500 Real-Time PCR system (Applied Biosystems, Waltham, MA, USA). Sample quality was evaluated using a high sensitivity DNA kit (Agilent Technologies, Santa Clara, CA, USA) with an Agilent 2100 Bioanalyzer as per the manufacturer's instructions. NGS with fusion detection was performed using a NextSeq 500/550 High Output v2 kit with a NextSeq 500 sequencer (Illumina, San Diego, CA, USA) for 302 cycles, with standing paired-end reads of 151 bp (average sequencing depth was ~2,164X, details in the **Supplementary File**).

The FASTQ reads were collapsed into unique observations based on barcodes using CASAVA (v1.8.2) software. Low-quality and adapter-contaminated reads were removed from the raw reads using Cutadapt (v1.12) and aligned to the Hg19 reference genome using the Burrow-Wheeler Aligner for short-read alignment (bwa aln; 0.7.12-r1039). Paired-end reads with hotspots were extracted from the paired-end alignment information (column 9th) in BAM format using Samtools (v0.1.19-44428cd), and the corresponding insert size information was extracted. Finally, the extracted paired-end reads were aligned to the Hg19 reference genome again using SOAP (2.21), and hotspot mutation fragment lengths and wild fragment lengths were calculated with the alignment mismatch information (column 11th) in the alignment files.

Digital Droplet PCR

EGFR-T790M, *EGFR*-L858R, *BRAF*-V600E, *PIK3CA*-E545K, *KRAS*-G12C, and *KRAS*-G12V mutant allele frequencies were determined using a Digital Droplet PCR system (Bio-Rad Laboratories, Inc., Hercules, CA, USA), with a droplet size of 1 nL in a total reaction volume of 20 μL, with ~20 ng of cfDNA library utilized. All primers and probes were synthesized by IDT (Skokie, IL, USA; **Table 1**). Droplet counts were determined using the QuantaSoft software (Bio-Rad).

For the *PIK3CA*-E545K (n = 1), *KRAS*-G12C (n = 3), *KRAS*-G12V (n = 1), *EGFR*-T790M (n = 5), and *EGFR*-L858R (n = 1) samples, amplified libraries were utilized prior to size selection to

TABLE 1 | Primers and probes used in droplet digital PCR experiments.

Mutation	Forward primer	Reverse primer	Wild probe	Mutation probe
T790M	GCCTGCTGGGCATCTG	TCTTTGTGTTCCCGGACATAGTC	VIC-ATGAGCTGCGTGATGAG-MGB-NFQ	FAM-ATGAGCTGCATGATGAG-MGB-NFQ
L858R	GCAGCATGTCAAGATCACAGATT	CCTCCTTCTGCATGGTATTCTTTCT	VIC-AGTTTGGCCAGCCCAA-MGB-NFQ	FAM-AGTTTGGCCCGCCCAA-MGB-NFQ
V600E	CATGAAGACCTCACAGTAAAAATAGGTGAT	TGGGACCCACTCCATCGA	VIC-CTAGCTACAGTGAAATC-MGB-NFQ	FAM-TAGCTACAGAGAAATC-MGB-NFQ
E545K	CACTTACCTCTGACTCCATAGAAAATCTT	AAAGCAATTACTACACGATATCCTCTCTC	HEX-TCCTGCTCAGTGATT-MGB-NFQ	FAM-CTCCTGCTTAGTGATT-MGB-NFQ
G12 G12V	AATTAGATGTATCGTCAAGGCACTCTT	GCTGAAAATGACTGAATATAAACTTGTGG	VIC-TACGCCACCAGCTC-MGB-NFQ	FAM-TACGCCAACAGCTC-MGB-NFQ
G12C				FAM-TACGCCACAAGCTCT-MGB-NFQ

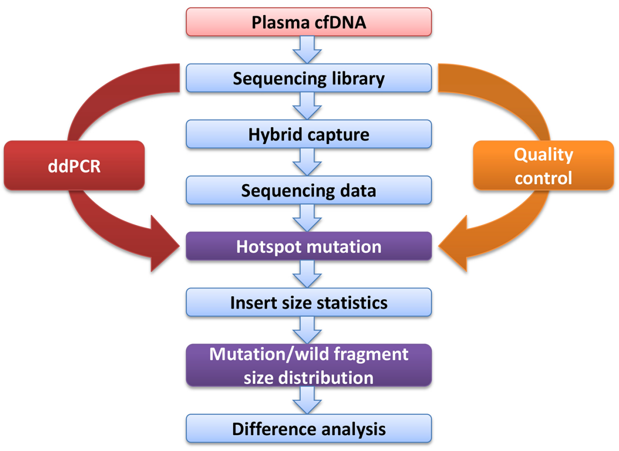


FIGURE 1 | Experimental design flow.

define gates for wild-type and mutation droplet populations. Libraries were constructed using the obtained DNA (~20 ng) and a Rapid DNA Lib Prep kit (ABclonal, Woburn, MA, USA). The obtained libraries (~1.2 mg) were then separated using 2% agarose gel electrophoresis and bands between 130–160 bp and 160–230 bp were extracted using a QIAquick Gel Extraction kit (Qiagen). All of the selected fragment size libraries were then validated using ddPCR as described above (**Supplementary File**).

RESULTS

Comparison of cfDNA Fragment Sizes in Cancer Patient Plasma Samples

Blood samples were obtained from cancer patients with defined driver gene hotspot mutations, including *EGFR*-T790M (n = 32), *EGFR*-L858R (n = 28), *BRAF*-V600E (n = 13), *PI3KCA*-E545K (n = 13), *KRAS*-G12C (n = 13), and *KRAS*-G12V (n = 6). The cfDNA-sequencing libraries were analyzed by both NGS and ddPCR to precisely detect the mutant allele frequencies of these six hotspots in each cancer patient (**Figure 1**). Some hotspot mutant allele frequencies were more variable, such as *EGFR*-T790M (0.11–74.75%), *EGFR*-L858R (0.15–35.77%), *PI3KCA*-E545K (0.10–21.67%), and *KRAS*-G12C (0.10–33.81%; **Table 2**). Furthermore, some hotspot allele frequencies within these driver genes were relatively low, including *BRAF*-V600E (0.10–0.30%) and *KRAS*-G12V (0.11–1.26%; **Table 2**), which could be explained by examining samples at different tumor stages and of different histological types collectively. Next, the cfDNA-sequencing libraries were sequenced, and size differences between plasma ctDNA and normal cfDNA were compared.

Mutant Alleles Have a Shorter Fragment Length Than the Wild-Type Alleles

In addition to examining cancer patient mutant allele frequencies, whole cfDNA fragment length distributions were globally observed. As expected, the mutant ctDNA fragments were generally shorter than the normal cfDNAs (**Figures 2** and **3**). In patients with a low mutation frequency, the ctDNA

TABLE 2 | Summary of the mutation frequencies based on next generation sequencing.

	T790M	*L858R*	*V600E*	*E545K*	*G12C*	*G12V*
Validation library number	32	28	13	13	13	6
Low mutation frequency 0.1–1%	16	11	13	11	4	5
Medium mutation frequency 1–10%	10	12	0	1	7	1
High-mutation frequency 10–100%	6	5	0	1	2	0
Mutation frequency distribution	0.11–74.75%	0.15–35.77%	0.10–0.30%	0.10–21.67%	0.10–33.81%	0.11–1.26%

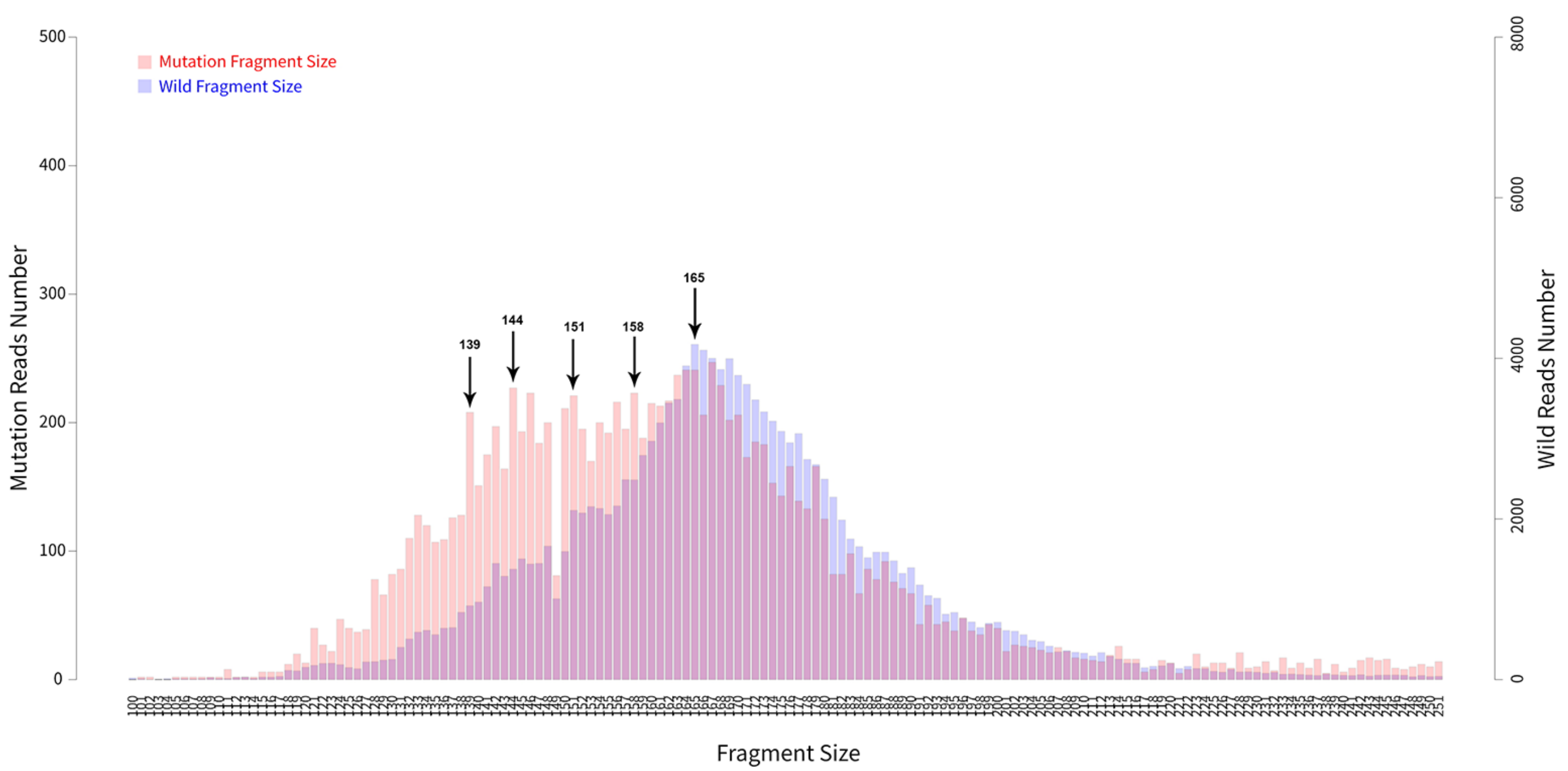


FIGURE 2 | Fragment length distributions of cfDNAs from 105 cancer patient blood samples.

fragment length was longer than the normal cfDNAs, such as *BRAF*-V600E (0.10–0.30%). However, this trend was not observed in the four other DNA fragment size distribution, including *EGFR*-T790M (0.11–74.75%), *EGFR*-L858R (0.15–35.77%), *PI3KCA*-E545K (0.10–21.67%), *KRAS*-G12V (0.11–1.26%) or *KRAS*-G12C (0.10–33.81%; **Figure 3**).

Longer Fragment Lengths in Mutant ctDNAs With a Low Mutation Frequency

Fragment size differences between cancer patient ctDNAs and normal cfDNAs were further examined in conjunction with a low, medium, or high mutant allele frequency. In fragments associated with a low mutant allele frequency, the ctDNA fragments were longer than the normal cfDNAs (**Figure 4**), such as *EGFR*-T790M (0.22 and 0.21%). However, in ctDNAs with a higher mutant allele frequency, such as *EGFR*-T790M (74.75%), or a medium frequency, such as *EGFR*-T790M (4.57%), fragment lengths were shorter than the normal cfDNAs (**Figure 5** and **Table 3**).

Low-Frequency Mutations Are Associated With Large Fragment Sizes

Different fragment sizes were observed among the mutant ctDNAs, including long ctDNA (longer than normal cfDNAs), normal ctDNA (comparable to normal cfDNA lengths), and short ctDNA (shorter than normal cfDNAs). Within these three groups, the mutant allele frequency distributions were examined and showed that a low mutation frequency was commonly associated with a long ctDNA fragment length, while normal and short ctDNAs were not (**Figure 6**).

Enrichment of Longer ctDNA Fragments Could Improve the Detection of Low-Frequency Mutations

After discovering that a low-frequency is associated with a longer ctDNA fragment size, this study aimed to determine if enriching longer cfDNA fragments could increase the mutation frequency in blood samples with a low mutant allele frequency. In one patient with a high frequency for *EGFR*-T790M (44.53%), cfDNA was extracted and different fragment sizes were obtained. To further detect the *EGFR*-T790M frequency, DNA libraries comprising two different DNA fragment sizes were examined using ddPCR. The *EGFR*-T790M frequency in a library with a fragment length between 160 and 230 bp (42.20%) was lower than the library with a fragment size between 130–160 bp (46.40%; **Figure 7A**). This was consistent with the findings presented above. Conversely, a cfDNA sample was obtained from a patient with a low *EGFR*-T790M frequency (0.54%) and different fragment sizes were collected and analyzed. In the library with fragment sizes between 160–230 bp, the *EGFR*-T790M frequency was increased (1.04%) when compared to the library with fragment sizes between 130–160 bp (0.30%; **Figure 7B**).

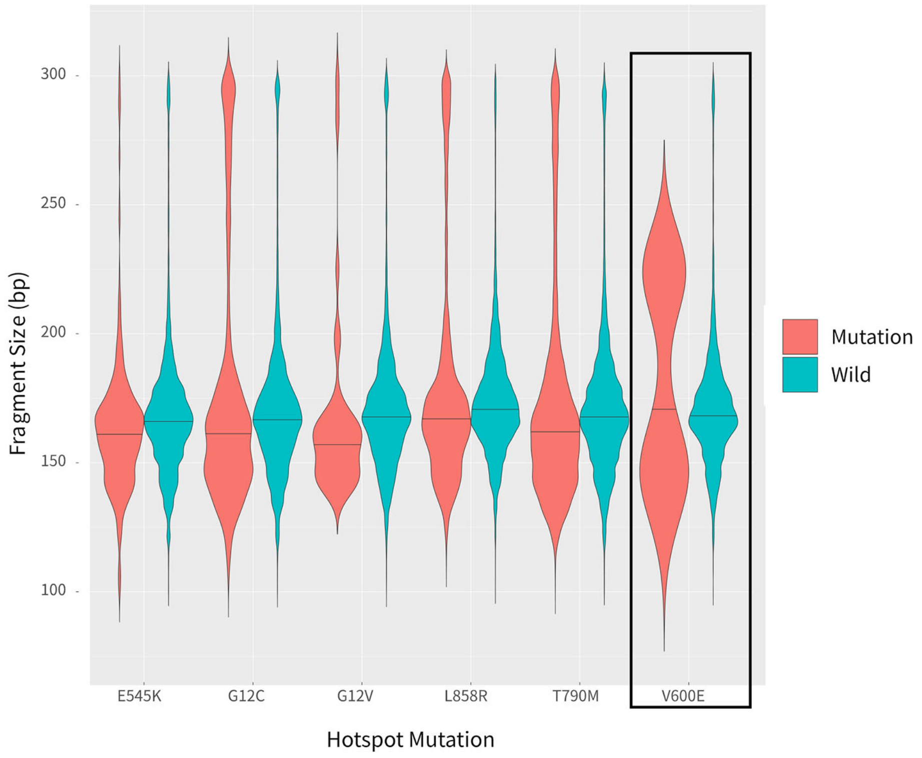


FIGURE 3 | Comparison of fragment length sizes between ctDNAs and normal cfDNAs.

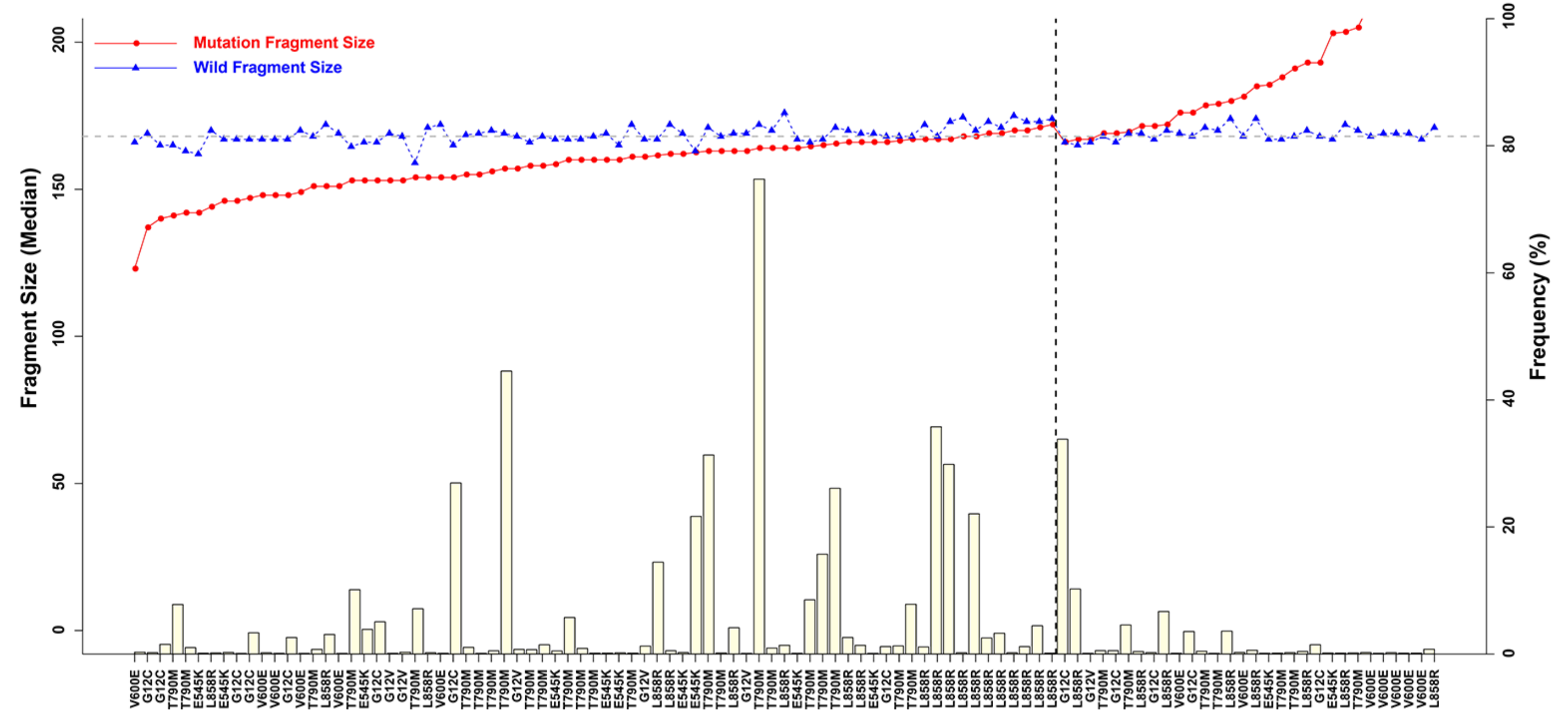


FIGURE 4 | Associations between mutant allele frequency and ctDNA fragment sizes. Gray horizontal dotted line is 168 bp, and black vertical dotted line is L858R.)

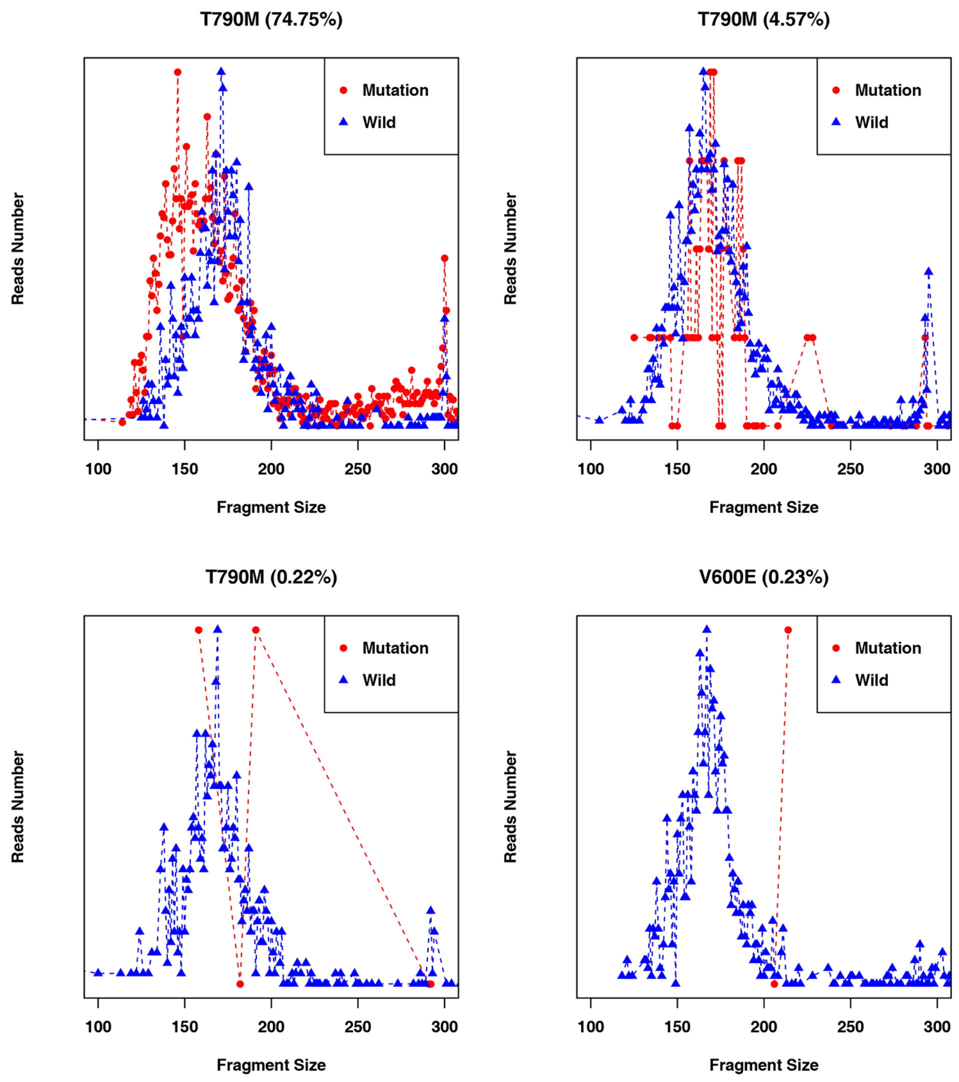


FIGURE 5 | Fragment length distributions of cancer patient ctDNAs and normal cfDNAs with high, medium, or low *EGFR-T790M* mutant allele frequencies.

TABLE 3 | Fragment length distributions of cancer patient ctDNAs and normal cfDNAs with high, medium, or low *EGFR-T790M* mutant allele frequencies.

Mutation type	NGS (%)	ddPCR (%)	Description	Mutation peak	Wild peak	Mutation fragment median	Wild fragment median
T790M	74.75	69.33	Short	146	171	164	172
T790M	4.57	5.55	Long	169/171	165	169.5	169
T790M	0.22	0.26	Other	158/191	169	191	168
V600E	0.23	0.17	Long	214	167	214	168

DISCUSSION

This study showed that a consistent fragment length difference occurs when comparing ctDNAs and normal cfDNA, with the mutant allele almost always associated with a shorter ctDNA fragment size, which is consistent with previous findings (Jiang et al., 2015; Underhill et al., 2016; Mouliere et al., 2018). However, some mutant ctDNAs were found to have a longer fragment size when compared to normal cfDNAs and were associated with a low mutant allele frequency, which has not been previously reported. Furthermore, this study showed that in cancer patient plasma samples, the ctDNA fragment length is associated with the mutant allele frequency and may even be affected by it.

Here, blood samples were obtained from 105 patients that contained different cancer driver gene mutations, such as NSCLC

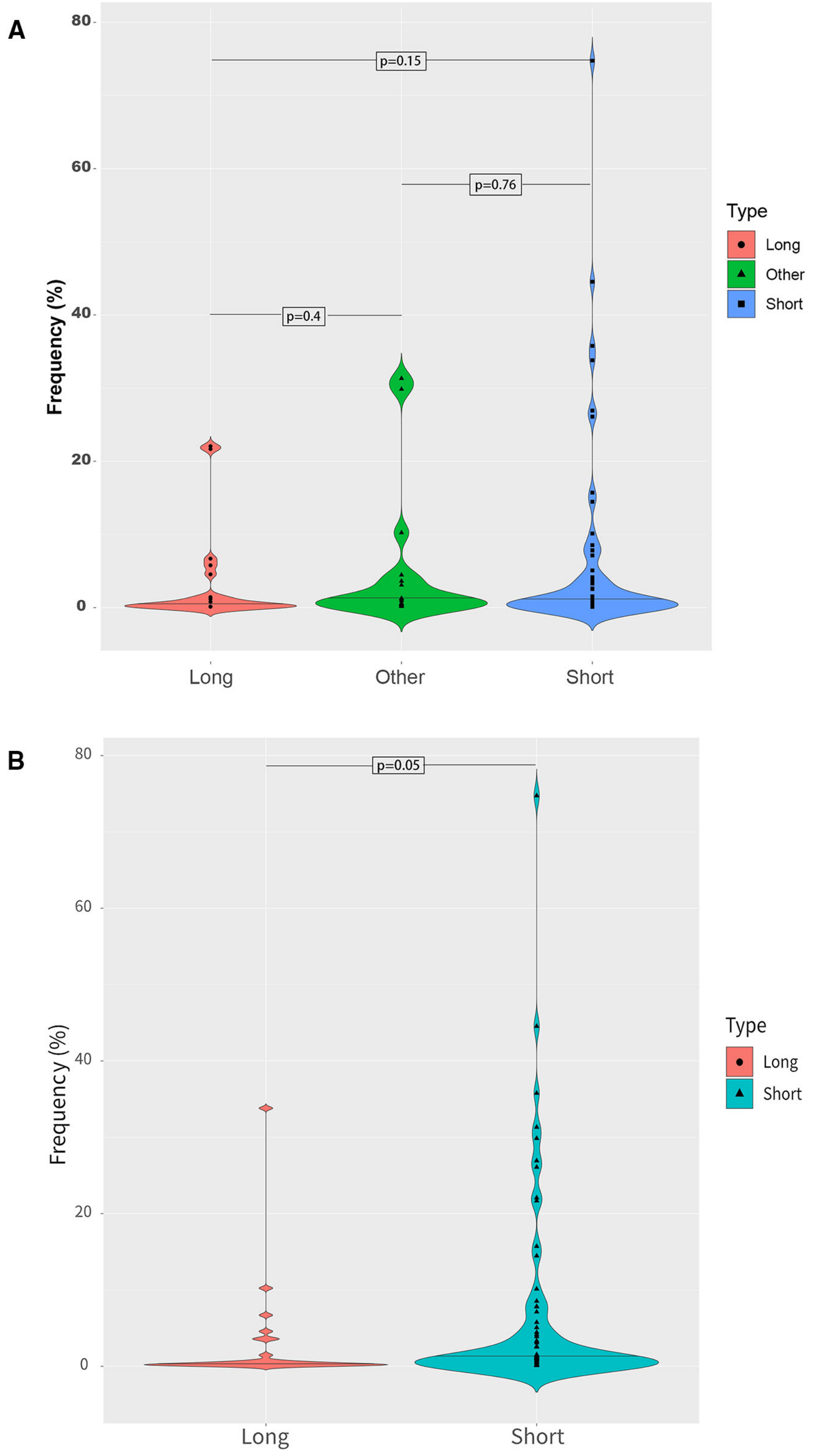


FIGURE 6 | Mutant allele frequency distributions based on ctDNA fragment length. **(A)** Relationship between the mutation fragment size peak and the mutation frequency. **(B)** Relationship between the median mutation fragment size peak and the mutation frequency.

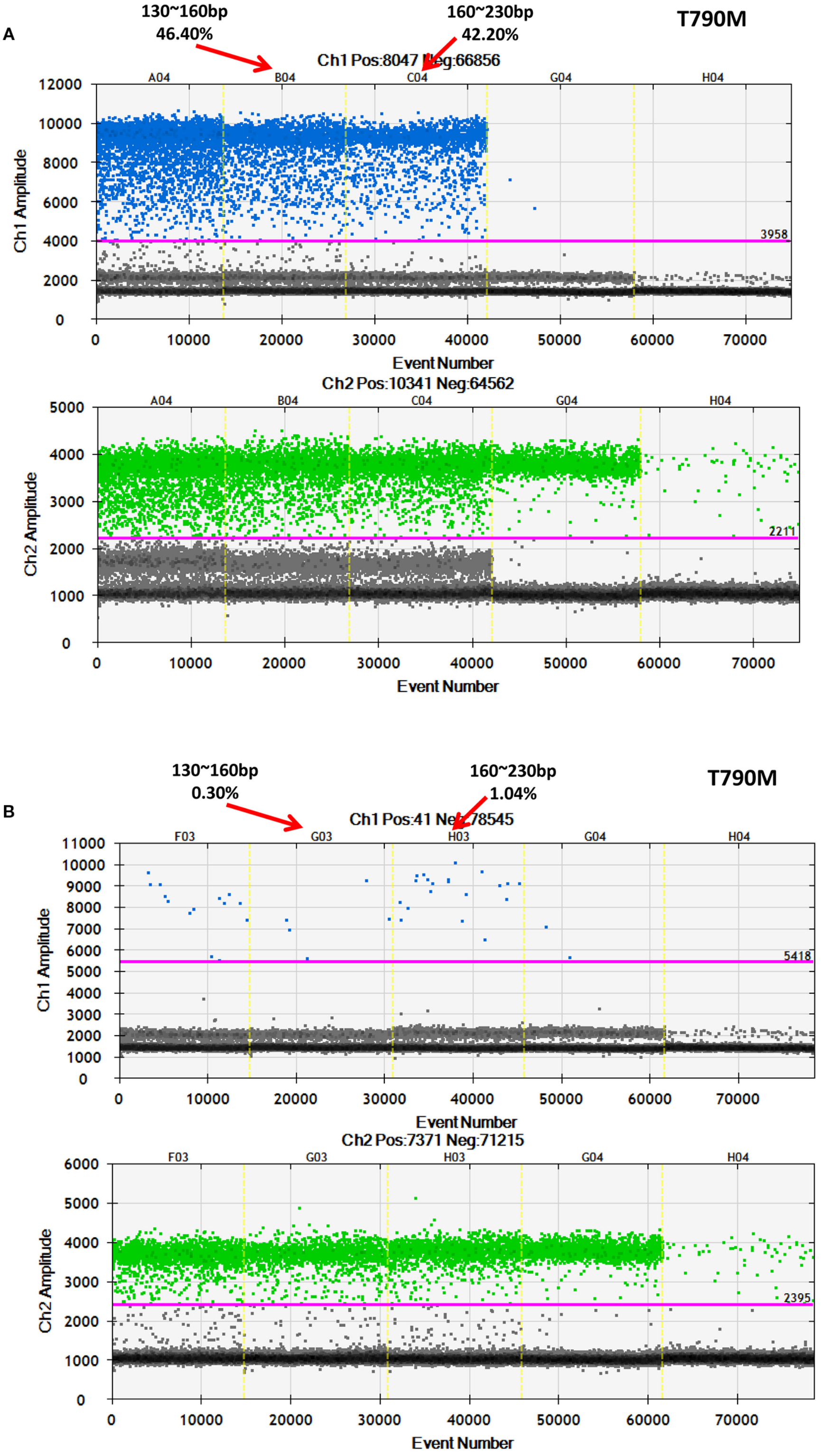


FIGURE 7 | Further validation of an association between fragment size and frequency using ddPCR. Examination of different fragment size libraries from a patient with **(A)** a high *EGFR-T790M* frequency (44.53%) and from a patient with **(B)** a low *EGFR-T790M* frequency (0.54%) using ddPCR.

patients with an *EGFR* gene mutation and colorectal cancer patients with a *BRAF* mutation. In general, mutant ctDNA fragments were found to be much shorter than normal cfDNA fragments regardless of the histological type or driver gene mutation. However, ctDNA fragments with a low mutant allele frequency were found to be longer than normal cfDNA fragments. In another study, longer mutant ctDNA fragments were also detected in cancer patient blood samples, but this phenomenon could not be explained at the time (Mouliere et al., 2018). The findings presented herein may partially explain the origin of these longer mutant ctDNA fragments.

In a previous study examining HCC plasma samples, ctDNAs with low fractional concentrations were also found to have a longer size distribution relative to the healthy controls (Jiang et al., 2015), which is similar to the observations in this study. However, the previous study only compared fragment length differences between cancer patients and healthy donors, and did not distinguish mutant ctDNA fragments from normal cfDNAs due to experimental design limitations. Taken together, these findings could suggest that early-stage tumors tend to release longer ctDNA fragments at a low-frequency, but this hypothesis requires further examination.

Mutant ctDNA fragments with a low allele frequency are hard to be accurately detected. Here, two advanced technologies to detect mutant ctDNA fragments and monitor mutant allele frequency were employed to overcome this obstacle. The cfDNA fragment sizes were accurately determined using deep-sequencing technologies, and the mutant allele frequencies were further confirmed using ddPCR. However, even these advanced technologies are susceptible to false positives.

Furthermore, the lost enrichment phenomenon of short fragments observed in this study may be related to factors such as the designed probe size (120 bp), cfDNA purification, and library construction. Moreover, the findings presented herein indicate that size selection can further improve the ctDNA detection rate and accuracy. Additionally, it would seem that when constructing a ctDNA library for early-stage cancer patients, a larger DNA fragment size (> 167 bp) should be enriched, while in later stages, enrichment of shorter DNA fragment size (< 167bp) is more beneficial.

In summary, this study demonstrates that plasma ctDNAs are generally shorter than normal cfDNAs. However, for cancer patients with a low mutant allele frequency or early tumor stage, mutant ctDNA fragments are longer than normal cfDNAs. These findings may potentially facilitate the accurate detection of cancer gene mutations when utilizing liquid biopsies, and improve the application of ctDNA detection in early cancer diagnoses.

DATA AVAILABILITY STATEMENT

FASTQ data files for this study can be found in the NCBI Sequence Read Archive (SRA) database (BioProject ID: PRJNA562379).

ETHICS STATEMENT

Our research was approved by the Medical Research Ethics Committee of Chifeng Municipal Hospital (Ethics [2018] No. 017). All patients provided informed written consent before de-identification.

AUTHOR CONTRIBUTIONS

GT, JL, and LZ designed the project and analyzed the data. XL, WTW, and JL wrote the manuscript. SL, YW, LP, YH, SY, GZ, SX, and HP collected the data. CQ, LS, and KZ did the ddPCR experiments. WWW, DY, YP, QL, and PB modified and reviewed the manuscript.

FUNDING

This research was funded by the Natural Science Foundation of China (Grant 61803151), the Chifeng Municipal Hospital In-Situ Science and Technology Plan Project (2017-KY012), the Natural Science Foundation of Hunan province (Grant 2018JJ3570) and the Project of Scientific Research Fund of Hunan Provincial Education Department (Grant 17A052).

ACKNOWLEDGMENTS

We thank Xu Chu for help adjusting the figures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00147/full#supplementary-material

REFERENCES

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., et al. (2018). Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*. 174 (5), 1293–1308.e36. doi: 10.1016/j.cell.2018.05.060

Bardelli, A., and Pantel, K. (2017). Liquid biopsies, what we do not know (Yet). *Cancer Cell* 31, 172–179. doi: 10.1016/j.ccell.2017.01.002

Diaz, L. A., and Bardelli, A. (2014). Liquid biopsies: genotyping circulating tumor DNA. *J. Clin. Oncol.* 32, 579–586. doi: 10.1200/JCO.2012.45.2011

Diehl, F., Schmidt, K., Choti, M. A., Romans, K., Goodman, S., Li, M., et al. (2008). Circulating mutant DNA to assess tumor dynamics. *Nat. Med.* 14, 985–990. doi: 10.1038/nm.1789

Donaldson, J., and Park, B. H. (2018). Circulating tumor DNA: measurement and clinical utility. *Annu. Rev. Med.* 69, 223–234. doi: 10.1146/annurev-med-041316-085721

Frenel, J. S., Carreira, S., Goodall, J., Roda, D., Perez-Lopez, R., Tunariu, N., et al. (2015). Serial next-generation sequencing of circulating cell-free DNA evaluating tumor clone response to molecularly targeted drug administration. *Clin. Cancer Res.* 21, 4586–4596. doi: 10.1158/1078-0432.CCR-15-0584

Hellwig, S., Nix, D. A., Gligorich, K. M., O'Shea, J. M., Thomas, A., Fuertes, C. L., et al. (2018). Automated size selection for short cell-free DNA fragments enriches for circulating tumor DNA and improves error correction during next generation sequencing. *PloS One* 13, e0197333. doi: 10.1371/journal.pone.0197333

Jiang, P., and Lo, Y. M. D. (2016). The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. *Trends Genet.* 32, 360–371. doi: 10.1016/j.tig.2016.03.009

Jiang, P., Chan, C. W. M., Chan, K. C. A., Cheng, S. H., Wong, J., Wong, V. W.-S., et al. (2015). Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci.* 112, E1317–E1325. doi: 10.1073/pnas.1500076112

Kimura, H. (2006). Detection of epidermal growth factor receptor mutations in serum as a predictor of the response to gefitinib in patients with non-small-cell lung cancer. *Clin. Cancer Res.* 12, 3915–3921. doi: 10.1158/1078-0432.CCR-05-2324

Lecomte, T., Berger, A., Zinzindohoué, F., Micard, S., Landi, B., Blons, H., et al. (2002). Detection of free-circulating tumor-associated DNA in plasma of colorectal cancer patients and its association with prognosis: plasma DNA in colorectal cancer patients. *Int. J. Cancer* 100, 542–548. doi: 10.1002/ijc.10526

Lo, Y. M. D., Chan, K. C. A., Sun, H., Chen, E. Z., Jiang, P., Lun, F. M. F., et al. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* 2, 61ra91–61ra91. doi: 10.1126/scitranslmed.3001720

Lun, F. M. F., Tsui, N. B. Y., Chan, K. C. A., Leung, T. Y., Lau, T. K., Charoenkwan, P., et al. (2008). Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. *Proc. Natl. Acad. Sci.* 105, 19920–19925. doi: 10.1073/pnas.0810373105

Mao, L., Hruban, R. H., Boyle, J. O., Tockman, M., and Sidransky, D. (1994). Detection of oncogene mutations in sputum precedes diagnosis of lung cancer. *Cancer Res.* 54, 1634–1637.

Mouliere, F., and Rosenfeld, N. (2015). Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc. Natl. Acad. Sci.* 112, 3178–3179. doi: 10.1073/pnas.1501321112

Mouliere, F., Robert, B., Arnau Peyrotte, E., Del Rio, M., Ychou, M., Molina, F., et al. (2011). High fragmentation characterizes tumour-derived circulating DNA. *PloS One* 6, e23418. doi: 10.1371/journal.pone.0023418

Mouliere, F., El Messaoudi, S., Gongora, C., Guedj, A.-S., Robert, B., Del Rio, M., et al. (2013). Circulating cell-Free DNA from colorectal cancer patients may reveal high KRAS or BRAF mutation load. *Transl. Oncol.* 6, 319–328. doi: 10.1593/tlo.12445

Mouliere, F., Chandrananda, D., Piskorz, A. M., Moore, E. K., Morris, J., Ahlborn, L. B., et al. (2018). Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci. Transl. Med.* 10, eaat4921. doi: 10.1126/scitranslmed.aat4921

Newman, A. M., Bratman, S. V., To, J., Wynne, J. F., Eclov, N. C. W., Modlin, L. A., et al. (2014). An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* 20, 548–554. doi: 10.1038/nm.3519

Siravegna, G., Marsoni, S., Siena, S., and Bardelli, A. (2017). Integrating liquid biopsies into the management of cancer. *Nat. Rev. Clin. Oncol.* 14, 531–548. doi: 10.1038/nrclinonc.2017.14

Stroun, M., Anker, P., Maurice, P., Lyautey, J., Lederrey, C., and Beljanski, M. (1989). Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* 46, 318–322. doi: 10.1159/000226740

Sun, K., Jiang, P., Wong, A. I. C., Cheng, Y. K. Y., Cheng, S. H., Zhang, H., et al. (2018). Size-tagged preferred ends in maternal plasma DNA shed light on the production mechanism and show utility in noninvasive prenatal testing. *Proc. Natl. Acad. Sci.* 115, E5106–E5114. doi: 10.1073/pnas.1804134115

Taly, V., Pekin, D., Benhaim, L., Kotsopoulos, S. K., Le Corre, D., Li, X., et al. (2013). Multiplex picodroplet digital PCR to detect KRAS mutations in circulating DNA from the plasma of colorectal cancer patients. *Clin. Chem.* 59, 1722–1731. doi: 10.1373/clinchem.2013.206359

Thierry, A. R., Mouliere, F., Gongora, C., Ollier, J., Robert, B., Ychou, M., et al. (2010). Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. *Nucleic Acids Res.* 38, 6159–6175. doi: 10.1093/nar/gkq421

Thierry, A. R., El Messaoudi, S., Gahan, P. B., Anker, P., and Stroun, M. (2016). Origins, structures, and functions of circulating DNA in oncology. *Cancer Metastasis Rev.* 35, 347–376. doi: 10.1007/s10555-016-9629-x

Tug, S., Helmig, S., Menke, J., Zahn, D., Kubiak, T., Schwarting, A., et al. (2014). Correlation between cell free DNA levels and medical evaluation of disease progression in systemic lupus erythematosus patients. *Cell. Immunol.* 292, 32–39. doi: 10.1016/j.cellimm.2014.08.002

Umetani, N., Giuliano, A. E., Hiramatsu, S. H., Amersi, F., Nakagawa, T., Martino, S., et al. (2006). Prediction of breast tumor progression by integrity of free circulating DNA in serum. *J. Clin. Oncol.* 24, 4270–4276. doi: 10.1200/JCO.2006.05.9493

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., et al. (2016). Fragment length of circulating tumor DNA. *PloS Genet.* 12, e1006162. doi: 10.1371/journal.pgen.1006162

US Food & Drug Administration. (2016). Premarket approval P150044 — Cobas EGFR MUTATION TEST V2. FDA http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpma/pma.cfm?id=P150044.

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., et al. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* 17, 223–238. doi: 10.1038/nrc.2017.7

Yu, S. C. Y., Chan, K. C. A., Zheng, Y. W. L., Jiang, P., Liao, G. J. W., Sun, H., et al. (2014). Size-based molecular diagnostics using plasma DNA for noninvasive prenatal testing. *Proc. Natl. Acad. Sci.* 111, 8583–8588. doi: 10.1073/pnas.1406103111

Conflict of Interest: Authors JL, WTW, YH, CQ, LS, SY, KZ, GZ, HP, QL, YP, WWW, DY, SX and GT were employed by the company Geneis (Beijing) Co. Ltd.

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

ORIGINAL RESEARCH
published: 05 March 2020
doi: 10.3389/fbioe.2020.00167

Identification and Analysis of Glioblastoma Biomarkers Based on Single Cell Sequencing

Quan Cheng[1,2†], *Jing Li*[3†], *Fan Fan*[1], *Hui Cao*[4], *Zi-Yu Dai*[1], *Ze-Yu Wang*[1] and *Song-Shan Feng*[1*]

[1] *Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha, China,* [2] *Department of Clinical Pharmacology, Xiangya Hospital, Central South University, Changsha, China,* [3] *Department of Rehabilitation, The Second Xiangya Hospital, Central South University, Changsha, China,* [4] *Department of Psychiatry, The Second People's Hospital of Hunan University of Chinese Medicine, Changsha, China*

Glioblastoma (GBM) is one of the most common and aggressive primary adult brain tumors. Tumor heterogeneity poses a great challenge to the treatment of GBM, which is determined by both heterogeneous GBM cells and a complex tumor microenvironment. Single-cell RNA sequencing (scRNA-seq) enables the transcriptomes of great deal of individual cells to be assayed in an unbiased manner and has been applied in head and neck cancer, breast cancer, blood disease, and so on. In this study, based on the scRNA-seq results of infiltrating neoplastic cells in GBM, computational methods were applied to screen core biomarkers that can distinguish the discrepancy between GBM tumor and pericarcinomatous environment. The gene expression profiles of GBM from 2343 tumor cells and 1246 periphery cells were analyzed by maximum relevance minimum redundancy (mRMR). Upon further analysis of the feature lists yielded by the mRMR method, 31 important genes were extracted that may be essential biomarkers for GBM tumor cells. Besides, an optimal classification model using a support vector machine (SVM) algorithm as the classifier was also built. Our results provided insights of GBM mechanisms and may be useful for GBM diagnosis and therapy.

Keywords: glioblastoma biomarkers, scRNA-seq, mRMR method, support vector machine, pericarcinomatous environment

OPEN ACCESS

Edited by:
Peilin Jia,
The University of Texas Health Science Center at Houston, United States

Reviewed by:
Liang Lan,
Hong Kong Baptist University, Hong Kong
Guohua Huang,
Shaoyang University, China

***Correspondence:**
Song-Shan Feng
fsssplendid@163.com

†These authors have contributed equally to this work

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

Received: 19 December 2019
Accepted: 19 February 2020
Published: 05 March 2020

Citation:
Cheng Q, Li J, Fan F, Cao H, Dai Z-Y, Wang Z-Y and Feng S-S (2020) Identification and Analysis of Glioblastoma Biomarkers Based on Single Cell Sequencing. Front. Bioeng. Biotechnol. 8:167. doi: 10.3389/fbioe.2020.00167

INTRODUCTION

Glioblastoma (GBM), with an annual incidence of 3.19 per 100,000, maintains the most common and aggressive primary adult brain tumor (Stupp et al., 2007, 2017; Chinot et al., 2014; Gilbert et al., 2014; Ostrom et al., 2016). Currently, the standard therapeutic regimen has been established, including surgical resection, followed by radiotherapy with concurrent chemotherapy (temozolomide), then followed by maintenance therapy (temozolomide for 6–12 months) (Stupp et al., 2005). However, the diffuse nature of GBMs makes it invariably recur after treatment, rendering local therapies invalid, because the migrating GBM cells outside of the neoplasm core are usually unaffected by local therapies and hence cause recurrence of GBMs (Darmanis et al., 2017). The mean disease-free survival is just over 6 months and the mean overall survival also remains gloomy, with an approximately 25% 2-year survival rate after diagnosis and a 5–10% 5-year survival rate (Stupp et al., 2005, 2017; Das and Marsden, 2013).

Tumor heterogeneity poses a great challenge to the treatment of GBM, which is determined by both heterogeneous GBM cells and a complex tumor microenvironment. It is critical important for researchers to understand how different types of GBM cells interact with neoplasm cells through profiling of different types of cell from cell population in paraneoplastic environment, as well as identifying the lineage and phenotypes (Darmanis et al., 2017). Verhaak et al. (2010) has proved bulk tumor sequencing methods were useful in generating classification schemas of GBM subtypes, but the heterogeneity of GBM was not unveiled in essence (Cancer Genome Atlas Research Network, 2008). Until recently, RNA profiling was limited to ensemble-based approaches, averaging over bulk cell populations. Therefore, the advent of single-cell RNA sequencing (scRNA-seq) enables the transcriptomes of great deal of individual cells to be assayed in an unbiased manner (Stegle et al., 2015) and has been applied in head and neck cancer (Puram et al., 2017), breast cancer (Bajikar et al., 2017), blood disease (Zhao et al., 2017), and so on. Patel et al. (2014) profiled 430 cells from five GBM patients using scRNA-seq and described inter-patient variation and molecular diversity of tumor cells within individual GBM patients. The diversities of GBM cells within tumors are responsible for cancer progression and finally result in treatment failure.

Currently, in order to improve future treatment options, an increasing number of researchers have focused on the targeted agents or genes (Liu et al., 2013; Xiao et al., 2014; Li et al., 2018). Furnari et al. (2007) have identified genetic molecular mechanisms in GBM patients: (1) dysregulation of growth factor signaling through amplification and mutational activation of receptor tyrosine kinase (RTK) genes; (2) activation of the phosphatidyl inositol 3-kinase (PI3K) pathway; and (3) deactivation of the p53 and retinoblastoma tumor suppressor pathways. Moreover, four distinct GBM subclasses, including neural, proneural (PGFRA/IDH1 events), classical (focal EGFR events), and mesenchymal (NF1 mutation and loss), were defined by gene expression studies from The Cancer Genome Atlas (TCGA) (Verhaak et al., 2010), which also found the majority of GBM neoplasms had abnormalities in the pathways (RB, TP53, and RTK) through projecting copy number and mutation data on these pathways, revealing that this is a crucial step for GBM pathogenesis. Apart from such researches focused on tumor or microenvironment, many studies analyzed the gene expression of immune cells in GBM via scRNA-seq. Muller et al. (2017) identified 66 new gene sets which can be applied as biomarkers (such as P2RY12, CD49D, and HLA-DRA) to distinguish the different lineages of the macrophage cell subsets.

In this study, based on the scRNA-seq results of infiltrating neoplastic cells in GBM, computational methods were applied to screen core biomarkers that can distinguish the discrepancy between GBM tumor and pericarcinomatous environment. The gene expression profiles of GBM from 2343 tumor cells and 1246 periphery cells were analyzed by maximum relevance minimum redundancy (mRMR) (Peng et al., 2005). Upon further analysis of the feature lists yielded by the mRMR method, 31 important genes were extracted that may be essential biomarkers for GBM tumor cells. Besides, an optimal classification model using a support vector machine (SVM) algorithm (Ding and Dubchak, 2001) as the classifier was also built.

MATERIALS AND METHODS

The Single Cell Gene Expression Profiles of Tumor and Surrounding Tissues

We download the single cell gene expression profiles of 2343 cells of tumor core and 1246 cells of surrounding tissue from Gene Expression Omnibus (GEO) with accession number of GSE84465 (Darmanis et al., 2017). 23,460 genes were measured using Illumina NextSeq 500. Within each sample, we counted the number of expressed genes, i.e., the number of genes with mapped reads. The average number of expressed genes in each sample was 2,581. Our goal is to discriminate the 2343 tumor cells (positive samples) and 1246 surrounding cells (negative samples).

The mRMR Ranking of Discriminative Genes

There have been many statistics methods for identifying the differentially expressed genes (DEGs). But these methods did not consider the relationships between genes. Usually, the number of DEGs was too large to apply as biomarker. Therefore, we adopted the information theory-based mRMR (minimal Redundancy Maximal Relevance) method (Peng et al., 2005) to overcome this problem. The mRMR method not only considers the associations between genes and samples, but also the redundancy between genes. If several genes are similar, only the most representative gene will be selected. This approach has been proven to be effective and has been widely used for many biomedical feature selection problems (Niu et al., 2013; Zhao et al., 2013; Zhou et al., 2015; Zhang et al., 2016; Liu et al., 2017), especially in single cell RNA-Seq analysis (Zhang et al., 2019). The sample size of single cell data was large and the gene expression was spare. It was easy to get too many redundant significant genes using traditional statistical based method, such as t-test. Therefore, the mRMR was suitable for analyzing single cell data to get small number of non-redundant biomarkers.

Let's describe the method mathematically. All genes, selected genes, to be selected genes can be represented as $\mathbf{\Omega}$, $\mathbf{\Omega}_s$, and $\mathbf{\Omega}_t$, respectively. The relevance of gene g from $\mathbf{\Omega}_t$ with tissue type t can be measured with mutual information (I) (Sun et al., 2012; Huang and Cai, 2013):

$$D = I(g, t). \tag{1}$$

And the redundancy R of the gene g with the selected genes in $\mathbf{\Omega}_s$ are

$$R = \frac{1}{m}\left(\sum_{g_i \in \mathbf{\Omega}_s} I(g, g_i)\right) \tag{2}$$

The goal of this algorithm is to get the gene g_j from $\mathbf{\Omega}_t$ that has maximum relevance with tissue type t and minimum redundancy with the selected genes in $\mathbf{\Omega}_s$, i.e., maximize the mRMR function

$$\max_{g_j \in \mathbf{\Omega}_t}\left[I(g_j, t) - \frac{1}{m}\left(\sum_{g_i \in \mathbf{\Omega}_s} I(g_j, g_i)\right)\right] \quad (j = 1, 2, \ldots, n) \tag{3}$$

The evaluation procedure will be continued for N rounds, and all the genes will be ranked as a list

$$S = \{g'_1, g'_2, \ldots, g'_h, \ldots, g'_N, \} \tag{4}$$

The index h reflects the trade-off between relevance with tissue type and redundancy with selected genes. The smaller index h is, the better discriminating power the gene has.

The Single Cell GBM Biomarker Optimization

Based on the top 100 mRMR genes, we constructed 100 SVM classifiers and applied an incremental feature selection (IFS) method (Jiang et al., 2013; Li et al., 2014; Shu et al., 2014; Zhang et al., 2014, 2015) to identify the optimal number of genes as biomarker. The svm function from R package e10171 was used to implement the SVM method. Each candidate gene set $S_k = \{g'_1, g'_2, \ldots, g'_k\}(1 \leq k \leq 100)$ included the top k genes in the mRMR list.

We used leave-one-out cross validation (LOOCV) (Cui et al., 2013; Yang et al., 2014) to evaluate the prediction performance of each SVM classifier. During LOOCV, all of the N samples were tested one-by-one. In each round, one sample was used for testing of the prediction model trained with all the other $N-1$ samples. After N rounds, all samples were tested one time, and the predicted tissue types were compared with the actual tissue types.

Since the positive and negative sample sizes were imbalance and Mathew's correlation coefficient (MCC) can consider both sensitivity and specificity (Huang et al., 2015), MCC was used in IFS optimization. MCC can be calculated as follows:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \tag{5}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative, respectively.

Based on the LOOCV MCC of each candidate gene set, an IFS curve can be plotted. The x-axis denoted the number of top genes that were used in the SVM classifier, and the y-axis denoted the LOOCV MCCs of the SVM classifiers. Based on the IFS curve, we can choose the right number of genes which had a good prediction performance as final biomarkers.

RESULTS AND DISCUSSION

The Discriminative Importance of Genes

We applied mRMR algorithm to evaluate the discriminative importance of features iteratively. We want to find the features that were strongly associated with samples groups and were not redundant with other selected features. Using the mRMR method, we identified the top 100 most important genes. These genes were listed in **Supplementary Table S1**.

The Optimal GBM Biomarker Genes Selected With IFS Method

After we got the top 100 mRMR genes, we still did not know how many genes should be selected. To optimize the selected biomarker genes, we adopted IFS method. Each time, we added one feature into the previous feature set and got a new feature set. Then SVM classifiers were built to predict each sample's labels during LOOCV. The IFS curve with the number of genes as x-axis and the prediction performance (LOOCV MCC) as y-axis were plotted in **Figure 1**. The peak MCC was 0.812 when 31 genes were used. These 31 genes were selected as optimal GBM biomarker genes. The 31 genes were listed in **Table 1**. The confusion matrix of the 31 genes were given in **Table 2**. The sensitivity, specificity, and accuracy were 0.948, 0.855, and 0.915, respectively.

Since the tumor tissues are usually a mixture of tumor cells and normal cells, the tumor purity may cause the misclassifications. To check this, **Figures 2A,B** showed the t-distributed stochastic neighbor embedding (t-SNE) plots of predicted GBM cells and predicted non-GBM cells, respectively. In **Figure 2A**, it can be seen that the false positive samples (red dots) and the true positive samples (black dots) were mixed and they were difficult to classify. Similarly, in **Figure 2B**, it can be seen that the false negative samples (black dots) and the true negative samples (red dots) were mixed. These t-SNE plots suggested that the GBM tissues may contain non-GBM cells and the non-GBM tissues may contain GBM cells, but most cells from the corresponding tissue were similar and the machine learning algorithm we used can get the robust single cell biomarkers even when there were tissue purity issues.

The Biological Functions of the Selected Genes

Upon analysis by the mRMR method, 31 important genes were extracted that may be essential biomarkers of GBM. We did Gene

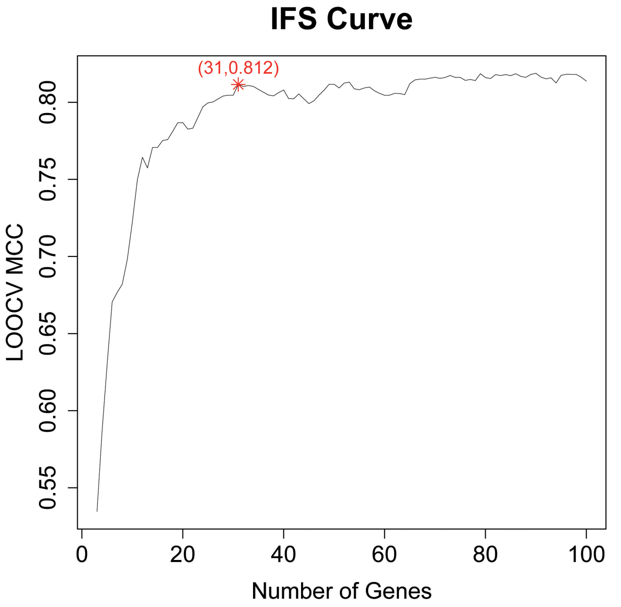


FIGURE 1 | The IFS curve of the top 100 mRMR genes. The x-axis was the number of genes and the y-axis was the prediction performance, i.e., LOOCV MCC. The peak MCC was 0.812 when 31 genes were used. These 31 genes were selected as optimal GBM biomarker genes.

TABLE 1 | The 31 selected GBM biomarker genes.

Rank	Gene	Rank	Gene
1	TMSB4X	17	VIM
2	IPCEF1	18	ATP1A2
3	MTSS1	19	RPL41
4	S100A10	20	EGR3
5	HTRA1	21	OMG
6	DHRS9	22	LDHA
7	TPI1	23	P2RY12
8	SNX22	24	SPOCK1
9	FCGBP	25	NAMPT
10	TMSB10	26	C1QL2
11	CCL3	27	PTN
12	SLC6A1	28	CCL4
13	SMOC1	29	PDZD2
14	SEC61G	30	LGALS1
15	TGFBI	31	CLDN10
16	CDR1		

TABLE 2 | The confusion matrix of the 31 selected genes.

	Predicted GBM	Predicted non-GBM
Actual GBM	2220	123
Actual non-GBM	181	1065

Ontology (GO) enrichment analysis of these 31 genes. The GO enrichment results were given in **Table 3**. It can be seen that their main function was cell adhesion and their main subcellular location was extracellular.

TABLE 3 | The GO enrichment results of the 31 selected genes.

GO term	FDR	*P*-value	Genes
GO:0007155 cell adhesion	0.0068	8.26*E*−07	EGR3, LGALS1, OMG, PTN, S100A10, CCL4, SPOCK1, TGFBI, CLDN10, MTSS1, PDZD2, P2RY12
GO:0022610 biological adhesion	0.0068	8.74*E*−07	EGR3, LGALS1, OMG, PTN, S100A10, CCL4, SPOCK1, TGFBI, CLDN10, MTSS1, PDZD2, P2RY12
GO:0031012 extracellular matrix	0.0029	1.57*E*−06	LGALS1, OMG, HTRA1, PTN, SPOCK1, TGFBI, VIM, SMOC1
GO:0005615 extracellular space	0.0107	1.56*E*−05	LGALS1, OMG, HTRA1, PTN, CCL3, CCL4, SPOCK1, TGFBI, TMSB4X, TPI1, NAMPT
GO:0005576 extracellular region	0.0107	1.87*E*−05	ATP1A2, LDHA, LGALS1, OMG, HTRA1, PTN, S100A10, CCL3, CCL4, SPOCK1, TGFBI, TMSB4X, TPI1, VIM, FCGBP, NAMPT, PDZD2, SMOC1, C1QL2
GO:0005578 proteinaceous extracellular matrix	0.0107	2.30*E*−05	LGALS1, OMG, PTN, SPOCK1, TGFBI, SMOC1
GO:0044421 extracellular region part	0.0108	2.89*E*−05	ATP1A2, LDHA, LGALS1, OMG, HTRA1, PTN, S100A10, CCL3, CCL4, SPOCK1, TGFBI, TMSB4X, TPI1, VIM, FCGBP, NAMPT, SMOC1

We compared the 31 genes with reported GBM signatures in GeneSigDB (Culhane et al., 2012) and found that the 31 genes were significantly overlapped with a signature called "Human

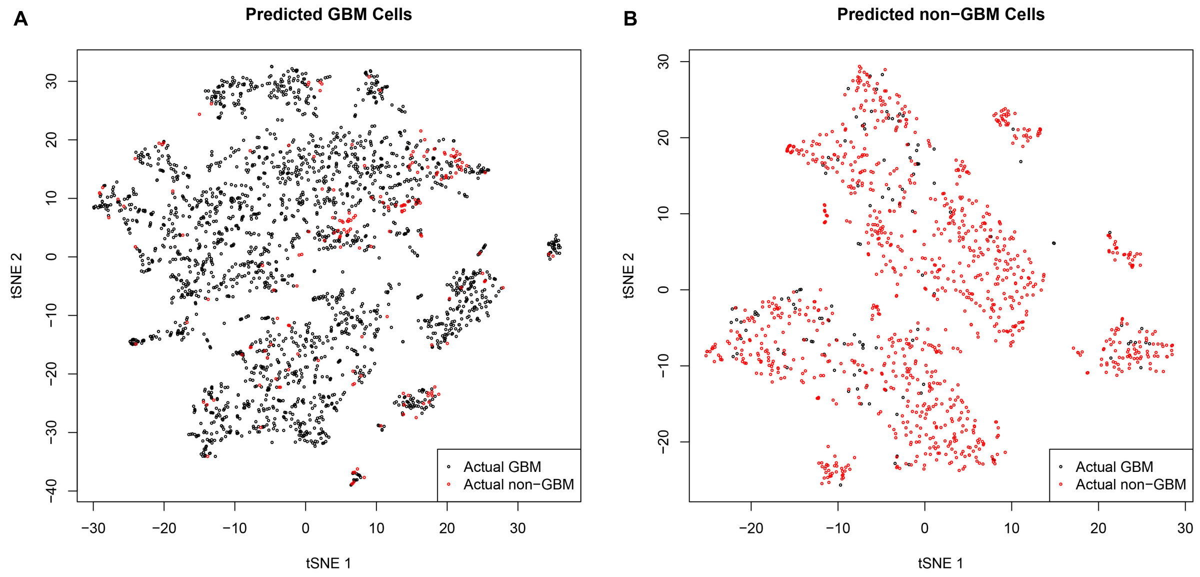


FIGURE 2 | The *t*-SNE plots of predicted GBM cells and predicted non-GBM cells. **(A)** The *t*-SNE plots of predicted GBM cells. It can be seen that the false positive samples (red dots) and the true positive samples (black dots) were mixed and they were difficult to classify. **(B)** The *t*-SNE plots of predicted non-GBM cells. It can be seen that the false negative samples (black dots) and the true negative samples (red dots) were mixed. These *t*-SNE plots suggested that the GBM tissues may contain non-GBM cells and the non-GBM tissues may contain GBM cells, but most cells from the corresponding tissue were similar and the machine learning algorithm we used can get the robust single cell biomarkers even when there were tissue purity issues.

Glioblastoma_Morandi08_22genes" which were from Table 1 of Morandi et al. (2008): the 22 up-regulated genes following camptothecin (CPT) treatment in both U87-MG and DBTRG-05 cells. The hypergeometric test *p*-value was 0.0157.

Among the 31 genes, several of them plays roles in tumor metastasis. Thymosin β4 (TMSB4X/Tβ4) is associated with tumor metastasis and progression which plays a role in cell proliferation, migration, and differentiation through a TGFβ/MRTF Signaling Axis (Morita and Hayashi, 2018). TMSB4X expression was associated with cancers in a stage- and histology-specific manner and could be an effective prognostic parameter and prognostic index. Thus far, the relationship between TMSB4X and GBM remain unknown. IPCEF1 is the C-terminal half of CNK3 which is required for HGF-dependent Arf6 activation and migration during cancer metastasis (Attar et al., 2012). MTSS1 plays an important role in cancer metastasis. Previous researches indicated that MTSS1 as a potential tumor biomarker and its reduced expression associated with bad prognosis in many cancers. In GBM, MTSS1was reported as a potential tumor suppressor and prognostic biomarker which could suppress cell migration and invasion (Zhang and Qi, 2015).

Several genes can facilitate cancer progression. S100A10 is a calcium binding protein which is found to be significantly correlated with poor survival in patients with gliomas (Sethi et al., 2012). S100A10 has been involved in cancer progression, but the unique function is not well understood (O'Connell et al., 2010). HTRA1 encodes a ubiquitously expressed serine protease with prominent expression in the vasculature. Inhibition of HTRA1 could deregulate angiogenesis in the tumor stroma which plays an important role in tumor progression (Chien et al., 2006; He et al., 2010; Klose et al., 2018).

There are several other reported tumor genes. DHRS9 is a member of the short-chain dehydrogenases/reductases (SDR) family. Recent research found that SDR family members have been involved in tumors (Hu et al., 2016). TPI1 encodes an enzyme, consisting of two identical proteins, which catalyzes the isomerization of glyceraldehydes-3-phosphate (G3P) and dihydroxy-acetone phosphate (DHAP) in glycolysis and gluconeogenesis. TPI1 was down-regulated in response to LLL12 treatment and validated using immunoblot (Jain et al., 2015). It may serve as potential therapeutic targets in GBM (Jain et al., 2015).

CONCLUSION

Glioblastoma is the most aggressive and incurable primary brain cancer in adults. The most common survival time after diagnosis is 12–15 months, with 5-year survival rate <5%. Symptoms of GBM are non-specific at early stage and the cause of GBM remains elusive. We analysis the data from 2343 tumor cells and 1246 periphery cells using mRMR and IFS method to characterize infiltrating tumor cells, and to define the cellular diversity.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84465.

AUTHOR CONTRIBUTIONS

S-SF and QC conceived and designed the study. QC, JL, Z-YD, and S-SF performed the data mining and statistical analyses. FF, HC, and Z-YW prepared the figures and tables. QC and JL drafted the initial manuscript. S-SF made critical comments and revision for the initial manuscript. S-SF, QC, and JL had primary responsibility for the final content. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 81703622), the China Postdoctoral Science Foundation (No. 2018M633002), the Hunan Provincial Natural Science Foundation of China (No. 2018JJ3838), and the Hunan Provincial Health Committee Foundation of China (No. C2019186).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00167/full#supplementary-material

TABLE S1 | The top 100 mRMR genes.

REFERENCES

Attar, M. A., Salem, J. C., Pursel, H. S., and Santy, L. C. (2012). CNK3 and IPCEF1 produce a single protein that is required for HGF dependent Arf6 activation and migration. *Exp. Cell Res.* 318, 228–237. doi: 10.1016/j.yexcr.2011.10.018

Bajikar, S. S., Wang, C. C., Borten, M. A., Pereira, E. J., Atkins, K. A., and Janes, K. A. (2017). Tumor-suppressor inactivation of GDF11 occurs by precursor sequestration in triple-negative breast cancer. *Dev. Cell* 43, 418–435.e13. doi: 10.1016/j.devcel.2017.10.027

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi: 10.1038/nature07385

Chien, J., Aletti, G., Baldi, A., Catalano, V., Muretto, P., Keeney, G. L., et al. (2006). Serine protease HtrA1 modulates chemotherapy-induced cytotoxicity. *J. Clin. Invest.* 116, 1994–2004. doi: 10.1172/JCI27698

Chinot, O. L., Wick, W., Mason, W., Henriksson, R., Saran, F., Nishikawa, R., et al. (2014). Bevacizumab plus radiotherapy-temozolomide for newly diagnosed glioblastoma. *N. Engl. J. Med.* 370, 709–722. doi: 10.1056/NEJMoa1308345

Cui, W., Chen, L., Huang, T., Gao, Q., Jiang, M., Zhang, N., et al. (2013). Computationally identifying virulence factors based on KEGG pathways. *Mol. Biosyst.* 9, 1447–1452. doi: 10.1039/c3mb70024k

Culhane, A. C., Schröder, M. S., Sultana, R., Picard, S. C., Martinelli, E. N., Kelly, C., et al. (2012). GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Res.* 40, D1060–D1066. doi: 10.1093/nar/gkr901

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., et al. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* 21, 1399–1410. doi: 10.1016/j.celrep.2017.10.030

Das, S., and Marsden, P. A. (2013). Angiogenesis in glioblastoma. *N. Engl. J. Med.* 369, 1561–1563. doi: 10.1056/NEJMcibr1309402

Ding, C. H., and Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358. doi: 10.1093/bioinformatics/17.4.349

Furnari, F. B., Fenton, T., Bachoo, R. M., Mukasa, A., Stommel, J. M., Stegh, A., et al. (2007). Malignant astrocytic glioma: genetics, biology, and paths to treatment. *Genes Dev.* 21, 2683–2710. doi: 10.1101/gad.1596707

Gilbert, M. R., Dignam, J. J., Armstrong, T. S., Wefel, J. S., Blumenthal, D. T., Vogelbaum, M. A., et al. (2014). A randomized trial of bevacizumab for newly diagnosed glioblastoma. *N. Engl. J. Med.* 370, 699–708. doi: 10.1056/NEJMoa1308573

He, X., Ota, T., Liu, P., Su, C., Chien, J., and Shridhar, V. (2010). Downregulation of HtrA1 promotes resistance to anoikis and peritoneal dissemination of ovarian cancer cells. *Cancer Res.* 70, 3109–3118. doi: 10.1158/0008-5472.CAN-09-3557

Hu, L., Chen, H. Y., Han, T., Yang, G. Z., Feng, D., Qi, C. Y., et al. (2016). Downregulation of DHRS9 expression in colorectal cancer tissues and its prognostic significance. *Tumour Biol.* 37, 837–845. doi: 10.1007/s13277-015-3880-6

Huang, T., and Cai, Y.-D. (2013). An information-theoretic machine learning approach to expression QTL analysis. *PLoS One* 8:e67899. doi: 10.1371/journal.pone.0067899

Huang, T., Wang, M., and Cai, Y.-D. (2015). Analysis of the preferences for splice codes across tissues. *Protein Cell* 6, 904–907. doi: 10.1007/s13238-015-0226-5

Jain, R., Kulkarni, P., Dhali, S., Rapole, S., and Srivastava, S. (2015). Quantitative proteomic analysis of global effect of LLL12 on U87 cell's proteome: an insight into the molecular mechanism of LLL12. *J. Proteomics* 113, 127–142. doi: 10.1016/j.jprot.2014.09.020

Jiang, Y., Huang, T., Chen, L., Gao, Y. F., Cai, Y., and Chou, K. C. (2013). Signal propagation in protein interaction network during colorectal cancer progression. *Biomed Res. Int.* 2013:287019. doi: 10.1155/2013/287019

Klose, R., Adam, M. G., Weis, E. M., Moll, I., Wustehube-Lausch, J., Tetzlaff, F., et al. (2018). Inactivation of the serine protease HTRA1 inhibits tumor growth by deregulating angiogenesis. *Oncogene* 37, 4260–4272. doi: 10.1038/s41388-018-0258-4

Li, B. Q., You, J., Huang, T., and Cai, Y. D. (2014). Classification of non-small cell lung cancer based on copy number alterations. *PLoS One* 9:e88300. doi: 10.1371/journal.pone.0088300

Li, Z., Guo, J., Ma, Y., Zhang, L., and Lin, Z. (2018). Oncogenic role of MicroRNA-30b-5p in glioblastoma through targeting proline-rich transmembrane protein 2. *Oncol. Res.* 26, 219–230. doi: 10.3727/096504017x14944585873659

Liu, J., Albrecht, A. M., Ni, X., Yang, J., and Li, M. (2013). Glioblastoma tumor initiating cells: therapeutic strategies targeting apoptosis and microRNA pathways. *Curr. Mol. Med.* 13, 352–357. doi: 10.2174/156652413805076830

Liu, L., Chen, L., Zhang, Y. H., Wei, L., Cheng, S., Kong, X., et al. (2017). Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection. *J. Biomol. Struct. Dyn.* 35, 312–329. doi: 10.1080/07391102.2016.1138142

Morandi, E., Severini, C., Quercioli, D., D'Ario, G., Perdichizzi, S., Capri, M., et al. (2008). Gene expression time-series analysis of camptothecin effects in U87-MG and DBTRG-05 glioblastoma cell lines. *Mol. Cancer* 7:66. doi: 10.1186/1476-4598-7-66

Morita, T., and Hayashi, K. (2018). Tumor progression is mediated by thymosin-beta4 through a TGFbeta/MRTF signaling axis. *Mol. Cancer Res.* 16, 880–893. doi: 10.1158/1541-7786.MCR-17-0715

Muller, S., Kohanbash, G., Liu, S. J., Alvarado, B., Carrera, D., Bhaduri, A., et al. (2017). Single-cell profiling of human gliomas reveals macrophage ontogeny as a basis for regional differences in macrophage activation in the tumor microenvironment. *Genome Biol.* 18:234. doi: 10.1186/s13059-017-1362-4

Niu, B., Huang, G., Zheng, L., Wang, X., Chen, F., Zhang, Y., et al. (2013). Prediction of substrate-enzyme-product interaction based on molecular descriptors and physicochemical properties. *BioMed Res. Int.* 2013:674215. doi: 10.1155/2013/674215

O'Connell, P. A., Surette, A. P., Liwski, R. S., Svenningsson, P., and Waisman, D. M. (2010). S100A10 regulates plasminogen-dependent macrophage invasion. *Blood* 116, 1136–1146. doi: 10.1182/blood-2010-01-264754

Ostrom, Q. T., Gittleman, H., Xu, J., Kromer, C., Wolinsky, Y., Kruchko, C., et al. (2016). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2009-2013. *Neuro Oncol.* 18(Suppl. 5), v1–v75. doi: 10.1093/neuonc/now207

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* 344, 1396–1401. doi: 10.1126/science.1254257

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1226–1238. doi: 10.1109/tpami.2005.159

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624.e24. doi: 10.1016/j.cell.2017.10.044

Sethi, M. K., Buettner, F. F., Ashikov, A., Krylov, V. B., Takeuchi, H., Nifantiev, N. E., et al. (2012). Molecular cloning of a xylosyltransferase that transfers the second xylose to O-glucosylated epidermal growth factor repeats of notch. *J. Biol. Chem.* 287, 2739–2748. doi: 10.1074/jbc.M111.302406

Shu, Y., Zhang, N., Kong, X., Huang, T., and Cai, Y. D. (2014). Predicting A-to-I RNA editing by feature selection and random forest. *PLoS One* 9:e110607. doi: 10.1371/journal.pone.0110607

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Stupp, R., Hegi, M. E., Gilbert, M. R., and Chakravarti, A. (2007). Chemoradiotherapy in malignant glioma: standard of care and future directions. *J. Clin. Oncol.* 25, 4127–4136. doi: 10.1200/JCO.2007.11.8554

Stupp, R., Mason, W. P., van den Bent, M. J., Weller, M., Fisher, B., Taphoorn, M. J., et al. (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* 352, 987–996. doi: 10.1056/NEJMoa043330

Stupp, R., Taillibert, S., Kanner, A., Read, W., Steinberg, D. M., Lhermitte, B., et al. (2017). Effect of tumor-treating fields plus maintenance temozolomide vs maintenance temozolomide alone on survival in patients with glioblastoma: a randomized clinical trial. *JAMA* 318, 2306–2316. doi: 10.1001/jama.2017.18718

Sun, L., Yu, Y., Huang, T., An, P., Yu, D., Yu, Z., et al. (2012). Associations between ionomic profile and metabolic abnormalities in human population. *PLoS One* 7:e38845. doi: 10.1371/journal.pone.0038845

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Xiao, S., Yang, Z., Lv, R., Zhao, J., Wu, M., Liao, Y., et al. (2014). miR-135b contributes to the radioresistance by targeting GSK3beta in human glioblastoma multiforme cells. *PLoS One* 9:e108810. doi: 10.1371/journal.pone.0108810

Yang, J., Chen, L., Kong, X., Huang, T., and Cai, Y. D. (2014). Analysis of tumor suppressor genes based on gene ontology and the KEGG pathway. *PLoS One* 9:e107202. doi: 10.1371/journal.pone.0107202

Zhang, G.-L., Pan, L.-L., Huang, T., and Wang, J.-H. (2019). The transcriptome difference between colorectal tumor and normal tissues revealed by single-cell sequencing. *J. Cancer* 10, 5883–5890. doi: 10.7150/jca.32267

Zhang, N., Huang, T., and Cai, Y. D. (2014). Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. *Mol. Genet. Genomics* 290, 343–352. doi: 10.1007/s00438-014-0922-5

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860(11 Part B), 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zhang, P. W., Chen, L., Huang, T., Zhang, N., Kong, X. Y., and Cai, Y. D. (2015). Classifying ten types of major cancers based on reverse phase protein array profiles. *PLoS One* 10:e0123147. doi: 10.1371/journal.pone.0123147

Zhang, S., and Qi, Q. (2015). MTSS1 suppresses cell migration and invasion by targeting CTTN in glioblastoma. *J. Neurooncol.* 121, 425–431. doi: 10.1007/s11060-014-1656-2

Zhao, T. H., Jiang, M., Huang, T., Li, B. Q., Zhang, N., Li, H. P., et al. (2013). A novel method of predicting protein disordered regions based on sequence features. *BioMed Res. Int.* 2013:414327. doi: 10.1155/2013/414327

Zhao, X., Gao, S., Wu, Z., Kajigaya, S., Feng, X., Liu, Q., et al. (2017). Single-cell RNA-seq reveals a distinct transcriptome signature of aneuploid hematopoietic cells. *Blood* 130, 2762–2773. doi: 10.1182/blood-2017-08-803353

Zhou, Y., Zhang, N., Li, B. Q., Huang, T., Cai, Y. D., and Kong, X. Y. (2015). A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. *J. Biomol. Struct. Dyn.* 33, 2479–2490. doi: 10.1080/07391102.2014.1001793

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 12 March 2020
doi: 10.3389/fbioe.2020.00187

The Joint Analysis of Multi-Omics Data Revealed the Methylation-Expression Regulations in Atrial Fibrillation

Ban Liu[1]†, Xin Shi[2]†, Keke Ding[3]†, Mengwei Lv[4,5], Yongjun Qian[6], Shijie Zhu[7], Changfa Guo[7]* and Yangyang Zhang[5]*

[1] Department of Cardiology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [2] Department of Pediatric Cardiovascular, Xin Hua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [3] Department of Cardiology, Shanghai Tongji Hospital, Tongji University School of Medicine, Shanghai, China, [4] Shanghai East Hospital of Clinical Medical College, Nanjing Medical University, Shanghai, China, [5] Department of Cardiovascular Surgery, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China, [6] Department of Cardiovascular Surgery, National Clinical Research Center for Geriatric, West China Hospital, Sichuan University, Chengdu, China, [7] Department of Cardiovascular Surgery, Zhongshan Hospital, Fudan University, Shanghai, China

Atrial fibrillation (AF) is one of the most prevalent heart rhythm disorder. The causes of AF include age, male sex, diabetes, hypertension, valve disease, and systolic/diastolic dysfunction. But on molecular level, its mechanisms are largely unknown. In this study, we collected 10 patients with persistent atrial fibrillation, 10 patients with paroxymal atrial fibrillation and 10 healthy individuals and did Methylation EPICBead Chip and RNA sequencing. By analyzing the methylation and gene expression data using machine learning based feature selection method Boruta, we identified the key genes that were strongly associated with AF and found their interconnections. The results suggested that the methylation of KIF15 may regulate the expression of *PSMC3*, *TINAG*, and *NUDT6*. The identified AF associated methylation-expression regulations may help understand the molecular mechanisms of AF from a multi-omics perspective.

Keywords: atrial fibrillation, methylation, multi-omics, feature selection, classification

OPEN ACCESS

Edited by:
Peilin Jia,
The University of Texas Health Science Center at Houston (UTHealth), United States

Reviewed by:
Yan Yao,
Chinese Academy of Medical Sciences, China
Jian Zhong Zhou,
The First Affiliated Hospital of Chongqing Medical University, China

***Correspondence:**
Changfa Guo
guo.changfa@zs-hospital.sh.cn
Yangyang Zhang
zhangyangyang_wy@vip.sina.com;
zhangyangyang_wy@tongji.edu.cn

†These authors have contributed equally to this work

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

Received: 19 December 2019
Accepted: 26 February 2020
Published: 12 March 2020

Citation:
Liu B, Shi X, Ding K, Lv M, Qian Y, Zhu S, Guo C and Zhang Y (2020) The Joint Analysis of Multi-Omics Data Revealed the Methylation-Expression Regulations in Atrial Fibrillation. Front. Bioeng. Biotechnol. 8:187. doi: 10.3389/fbioe.2020.00187

INTRODUCTION

Atrial fibrillation (AF), one of the most prevalent heart rhythm disorders, is a potential cause of heart failure and ischemic stroke with high morbidity and mortality (Ogawa et al., 2017; Asmarats et al., 2019). The cause of AF is multifactorial which include age, male sex, diabetes, hypertension, valve disease, and systolic/diastolic dysfunction (Schnabel et al., 2009; Lip et al., 2013; Voukalis et al., 2016). Depends on how often atrial fibrillation occurs and how it responds to treatment, AF is roughly divided into two major subtypes-paroxysmal atrial fibrillation (PAF) and persistent atrial fibrillation (PeAF). In the treatment of AF, drugs were the first choice, non-drug therapies were used only when drug therapy failed or patients could not tolerate the medication. In contrast to the extensive knowledge of etiology, the underlying mechanism of AF remains elusive. Further study of the potential mechanisms of AF could provide novel strategies for the treatment and management to increase quality of life and reduce economic burden of social (Chugh et al., 2001).

With the development of next generation sequencing (NGS) technologies, growing evidence have demonstrated that AF is a disease with a significant genetic contribution. Previous studies have filtered novel genetic variants and candidate genes including transcriptional factor genes (*PITX2, PRRX1, ZFHX3, NKX2.5*, *TBX5*), ion channel genes (*KCNN3*, *HCN4*, *CACNA1C*, *SCN5A*, *KCNQ1*, *KCNH2*), and caveolin genes (*CAV1* and *CAV2*) (Ellinor et al., 2010; Olesen et al., 2014; Sinner et al., 2014; Ma et al., 2016; Nielsen et al., 2018). However, these genes explain only a small fraction of the biology and genetic underpinnings of AF.

Epidemiological studies have demonstrated that genetic, environmental, behavioral, and clinical factors contribute to AF pathogenic mechanism (Zhong et al., 2016). Emelia J. B performed genome-wide methylation using whole blood samples from 183 prevalent AF and 220 incident AF cases. They examined the association between DNA methylation and GWAS loci, suggesting DNA methylation might be a possible mechanism through which AF-specific genetic variations affect gene regulation (Lin et al., 2017). To date, only a few studies have investigated differential DNA methylation as a predictor biomarker at specific candidate loci that were previously associated with AF.

Therefore, we applied DNA methylation profiling study to identify the likely rare damaging variants and putative candidate genes from 10 patients with persistent atrial fibrillation (PeAF), 10 patients with paroxymal atrial fibrillation (PAF) and 10 healthy individuals. Interestingly, we identified top 10 genes (KIF15, ABCA3, FOXG1, VGF, PDE4D, EIF3C, CNTNAP5, SHOX2, VGF, TRIM59) as functional candidate genes and the expression level are significantly increased in PeAF and PAF patients than control. Given the importance of DNA methylation to gene expression, we investigated the gene expression of the same participants using RNA sequencing. We also defined top 10 genes (*EPN3*, *EMD*, *SMCO4*, *F2RL2*, *TMED1*, *PSMC3*, *PDZD11*, *NUDT6*, *TINAG*, *GALNT5*) in gene expression data and the expression pattern of these genes was significantly different between PeAF and PAF. These results have improved our understanding of the underlying mechanism and offer new insights into the potential pathway of AF, which could provide novel therapeutic option for this disease.

MATERIALS AND METHODS

Atrial Fibrillation Patients

Ten patients with paroxymal atrial fibrillation (g1), 10 patients with persistent atrial fibrillation (continuous atrial fibrillation lasting more than 12 months) (g2) and 10 healthy individuals (g3) were enrolled in this study (**Table 1**). All patients were subjected to detailed medical evaluation, which included medical history, physical examination, electrocardiography (ECG), and echocardiography. Patients with chronic heart failure, coronary heart disease, cardiomyopathy, hyperthyroidism or chronic pulmonary heart disease were excluded.

The study was conducted in accordance with the Declaration of Helsinki, and the protocol used to collect human heart tissue was approved by the Ethics Committee of Shanghai East Hospital (DI: 0402017).

Written informed consents to participate in this study were provided by all the enrolled patients before operation of fibrillation ablation. The left atrial appendage tissues which were abandoned during isolated surgical ablation were collected. Normal left atrial appendages were collected from healthy male donors.

The Methylation Profiles

The DNA methylation status of 850K probes in the 30 samples was measured using Methylation EPICBead Chip. The raw data was quality controlled and preprocessed using R/Bioconductor package minfi[1] (Aryee et al., 2014). The beta value ranged from 0 to 1 was calculated to represent how each position was methylated. 1 meant high methylation and 0 meant low methylation.

The RNA Sequencing Profiles

The total RNAs were extracted using RNeasy Mini Kit (Cat#74106, Qiagen) and the RNA integrity was checked using Agilent Bioanalyzer 2100 (Agilent technologies, Santa Clara, CA, United States). Qualified total RNA was further purified by RNAClean XP Kit (Cat A63987, Beckman Coulter Inc., Kraemer Boulevard, Brea, CA, United States) and RNase-Free DNase Set (Cat#79254, QIAGEN, GmBH, Germany). Pair-end sequencing reads were generated using Illumina data collection software. First, the reads were mapped onto human reference genome GRCh38 using Hisat2 (version:2.0.4[2]) (Kim et al., 2015). Then, Stringtie (version:1.3.0[3]) (Pertea et al., 2015) was used to calculate the FPKM (Fragments Per Kilobase of exon model per Million mapped reads).

Feature Selection Algorithm

There were 866,091 methylation probes and 50,868 RNAs. The number of features were extremely large. It was difficult to select key features using traditional statistical methods. Therefore, we adopted the latest machine learning based feature selection method Boruta to get the key methylation probes and RNAs.

Boruta (Kursa and Rudnicki, 2010) is a feature selection method based on random forest. It can select sample group relevant features effectively. First, it will shuffle the features to create many permuted datasets. Then, it will evaluate the importance score of each feature in the original actual dataset and permuted datasets. Then, it will compare the actual importance score with permuted scores and find the features with significantly higher actual importance scores than permuted scores. After multiple iterations, it will select all the sample group relevant features. The python code from https://github.com/scikit-learn-contrib/boruta_py was used to apply the Boruta feature selection algorithm.

[1] https://bioconductor.org/packages/minfi/

[2] http://ccb.jhu.edu/software/hisat2/

[3] http://ccb.jhu.edu/software/stringtie/

TABLE 1 | Demographic characteristics of AF patients.

No.	Age (years)	Gender	Weight (Kg)	Height (cm)	Smoking	Hypertension	Diabetes mellitus	Coronary angiography or CTA	LVEF (%)	Left atrial diameter (mm)	Duration of AF (years)
1	69	Male	76	169	No	Yes	No	Negative	70	40	–
2	63	Male	64	170	No	No	No	Negative	59	46	–
3	63	Male	70	170	No	No	No	Negative	66	39	–
4	69	Male	67	173	No	Yes	No	Negative	67	46	–
5	69	Male	75	165	No	No	No	Negative	70	36	–
6	61	Male	76	176	No	Yes	Yes	Negative	60	42	–
7	64	Male	52	168	Yes	No	No	Negative	64	40	–
8	64	Male	71	181	No	Yes	Yes	Negative	63	39	–
9	61	Male	87	167	Yes	Yes	No	Negative	62	37	–
10	66	Male	82	173	No	Yes	No	Negative	63	42	–
11	63	Male	86	176	No	Yes	No	Negative	57	46	2.5
12	63	Male	80	178	No	No	No	Negative	68	55	3
13	64	Male	70	170	No	No	No	Negative	67	41	4
14	64	Male	84	164	No	Yes	No	Negative	55	48	2
15	65	Male	73	169	No	Yes	No	Negative	69	55	3.5
16	66	Male	66	168	No	Yes	No	Negative	64	45	4
17	67	Male	80	175	No	Yes	Yes	Negative	59	47	2.5
18	67	Male	73	165	Yes	Yes	No	Negative	59	47	3
19	63	Male	61	164	No	No	No	Negative	73	49	2
20	67	Male	90	178	No	Yes	No	Negative	70	58	2.5

LVEF, left ventricular ejection fraction; AF, atrial fibrillation; CTA, CT angiography.

Classification Predictor

To evaluate how well the selected features can classify the samples, we built an SVM (Support Vector Machine) classifier using the methylation data and another RNA-Seq data based SVM classifier. The svm function in R package e10171[4] (Chang and Lin, 2011) was used to apply the SVM classification algorithm.

LOOCV (leave-one-out cross validation) was used to objectively evaluate the classification performance. Each time, one sample was treated as test sample while all the other samples were used to train the model. After 30 rounds, all samples had been tested once and the overall accuracy was calculated based on the confusion matrix. In confusion matrix, the actual sample groups were compared with predicted sample groups.

RESULTS

The Key Methylation Features Identified With Boruta

We ran Boruta feature selection algorithm on the methylation data and got 10 key methylation features. These 10 key methylation features were listed in **Table 2**. The probes were annotated to genome positions (Genome Build 37) and genes using the official annotation file from Illumina. Sometime, one position may be associated with multiple genes. Therefore, the 10 methylation probes can be mapped onto 15 gene symbols. We checked the GO annotation of these genes and found that cg16703882 (*SHOX2*) was associated with GO:0007507: heart development which was closely relevant to AF.

We plotted the heatmap of these 10 key methylation features in **Figure 1**. It can be seen that most of the 10 patients with paroxymal atrial fibrillation (g1), 10 patients with persistent atrial fibrillation (g2) and 10 healthy individuals (g3) were cluster into the right groups.

The Key Gene Expression Features Identified With Boruta

Similarly, we ran Boruta feature selection algorithm on the RNA-Seq gene expression data and got 10 key gene expression features.

TABLE 2 | The 10 key methylation features identified by Boruta.

ILMNID	Chromosome	Position	Strand	UCSC Ref gene name
cg00702638	3	44803293	R	KIF15; KIAA1143
cg02331561	16	2391081	F	ABCA17P; ABCA3
cg02991338	14	29236017	R	FOXG1
cg04084157	7	100809049	F	VGF
cg05995159	5	59325256	R	PDE4D
cg06357615	16	28403195	R	MIR6862-2; MIR6862-1; EIF3CL; EIF3C
cg11344566	2	124782885	F	CNTNAP5
cg16703882	3	157823479	R	SHOX2
cg21186299	7	100808810	R	VGF
cg26856080	3	160167746	R	TRIM59

[4] https://CRAN.R-project.org/package=e1071

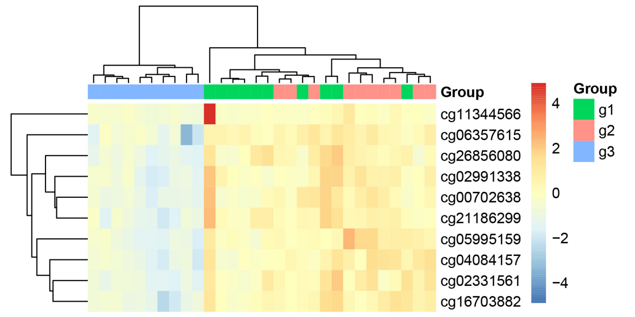


FIGURE 1 | The heatmap of the key methylation features. The 30 samples were from three groups: 10 patients with paroxymal atrial fibrillation (g1), 10 patients with persistent atrial fibrillation (g2), and 10 healthy individuals (g3). It can be seen that most samples were cluster into the right groups.

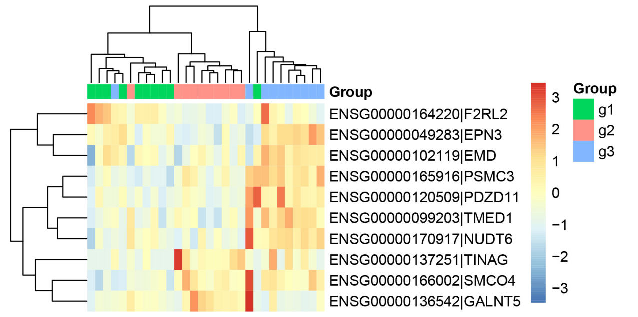


FIGURE 2 | The heatmap of the key gene expression features. The 30 samples were from three groups: 10 patients with paroxymal atrial fibrillation (g1), 10 patients with persistent atrial fibrillation (g2), and 10 healthy individuals (g3). It can be seen that most samples were cluster into the right groups.

The 10 key gene expression features were given in **Table 3**. We also plotted their heatmap in **Figure 2**. The clusters were also largely correct.

The Classification Performance of Methylation and Gene Expression

From **Figures 1**, **2**, we can see that both methylation and gene expression features can correctly cluster most samples. But we would like to evaluate their performance objectively and quantitatively. Therefore, we applied LOOCV to test the SVM classifiers of methylation and gene expression features. The confusion matrixes of methylation features and gene expression features were listed in **Tables 4**, **5**, respectively.

From **Table 3**, we can see that the AF patients (g1 + g2) and healthy individuals (g3) were perfectly classified using the methylation features, but the methylation data did not have great performance on classifying the subtype of AF (g1 and g2). From **Table 4**, we can see that the two subtype of AF, paroxysmal atrial fibrillation (g1) and persistent atrial fibrillation (g2), had very different gene expression pattern. In other words, the methylation data and gene expression data complement each other. The methylation data can be used to predict the AF and the gene expression data can be used to classify the subtypes of AF.

TABLE 4 | The confusion matrix of key methylation features.

	Predicted g1	Predicted g2	Predicted g3
Actual g1	8	2	0
Actual g2	3	7	0
Actual g3	0	0	10

TABLE 5 | The confusion matrix of key gene expression features.

	Predicted g1	Predicted g2	Predicted g3
Actual g1	9	0	1
Actual g2	0	9	1
Actual g3	1	0	9

We checked the wrongly predicted samples in **Tables 3**, **4**. They were different. Within the 30 samples, 22 samples had the same predicted labels by expression and methylation. All these 22 samples were correctly predicted. For the 8 inconsistent samples between expression and methylation predictions, at least one of the two predictions (the expression-based prediction and the methylation-based prediction) was correct. In other words, all the samples can be corrected classified based on either expression or methylation. The expression-based prediction and the methylation-based prediction were complementary.

TABLE 3 | The 10 key gene expression features identified by Boruta.

Gene ID	Name	Description	GRCh38 locus
ENSG00000049283	EPN3	Epsin 3	17:50532543-50543750
ENSG00000102119	EMD	Emerin	X:154379197-154381523
ENSG00000166002	SMCO4	Single-pass membrane protein with coiled-coil domains 4	11:93478472-93543508
ENSG00000164220	F2RL2	Coagulation factor II (thrombin) receptor-like 2	5:76615482-76623434
ENSG00000099203	TMED1	Transmembrane p24 trafficking protein 1	19:10832438-10836318
ENSG00000165916	PSMC3	Proteasome 26S subunit, atpase 3	11:47418769-47426473
ENSG00000120509	PDZD11	PDZ domain containing 11	X:70286595-70290514
ENSG00000170917	NUDT6	Nudix hydrolase 6	4:122888697-122922968
ENSG00000137251	TINAG	Tubulointerstitial nephritis antigen	6:54307859-54390152
ENSG00000136542	GALNT5	Polypeptide N-acetylgalactosaminyltransferase 5	2:157257598-157314211

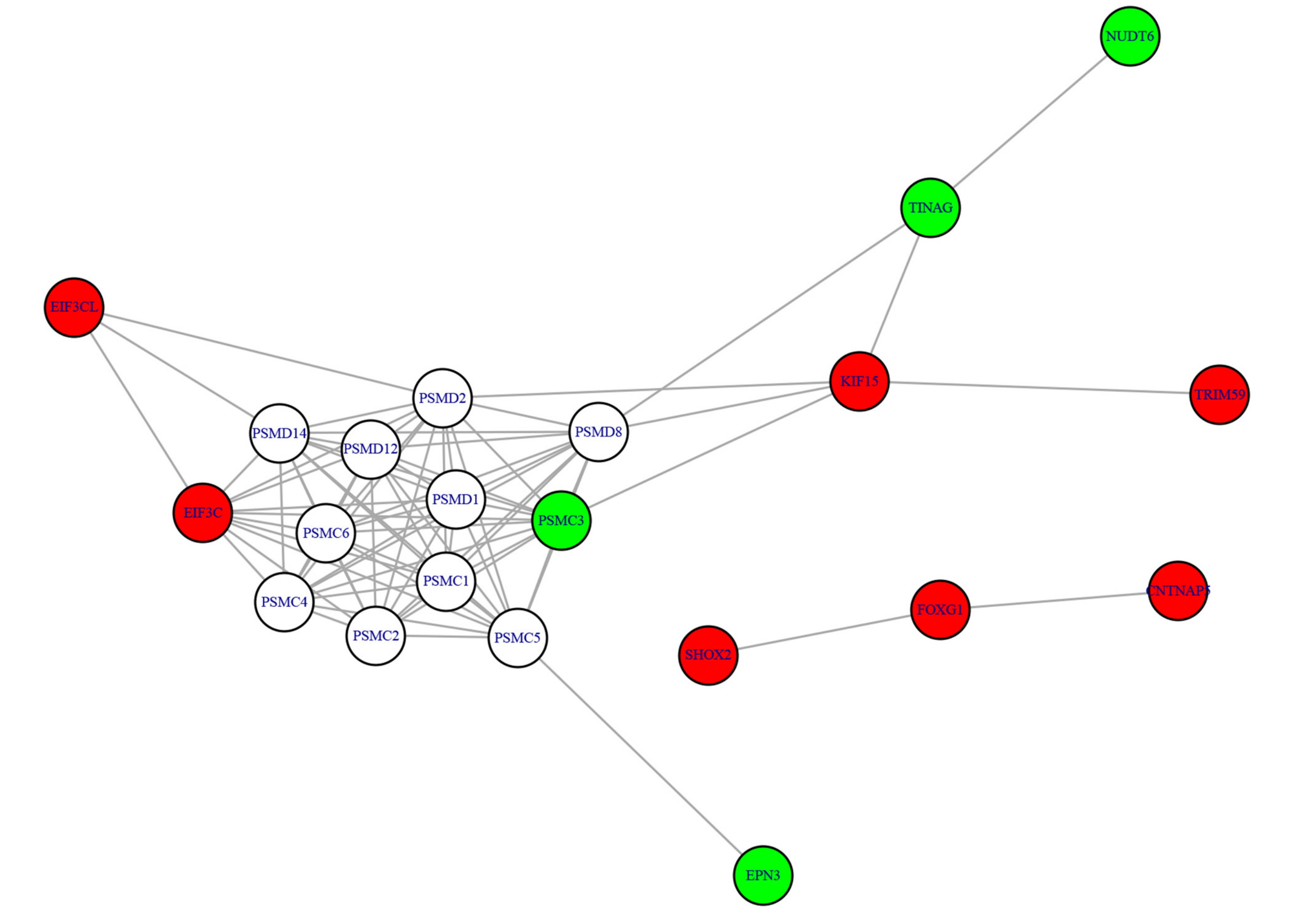


FIGURE 3 | The methylation-expression regulation network. The red and green nodes were methylation and expression genes, respectively. The methylation genes located in three clusters: EIF3CL-EIF3C, KIF15-TRIM59, and SHOX2-FOXG1-CNTNAP5. KIF15 can directly or indirectly regulate the expression genes and may play important roles.

The Methylation-Expression Regulation Network

We mapped the genes of methylation features and gene expression features to the STRING network (Version 11.0[5]) (Szklarczyk et al., 2018) and visualized the network using R package igraph (Csardi and Nepusz, 2006)[6] to identify the potential relationship between two candidate genes sets. The methylation-expression regulation network was shown in **Figure 3**. In the network, the methylation and expression genes were marked in red and green. The methylation genes located in three clusters: *EIF3CL-EIF3C*, *KIF15-TRIM59*, and *SHOX2-FOXG1-CNTNAP5*. The three expression genes (*PSMC3*, *TINAG*, and *NUDT6*) were connected with methylation gene *KIF15*. Even *EPN3* can be indirectly connected to *KIF15*. That made *KIF15* at the center of the network. These results suggested that *KIF5* may play important roles in the pathogenesis of AF.

DISCUSSION

DNA methylation, a pre-transcriptional modification characterized by the addition of methyl groups to specific nucleotides, regulates the stability of gene expression states and maintains genome integrity by collaborating with proteins that modify nucleosomes (Ma et al., 2014; Tao et al., 2016; Shen et al., 2017). Previous studies considered that changes in DNA methylation states contribute to the regulation of biological processes underlying AF, such as fibrosis, atrial dilatation, atrial fibroblast proliferation and differentiation from fibroblasts into myofibroblasts (Zhao et al., 2017). To further enhance the biological understanding of the atrial fibrillation, our study focused on DNA methylation, particularly with respect to how it relates to mRNA expression. Among our two gene sets of top 10 genes, we found PDED4, SHOX2, and EMD were the most important genes for AF which have been reported associated with AF in previous reference.

Atrial fibrillation is reported to be associated with a profound remodeling of membrane receptors and alterations in cAMP

[5] http://string-db.org

[6] https://CRAN.R-project.org/package=igraph

dependent regulation of Ca2$^+$ handling. PDE4 is expressed in human atrial myocytes and accounts for approximately 15% of PDE (phosphodiesterase) activity (Molina et al., 2012). PDE4D encoded protein has 3',5'-cyclic-AMP phosphodiesterase activity and degrades cAMP, which acts as a signal transduction molecule in multiple cell types and represents the major PDE4 subtype (Berk et al., 2016). The activity of PDEs decreased with age, and the relative PDED4 activity was lower in patients with permanent atrial fibrillation than in age-matched sinus rhythm controls (Milton et al., 2011). Previous study provided evidence that patients with pAF were found to have a decreased PDE4 activity as compared with patients in sinus rhythm (Yeh et al., 2007).

Short Stature Homeobox 2 (*SHOX2*) is a member of the homeobox family of genes in which mutations associated with early-onset and familial AF (Hoffmann et al., 2019). *SHOX2* is considered as a key regulator of sinus node development of which deficiency could lead to bradycardia in animal models (Vicente-Steijn et al., 2017). Previous study demonstrated *SHOX2* was susceptible for SND and AF by screening 98 SND patients and 450 individuals with AF. In the heart development of mouse and zebrafish, they also proved *SHOX2* plays an important role, the mutation of SHOX2 could lead to severe bradycardia (Blaschke et al., 2007; Ye et al., 2015).

Emerin (EMD) encodes a serine-rich nuclear membrane protein which located on the cytoplasmic surface of the inner nuclear membrane and related to X-linked Emery-Dreifuss muscular dystrophy (EDMD) (Capanni et al., 2009). Previous study found a nonsense mutation in EMD from two EDMD families which is associated with X-linked recessive inheritance, result in serious cardiac complication, including AF (Sakata et al., 2005). Cardiologic assessment revealed slow atrial fibrillation in a recent case of a 65-year-old male patient with a hemizygous duplication of 5 bases in exon 6 of the EMD, gene on the X chromosome (Kissel et al., 2009; Zhao et al., 2014; Brisset et al., 2019).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Shanghai East Hospital. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this manuscript.

AUTHOR CONTRIBUTIONS

YZ and CG conceived and designed the experiments. BL, XS, and KD performed the experiments. ML, YQ, and SZ analyzed the data. BL, XS, and YZ wrote the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

The present study was funded by the National Key Research and Development Program (Grant No. 2018-YFC-1312505 to YZ), the National Natural Science Foundation of China (Grant No. 81770408 to CG), and (Grant No. 81900280 to XS).

REFERENCES

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. *Bioinformatics* 30, 1363–1369. doi: 10.1093/bioinformatics/btu049

Asmarats, L., Cruz-Gonzalez, I., Nombela-Franco, L., Arzamendi, D., Peral, V., Nietlispach, F., et al. (2019). Recurrence of device-related thrombus after percutaneous left atrial appendage closure. *Circulation* 140, 1441–1443. doi: 10.1161/CIRCULATIONAHA.119.040860

Berk, E., Christ, T., Schwarz, S., Ravens, U., Knaut, M., and Kaumann, A. J. (2016). In permanent atrial fibrillation, PDE3 reduces force responses to 5-HT, but PDE3 and PDE4 do not cause the blunting of atrial arrhythmias. *Br. J. Pharmacol.* 173, 2478–2489. doi: 10.1111/bph.13525

Blaschke, R. J., Hahurij, N. D., Kuijper, S., Just, S., Wisse, L. J., Deissler, K., et al. (2007). Targeted mutation reveals essential functions of the homeodomain transcription factor Shox2 in sinoatrial and pacemaking development. *Circulation* 115, 1830–1838. doi: 10.1161/CIRCULATIONAHA.106.637819

Brisset, M., Ben Yaou, R., Carlier, R. Y., Chanut, A., Nicolas, G., Romero, N. B., et al. (2019). X-linked Emery-Dreifuss muscular dystrophy manifesting with adult onset axial weakness, camptocormia, and minimal joint contractures. *Neuromuscul. Disord.* 29, 678–683. doi: 10.1016/j.nmd.2019.06.009

Capanni, C., Del Coco, R., Mattioli, E., Camozzi, D., Columbaro, M., Schena, E., et al. (2009). Emerin-prelamin a interplay in human fibroblasts. *Biol. Cell* 101, 541–554. doi: 10.1042/BC20080175

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Trans. on Intell. Syst. Technol.* 2, 1–27. doi: 10.1145/1961189.1961199

Chugh, S. S., Blackshear, J. L., Shen, W. K., Hammill, S. C., and Gersh, B. J. (2001). Epidemiology and natural history of atrial fibrillation: clinical implications. *J. Am. Coll. Cardiol.* 37, 371–378. doi: 10.1016/s0735-1097(00)01107-4

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Inter J. Complex Syst.* 1695, 1–9.

Ellinor, P. T., Lunetta, K. L., Glazer, N. L., Pfeufer, A., Alonso, A., Chung, M. K., et al. (2010). Common variants in KCNN3 are associated with lone atrial fibrillation. *Nat. Genet.* 42, 240–244. doi: 10.1038/ng.537

Hoffmann, S., Paone, C., Sumer, S. A., Diebold, S., Weiss, B., Roeth, R., et al. (2019). Functional characterization of rare variants in the shox2 gene identified in sinus node dysfunction and atrial fibrillation. *Front. Genet.* 10:648. doi: 10.3389/fgene.2019.00648

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kissel, J. T., Dimberg, E. L., Emslie-Smith, A. M., Selcen, D., and Keegan, B. M. (2009). A 49-year-old man with contractures, weakness, and cardiac arrhythmia. *Neurology* 72, 2036–2043. doi: 10.1212/01.wnl.0000350522.42635.09

Kursa, M., and Rudnicki, W. (2010). Feature selection with the boruta package. *J. Statist. Softw. Artic.* 36, 1–13. doi: 10.18637/jss.v036.i11

Lin, H., Yin, X., Xie, Z., Lunetta, K. L., Lubitz, S. A., Larson, M. G., et al. (2017). Methylome-wide association study of atrial fibrillation in framingham heart study. *Sci. Rep.* 7:40377. doi: 10.1038/srep40377

Lip, G. Y. H., Lane, D. A., Buller, H., and Apostolakis, S. (2013). Development of a novel composite stroke and bleeding risk score in patients with atrial fibrillation: the AMADEUS study. *Chest* 144, 1839–1847. doi: 10.1378/chest.13-1635

Ma, B., Wilker, E. H., Willis-Owen, S. A., Byun, H. M., Wong, K. C., Motta, V., et al. (2014). Predicting DNA methylation level across human tissues. *Nucleic Acids Res.* 42, 3515–3528. doi: 10.1093/nar/gkt1380

Ma, J. F., Yang, F., Mahida, S. N., Zhao, L., Chen, X., Zhang, M. L., et al. (2016). TBX5 mutations contribute to early-onset atrial fibrillation in chinese and caucasians. *Cardiovasc. Res.* 109, 442–450. doi: 10.1093/cvr/cvw003

Milton, A. G., Aykanat, V. M., Hamilton-Bruce, M. A., Nezic, M., Jannes, J., and Koblar, S. A. (2011). Association of the phosphodiesterase 4D (PDE4D) gene and cardioembolic stroke in an Australian cohort. *Int. J. Stroke* 6, 480–486. doi: 10.1111/j.1747-4949.2011.00616.x

Molina, C. E., Leroy, J., Richter, W., Xie, M., Scheitrum, C., Lee, I. O., et al. (2012). Cyclic adenosine monophosphate phosphodiesterase type 4 protects against atrial arrhythmias. *J. Am. Coll. Cardiol.* 59, 2182–2190. doi: 10.1016/j.jacc.2012.01.060

Nielsen, J. B., Thorolfsdottir, R. B., Fritsche, L. G., Zhou, W., Skov, M. W., Graham, S. E., et al. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet.* 50, 1234–1239. doi: 10.1038/s41588-018-0171-3

Ogawa, H., Senoo, K., An, Y., Shantsila, A., Shantsila, E., Lane, D. A., et al. (2017). Clinical features and prognosis in patients with atrial fibrillation and prior stroke: comparing the fushimi and darlington AF registries. *eBio Med.* 18, 199–203. doi: 10.1016/j.ebiom.2017.03.022

Olesen, M. S., Nielsen, M. W., Haunso, S., and Svendsen, J. H. (2014). Atrial fibrillation: the role of common and rare genetic variants. *Eur. J. Hum. Genet.* 22, 297–306. doi: 10.1038/ejhg.2013.139

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Sakata, K., Shimizu, M., Ino, H., Yamaguchi, M., Terai, H., Fujino, N., et al. (2005). High incidence of sudden cardiac death with conduction disturbances and atrial cardiomyopathy caused by a nonsense mutation in the STA gene. *Circulation* 111, 3352–3358. doi: 10.1161/CIRCULATIONAHA.104.527184

Schnabel, R. B., Sullivan, L. M., Levy, D., Pencina, M. J., Massaro, J. M., D'Agostino, R. B., et al. (2009). Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 373, 739–745. doi: 10.1016/S0140-6736(09)60443-8

Shen, K., Tu, T., Yuan, Z., Yi, J., Zhou, Y., Liao, X., et al. (2017). DNA methylation dysregulations in valvular atrial fibrillation. *Clin. Cardiol.* 40, 686–691. doi: 10.1002/clc.22715

Sinner, M. F., Tucker, N. R., Lunetta, K. L., Ozaki, K., Smith, J. G., Trompet, S., et al. (2014). Integrating genetic, transcriptional, and functional analyses to identify 5 novel genes for atrial fibrillation. *Circulation* 130, 1225–1235. doi: 10.1161/CIRCULATIONAHA.114.009892

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). 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

Tao, H., Shi, K. H., Yang, J. J., and Li, J. (2016). Epigenetic mechanisms in atrial fibrillation: new insights and future directions. *Trends Cardiovasc. Med.* 26, 306–318. doi: 10.1016/j.tcm.2015.08.006

Vicente-Steijn, R., Kelder, T. P., Tertoolen, L. G., Wisse, L. J., Pijnappels, D. A., Poelmann, R. E., et al. (2017). RHOA-ROCK signalling is necessary for lateralization and differentiation of the developing sinoatrial node. *Cardiovasc. Res.* 113, 1186–1197. doi: 10.1093/cvr/cvx104

Voukalis, C., Lip, G. Y., and Shantsila, E. (2016). Emerging Tools for Stroke Prevention in Atrial Fibrillation. *eBio Med.* 4, 26–39. doi: 10.1016/j.ebiom.2016.01.017

Ye, W., Wang, J., Song, Y., Yu, D., Sun, C., Liu, C., et al. (2015). A common Shox2-Nkx2-5 antagonistic mechanism primes the pacemaker cell fate in the pulmonary vein myocardium and sinoatrial node. *Development* 142, 2521–2532. doi: 10.1242/dev.120220

Yeh, Y. H., Ehrlich, J. R., Qi, X., Hebert, T. E., Chartier, D., and Nattel, S. (2007). Adrenergic control of a constitutively active acetylcholine-regulated potassium current in canine atrial cardiomyocytes. *Cardiovasc. Res.* 74, 406–415. doi: 10.1016/j.cardiores.2007.01.020

Zhao, G., Zhou, J., Gao, J., Liu, Y., Gu, S., Zhang, X., et al. (2017). Genome-wide DNA methylation analysis in permanent atrial fibrillation. *Mol. Med. Rep.* 16, 5505–5514. doi: 10.3892/mmr.2017.7221

Zhao, J., Liu, T., and Li, G. (2014). Relationship between two arrhythmias: sinus node dysfunction and atrial fibrillation. *Arch. Med. Res.* 45, 351–355. doi: 10.1016/j.arcmed.2014.04.005

Zhong, J., Agha, G., and Baccarelli, A. A. (2016). The role of DNA Methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. *Circ. Res.* 118, 119–131. doi: 10.1161/CIRCRESAHA.115.305206

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: 16 April 2020
doi: 10.3389/fgene.2020.00270

Identifying Shared Risk Genes for Asthma, Hay Fever, and Eczema by Multi-Trait and Multiomic Association Analyses

*Hongping Guo[1,2], Jiyuan An[3] and Zuguo Yu[1,4]**

[1] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan, China, [2] School of Mathematics and Computer Science, Hanjiang Normal University, Hubei, China, [3] Centre for Tropical Crops and Biocommodities, Queensland University of Technology, Brisbane, QLD, Australia, [4] School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD, Australia

OPEN ACCESS

Edited by:
Tao Huang,
Shanghai Institutes for Biological Sciences (CAS), China

Reviewed by:
Xuehai Hu,
Huazhong Agricultural University, China
Liqian Zhou,
Hunan University of Technology, China
Yuan Chen,
Hunan Agricultural University, China
Weiyang Chen,
Qilu University of Technology, China

***Correspondence:**
Zuguo Yu
yuzuguo@aliyun.com

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

Received: 19 December 2019
Accepted: 05 March 2020
Published: 16 April 2020

Citation:
Guo H, An J and Yu Z (2020) Identifying Shared Risk Genes for Asthma, Hay Fever, and Eczema by Multi-Trait and Multiomic Association Analyses. Front. Genet. 11:270. doi: 10.3389/fgene.2020.00270

Asthma, hay fever and eczema are three comorbid diseases with high prevalence and heritability. Their common genetic architectures have not been well-elucidated. In this study, we first conducted a linkage disequilibrium score regression analysis to confirm the strong genetic correlations between asthma, hay fever and eczema. We then integrated three distinct association analyses (metaCCA multi-trait association analysis, MAGMA genome-wide and MetaXcan transcriptome-wide gene-based tests) to identify shared risk genes based on the large-scale GWAS results in the GeneATLAS database. MetaCCA can detect pleiotropic genes associated with these three diseases jointly. MAGMA and MetaXcan were performed separately to identify candidate risk genes for each of the three diseases. We finally identified 150 shared risk genes, in which 60 genes are novel. Functional enrichment analysis revealed that the shared risk genes are enriched in inflammatory bowel disease, T cells differentiation and other related biological pathways. Our work may provide help on treatment of asthma, hay fever and eczema in clinical applications.

Keywords: asthma, hay fever, eczema, association studies, shared genes, multi-trait, multiomic

INTRODUCTION

Asthma is a bronchial disease characterized by chronic inflammation and narrowing of the airways. It results in recurring coughing, periods of wheezing, chest tightness, and mucus production (Moffatt et al., 2010; Vicente et al., 2017; Pividori et al., 2019). Hay fever (allergic rhinitis) is an inflammation disease of the nasal mucous membranes. Its symptoms include sneezing, nasal congestion, rhinorrhea, and itching (Ramasamy et al., 2011; Bunyavanich et al., 2014; Ferreira et al., 2014). Eczema (atopic dermatitis) is a form of dermatitis. Its manifestations include itching and dryness, recurring skin rashes with redness, blistering and skin edema (Sun et al., 2011; Weidinger et al., 2013; Paternoster et al., 2015). The three diseases have high global prevalence. Nearly 15% of the world population are affected by asthma (Vicente et al., 2017), 10∼20% by hay fever (Ober and Yao, 2011), 15∼30% of children and 5∼10% of adults are affected by eczema (Waage et al., 2018). Poor life quality and substantial medical expenditure bother the patients (Ober and Yao, 2011; Waage et al., 2018). Moreover, the three diseases have significant genetic contributions in different patients. The heritability ranges from 35% to 95% for asthma, from 33% to 91% for hay fever and from 71% to 84% for eczema (Ober and Yao, 2011; Zhu et al., 2018; Johansson et al., 2019).

Genome-wide association studies (GWAS) are the most powerful tools to identify the disease-associated variants. GWAS have been carried out separately for asthma, hay fever and eczema in the last two decades (Moffatt et al., 2010; Paternoster et al., 2015; Waage et al., 2018). To date (2019.11), hundreds of statistically significant single-nucleotide polymorphisms (SNPs) have been identified to be associated with each of three diseases according to GWAS-catalog database (MacArthur et al., 2017).

Clinical and epidemiological studies have found that the three diseases often co-occur in the same person or different members from the same family (Ober and Yao, 2011; Ferreira et al., 2017). Up to 90% of asthmatics suffer from allergic diseases such as hay fever and eczema (Leynaert et al., 2000; Zhu et al., 2018). Furthermore, eczema was demonstrated to be a major risk factor for the development of asthma and hay fever (Spergel, 2010). About 30% eczema patients were affected by asthma, and approximately 66% eczema patients were affected by hay fever (Ober and Yao, 2011). Similarly, 19∼38% hay fever patients were affected by asthma simultaneously (Ober and Yao, 2011). These phenomena indicate potential genetic pleiotropy and co-morbidity between asthma, hay fever and eczema. Therefore, identifying shared risk genes between these three diseases can broaden our knowledge of the underlying shared genetic causes, as well as lead the way to prevention and treatments based on the molecular mechanisms (Marenholz et al., 2013; Ferreira et al., 2017; Zhu et al., 2018).

In the past 3 years, several large-scale GWAS focused on unraveling the shared genetic architectures between asthma, hay fever and eczema based on data from UK Biobank (Sudlow et al., 2015; Ferreira et al., 2017; Zhu et al., 2018; Johansson et al., 2019). Researchers (Ferreira et al., 2017) performed meta-analysis of allergic diseases (asthma and/or hay fever and/or eczema) based on GWAS results from 13 studies by using METAL (Willer et al., 2010) software to identify the associations, and used GeneNetwork (Fehrmann et al., 2015) to identify biological processes enriched among the genes. Finally the reason why asthma, hay fever and eczema partly coexist was revealed, i.e., they share many genetic variations that dysregulate the expression of immune-related genes. Subsequently, another study (Zhu et al., 2018) applied cross-trait GWAS meta-analysis by using R package ASSET (Bhattacharjee et al., 2012) to combine the associations for asthma and allergic diseases (hay fever and/or eczema) at individual variants. They demonstrated that shared risk loci not only influence immune/inflammatory systems but also tissues with epithelium cells. A recent work showed that these three diseases shared a large amount of genetic contributions, but part of which is more disease specific (Johansson et al., 2019). However, these studies did not make strict distinction between the three diseases in phenotypic definition. Either they used a broad allergic disease defined as asthma and/or hay fever and/or eczema, or a slightly more narrow definition which distinguished asthma from allergic diseases, i.e., asthma and allergic diseases (hay fever and/or eczema). This may cause inaccurate conclusions. Moreover, the pleiotropic effect between each gene (including multiple variants) and these three correlated diseases jointly were not taken into account, which may lead to low statistical power or small percentage of explainable genetic variance. Multi-trait association study method metaCCA (Cichonska et al., 2016) enables the pleiotropy to be resolved effectively. It has been applied to identify shared pleiotropic genes for three correlated diseases (type 2 diabetes, obesity and dyslipidemia) (Chen et al., 2018) and five major psychiatric disorders (Jia et al., 2019), respectively. However, the sample sizes in the above-mentioned two studies were not large enough (several tens of thousands), and only genome data was used, resulting in only 25 and 66 shared risk genes obtained, separately.

In this study, we firstly performed a linkage disequilibrium (LD) score regression to evaluate genetic correlations between asthma, hay fever and eczema. We then integrated three distinct association analyses (metaCCA multi-trait association analysis, MAGMA genome-wide and MetaXcan transcriptome-wide gene-based tests) to identify shared risk genes based on the large-scale GWAS results in GeneATLAS database (Canela-Xandri et al., 2018). MetaCCA can detect pleiotropic genes jointly associated with these three diseases (Cichonska et al., 2016). MAGMA (de Leeuw et al., 2015) considers the correlations between genes and each disease, and MetaXcan (Gamazon et al., 2015) merges the gene expression information to identify candidate risk genes for each of the three diseases. Through these three different analyses, we obtained the potential shared risk genes associated with these three diseases. Finally we verified them by GWAS-catalog analysis, enrichment analysis and protein–protein interaction (PPI) network analysis to provide biology insights.

MATERIALS AND METHODS

GWAS Result Datasets

We downloaded the GWAS results from a publicly accessible database GeneATLAS (Canela-Xandri et al., 2018), including asthma (N_{cases} = 52269, $N_{controls}$ = 399995), hay fever (N_{cases} = 25473, $N_{controls}$ = 426791) and eczema (N_{cases} = 11552, $N_{controls}$ = 440712). The total 452264 samples are all European-ancestry individuals from UK Biobank. In this study, we used the same 623944 genotyped variants in each sample that passed quality control in GeneATLAS.

Methods

LD Score Regression Analysis

We applied linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015) to estimate genetic correlations, as well as SNP heritability and LD-score intercept for asthma, hay fever and eczema, respectively. We used the reference panel from European-ancestry population of 1000 Genome Project Phase 3 (The 1000 Genomes Project Consortium, 2015).

Multi-Trait Association Analysis

After estimating genetic correlations between asthma, hay fever and eczema, we used metaCCA multi-trait GWAS approach to identify pleiotropic genes associated equally with the three diseases. MetaCCA enables the measure of correlation between the gene (including multiple variants) and multiple traits using canonical correlation analysis (CCA) (Cichonska et al., 2016).

This takes into consideration that there exist dependencies (i.e., covariances) between genotypic and phenotypic variables, and the cross-covariance between all genotypic and phenotypic variables is made of univariate regression coefficients in linear model.

In order to reduce the computation time and memory, we first conducted gene annotation by referring NCBI human genome build 37 (including 19427 gene locations), and found that 301949 (48.39%) of the total 623944 SNPs are mapped to 17446 genes. Then we performed linkage disequilibrium (LD) based pruning to filter SNPs using PLINK software (version: 1.90b) with parameters (–indep-pairwise 50 5 0.2) (Jia et al., 2019), i.e., calculating LD between each pair of SNPs in a window of 50 SNPs, removing one of a pair of SNPs if the LD is greater than 0.2, shifting the window of 5 SNPs forward and repeating the procedure until no pairs of SNPs with high LD remain. We selected those SNPs which overlap with variants from the European population in HapMap3. After pruning, 24946 of the input 301949 SNPs are mapped to 6575 genes. We used 24946 SNPs to estimate genotypic correlation structure. 301949 SNPs were applied to estimate phenotypic correlation structure due to the fact that the larger number of variants, the higher the estimation accuracy (Cichonska et al., 2016). The covariance matrix between all genotypic and phenotypic variables is made up of regression coefficients in the GWAS results. The majority of the CPU memory in metaCCA is spent on estimating the covariance between genotypic variables. The space complexity is O(n^2), where n is the number of SNPs, and it used about 6.3 gb memory for 24946 SNPs. MetaCCA mainly uses CPU time in estimation of genotypic correlation structure and canonical correlations. In our study, metaCCA took about 4 h for multi-trait gene test of the three diseases. We performed the operations on a computer of Intel Xeon E5-2640 CPU 2.40 GHz.

To determine significant loci ($p < 5 \times 10^{-8}$) that are independent from each other, we used the clump procedure of PLINK software (Purcell et al., 2007). We set parameters (–clump-p1 5×10^{-8} –clump-p2 1×10^{-5} –clump-r2 0.2 –clump-kb 500) (Zhu et al., 2018) indicating the SNPs with a p-value less than 1×10^{-5}, LD statistic r^2 more than 0.2, and within 500 kb distance from the peak, will be assigned to that peak's clump.

Genome-Wide Gene-Based Analysis

Gene-based analysis is a statistical method for simultaneous analysis of multiple genetic variations to determine their joint effect. MAGMA, a genome-wide gene-based association method based on a multiple linear principal components regression model (de Leeuw et al., 2015), was used to identify significant genes using the GWAS results for asthma, hay fever and eczema, respectively. We regarded the individual-level genotype data from European-ancestry population of 1000 Genomes Project Phase 3 as reference. 19427 genes in the whole genome were used to determine the significance threshold in Bonferroni correction. The space complexity of MAGMA is O(k^2), where k is the number of genes. For a human genome, the required memory is about 5 gb. In MAGMA, the majority of the CPU time is spent on the ordinary least squares method, the time complexity is O($k^2 \times (n + k)$), where k is the number of genes and n is the number of SNPs. In our study, MAGMA took about 1 min to analyze each disease.

Transcriptome-Wide Gene-Based Analysis

We used the MetaXcan framework to integrate expression quantitative trait loci (eQTL) information with GWAS results and map genes associated with disease traits. MetaXcan is a transcriptome-wide gene-based association approach that estimates tissue-specific gene expression profiles from GWAS results using prediction models trained in large reference databases, and correlates predicted expression levels with diseases (such as asthma) to detect potential disease-associated genes (Barbeira et al., 2018). It has high concordance (correlation coefficient: $R^2 > 0.999$) with the individual-level version PrediXcan (Gamazon et al., 2015). Training sets are reference transcriptome datasets from the Genotype-Tissue Expression Project (GTEx: version 7) (GTEx Consortium, 2017), the weights and covariances of prediction model for different tissues are available from PredictDB (http://predictdb.org/).

In order to reduce multiple-testing burden, we analyzed 10 of the total 48 tissues, 4 obvious tissues (Whole Blood, Lung, Skin Sun Exposed and Skin Not Sun Exposed) plus 6 other relevant tissues (Cells EBV-transformed lymphocytes, Cells Transformed fibroblasts, Esophagus Gastroesophageal Junction, Esophagus Mucosa, Esophagus Muscularis and Vagina) reported in previous studies (Ferreira et al., 2017; Zhu et al., 2018). The total number of genes (27314) in the 10 tissues was used to determine the Bonferroni correction threshold. We ran MetaXcan separately in asthma, hay fever and eczema, each with the same 10 tissues, and used per SNP p-value from GWAS results after correction for the LD-score intercept. MetaXcan uses a small amount of memory and very little CPU time. MetaXcan's CPU time is primarily spent on the calculation of covariance of the gene matrix. The space and time complexity are O(k^2) and O(k^3) respectively, where k is the number of genes in the tissue. In our study, 18 min were spent on MetaXcan's analysis of 10 tissues for each disease.

GWAS-Catalog Analysis, Enrichment Analysis and PPI Network Analysis

To understand whether the identified genes have been reported in the previous GWAS studies for asthma, hay fever and eczema, we downloaded the corresponding GWAS catalog from NHGRI-EBM (3 November, 2019), and searched the genes one by one. To gain biology insights from the shared risk genes, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway analysis using the Enrichr web server (Kuleshov et al., 2016) from http://amp.pharm.mssm.edu/Enrichr. The significant criterion is that the adjusted p-value is less than 0.05. In addition, we used STRING v10 (Szklarczyk et al., 2015) from https://string-db.org/ to analyze the PPI network.

A flow chart of our work is shown in **Figure 1**. That is, we integrated three association studies (metaCCA multi-trait association analysis, MAGMA genome-wide and MetaXcan transcriptome-wide gene-based tests) to identify candidate risk

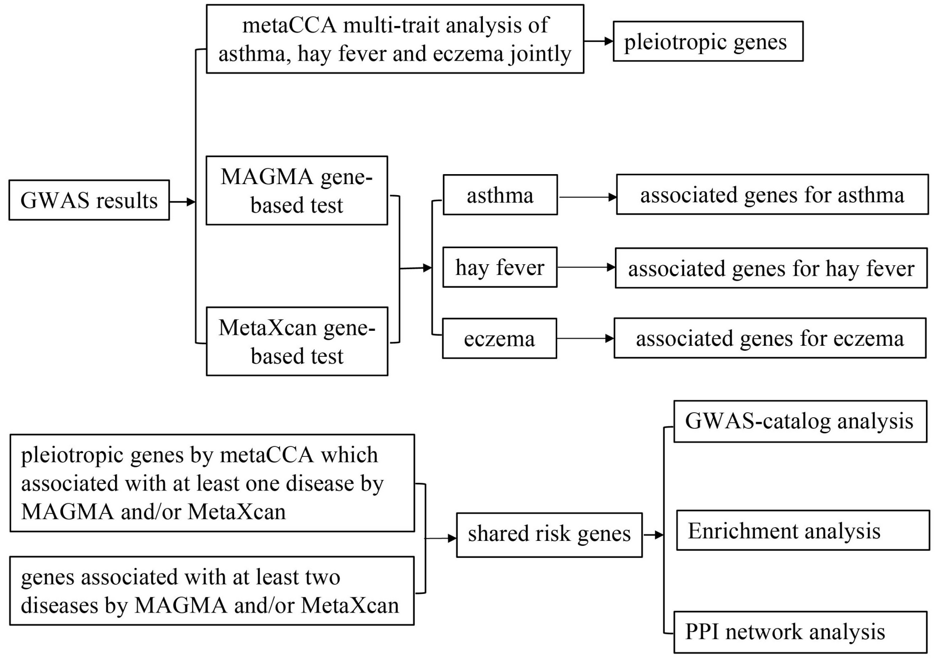

metaCCA multi-trait analysis of asthma, hay fever and eczema jointly
pleiotropic genes
GWAS results
MAGMA gene-based test
MetaXcan gene-based test
asthma
hay fever
eczema
associated genes for asthma
associated genes for hay fever
associated genes for eczema
pleiotropic genes by metaCCA which associated with at least one disease by MAGMA and/or MetaXcan
genes associated with at least two diseases by MAGMA and/or MetaXcan
shared risk genes
GWAS-catalog analysis
Enrichment analysis
PPI network analysis

FIGURE 1 | Flow chart of the present work.

genes, and then conducted GWAS-catalog analysis, enrichment analysis and PPI network analysis to the shared risk genes.

RESULTS

Genetic Correlation Between Asthma, Hay Fever and Eczema

We evaluated the genetic correlation between asthma, hay fever and eczema using LD score regression (LDSC). Genetic correlation between asthma and hay fever (r_g = 0.665, SE = 0.0457, P = 5.26 × 10^{-48}) is the strongest, followed by the correlation between asthma and eczema (r_g = 0.4519, SE = 0.0577, P = 4.93 × 10^{-15}), then between hay fever and eczema (r_g = 0.3297, SE = 0.0714, P = 3.85 × 10^{-6}) (**Table 1**). In summary, significant genetic correlations are observed between any pair of the three diseases. Additionally, estimates of SNP heritability (h^2) on the liability scale (assuming 15% disease prevalence) is 11.85% (SE = 1.15%) for asthma, 4.65% (SE = 0.41%) for hay fever and 2.36% (SE = 0.53%) for eczema. Furthermore, the LD score intercepts for asthma, hay fever and eczema are 1.043 (SE = 0.0143), 1.0195 (SE = 0.0102) and 1.0085 (SE = 0.0105), respectively, indicating most of the inflation is due to polygenic effect rather than population structure or sample overlap (An et al., 2019).

Pleiotropic Genes Identified by Multi-Trait Association Study

We performed metaCCA multi-trait association study to identify pleiotropic genes that are associated jointly with asthma, hay fever and eczema. There were 66 pleiotropic genes that reached the significant threshold ($P_{metaCCA}$ < 7.6 × 10^{-6}) after the Bonferroni correction of the LD pruned 6575 genes, the canonical correlations of which ranged from 0.0077 to 0.0302. The results for the metaCCA gene-based test are shown in **Supplementary Data 1**.

TABLE 1 | Genetic correlation between asthma, hay fever, and eczema.

Diseases[1]	Asthma	Hay fever	Eczema
Asthma	1	0.665 (0.0457)	0.4519 (0.0577)
Hay fever	5.256 × 10^{-48}	1	0.3297 (0.0714)
Eczema	4.930 × 10^{-15}	3.848 × 10^{-6}	1

[1] *Element in upper off-diagonal is the genetic correlation r_g (standard deviation SE), element in lower off-diagonal is the corresponding genetic correlation P-value.*

Genes Identified by Genome-Wide and Transcriptome-Wide Studies

We conducted MAGMA genome-wide gene-based analysis to identify genes associated with asthma, hay fever and eczema, respectively. 287, 80, and 57 significant genes (P_{MAGMA} < 2.57 × 10^{-6}) were identified after Bonferroni correction of the total 19427 genes (**Supplementary Data 2**). Moreover, we carried out MetaXcan transcriptome-wide gene-based analysis, and detected 204, 48, and 53 genes that were above the significance level ($P_{MetaXcan}$ < 1.84 × 10^{-6}) determined by 27314 genes in 10 relevant tissues (**Supplementary Data 3–5**).

Noticing that some overlapping genes exist for the same gene-based test, we took the results in MAGMA as an

TABLE 2 | Number of genes identified by MAGMA and MetaXcan.

Methods	Asthma	Hay fever	Eczema	Asthma and Hay fever	Asthma and Eczema	Hay fever and Eczema	Asthma and Hay fever and Eczema
MAGMA	287	80	57	65	36	19	17
MetaXcan	204	48	53	37	33	5	4
Combined[1]	397	109	91	94	59	24	23

[1] *Number of genes identified by MAGMA and/or MetaXcan.*

TABLE 3 | Details of overlapping genes in Type I and II of shared risk genes.

Genes[1]	$P_{metaCCA}$	Asthma		Hay fever		Eczema		Literature PMID
		P_{MAGMA}	$P_{MetaXcan}$	P_{MAGMA}	$P_{MetaXcan}$	P_{MAGMA}	$P_{MetaXcan}$	
TNXB†	7.12e-29	3.51e-35		1.39e-10		1.20e-09		23886662
C6orf10‡	1.60e-18	1.59e-22		1.01e-12		9.84e-10		21804548, 23042114
CLEC16A*	8.26e-16	4.24e-22		3.51e-10		5.92e-11		31036433, 30013184, 26482879
C2*	1.84e-06	1.31e-14	3.51e-21	1.08e-13		5.06e-08	1.45e-08	29551627, 25085501, 26542096
WDR36*	1.95e-26	1.61e-24	5.68e-14	2.58e-14	2.52e-08			30929738, 24388013, 30595370
PSORS1C2	3.54e-15	3.77e-13		6.80e-07				
HLA-DMB	7.72e-14		6.67e-14	3.34e-07				
BTNL2†	1.14e-12	1.03e-59		5.08e-09				29273806
BAG6	5.69e-11	9.44e-19	6.64e-15	7.05e-10	6.03e-18			
SLC25A46*	2.79e-09	1.35e-09		9.52e-09				31036433, 22036096, 30595370
CAMK4‡	2.31e-08	5.12e-11		1.28e-08				29785011, 30013184
MUC22	8.56e-07	1.56e-13		2.01e-11				
PLCL1‡	6.08e-06	2.23e-06		9.73e-12				30013184, 30595370
RNF5	4.31e-17	1.75e-12	3.39e-13			2.06e-10	5.84e-11	
KIF3A‡	7.05e-16	5.35e-13	6.57e-14			7.51e-08		31036433, 26542096
DDAH2	1.78e-07		1.43e-08				3.10e-08	
RAD50*	4.05e-06	6.20e-29	9.21e-31			6.36e-07		30929738, 30013184, 26482879

[1] *Symbol †, ‡, and * behind the genes represents 1, 2, and 3 associated diseases (asthma, hay fever, eczema) reported in GWAS-catalog, respectively. PMID, PubMed unique identifier. The blank cells are non-significant p-values or no supporting literature.*

example, there are 65 overlapping genes between asthma and hay fever, 36 between asthma and eczema, 19 between hay fever and eczema, and 17 among the three diseases. Similarly, some genes detected by both MAGMA and MetaXcan for the same disease, such as 94 overlapping genes are identified in asthma. We combined the genes identified by MAGMA and/or MetaXcan, and obtained 397, 109, and 91 significant genes for asthma, hay fever and eczema, respectively. The numbers of genes identified by the two approaches are shown in **Table 2**.

Shared Risk Genes for Asthma, Hay Fever, and Eczema

We considered the shared risk genes from two types. Type I includes the pleiotropic genes by metaCCA which were

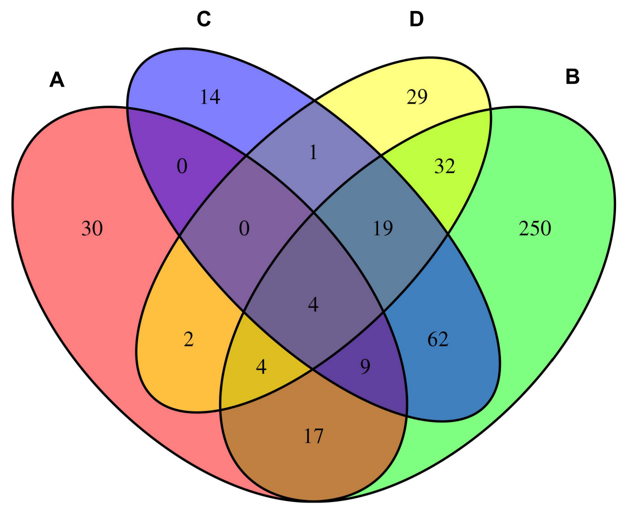


FIGURE 2 | Venn diagram of the pleiotropic genes identified by metaCCA **(A)** and the combined genes identified by MAGMA and/or MetaXcan for asthma **(B)**, hay fever **(C)** and eczema **(D)**.

associated with at least one disease by MAGMA and/or MetaXcan, it is inspired by these two studies (Chen et al., 2018; Jia et al., 2019); Type II includes the pleiotropic genes associated with at least two diseases by MAGMA and/or MetaXcan. We found that type I includes 36 genes ($P_{metaCCA} < 7.6 \times 10^{-6}$, $P_{MAGMA} < 2.57 \times 10^{-6}$, and/or $P_{MetaXcan} < 1.84 \times 10^{-6}$ in at least one of asthma, hay fever and eczema), and type II contains 131 genes ($P_{MAGMA} < 2.57 \times 10^{-6}$ and/or $P_{MetaXcan} < 1.84 \times 10^{-6}$ in at least two of asthma, hay fever and eczema). After removing the repetitions in these two types, 150 shared risk genes were obtained (**Supplementary Data 6**). Here we only showed the details of the 17 overlapping genes in type I and II in **Table 3**. A Venn diagram (**Figure 2**) shows the pleiotropic genes identified by metaCCA and the combined genes identified by MAGMA and/or MetaXcan for asthma, hay fever and eczema. We can see that four overlap genes can not only be detected by metaCCA but also associated with all of the three diseases by MAGMA and/or MetaXcan analyses.

GWAS-Catalog Analysis, Enrichment Analysis and PPI Network Analysis

To see whether the 150 shared risk genes have been reported previously, GWAS-catalog analysis was carried out for each gene. We found 23 genes have been reported to be associated with all of the three diseases, 31 genes have been reported to be associated with two diseases, and 36 genes have been reported to be associated with one disease. Furthermore, 60 genes have never been reported, suggesting that these are novel ones. Gene names involved in these four different classes are listed in **Table 4**, their corresponding PubMed IDs of supporting literatures are shown in **Supplementary Data 7**. Among the 90 genes which have been reported as associated with diseases before, 85, 31, and 51 of them have been reported as associated with asthma, hay fever and eczema (**Supplementary Data 7**), respectively. Some genes are only detected by metaCCA. CGN has been reported associated with asthma, but it was not detected by MAGMA and/or MetaXcan for asthma data; RAD50 has been reported as associated with hay fever, but it was not detected by MAGMA and/or MetaXcan for hay fever data; eight genes (AHI1, IL2, MICB, NDFIP1, PLCL1, PRKCQ, SLC25A46, and WDR36) have been reported as associated with eczema, but they were not detected by MAGMA and/or MetaXcan for eczema data (**Supplementary Data 6, 7**). Similarly, there are also some reported genes that can only be detected by MAGMA and/or MetaXcan. 67 of the reported genes which are associated with asthma can only be successfully identified by MAGMA and/or MetaXcan, but not by metaCCA. For hay fever and eczema, gene numbers of this class are 22 and 15 (**Supplementary Data 7**), respectively. In addition, there are 5 genes (C2, CLEC16A, RAD50, SLC25A46, and WDR36) have been reported to be associated with all of the three diseases for the 66 pleiotropic genes by metaCCA (**Supplementary Data 1**). For the 424 genes (287 for asthma, 80 for hay fever, 57 for eczema) detected by MAGMA, there are 141, 23, and 24 that have been reported associated with asthma, hay fever and eczema in the GWAS-catalog (**Supplementary Data 2**), respectively.

TABLE 4 | List of 150 shared risk genes divided into four categories.

Related diseases[1]	Gene names
3	BACH2, C11orf30, C2, CLEC16A, GSDMA, HLA-B, HLA-C, HLA-DQA1, IKZF3, IL13, IL18R1, IL1RL1, IL2, IL2RA, IL7R, LPP, RAD50, SLC25A46, SMAD3, TLR1, TNF, TSLP, WDR36
2	AAGAB, ADAD1, C6orf10, CAMK4, CD247, D2HGDH, ERBB3, FLG, GSDMB, HLA-DQB1, HLA-DRB1, IL18RAP, IL1R1, IL33, KIAA1109, KIF3A, MICA, MICB, NDFIP1, PBX2, PLCL1, PRKCQ, PRR5L, RORC, RPS26, RTEL1, SMARCE1, STAT6, TLR10, TMEM232, ZBTB46
1	AHI1, BRD2, BTNL2, C4A, CGN, FAM114A1, GAL3ST2, GLDC, GPSM3, HLA-DPA1, HLA-DQA2, HLA-DQB2, HLA-DRA, HLA-DRB5, HLA-DRB6, HLA-DRB9, IKZF4, IL21R, ITPR3, LCE3D, MRVI1, NOTCH4, ORMDL3, PSORS1C1, S100A1, SLC22A4, SLC22A5, SLC9A2, SLC9A4, SPRR2D, SUOX, TAP2, TLR6, TNXB, TRIM26, ZGPAT
0	AGER, AGPAT1, AIF1, ARNT, ATF6B, BAG6, BAK1, C4B, C6orf25, C6orf47, C6orf48, CCHCR1, CFB, CXXC11, CYP21A2, DDAH2, DIS3L, DOCK3, DPP4, DXO, EGFL8, EHMT2, FKBPL, GNL1, HCG27, HCG4B, HLA-DMB, HSPA1B, HSPA1L, HSPA4, KPRP, LEMD2, LINGO4, LOC101929163, LST1, MRPL9, MSH5, MUC21, MUC22, NELFE, PGLYRP4, PPT2, PRRC2A, PRRT1, PRUNE, PSMD4, PSORS1C2, RNF5, S100A2, SAPCD1, SEMA6C, SKIV2L, SLC44A4, STK19, TAP1, TCF19, TNXA, VWA7, ZBTB12, ZKSCAN3

[1] *The digit in the first column means the number of associated diseases (asthma, hay fever, eczema) reported in GWAS-catalog.*

Before conducting enrichment analysis, we excluded the genes in the major histocompatibility complex (MHC) region (Zhu et al., 2019). On the one hand, a majority of genes in MHC region are related to immune response which may bring false positives (Pividori et al., 2019); on the other hand, for asthma and allergy diseases, MHC region was reported as containing some of the strongest association signals such as HLA-DQB and HLA-B (Waage et al., 2018). We expected to find other biological pathways besides immunity. KEGG pathway enrichment analysis by Enrichr web server (http://amp.pharm.mssm.edu/Enrichr) shows that 6 biological pathways were significantly enriched (**Supplementary Data 8**). They are inflammatory bowel disease (IBD) (hsa05321), Th17 cell differentiation (hsa04659), cytokine–cytokine receptor interaction (hsa04060), Th1 and Th2 cell differentiation (hsa04658), JAK-STAT signaling pathway (hsa04630) and chagas disease (American trypanosomiasis) (hsa05142). The most strongly enriched one is IBD pathway (hsa05321) including 8 enriched genes (IL18RAP, SMAD3, IL13, RORC, IL21R, STAT6, IL2, IL18R1). A bubble chart shows the result of KEGG pathway analysis (**Figure 3**).

To understand the interactions between shared risk genes (excluding those in MHC region), we conducted PPI network analysis using STRING tool. There are in total 168 pairs of interaction in PPI network (**Supplementary Data 9**), all the interacting genes have combined scores of no less than 0.4, in which 9 pairs of genes (IL2RA-IL2, IL33-IL1RL1, TSLP-IL7R, IL18R1-IL18RAP, IL13-STAT6, IKZF3-IL2, CD247-IL2, LCE3D-SPRR2D, TLR6-TLR1) with scores $\geq$ 0.95. The 10 hub genes (degree $\geq$ 10) that interact extensively with other genes in PPI network are IL2, IL13, TSLP, IL2RA, IL33, STAT6, ORMDL3, IL1R1, IL1RL1 and IL7R. The PPI network for shared risk genes are shown in **Figure 4**.

DISCUSSION

Two-thirds of our identified shared risk genes were reported to associate with at least one of the three diseases, asthma, hay fever and eczema. Results obtained by Enrichment analysis are mostly consistent with the findings in previous researches. For example, we found substantial shared genes in the HLA region, which was highlighted by their prominent role in immune response (Pividori et al., 2019), and immune response is one of the major factors influencing asthma, hay fever and eczema (Ferreira et al., 2017; Zhu et al., 2018). Additionally, IBD pathway (hsa05321) is the most strongly enriched pathway in our study, which was demonstrated to share susceptibility genes with allergic disease (Kreiner et al., 2017). Moreover, there are also some T cell (including TH17, TH1, TH2) related pathways enriched, involving Th17 cell differentiation (hsa04659), Th1 and Th2 cell differentiation (hsa04658). This conclusion supports that of a previous study which widely documented contribution of these T cell subsets to allergic responses (Farh et al., 2015).

We found four genes (C2, CLEC16A, C6orf10, TNXB) which have statistical significance in metaCCA, MAGMA and MetaXcan association studies for the three diseases. C2 and CLEC16A have been reported to associate with all the three diseases (Waage et al., 2018; Zhu et al., 2018; Kichaev et al., 2019).

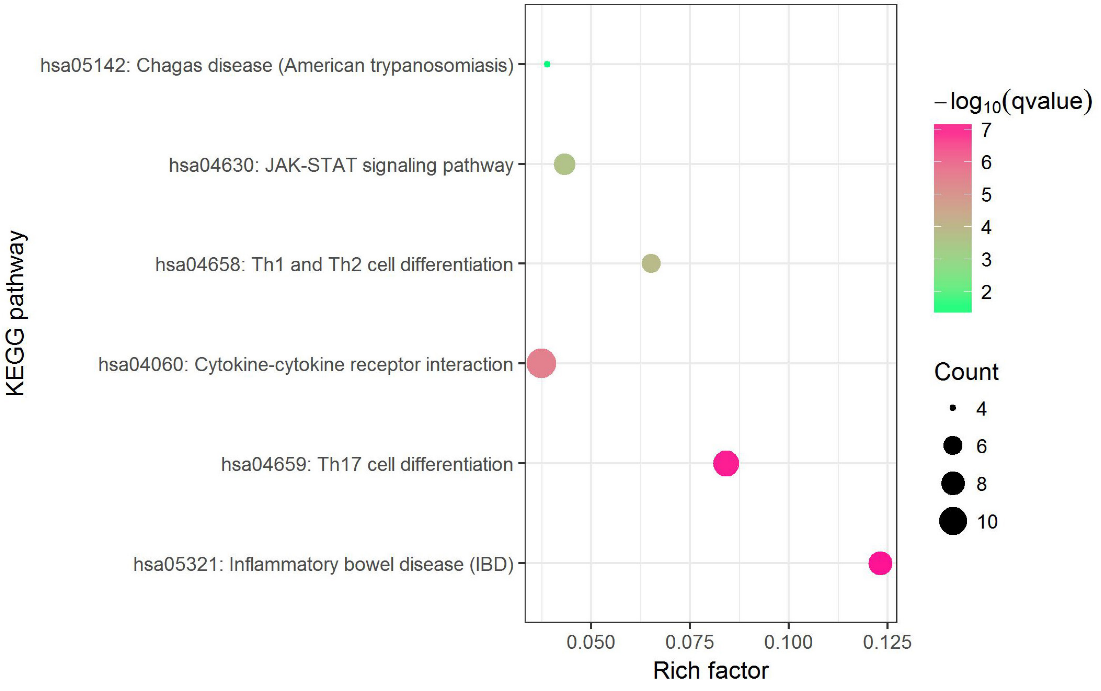


FIGURE 3 | Bubble chart of enrichment analysis of shared risk genes (excluding those in MHC region).

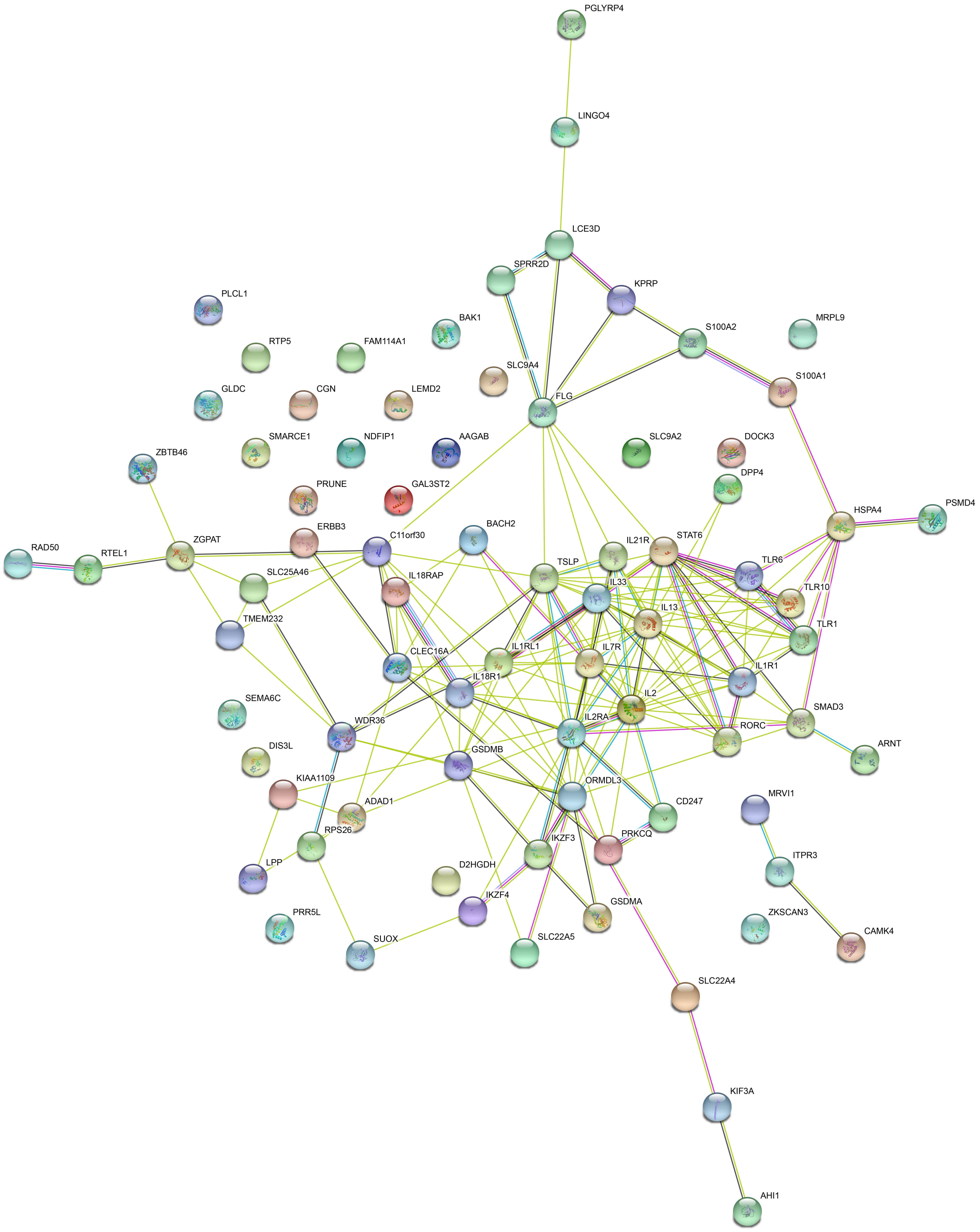


FIGURE 4 | PPI network for shared risk genes (excluding those in MHC region).

Although TNXB has only been reported to associate with eczema (Baurecht et al., 2015), it may be very important for asthma and hay fever. Among the 17 overlapping genes from types I and II of shared risk genes, six genes (PSORS1C2, HLA-DMB, BAG6, MUC22, RNF5, DDAH2) have never been reported before. Furthermore, cytokine-cytokine receptor interaction (hsa04060), JAK-STAT signaling pathway (hsa04630) and chagas disease (American trypanosomiasis) (hsa05142) also enriched in our study. These findings may be helpful in pathological diagnosis studies.

From the single-trait GWAS results of asthma, hay fever and eczema, only one independent loci (rs61893460) is found to associate with these three diseases. rs61893460 locates in C11orf30-LRRC32 region on chromosome 11 and was reported associated with total serum IgE levels (Li et al., 2012). IgE is released from the immune system and travels to local organs or tissues to type 2 cytokines, which can further cause asthma, hay fever and eczema (Ferreira et al., 2017). However, metaCCA multi-trait analysis identifies 66 pleiotropic genes, which implies stronger statistical power. We did not regard all of the 66 pleiotropic genes as shared risk genes, but refined them under a restraint, that is, they must be associated with at least one of the three diseases by MAGMA/MetaXcan. This idea derives from the two studies (Chen et al., 2018; Jia et al., 2019).

Using multi-trait analysis, we only identified five genes which have been reported associated with the three diseases, while 23 reported genes are detected by integrating multi-trait and multiomic methods. In addition, among the 90 genes which have been reported, some cannot be detected by a single method. Take gene RAD50 for example, it was reported to be associated with the three diseases in GWAS-catalog and can be identified by multi-trait method (metaCCA), but it cannot be detected by multiomic methods (MAGMA and/or MetaXcan) for hay fever disease. RAD50 promotes the development of asthma by inducing inflammatory factors secreted by Th2 cell (Li et al., 2010), and it was found to be associated with hay fever (Waage et al., 2018). These results imply the benefits of integration.

Note that 73 of 136 independent risk variants are novel in Ferreira et al. (2017), 41 of 141 loci are novel in Johansson et al. (2019), and 60 of 150 shared risk genes are novel in our study. Besides the different phenotypic definitions which we have explained in the Introduction section, the determining of novel status is also different. The novel variants not only included those risk loci that never reported to associate with any of the three diseases in GWAS-catalog, but also contained the variants that had LD statistic $r^2 < 0.05$ with all reported variants (Ferreira et al., 2017). Moreover, the novel loci were composed of variants if the locus was distanced >1 Mb from any of the previously reported loci for any of the three diseases in GWAS-catalog, PubMed or bioRxiv, as well as those variants if $r^2 < 0.05$ between the identified variant and previously reported variants (Johansson et al., 2019). Both of the definitions of "novel" in these two studies are broader than ours. In addition, we investigated genetic overlap on gene level rather than genetic variant level.

Compared with the previous studies, our work has some achievements. First, we confirmed strong genetic correlations between the three diseases. Second, we considered the pleiotropic effects via multi-trait association analysis, which yields a statistical power advantage compared to single-trait modeling strategies. Third, we identified more shared risk genes from multi-omic (genome-wide and transcriptome-wide) perspective.

Limitations

First, our results cannot be used to represent the worldwide population or children, because the samples are of European-ancestry individuals aged between 40 and 69 years old from UK Biobank. Second, association studies results in our work mean potential shared risk genes, they do not represent the causative genes. Mendelian randomization analysis can be used to reveal the causality (Verbanck et al., 2018), and fine mapping is helpful in detecting the pathogenic variants and genes (Marenholz et al., 2013; Farh et al., 2015). Third, the functions of novel shared risk genes are still unknown. There is a long way to go in understanding the gene functions and their roles in disease pathophysiology. Further studies should also highlight and explore the biological interpretation and try to translate the findings to clinical research or practice.

CONCLUSION

We confirmed strong genetic correlations between asthma, hay fever and eczema. Three different association studies are integrated to identify the shared risk genes between these three diseases. One is metaCCA multi-trait association analysis considering the joint effect, another two are MAGMA and MetaXcan gene-based tests using genome-wide and transcriptome-wide data referring to 1000 Genomes and GTEx project, respectively. We identified 150 shared risk genes, in which 60 are novel. Functional enrichment analysis reveals that the shared risk genes are enriched in inflammatory bowel disease (IBD), T cells differentiation and other related biological pathways. Our work may provide help on treatment of asthma, hay fever and eczema in clinical application.

DATA AVAILABILITY STATEMENT

The GWAS result datasets analyzed for this study can be found in the GeneALTAS http://geneatlas.roslin.ed.ac.uk/.

AUTHOR CONTRIBUTIONS

HG conceived the project, performed the data analysis, and wrote the manuscript. JA participated in guidance and discussion. ZY contributed to guidance and supervised the project. All authors read and approved the final manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant No. 11871061); Collaborative Research project for Overseas Scholars (including Hong Kong and Macau) of National Natural Science Foundation of China (Grant No. 61828203); Chinese Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) (Grant No. IRT_15R58); Hunan Provincial Innovation Foundation for Postgraduate (Grant No. CX2018B375); Project for Excellent Young and Middle-aged Science and Technology Innovation Team of Hubei Province (Grant No. T201731).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00270/full#supplementary-material

REFERENCES

An, J., Gharahkhani, P., Law, M. H., Ong, J.-S., Han, X., Olsen, C. M., et al. (2019). Gastroesophageal reflux GWAS identifies risk loci that also associate with subsequent severe esophageal diseases. *Nature Communications* 10, 4219. doi: 10.1038/s41467-019-11968-2

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* 9, 1825. doi: 10.1038/s41467-018-03621-1

Baurecht, H., Hotze, M., Brand, S., Buning, C., Cormican, P., Corvin, A., et al. (2015). Genome-wide comparative analysis of atopic dermatitis and psoriasis gives insight into opposing genetic mechanisms. *American Journal of Human Genetics* 96, 104–120. doi: 10.1016/j.ajhg.2014.12.004

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., et al. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *American Journal of Human Genetics* 90, 821–835. doi: 10.1016/j.ajhg.2012.03.015

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 47, 291–295. doi: 10.1038/ng.3211

Bunyavanich, S., Schadt, E. E., Himes, B. E., Lasky-Su, J., Qiu, W., Lazarus, R., et al. (2014). Integrated genome-wide association, coexpression network, and expression single nucleotide polymorphism analysis identifies novel pathway in allergic rhinitis. *BMC Medical Genomics* 7:48. doi: 10.1186/1755-8794-7-48

Canela-Xandri, O., Rawlik, K., and Tenesa, A. (2018). An atlas of genetic associations in UK Biobank. *Nature Genetics* 50, 1593–1599. doi: 10.1038/s41588-018-0248-z

Chen, Y., Xu, C., Zhang, J., Zeng, C., Wang, X., Zhou, R., et al. (2018). Multivariate analysis of genomics data to identify potential pleiotropic genes for type 2 diabetes, obesity and dyslipidemia using Meta-CCA and gene-based approach. *PLoS ONE* 13:e0201173. doi: 10.1371/journal.pone.0201173

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A. J., Soininen, P., Lehtimaki, T., et al. (2016). metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* 32, 1981–1989. doi: 10.1093/bioinformatics/btw052

de Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. *PLoS Computational Biology* 11:e1004219. doi: 10.1371/journal.pcbi.1004219

Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343. doi: 10.1038/nature13835

Fehrmann, R. S. N., Karjalainen, J. M., Krajewska, M., Westra, H.-J., Maloney, D., Simeonov, A., et al. (2015). Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics* 47, 115–125. doi: 10.1038/ng.3173

Ferreira, M. A., Matheson, M. C., Tang, C. S., Granell, R., Ang, W., Hui, J., et al. (2014). Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *Journal of Allergy and Clinical Immunology* 133, 1564–1571. doi: 10.1016/j.jaci.2013.10.030

Ferreira, M. A., Vonk, J. M., Baurecht, H., Marenholz, I., Tian, C., Hoffman, J. D., et al. (2017). Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nature Genetics* 49, 1752–1757. doi: 10.1038/ng.3985

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. *Nature Genetics* 47, 1091–1098. doi: 10.1038/ng.3367

GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Jia, X., Yang, Y., Chen, Y., Cheng, Z., Du, Y., Xia, Z., et al. (2019). Multivariate analysis of genome-wide data to identify potential pleiotropic genes for five major psychiatric disorders using MetaCCA. *Journal of Affective Disorders* 242, 234–243. doi: 10.1016/j.jad.2018.07.046

Johansson, A., Rask-Andersen, M., Karlsson, T., and Ek, W. E. (2019). Genome-wide association analysis of 350000 caucasians from the UK Biobank identifies novel loci for asthma, hay fever and eczema. *Human Molecular Genetics* 28, 4022–4041. doi: 10.1093/hmg/ddz175

Kichaev, G., Bhatia, G., Loh, P.-R., Gazal, S., Burch, K. K., Freund, M., et al. (2019). Leveraging polygenic functional enrichment to improve GWAS power. *American Journal of Human Genetics* 104, 65–75. doi: 10.1016/j.ajhg.2018.11.008

Kreiner, E., Waage, J., Standl, M., Brix, S. H., Pers, T., Alves, A. C., et al. (2017). Shared genetic variants suggest common pathways in allergy and autoimmune diseases. *Journal of Allergy and Clinical Immunology* 140, 771–781. doi: 10.1016/j.jaci.2016.10.055

Kuleshov, M., Jones, M., Rouillard, A., Fernandez, N., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* 44, W90–W97. doi: 10.1093/nar/gkw377

Leynaert, B., Neukirch, F., Demoly, P., and Bousquet, J. (2000). Epidemiologic evidence for asthma and rhinitis comorbidity. *Journal of Allergy and Clinical Immunology* 106, S201–S205. doi: 10.1067/mai.2000.110151

Li, X., Howard, T. D., Zheng, S. L., Haselkorn, T., Peters, S. P., Meyers, D. A., et al. (2010). Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. *Journal of Allergy and Clinical Immunology* 125, 328–335. doi: 10.1016/j.jaci.2009.11.018

Li, X. J., Ampleford, E. D., Howard, T. C., Moore, W., Li, H. W., Busse, W., et al. (2012). The C11orf30-LRRC32 region is associated with total serum IgE levels in asthma. *Journal of Allergy and Clinical Immunology* 129, 575–578. doi: 10.1016/j.jaci.2011.09.040

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research* 45, D896–D901. doi: 10.1093/nar/gkw1133

Marenholz, I., Esparza-Gordillo, J., and Lee, Y.-A. (2013). Shared genetic determinants between eczema and other immune-related diseases. *Current Opinion in Allergy and Clinical Immunology* 13, 478–486. doi: 10.1097/ACI.0b013e328364e8f7

Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Heath, S., et al. (2010). A large-scale, consortium-based genomewide association study of asthma. *The New England Journal of Medicine* 363, 1211–1221. doi: 10.1056/NEJMoa0906312

Ober, C., and Yao, T.-C. (2011). The genetics of asthma and allergic disease: a 21st century perspective. *Immunological Reviews* 242, 10–30. doi: 10.1111/j.1600-065X.2011.01029.x

Paternoster, L., Standl, M., Waage, J., Baurecht, H., Hotze, M., Strachan, D. P., et al. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics* 47, 1449–1456. doi: 10.1038/ng.3424

Pividori, M., Schoettler, N., Nicolae, D. L., Ober, C., and Im, H. K. (2019). Shared and distinct genetic risk factors for childhood-onset and adult-onset asthma: genome-wide and transcriptome-wide studies. *The Lancet Respiratory Medicine* 7, 509–522. doi: 10.1016/S2213-2600(19)30055-4

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics* 81, 559–575. doi: 10.1086/519795

Ramasamy, A., Curjuric, I., Coin, L. J., Kumar, A., McArdle, W. L., Imboden, M., et al. (2011). A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. *Journal of Allergy and Clinical Immunology* 128, 996–1005. doi: 10.1016/j.jaci.2011.08.030

Spergel, J. (2010). Epidemiology of atopic dermatitis and atopic march in children. *Immunology and allergy clinics of North America* 30, 269–280. doi: 10.1016/j.iac.2010.06.003

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine* 12:e1001779. doi: 10.1371/journal.pmed.1001779

Sun, L., Xiao, F., Li, Y., Zhou, W., Tang, H., Tang, X., et al. (2011). Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population. *Nature Genetics* 43, 690–694. doi: 10.1038/ng.851

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research* 43, D447–D452. doi: 10.1093/nar/gku1003

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics* 50, 693–698. doi: 10.1038/s41588-018-0099-7

Vicente, C. T., Revez, J. A., and Ferreira, M. A. R. (2017). Lessons from ten years of genome-wide association studies of asthma. *Clinical & Translational Immunology* 6, e165. doi: 10.1038/cti.2017.54

Waage, J., Standl, M., Curtin, J. A., Jessen, L. E., Thorsen, J., Tian, C., et al. (2018). Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nature Genetics* 50, 1072–1080. doi: 10.1038/s41588-018-0157-1

Weidinger, S., Willis-Owen, S. A., Kamatani, Y., Baurecht, H., Morar, N., Liang, L., et al. (2013). A genome-wide association study of atopic dermatitis identifies loci with overlapping effects on asthma and psoriasis. *Human Molecular Genetics* 22, 4841–4856. doi: 10.1093/hmg/ddt317

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Zhu, Z., Lee, P. H., Chaffin, M. D., Chung, W., Loh, P.-R., Lu, Q., et al. (2018). A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nature Genetics* 50, 857–864. doi: 10.1038/s41588-018-0121-0

Zhu, Z., Lin, Y., Li, X., Driver, J. A., and Liang, L. (2019). Shared genetic architecture between metabolic traits and alzheimers disease: a large-scale genome-wide cross-trait analysis. *Human Genetics* 138, 271–285. doi: 10.1007/s00439-019-01988-9

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