SYSTEM BIOLOGY METHODS AND TOOLS FOR INTEGRATING OMICS DATA - VOLUME II

EDITED BY: Liang Cheng, Lei Deng and Mingxiang Teng

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-915-5
DOI 10.3389/978-2-88976-915-5

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

SYSTEM BIOLOGY METHODS AND TOOLS FOR INTEGRATING OMICS DATA - VOLUME II

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

Citation: Cheng, L., Deng, L., Teng, M., eds. (2022). System Biology Methods and Tools for Integrating Omics Data - Volume II. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-915-5

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 April 2021
doi: 10.3389/fgene.2021.686116

Detection of circRNA Biomarker for Acute Myocardial Infarction Based on System Biological Analysis of RNA Expression

*Wen Yang[1], Li Sun[2], Xun Cao[1], Luyifei Li[1], Xin Zhang[1], Jianqian Li[1], Hongyan Zhao[3], Chengchuang Zhan[1], Yanxiang Zang[1], Tiankai Li[1], Li Zhang[1], Guangzhong Liu[1] and Weimin Li[1]**

[1] Department of Cardiology, The First Affiliated Hospital, Harbin Medical University, Harbin, China, [2] Department of Cardiology, The First Affiliated Hospital, China University of Science and Technology, Hefei, China, [3] Department of Cardiology, The People's Hospital of Liaoning Province, Shenyang, China

Acute myocardial infarction (AMI) is myocardial necrosis caused by the persistent interruption of myocardial blood supply, which has high incidence rate and high mortality in middle-aged and elderly people in the worldwide. Biomarkers play an important role in the early diagnosis and treatment of AMI. Recently, more and more researches confirmed that circRNA may be a potential diagnostic biomarker and therapeutic target for cardiovascular diseases. In this paper, a series of biological analyses were performed to find new effective circRNA biomarkers for AMI. Firstly, the expression levels of circRNAs in blood samples of patients with AMI and those with mild coronary stenosis were compared to reveal circRNAs which were involved in AMI. Then, circRNAs which were significant expressed abnormally in the blood samples of patients with AMI were selected from those circRNAs. Next, a ceRNA network was constructed based on interactions of circRNA, miRNA and mRNA through biological analyses to detect crucial circRNA associated with AMI. Finally, one circRNA was selected as candidate biomarker for AMI. To validate effectivity and efficiency of the candidate biomarker, fluorescence in situ hybridization, hypoxia model of human cardiomyocytes, and knockdown and overexpression analyses were performed on candidate circRNA biomarker. In conclusion, experimental results demonstrated that the candidate circRNA was an effective biomarker for diagnosis and therapy of AMI.

Keywords: circRNA1, AMI 2, microarray 3, bioinformatics 4, circRNA_1047615

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Yi Xiong,
Shanghai Jiao Tong University, China

****Correspondence:***
Weimin Li
liweimin_2009@163.com

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

Received: *26 March 2021*
Accepted: *12 April 2021*
Published: *30 April 2021*

Citation:
Yang W, Sun L, Cao X, Li L, Zhang X, Li J, Zhao H, Zhan C, Zang Y, Li T, Zhang L, Liu G and Li W (2021) Detection of circRNA Biomarker for Acute Myocardial Infarction Based on System Biological Analysis of RNA Expression. Front. Genet. 12:686116. doi: 10.3389/fgene.2021.686116

INTRODUCTION

AMI is myocardial necrosis induced by sudden occlusion of a coronary artery (Anderson and Morrow, 2017). In the past few decades, AMI has become a significant cause of emergency medical care, hospitalization, and death in China (Gao et al., 2008; Dai et al., 2017). Globally, the incidence of AMI is increasing year by year with a serious threat to human health and survival quality (Roger et al., 2012). Early diagnosis of AMI is critical for the appropriate initiation of life-saving treatment (Jeong et al., 2020). Biomarkers, such as creatine kinase isoenzyme (CKMB)

and troponin I (TnI), are considered the gold standard for AMI. However, early diagnosis of AMI with borderline values of cardiac enzymes or waiting for serial changes could be challenging (Hajar, 2016). Therefore, a better understanding of the pathophysiological mechanisms of AMI and identifying new biomarkers for accurate and specific diagnosis are valuable.

Circular RNA (circRNA) is a type of single-stranded RNA that differs from well-known linear RNA by forming a covalently closed continuous loop (Zeng et al., 2017; Lu et al., 2019). They are generated by back-splicing of pre-mRNA transcripts, in which an upstream splice acceptor is connected to a downstream splice donor (Chen et al., 2019). The closed circular RNA was often considered a by-product of splicing error with little functional potential (Cocquerelle et al., 1993). However, based on the development of high-throughput sequencing, circRNAs have been found abundant, conserved, and specific, implying that they may possess biological and regulatory functions in the cytoplasm (Chen et al., 2019). Currently, circRNAs are found to have the following functions: they can modulate gene expression at the transcriptional or post-transcriptional level by sponging microRNAs (miRNAs) (Hansen et al., 2013; Wei et al., 2014); they can interact with RNA-binding proteins (Du et al., 2016; Wei et al., 2017); and they also have been shown to code for proteins (Begum et al., 2018). circRNA can serve as efficient miRNA sponges, interacting with miRNA to regulate mRNA expression. These specific functions and features of circRNAs suggest that they may be the ideal biomarkers to diagnose some human diseases rapidly.

Circular RNAs have been confirmed to be involved in the development of a variety of diseases (Zeng et al., 2020), including tumor system diseases (Xu et al., 2020), neurological disorders (Rybak-Wolf et al., 2015), endocrine system diseases (Gu et al., 2017), rheumatic system diseases (Luo et al., 2020), and cardiovascular diseases (CVD) (Geng et al., 2016; Wang et al., 2016) observed that circRNA HRCR acted as an endogenous miR-223 sponge to inhibit cardiac hypertrophy and heart failure. (Geng et al., 2016) found that over-expression of circRNA CDR1 *in vivo* increased the cardiac infarct size and suggested the potential of CDR1 was used as a new therapeutic target. These studies implied that circRNA may be a potential diagnostic biomarker and therapeutic target for CVD. However, few studies focused on the effect of circRNA on AMI. This study aimed to investigate the relationship between differentially expressed circRNA and AMI, and reveal the potential mechanisms via circRNA overexpression and knockdown. The ultimate goal was to provide new biomarkers for AMI diagnosis and new target for clinical treatment.

In this study, we performed a series of system biological analysis on RNA expression to find new effective circRNA biomarkers for AMI. The Arraystar Human Circular RNA Microarray Version 2.0 system was employed to detect the differential expression of circular RNAs in the whole blood of 8 patients (4 with acute myocardial infarction (AMI) and 4 with mild coronary artery stenosis). A total of 64 up-regulated and 90 down-regulated circRNAs were identified using traditional statistical methods such as Student two-sample *t* test and fold change. Therefore, five typical down-regulated circRNAs were chosen for RT-qPCR validation. The relative expression levels of 3 circRNAs (068655, 104761, and 104765) were consistent with the results of the microarray. TargetScan and miRanda databases were used to predict interactions between circRNAs and miRNAs. Furthermore, the circRNA-microRNA-mRNA network was constructed. The prediction suggests that the circRNA_104761 can sponge microRNA-449 and microRNA-34a, which are closely correlated with AMI. A larger scale sample experiment observed that the expression of circRNA_104761 was the highest in healthy volunteers, the second highest in mild coronary artery stenosis patients, and the lowest in AMI patients. The area under the receiver operating characteristic (ROC) curve for circRNA_104761 is 0.89, implying a satisfactory prediction accuracy for AMI. To further verify the role of circRNA_104761 in AMI, the hypoxia model of human cardiomyocytes AC16 was established. All the experimental results demonstrated that circRNA_104761 could not only be an effective biomarker for AMI diagnosis, but also differentiate normal coronary artery, mild coronary artery stenosis, and AMI. Furthermore, circRNA_104761 may become a potential therapeutic target.

MATERIALS AND METHODS

Overall Strategy

Abnormally expressed circRNAs often affect the occurrence and development of diseases. To discover the circRNAs related to acute myocardial infarction (AMI), the expression levels of circRNA in blood samples of patients with AMI and those with mild coronary stenosis are firstly analyzed to find abnormally expressed (up-regulated or down-regulated) circRNAs. Next, differentially expressed circRNAs between AMI patients and mild coronary artery stenosis patients were analyzed with hierarchical clustering to find out the similarity of these whole blood samples. Then, RT-qPCR was performed to detect expression levels of circRNAs that were significantly abnormally expressed in blood samples of AMI patients. Afterward, TargetScan and miRanda databases were applied to obtain the data of circRNAs-miRNAs interaction to construct a ceRNA network involving three candidate circRNA biomarkers. Finally, according to the reported data of miRNA regulation of AMI, the circRNA involved in relevant regulation progression was identified and selected circRNA_104761 as candidate biomarker.

To determine the diagnostic potential of the circRNA biomarker selected by above methods for AMI, a series of biochemical experiments were performed. First of all, the expression levels of candidate circRNA biomarker in blood samples of AMI patients, mild coronary artery stenosis patients and normal coronary artery volunteers were detected. The expression levels of candidate circRNA were significantly different in these three groups, and it indicated that the circRNA biomarker was sensitive to AMI and can be used as diagnostic marker. Secondly, hypoxia is a direct consequence of AMI and an important factor leading to death. Subsequently, the expression levels of candidate circRNA biomarker in human cardiomyocytes under different hypoxia conditions were analyzed. The expression levels of candidate circRNA were

significantly different under different hypoxia conditions, and it would suggest the circRNA biomarker can be used as molecular marker to determine the pathogenesis of AMI. Finally, the expression levels of candidate circRNA biomarker were intervened by either knockdown or overexpression, identified the influence on the occurrence and development of AMI. The overall strategy was illustrated in **Figure 1A**.

Collection of Patient Samples and Ethics Statement

Whole blood samples were collected from 34 AMI patients, 34 patients with mild coronary artery stenosis, and 30 volunteers with normal coronary arteries who attended the First Affiliated Hospital of Harbin Medical University (Harbin, China) in 2019. AMI patients were diagnosed based on acute ischaemic-type chest pain, electrocardiogram (ECG), cardiac enzyme, and coronary angiography, etc. The patients with mild coronary artery stenosis and volunteers with normal coronary arteries were diagnosed by coronary CTA. Patients were excluded from malignant arrhythmia, cardiomyopathy, valvular heart disease, malignant tumors and rheumatic immune system diseases. Blood samples from the AMI patients were collected in 10 min when they arrived at the hospital before taking any medications. Patients with mild coronary artery stenosis and normal coronary artery volunteers were recruited at the time of the fasting blood in the morning. The clinical specimens were obtained from patients who gave informed consent. In the microarray experiment, blood samples from 4 AMI patients and 4 patients with mild coronary artery stenosis were collected. circRNA_104761 was further validated in 30 AMI patients, 30 patients with mild coronary artery stenosis, and 30 volunteers with normal coronary arteries using RT-qPCR. All patients were males, and their age was recorded. The clinical characteristics of the study populations are shown in **Supplementary Table 1**. The Harbin Medical University ethics committee approved all experimental protocols for the use of human samples, and the methods were carried out in accordance with the approved guidelines.

Handing and Extraction Total RNA From Human Blood Samples

Whole blood samples (1 mL per patient) were drawn from the study donors via direct venous puncture into 2.0 mL siliconized vacuum tubes containing K2 ethylene diamine tetraacetic acid (EDTA) for Microarray analysis and RT-qPCR. After blood collection, the blood samples were immediately placed into a liquid nitrogen tank and quickly transferred to an ultra-low temperature freezer at −80°C for storage until use. This study extracted total RNA using TRI Reagent BD (Molecular Research Center, OH, United States).

Microarray Hybridization and Data Analysis

In this study, blood samples from 4 AMI patients and 4 patients with mild coronary artery stenosis were analyzed by the Arraystar Human circRNA Microarray version 2.0 system (Arraystar Inc, Rockville, MD, United States). Total RNA from each sample was quantified using the NanoDrop ND-1000. All samples' preparation and microarray hybridization were conducted based on the Arraystar's standard protocols. Briefly, total RNAs were digested with Rnase R (Epicentre, Inc.) to remove linear RNAs and enrich circular RNAs. The enriched circular RNAs were then amplified and transcribed into fluorescent cRNA utilizing a random priming method (Arraystar Super RNA Labeling Kit; Arraystar, MD, United States). The labeled cRNAs were hybridized onto the Arraystar Human circRNA version 2.0 (8x15K, Arraystar). After having washed the slides, the arrays were scanned by the Agilent Scanner G2505C. Agilent Feature Extraction software (version 11.0.1.1) was used to analyze acquired array images. Quantile normalization and subsequent data processing were performed using the R software. Before being used for the cluster analysis, the data were converted to standards. The function of dist and hclust were used to calculate distance and cluster, respectively. Hierarchical Clustering was performed to show the distinguishable expression profile of circRNAs between two groups. Differentially expressed circRNAs with statistical significance between the two groups were identified through Volcano Plot filtering. Differentially expressed circRNAs between two samples were identified through Fold Change filtering.

RT-qPCR

Total RNA was isolated from 1 mL whole blood using a phenol-chloroform extraction procedure (Jiang et al., 2019; Liu et al., 2019), and RNA extraction process was performed as previously described (Section 2.2). circRNAs' relative expression level was detected by TB Green Premix Ex Taq II (TaKaRa Bio, Shiga, Japan) with β-actin as an internal control. The validation of all the circRNAs by qPCR was performed ViiA 7 Real-time PCR System (Applied Biosystems). $2^{-\Delta\Delta}$ Ct method was used to analyze the RT-qPCR data. Primers used in RT-qPCR for validation are shown in **Supplementary Table 2**.

Cell Culture and Hypoxia Models

Human myocardial cell AC16 was purchased from Shenzhen Haodi Huatuo Biotechnology Co Ltd. The basal medium was DMEM medium supplemented with 10% fetal bovine serum (Gibco) and 1% double antibody (100 U/ml penicillin and 100 μg/ml streptomycin, Invitrogen) at 37°C and 5% CO_2 atmosphere. For the hypoxia experiments, the cells were seeded in the anaerobic mode (oxygen concentration less than 0.1%, carbon dioxide is 5%, and the rest is nitrogen), and then used after 6, 12, and 24 h of treatment.

Fluorescence *in situ* Hybridization (FISH)

First, AC16 cell climbing slices were fixed in 4% paraformaldehyde (DEPC) for 20 min, shaken, and washed 3 times with PBS (pH 7.4) on a decolorizing shaker, and digested by dropping proteinase K (20 ug/ml) for 8 min. Then, pre-hybridization and hybridization were performed, blocking serum BSA, mouse anti-digoxigenin labeled peroxidase (anti-DIG-HRP), and CY3-TSA was instilled in sequence, and stained with DAPI. Finally, we observed and collected images under fluorescence microscopy.

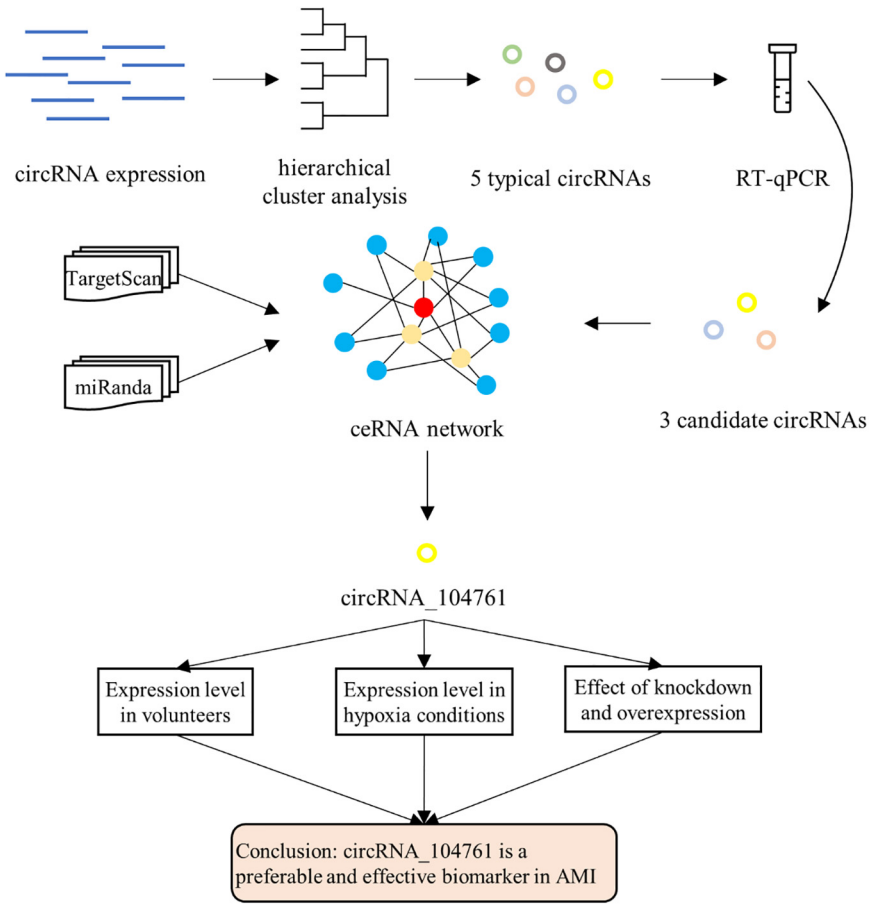

FIGURE 1 | Overall strategy.

Knockdown and Overexpression of circRNA_ 104761

The sequences of siRNA for circRNA_ 104761 used in this study were all synthesized by General Biological System (Anhui) Co Ltd, and three pairs of down-regulation primers were designed for circRNA_ 104761 at the same time. The sequences of primers are shown in **Supplementary Table 3**. The plasmid vector of circRNA_104761 overexpression model was also synthesized by General Biological System (Anhui) Co Ltd. Cell transfection was performed according to the manufacturer's instructions.

CCK-8 Assay, LDH Assay, and Apoptosis Assay

Cell Counting Kits (CCK-8 Kits) were purchased from Tongren Chemical (item number: CK04) to detect cell activity. LDH Assay Kits were purchased from Biyuntian Biotechnology Company (item number: C0017C0017). AnnexinV-FITC/PI Apoptosis Detection Kits were purchased from BD Company (item number: 556547). CCK-8 assay, LDH assay, and Apoptosis assay were all performed according to the kit's instructions.

Statistical Analysis

Statistical significance between groups was calculated by Student two-sample *t*-test. The diagnostic value of circRNAs was assessed by receiver operating characteristic (ROC) curves. SPSS statistics version 16.0 software (SPSS Inc, Chicago, IL, United States) was used to do the statistical analysis. A p-value < 0.05 was considered to be significant. The Student two-sample t-test is expressed as:

$$t(\upsilon(i)) = \frac{m_2(i) - m_1(i)}{\sqrt{\frac{s_1^2(i)}{n_1} + \frac{s_2^2(i)}{n_2}}}, \quad (1)$$

where i refers to the ith circRNA. n_1 and n_2 correspond to sample size of two groups. $m_1(i)$ and $m_2(i)$ represent the mean values of i within the samples in each group. $s_1^2(i)$ and $s_2^2(i)$ denotes the corresponding sample variances. $\upsilon(i)$ refers to the freedom. That is:

$$\upsilon(i) = \frac{(s_1^2(i)/n_1 + s_2^2(i)/n_2)^2}{s_1^4(i)/[n_1^2 \cdot (n_1 - 1)] + s_2^4(i)/[n_2^2 \cdot (n_2 - 1)]}. \quad (2)$$

In order to obtain ROC curves and the area under it (Wang et al., 2013; Zhao et al., 2015, 2017; Yang et al., 2020, 2021; Zhai et al., 2020), a certain classifier needs to be assigned. Here, we utilize

Fisher's linear discriminative analysis. A direction vector $\boldsymbol{w}$ is to be determined, where data $\boldsymbol{x}$ is to be projected to obtain a value y. That is $y = \boldsymbol{w}^t\boldsymbol{x}$. The means of the two groups can be expressed as:

$$\boldsymbol{m_1} = \frac{1}{\mathrm{n_1}}\sum_{\mathrm{x}\in \mathrm{D_1}} \mathbf{x},\ \boldsymbol{m_2} = \frac{1}{\mathrm{n_2}}\sum_{\mathrm{x}\in \mathrm{D_2}} \mathbf{x},\ \mathrm{m_1} = \frac{1}{\mathrm{n_1}}\sum_{\mathrm{x}\in \mathrm{D_1}} \mathbf{w}^{\mathrm{t}}\mathbf{x}$$
$$= \mathbf{w}^t m_1,\ m_2 = \frac{1}{n_2}\sum_{x\in D_2} \boldsymbol{w}^t\mathbf{x} = w^t\boldsymbol{m}_2, \tag{3}$$

where m_1, m_2, m_1, and m_2 correspond to mean vectors and mean values of the two sample groups before and after projection. Also, we can get:

$$|m_1 - m_2| = |\boldsymbol{w}^{\mathrm{t}}(\boldsymbol{m_1} - \boldsymbol{m_2})|. \tag{4}$$

The covariance matrix of samples between classes is:

$$\boldsymbol{S}_B = (\boldsymbol{m_1} - \boldsymbol{m_2})(\boldsymbol{m_1} - \boldsymbol{m_2})^t. \tag{5}$$

From equation (4) and equation (5), we can get:

$$(m_1 - m_2)^2 = (\boldsymbol{w}^t\boldsymbol{m_1} - \boldsymbol{w}^t\boldsymbol{m_2})^2 = \boldsymbol{w}^t(\boldsymbol{m_1} - \boldsymbol{m_2})(\boldsymbol{m_1} - \boldsymbol{m_2})^t\boldsymbol{w}$$
$$= w^t\boldsymbol{S}_B\boldsymbol{w}. \tag{6}$$

Correspondingly, the covariance matrix of within class samples can be expressed as:

$$\boldsymbol{S}_w = \sum_{x\in D_1}(\boldsymbol{x} - \boldsymbol{m_1})(\boldsymbol{x} - \boldsymbol{m_1})^t + \sum_{x\in D_2}(\boldsymbol{x} - \boldsymbol{m_2})(\boldsymbol{x} - \boldsymbol{m_2})^t. \tag{7}$$

From equation (7), we can get:

$$\begin{aligned} s_1^2 + s_2^2 &= \sum_{y\in D_1}(y - m_1)^2 + \sum_{y\in D_2}(y - m_2)^2 \\ &= \sum_{x\in D_1}(\boldsymbol{w}_t\mathrm{x}\text{-}\mathrm{w}_t\boldsymbol{m_1})^2 + \sum_{x\in D_2}(\boldsymbol{w}_t\mathrm{x}\text{-}\mathrm{w}_t\boldsymbol{m_2})^2 \\ &= \sum_{x\in D_1}\boldsymbol{w}^t(\boldsymbol{x} - \boldsymbol{m_1})(\boldsymbol{x} - \boldsymbol{m_1})^t\boldsymbol{w} \\ &\quad + \sum_{x\in D_2}\boldsymbol{w}^t(\boldsymbol{x} - \boldsymbol{m_2})(\boldsymbol{x} - \boldsymbol{m_2})^t\boldsymbol{w} = \boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w} \end{aligned} \tag{8}$$

From equation (6) and equation (8), the Optimization function is expressed as:

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{\boldsymbol{w}^t\boldsymbol{S}_B\boldsymbol{w}}{\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w}} \tag{9}$$

Correspondingly, the best direction for projection can be obtained using following derivation. That is:

$$\begin{aligned} &\frac{\partial J}{\partial \boldsymbol{w}} = \frac{(\boldsymbol{S}_B + \boldsymbol{S}_B^t)\boldsymbol{w}}{\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w}} - \frac{\boldsymbol{w}^t\boldsymbol{S}_B\boldsymbol{w}[(\boldsymbol{S}_w + \boldsymbol{S}_w^t)\boldsymbol{w}]}{(\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w})^2} = 0 \Leftrightarrow \\ &\frac{2(\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w})\boldsymbol{S}_B\boldsymbol{w} - 2\boldsymbol{w}^t\boldsymbol{S}_B\boldsymbol{w}(\boldsymbol{S}_w)\boldsymbol{w}}{(\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w})^2} = 0 \Leftrightarrow \\ &\boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w}\boldsymbol{S}_B\boldsymbol{w} = \boldsymbol{w}^t\boldsymbol{S}_B\boldsymbol{w}\boldsymbol{S}_w\boldsymbol{w} \Leftrightarrow \\ &\lambda\boldsymbol{S}_w^{-1}\boldsymbol{S}_B\boldsymbol{w} = \boldsymbol{w},\ \textit{where}\ \boldsymbol{w}^t\boldsymbol{S}_w\boldsymbol{w}/\boldsymbol{w}^t\boldsymbol{S}_B\boldsymbol{w} = \lambda \Leftrightarrow \\ &\lambda\boldsymbol{S}_w^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2})(\boldsymbol{m_1} - \boldsymbol{m_2})^t\boldsymbol{w} = \boldsymbol{w} \Leftrightarrow \\ &\boldsymbol{w} = \lambda^{'}\boldsymbol{S}_w^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2}),\ \text{where}\ \lambda^{'} = \lambda(\boldsymbol{m_1} - \boldsymbol{m_2})^t\boldsymbol{w} \end{aligned} \tag{10}$$

Regarding $\lambda^{'}$ as a scalar which can be omitted, the final direction vector w can be expressed as:

$$\boldsymbol{w} = \boldsymbol{S}_w^{-1}(\boldsymbol{m_1} - \boldsymbol{m_2}). \tag{11}$$

Therefore, Fisher's linear discriminative analysis can be expressed as:

$$\boldsymbol{w}^t\boldsymbol{x} + \boldsymbol{w}_0 = \mathbf{0} \tag{12}$$

where $\boldsymbol{w}_0 = \boldsymbol{w}\,(\boldsymbol{m_1} + \boldsymbol{m_2})\,/2$.

RESULTS

circRNA Expression Profiles of AMI and Mild Coronary Artery Stenosis Patients

In this study, 64 up-regulated and 90 down-regulated circRNAs were identified in 4 AMI patients compared with 4 mild coronary artery stenosis patients (fold change > 2.0) by microarray analysis (GSE169594), indicating these circRNAs were dysregulation. Differentially expressed circRNAs between AMI patients and mild coronary artery stenosis patients were subjected to hierarchical clustering analysis, suggesting the similarity of these whole blood samples. Hierarchical clustering revealed that the circRNA expression levels were distinguishable in the associated heat map (**Figure 2A**). A shorter distance generally indicates a high similarity. Therefore, **Figure 2A** shows that the circRNAs in the AMI patient group, and circRNAs in the mild coronary artery stenosis patient group had a relatively higher similarity. Box plot view (**Figure 2B**) shows the distribution of the hybridization data and degree of dispersion in AMI patients and mild coronary artery stenosis patients. The box plot shows that after log2 normalization, no abnormal distributions of data were observed in the 8 samples. The scatter and volcano plots shows varied circRNA expressions between the AMI and mild coronary artery stenosis samples (**Figures 2C,D**). In addition, a volcano plot identified differentially expressed circRNAs at different *p*-values and fold-changes between the two groups.

In situ Validation of the Differentially Expressed circRNAs by RT-qPCR

In terms of the microarray results, circRNAs were down-regulated greater than up-regulate, so the down-regulated circRNAs were selected for continued validation. According to circRNA fold change values, P value magnitude, basic intensity of raw signal value (RawIntensity) (recommended above 200), number and sites of circRNA-bound miRNAs and current research status of bound miRNAs, five typical down-regulated circRNAs (hsa_circRNA068655, 089763, 103149, 104761, and 104765) were chosen (shown in **Table 1**) for further RT-qPCR validation. As shown in **Figure 3**, the relative expression levels of 4 circRNAs (068655, 089763, 104761, and 104765) were down-regulated in 4 AMI patients, which were consistent with the results of the microarray. However, the *p*-value of circRNA_089763 was over 0.05, which implied a nonsignificant expression difference of circRNA_089763 in two groups. In

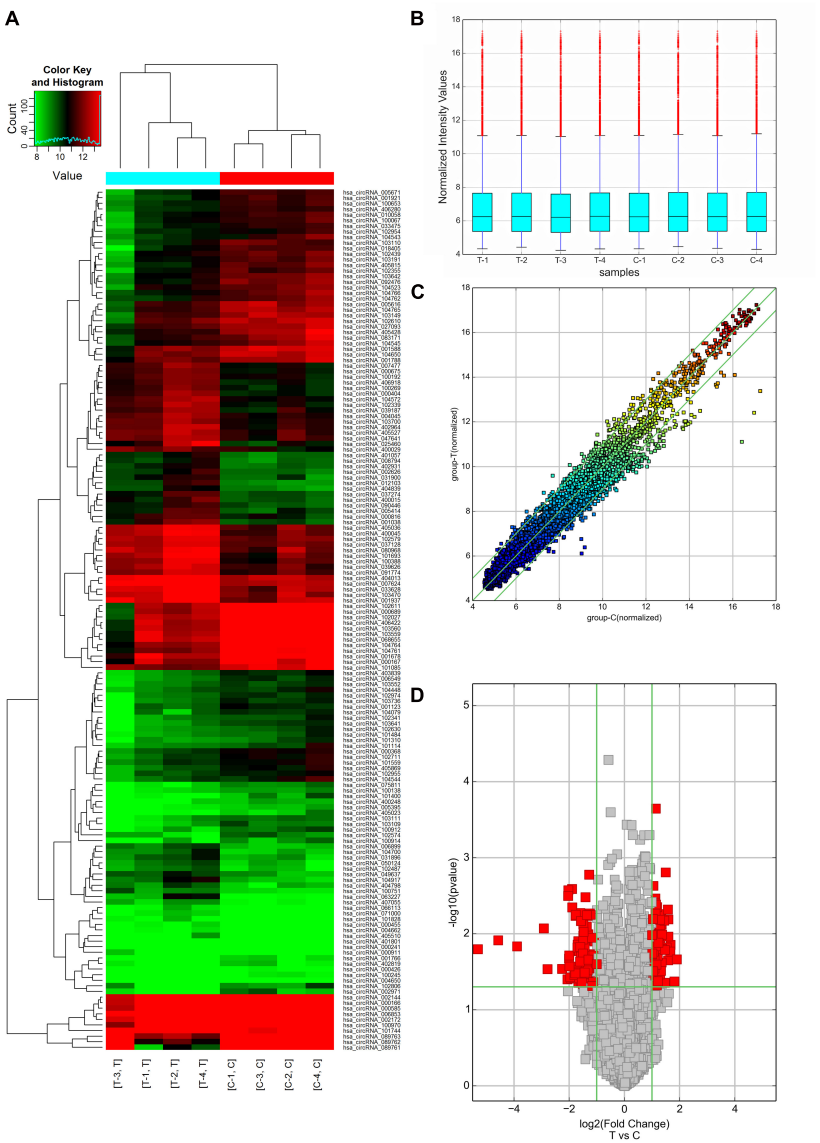

FIGURE 2 | circRNA expression profiles of AMI patients (Group T: T1, T2, T3, T4, and *n* = 4) and mild coronary artery stenosis patients (Group C: C1, C2, C3, C4, and *n* = 4) screened by microarray analysis. **(A)** Heat Map showing a distinguishable expression profile of circRNAs between two groups. Black stands for 0, indicating no change in gene expression; red represents up-regulation, and green represents down-regulation. **(B)** Boxplot view showed the distribution of normalized expression intensity values for two groups. **(C)** Scatter plot indicated the variation of circRNA expression in AMI patients (*y*-axis) and mild coronary artery stenosis patients (*x*-axis). **(D)** Volcano plots visualizing differential circRNA expression between the two groups. The vertical lines correspond to a 2.0-fold change (FC) (log2 scaled) (up-regulation and down-regulation, respectively).

addition, the relative expression level of circRNA_103149 showed the opposite results of microarray analysis. Therefore, the circRNA_068655, circRNA_104761, and circRNA_104765 could be the potential biomarkers for AMI diagnosis and potential target for AMI treatment.

Detailed Annotation for Interaction Between circRNA and miRNA

This study applied TargetScan and miRanda to find the target miRNA which interacted with circRNAs (068655, 104761, and 104765) and predict the potential biological process in which the discovered circRNAs may participate in. The interaction between circRNAs (068655, 104761, and 104765) and corresponding miRNAs are shown in **Table 2**. circRNA _104761 may bind potential target miRNAs (hsa-miR-34c-5p, hsa-miRNA-449a, hsa-miRNA 449b-5p, hsa-miRNA-449c-3p, hsa-miR-370-3p) and the secondary structure of the binding site were predicted in **Figure 4A**. A circRNA-miRNA-mRNA network was built by Cytoscape_3.7.0 as shown in **Figure 4B**. circRNA_104761 may sponge Hsa-miRNA-34a-5p, Hsa-miR-34b-5p, Hsa-miR-34c-5p, Hsa-miRNA-449a,

TABLE 1 | Five typical down-regulated circRNAs in AMI patients identified by microarray analysis.

circRNAs (has_circRNA)	Alias (has_circRNA)	Fold change	*P*-value	FDR	Regulation	circRNA type	Gene symbol
089763	0089763	14.83	0.014599771	0.684403672	Down	exonic	JA760600
104761	0001847	4.09	0.002673316	0.665381639	Down	exonic	UBAP2
068655	0068655	3.43	0.035551229	0.710909208	Down	exonic	UBXN7
104765	0001850	3.40	0.00561669	0.665381639	Down	exonic	UBAP2
103149	0002903	3.39	0.026855105	0.684403672	Down	exonic	PCNT

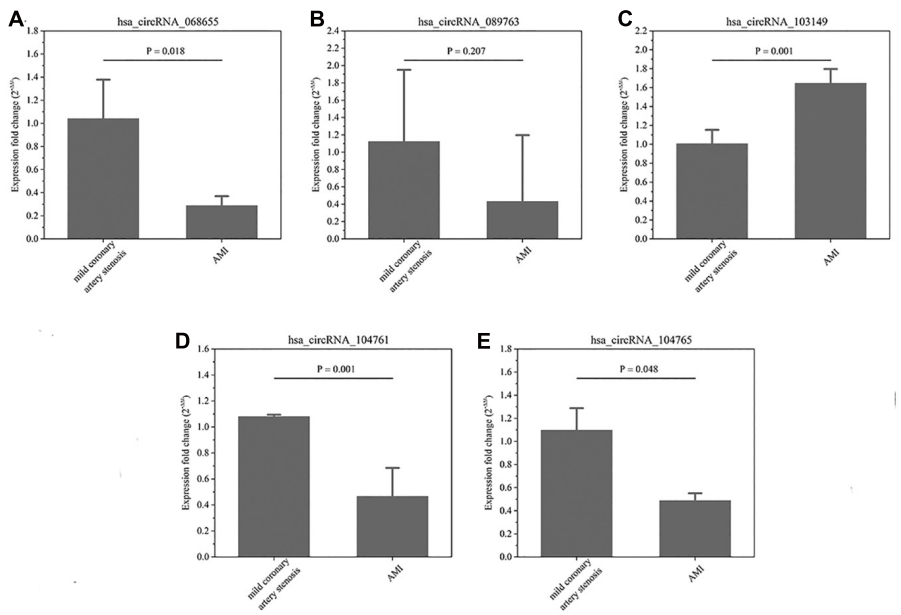

FIGURE 3 | (A–E) The relative expression of circRNAs. In-situ verification of the relative expression level of 5 down-regulated circRNAs (068655, 089763, 103149, 104761, and 104765) by RT-qPCR. Results are represented as means ± standard deviation (SD). Data shown in the graphic was analyzed by independent sample *t* tests with a significance level of 95%.

Hsa-miRNA-449b-5p, and Hsa-miRNA-449c-3p. The circRNA_104761 can sponge microRNA-449 and microRNA-34a which is correlated with AMI (Fan et al., 2013; Zhang et al., 2019). Therefore, circRNA_104761 was selected for further validation in subsequent assays.

Expression Levels of circRNA_104761 in the Whole Blood of 90 Volunteers

The differential expressed circRNA_104761 was further validated in 30 AMI patients, 30 mild coronary artery stenosis patients, and 30 normal coronary artery volunteers. As shown in **Figure 5A**, the average expression level of circRNA_104761 was significantly lower (18%) in AMI patients (0.639 ± 0.217)) than mild coronary artery stenosis patients (0.824 ± 0.216, p = 0.002), suggesting that the predicated circRNA by microarray was effective in a larger scale sample. Besides, it is worthy to note that the expression difference of circRNA_104761 between mild coronary artery stenosis group (0.824 ± 0.216) and normal coronary artery group (1.012 ± 0.235) was also significant (p = 0.002), which implied the expression of circRNA_104761 in mild coronary artery stenosis patients had been inhibited. The expression of circRNA_104761 was the highest in normal coronary artery volunteers, the second highest in mild coronary artery stenosis patients, and the lowest in AMI patients. The median cycle threshold (Ct) value for circRNA_104761 in 90 samples was 29.027, ranging from 27.623 to 32.971. These results suggest that circRNA_104761 is sensitive and abundant in human blood.

ROC Analysis of Validated circRNAs in AMI Patients and FISH

The receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) were used to confirm the relationship between circRNA_104761 and AMI. As shown in **Figure 5B**, the AUC value of the ROC curve for circRNA_104761 was 0.890 (95% confidence interval [CI] = 0.807–0.973). Meanwhile, the sensitivity and specificity of the circRNA_104761 ROC curve were 0.867 and 0.800, respectively. These results indicated that circRNA_104761 can be considered a preferable and effective biomarker for the diagnosis of AMI. RNA-FISH assay reveals that circRNA_104761

TABLE 2 | TargetScan and miRanda predicted the interaction between circRNA and microRNA.

circRNAs (has_circRNA)	T/C	*p*-value	MRE1	MRE2	MRE3	MRE4	MRE5
068655	0.27	0.00441	hsa-miR-3140-3p	hsa-miR-4539	hsa-miR-3660	hsa-miR-4260	hsa-miR-3118
104761	0.43	0.03983	hsa-miR-34c-5p	hsa-miR-449a	hsa-miR-449b-5p	hsa-miR-449c-5p	hsa-miR-370-3p
104765	0.44	0.04856	hsa-miR-532-5p	hsa-miR-496	hsa-miR-767-5p	hsa-miR-589-5p	hsa-miR-188-3p

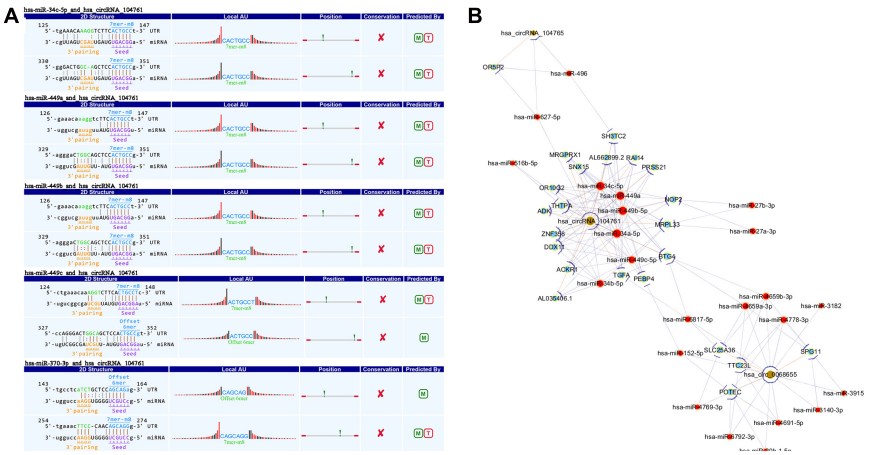

FIGURE 4 | The predicted miRNAs. **(A)** TargetScan and miRanda predicted the 5 most potential target miRNAs that may bind to circRNA _104761 and the binding site's secondary structure. **(B)** Network of circRNA (068655, 104761, and 104765) -miRNA-mRNA.

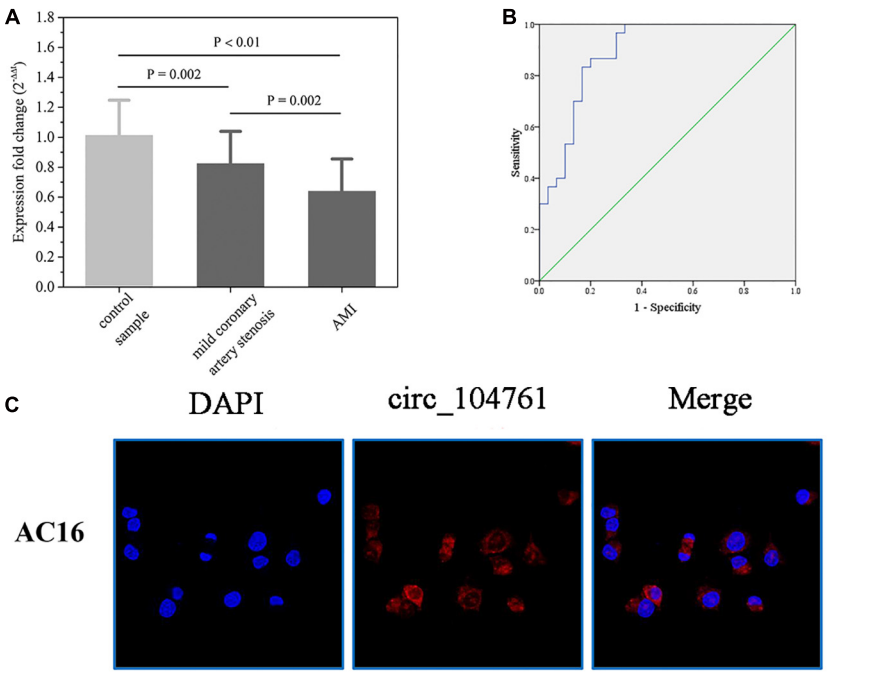

FIGURE 5 | The relative expression of circRNA _104761 in larger scale and the ROC curve analysis. **(A)** Verification of the relative expression level of down-regulated circRNA _104761 by RT-qPCR in expanded sample test. **(B)** ROC curve analysis evaluating the diagnostic value of validated circRNA_104761 for AMI. **(C)** RNA-FISH assays determined the subcellular localization of circRNA_104761 in AC16. Results are represented as means ± standard deviation (SD). Data shown in the graph was analyzed by independent sample *t*-tests with a significance level of 95%.

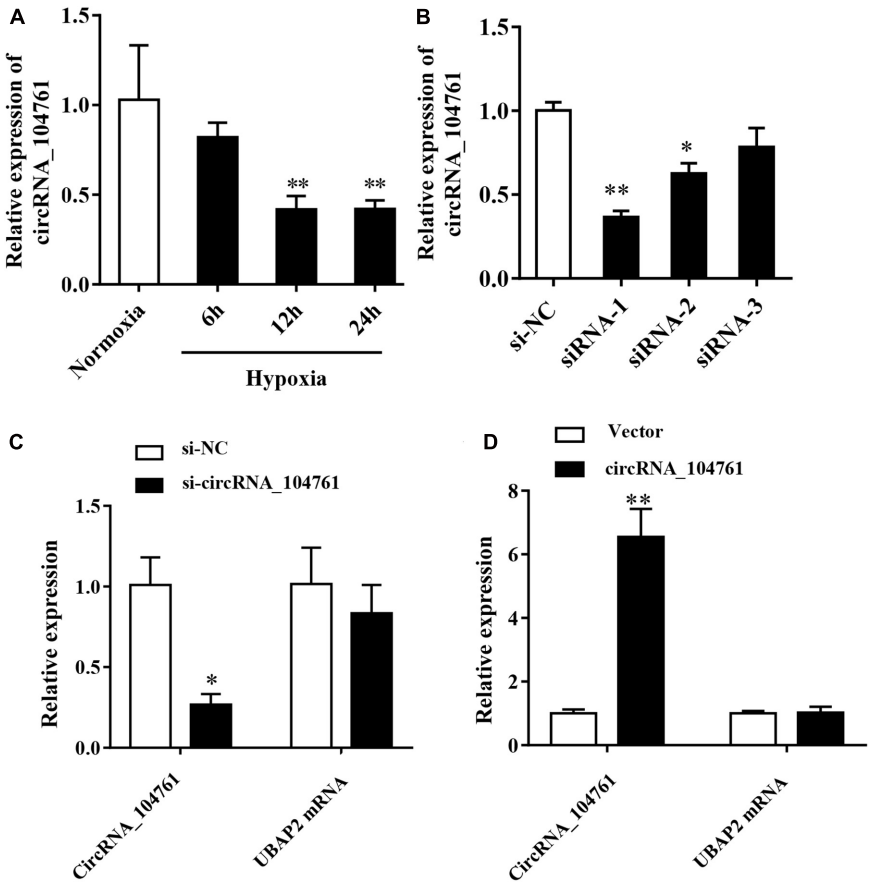

FIGURE 6 | Expression of circRNA_104761 in hypoxic human cardiomyocytes AC16 and the circRNA_104761 editing efficiency. **(A)** Expression of circRNA _ 104761 in AC16. **(B)** The down-regulation efficiency of three siRNAs (siRNA-1, siRNA-2, and siRNA-3). **(C)** The effect of siRNA-1 on endogenous circRNA_104761 and its source gene UBAP2. **(D)** The over-expression efficiencies of circRNA_104761 and UBAP2. $^{*}p < 0.05$ and $^{**}p < 0.01$.

(FAM-labeled) distinctly distributed in the cytoplasm of AC16 cells (**Figure 5C**).

Expression of circRNA_104761 in Hypoxic Human Cardiomyocytes AC16 and the circRNA_104761 Editing Efficiency

Hypoxia is an important factor causing myocardial injury. After cardiomyocytes AC16 were treated with hypoxia condition for 6 h (6 h group), 12 h (12 h group), and 24 h (24 h group), and the expression of circRNA_104761 was detected by RT-qPCR. The results showed that the expression level of circRNA_104761 in hypoxic cardiomyocytes AC16 (Hypoxia, 6, 12, and 24h) was inhibited compared with normal AC16 (Normoxia), and the expression level of circRNA_104761 in hypoxic cardiomyocytes AC16 (12h, 24h) was significantly decreased (**Figure 6A**, $p < 0.01$). Hypoxia treatment for 12 h was chosen for subsequent experiments. To further investigate the effect of circRNA_104761 on cardiomyocytes, we constructed siRNAs that could knockdown the expression of circRNA_104761. This study designed three pairs of siRNAs (siRNA-1, siRNA-2, and siRNA-3) to verify the down-regulation efficiency by RT-qPCR. The results showed that siRNA-1 and siRNA-2 down-regulation efficiency was significant (**Figure 6B**). Moreover, siRNA-1 was selected for further experiments. The source gene of circRNA_104761 was UBAP2 mRNA by circbase query, and specific siRNA-1 was able to significantly interfere with endogenous circRNA_104761, but had no significant effect on its source gene (UBAP2) (**Figure 6C**). Furthermore, we constructed a plasmid vector to overexpress circRNA_104761 in AC 16, and RT-qPCR was used to verify the overexpression efficiency. The results showed that the overexpression efficiency of the constructed plasmid vector was significant. The specific plasmid vector could significantly overexpress circRNA_104761, but had no significant effect on its source gene (UBAP2) (**Figure 6D**). Therefore, this plasmid vector was used for the following experiments.

Effect of circRNA_104761 Knockdown and Overpression in Human Cardiomyocytes AC16

After knockdown of circRNA_104761 with siRNA, LDH assay demonstrated LDH activity increased after hypoxia treatment,

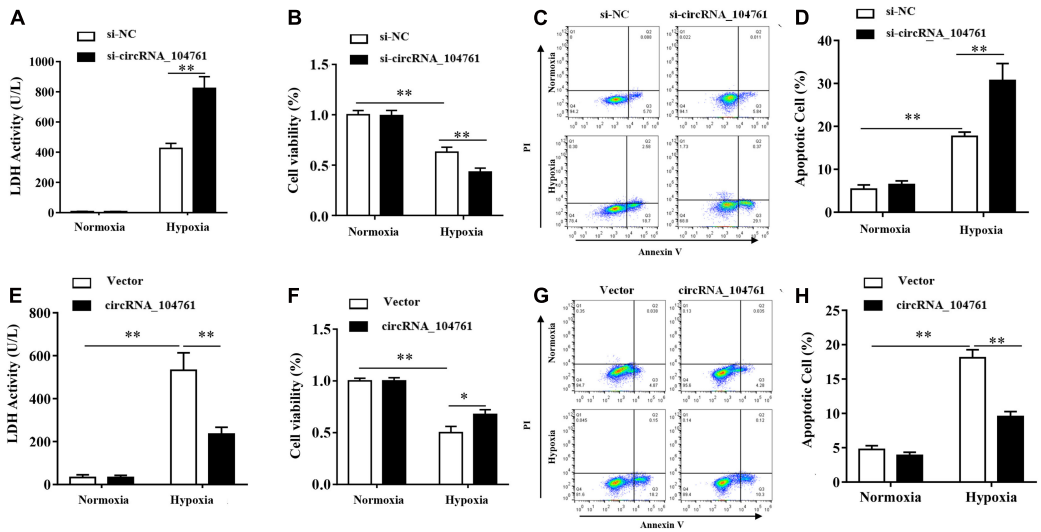

FIGURE 7 | Effect of circRNA_104761 knockdown and overpression in human cardiomyocytes AC16. **(A–D)** Effect of circRNA_104761 knockdown on AC16. **(A)** LDH activity test (Normoxia and Hypoxia). **(B)** Cell viability was determined by CCK8 assay. **(C,D)** Flow data analysis of apoptosis. **(E–H)** Effect of circRNA_104761 overexpression on AC16. **(E)** LDH activity test (Normoxia and Hypoxia). **(F)** Cell viability was determined by CCK8 assay. **(G,H)** Flow data analysis of apoptosis. Results are represented as means ± standard deviation (SD). Data shown in this graph was analyzed by independent sample *t*-tests with a significance level of 95%. $^{**}p$-value < 0.01 and $^{*}p$-value < 0.05.

and that down-regulation of circRNA_104761 significantly increased the release of lactate dehydrogenase in AC16 cell after hypoxia treatment (**Figure 7A**, p-value < 0.01). CCK8 assay demonstrated cell viability was significantly lower after hypoxia treatment, and that down-regulation of circRNA_104761 significantly reduced AC16 cell viability after hypoxia condition (**Figure 7B**), which enhanced AC16 cell death. Flow cytometry data demonstrated apoptosis was significantly increased after hypoxia, and that down-regulation of circRNA_104761 significantly exacerbated the early apoptotic level of AC16 cells after hypoxia treatment (**Figures 7C,D**).

When the expression of circRNA_104761 was overexpressed by the constructed plasmid vector, LDH activity was significantly increased after hypoxia, and the results of LDH tests demonstrated that exogenous overexpression of circRNA_104761 significantly decreased the release of lactate dehydrogenase in AC16 cell after hypoxia treatment (**Figure 7E**, $p < 0.01$). CCK8 assay demonstrated cell viability was significantly lower after hypoxia treatment, and that the overexpression of circRNA_104761 significantly increased AC16 cell viability after hypoxia treatment (**Figure 7F**). Flow cytometry data demonstrated apoptosis significantly increased after hypoxia treatment, and that overexpression of circRNA_104761 significantly alleviated the early apoptotic level of AC16 cell after hypoxia treatment (**Figures 7G,H**).

DISCUSSION

For the first time, our study applied microarray to identify the differences in circRNA expression levels between AMI patients and mild coronary artery stenosis patients. Results of microarray analysis were validated by RT-qPCR in larger samples (30 AMI patients, 30 mild coronary artery stenosis patients, and 30 normal coronary artery volunteers) and in human cardiomyocytes AC16, which implied that the expression of circRNA_104761 was an effective biomarker for AMI diagnosis. Given that circRNAs and miRNAs interact each other, circRNAs may be involved in the biological process of AMI through sponge miRNAs (Tang et al., 2018; Faiza et al., 2019; Chowdhury et al., 2020; Khan et al., 2020). The target microRNAs of circRNA_104761 were predicted by TargetScan, miRanda, and circRNA–microRNA-mRNA network, and we found that circRNA_104761 could sponge microRNA-449 and microRNA-34a. It is worth noting that miRNA-449 and miRNA-34a are closely linked to AMI.

circRNA_ 104761 may promote cardiomyocyte apoptosis through sponging miR-449. In the study of Zhang et al. (2019), they found that knocking down lncRNAX inactivation specific transcript (XIST) in the AMI rat model could down-regulate the level of miRNA-449 and inhibit rat cardiomyocyte apoptosis, suggesting that miRNA-449 is directly involved in the regulation of MI. MiRNA-449 regulate gene expression post-transcriptionally through mRNA degradation or translational repression (Esquela-Kerscher and Slack, 2006). MiRNA-449 is down-regulated in various cancers and is a strong inducer of cell cycle arrest (including senescence) and apoptosis in tumor cell lines (Bou Kheir et al., 2011). MiRNA-449 regulates various pathways (Lize et al., 2011), including Notch (Marcet et al., 2011), p53, E2F1 (Lize et al., 2010; Noonan et al., 2010),Wnt (Iliopoulos et al., 2009), and cell cycle (Bou Kheir et al., 2011). Among them, miRNA-449 provides negative feedback on E2F pathway and positive feedback on the p53 pathway, strengthening E2F1-p53 interdependence (Lize et al., 2010). In response to DNA damage,

the transcription p53 and E2F1 deregulated in cancer, and then they were activated to induce pro-apoptotic genes, which directly promote apoptosis (Lize et al., 2011). circRNA_104761 is down-regulated in AMI, which may reduce sponge miRNA-449 and change p53 and E2F-1 pathways,increasing myocardial apoptosis.

Furthermore, circRNA_ 104761 may promote cardiomyocyte apoptosis through sponging miR-34a. Fan et al. (2013) observed that miRNA-34a promoted cardiomyocyte apoptosis by negatively regulating aldehyde dehydrogenase 2 (ALDH2), which was increased in circulation under myocardial infarction (MI) conditions. In addition, the miR-34a can act as p53-responsive genes, which can induce apoptosis and cell cycle arrest in tumor cell lines (Rockenfeller et al., 2010). MiRNA-34a regulates many target proteins, which induce cell apoptosis in p53-dependent manner, including bcl-2 (Cole et al., 2008), YY1 (Chen et al., 2011), Notch (Li et al., 2009), MAPK (Tivnan et al., 2011), and DLL1 (Lewis et al., 2003), or independent manner. In AMI, the expression of circRNA_104761 is down-regulated, which may reduce the sponge function of circRNA_104761 on miRNA-34a, resulting in changes in the p53-miRNA-34a axis, causing myocardial apoptosis.

There are several limitations in our study, which cannot be ignored. First, the number of subjects is not large enough, limiting the clinical value of circRNA_104761 as a potential biomarker. Also, a more diverse control group is needed, such as patients with moderate coronary artery stenosis and patients with severe coronary artery stenosis. Second, to further illustrate the application value of circRNA_104761 in AMI, animal models with knockdown or overexpression of circRNA_104761 are needed, and this experiment only carried out cell verification. In addition, due to limited funding, we only speculated the mechanism of circRNA_104761 via miRNA-449 and miRNA-34a to cause AMI by functional analysis. The relationship between circRNA_104761 and miRNA-499/miRNA-34a needs further investigation and verification. Finally, restricting the population to Asian males limited the generalizability of the findings to females and other races.

In summary, our results demonstrated that circRNA_104761 could not only be an effective biomarker for AMI diagnosis, but also differentiate normal coronary artery, mild coronary artery stenosis,and AMI. This study also identified that knockdown of circRNA_104761 with siRNA aggravated hypoxia-induced cardiomyocytes injury in AC16, and overexpression of circRNA_104761 alleviated hypoxia-induced injury. Therefore, circRNA_104761 may be considered a potential therapeutic target.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Harbin Medical University Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WY, LS, XC, LL, XZ, and WL conceived, designed the study, and revised the manuscript. WY, JL, TL, LZ, and GL collected samples. WY, HZ, CZ, and YZ performed the experiments and analyzed the data. All authors approved the final version of the manuscript.

FUNDING

This work was supported by Heilongjiang Research Projects of Basic Scientific Research (No. 2018KYYWF-0492).

ACKNOWLEDGMENTS

The authors would like to express their gratitude to EditSprings (https://www.editsprings.com/) for the expert linguistic services provided.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.686116/full#supplementary-material

REFERENCES

Anderson, J. L., and Morrow, D. A. (2017). Acute myocardial infarction. *N. Engl. J. Med.* 376, 2053–2064.

Begum, S., Yiu, A., Stebbing, J., and Castellano, L. (2018). Novel tumour suppressive protein encoded by circular RNA, circ-SHPRH, in glioblastomas. *Oncogene* 37, 4055–4057. doi: 10.1038/s41388-018-0230-3

Bou Kheir, T., Futoma-Kazmierczak, E., Jacobsen, A., Krogh, A., Bardram, L., Hother, C., et al. (2011). miR-449 inhibits cell proliferation and is down-regulated in gastric cancer. *Mol. Cancer* 10:29. doi: 10.1186/1476-4598-10-29

Chen, Q. R., Yu, L. R., Tsang, P., Wei, J. S., Song, Y. K., Cheuk, A., et al. (2011). Systematic proteome analysis identifies transcription factor YY1 as a direct target of miR-34a. *J. Proteome Res.* 10, 479–487. doi: 10.1021/pr1006697

Chen, Z., Ren, R., Wan, D., Wang, Y., Xue, X., Jiang, M., et al. (2019). Hsa_circ_101555 functions as a competing endogenous RNA of miR-597-5p to promote colorectal cancer progression. *Oncogene* 38, 6017–6034. doi: 10.1038/s41388-019-0857-8

Chowdhury, M. R., Basak, J., and Bahadur, R. P. (2020). Elucidating the functional role of predicted miRNAs in Post-transcriptional gene regulation along with symbiosis in *Medicago truncatula*. *Curr. Bioinform.* 15, 108–120. doi: 10.2174/1574893614666191003114202

Cocquerelle, C., Mascrez, B., Hetuin, D., and Bailleul, B. (1993). Mis-splicing yields circular RNA molecules. *FASEB J.* 7, 155–160. doi: 10.1096/fasebj.7.1.7678559

Cole, K. A., Attiyeh, E. F., Mosse, Y. P., Laquaglia, M. J., Diskin, S. J., Brodeur, G. M., et al. (2008). A functional screen identifies miR-34a as a candidate neuroblastoma tumor suppressor gene. *Mol. Cancer Res.* 6, 735–742. doi: 10.1158/1541-7786.mcr-07-2102

Dai, Y., Yang, J., Gao, Z., Xu, H., Sun, Y., Wu, Y., et al. (2017). Atrial fibrillation in patients hospitalized with acute myocardial infarction: analysis of the china acute myocardial infarction (CAMI) registry. *BMC Cardiovasc. Disord.* 17:2. doi: 10.1186/s12872-016-0442-9

Du, W. W., Yang, W., Liu, E., Yang, Z., Dhaliwal, P., and Yang, B. B. (2016). Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Res.* 44, 2846–2858. doi: 10.1093/nar/gkw027

Esquela-Kerscher, A., and Slack, F. J. (2006). Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer* 6, 259–269. doi: 10.1038/nrc1840

Faiza, M., Tanveer, K., Fatihi, S., Wang, Y., and Raza, K. (2019). Comprehensive overview and assessment of microRNA target prediction tools in Homo sapiens and *Drosophila melanogaster*. *Curr. Bioinform.* 14, 432–445. doi: 10.2174/1574893614666190103101033

Fan, F., Sun, A., Zhao, H., Liu, X., Zhang, W., Jin, X., et al. (2013). MicroRNA-34a promotes cardiomyocyte apoptosis post myocardial infarction through down-regulating aldehyde dehydrogenase 2. *Curr. Pharm. Des.* 19, 4865–4873. doi: 10.2174/1381612811319990325

Gao, R., Patel, A., Gao, W., Hu, D., Huang, D., Kong, L., et al. (2008). Prospective observational study of acute coronary syndromes in China: practice patterns and outcomes. *Heart* 94, 554–560. doi: 10.1136/hrt.2007.119750

Geng, H. H., Li, R., Su, Y. M., Xiao, J., Pan, M., Cai, X. X., et al. (2016). The circular RNA Cdr1as promotes myocardial infarction by mediating the regulation of miR-7a on its target genes expression. *PLoS One* 11:e0151753. doi: 10.1371/journal.pone.0151753

Gu, Y., Ke, G., Wang, L., Zhou, E., Zhu, K., and Wei, Y. (2017). Altered expression profile of circular RNAs in the serum of patients with diabetic retinopathy revealed by microarray. *Ophthalm. Res.* 58, 176–184. doi: 10.1159/000479156

Hajar, R. (2016). Evolution of myocardial infarction and its biomarkers: a historical perspective. *Heart Views* 17, 167–172. doi: 10.4103/1995-705x.201786

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

Iliopoulos, D., Bimpaki, E. I., Nesterova, M., and Stratakis, C. A. (2009). MicroRNA signature of primary pigmented nodular adrenocortical disease: clinical correlations and regulation of Wnt signaling. *Cancer Res.* 69, 3278–3282. doi: 10.1158/0008-5472.can-09-0155

Jeong, J. H., Seo, Y. H., Ahn, J. Y., Kim, K. H., Seo, J. Y., Chun, K. Y., et al. (2020). Performance of copeptin for early diagnosis of acute myocardial infarction in an emergency department setting. *Ann. Lab. Med.* 40, 7–14. doi: 10.3343/alm.2020.40.1.7

Jiang, X.-W., Liu, Y., Huang, T.-S., and Zhu, X.-Y. (2019). MGB block ARMS Real-time PCR for diagnosis of CYP2C19 mutation in a chinese population. *Curr. Bioinform.* 14, 391–399. doi: 10.2174/1574893614666190109154252

Khan, A., Zahra, A., Mumtaz, S., Fatmi, M. Q., and Khan, M. J. (2020). Integrated In-silico analysis to study the role of microRNAs in the detection of chronic kidney diseases. *Curr. Bioinform.* 15, 144–154. doi: 10.2174/1574893614666190923115032

Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian MicroRNA targets. *Cell* 115, 787–798. doi: 10.1016/s0092-8674(03)01018-3

Li, Y., Guessous, F., Zhang, Y., Dipierro, C., Kefas, B., Johnson, E., et al. (2009). MicroRNA-34a inhibits glioblastoma growth by targeting multiple oncogenes. *Cancer Res.* 69, 7569–7576. doi: 10.1158/0008-5472.can-09-0529

Liu, W., Jiang, X., Liu, Y., and Ma, Q. (2019). Bioinformatics analysis of quantitative PCR and reverse transcription PCR in detecting HCV RNA. *Curr. Bioinform.* 14, 400–405. doi: 10.2174/1574893613666180703103328

Lize, M., Klimke, A., and Dobbelstein, M. (2011). MicroRNA-449 in cell fate determination. *Cell Cycle* 10, 2874–2882. doi: 10.4161/cc.10.17.17181

Lize, M., Pilarski, S., and Dobbelstein, M. (2010). E2F1-inducible microRNA 449a/b suppresses cell proliferation and promotes apoptosis. *Cell Death Differ.* 17, 452–458. doi: 10.1038/cdd.2009.188

Lu, Y., Deng, X., Xiao, G., Zheng, X., Ma, L., and Huang, W. (2019). circ_0001730 promotes proliferation and invasion via the miR-326/Wnt7B axis in glioma cells. *Epigenomics* 11, 1335–1352. doi: 10.2217/epi-2019-0121

Luo, Q., Zeng, L., Zeng, L., Rao, J., Zhang, L., Guo, Y., et al. (2020). Expression and clinical significance of circular RNAs hsa_circ_0000175 and hsa_circ_0008410 in peripheral blood mononuclear cells from patients with rheumatoid arthritis. *Int. J. Mol. Med.* 45, 1203–1212.

Marcet, B., Chevalier, B., Luxardi, G., Coraux, C., Zaragosi, L. E., Cibois, M., et al. (2011). Control of vertebrate multiciliogenesis by miR-449 through direct repression of the Delta/Notch pathway. *Nat. Cell Biol.* 13, 693–699. doi: 10.1038/ncb2241

Noonan, E. J., Place, R. F., Basak, S., Pookot, D., and Li, L. C. (2010). miR-449a causes Rb-dependent cell cycle arrest and senescence in prostate cancer cells. *Oncotarget* 1, 349–358. doi: 10.18632/oncotarget.167

Rockenfeller, P., Ring, J., Muschett, V., Beranekt, A., Büttner, S., Gutierrez, D., et al. (2010). Fatty acids trigger mitochondrion-dependent necrosis. *Cell Cycle* 9, 2908–2914. doi: 10.4161/cc.9.14.12346

Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B., et al. (2012). Executive summary: heart disease and stroke statistics–2012 update: a report from the American heart association. *Circulation* 125, 188–197.

Rybak-Wolf, A., Stottmeister, C., Glazar, P., Jens, M., Pino, N., Giusti, S., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* 58, 870–885. doi: 10.1016/j.molcel.2015.03.027

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

Tivnan, A., Tracey, L., Buckley, P. G., Alcock, L. C., Davidoff, A. M., and Stallings, R. L. (2011). MicroRNA-34a is a potent tumor suppressor molecule in vivo in neuroblastoma. *BMC Cancer* 11:33. doi: 10.1186/1471-2407-11-33

Wang, G., Qi, K., Zhao, Y., Li, Y., Juan, L., Teng, M., et al. (2013). Identification of regulatory regions of bidirectional genes in cervical cancer. *BMC Med. Genom.* 6(Suppl. 1):S5. doi: 10.1186/1755-8794-6-S1-S5

Wang, K., Long, B., Liu, F., Wang, J. X., Liu, C. Y., Zhao, B., et al. (2016). A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. *Eur. Heart J.* 37, 2602–2611. doi: 10.1093/eurheartj/ehv713

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE ACM Trans. Comput. Biol. Bioinform.* 11, 192–201. doi: 10.1109/tcbb.2013.146

Wei, L., Tang, J., and Zou, Q. (2017). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026

Xu, Z., Tie, X., Li, N., Yi, Z., Shen, F., and Zhang, Y. (2020). Circular RNA hsa_circ_0000654 promotes esophageal squamous cell carcinoma progression by regulating the miR-149-5p/IL-6/STAT3 pathway. *IUBMB Life* 72, 426–439. doi: 10.1002/iub.2202

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inform. Fusion* [Epub ahead of print]. doi: 10.1016/j.inffus.2021.02.015

Yang, Y. H., Ma, C., Wang, J. S., Yang, H., Ding, H., Han, S. G., et al. (2020). Prediction of N7-methylguanosine sites in human RNA based on optimal sequence features. *Genomics* 112, 4342–4347. doi: 10.1016/j.ygeno.2020.07.035

Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Computat. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform.* 21, 1425–1436. doi: 10.1093/bib/bbz080

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying antioxidant proteins by using amino acid composition and protein-protein interactions. *Front. Cell Dev. Biol.* 8:591487. doi: 10.3389/fcell.2020.591487

Zhang, M., Liu, H. Y., Han, Y. L., Wang, L., Zhai, D. D., Ma, T., et al. (2019). Silence of lncRNA XIST represses myocardial cell apoptosis in rats with acute myocardial infarction through regulating miR-449. *Eur. Rev. Med. Pharmacol. Sci.* 23, 8566–8572.

Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed. Res. Int.* 2017:7049406.

Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed. Res. Int.* 2015:861402.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 June 2021
doi: 10.3389/fgene.2021.705038

Epigenetic Marks and Variation of Sequence-Based Information Along Genomic Regions Are Predictive of Recombination Hot/Cold Spots in *Saccharomyces cerevisiae*

Guoqing Liu[1,2], Shuangjian Song[1], Qiguo Zhang[1], Biyu Dong[1], Yu Sun[3], Guojun Liu[1,2] and Xiujuan Zhao[1,2]*

[1] School of Life Sciences and Technology, Inner Mongolia University of Science and Technology, Baotou, China, [2] Inner Mongolia Key Laboratory of Functional Genomics and Bioinformatics, Inner Mongolia University of Science and Technology, Baotou, China, [3] School of Life Sciences, Inner Mongolia University, Hohhot, China

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Meng Zhou,
Wenzhou Medical University, China
Bingqiang Liu,
Shandong University, China

****Correspondence:***
Guoqing Liu
gqliu1010@163.com

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

Received: *04 May 2021*
Accepted: *07 June 2021*
Published: *29 June 2021*

Citation:
Liu G, Song S, Zhang Q, Dong B, Sun Y, Liu G and Zhao X (2021) Epigenetic Marks and Variation of Sequence-Based Information Along Genomic Regions Are Predictive of Recombination Hot/Cold Spots in Saccharomyces cerevisiae. Front. Genet. 12:705038. doi: 10.3389/fgene.2021.705038

Characterization and identification of recombination hotspots provide important insights into the mechanism of recombination and genome evolution. In contrast with existing sequence-based models for predicting recombination hotspots which were defined in a ORF-based manner, here, we first defined recombination hot/cold spots based on public high-resolution Spo11-oligo-seq data, then characterized them in terms of DNA sequence and epigenetic marks, and finally presented classifiers to identify hotspots. We found that, in addition to some previously discovered DNA-based features like GC-skew, recombination hotspots in yeast can also be characterized by some remarkable features associated with DNA physical properties and shape. More importantly, by using DNA-based features and several epigenetic marks, we built several classifiers to discriminate hotspots from coldspots, and found that SVM classifier performs the best with an accuracy of ∼92%, which is also the highest among the models in comparison. Feature importance analysis combined with prediction results show that epigenetic marks and variation of sequence-based features along the hotspots contribute dominantly to hotspot identification. By using incremental feature selection method, an optimal feature subset that consists of much less features was obtained without sacrificing prediction accuracy.

Keywords: recombination hotspots, DNA physical property, classifier, epigenetic mark, optimal feature set

INTRODUCTION

Meiotic recombination is crucial to gametogenesis as it helps the faithful separation of homologous chromosomes into gametes by forming chiasma (Coop and Przeworski, 2007). Abnormal or no recombination between homologous chromosomes would cause aneuploidy in gametes and affect health in offspring. For example, 10–30% of zygotes are aneuploid and approximately 30% of

maternally derived cases with chromosome mis-segregation are associated with failure of crossover formation (MacLennan et al., 2015). Recombination also attracts researchers' attention because it drives genome evolution by producing genetic diversity (Webster and Hurst, 2012).

During meiosis, DNA double-strand break initiates recombination at leptotene stage of first round of meiotic division (MI) (Baudat et al., 2013). Only a few of DSB sites across a chromosome are selected to designate cross-over (CO) that is followed by CO maturation (Wang et al., 2017). DSB hot sites are strongly correlated with recombination rate, and hence are used to indicate recombination hotspots. In contrast with hotspots, coldspots refer to the genomic regions undergo no or extremely low level of recombination. Recombination rate is unevenly distributed along chromosomes, but it is still unclear that how hotspots are arranged across the genome. DNA sequence features like PRDM9-binding motif (Myers et al., 2008), GC content (Galtier et al., 2001), GC-skew (Smagulova et al., 2011), SNP pattern (Pratto et al., 2014), and dinucleotide bias (Liu and Li, 2008) were known to correlate recombination rate, but the effects of DNA physical properties and DNA shape on recombination need further investigation.

Computational identification of recombination hotspots may help people get quick information about recombination and relieve the time-consuming experimental determination of hotspots with high cost. As reviewed in Yang et al., 2020, there are some existing models for hotspot identification at present (Zhou et al., 2006; Jiang et al., 2007; Liu et al., 2012, 2017; Chen et al., 2013; Li et al., 2014; Qiu and Xiao, 2014; Jani et al., 2018; Zhang and Kong, 2019; Khan et al., 2020). Almost all of the models were DNA sequence dependent and epigenetic marks that have been increasingly freely available were not considered. For example, nucleosome depletion (Pan et al., 2011) and H3K4me3 mark (Borde et al., 2009) were not considered in the models. Although in our previous study, we attempted to include the effect of nucleosome occupancy (Zhang and Liu, 2014), the use of MNase-seq data derived from non-meiotic cells may not provide reliable information. In fact there are more and more chromatin level factors and DNA-protein binding have been shown to affect recombination (Getun et al., 2010; Zhang et al., 2011; Cesarini et al., 2012; de Castro et al., 2012; Sommermeyer et al., 2013; Yamada et al., 2013; Gittens et al., 2019; Pyatnitskaya et al., 2019; Heldrich et al., 2020; Karányi et al., 2020; Paiano et al., 2020; Serrano-Quílez et al., 2020). In addition, DNA shape and physical properties were also shown to affect recombination hotspot identification (Chen et al., 2013), but the importance of individual DNA shape feature is unclear because they were implicitly incorporated in the model in the form of pseudo nucleotide composition. Furthermore, as far as we know, DNA shape parameter sets derived from different groups differ a lot (Liu et al., 2016), suggesting that the accuracy of the parameter estimation is unclear. In this aspect, it is also worth noting that the DNA shape parameters are sequence context-dependent (Zhou et al., 2013), and context-dependent estimation of DNA shape parameters may assist hotspot prediction. Indeed, DNA shape features were used in the prediction of DSB sites (not meiotic DSB sites) in human cell lines (Mourad et al., 2018).

In this study, we first characterized the recombination hot/cold spots with regard to DNA sequence-based features and some other features like histone modification and Top2 binding signal, and then developed several classifiers to discriminate recombination hotspots from coldspots. Comparison with other models demonstrated the good performance of our model.

MATERIALS AND METHODS

Benchmark Datasets

Benchmark datasets here include two datasets: positive and negative dataset. Positive dataset consists of 3,600 recombination hotspots defined by other group based on high-resolution Spo11-oligo sequencing data (Pan et al., 2011). Generally speaking, the construction of negative dataset is much trickier than positive one in binary classification, because the negative samples are much more enriched than positive samples, leading to unbalance between positive and negative dataset. Moreover, negative samples selected to represent non-positive samples may include a big noise. For example, there is a tremendous number of "non-hotspot" regions in the genome, but recombination rate at those regions are not necessarily low because they are just undetected by peak calling algorithm for hotspot identification. To address this problem, we defined negative dataset of recombination coldspots as the genomic regions of at least 500 bp long with no Spo11-oligo signal (zero value) based on the full Spo11-oligo map (Pan et al., 2011). Defining coldspots in this way, we focus on relatively large cold regions with low recombination, which may not result from the noise or limited sequencing depth in Spo11-oligo seq. To give a visual inspection, a plot of hot/cold spot regions along with Spo11-oligo signal is shown (**Figure 1**). The final benchmark consists of 3,600 hotspots and 2,538 coldspots. The length distribution of the hot/cold spots sequences was provided in Supplementary Information (**Supplementary Figure 1**). All datasets used in this study were provided in Supplementary data (**Supplementary Table 1**).

It should be highlighted that the hotspots and coldspots used in this study are not defined as in previous models in ORF-based way (Zhou et al., 2006; Jiang et al., 2007; Liu et al., 2012, 2017; Chen et al., 2013; Li et al., 2014; Qiu and Xiao, 2014; Jani et al., 2018; Zhang and Kong, 2019; Khan et al., 2020), but are based on the high-resolution Spo11-oligo seq data. In this way we train our models on "true" hotspots, rather than on hot/cold ORFs that are not necessarily equivalent to "true" hotspots.

Feature Extraction

Three types of features are used in our prediction (**Table 1**): sequence compositional information, DNA physical properties and non-DNA features. Features that indicates sequence compositional information includes: GC content, GC-skew, mutual information and k-mer composition. Features used to reflect DNA physical properties include DNA shape parameters (Zhou et al., 2013), DNA rigidity, etc. Non-DNA features we used include some epigenetic marks (H3K4me3 and H3K56ac),

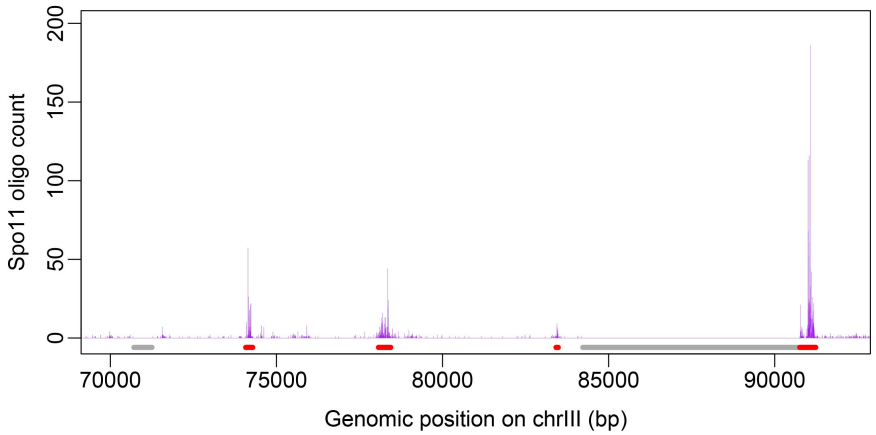

FIGURE 1 | Distribution of hot/cold spots along chromosome is shown with Spo11-oligo signal taken from Pan et al. (2011). Genomic regions marked in red denote hotspots and gray represent coldspots defined in this study.

MNase-seq signal, and Top2 binding signal. These features are calculated in the following way.

$$p_t(k) = \begin{cases} \frac{N_t}{\sum_{t=1}^{4^k} N_t} & k = 1, 2 \\ \frac{N_t+1}{\sum_{t=1}^{4^k} N_t+4^k} & k = 3, 4, 5, 6 \end{cases} \quad (1)$$

$$GC{-}content = \frac{N_G + N_C}{N_A + N_T + N_G + N_C} \quad (2)$$

$$GC\text{-}skew = \frac{N_G - N_C}{N_G + N_C} \quad (3)$$

$$MI = \sum_{i,j} p_{ij} \log_2 \frac{p_{ij}}{p_i p_j} \quad (4)$$

where N_i represents the occurrence number of nucleotide i in a DNA sequence; p_i or p_j $(i, j = A, G, C, T)$ is the fraction of nucleotide i or j and p_{ij} is the fraction of dinucleotide ij in a sequence. Mutual information (MI) describes the overall deviation of observed probabilities of dinucleotides from those expected from mononucleotide probabilities (Luo et al., 1998). $p_t(k)$represents the composition of t-th k-mer (oligonucleotide of k bp in length) in a sequence, which refers to the occurrence probability of the k-mer counted by a sliding step of 1 bp along the sequence. To avoid the shortcoming caused by small sequence length in the calculation of k-mer compositional probability, Laplacian correction was done for k-mers where $k > 2$ [see eq. (1)].

DNA shape parameters were calculated at base pair step resolution using R package DNAshapeR (Zhou et al., 2013). With respect to DNA physical property, we also used the parameter set collected in a previous study (Chen et al., 2012), three DNA thermodynamic property parameters including Gibbs free energy, entropy and enthalpy (Ignatova et al., 2008), DNA rigidity (Scipioni et al., 2002; Liu et al., 2018), and parameter set including equilibrium base-pair step parameters (**Supplementary Figure 2**) and force constants which were estimated in our previous study by using crystal structure of protein-DNA complexes (Liu et al., 2019, 2021). The values of the parameters were listed in **Supplementary Tables 2–4**.

Sequence-based features including sequence-compositional information, DNA shape features, and DNA physical properties were calculated by merely using the DNA sequence as input. At first, we retrieved 1000-bp sequence for each hot/cold spot from the genome of *Saccharomyces cerevisiae* (SacCer3). Then, sequence-based features were calculated. GC-content, GC-skew, and MI were calculated along the sequence by using a sliding window of 100, 100, and 200 bp, respectively. K-mer composition was calculated for central 300-bp (or 150- and 500-bp) regions of the sequences. Other sequence-based parameters (DNA shape features and DNA physical properties) were calculated at each base-pair step and smoothed with a 10-bp average. Based on these data, distribution profile plots for the features (e.g., **Figures 2–4**) were generated. Finally, mean and variance of the sequence-based parameters along the central 300 bp were calculated and used as final features in the prediction. Calculated variance hear measures the variation of sequence-based parameters along the sequence. Utilizing the processed data available online, non-DNA features were calculated by averaging the signals within 300 bp regions at hot/cold spots. Variance was not calculated for non-DNA features.

Classifiers

Random Forest

Random Forest (RF) is one of the widely used ensemble learning algorithms (Breiman, 2001). It generates numerous decision trees based on the training set and then majority voting strategy is used to label the class of the sequences in the test set. Its success in various fields is ascribed partially to de-correlating the bootstrap sampling decision trees by random sampling sub-sized features from the whole feature space at each splitting node. A RF-based model was developed to classify recombination hot/cold spots by using R package "randomForest". To be specific, after the

TABLE 1 | Features used in this study.

Feature type	Features	Feature extraction manner	Feature number ($96 + 4^k$)	References
DNA composition	GC content	Mean + var	2	–
	GC-skew	Mean + var	2	–
	MI	Mean + var	2	Luo et al., 1998
	K-mer composition	Overall	4^k	–
DNA shape	MGW, HelT, rise, roll, shift, slide, tilt, buckle, opening, ProT, shear, stagger, and stretch	Mean + var	13 × 2	Zhou et al., 2013
DNA physical properties	EP	Mean + var	2	Zhou et al., 2013
	Rigidity	Mean + var	2	Scipioni et al., 2002
	Gibbs free energy	Mean + var	2	Ignatova et al., 2008
	Enthalpy	Mean + var	2	Ignatova et al., 2008
	Entropy	Mean + var	2	Ignatova et al., 2008
	Parameter set (Chen)	Mean + var	12 × 2	Chen et al., 2012
	Parameter set (Liu)	Mean + var	12 × 2	Liu et al., 2021
Non-DNA features	H3K4me3 (GSE11004)	Mean	1	Borde et al., 2009
	H3K56ac (GSE37487)	Mean	1	Karányi et al., 2020
	H3K4me3 (GSE59005)	Mean	1	Hu et al., 2015
	H3K56ac (GSE59005)	Mean	1	Hu et al., 2015
	MNase-seq (GSE59005)	Mean	1	Hu et al., 2015
	Top2-CC-seq (GSE136675)	Mean	1	Gittens et al., 2019

MGW, minor groove width; ProT, propeller twist; HelT, helical twist; EP, electrostatic potential; Parameter set (Chen) include 12 features collected in Chen et al. (2012); Parameter set (Liu) include force constants and equilibrium structure parameters for 10 unique dinucleotides presented in Liu et al. (2021). Data of Top2 CC-seq used here refers to the processed data of VP16-treated sample (RA7-RA13_Cer3H4L2_MJ551_pdr1mre11_VP16.FullMap); H3K4me3, H3K56ac, and MNase-seq data were derived from meiotic cells at 4 h during sporulation when recombination initiates. For some non-DNA features, data resolution is not high enough (e.g., H3K4me3_GSE11004), which would impede us to obtain reliable high-resolution variation patterns of the features at hot/cold spots. Therefore, variances of non-DNA features along hot/cold spots were not considered.

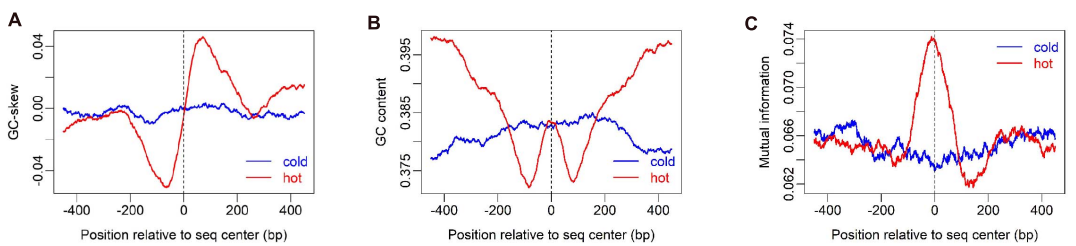

FIGURE 2 | Sequence compositional feature profile of hot/cold spots. GC-skew **(A)**, GC-content **(B)**, and Mutual information **(C)** were calculated along sequences by using a sliding window of 100 bp, 100 bp, and 200 bp respectively.

benchmark dataset was prepared, we characterized each sequence and prepared feature matrix for benchmark dataset. The number of features sampled from the feature space at each splitting point was set to $\log_2 m$ where m is total number of features in feature space. Optimal number of decision trees generated in the RF was set to 130 by inspecting Error-tree plot. Five-fold cross-validation was performed to evaluate the model.

Support Vector Machine

Support vector machine (Cortes and Vapnik, 1995) is an efficient classifier which has been widely used to solve classification and regression tasks. In SVM algorithm, input data (feature data) is mapped to a new feature space with higher dimension by using a kernel function and then optimal separating hyperplane is determined in the new feature space. In the current study, linear kernel was used to implement SVM-based classification using R package "e1071" with default values for all other parameters.

Logistic Regression

Logistic regression model is a generalized linear model that is used to predict the probability of a binary (yes/no) event occurring based on a set of independent variables (Collins et al., 2004; Nick and Campbell, 2007). In brief, the model the outcome of multiple regression is mapped to logistic function (sigmoid function), which is then transformed to eq. (5) by logit transform and the result of a binary event is predicted based on a

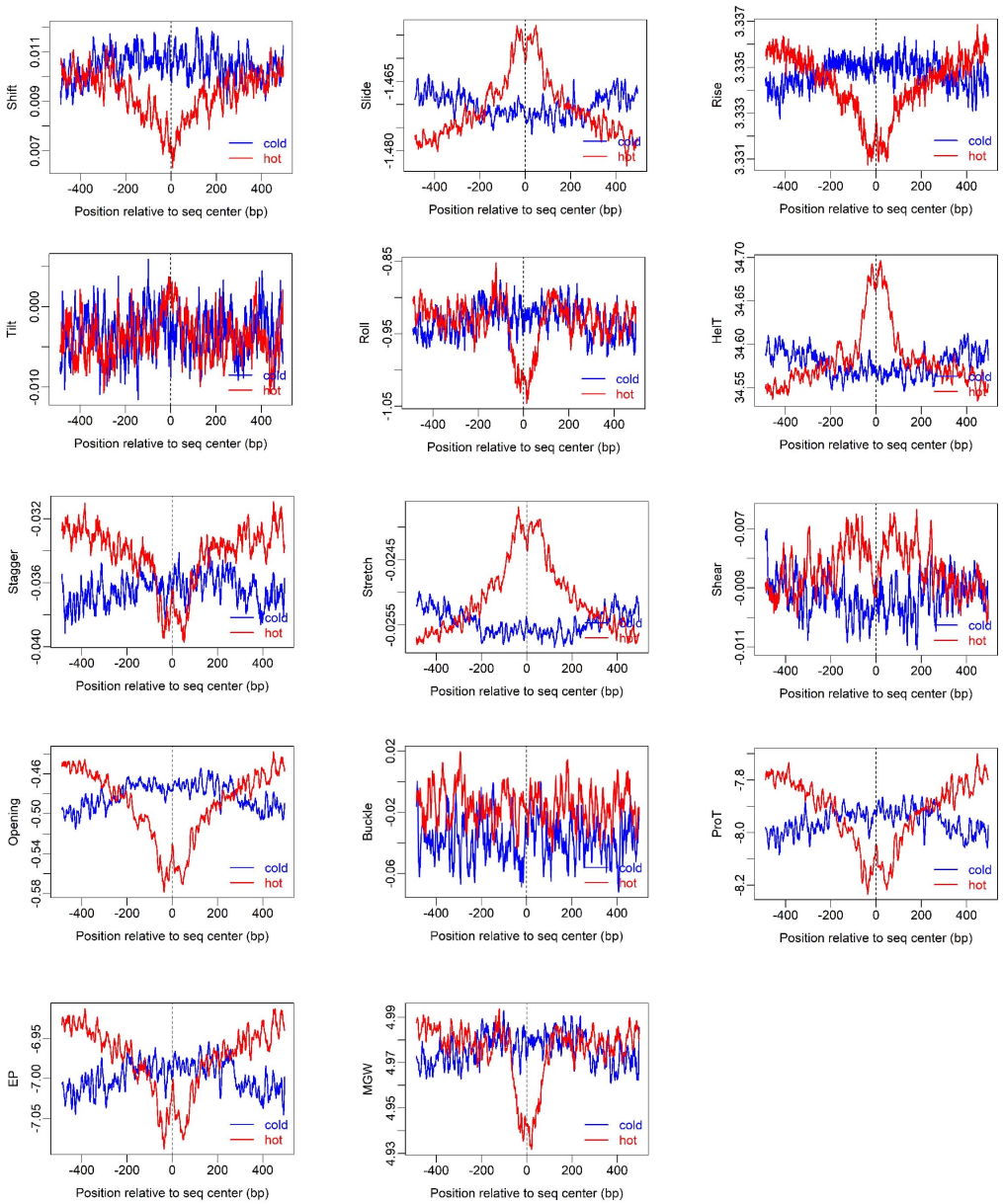

FIGURE 3 | Distribution of DNA shape and physical properties at hot/cold spots. The plots were smoothed with a 10-bp moving average.

threshold value (e.g., 0.5). In our model, independent variables are sample features, and the dependent variable is the label of the sample (e.g., hotspot or coldspot). The regression coefficients are estimated based on train dataset, and the outcomes of test samples are predicted.

$$\text{logit(p)} = \ln\frac{p}{1-p} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \quad (5)$$

Naive Bayesian Classifier

Naive Bayesian classifier is a simple and fast classification algorithm (Friedman et al., 1997), which has been successfully used for many machine learning purposes and works particularly well in text classification. It uses Bayes' Theorem to predict the label of a sample. "Naive" means the assumption that the occurrence of features is independent with each other, and thus likelihood $P(x|c)$ is calculated as the product of each feature's likelihood $P(x_i|c)$ as indicated in eq. (6). Likelihood probability for each feature is estimated by a Gaussian model. Then two posterior probabilities are calculated for each test sample by using Bayes theorem and the larger probability indicates the class (label) of the sample.

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)}\prod_i P(x_i|c) \quad (6)$$

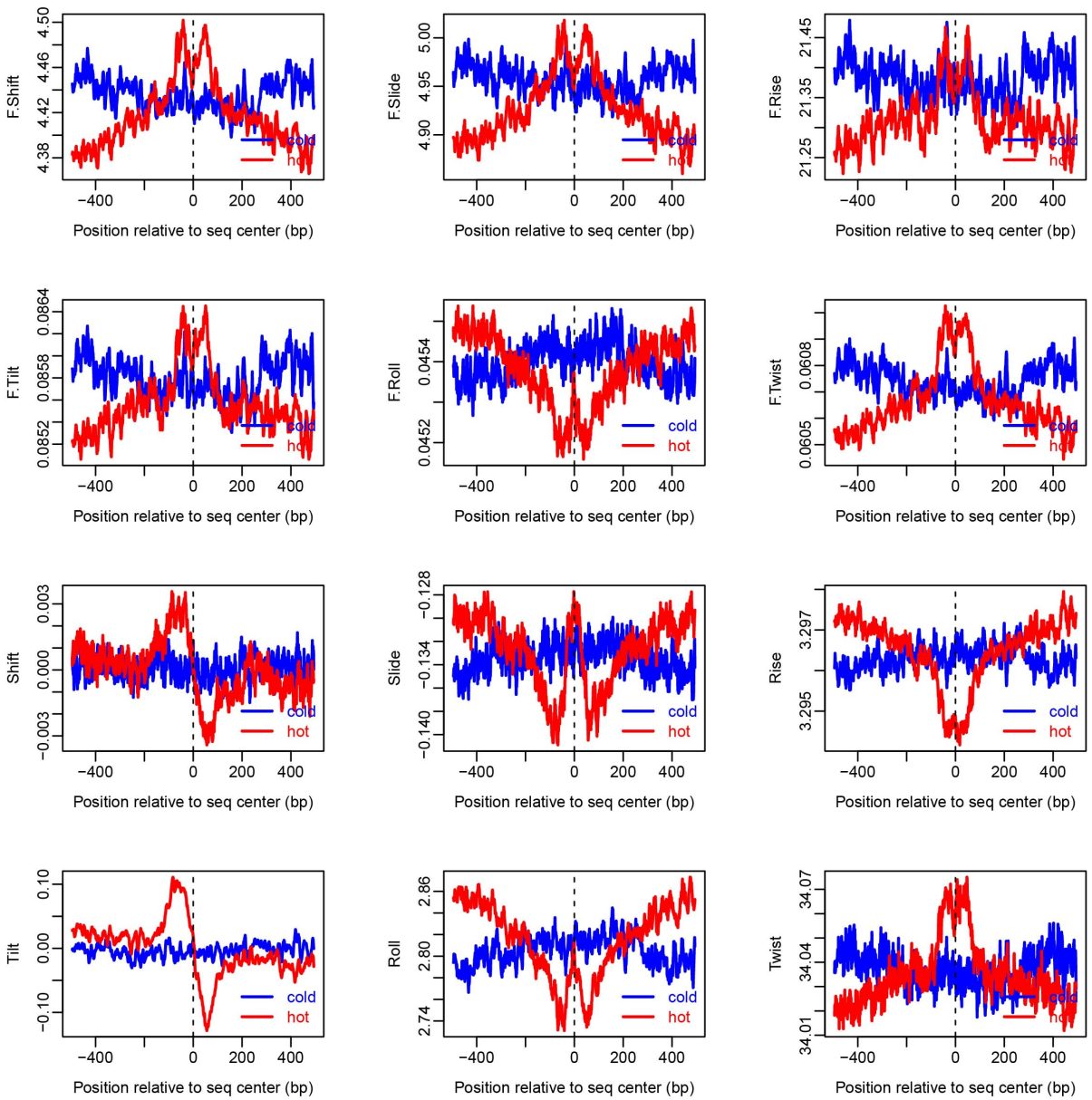

FIGURE 4 | Distribution of DNA base-pair-step parameters at hot/cold spots. The plots were smoothed with a 10-bp moving average. The base-pair-step parameters were taken from Liu et al., 2021 (**Supplementary Table 2**).

Where$P(c|x)$is posterior probability that represents the probability of observing class c (c = hotspot or coldspot) given feature set x, $P(c)$ is prior probability, and $P(x|c)$ is class-conditional probability (likelihood).

Decision Tree

Decision tree describes the classification process of samples based on features (Quinlan, 1986). In other words, it consists of a series of decision rules that divide samples contained in each node into two or more subsets according to a feature-based decision rule. Decision tree begins with a root node representing training samples, and recursively generates new branches and nodes by using feature-based "if-then" rule until the node cannot be further classified. The final nodes are called leaf nodes. At each decision step, the best feature is used. Best feature for each node (root node or internal decision node) can be selected by a quantitative measurement method such as Gini index or Information Gain. Based on training data-based decision tree, the labels of test samples are predicted. The typical algorithm of decision tree is CART (Breiman et al., 1984), and we used R package "rpart" to develop CART-based decision tree classifier (parameters used in rpart function: method = "class," cp = 0.000001).

Assessment of Model Performance

Five-fold cross-validation was performed for each of the five classifiers introduced above, and overall performances were reported. The performance of classification model is quantified by widely used metrics including Sensitivity (*SN*), Specificity (*SP*), Accuracy (*ACC*), *F-measure*, and Area Under ROC curve (AUC)

$$SN = \frac{TP}{TP + FN} \tag{7}$$

$$SP = \frac{TN}{TN + FP} \tag{8}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$F-\text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FN + FP} \tag{10}$$

where *TP*, *FN*, *TN*, and *FP* denote, respectively, the numbers of true positive, false negative, true negative, and false positive samples. *F-measure* is the harmonic mean of the precision and recall.

RESULTS AND DISCUSSION

Characterization of Hotspots

To show how DNA-based features distribute at hot/cold spots, we plotted the average profile of DNA-based parameters at hot/cold spots (**Figures 2–4**). It is apparent that some of the parameters exhibit a clear characteristic pattern at hotspots, contrasting with random distributions at coldspots. For example, GC-skew shows a characteristic reversed skew between the two sides of the hotspot center, probably due to mutational bias (Smagulova et al., 2011); Mutual information has a dramatic peak at hotspot center (**Figure 2**), suggesting the possible biased usage of dinucleotides (Liu and Li, 2008); DNA shape parameters such as slide, shift, rise, helical twist, roll, stretch, opening, propeller twist, and minor groove width (MGW) show a peak or dip at the hotspot center (**Figure 3**). The force constants reflecting the deformation rigidity with regard to corresponding degrees of freedom also differ between hotspots and coldspots (**Figure 4**). It is worth noting that some of the distribution patterns of base-pair step parameters calculated based on our previously estimated parameter set (**Figure 4**) differ from DNAshapeR-based results (**Figure 3**). For example, both tilt and shift exhibit an anti-symmetric pattern with respect to hotspot center in **Figure 4**, while this pattern is absent for DNAshapeR-based results (**Figure 3**). It would be interesting if the specific patterns observed in **Figure 4** represent an intrinsic feature of recombination hotspots. We also presented the distribution patterns of some other DNA physical properties at hot/cold spots (**Supplementary Figure 3**).

In addition, we also analyzed the difference of several epigenetic signals between hotspots and coldspots (**Figure 5**). The results show that H3K4me3, H3K56ac, MNase-seq signal, and Top2 binding signal differ between hotspots and coldspots. High levels of H3K4me3 and H3K56ac and reduced MNase-seq signal at hotspot center are usually used to indicate high chromatin accessibility. The enrichment of top2 binding at hotspots was reported previously (Gittens et al., 2019). It is unexpected that two H3K56ac datasets show different enrichment patterns (**Figure 5**), and the reason for the discrepancy is unclear.

Performances of Classification Models

DNA-Based Prediction

We first focus on DNA-based prediction as many others done before. DNA-based features can be divided into two types: DNA compositional features and DNA physical properties. Let's start with DNA compositional features.

Our previous study as well as others' show that k-mer composition is related to recombination hotspots (Liu et al., 2012). To gain knowledge about which size of k-mer (k = 1–6) has the best predictive ability to discriminate hotspots from coldspots, we trained classifiers on k-mer probability features, where k ranges from 1 to 6, and predicted the class of test set samples. Our results based on five-fold cross-validation show that 4-mer composition is the best predictor (**Supplementary Table 5** and **Figure 6**), achieving an accuracy of ∼83.7% by SVM-based classifier. Among the five classifiers, SVM performs the best, followed by logistic regression and RF. Naive Bays classifier is unstable when k is larger than four, which might be caused by inadequate sampling of k-mers in short sequences (300-bp) we used. Because many k-mers when k is 4–6 have zero occurrence in a short DNA sequence, and the derived probability of zero for the k-mers does not represent true case. Even if we introduced pseudo-count to smooth the k-mer probability, Naive Bays classifier still performs badly. Particularly for Naive Bays classifier, Gaussian distribution-based maximum likelihood estimate of posterior probability is unreliable, or even un-computable, because many zero values of k-mer occurrence (or homogeneous value of smoothed probability) may result in the variance of zero for a particular k-mer feature in feature space (4^k features), making the Gaussian probability density used in maximum likelihood estimate of posterior probability un-computable. In addition, predictions based on sequences shorter or longer than 300 bp (e.g., 150 and 500-bp) could not generate improved accuracy, suggesting that 300 bp is a proper window size for hotspot prediction (**Supplementary Table 5**).

The second class of DNA-based features is DNA physical properties, which impact DNA deformation such as DNA bending, stretching, base-paring and stacking. DNA shape parameters were included in this category. When predicting hot/cold spots based on this feature set, a worse prediction accuracy (**Supplementary Table 6**, SVM: ACC = 80.3%) than the 4-mer compositional features (**Supplementary Table 5**, SVM: ACC = 83.7%) was obtained (**Figure 7**). Again, predictions based on 300-bp window-based feature extraction are better than 150- and 500-bp window (**Supplementary Table 6**).

We then ask if the variation of sequence-based parameters along the sequences (see **Figures 2–4**) contributes to hot/cold spot classification. To test this, we included the variance of

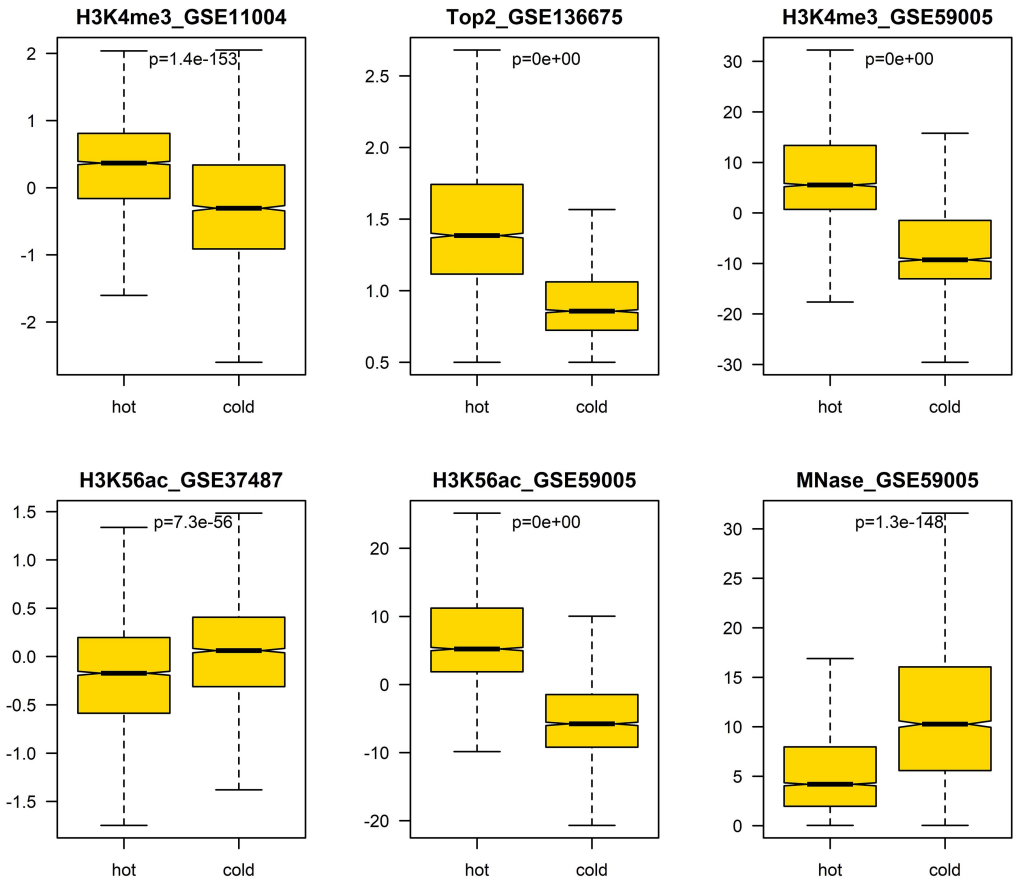

FIGURE 5 | Difference of epigenetic marks between hot/cold spots. The *p*-values were obtained via *t*-test.

the sequence-based features along the sequences in feature set, and made predictions. The results show that the variation of the parameters indeed remarkably improved the prediction performance (**Supplementary Table 7** and **Figure 7**, ACC = 85.4 vs. 80.3%). Combination of all the DNA-based features produced a prediction accuracy of 85.6% (**Figure 7** and **Supplementary Table 8**).

Non-DNA Features Are a Strong Predictor of Recombination Hotspots

After evaluating the influence of DNA sequence information on discriminating hotspots and coldspots, we then sought to uncover how non-DNA features affect the identification of hotspots. Based on prior knowledge discovered in other experimental studies, we considered several types of non-DNA features: MNase-seq signal, histone modification signals (H3K4me3 and H3K56ac), and Top2 signal. It is apparent that this non-DNA feature set is capable of classifying hot/cold spots with a much higher accuracy (**Figure 6J**, AUC = 0.969) than DNA sequence-based features (**Figure 6I**, AUC = 0.922). It is unexpected that H3K56ac signal difference between hotspots and coldspots differs between two independent studies from which we obtained H3K56ac data (**Figure 5**). But in both studies (Hu et al., 2015; Karányi et al., 2020), H3K56ac was claimed to have positive contribution to recombination, probably due to H3K56ac-promoted chromatin accessibility which favors the binding of recombination machinery to hotspots. We therefore carried on prediction after removing the unexpected H3K56ac feature (H3K56ac_GSE37487) as well as one of redundant H3K4me3 features (H3K4me3_GSE11004) from our feature space. We see that even based on the only four non-DNA features, we still obtained high prediction accuracy (**Figure 6K** and **Supplementary Table 9**). Non-DNA features obtained from 150-bp regions led to almost the same prediction accuracy than features based on 300-bp span (**Supplementary Table 9**).

It is interesting that among the five classifiers used in this study, RF performs best when using non-DNA features, but SVM is the best when prediction is based on DNA features (**Supplementary Table 9**). This suggests that prediction performance is determined by the combinatorial effect of features and classification algorithm. Overall, SVM works the best with the whole feature set which consists of DNA-based features and non DNA features (**Supplementary Table 10**). The feature matrices for hot/cold spots were available at https://github.com/gqliu1010/Rec_hotspots.

Effect of Hot/Cold Spot Length on Prediction Performance

We carried out our prediction above on the whole hot/cold spots dataset by calculating features from equally sized regions

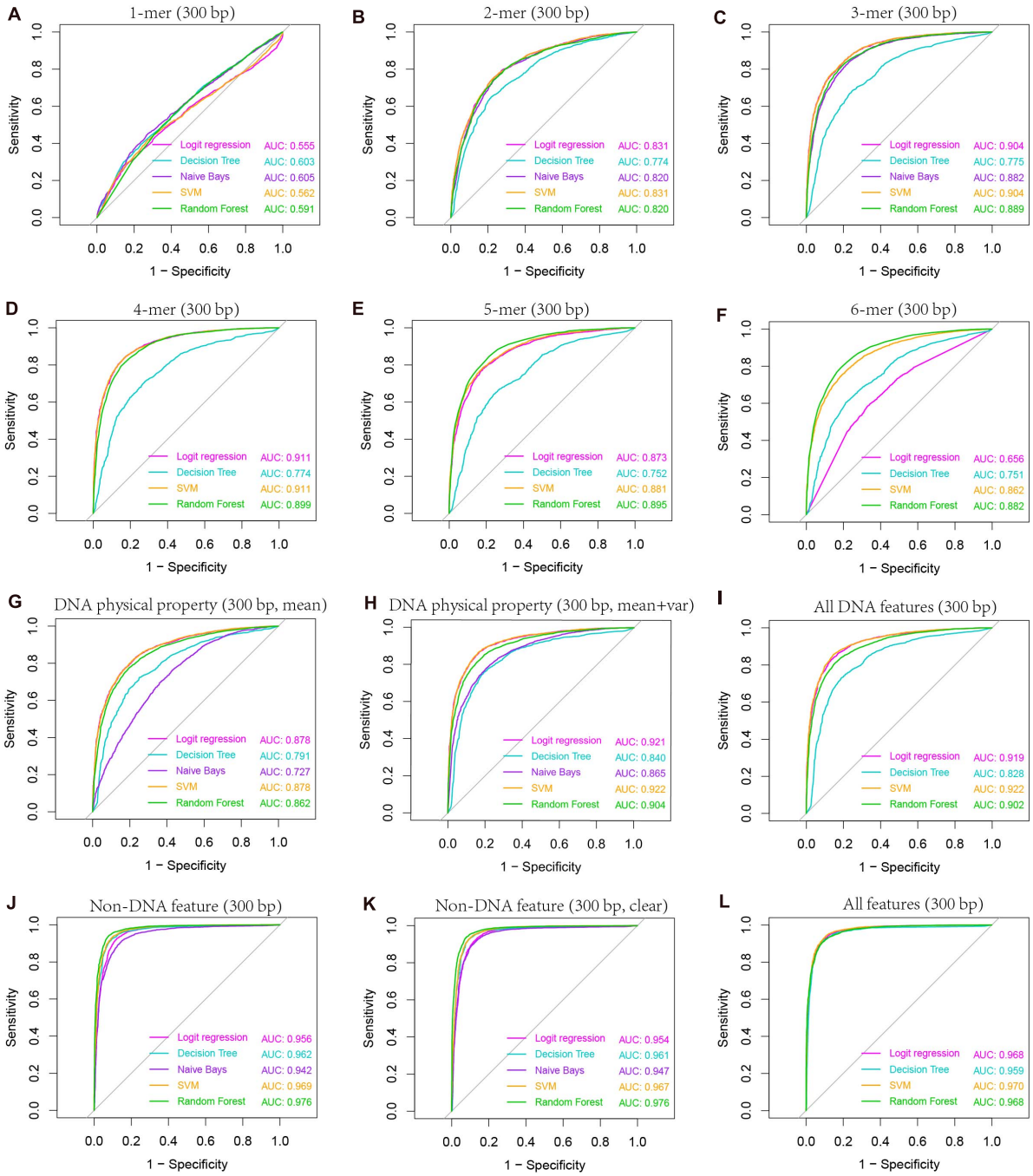

FIGURE 6 | AUC-based comparison of prediction performance between different classification models. Results are based on combined decision values inferred from five-fold cross validation. Features including k-mer composition **(A–F)**, DNA physical properties **(G,H)**, and several non-DNA features (H3K4me3, H3K56ac, MNase-seq signal, and Top2 binding signal) were obtained from 300-bp regions centered at hot/cold spots. Mean and variance were calculated for DNA physical property features by averaging across the 300-bp genomic regions for each hot/cold spot. In non-DNA features **(J)**, predictions were based on six features (H3K4me3_GSE11004, H3K4me3_GSE59005, H3K56ac_GSE37487, H3K56ac_GSE59005, MNase_GSE59005, and Top2_GSE136675), and those excluding two redundant features (H3K4me3_GSE11004 and H3K56ac_GSE37487) were denoted as "clear" **(K)**. All DNA features **(I)** include 4-mer composition, GC-content, GC-skew, mutual information, DNA physical property features listed in **Table 1**. Note that DNA physical property features here include DNA physical properties and DNA shape parameters. All features include all DNA features and clear non-DNA features **(L)**.

(e.g., 300-bp regions), without considering the potential effect of hot/cold spot length. Given the variable size of hot/cold spots, it is conceivable that features are also size-dependent. To investigate size-related effect, we selected the hot/cold spots that are larger than 300 bp, and re-examined if prediction accuracy is affected in this case. Our results show that both DNA-based and non-DNA

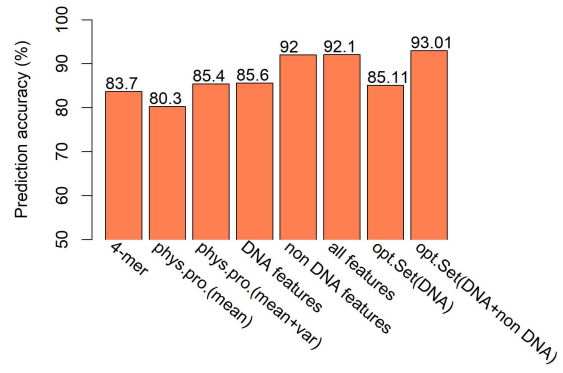

FIGURE 7 | Comparison of SVM-based prediction accuracy between various feature sets. Feature range used in the prediction is 300 bp. Results are based on combined decision values inferred from five-fold cross validation. "phys.pro." denotes physical property-based prediction, and "opt.Set" denotes optimal feature set.

feature-based prediction got increased accuracy (**Figure 8** vs. **Figure 6**), indicating that longer hot/cold spots are more predictable as their underlying DNA sequence and epigenetic information are more informative than shorter hot/cold spots.

Comparison With Existing Models

In order to assess the performance of models presented in this study, we compared with some other existing computational models designed to predict hot/cold spots. Hold-out validation is used for prediction: randomly sampled 70% of the whole benchmark dataset is used to train models and the remaining 30% is used as test set. All the compared models made predictions on the same test set. As far as we know, previously developed models for recombination hot/cold spot classification are all based on DNA-based features. Hence, in order to make comparison more objective, we compared our DNA-based models with existing models.

The results show that our model achieved similar level of prediction accuracy (**Table 2**, SVM: ACC = 85.1%) as aforementioned five-fold cross-validation (**Supplementary Table 8**, SVM: ACC = 85.6%). However, applying the webservers for two other start-of-art models to the same test dataset, we obtained prediction accuracy of ∼60%, which is worse than our models. Why do the start-of-art models have so poor power to discriminate hot/cold spots? It is most likely because those models were trained on ORF sequences with high DSB frequency, while hotspots and coldspots in this study were rigorously defined based on high resolution Spo11 oligo-seq data. Although it was reported that recombination hotspots in budding yeast prefer promoter regions and may have overlap with coding region (Mancera et al., 2008), it is inappropriate to represent a hotspot with its adjacent ORF as coding regions and non-coding genomic regions differ a lot in terms of composition, structure and function. Thus, ORF-based training is not the best choice in computational models and may fail to predict rigorously defined hot/cold spots. Indeed, an IDQD model (Liu et al., 2012) trained on the hot/cold spots defined in this study achieved a much successful prediction (**Table 2**).

Feature Importance and Optimal Feature Set

To give information about what features weigh much in our computational model, we first sorted the features according to Gini index that has been widely used to measure feature importance. The feature importance was inferred from the RF model trained on the whole benchmark dataset. We see that in DNA features, the variations of the DNA-based parameters along sequences rank high and composed the majority of the top 30 features (**Figure 9B**). Stretch and mutual information rank in the top 30. In addition, the list of top 30 4-mers (**Figure 9A**) indicates that oligomers such as AAAA/TTTT, TATA, and CGCG are important in hot/cold spot classification.

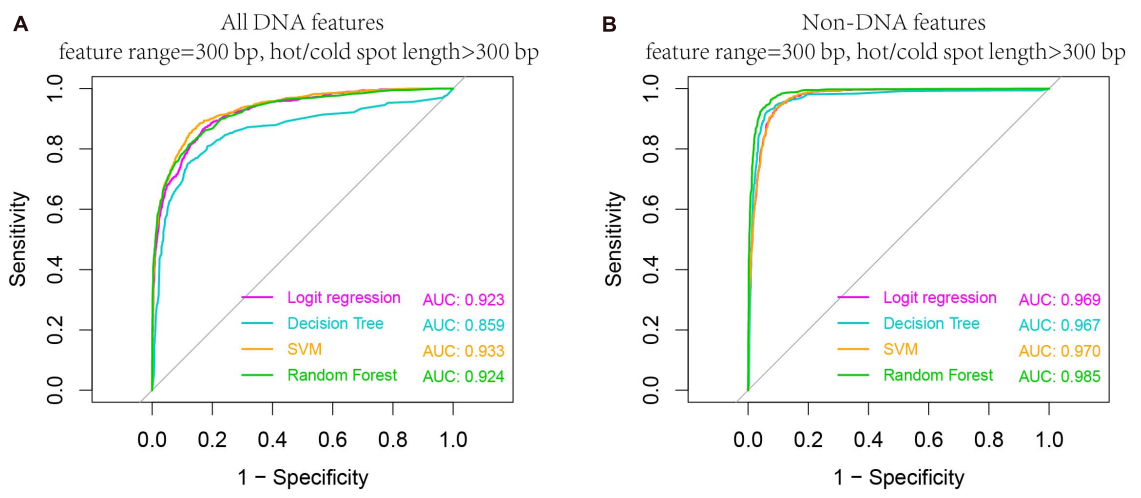

FIGURE 8 | Higher AUC values are obtained when predicting larger hotspots (>300 bp). Results are based on combined decision values inferred from five-fold cross validation by using all DNA features **(A)** and non-DNA features **(B)**. See **Table 1** for feature details.

TABLE 2 | The performances of several models in discriminating recombination hot/cold spots (feature range = 300 bp).

Method	Feature	*SN* (%)	*SP* (%)	*TA* (%)	*F-measure*
iRSpot-PseDNC[a]	PseDNC	47.3	56.9	51.3	53.2
iRecSpot-EF[b]	DNA-based features	38.8	71.5	51.8	49.3
IDQD	4-mer	82.8	83.3	83.0	85.1
SVM (current study)	All DNA features	85.1	85.0	85.1	86.8
RF (current study)	All DNA features	87.0	79.0	83.6	86.0
Logistic regression (current study)	All DNA features	86.2	81.1	84.0	86.2

[a] *Prediction from Chen et al., 2013.*
[b] *Prediction from Jani et al., 2018.*

Feature selection is crucial in machine learning, because the high dimension of feature space often cause high risk of over-fitting and make the prediction model computationally expensive. There are several feature selection strategies, such as filter, wrapper and embedding. We used IFS method (Zhang et al., 2021), which is a filter-based approach, to obtain an optimal feature set which can give best prediction. In the IFS method, analysis of variance (ANOVA) was used to assess feature importance. The features were sorted according to the decreasing order of the ratio between inter-group variance and intra-group variance. The higher the ratio is, the more powerful the feature is in discriminating the two groups of samples (hotspots and coldspots). Then the features were added one by one to feature space in the descending order of feature importance. For each turn of feature addition, SVM classifier was trained by using the new feature set, and average prediction accuracy of five-fold cross validation was reported (**Figure 10A**). If the addition of a feature increases the average prediction accuracy, the feature was retained in the feature set, otherwise it was removed. Optimal sets were sought, respectively, in DNA-based feature space and the whole feature space. We show that our model based on the optimal feature set which consists of only 62 features achieved a slightly improved accuracy than all-feature-based model (**Figure 7**, 93 vs. 92.1%). In addition, we also examined the overlap between top 50 features determined, respectively, by Gini index and ANOVA. Most of them (80%) occur in both feature

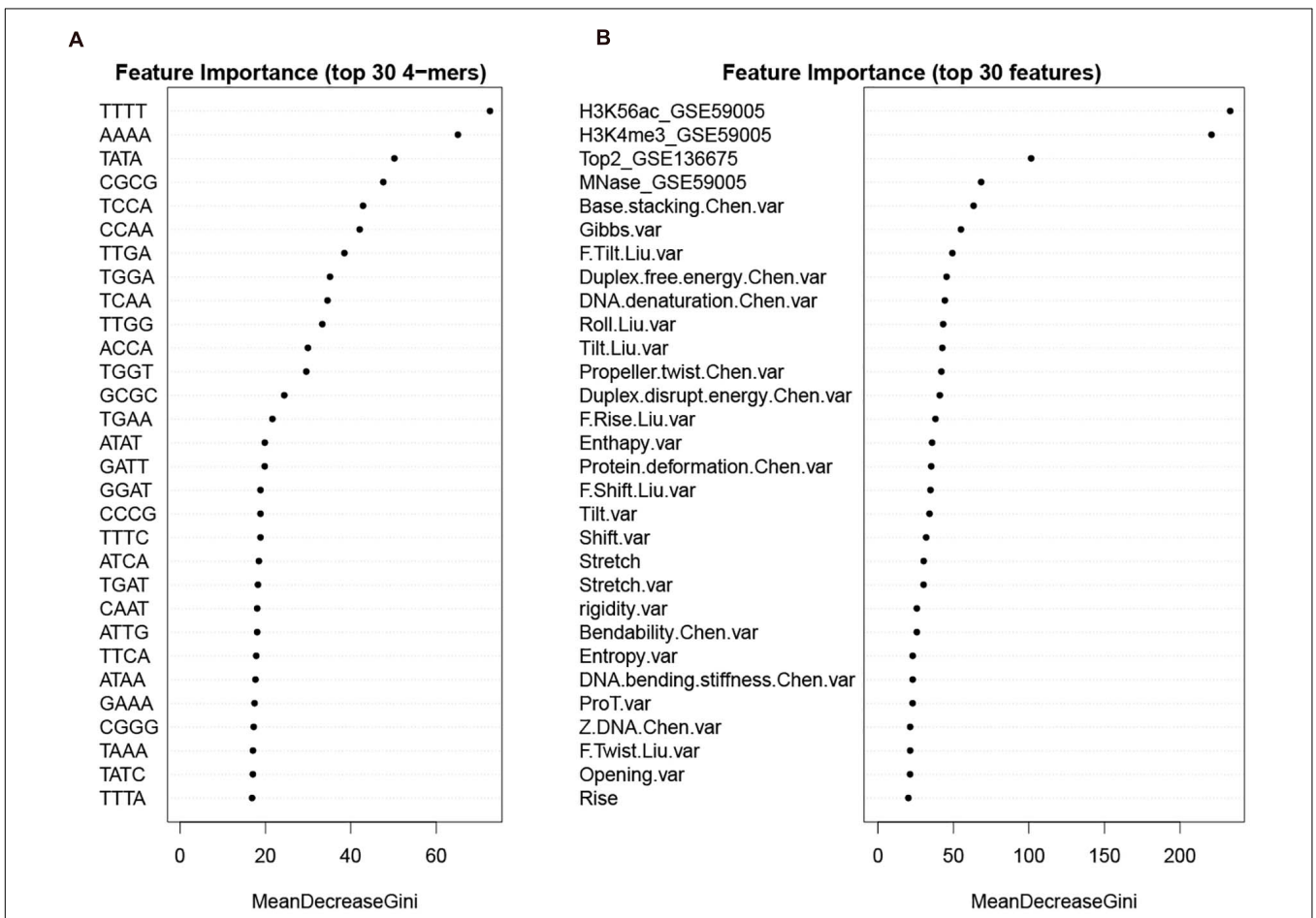

FIGURE 9 | Feature importance sorted by Gini index. **(A)** top 30 k-mers; **(B)** top 30 features selected from the whole feature set. Examples of feature name illustration: "F.Slide.Liu.var" represents the variance of force constant for slide taken from Liu et al. (2021); "Duplex.free.energy.Chen.var" represents the variance of "Duplex free energy" taken from Chen et al. (2012); "Stretch" is DNA shape parameter calculated by using DNAshapeR. See **Supplementary Tables 2–4** for more details about the features.

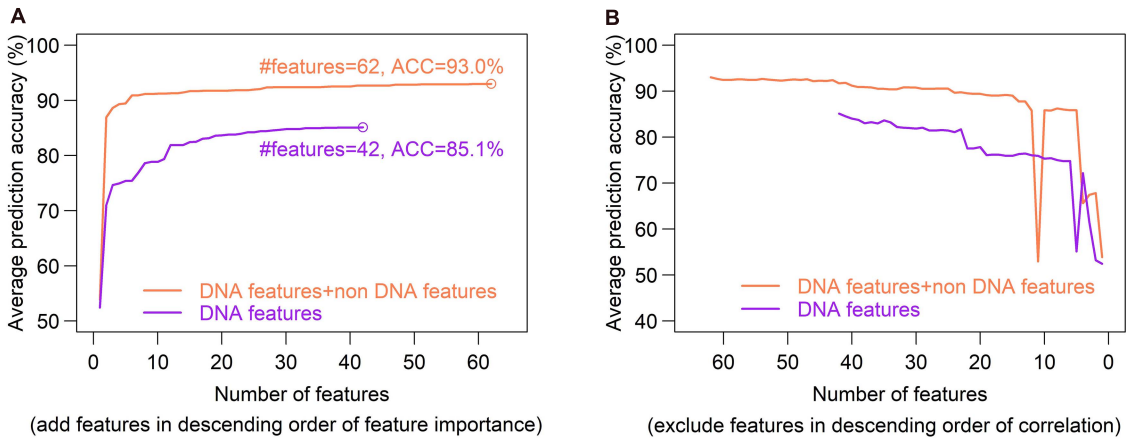

FIGURE 10 | Optimal set of DNA features is determined through a SVM-based Incremental Feature Selection method (IFS). **(A)**, In the IFS method, ANOVA was used to sort feature importance and then features were added one by one to feature space in the descending order of feature importance. For each turn of feature addition, SVM classifier was trained on the updated feature set, and average prediction accuracy of five-fold cross validation was reported. If the addition of a feature increases the average prediction accuracy, the feature was retained in the feature set, otherwise it was removed. Optimal sets were sought, respectively, in DNA-based feature space (DNA features) and the whole feature space (DNA features + non DNA features). Two optimal feature sets composed of 45 features and 44 features were obtained. **(B)**, Inter-correlated features were excluded sequentially from the feature sets obtained in figure **(A)**. During the feature-excluding process, no new peak was observed for prediction accuracy, and thus the optimal feature set determined through IFS remain unchanged.

sets, suggesting the consistency of feature importance between the two methods (**Supplementary Figure 4**). The consistent features occurred in both top feature sets might represent the most important features (**Supplementary Table 11**).

Excluding redundant features is another way to reduce feature dimensionality with no or little sacrifice in prediction accuracy. If two features strongly correlate with each other, it is possible that only one of them is sufficient for prediction. We used a recursive redundant-feature-excluding method, in which highly correlated features are excluded one by one from the optimal feature set according to the descending order of Pearson's correlation coefficients between features. One of the two correlated features, performing worse in univariate classification, was removed at each round, and then the model was re-trained on the updated feature set of training dataset, followed by a five-fold cross validation. The univariate classification means individual feature-based classification. During the feature-excluding process, no new peak was observed for prediction accuracy, and thus the optimal feature set determined through IFS remained unchanged (**Figure 10B**). We can also see that the earliest removal of features which represent the exclusion of highly correlated (redundant) features has little impact on prediction accuracy, while the later-removal of features affect prediction accuracy remarkably (**Figure 10B**).

CONCLUSION

In summary, firstly we defined a reliable set of recombination cold spots based on high-resolution Spo11-oligo sequencing data; secondly, we characterized recombination hot/cold spots in terms of sequence-derived features and epigenetic marks; thirdly, we performed binary predictions based on five classification algorithms. Our results show that, overall, SVM classifier performs the best in hot/cold spot classification, and also outperforms other existing methods. Importantly, our results indicate that variance in sequence-based feature profile and epigenetic marks are able to assist remarkably the identification of recombination hotspots.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

GQL developed the model, carried out the analysis, and wrote the manuscript. SS and YS carried out the partial calculation of DNAshape parameters. SS, QZ, BD, YS, GJL, and XZ participated in the data analysis and discussion. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the National Natural Science Foundation of China (31660322 and 21767020), and Inner Mongolia Natural Science Foundation of China (2018LH03023 and 2019MS02021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.705038/full#supplementary-material

REFERENCES

Baudat, F., Imai, Y., and de Massy, B. (2013). Meiotic recombination in mammals: localization and regulation. *Nat. Rev. Genet.* 14, 794–806. doi: 10.1038/nrg3573

Borde, V., Robine, N., Lin, W., Bonfils, S., Géli, V., and Nicolas, A. (2009). Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *EMBO J.* 28, 99–111. doi: 10.1038/emboj.2008.257

Breiman, L. (2001). Random forest. *Machine Learn.* 45, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Cesarini, E., D'Alfonso, A., and Camilloni, G. (2012). H4K16 acetylation affects recombination and ncRNA transcription at rDNA in Saccharomyces cerevisiae. *Mol. Biol. Cell.* 23, 2770–2781. doi: 10.1091/mbc.e12-02-0095

Chen, W., Feng, P. M., Lin, H., and Chou, K. C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res.* 41:e68. doi: 10.1093/nar/gks1450

Chen, W., Lin, H., Feng, P. M., Ding, C., Zuo, Y. C., and Chou, K. C. (2012). iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. *PLoS One* 7:e47843. doi: 10.1371/journal.pone.0047843

Collins, M., Schapire, R., and Singer, Y. (2004). Logistic regression, AdaBoost and bregman distances. *Machine Learn.* 48, 253–285.

Coop, G., and Przeworski, M. (2007). An evolutionary view of human recombination. *Nat. Rev. Genet.* 8, 23–34. doi: 10.1038/nrg1947

Cortes, C., and Vapnik, V. N. (1995). Support vector networks. *Machine Learn.* 20, 273–297.

de Castro, E., Soriano, I., Marín, L., Serrano, R., Quintales, L., and Antequera, F. (2012). Nucleosomal organization of replication origins and meiotic recombination hotspots in fission yeast. *EMBO J.* 31, 124–137. doi: 10.1038/emboj.2011.350

Friedman, N., Geiger, D., and Pazzanzi, M. (1997). Bayesian network classifiers. *Machine Learn.* 29, 131–163.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-Content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907–911. doi: 10.1093/genetics/159.2.907

Getun, I. V., Wu, Z. K., Khalil, A. M., and Bois, P. R. (2010). Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. *EMBO Rep.* 11, 555–560. doi: 10.1038/embor.2010.79

Gittens, W. H., Johnson, D. J., Allison, R. M., Cooper, T. J., Thomas, H., and Neale, M. J. (2019). A nucleotide resolution map of Top2-linked DNA breaks in the yeast and human genome. *Nat. Commun.* 10:4846.

Heldrich, J., Sun, X., Vale-Silva, L. A., Markowitz, T. E., and Hochwagen, A. (2020). Topoisomerases modulate the timing of meiotic DNA breakage and chromosome morphogenesis in saccharomyces cerevisiae. *Genetics* 215, 59–73. doi: 10.1534/genetics.120.303060

Hu, J., Donahue, G., Dorsey, J., Govin, J., Yuan, Z., Garcia, B. A., et al. (2015). H4K44 acetylation facilitates chromatin accessibility during meiosis. *Cell Rep.* 13, 1772–1780. doi: 10.1016/j.celrep.2015.10.070

Ignatova, Z., Martinez-Perez, I., and Zimmermann, K. H. (2008). *DNA Computing Models*. New York, NY: Springer.

Jani, M. R., Khan Mozlish, M. T., Ahmed, S., Tahniat, N. S., Farid, D. M., and Shatabda, S. (2018). iRecSpot-EF: effective sequence based features for recombination hotspot prediction. *Comput. Biol. Med.* 103, 17–23. doi: 10.1016/j.compbiomed.2018.10.005

Jiang, P., Wu, H., Wei, J., Sang, F., Sun, X., and Lu, Z. (2007). RF-DYMHC: detecting the yeast meiotic recombination hotspots and coldspots by random forest model using gapped dinucleotide composition features. *Nucleic Acids Res.* 35, W47–W51.

Karányi, Z., Hornyák, L., and Székvölgyi, L. (2020). Histone H3 lysine 56 acetylation is required for formation of normal levels of meiotic DNA breaks in S. cerevisiae. *Front. Cell Dev. Biol.* 7:364. doi: 10.3389/fcell.2019.00364

Khan, F., Khan, M., Iqbal, N., Khan, S., Muhammad Khan, D., Khan, A., et al. (2020). Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach. *Front. Genet.* 11:539227. doi: 10.3389/fgene.2020.539227

Li, L., Yu, S., and Xiao, W. (2014). Sequence-based identification of recombination spots using pseudo nucleic acid representation and recursive feature extraction by linear kernel SVM. *BMC Bioinform.* 15:340. doi: 10.1186/1471-2105-15-340

Liu, B., Wang, S., and Long, R. (2017). iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 33, 35–41. doi: 10.1093/bioinformatics/btw539

Liu, G., and Li, H. (2008). The correlation between recombination rate and dinucleotide bias in Drosophila melanogaster. *J. Mol. Evol.* 67, 358–367. doi: 10.1007/s00239-008-9150-0

Liu, G., Liu, G. J., Tan, J. X., and Lin, H. (2019). DNA physical properties outperform sequence compositional information in classifying nucleosome-enriched and -depleted regions. *Genomics* 111, 1167–1175. doi: 10.1016/j.ygeno.2018.07.013

Liu, G., Liu, J., Cui, X., and Cai, L. (2012). Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. *J. Theor. Biol.* 293, 49–54. doi: 10.1016/j.jtbi.2011.10.004

Liu, G., Ma, Q., and Xu, Y. (2018). Physical properties of DNA may direct the binding of nucleoid-associated proteins along the E. coli genome. *Math. Biosci.* 301, 50–58. doi: 10.1016/j.mbs.2018.03.026

Liu, G., Xing, Y., Zhao, H., Wang, J., Shang, Y., and Cai, L. (2016). A deformation energy-based model for predicting nucleosome dyads and occupancy. *Sci. Rep.* 6:24133.

Liu, G., Zhao, H., Meng, H., Xing, Y., and Cai, L. (2021). A deformation energy model reveals sequence-dependent property of nucleosome positioning. *Chromosoma* 130, 27–40. doi: 10.1007/s00412-020-00750-9

Luo, L., Lee, W., Jia, L., Ji, F., and Tsai, L. (1998). Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E.* 58, 861–871. doi: 10.1103/physreve.58.861

MacLennan, M., Crichton, J. H., Playfoot, C. J., and Adams, I. R. (2015). Oocyte development, meiosis and aneuploidy. *Semin. Cell Dev. Biol.* 45, 68–76. doi: 10.1016/j.semcdb.2015.10.005

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454, 479–485. doi: 10.1038/nature07135

Mourad, R., Ginalski, K., Legube, G., and Cuvier, O. (2018). Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biol.* 19:34.

Myers, S., Freeman, C., and Auton, A. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* 40, 1124–1129. doi: 10.1038/ng.213

Nick, T. G., and Campbell, K. M. (2007). Logistic regression. *Methods Mol. Biol.* 404, 273–301.

Paiano, J., Wu, W., and Yamada, S. (2020). ATM and PRDM9 regulate SPO11-bound recombination intermediates during meiosis. *Nat. Commun.* 11:857.

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., et al. (2011). A hierarchical combination of factors shapes the genomewide topography of yeast meiotic recombination initiation. *Cell* 144, 719–731. doi: 10.1016/j.cell.2011.02.009

Pratto, F., Brick, K., Khil, P., Smagulova, F., Petukhova, G. V., and Camerini-Otero, R. D. (2014). Recombination initiation maps of individual human genomes. *Science* 346:1256442. doi: 10.1126/science.1256442

Pyatnitskaya, A., Borde, V., and De Muyt, A. (2019). Crossing and zipping: molecular duties of the ZMM proteins in meiosis. *Chromosoma* 128, 181–198. doi: 10.1007/s00412-019-00714-8

Qiu, W. R., and Xiao, X. (2014). iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. *Int. J. Mol. Sci.* 15, 1746–1766. doi: 10.3390/ijms15021746

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learn.* 1, 81–106.

Scipioni, A., Anselmi, C., Zuccheri, G., Samori, B., and De Santis, P. (2002). Sequence-dependent DNA curvature and flexibility from scanning force microscopy images. *Biophys. J.* 83, 2408–2418. doi: 10.1016/s0006-3495(02)75254-5

Serrano-Quílez, J., Roig-Soucase, S., and Rodríguez-Navarro, S. (2020). Sharing marks: H3K4 methylation and H2B ubiquitination as features of meiotic recombination and transcription. *Int. J. Mol. Sci.* 21:4510. doi: 10.3390/ijms21124510

Smagulova, F., Gregoretti, I. V., Brick, K., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472, 375–378. doi: 10.1038/nature09869

Sommermeyer, V., Béneut, C., Chaplais, E., Serrentino, M. E., and Borde, V. (2013). Spp1, a member of the Set1 Complex, promotes meiotic DSB formation in promoters by tethering histone H3K4 methylation sites to chromosome axes. *Mol. Cell.* 49, 43–54. doi: 10.1016/j.molcel.2012.11.008

Wang, S., Hassold, T., Hunt, P., White, M. A., Zickler, D., Kleckner, N., et al. (2017). Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. *Cell* 168, 977–989. doi: 10.1016/j.cell.2017.02.002

Webster, M. T., and Hurst, L. D. (2012). Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet.* 28, 101–109. doi: 10.1016/j.tig.2011.11.002

Yamada, S., Ohta, K., and Yamada, T. (2013). Acetylated histone H3K9 is associated with meiotic recombination hotspots, and plays a role in recombination redundantly with other factors including the H3K4 methylase Set1 in fission yeast. *Nucleic Acids Res.* 41, 3504–3517. doi: 10.1093/nar/gkt049

Yang, H., Yang, W., Dao, F. Y., Lv, H., Ding, H., Chen, W., et al. (2020). A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief Bioinform.* 21, 1568–1580. doi: 10.1093/bib/bbz123

Zhang, B. J., and Liu, G. Q. (2014). Predicting recombination hotspots in yeast based on DNA sequence and chromatin structure. *Curr. Bioinform.* 9, 28–33. doi: 10.2174/15748936089991401091214444

Zhang, D., Xu, Z. C., Su, W., Yang, Y. H., Lv, H., Yang, H., et al. (2021). iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics* 37, 171–177. doi: 10.1093/bioinformatics/btaa702

Zhang, L., and Kong, L. (2019). iRSpot-PDI: identification of recombination spots by incorporating dinucleotide property diversity information into Chou's pseudo components. *Genomics* 111, 457–464. doi: 10.1016/j.ygeno.2018.03.003

Zhang, L., Ma, H., and Pugh, B. F. (2011). Stable and dynamic nucleosome states during a meiotic developmental process. *Genome Res.* 21, 875–884. doi: 10.1101/gr.117465.110

Zhou, T., Weng, J., Sun, X., and Lu, Z. (2006). Support vector machine for classification of meiotic recombination hotspots and coldspots in Saccharomyces cerevisiae based on codon composition. *BMC Bioinform.* 7:223. doi: 10.1186/1471-2105-7-223

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., et al. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.* 41, W56–W62.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fgene.2021.690349

Identification of Causal Genes of COVID-19 Using the SMR Method

*Yan Zong and Xiaofei Li**

Department of Infectious Diseases, Yiwu Central Hospital, Jinhua, China

Since the first report of COVID-19 in December 2019, more than 100 million people have been infected with SARS-CoV-2. Despite ongoing research, there is still limited knowledge about the genetic causes of COVID-19. To resolve this problem, we applied the SMR method to analyze the genes involved in COVID-19 pathogenesis by the integration of multiple omics data. Here, we assessed the SNPs associated with COVID-19 risk from the GWAS data of Spanish and Italian patients and lung eQTL data from the GTEx project. Then, GWAS and eQTL data were integrated by summary-data-based (SMR) methods using SNPs as instrumental variables (IVs). As a result, six protein-coding and five non-protein-coding genes regulated by nine SNPs were identified as significant risk factors for COVID-19. Functional analysis of these genes showed that UQCRH participates in cardiac muscle contraction, PPA2 is closely related to sudden cardiac failure (SCD), and OGT, as the interacting gene partner of PANO1, is associated with neurological disease. Observational studies show that myocardial damage, SCD, and neurological disease often occur in COVID-19 patients. Thus, our findings provide a potential molecular mechanism for understanding the complications of COVID-19.

Keywords: SMR, COVID-19, eQTL, GWAS, UQCRH, PPA2, OGT, PANO1

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Juan Wang,
Inner Mongolia University, China
Sheng Li,
Wuhan University, China

***Correspondence:**
Xiaofei Li
xiaofeili2000@163.com

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

Received: *02 April 2021*
Accepted: *07 May 2021*
Published: *05 July 2021*

Citation:
Zong Y and Li X (2021) Identification of Causal Genes of COVID-19 Using the SMR Method. Front. Genet. 12:690349. doi: 10.3389/fgene.2021.690349

INTRODUCTION

In December 2019, SARS-CoV-2 was first reported to lead to the respiratory disease coronavirus disease 2019 (COVID-19) (Zhou et al., 2020). Subsequently, COVID-19 quickly spread to all parts of the world and became a worldwide public health event. As of February 17, 2021, more than 100 million people had been infected, and more than 2.4 million people had died of COVID-19. At present, a total of seven types of coronaviruses that can infect humans have been discovered, including SARS-CoV, SARS-CoV-2, and MERS-CoV, which have high case fatality rates (CFRs) (Gussow et al., 2020). The other four coronaviruses, HCoV-HKU1, HCoV-NL63, HCoV-OC43, and HCoV-229E, only cause mild symptoms in humans. Although the CFR of SARS-CoV-2 is relatively lower than those of SARS-CoV and MERS-CoV, it is still highly infectious.

Exploring the origin of the virus would help to increase the understanding of SARS-CoV-2 (Cui et al., 2019; Narang et al., 2019; Benetti et al., 2020; Meng et al., 2020; Wan et al., 2020; Qi et al., 2021). Previous studies have shown that bats are the natural host of the evolved coronavirus (Jiang et al., 2020; Kwon et al., 2020). Based on the alignment of the reference genome sequence, a phylogenetic tree was constructed, indicating that the genes are very similar between SARS-CoV-2 and members of the bat Sarbecovirus subgenus Betacoronavirus (Wu et al., 2020). According to sequence mapping, the whole-genome sequence of SARS-CoV-2 has the highest similarity with SARS-CoV BatCoV RaTG13, reaching over 96%. The similarity between SARS-CoV-2 and the

coronaviruses SARS-CoV and MERS-CoV is only 79 and 50%, respectively (Andersen et al., 2020). Although the specific route of transmission of SARS-CoV-2 from its natural host to humans is not yet clear, researchers have discovered that the key functional sites of the SARS-CoV-2 spike protein are almost the same as those of the virus isolated from pangolins. Therefore, a pangolin coronavirus may have provided part of the spike gene for SARS-CoV-2.

Computational methods with omics data have shown strong power in identifying disease-related genes (Zhao T. et al., 2020c). SARS-CoV-2 is a single-stranded RNA virus (Chen et al., 2020). The reference genome shows that it has 29,903 nucleic acid base pairs, including seven conserved unstructured protein domains and four structural protein domains, including the spike protein. The sequence of SARS-CoV-2 has mutated since its discovery. As early as March 2020, researchers analyzed 160 early virus strains and found three important single-point mutations, T8782C, C28144T, and G26144T (Forster et al., 2020). Among them, the mutations T8782C and C28144T are used to distinguish between type A and B viruses, and the mutation G26144T represents a newer virus type (type C). In May 2020, Cheng et al. analyzed more than 1,800 strains of viruses in the Americas, Europe, and Asia and verified that type B viruses are more infectious and have almost replaced the type A viruses in current circulation (Cheng et al., 2021). Through comparison with the reference genome, it was found that each virus mutated at approximately 1.75 sites per month. The overall mutations were silent mutations and would not cause major functional changes in the virus. Cheng et al. proposed to analyze the mutation rules of the virus over time and identified seven dominant mutations. According to functional bioinformatics analysis, these mutations likely caused the virus to decrease in toxicity and increase in infectivity. In addition, mutation clustering information (Zhu X. et al., 2019; Qi et al., 2020; Zhao T. et al., 2020b; Zhao X. et al., 2020; Zou et al., 2020; Zhao et al., 2021) indicated that the virus circulating in the Americas is more similar to RaTG13.

By searching for the origin of SARS-CoV-2, researchers also found that SARS-CoV-2 and SARS-CoV use the same receptor, angiotensin-converting enzyme II (ACE2). SARS-CoV-2 invades human cells by binding to ACE2. By interacting with proteins in human cells, SARS-CoV-2 affects human health. In August 2020, researchers found hundreds of interactions between human proteins and SARS-CoV-2 proteins. Although many proteins that interact with SARS-CoV-2 have been discovered so far, the genes associated with SARS-CoV-2 pathogenesis are still unknown. Since genomics data of many COVID-19 patients have been reported, there is an opportunity to find potential risk genes through analysis and integration of omics data of susceptible populations (Cheng, 2019; Cheng et al., 2019; Li F. et al., 2020; Wang et al., 2020).

SMR is a method used to determine the causal association between genetically determined traits and diseases, and eQTLs are genetic variations related to the expression of traits. Since eQTL data are tissue specific, it is possible to correlate the eQTL data of disease-related tissues with disease GWAS data and use the SMR method to find causal genes for diseases. Therefore, in our study, we used lung eQTL data of GTEx and GWAS data of COVID-19 patients to identify the genes related to COVID-19 pathogenesis using the SMR method.

MATERIALS AND METHODS

Risk SNPs for Severe COVID-19

According to current knowledge, some people are more susceptible than others to COVID-19. To assess the impact of

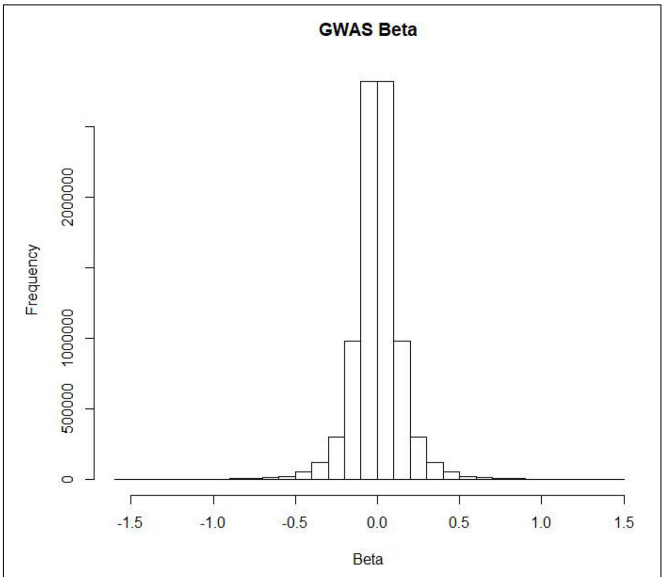

FIGURE 1 | The distribution of GWAS beta in severe COVID-19 patients.

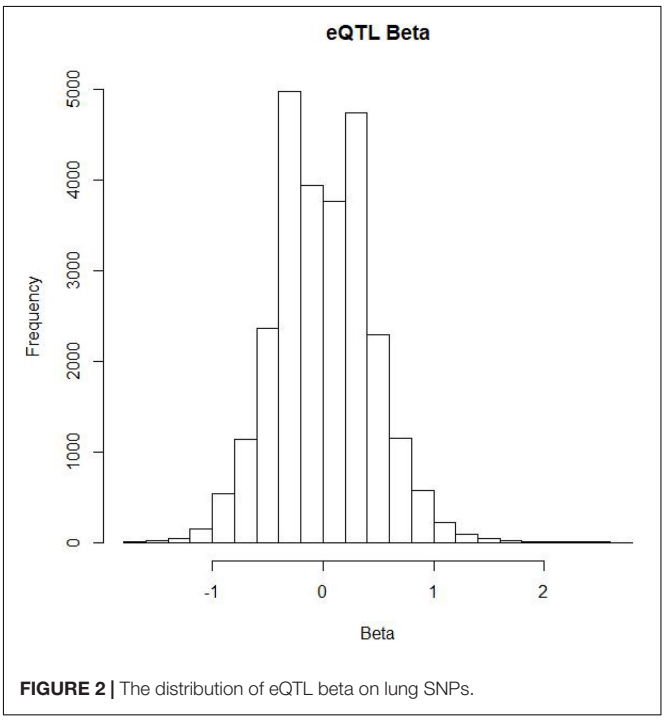

FIGURE 2 | The distribution of eQTL beta on lung SNPs.

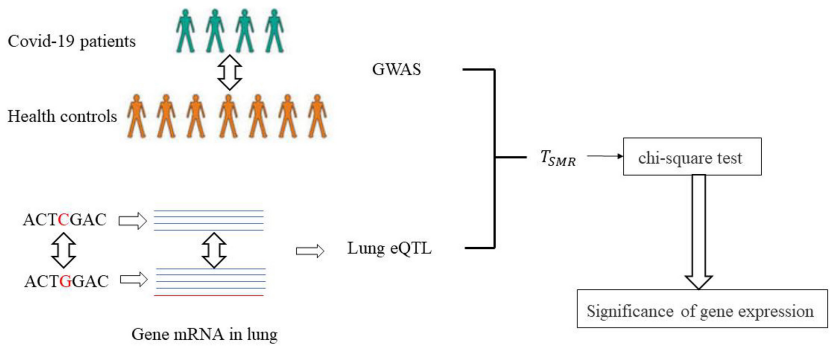

FIGURE 3 | The workflow of identifying causal genes of COVID-19 using the SMR method.

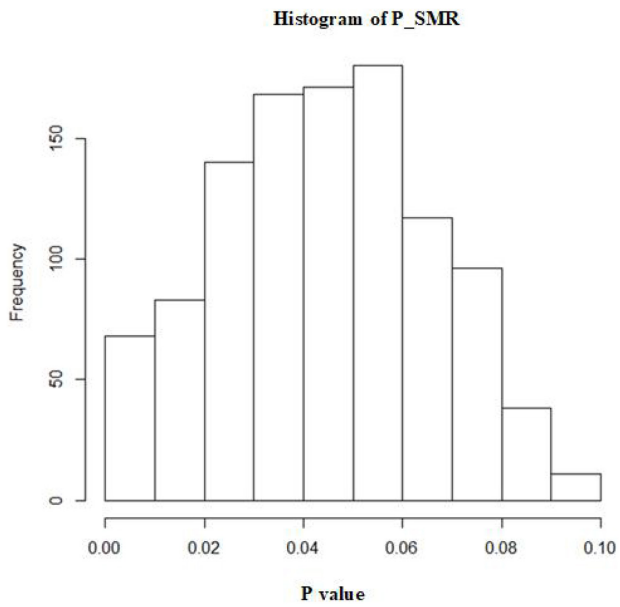

FIGURE 4 | The distribution of SMR *p*-values of SNPs.

SNPs on the risk of COVID-19, researchers conducted a genome-wide association study (GWAS) on two groups of European patients in seven hospitals in Italy and Spain and performed a meta-analysis of the results of the two groups (Ellinghaus et al., 2020). First, quality control was conducted on 1980 severe COVID-19 patients, and 1,610 patients remained after removing population outliers. Then, the researchers performed a GWAS on 835 Italian patients and 1,255 control group members (Sun et al., 2019), as well as 775 Spanish patients and 950 control group members. As a result, they obtained severe COVID-19 risk data for 8,582,968 SNPs. Finally, the two sets of experimental results were meta-analyzed to obtain the final risk SNPs. **Figure 1** shows the distribution of risk betas of SNPs in COVID-19 patients. Most SNPs had no impact on the risk of COVID-19. In addition, researchers found that COVID-19 high-risk SNPs clustered on 3p21.31.

Lung eQTL

To date, many GWAS have been performed. These studies have identified thousands of disease-related risk SNPs. Since more than 90% of SNP sites exist outside of protein-coding genes, it is difficult to understand the mechanisms by which these SNPs affect diseases. To this end, researchers investigate expressed quantitative trait loci (eQTLs) to reveal the genes regulated by SNPs in the blood, lung and other tissues (Zhao T. et al., 2020a). Therefore, the NIH launched the gene type-tissue expression (GTEx) project, with the goal of establishing the relationships between SNPs and gene expression in different tissues. Currently, the project has accepted more than 900 post-mortem donors. Sequencing of different tissues in the donors has identified a large number of SNP-regulated genes. Summarized GTEx data could be obtained from the project's website. **Figure 2** shows the distribution of beta values of lung SNPs on gene expression.

Identification of Potential Causal Genes of COVID-19 Based on the SMR Method

In the domain of biomedicine, many causal disease associations have been discovered through observational research, such as the association between smoking and lung cancer (Ghosh and Yan, 2020; Li J. et al., 2020; Yuan et al., 2020). However, the associations between phenotypes and diseases found in these observational studies cannot reflect causality. However, because observational studies are usually disturbed by external factors and often face practical problems related to long time frames and high costs, there are large errors in the analysis of pathogenic factors of diseases. Mendelian randomization follows the Mendelian inheritance law of allele separation and free recombination of nonalleles and makes causal inferences based on genetic variation, which does not change with environment or age, so this method is widely used in causal inference of pathogenic factors. Sometimes, due to the influence of confounding factors, the correlation found is not accurate. This greatly limits the development of the field. To solve this problem, the statistician Katan (1986) introduced the concept of Mendelian randomization (MR) in 1986 to study whether low serum cholesterol levels can increase the risk of cancer. MR uses genotype as an instrumental variable (IV) and applies the two-stage least squares method to infer the pathogenicity of diseases.

With the gradual deepening of GWAS research, this method has been widely applied using SNPs as IVs. This method assumes that Z is an instrumental variable (SNP), X represents exposure factors or gene expression levels, and Y represents the disease. According to the two-stage least squares method, the effect of X on Y is evaluated as follows.

$$Beta_{XY} = Beta_{ZX} / Beta_{ZY} \quad (1)$$

Here, $Beta_{ZX}$ represents the least-squares estimate of $\underline{Z}$ on X, and $Beta_{ZY}$ is the least-squares estimate of Z on Y. Then, we estimate the significance of X on Y as T_{MR}.

$$T_{MR} = (Beta_{XY})^2 / var(Beta_{XY}) \quad (2)$$

Theoretically, GWAS and eQTL need to target the same sample, but GWAS and eQTL are performed independently, so we use the two-sample MR method, and T_{SMR} is obtained as the final evaluation value (Zhao S. et al., 2019; Zhao T. et al., 2019):

$$T_{SMR} = {Z_{ZY}}^{2*} {Z_{ZX}}^2 / ({Z_{ZY}}^2 {Z_{ZX}}^2) \quad (3)$$

where Zzy represents the Z statistics from the GWAS, and Zzx represents the z statistics from the eQTL study. Since the true values of $Beta_{ZX}$ and $Beta_{ZY}$ cannot be obtained, we can use estimations to replace them. The T_{SMR} yields an approximate Chi-square test statistic.

Here, we used the SMR method to evaluate the mechanism by which SNPs affect COVID-19 and identify potential pathogenic genes in the lungs. The specific process is shown in **Figure 3**. The GWAS data of COVID-19 patients and the lung eQTL data were obtained from a public dataset. T_{SMR} was calculated based on the two-sample MR method, which was then further evaluated using the chi-square test to identify significant SNPs and genes.

RESULTS

Causal SNPs and Genes of COVID-19

A total of 1,072 SNPs appeared in both GWAS and eQTL data. After application of the SMR method, 1,072 SNPs were determined to regulate the P value distribution of COVID-19 through genes (**Figure 4**). The P values of SMR for most SNPs were concentrated in the range of 0.02–0.08. Here, the threshold was set as 0.003, and then 11 genes (UQCRH,

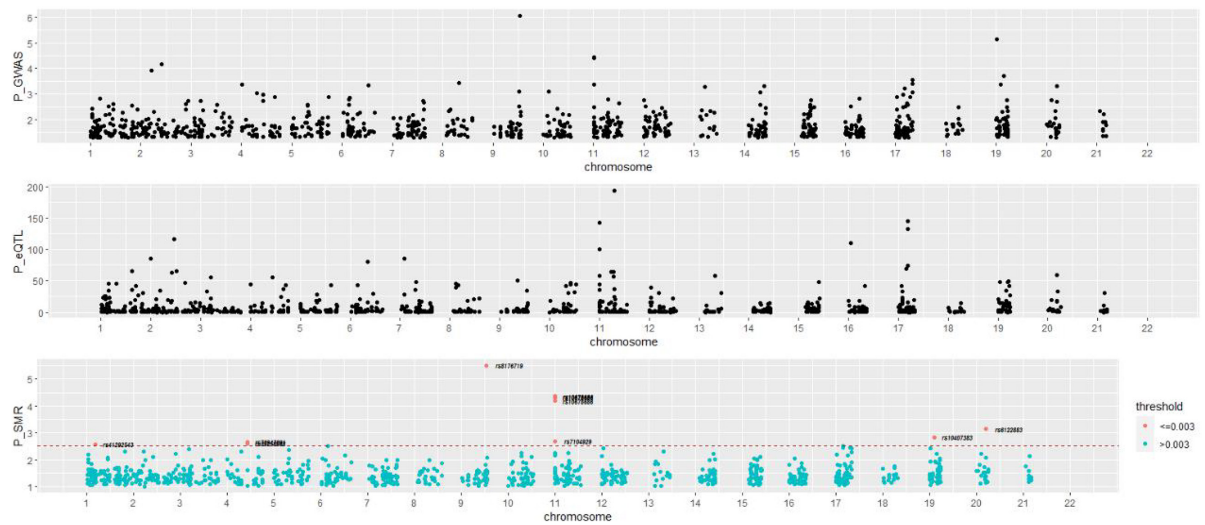

FIGURE 5 | The experimental results based on SMR.

TABLE 1 | Nine Causal SNPs and eleven pathogenic genes of COVID-19.

SNP	P_SMR	Gene	chr	pos	P_GWAS	P_eQTL
rs41292543	0.002574	UQCRH	1	46309111	0.001553	$2.65E-16$
rs35258888	0.002419	PPA2	4	105355205	0.001035	$2.61E-10$
rs70947091	0.002154	PAPSS1	4	107694523	0.001899	$2.28E-56$
rs8176719	3.07E-06	ABO	9	133257521	$8.76E-07$	$1.27E-34$
rs10678686	4.68E-05	AP006621.5	11	780321	$3.96E-05$	$1.28E-100$
rs10678686	4.34E-05	CMB9-55F22.1	11	780321	$3.96E-05$	$9.62E-143$
rs10678686	6.40E-05	AP006621.6	11	780321	$3.96E-05$	$1.12E-44$
rs7104929	0.001973	PANO1	11	784340	$3.73E-05$	0.0111629
rs10407383	0.001423	CTD-2027I19.2	19	24134099	0.000412	$4.61E-09$
rs6122883	0.000702	LINC01273	20	50172836	0.000494	$3.43E-34$
rs6151429	0.000645	ARSA	22	50625049	0.000522	$4.70E-50$

PPA2, PAPSS1, ABO, AP006621.5, CMB9-55F22.1, AP006621.6, PANO1, CTD-2027I19.2, LINC01273, and ARSA) regulated by nine SNPs (rs41292543, rs35258888, rs70947091, rs8176719, rs10678686, rs10678686, rs10678686, rs7104929, rs10407383, rs6122883, and rs6151429) with P values lower than the threshold were considered to increase the risk of COVID-19 (**Table 1**). **Figure 5** shows the GWAS P value, eQTL P value, and SMR *P*-value of all SNPs.

Functional Analysis of Causal Genes

Among the identified pathogenic genes, there were a total of six protein-coding genes (UQCRH, PPA2, PAPSS1, ABO, PANO1, and ARSA) and five noncoding genes. We then performed functional analysis of these six protein-coding genes to identify their functions, related diseases and pathways.

UQCRH participates in cardiac muscle contraction. A large number of studies have found that approximately 8–12% of COVID-19 patients have myocardial damage (Lippi et al., 2020). Although heart problems are usually not the most prominent or deadly feature of COVID-19, they are common and are severe enough that most people admitted to the hospital for COVID-19 are now being screened for myocardial damage. There are many potential causes of COVID-19-related myocardial damage, but it is often difficult to determine the specific cause in a specific individual. The UQCRH gene found here may be a potential cause. In addition, according to Gene Ontology annotation, UQCRH has ubiquinol-cytochrome-c reductase activity.

PPA2 is closely related to sudden cardiac failure (SCD). At present, there have been reports of COVID-19 patients who have died of SCD. In July 2020, Samira et al. diagnosed three patients with COVID-19 according to the reverse transcriptase-polymerase chain reaction of nasopharyngeal swabs and radiological examinations (Shirazi et al., 2021) who eventually died of SCD. Thus, the authors recommended that it is necessary to monitor the heart conditions of COVID-19 patients. Although there is no direct causal link between SCD and COVID-19, analysis of current data shows that there is a reasonable link between them. According to the latest studies, the incidence of SCD in the community and hospital environment has increased since the outbreak of COVID-19 (Yadav et al., 2020). Based on our findings, PPA2 can increase the risk of COVID-19, so PPA2 may be a potential factor that results in SCD in COVID-19 patients.

Interaction Between the Causal Gene PANO1 and OGT

We searched for the interactions between causal genes of COVID-19 and other genes. As a result, we found that OGT can interact with PANO1. Then, we further investigated the function of OGT to explore the potential mechanism of PANO1 in the risk of COVID-19. At the start of the COVID-19 epidemic, some patients experienced neurological symptoms, such as feeling confused, being unable to discern direction, and feeling restless (Marshall, 2020). A total of 0.2% of the patients of two other SARS-CoV-2-related coronaviruses, SARS-CoV and MERS-CoV, have neurological disease. Given the number of COVID-19 patients, hundreds of thousands of patients may have neurological complications. As genes related to neurological disease and risk genes for COVID-19, OGT, and PANO1 must be considered further.

CONCLUSION

In this article, we used the SMR method to analyze the genes involved in COVID-19 pathogenesis. Here, the risk SNPs for COVID-19 were derived from the GWAS data of Spanish and Italian patients. Lung eQTL data were acquired from the GTEx project. In the postgenomic era, MR and SMR methods have been widely used (Liu et al., 2018). Currently, a large number of pathogenic phenotypes and genes have been identified based on these methods. Through SMR, this article discovered six protein-coding genes and five noncoding genes that can increase the risk of COVID-19. Finally, nine SNPs that met the threshold conditions were identified, and the SMR method was used to determine that these SNPs regulated 11 disease-causing genes that could increase the risk of COVID-19. Then, disease pathway enrichment analysis was performed on these genes.

Through functional analysis, we found that UQCRH participates in cardiac muscle contraction, PPA2 is closely related to SCD, and myocardial damage and SCD occurred in patients with COVID-19. Therefore, our findings provide a potential molecular mechanism for these processes. Further analysis revealed an interaction between OGT and PANO1. OGT is associated with neurological disease. This may explain the neurological complications in COVID-19 patients.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YZ and XL wrote the manuscript and did the experiments. XL provided ideas of this work. YZ analyzed the data. Both authors approved the submitted version.

REFERENCES

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9

Benetti, E., Tita, R., Spiga, O., Ciolfi, A., Birolo, G., Bruselles, A., et al. (2020). ACE2 gene variants may underlie interindividual variability and susceptibility to COVID-19 in the Italian population. *Eur. J. Hum. Genet.* 28, 1602–1614. doi: 10.1038/s41431-020-0691-z

Chen, W., Feng, P., Liu, K., Wu, M., and Lin, H. (2020). Computational identification of small interfering RNA targets in SARS-CoV-2. *Virol. Sin.* 35, 359–361. doi: 10.1007/s12250-020-00221-6

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/1566523219041910221133O7

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi: 10.1093/bib/bbab042

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational methods for identifying similar diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Cui, J., Li, F., and Shi, Z. L. (2019). Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* 17, 181–192. doi: 10.1038/s41579-018-0118-9

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., et al. (2020). Genomewide association study of severe Covid-19 with respiratory failure. *N. Engl. J. Med.* 383, 1522–1534. doi: 10.1056/nejmoa2020283

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9241–9243. doi: 10.1073/pnas.2004999117

Ghosh, A., and Yan, H. (2020). Stability analysis at key positions of EGFR related to non-small cell lung cancer. *Curr. Bioinform.* 15, 260–267. doi: 10.2174/1574893614666191212112026

Gussow, A. B., Auslander, N., Faure, G., Wolf, Y. I., Zhang, F., and Koonin, E. V. (2020). Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci. U.S.A.* 117, 15193–15199. doi: 10.1073/pnas.2008176117

Jiang, S., Du, L., and Shi, Z. (2020). An emerging coronavirus causing pneumonia outbreak in Wuhan, China: calling for developing therapeutic and prophylactic strategies. *Emerg. Microbes Infect.* 9, 275–277. doi: 10.1080/22221751.2020.1723441

Katan, M. B. (1986). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* 1, 507–508. doi: 10.1016/s0140-6736(86)92972-7

Kwon, E., Cho, M., Kim, H., and Son, H. S. (2020). a study on host tropism determinants of influenza virus using machine learning. *Curr. Bioinform.* 15, 121–134. doi: 10.2174/1574893614666191104160927

Li, F., Luo, M., Zhou, W., Li, J., Jin, X., Xu, Z., et al. (2020). Single cell RNA and immune repertoire profiling of COVID-19 patients reveal novel neutralizing antibody. *Protein Cell* 1–5. doi: 10.1007/s13238-020-00807-6

Li, J., Chang, M., Gao, Q., Song, X., and Gao, Z. (2020). Lung cancer classification and gene selection by combining affinity propagation clustering and sparse group lasso. *Curr. Bioinform.* 15, 703–712. doi: 10.2174/1574893614666191017103557

Lippi, G., Lavie, C. J., and Sanchis-Gomar, F. (2020). Cardiac troponin I in patients with coronavirus disease 2019 (COVID-19): evidence from a meta-analysis. *Prog. Cardiovasc. Dis.* 63, 390–391. doi: 10.1016/j.pcad.2020.03.001

Liu, G. Y., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., et al. (2018). Circulating vitamin E levels and Alzheimer's disease: a Mendelian randomization study. *Neurobiol. Aging* 72, 189.e1-189.e9. doi: 10.1016/j.neurobiolaging.2018.08.008

Marshall, M. (2020). How COVID-19 can damage the brain. *Nature* 585, 342–343. doi: 10.1038/d41586-020-02599-5

Meng, C., Zhang, J., Ye, X., Guo, F., and Zou, Q. (2020). Review and comparative analysis of machine learning-based phage virion protein identification methods. *Biochim. Biophys. Acta Proteins Proteom.* 1868:140406. doi: 10.1016/j.bbapap.2020.140406

Narang, P., Dangi, M., Sharma, D., Khichi, A., and Chhillar, A. K. (2019). An integrated Chikungunya virus database to facilitate therapeutic analysis: ChkVDb. *Curr. Bioinform.* 14, 323–332. doi: 10.2174/1574893613666181029124848

Qi, C., Wang, P., Fu, T., Lu, M., Cai, Y., Chen, X., et al. (2021). A comprehensive review for gut microbes: technologies, interventions, metabolites and diseases. *Brief. Funct. Genomics* 20, 42–60. doi: 10.1093/bfgp/elaa029

Qi, R., Ma, A., Ma, Q., and Zou, Q. (2020). Clustering and classification methods for single-cell RNA-sequencing data. *Brief. Bioinform.* 21, 1196–1208. doi: 10.1093/bib/bbz062

Shirazi, S., Mami, S., Mohtadi, N., Ghaysouri, A., Tavan, H., and Nazari, A. (2021). Sudden cardiac death in COVID-19 patients, a report of three cases. *Future Cardiol.* 17, 113–118. doi: 10.2217/fca-2020-0082

Sun, L., Liu, G., Su, L., and Wang, R. (2019). HS-MMGKG: a fast multi-objective harmony search algorithm for two-locus model detection in GWAS. *Curr. Bioinform.* 14, 749–761. doi: 10.2174/1574893614666190409110843

Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. (2020). Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS Coronavirus. *J. Virol.* 94, e127–e120.

Wang, P., Jin, X., Zhou, W., Luo, M., Xu, Z., Xu, C., et al. (2020). Comprehensive analysis of TCR repertoire in COVID-19 using single cell sequencing. *Genomics* 113, 456–462. doi: 10.1016/j.ygeno.2020.12.036

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.

Yadav, R., Bansal, R., Budakoty, S., and Barwad, P. (2020). COVID-19 and sudden cardiac death: a new potential risk. *Indian Heart J.* 72, 333–336. doi: 10.1016/j.ihj.2020.10.001

Yuan, F., Lu, L., and Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochim. Biophys. Acta Mol. Basis Dis.* 1866:165822. doi: 10.1016/j.bbadis.2020.165822

Zhao, S., Jiang, H., Liang, Z. H., and Ju, H. (2019). Integrating multi-omics data to identify novel disease genes and single-neucleotide polymorphisms. *Front. Genet.* 10:1336. doi: 10.3389/fgene.2019.01336

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet.* 10:1021. doi: 10.3389/fgene.2019.01021

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 21:43. doi: 10.1186/s12859-020-3388-y

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270–273. doi: 10.1038/s41586-020-2012-7

Zhu, X., Li, H.-D., Guo, L., Wu, F.-X., and Wang, J. (2019). Analysis of single-cell RNA-seq data by clustering approaches. *Curr. Bioinform.* 14, 314–322. doi: 10.2174/1574893614666181120095038

Zou, Q., Lin, G., Jiang, X., Liu, X., and Zeng, X. (2020). Sequence clustering in bioinformatics: an empirical study. *Brief. Bioinform.* 21, 1–10. doi: 10.1093/bib/bby090

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 July 2021
doi: 10.3389/fgene.2021.712164

Identification of Parkinson's Disease-Causing Genes via Omics Data

*Xinran Cui†, Chen Xu†, Liyuan Zhang and Yadong Wang**

Center for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

Parkinson's disease (PD) is the second most frequent neurogenic disease after Alzheimer's disease. The clinical manifestations include mostly motor disorders, such as bradykinesia, myotonia, and static tremors. Since the cause of this pathological features remain unclear, there is currently no radical treatment for PD. Environmental and genetic factors are thought to contribute to the pathology of PD. To identify the genetic factors, some studies employed the Genome-Wide Association Studies (GWAS) method and detected certain genes closely related to PD. However, the functions of these gene mutants in the development of PD are unknown. Combining GWAS and expression Quantitative Trait Loci (eQTL) analysis, the biological meaning of mutation could be explained to some extent. Therefore, the present investigation used Summary data-based Mendelian Randomization (SMR) analysis to integrate of two PD GWAS datasets and four eQTL datasets with the objective of identifying casual genes. Using this strategy, we found six Single Nucleotide Polymorphism (SNP) loci which could cause the development of PD through altering the susceptibility gene expression, and three risk genes: Synuclein Alpha (SNCA), Mitochondrial Poly(A) Polymerase (MTPAP), and RP11-305E6.4. We proved the accuracy of results through case studies and inferred the functions of these genes in PD. Overall, this study provides insights into the genetic mechanism behind PD, which is crucial for the study of the development of this disease and its diagnosis and treatment.

Keywords: Parkinson's disease, SMR analysis, GWAS summary data, eQTL summary data, risk genes

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Yuansong Zhao,
University of Texas Health Science Center at Houston, United States
Sheng Li,
Zhongnan Hospital, Wuhan University, China

***Correspondence:**
Yadong Wang
ydwang@hit.edu.cn

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

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

Received: *20 May 2021*
Accepted: *02 July 2021*
Published: *28 July 2021*

Citation:
Cui X, Xu C, Zhang L and Wang Y (2021) Identification of Parkinson's Disease-Causing Genes via Omics Data. Front. Genet. 12:712164. doi: 10.3389/fgene.2021.712164

INTRODUCTION

Parkinson's disease (PD) is the second most common degenerative disorder of the nervous system. As the incidence of this disease is strongly linked to age, approximately 1% of 65-year-olds has this disease, rising to 4–5% among aged 85 (Trinh and Farrer, 2013). Statistically, the rate of PD increases 5–10 times from the age of 60 to the age of 90 (Poewe et al., 2017). The main clinical manifestations are involuntary limb tremor, bradykinesia, walking difficulty, and stiff limbs, and these symptoms become aggravated with time. This causes that PD patients are peculiarly prone to falls on routine activities. The incidence of falls could reach 40–70% (Kerr et al., 2010). Falling can lead to injury and the decrease of survival for PD patients. Thus, the health and life of human beings, especially the elderly, are threatened by this disease.

There are two main pathologic characteristics of PD. One is the degeneration and death of dopamine neurons in the substantia nigra pars compacta of the midbrain and the consequent depletion of dopamine in the striatum (Ammal Kaidery and Thomas, 2018). Dopamine is synthesized by dopamine neurons, and then delivered to the striatum to regulate somatic motor (Schwarz and Peever, 2011). The other feature is the formation of acidophilic inclusions known as Lewy bodies in the cytoplasm of the remaining neurons in the substantia nigra (Feng et al., 2020). Although many studies have explored the pathological mechanism of PD, there is still no sufficient evidence to explain the degeneration of dopaminergic neurons and the formation of Lewy bodies, thus the current PD treatment approaches can only relieve symptoms with medication but cannot reverse this disease progression (Ball et al., 2019). Moreover, long-term treatment results in the development of drug resistance, and the available drugs have significant adverse effects. Therefore, it is urgent to understand the biological process leading to these pathological changes for the cure of PD.

Some studies revealed that both environmental and genetic factors contribute to the pathological features of PD. The main biological mechanism by which environmental factors can damage dopaminergic neurons involves the inhibition of the activity of mitochondrial complex enzymes and the mitochondrial respiratory chain (Holper et al., 2019). In addition to environmental factors, some patients may have inherited certain particular mutated gene that lead to the development of PD (Antony et al., 2013). Understanding what these genes do has crucial implications for understanding the changes in the biological processes underlying PD. Thus, it is necessary to reveal these mutant genes for conquering this disease.

A genome-wide association study (GWAS) identified more than 10 PD-causing genes (Nalls et al., 2014). GWAS is a strategy for identifying common genetic variants [Single Nucleotide Polymorphism (SNP)] significantly associated with a complex trait or a disease in the whole human genome, thus recognizing the disease-related genes (Cannon and Mohlke, 2018). In the results of GWAS analysis, most significant SNP sites were located in non-coding regions, making it difficult to directly explore the regulatory mechanism of these sites (Rojano et al., 2016). Thus, there are some PD risk genes revealed in many GWAS studies, what role do these genes play in the development of PD remains unknown.

The combined analysis of GWAS and expression Quantitative Trait Loci (eQTL) has become an important means to reveal the function of significant variants. Although GWAS have identified thousands of variants associated with complex traits, their biological explanation is often still unclear. Most of these variants overlap with eQTL, suggesting that they may be involved in the regulation of gene expression (Zhu et al., 2016). Genes associated with these variants could be regarded as PD causing-genes. This study exploited Summary data-based Mendelian Randomization analysis (SMR) to integrate and analyze the PD summary data of the GWAS with the summary data of eQTL, to explore the genetic mechanism by which certain disease-causing genes contribute to PD. The SMR analysis method does not require the data with both genotype and gene expression and a massive size. Thus, this approach could leverage the published data to a large extent. The statistical analysis of the relationship between a single SNP and gene expression is called the eQTL analysis (Shabalin, 2012). If the expression of a gene is affected by a SNP, then this genetic variant is considered as an eQTL locus. Since SNP is the subject of study in both GWAS and eQTL, SNP is used as an instrumental variable in the SMR method to determine which genes expression changes could lead to the occurrence of PD. Thus, the SMR analysis results may provide a direction for the treatment of PD.

MATERIALS AND METHODS

Data Acquisition

Two public summary datasets of GWAS for PD were downloaded from the GWAS catalog website. One of the GWAS datasets represented data from 282,871 white British inpatient samples reported by the UK Biobank. The UK Biobank is a cohort study collecting, physical, and health data of approximately 500,000 British individuals. For the purpose of this analysis, this dataset was named "GUB" (Bi et al., 2020). The other GWAS dataset is based on the genetic data of 28,568 PD patients obtained from International Parkinson's Disease Genomics Consortium and was named "GIPD" (Blauwendraat et al., 2019). This data consists of Parkinson's patients from European countries such as United Kingdom, Dutch, Finnish, and German. The present study also employed four eQTL datasets. EQTL data are generally collected from peripheral blood, thus one dataset is the summary level statistics of eQTL from the Consortium for the Architecture of Gene Expression (CAGE) data. It provides the measurements of the level of gene expression in peripheral blood. This dataset contains more than 3 million SNPs, identified by 33,323 probes. Since PD is a neurodegenerative disease, the second dataset includes the level of gene expression in brain tissue and includes information on 28,522 probes and more than 13 million SNPs. To explore whether PD is associated with other factors such as reduced immunity, we selected two sets of eQTL data sets for T cells. The remaining two datasets list the level of gene expression in CD4- and CD8-positive cells, respectively. CD4 and CD8 are both markers of T lymphocytes. The CD4 dataset includes more than five hundred thousand SNPs, measured by 7,350 probes, while the CD8 dataset includes more than three hundred thousand SNPs through using 5,829 probes.

SMR Analysis

Both GWAS and eQTL were used to investigate the relationship between SNP and traits or gene expression through linear regression analysis. In the regression analysis, the effect size (beta-value) corresponds to the value of the regression coefficient, while SE stands for the standard error of the regression coefficient. Then, the GWAS and eQTL data were standardized using the Z-score method, in which the Z was calculated as the quotient of the beta-value and the SE-value.

After computing the Z-score, we performed SMR analysis on an eQTL dataset and a GWAS dataset (**Figure 1**). Since SNPs are regarded as instrumental variables in SMR analysis, we identified the overlapped SNPs between an eQTL dataset and a GWAS

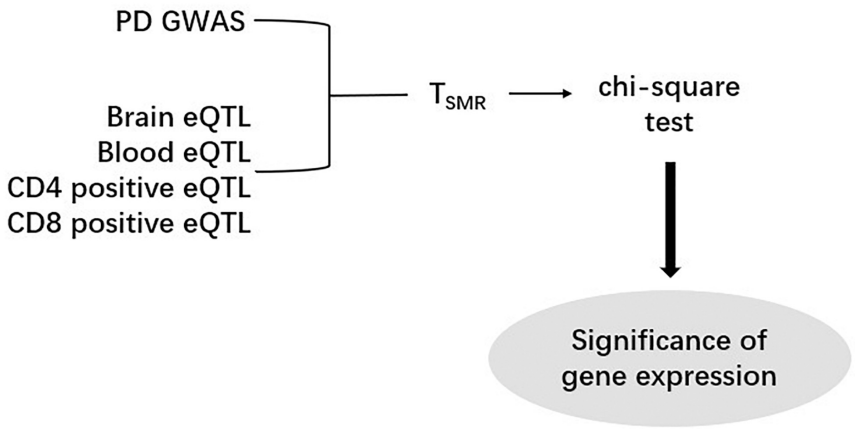

FIGURE 1 | The Summary data-based Mendelian Randomization (SMR) analysis flow chart. One PD GWAS was integrated with each of eQTL data. After calculating T_{SMR} and chi-square test, the significance of gene expression were discovered.

dataset and then generated a new dataset including all eQTL and GWAS data of the overlapped SNPs. As a result, eight new datasets were obtained by this approach, and each of them was subsequently subjected to the SMR analysis. According to the formula,

$$T_{SMR} \approx \frac{Z_{GWAS}^2 \times Z_{eQTL}^2}{Z_{GWAS}^2 + Z_{eQTL}^2}$$

the Z values of GWAS and eQTL were used to calculate the T_{SMR} value. The chi-square test was applied to the T_{SMR} values to calculate their *P* values (P_{SMR}). Each SMR dataset has a specific threshold. This threshold is calculated by dividing 0.05 by the number of probes in the corresponding eQTL dataset. If we find that some P_{SMR} values are less than the threshold value of their data set, the genes corresponding to these P_{SMR} can be considered as risk genes.

RESULTS

Identification of Overlapping SNPs

Since SNP is an instrumental variable, we searched for the same SNPs between a GWAS dataset and an eQTL dataset, generating eight new datasets (**Table 1** and **Supplementary Figures 1, 2**). The results showed that both GUB_Brain and GIPD_Brain datasets contain more than 10 million overlapped SNPs and about 28, 000 genes. There are over two million SNPs and about 20,000 in GUB_Cage and the GIPD_Cage datasets. The other four data have fewer than half a million SNPs and about four to five thousand genes.

Subsequently, we compared the two datasets generated by using the same eQTL data and found that the genes overlapping between the two datasets accounted for about 98–100% of genes in each dataset, implying that the PD-associated genes identified in the two GWAS datasets are highly similar (**Figure 2**). Additionally, we analyzed the four datasets generated by the same GWAS dataset. This analysis showed that the overlap rate of the genes was not high among these datasets, with the largest overlap being less than 50% (**Figure 3**). Even though CD4 and CD8 are both markers of T lymphocytes, the gene overlap rate between the GUB_CD4 and GUB_CD8 datasets or between the GIPD_CD4 and GIPD_CD8 datasets were around 36%. This indicates a low degree of correlation between these eQTL datasets, which may be caused by the large differences in the number of SNPs found.

SMR Analysis

The SMR approach was employed to analyze the eight datasets. The P_{SMR} value calculated was compared with the corresponding threshold value, and finally three risk genes were found (**Tables 2**, **3**). As **Supplementary Figure 3** shown, whether GUB GWAS or GIPD GWAS, the significant SNPs were only located on chromosome 4. However, the results of the SMR analysis showed that that GUB_Brain identified one target SNP locus associated with two genes: Mitochondrial Poly(A) Polymerase (MTPAP), RP11-305E6.4, which is located on chromosome 10. Coincidentally, GUB_Cage identified only this SNP locus, and the MTPAP gene associated with this SNP was also identified. Additionally, GIPD_Cage identified the highest number of the SNP sites of interest. These five SNPs were located on chromosome 4 and corresponded to the same Synuclein Alpha (SNCA) gene. However, the risk genes were not detected in the GIPD_Brain dataset. Similarly, based on **Supplementary Figures 3E–H**, no significant gene was identified

TABLE 1 | The new dataset list for SMR analysis.

eQTL	GWAS			
	Brain	**Cage**	**CD4**	**CD8**
GUB	GUB_Brain	GUB_Cage	GUB_CD4	GUB_CD8
GIPD	GIPD_Brain	GIPD_Cage	GIPD_CD4	GIPD_CD8

Each GWAS data combined with each eQTL data were analyzed to produce a new SMR data. Thus, there were eight SMR datasets.

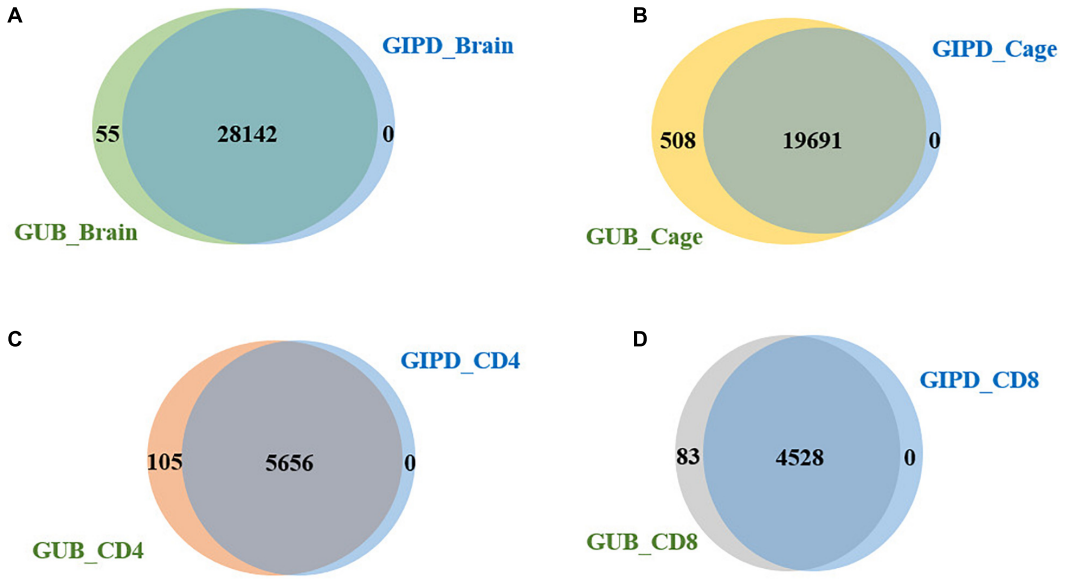

FIGURE 2 | (A) The overlap genes of the GUB_Brain and GIPD_Brain data. 28,142 genes were overlapped. **(B)** The overlap genes of the GUB_Cage and GIPD_Cage data. There were 19,691 overlapped genes. **(C)** The overlap genes of the GUB_CD4 and GIPD_CD4 data. These two datasets have 5,656 identical genes. **(D)** The overlap genes of the GUB_CD8 and GIPD_CD8 data. These two datasets have 4,528 identical genes.

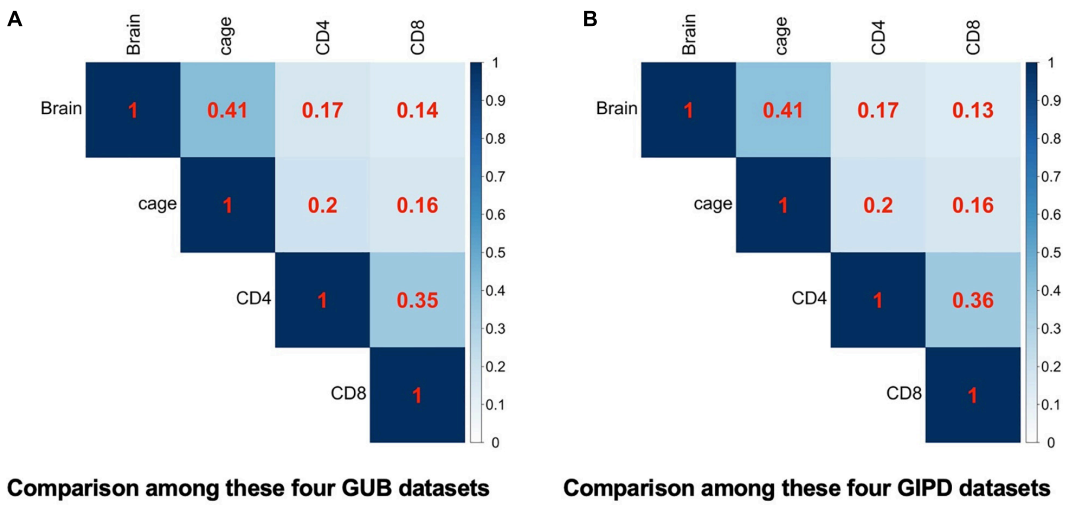

FIGURE 3 | (A) The heatmap of the four GUB data. **(B)** The heatmap of the four GIPD data. The correlation of these four different eQTL datasets were compared under the background of the same GWAS dataset.

for GUB _CD4, GUB _CD8, GIPD_ CD4, and GIPD_CD8 datasets, likely due to significantly lower *p*-values of CD4 and CD8 eQTL than those of the other two eQTL datasets. Thus, the SMR analysis of the 8 datasets identified a total of six candidate SNP loci and three genes, and the expression level of these three genes can affect the occurrence of PD.

Gene Function Analysis

Three candidate genes were SNCA, MTPAP, and RP11-305E6.4. To investigate how the expression of the three genes identified using SMR analysis contributes to the development of PD, we searched for their function in the KEGG database and related publications. The main pathological features of PD consist of the formation of Lewy bodies. The main component of the Lewy body is α-synuclein (α-syn) encoded by the SNCA gene. Mutation of this gene can cause the overexpression of α-syn, leading to the formation of Lewy bodies and hence the development of PD. The MTPAP gene encodes a nuclear polymerase responsible for generating homopolymerized (A) tails on mitochondrial mRNA. Although the search results did not reveal an evident relationship between this gene and PD, some studies show that mitochondrial dysfunction is involved in

TABLE 2 | P_{SMR} threshold for these eight SMR datasets.

SMR dataset	GUB_Brain	GUB_Cage	GUB_CD4	GUB_CD8
P_{SMR} threshold	1.8×10^{-6}	1.5×10^{-6}	6.8×10^{-6}	8.6×10^{-6}
SMR dataset	GIPD_Brain	GIPD_Cage	GIPD_CD4	GIPD_CD8
P_{SMR} threshold	1.8×10^{-6}	1.5×10^{-6}	6.8×10^{-6}	8.6×10^{-6}

The same eQTL datasets have the same threshold.

TABLE 3 | Discovery of PD causative gene by the Summary data-based Mendelian Randomization (SMR) analysis for these eight datasets.

SMR datasets	Number of discovered genes	Gene name	P value
GUB_Brain	2	MTPAP	7.215×10^{-7}
		RP11-305E6.4	1.624×10^{-6}
GUB_Cage	1	MTPAP	4.191×10^{-7}
GUB_CD4	0	–	–
GUB_CD8	0	–	–
GIPD_Brain	0	–	–
GIPD_Cage	1	SNCA	4.671×10^{-7}
GIPD_CD4	0	–	–
GIPD_CD8	0	–	–

This list showed the number of genes found in different datasets, their names, and the value of P_{SMR} calculated.

the pathogenesis of many neurodegenerative diseases, including PD. Abnormal mitochondrial structure or function has been found to induce a progressive loss of dopaminergic neurons and even trigger PD symptoms. Although the exact function of the polyadenylation of mitochondrial mRNA is unknown, the process is essential for maintaining correct mRNA expression in the mitochondria, and its disruption can lead to mitochondrial dysfunction. Therefore, the mutation of this gene may block the expression of MTPAP, causing mitochondrial dysfunction and leading to PD. The MTPAP gene is also known as PAPD1 or RP11-305E6.3, and another gene was found to be RP11-305E6.4. Both MTPAP and RP11-305E6.4 gene corresponds to the same SNP locus in this SMR analysis. This SNP may be localized in a non-coding gene regulatory region between these two adjacent genes. However, the function of this gene has not been identified yet, and its impact on PD remains unknown.

DISCUSSION

The SMR method was employed to integrate two GWAS datasets and four eQTL datasets. This approach identified six SNP candidate loci and three risk genes whose expression can significantly influence on the development of PD. The SNCA gene is the first confirmed pathogenic gene for PD (Lunati et al., 2018). It is located on chromosome 4 and contains six exons. The SNCA gene encodes the α-syn protein that is the main component of Lewy bodies (Mehra et al., 2019). α-syn is abundant in the brain and is also expressed in the heart, skeletal muscle, and other tissues. In the brain, α-syn is found primarily in presynaptic terminals, which release neurotransmitters essential for normal brain function. Mutation of the SNCA gene can cause the overexpression of the α-syn protein, leading to the formation of Lewy bodies and the development of PD. Different types of variations in the coding region and non-coding regions of the SNCA gene can increase its transcription and translation. The level of α-syn protein can also be increased by point mutation or copy number duplication of the SNCA gene (Kim, 2013). Moreover, in SNCA copy number repeat variation, the disease was more severe in the presence of the triploid type than the diploid type, indicating that the expression of α-syn may positively correlate with the severity of PD (Nussbaum, 2018). Therefore, the SNCA gene can be considered to be an effective target for the treatment of PD. SNCA is also believed to be involved in various other neurodegenerative diseases, such as Alzheimer's disease, Lewy body disease, and muscular atrophy. Thus, the development of methods to inhibit SNCA gene mutations and decrease the formation of aggregates are of great clinical relevance.

Additionally, the polyadenylation of mRNA by the nuclear DNA-encoded mitochondrial poly(A) RNA polymerase is crucial for maintaining gene expression in human mitochondria (Lapkouski and Hällberg, 2015). Although the exact function of mitochondria mRNA transcription of adenosine acidification is not yet fully understood, the process is essential for ensuring correct mRNA expression in the mitochondria. MTPAP mutant proteins can shorten polyadenylation of mitochondrial mRNA, resulting in post-transcriptional downregulation of the expression of components of the respiratory chain complex and the impairment of an essential mitochondrial function (Wilson et al., 2014). Mitochondrial dysfunction is involved in many processes and diseases, including aging, cancer, diabetes, and neurodegenerative diseases such as PD and Alzheimer's disease (Larsen et al., 2018). Among several mechanisms responsible for the pathogenesis of PD, mitochondrial dysfunction may be related to the death of dopaminergic neurons. Many of the PD-associated gene mutations result in an abnormal mitochondrial function and, eventually, neuronal damage, which is a critical component of the onset and development of the disease. Thus, compounds that target mitochondria and improve their function represent potential therapeutic options for delaying and treating degenerative diseases of the central nervous system.

RP11-305E6.4 is a long non-coding RNA (lncRNA) gene. LncRNAs are non-coding RNA molecules with a length of more than 200 nucleotides, which can govern gene expression, transcription, and post-transcription. Currently, there are few studies on this gene, and it is still unknown which genes are regulated by this lncRNA to influence the occurrence of PD. Nevertheless, the transcript of RP11-305E6.4 overlaps with that of MTPAP, and further studies could be conducted to determine whether this gene can regulate MTPAP to affect the occurrence of PD in the future.

In conclusion, we verified that the abnormal expression of the SNCA gene could lead to PD and found that the abnormal expression of the MTPAP and RP11-305E6.4 genes may also cause PD. This study further demonstrates that the design of drugs targeting SNCA gene is conducive to inhibit the formation of Lewy bodies, and to completely cure PD. Moreover, we have identified two new candidate genes for PD. This provides a research direction for understanding the biological significance behind the pathological features of PD.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST007001-GCST008000/GCST007780/; http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST010001-GCST011000/GCST010765/.

AUTHOR CONTRIBUTIONS

XC, CX, and LZ contributed to the design and implementation of the research, to the analysis of the results, and to the writing of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been supported by the National Key Research and Development Program of China (2017YFC0907503).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.712164/full#supplementary-material

Supplementary Figure 1 | (A) The Manhattan diagram of GUB GWAS data. **(B)** The Manhattan diagram of GIPD GWAS data.

Supplementary Figure 2 | (A) The Manhattan diagram of Brain eQTL data. **(B)** The Manhattan diagram of Cage eQTL data. **(C)** The Manhattan diagram of CD4 eQTL data. **(D)** The Manhattan diagram of CD8 eQTL data.

Supplementary Figure 3 | (A) The Summary data-based Mendelian Randomization (SMR) analysis results of GUB_Brain. **(B)** The SMR analysis results of GIPD_Brain. **(C)** The SMR analysis results of GUB_Cage. **(D)** The SMR analysis results of GIPD_Cage. **(E)** The SMR analysis results of GUB_CD4. **(F)** The SMR analysis results of GIPD_CD4. **(G)** The SMR analysis results of GUB_CD8. **(H)** The SMR analysis results of GIPD_CD8.

REFERENCES

Ammal Kaidery, N., and Thomas, B. (2018). Current perspective of mitochondrial biology in Parkinson's disease. *Neurochem. Int.* 117, 91–113. doi: 10.1016/j.neuint.2018.03.001

Antony, P. M., Diederich, N. J., Krüger, R., and Balling, R. (2013). The hallmarks of Parkinson's disease. *FEBS J.* 280, 5981–5993. doi: 10.1111/febs.12335

Ball, N., Teo, W. P., Chandra, S., and Chapman, J. (2019). Parkinson's disease and the environment. *Front. Neurol.* 10:218. doi: 10.3389/fneur.2019.00218

Bi, W., Fritsche, L. G., Mukherjee, B., Kim, S., and Lee, S. (2020). A fast and accurate method for genome-wide time-to-event data analysis and its application to UK biobank. *Am. J. hum. Genet.* 107, 222–233. doi: 10.1016/j.ajhg.2020.06.003

Blauwendraat, C., Heilbron, K., Vallerga, C. L., Bandres-Ciga, S., von Coelln, R., Pihlstrøm, L., et al. (2019). Parkinson's disease age at onset genome-wide association study: defining heritability, genetic loci, and α-synuclein mechanisms. *Mov. Disord.* 34, 866–875. doi: 10.1002/mds.27659

Cannon, M. E., and Mohlke, K. L. (2018). Deciphering the emerging complexities of molecular mechanisms at GWAS Loci. *Am. J. hum. Genet.* 103, 637–653. doi: 10.1016/j.ajhg.2018.10.001

Feng, Y. S., Yang, S. D., Tan, Z. X., Wang, M. M., Xing, Y., Dong, F., et al. (2020). The benefits and mechanisms of exercise training for Parkinson's disease. *Life Sci.* 245:117345. doi: 10.1016/j.lfs.2020.117345

Holper, L., Ben-Shachar, D., and Mann, J. J. (2019). Multivariate meta-analyses of mitochondrial complex I and IV in major depressive disorder, bipolar disorder, schizophrenia, Alzheimer disease, and Parkinson disease. *Neuropsychopharmacology* 44, 837–849. doi: 10.1038/s41386-018-0090-0

Kerr, G. K., Worringham, C. J., Cole, M. H., Lacherez, P. F., Wood, J. M., and Silburn, P. A. (2010). Predictors of future falls in Parkinson disease. *Neurology* 75, 116–124. doi: 10.1212/WNL.0b013e3181e7b688

Kim, H. J. (2013). Alpha-synuclein expression in patients with Parkinson's disease: a Clinician's perspective. *Exp. Neurobiol.* 22, 77–83. doi: 10.5607/en.2013.22.2.77

Lapkouski, M., and Hällberg, B. M. (2015). Structure of mitochondrial poly(A) RNA polymerase reveals the structural basis for dimerization, ATP selectivity and the SPAX4 disease phenotype. *Nucleic Acids Res.* 43, 9065–9075. doi: 10.1093/nar/gkv861

Larsen, S. B., Hanss, Z., and Krüger, R. (2018). The genetic architecture of mitochondrial dysfunction in Parkinson's disease. *Cell Tissue Res.* 373, 21–37. doi: 10.1007/s00441-017-2768-8

Lunati, A., Lesage, S., and Brice, A. (2018). The genetic landscape of Parkinson's disease. *Rev. neurol.* 174, 628–643. doi: 10.1016/j.neurol.2018.08.004

Mehra, S., Sahay, S., and Maji, S. K. (2019). α-Synuclein misfolding and aggregation: Implications in Parkinson's disease pathogenesis. *Biochim. Biophys. Acta Proteins Proteom.* 1867, 890–908. doi: 10.1016/j.bbapap.2019.03.001

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., et al. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* 46, 989–993. doi: 10.1038/ng.3043

Nussbaum, R. L. (2018). Genetics of synucleinopathies. *Cold Spring Harb. Perspect. Med.* 8:a024109. doi: 10.1101/cshperspect.a024109

Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., et al. (2017). Parkinson disease. *Nat. Rev. Dis. Primers* 3:17013. doi: 10.1038/nrdp.2017.13

Rojano, E., Ranea, J. A., and Perkins, J. R. (2016). Characterisation of non-coding genetic variation in histamine receptors using AnNCR-SNP. *Amino Acids* 48, 2433–2442. doi: 10.1007/s00726-016-2265-5

Schwarz, P. B., and Peever, J. H. (2011). Dopamine triggers skeletal muscle tone by activating D1-like receptors on somatic motoneurons. *J. Neurophysiol.* 106, 1299–1309. doi: 10.1152/jn.00230.2011

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* 28, 1353–1358. doi: 10.1093/bioinformatics/bts163

Trinh, J., and Farrer, M. (2013). Advances in the genetics of Parkinson disease. *Nat.Rev. Neurol.* 9, 445–454. doi: 10.1038/nrneurol.2013.132

Wilson, W. C., Hornig-Do, H. T., Bruni, F., Chang, J. H., Jourdain, A. A., Martinou, J. C., et al. (2014). A human mitochondrial poly(A) polymerase mutation reveals the complexities of post-transcriptional mitochondrial gene expression. *Hum. Mol. Genet.* 23, 6345–6355. doi: 10.1093/hmg/ddu352

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

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fgene.2021.726599

Genetic Mechanism Revealed of Age-Related Macular Degeneration Based on Fusion of Statistics and Machine Learning Method

*Yongyi Du[1], Ning Kong[1] and Jibin Zhang[2]**

[1] Department of Ophthalmology, Panyu Central Hospital, Guangzhou, China, [2] Department of Stomatology, Panyu Central Hospital, Guangzhou, China

Age-related macular degeneration (AMD) is the most common cause of irreversible vision loss in the developed world which affects the quality of life for millions of elderly individuals worldwide. Genome-wide association studies (GWAS) have identified genetic variants at 34 loci contributing to AMD. To better understand the disease pathogenesis and identify causal genes for AMD, we applied random walk (RW) and support vector machine (SVM) to identify AMD-related genes based on gene interaction relationship and significance of genes. Our model achieved 0.927 of area under the curve (AUC), and 65 novel genes have been identified as AMD-related genes. To verify our results, a statistics method called summary data-based Mendelian randomization (SMR) has been implemented to integrate GWAS data and transcriptome data to verify AMD susceptibility-related genes. We found 45 genes are related to AMD by SMR. Among these genes, 37 genes overlap with those found by SVM-RW. Finally, we revealed the biological process of genetic mutations leading to changes in gene expression leading to AMD. Our results reveal the genetic pathogenic factors and related mechanisms of AMD.

Keywords: AMD, GWAS, eQTL, SNPs, disease susceptibility

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China
Hui Ding,
University of Electronic Science and Technology of China, China

****Correspondence:***
Jibin Zhang
kn13@tom.com

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

Received: *17 June 2021*
Accepted: *13 July 2021*
Published: *05 August 2021*

Citation:
Du Y, Kong N and Zhang J (2021) Genetic Mechanism Revealed of Age-Related Macular Degeneration Based on Fusion of Statistics and Machine Learning Method. Front. Genet. 12:726599. doi: 10.3389/fgene.2021.726599

INTRODUCTION

Age-related macular degeneration (AMD) is the most common cause of irreversible blindness with limited therapeutic options in the elderly in many countries (Lim et al., 2012). AMD causes decreased photoreceptor function in the macular area of the retina (Fritsche et al., 2014). Researchers have found many factors which are related to the development and severity of AMD.

Genetic factors are significantly related to AMD. In 2005, Klein et al. found that CFH gene was related to AMD, which was the first discovered AMD-related gene (Haines et al., 2005). This gene is significantly expressed in retinal pigment epithelial cells. Y402H mutation of CFH impairs the complement pathway regulation function of CFH gene (Landowski et al., 2019). Subsequently, the ARMS2 gene cluster was also found to be related to AMD. Multiple studies have shown that there is a strong correlation between multiple genetic variants in this gene cluster and AMD (Johnson et al., 2001). Recently, it has been discovered that the apolipoprotein E (APOE) gene

has a strong correlation with AMD (Fernández-Vega et al., 2020). The APOE gene plays a role in transporting lipids and cholesterol in the central nervous system, and multiple studies have shown that this gene is associated with neurodegenerative diseases such as Alzheimer's disease and stroke (Feher et al., 2006; Zhao et al., 2019, 2020d). The gene is expressed on photoreceptor cells, retinal ganglion cells, retinal pigment epithelial cells, Bruch's membrane, and the choroid. Most studies have proved APOE can prevent AMD (Pang et al., 2000). The genetic risk of advanced AMD is increased (Heiba et al., 1994). Researchers have found that the heritability estimate for twin studies is 0.45 for early AMD (Hammond et al., 2002) but 0.71 for late AMD (Seddon et al., 2005).

Computational methods have been widely used to discover functions of biological molecules (Zhao et al., 2020a,b, 2021a). AMD-related genome-wide association studies (GWAS) analyses have identified a strong association of 52 independent single-nucleotide polymorphisms (SNPs) at 34 genetic loci accounting for over 50% of the genetic heritability (Fritsche et al., 2016). Machine learning methods can help researchers find disease-related information on a large scale. However, these methods cannot explain the genetic mechanism of the results. GWAS studies are a valuable resource for understanding disease pathologies, but they may not precisely point out the causal genes responsible for the disease of interest. Besides, there have been studies that reported that causal genes are distinct from the nearest genes discovered by GWAS (Smemo et al., 2014; Claussnitzer et al., 2015). However, The gene expression is related to the genetic variant so the gene expression levels are different in different genotypes (Zhao et al., 2020c). Expression quantitative trait locus (eQTL) mapping offers a powerful approach to elucidate the genetic component underlying altered gene expression. Gene expression is vital for complex diseases (Zhao et al., 2021b) and is also differentially regulated across tissues, such as the brain, heart, and pancreas. Ratnapriya et al. (2019) have found potential causal genes in six AMD GWAS loci from human retinal samples. However, that analysis only considered retinal samples and was not comprehensive since it is difficult to obtain multiple living tissues and most eQTL studies so far have been performed with RNA isolated from immortalized lymphoblasts or lymphocytes. In this study, we fused random walk (RW) with support vector machine (SVM) to identify AMD-related genes. Since many GWAS and eQTL studies have been made public, to verify our results, AMD GWAS data and blood eQTL studies are integrated to further find expression of the genes related to AMD. In this method, we referred to the concept of Mendelian randomization (MR) analysis (Davey Smith and Ebrahim, 2003; Katan, 2004), where a genetic variant (such as a SNP) is considered as an instrumental variable (such as gene expression) to validate for the causative effect of an exposure on an outcome (such as a phenotype). Based on this assumption, we can obtain AMD-related genes based on MR. We collected eQTL data from the GTEx database and collected GWAS datasets including 12,711 advanced AMD cases and 14,590 controls of European descent from a study by Han et al. (2020); 707 Caucasian AMD patients and 2,014 controls from a study by Yan et al. (2018); and 14,034 cases, 91,214 controls, and 11 sources of data including the International AMD Genomics Consortium, IAMDGC, and United Kingdom Biobank (UKBB) from a study by Winkler et al. (2020). Based on these GWAS studies and eQTL dataset, we can not only identify genes related to AMD but also speculate on their biological processes.

MATERIALS AND METHODS

Encoding Gene Interaction Network by Random Walk

The RW algorithm is a method that is simple to operate but not easy to fall into a local minimum. We constructed a gene interaction network by known AMD-related genes and a string database. Then, we implemented RW on the gene interaction network.

$f(x)$ is a multivariate function with n variables; $x = (x_1, x_2, \ldots, x_n)$ is an n dimension vector.

Step 1: Given the initial iteration point x, λ is the first walking step length, and ϵ is the control accuracy (ϵ is a very small positive number, used to control the end of the algorithm).

Step 2: Given the number of iterations control N, k is the current iteration number; set $k = 1$.

Step 3: When $k < N$, randomly generate an n-dimensional vector between $(-1, 1)$. $u = (u_1, u_2, \ldots, u_n)$, $(-1 < u_i < 1, i = 1, 2,\ldots, n)$, and standardize it to get u'.

$$u' = \frac{u}{\sqrt{\sum u_i^2}}$$

Let $x1 = x + \lambda u'$ to complete the first step of walking.

Step 4: Calculate the value of the function, if $f(x1) < f(x)$, which is a better point than the initial value, then reset k to 1, change $x1$ to x, and go back to step 2; otherwise, $k = k + 1$. Go back to step 3.

Step 5: If no better value can be found for N consecutive times, it is considered that the optimal solution is within the N-dimensional sphere with the current optimal solution as the center and the current step as the radius (if it is three-dimensional, it just happens to be in the space sphere). At this point, if $\lambda < \epsilon$, the algorithm ends; otherwise, let $\lambda = \lambda 2$, go back to step 1, and start a new round of walking.

Finally, we can get the gene feature after encoding the gene network.

Classification by Support Vector Machine

We obtained the gene feature in the last section. Then, we can input the gene feature and label into SVM to get the relationship between the gene and AMD. The workflow of SVM is shown in **Figure 1**.

First, we used Z-score normalization to process the gene feature. Then, we constructed a Lagrangian function to obtain the values and dualized the original problem. Sequential minimal optimization (SMO) algorithm was used to solve the dualization

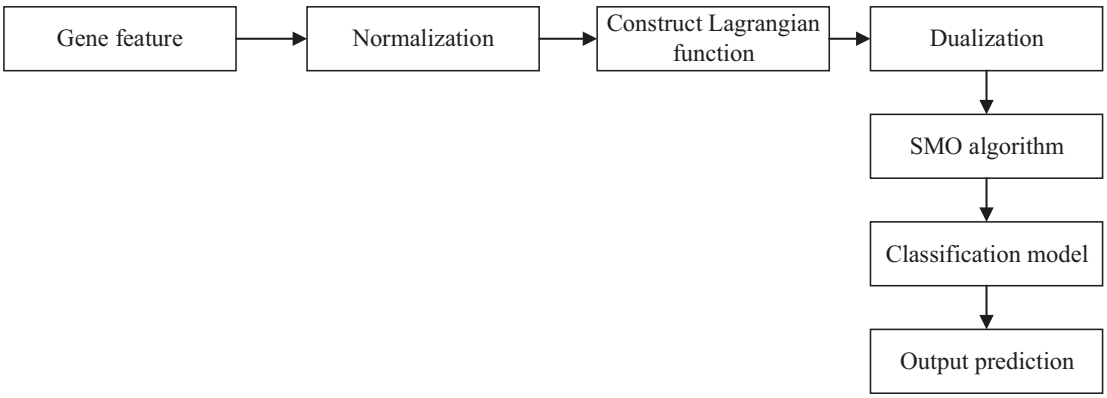

FIGURE 1 | Workflow of SVM. SVM, support vector machine; SOM, sequential minimal optimization.

problem. Finally, we can obtain the classification model and output the prediction results.

RESULTS

AMD-Related Genes Identification by SVM-RW

We obtained 34 known AMD-related genes from GWAS data. We constructed a gene network which has 239 nodes (genes). We did 10-cross validation by SVM-RW and tested the performance of SVM-RW. The area under the curve (AUC) of SVM-RW is shown in **Figure 2**.

SVM-RW achieved AUC of 0.927 in identifying AMD-related genes. We compared the results of SVM-RW with several other methods. The results are shown in **Table 1**.

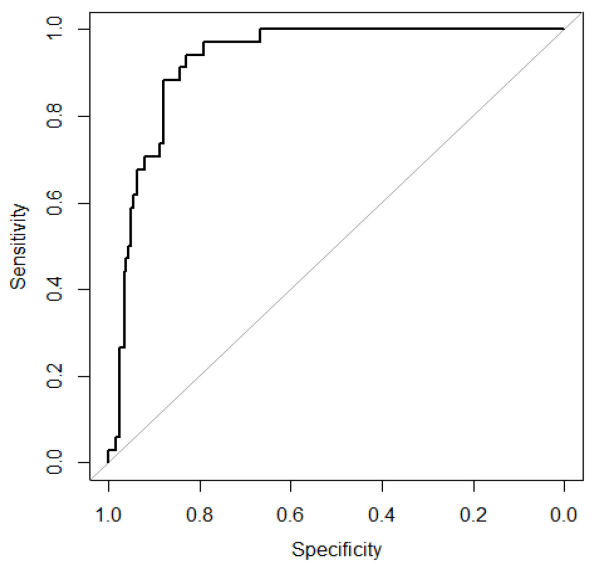

FIGURE 2 | ROC curve of SVM-RW. ROC, receiver–operator characteristic; SVM-RW, support vector machine and random walk.

After verifying the effectiveness of SVM-RW, we randomly selected 34 genes as negative samples and built a final SVM model. SVM-RW predicted 65 novel genes as AMD-related genes.

Verify SVM-RW Results by Summary Data Level-Mendelian Randomization Analysis

If we use g to denote a genetic variant (such as a SNP), x as the expression level of a gene, and y as the trait, then the two-step least-squares (2SLS) estimate of the effect of x on y from an MR analysis can be denoted as:

$$\hat{E}_{xy} = \hat{E}_{zy}/\hat{E}_{zx} \quad (1)$$

where $\hat{E}_{zy}$ and $\hat{E}_{zx}$ indicate the least-squares estimates of y and x on z, respectively, and E_{xy} indicates the effect size of x on y free of confounding from non-genetic factors. Then the sampling variance of the 2SLS estimate of E_{xy} can be denoted as:

$$\mathrm{var}\left(\hat{E}_{xy}\right) = \left[\mathrm{var}\,(y)\ (1 - P_{xy}^2\right] / \left[n\ \mathrm{var}(x) P_{zy}^2\right] \quad (2)$$

where n denotes the sample size, P_{xy}^2 indicates the proportion of variance in the explanation of y by x, and P_{zy}^2 is the proportion of variance in the explanation of y by z. Therefore, we use the

TABLE 1 | Comparison results.

Algorithm	AUC	AUPR
SVM-RW	**0.927**	**0.781**
Random forest-RW	0.852	0.645
Naive Bayes-RW	0.711	0.586
Backpropagation-artificial neural network-RW	0.823	0.692
Logistic regression-RW	0.691	0.531

AUC, area under the curve; AUPR, area under the precision-recall curve; RW, random walk.
Bold values highlight the result of SVM-RW.

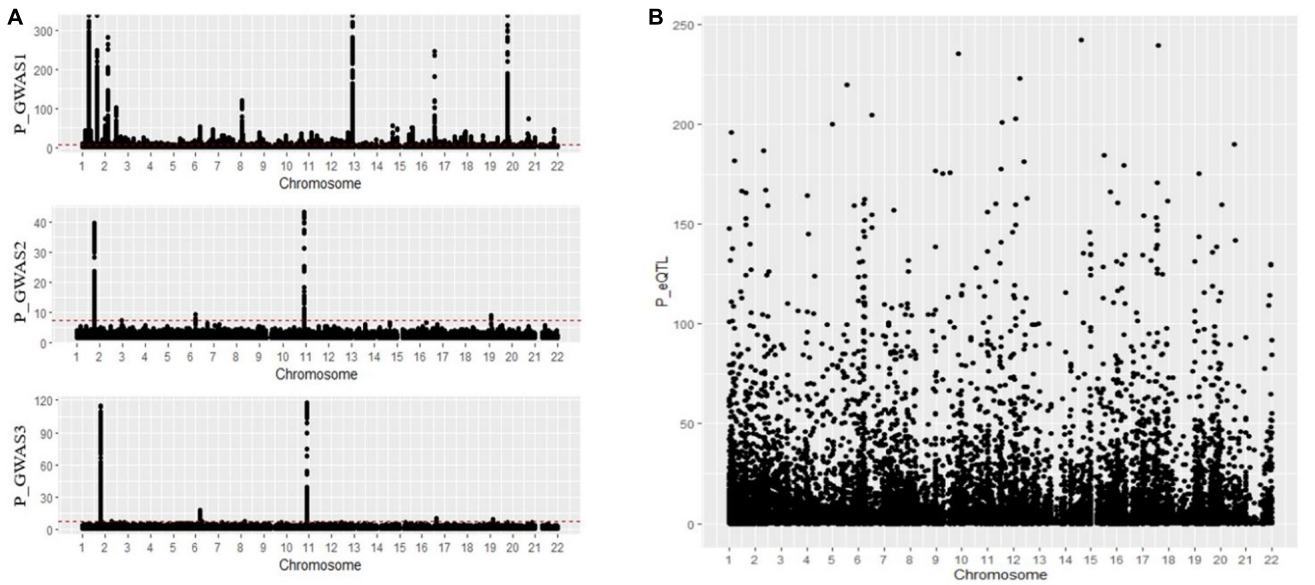

FIGURE 3 | (A,B) Distribution of GWAS summary dataset and blood eQTL dataset. GWAS, genome-wide association studies; eQTL, expression quantitative trait locus.

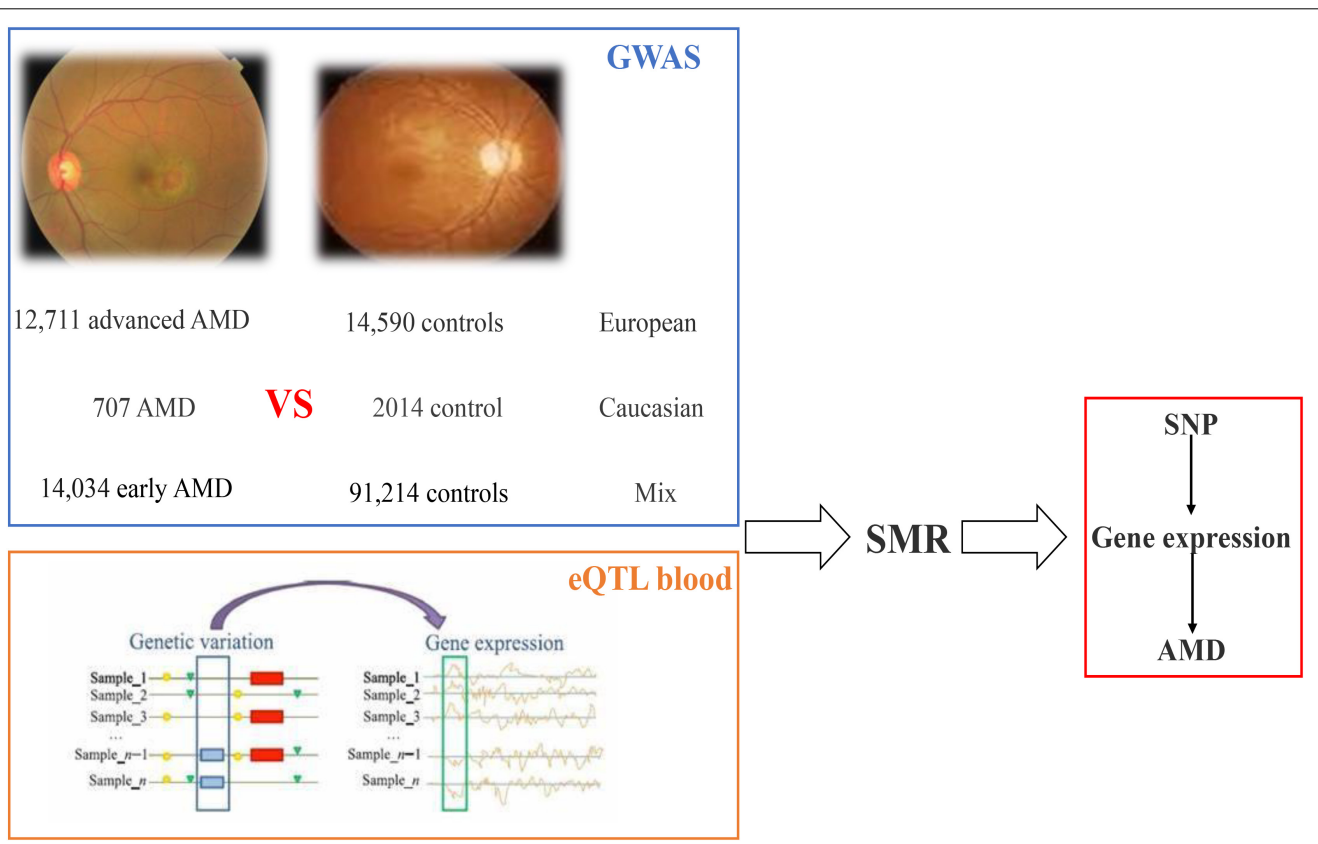

FIGURE 4 | Workflow of SMR on AMD based on GWAS and eQTL datasets. AMD, age-related macular degeneration; GWAS, genome-wide association studies; eQTL, expression quantitative trait locus; SMR, summary data-based Mendelian randomization.

statistic T_{MR} to test the significance of E_{xy}; T_{MR} can be denoted as:

$$T_{MR} = \hat{E}_{xy}^2 / var(\hat{E}_{xy})$$

where $T_{MR}=\chi_1^2$.

Based on the suggestion that the power of detecting E_{xy} can be significantly increased using a two-sample MR analysis (Inoue and Solon, 2010; Pierce and Burgess, 2013), if GWAS and eQTL datasets share the same population, we can use unbiased estimates $\hat{\epsilon}_{zx}$ to replace E_{zx}. We therefore have

$$\hat{E}_{xy} = \hat{E}_{zy} / \hat{\epsilon}_{zx} \quad (3)$$

where $\hat{E}_{zy}$ is the estimate of a SNP effect from a GWAS for a trait, and $\hat{\epsilon}_{zx}$ is the estimate of a SNP effect on the expression level of a gene from an eQTL study. The sampling variance of $\hat{E}_{xy}$ can be approximately computed by the Delta method (Lynch and Walsh, 1998) as:

$$\text{var}\left(\hat{E}_{xy}\right) \approx \frac{E_{zy}^2}{\epsilon_{zx}^2}\left[\frac{\text{var}(\hat{\epsilon}_{zx})}{\epsilon_{zx}^2} + \frac{\text{var}(\hat{E}_{zy})}{E_{zy}^2} - \frac{2\text{cov}(\hat{\epsilon}_{zx}, \hat{E}_{zy})}{\epsilon_{zx} E_{zy}}\right] \quad (4)$$

where $\text{cov}(\hat{\epsilon}_{zx}, \hat{E}_{zy})$ is 0. Based on experience, we can replace them by their estimates in practice, indicated as an approximate χ^2 test statistic of:

$$T_{SMR} = \frac{\hat{E}_{xy}^2}{\text{var}\left(\hat{E}_{zy}\right)} \approx \frac{z_{zy}^2 z_{zx}^2}{z_{zy}^2 + z_{zx}^2} \quad (5)$$

where z_{xy} is the z statistic of the GWAS and z_{zx} is the z statistic of the eQTL study.

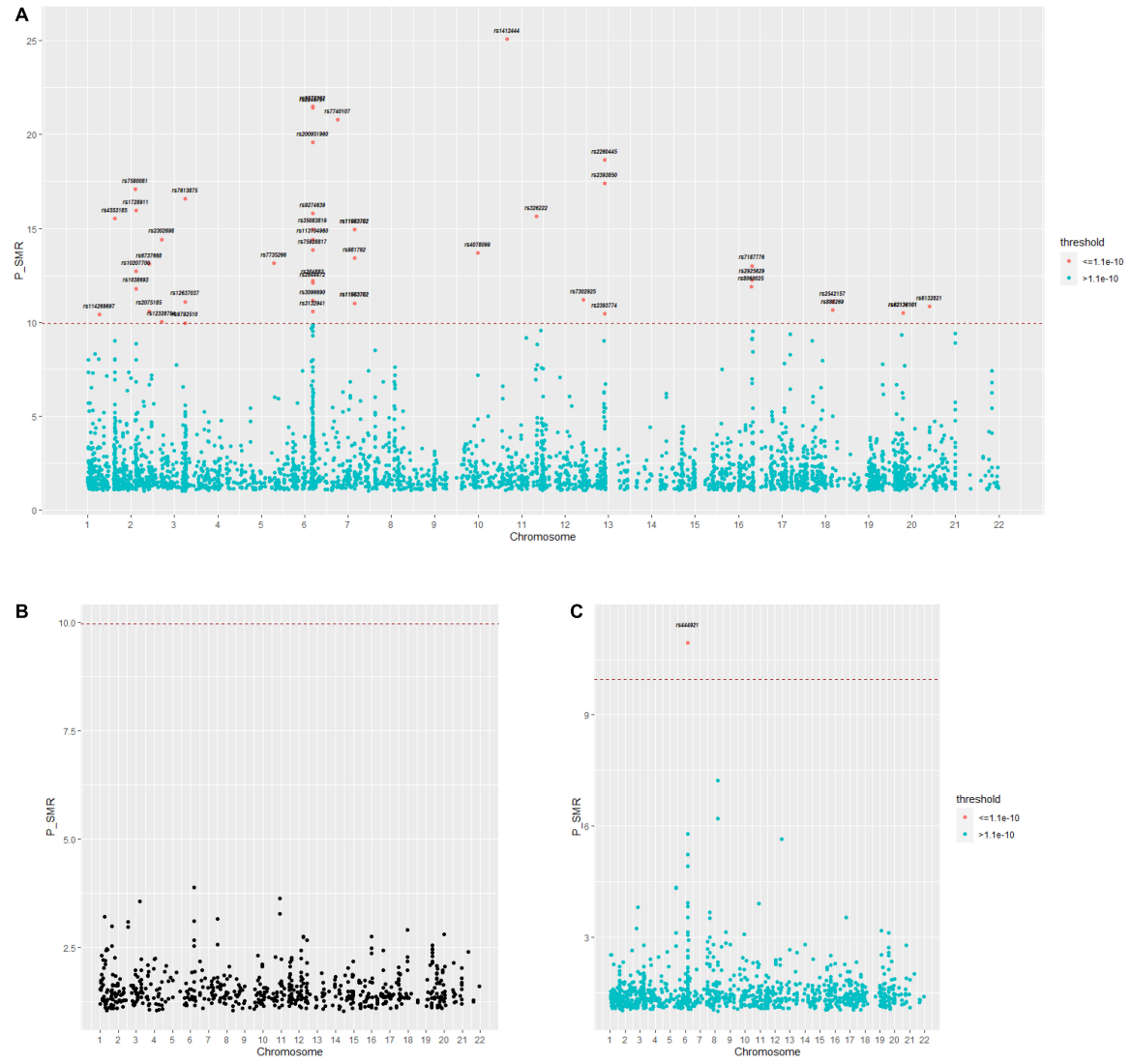

FIGURE 5 | (A–C) Distribution of *p*-value calculated by SMR for three GWAS datasets. GWAS, genome-wide association studies; SMR, summary data-based Mendelian randomization.

In an MR analysis, E_{xy} is interpreted as the effect of a phenotype on the gene expression without considering non-genetic confounders. We first collected GWAS summary data and blood eQTL data from available online studies. We first collected a GWAS summary dataset composed of 12,711 advanced AMD cases and 14,590 controls of European descent from the study by Han et al. (2020); 707 Caucasian AMD patients and 2,014 controls from the study by Yan et al. (2018); and 14,034 cases, 91,214 controls, and 11 sourced from datasets including the International AMD Genomics Consortium, IAMDGC, and UKBB from the study by Winkler et al. (2020). The distribution of the above datasets is shown in **Figures 3A,B**.

Then summary data-based Mendelian randomization (SMR) analysis is implemented on the blood eQTL data and GWAS data; in this paper, we identified 48 SNPs regulating 45 genes (including 41 coding genes and four non-coding genes) resulting in AMD susceptibility. The workflow is shown in **Figure 4**.

For the first GWAS datasets consisting of 12,711 AMD cases and 14,590 controls from European cohorts, in total we found 3,872 SNPs coexist in both GWAS data and eQTL data; 43 of 3,872 SNPs are significant and regulate 44 genes in gene expression level. In the second GWAS dataset, we found 714 SNPs coexist in both GWAS dataset and eQTL dataset, with none significant. In the third GWAS dataset, we found 1,149 SNPs coexist both in GWAS dataset and eQTL dataset, with one significant regulating one gene in gene expression level. The distribution of the *p*-value of SNPs regulating genes tested by SMR is shown in **Figures 5A–C**. A **Supplementary Table 1** indicates the *p*-values of significant SNPs regulating genes tested by SMR; the last line resulted from GWAS dataset 3, and the rest resulted from GWAS dataset 1.

Case Study

Age-related macular degeneration has been described as a partly genetic disease (Heiba et al., 1994; Stone et al., 2004). Recently, a unifying hypothesis is that immune response gene polymorphisms modulate susceptibility to AMD. Human leukocyte antigen (HLA) polymorphisms, encoded within the major histocompatibility complex (MHC), are the most polymorphic within the human genome. In AMD, researchers detected intense HLA-DR immunoreactivity in not only soft but also hard drusen (Mullins et al., 2000). In the study of Goverdhan et al. (2005), considering the effect of smoking, age, and body mass index (BMI), HLA alleles B*4001, DRB1*1301, and Cw*0701 were found to be related to AMD, which is consistent with our results displayed in **Table 1**.

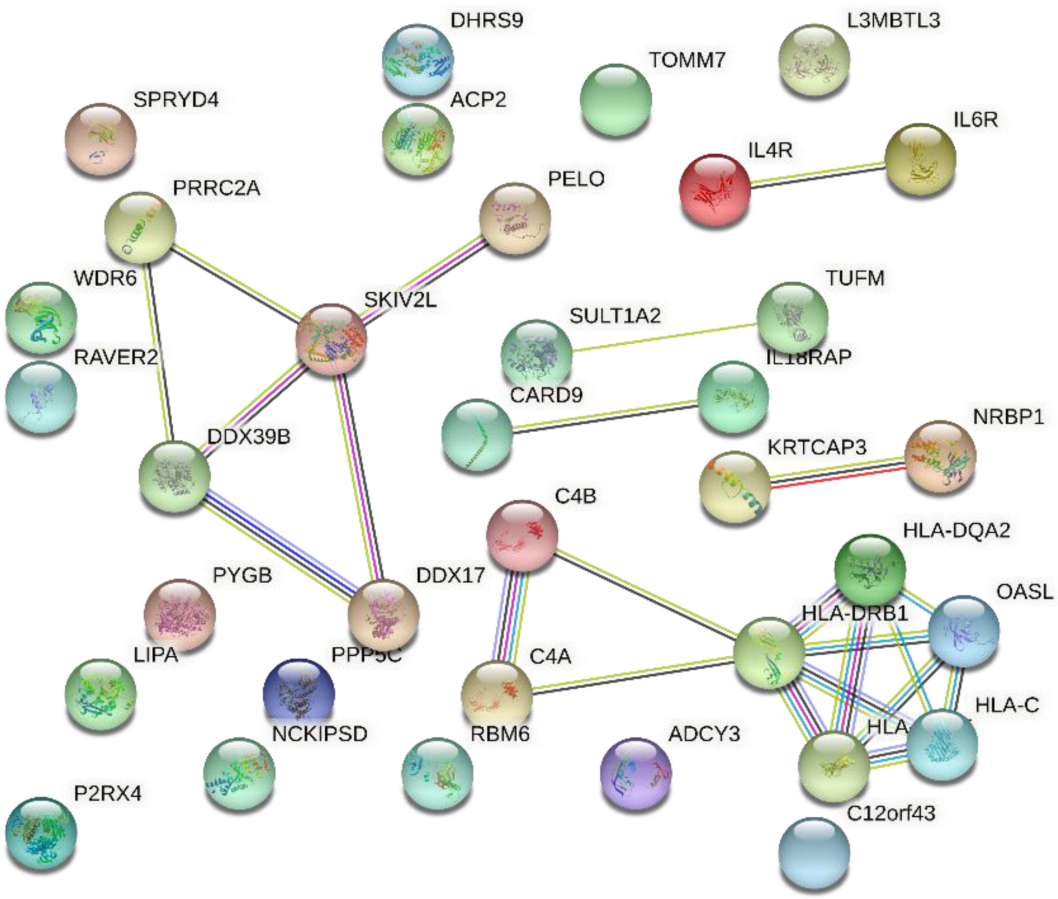

FIGURE 6 | Gene interaction network obtained from 45 genes.

In a study by Gu et al. (2013), they researched P2RX7 and P2RX4 genes in 744 AMD patients and 557 Caucasian controls and reached a conclusion that a rare functional haplotype of the P2RX4 leads to loss of innate phagocytosis and confers increased risk of AMD. P2RX7 and P2RX4 damage the normal scavenger function of macrophages and microglia through interaction, making individuals susceptible to AMD.

Gene Interaction Network Based on AMD

Figure 6 shows the gene interaction network produced from the results of SMR on AMD. Based on the interaction network, the HLA class intensively interacted and is significantly associated with AMD.

The cluster consisting of DDX39B (aka BAT1), PRRC2A (aka BAT2), and SKIV2L are genes found in the class III region of the MHC (MHC Class III). These genes encode RNA-binding proteins with clear roles in post-transcriptional gene regulation and RNA surveillance. They are likely to have important functions in immunity and are associated with autoimmune diseases (Schott and Garcia-Blanco, 2020). Early work by immunologists have shown that DDX39B promoted gene expression of anti-inflammatory pathways (Allcock et al., 2001). Therefore, understanding the genes interactions may help speculate on the proposed AMD mechanisms and immunotherapy.

CONCLUSION

We applied the SMR method on AMD to test the gene–AMD associations based on GWAS summary data and blood eQTL data. From a total of 27,452 AMD cases and 107,818 controls, we obtained 44 SNPs regulating 45 genes significantly associated with AMD. Among the results, HLA class genes have been proved to be associated with immunologically mediated diseases because of the critical role of HLA in mediating the immune response, and genes from MHC Class III are also associated with autoimmune diseases. These genes may play important roles in causing AMD susceptibility and need to be further verified with experiments. Since AMD has been considered as a genetic disease, from this perspective, it is helpful in understanding the disease from gene-expression level to speculate about the AMD mechanisms and pathology and propose future treatment options for AMD.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YD, NK, and JZ participated in its design, analyzed the data, and wrote the manuscript. All authors read and approved the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.726599/full#supplementary-material

REFERENCES

Allcock, R. J., Williams, J. H., and Price, P. (2001). The central MHC gene, BAT1, may encode a protein that down-regulates cytokine production. *Genes Cells* 6, 487–494. doi: 10.1046/j.1365-2443.2001.00435.x

Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907. doi: 10.1056/nejmoa1502214

Davey Smith, G., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* 32, 1–22. doi: 10.1093/ije/dyg070

Feher, J., Kovacs, I., Artico, M., Cavallotti, C., Papale, A., and Gabrieli, C. B. (2006). Mitochondrial alterations of retinal pigment epithelium in age-related macular degeneration. *Neurobiol. Aging* 27, 983–993. doi: 10.1016/j.neurobiolaging.2005.05.012

Fernández-Vega, B., García, M., Olivares, L., Álvarez, L., González-Fernández, A., Artime, E., et al. (2020). The association study of lipid metabolism gene polymorphisms with AMD identifies a protective role for APOE-E2 allele in the wet form in a Northern Spanish population. *Acta Ophthalmol.* 98, e282–e291. doi: 10.1111/aos.14280

Fritsche, L. G., Fariss, R. N., Stambolian, D., Abecasis, G. R., Curcio, C. A., and Swaroop, A. (2014). Age-related macular degeneration: genetics and biology coming together. *Annu. Rev. Genomics Hum. Genet.* 15, 151–171. doi: 10.1146/annurev-genom-090413-025610

Fritsche, L. G., Igl, W., Bailey, J. N. C., Grassmann, F., Sengupta, S., Bragg-Gresham, J. L., et al. (2016). A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* 48, 134–143. doi: 10.1038/ng.3448

Goverdhan, S. V., Howell, M. W., Mullins, R. F., Osmond, C., Hodgkins, P. R., Self, J., et al. (2005). Association of HLA class I and class II polymorphisms with age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* 46, 1726–1734. doi: 10.1167/iovs.04-0928

Gu, B. J., Baird, P. N., Vessey, K. A., Skarratt, K. K., Fletcher, E. L., Fuller, S. J., et al. (2013). A rare functional haplotype of the P2RX4 and P2RX7 genes leads to loss of innate phagocytosis and confers increased risk of age-related macular degeneration. *FASEB J.* 27, 1479–1487. doi: 10.1096/fj.12-215368

Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., et al. (2005). Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308, 419–421. doi: 10.1126/science.1110359

Hammond, C. J., Webster, A. R., Snieder, H., Bird, A. C., Gilbert, C. E., and Spector, T. D. (2002). Genetic influence on early age-related maculopathy: a twin study. *Ophthalmology* 109, 730–736. doi: 10.1016/s0161-6420(01)01049-1

Han, X., Ong, J.-S., An, J., Hewitt, A. W., Gharahkhani, P., and MacGregor, S. (2020). Using mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration. *Eur. J. Epidemiol.* 35, 139–146. doi: 10.1007/s10654-019-00598-z

Heiba, I. M., Elston, R. C., Klein, B. E., and Klein, R. (1994). Sibling correlations and segregation analysis of age-related maculopathy: the beaver dam eye study. *Genet. Epidemiol.* 11, 51–67. doi: 10.1002/gepi.1370110106

Inoue, A., and Solon, G. (2010). Two-sample instrumental variables estimators. *Rev. Econ. Stat.* 92, 557–561. doi: 10.1162/rest_a_00011

Johnson, L. V., Leitner, W. P., Staples, M. K., and Anderson, D. H. (2001). Complement activation and inflammatory processes in drusen formation and age related macular degeneration. *Exp. Eye Res.* 73, 887–896. doi: 10.1006/exer.2001.1094

Katan, M. B. (2004). Apolipoprotein E isoforms, serum cholesterol, and cancer. *Int. J. Epidemiol.* 33:9. doi: 10.1093/ije/dyh312

Landowski, M., Kelly, U., Klingeborn, M., Groelle, M., Ding, J.-D., Grigsby, D., et al. (2019). Human complement factor H Y402H polymorphism causes an age-related macular degeneration phenotype and lipoprotein dysregulation in mice. *Proc. Natl. Acad. Sci. U.S.A.* 116, 3703–3711. doi: 10.1073/pnas.1814014116

Lim, L. S., Mitchell, P., Seddon, J. M., Holz, F. G., and Wong, T. Y. (2012). Age-related macular degeneration. *Lancet* 379, 1728–1738. doi: 10.1016/S0140-6736(12)60282-7

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates Inc.

Mullins, R. F., Russell, S. R., Anderson, D. H., and Hageman, G. S. (2000). Drusen associated with aging and age-related macular degeneration contain proteins common to extracellular deposits associated with atherosclerosis, elastosis, amyloidosis, and dense deposit disease. *FASEB J.* 14, 835–846. doi: 10.1096/fasebj.14.7.835

Pang, C., Baum, L., Chan, W., Lau, T., Poon, P., and Lam, D. (2000). The apolipoprotein E ε4 allele is unlikely to be a major risk factor of age-related macular degeneration in Chinese. *Ophthalmologica* 214, 289–291. doi: 10.1159/000027506

Pierce, B. L., and Burgess, S. (2013). Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* 178, 1177–1184. doi: 10.1093/aje/kwt084

Ratnapriya, R., Sosina, O. A., Starostik, M. R., Kwicklis, M., Kapphahn, R. J., Fritsche, L. G., et al. (2019). Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* 51, 606–610. doi: 10.1038/s41588-019-0351-9

Schott, G., and Garcia-Blanco, M. A. (2020). MHC class III RNA binding proteins and immunity. *RNA Biol.* 18, 640–646. doi: 10.1080/15476286.2020.1860388

Seddon, J. M., George, S., Rosner, B., and Rifai, N. (2005). Progression of age-related macular degeneration: prospective assessment of C-reactive protein, interleukin 6, and other cardiovascular biomarkers. *Arch. Ophthalmol.* 123, 774–782. doi: 10.1001/archopht.123.6.774

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371–375. doi: 10.1038/nature13138

Stone, E. M., Braun, T. A., Russell, S. R., Kuehn, M. H., Lotery, A. J., Moore, P. A., et al. (2004). Missense variations in the fibulin 5 gene and age-related macular degeneration. *N. Engl. J. Med.* 351, 346–353. doi: 10.1056/nejmoa040833

Winkler, T. W., Grassmann, F., Brandl, C., Kiel, C., Günther, F., Strunz, T., et al. (2020). Genome-wide association meta-analysis for early age-related macular degeneration highlights novel loci and insights for advanced disease. *BMC Med. Genomics* 13:120. doi: 10.1186/s12920-020-00760-7

Yan, Q., Ding, Y., Liu, Y., Sun, T., Fritsche, L. G., Clemons, T., et al. (2018). Genome-wide analysis of disease progression in age-related macular degeneration. *Hum. Mol. Genet.* 27, 929–940. doi: 10.1093/hmg/ddy002

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 9:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Valsdottir, L. R., Zang, T., and Peng, J. (2021a). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2019). Identifying alzheimer's disease-related proteins by LRRGD. *BMC Bioinformatics* 20:570. doi: 10.1186/s12859-019-3124-7

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci. U.S.A.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2020d). Identifying protein biomarkers in blood for alzheimer's disease. *Front. Cell Dev. Biol.* 8:472. doi: 10.3389/fcell.2020.00472

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 19 August 2021
doi: 10.3389/fgene.2021.731515

PanSVR: Pan-Genome Augmented Short Read Realignment for Sensitive Detection of Structural Variations

Gaoyang Li[1†], Tao Jiang[1†], Junyi Li[1,2] and Yadong Wang[1]*

[1] Center for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China,
[2] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China

The comprehensive discovery of structure variations (SVs) is fundamental to many genomics studies and high-throughput sequencing has become a common approach to this task. However, due the limited length, it is still non-trivial to state-of-the-art tools to accurately align short reads and produce high-quality SV callsets. Pan-genome provides a novel and promising framework to short read-based SV calling since it enables to comprehensively integrate known variants to reduce the incompleteness and bias of single reference to breakthrough the bottlenecks of short read alignments and provide new evidences to the detection of SVs. However, it is still an open problem to develop effective computational approaches to fully take the advantage of pan-genomes. Herein, we propose Pan-genome augmented Structure Variation calling tool with read Re-alignment (PanSVR), a novel pan-genome-based SV calling approach. PanSVR uses several tailored methods to implement precise re-alignment for SV-spanning reads against well-organized pan-genome reference with plenty of known SVs. PanSVR enables to greatly improve the quality of short read alignments and produce clear and homogenous SV signatures which facilitate SV calling. Benchmark results on real sequencing data suggest that PanSVR is able to largely improve the sensitivity of SV calling than that of state-of-the-art SV callers, especially for the SVs from repeat-rich regions and/or novel insertions which are difficult to existing tools.

Keywords: structure variation calling, pan-genome, read re-alignment, high-throughput sequencing data, repeat-rich region variation

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Bo Jin,
Dalian University of Technology, China

****Correspondence:***
Yadong Wang
ydwang@hit.edu.cn

† These authors have contributed equally to this work

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

Received: *27 June 2021*
Accepted: *26 July 2021*
Published: *19 August 2021*

Citation:
Li G, Jiang T, Li J and Wang Y (2021) PanSVR: Pan-Genome Augmented Short Read Realignment for Sensitive Detection of Structural Variations. Front. Genet. 12:731515. doi: 10.3389/fgene.2021.731515

INTRODUCTION

Structural variants (SVs) are the genomic variations usually defined as genome rearrangement longer than 50 base pairs (bps), which alter a large number of bases in human genomes, although they are fewer than that of single nucleotide variants (SNVs) and short indels. Previous studies have demonstrated that there are many associations between SVs and human phenotypes and diseases (Weischenfeldt et al., 2013; Sudmant et al., 2015; Chiang et al., 2017), thus the comprehensive discovery of SVs in human genomes is fundamental to many genomics studies.

High throughput sequencing (HTS) technologies are rapidly developing and ubiquitously used in human genome re-sequencing projects. Especially, the short reads produced by mainstream

platforms like Illumina sequencers play important roles to the detection of various types of genomic variations including SNVs, indels and SVs (Collins et al., 2019). Due to the high sequencing quality, short reads are feasible to call SNVs and indels and they have demonstrated their ability in many large-scale genomic studies to build the variation maps of various populations (Durbin et al., 2010; The 1000 Genomes Project Consortium, 2012; The UK 10K Consortium, 2015; Cong et al., 2021). However, due to the limited read length, short read had lower ability in SV calling theoretically and practically, comparing to that of the data produced by long reads sequencing platforms such as PacBio or ONT sequencers (Ebert et al., 2020; Beyter et al., 2021). For example, a previous study (Ebert et al., 2021) indicated that, on average 9,320 SVs per sample were called with short reads by three SV calling pipelines, however, this is still less than half of the number of SVs called by long reads. Many of SV calling tools designed for TGS long reads [for example sniffles (De Coster et al., 2019), cuteSV (Jiang et al., 2020), and svim (Heller and Vingron, 2019, 2020)], have the ability to call over 20,000 SVs per individual." Therefore, it is important to develop novel approaches to improve the ability of SV calling with short reads since the sequencing cost of short reads is still much lower.

Many efforts have been made to develop short read-based SV calling approaches. Most of state-of-the-art SV callers [for example delly (Rausch et al., 2012), lumpy (Layer et al., 2014), manta (Chen et al., 2016), and CNVnator (Abyzov et al., 2011)] extract one or multiple kinds of signatures from read alignments, such as discordant read pair, split read, read depth, and local assembly, as evidences to detect SVs. However, all these kinds of signatures could be less effective in practice due to the shortcomings of read aligners which it is still non-trivial to produce the accurate and confident alignments around the breakpoints of SVs (Zook et al., 2020). Most of state-of-the-art read aligners, such as BWA-MEM (Li, 2013), NovoAlign, Bowtie2 (Langmead and Salzberg, 2012), and deBGA (Liu et al., 2016), use seed-and-extension approach. They usually neglect the highly repetitive seeds occurring many times in the reference, however, this could map the reads from repeat-rich regions incorrectly and further affect SV calling. Meanwhile, reads from long novel insertions cannot be correctly aligned in theory, since the abundance of the inserted sequences in reference. Thus, it could extract very few evidences for those insertion events from the alignment results.

With the increasing numbers of sequences samples and known genomic variations (Chaisson et al., 2019), pan-genome-based methods are promising to break through the bottlenecks to the alignment of short reads and provide new opportunities to solve the problems in SV calling. Pan-genome is the ensemble of all the genomes from a species (Sherman and Salzberg, 2020), and in practice it is usually composed by the genomes of multiple samples of the same species or a reference genome plus a set of genomic variations of a population. It has advantages to use a pan-genomes as reference instead of a single genome in read alignment since pan-genome enables to integrate much more reference information to help the alignment of SV-spanning reads. For example, with the integration of known SVs, pan-genome has less bias during the seeding process, so that aligners can locate reads to SV regions with more confidence. Moreover, the sequences of integrated SVs also help the aligners to implement full-length read alignments with high scores instead of the chimeric alignments with plenty of clippings, split alignments and discordant pairs under the circumstance of a single reference. Further, the alignments between reads and integrated SVs can also be used as the evidences of SVs in donor genomes.

However, it is still an open problem to well-organize pan-genome and take its advantage to implement effective and efficient read alignment and SV calling. Efforts have been made to the construction and organization of pan-genome (Sirén et al., 2011, 2020a; Paten et al., 2018; Rakocevic et al., 2019). Moreover, several read alignment and genotyping approaches have been proposed. VG (Garrison et al., 2018; Hickey et al., 2020), giraffe (Sirén et al., 2020b), minigraph (Li et al., 2020) are designed for aligning short reads and GraphAligner (Rautiainen and Marschall, 2020) is designed for aligning long reads. They show higher ability to read alignment and genotyping comparing to the traditional pipelines using single reference. However, most of them are not tailored to SV calling. Especially, these approaches still do not fully consider the divergences between known SVs and the SVs in donor genome, so that they could still have lowered ability to handle newly sequenced samples. Thus, novel computational approaches are still on demand. Moreover, the extraction and analysis of SV signatures is largely different between traditional and pan-genome-based approaches, and they could also be complementary to each other. However, it is also another open problem to integrate various approaches to achieve highest yields in SV calling tasks.

Herein, we propose a novel approach, i.e., Pan-genome augmented Structure Variation calling tool with read Re-alignment (PanSVR). PanSVR focuses to well-handle the potential SV-spanning reads under pan-genome framework to implement more sensitive SV calling. Mainly, it collects known SV information to build pan-genome SV reference and use it as anchors to precisely re-align chimeric reads and find the evidences of SVs with the improved alignments of the reads against pan-genome. Benchmark results on real sequencing data suggest that PanSVR enable to largely improve the sensitivity of SV calling than that of state-of-the-art SV callers, especially for the SVs from repeat-rich regions and/or novel insertions which are difficult to existing tools.

MATERIALS AND METHODS

Overview of PanSVR Approach

The motivation of PanSVR is to take the advantages of known SVs as anchors to improve the sensitivity and accuracy of the alignment of SV-spanning reads to breakthrough the bottleneck of commonly used short read aligners. Moreover, with the improved read alignments, more homogeneous SV signatures can be captured and higher numbers of supporting reads can be found to facilitate the detection of SVs.

Pan-genome augmented structure variation calling tool with read re-alignment uses several tailored methods to implement this approach. Mainly, it is composed by two parts. Firstly,

PanSVR integrates known SVs into commonly used reference genome to build an augmented pan-genome SV reference. The SV reference consists of the sequences around SV sites including the sequences of novel insertions which do not exist in the original reference. This reference is used as anchors to provide additional information for read aligners to improve the reads having clippings, split alignments or discordantly placed which are potentially SV-spanning reads. Secondly, PanSVR collects potential SV-spanning reads and employs short read aligner to re-align those reads against the SV reference. The newly supplied alignments have fewer large divergences such as clippings and split-alignments but more homogenous and confident alignments with the anchors, i.e., the sequences around SV sites. Thus, more homogeneous SV evidences can be collected by PanSVR to further use them to infer accurate SV events. Mainly, PanSVR approach have three main steps as following (**Figure 1**).

(1) Given a set of known SV events (in VCF format), PanSVR converts each of them as an anchor sequence. The generated anchor sequences are then concatenated to build the SV reference and further being indexed by a de Bruijn graph-based genome indexing (RdBG-index) approach (Liu et al., 2016).
(2) Given a set of aligned sequencing reads (in BAM/CRAM format), PanSVR extract the reads having SV signatures (such as clippings and split alignments) and re-align them against the SV reference with the help of RdBG-index and a tailored realignment method. The results are filtered based on the new and original alignments of the same reads and PanSVR clusters them based on their mapping coordinates.
(3) PanSVR separately assemble the reads for all the clusters to generate consensus sequences. Each of the generated sequence is precisely aligned to local region around SV sites in the original reference. The alignment results are used as evidences to infer SVs.

The Construction of SV Reference

Initially, an SV related pan-genome reference ("SV reference") is built from known SVs. Using a reference and a set of SVs records in VCF format as inputs, PanSVR extracts the sequences around the breakpoints of known SVs and stores them in a FASTA format file. It is also worth noting that the current version of PanSVR accepts only one VCF file to build SV reference. However, SV merging tools like SURVIVOR (Jeffares et al., 2017) are feasible to merge multiple SV sets before the construction of SV reference. By default, the sequences of 250 bp flanking SV breakpoints are extracted to construct SV reference as they are long enough to align the short reads produced by mainstream platforms. In details, PanSVR constructs SV reference by the following methods:

(1) For each of the deletions, genomic sequences upstream the first breakpoint and downstream the second breakpoint are directly concatenated together to make the SV anchor sequence;
(2) For each of the insertions and duplications, the inserted (or duplicated) sequences recorded in the ALT field of VCF file are extracted, and the SV anchor sequence is produced by concatenating the local reference sequence upstream the breakpoint, the inserted sequences and the local reference sequence downstream the breakpoint.

Structure variation reference is generated by concatenating all the generated SV anchor sequences. Further, PanSVR employs a de Bruijn graph-based indexing approach to index SV reference (the default value of k-mer is 22 bp) for the realignment of potential SV-spanning reads.

The Realignment and Clustering of Potential SV-Spanning Reads

Pan-genome augmented structure variation calling tool with read re-alignment recognizes the reads potentially spanning SV sites according to their alignments against original reference, and realigns them against the SV reference. Especially, the reads are handled by two steps, i.e., single end read mapping and mate pairing. Further, the realigned reads are clustered by their coordinates and SV signals for SV inference. The method is implemented in four sub-steps as following.

Chimeric Reads Extraction

Reads with chimeric alignments are initially extracted from original SAM/BAM/CRAM files and stored as FASTQ format. Pair-end reads are re-paired by their names if the input file is sorted BAM/CRAM file. In details, PanSVR rejects the read-pairs being perfectly aligned to the reference, i.e., no more than one mismatch for any end in a read-pair and other reads are extracted. This is a restrict condition since SNPs and indel are also useful for SV detection if the reads are mapped to highly repetitive regions, such as VNTRs or STRs. The alignment information related to SV calling is extracted, including alignment position, alignment score, CIGAR, MAPQ, and ISIZE if available. The information is further recorded in the comment field of the converted FASTQ file.

Single-End Read Realignment

The extracted reads are re-aligned to SV reference using a seeding-chaining-and-extension approach (**Figure 2**). To reduce computational cost, PanSVR selects unique k-mers in a read as seeds (default value of k: 20), unlike traditional seeding methods. This design is to handle repetitive k-mers within STR or VNTR regions which could appear hundreds and thousands of times in reference and consume plenty of time during the seeding and chaining process. Other than unique seeds, the seeds from repetitive regions are also employed, if they are placed at either end of the reads (**Figure 2A**).

A two-phase chaining method is used for chaining the seeds. In the first phase, seeds are chained within the unitigs of RdBG-index of SV reference to generate longer match blocks from the shorter seeds. The match blocks are then mapped back to original reference as long seeds. If a match block is highly repetitive, i.e., it can be mapped to over 1000 genomic positions, 1000 positions are randomly selected for further processing. In the second

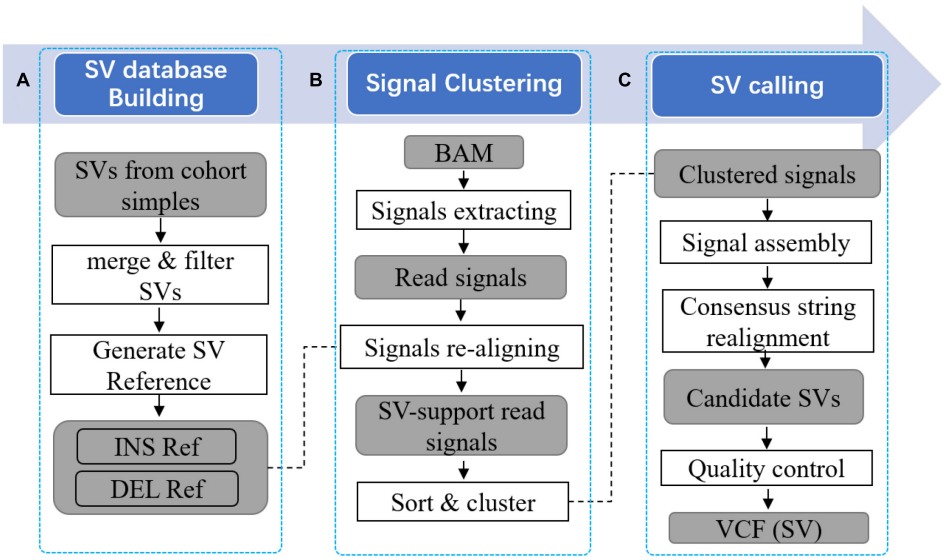

FIGURE 1 | Overview of PanSVR SV calling process. Three main steps of PanSVR SV calling process. **(A)** In the first step, SV reference is built from known SVs; **(B)** In the second step, read signals are extracted from original BAM files and mapped to the SV reference; **(C)** Finally, read signals clustered around SV breakpoints are assembled and SV results generated from consensus strings.

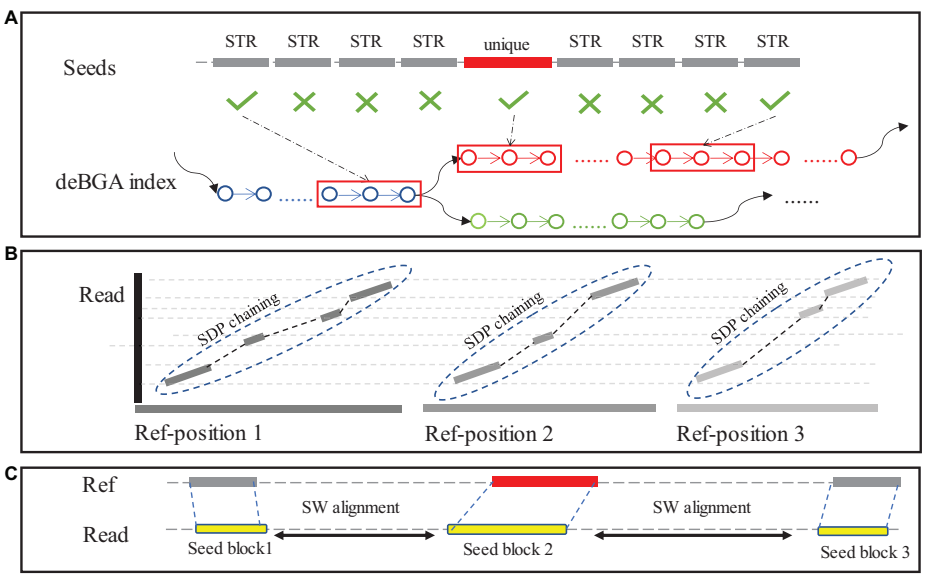

FIGURE 2 | The seeding-chaining-and-extension in the alignment step. **(A)** The seeds generated in "unique region" of reads are located in reference using deBGA index. **(B)** Seeds within UNITIG of deBGA index will be greedy chained to longer blocks, then those blocks will be mapped to reference and chained again using SDP. **(C)** Sequence between chained blocks will be aligned using NW algorithm.

phase, the long seeds are chained by using a sparse dynamic programming (SDP)-based method with following functions:

$$f\left(LS_p\right) = \max\left\{\max_{p>q\geq 1}\left\{f\left(LS_q\right) + \mathrm{L}\left(LS_{pq}\right) - \theta\left(p,q\right)\right\}, L\left(LS_p\right)\right\} \quad (1)$$

$$\theta\left(p,q\right) = 0.125 \times \left(\left(LS_p^R - LS_q^R\right) - \left(LS_p^r - LS_q^r\right)\right) + 3 \quad (2)$$

where LS_p and LS_q are the p-th and q-th long seeds (sorted by coordinates in reference); $\mathrm{L}\left(LS_p\right)$ is the length of long seed p and $\mathrm{L}\left(LS_{pq}\right)$ is the length of LS_p (only consider the part that not overlap with LS_q). LS_p^r is the position of LS_p on the reference, and LS_p^r is the position of LS_p on the read; $f\left(\mathrm{LS}_p\right)$ is the scoring function for the LS_p, and $\theta\left(p,q\right)$ is the penalty score for the two chained long seeds LS_p and LS_q.

In extension step, a traditional Smith-Waterman alignment is implemented for the top 12 seed chains with highest scores using ksw2 library (Li, 2018; Suzuki and Kasahara, 2018). The results are recorded in a list as single end alignment.

Mate Read Pairing

For a read pair, PanSVR uses the single end alignments for the both two ends of a read pair and their original alignments to compose a concordant pair-end alignment and compute the score of the refined alignment. Since the coordinates in the SV reference are not always same to the coordinates in original reference, the two coordinates of the original alignments could be divided to two different values that one of them is not changed and the other is adjusted by the length of the corresponding SV. Both two the values can be used as its coordinates. The score of a read pair is defined as the sum of alignment scores for both ends. When the two ends in a pairing condition have right directions and the ISIZE is within 1.5 times standard deviations of mean ISIZE, an additional score is added. The final score of a read pair is calculated using the following functions:

$$S\left(RP_i\right) = \max_{\substack{N_i > p \geq 1 \\ M_i > q \geq 1}} \left\{ s\left(R1_p\right) + \mathrm{s}\left(\mathrm{R2}_q\right) + \theta\left(p, q\right) \right\} \quad (3)$$

$$\theta\left(p, q\right) = \begin{cases} \mathrm{K} \ \textit{if} \ R1_p \ \textit{paied with} \ \mathrm{R2}_q \ \textit{properly} \\ 0 \qquad\qquad\qquad \textit{otherwise} \end{cases} \quad (4)$$

where $S\left(RP_i\right)$ is the final score of the i-th read pair; N_i is the number of single end alignment results for the first read in read pair and M_i is the number for results for the second read; $s\left(R1_p\right)$ is the score of the p-th single end alignment result for first read in a read pair, and $s\left(\mathrm{R2}_q\right)$ is the score of the q-th single end alignment result for the second read in that read pair; $\theta\left(p, q\right)$ is the additional score be added when the two results pairing properly.

All pairing conditions are sorted by the scores and the one with the highest score is output as paired alignment result. It is also worth noting that the alignment result is discarded and the corresponding read-pair is recorded as unmapped if its alignment result (or one of the multiple results with equal scores) is not made by PanSVR but the original aligner. All the remaining alignment results are stored in SAM format. An additional tag that records the ID of SV anchor sequence is added in the SAM optional field, and it will be used to cluster the read in the following steps.

Read Clustering

All the SAM records of the improved alignments are sort by their positions in the SV reference. Since there could be multiple known SVs in highly repetitive regions and some of various known SVs could overlap with each other, the chimeric reads could be mistakenly assigned during read clustering. To address this issue, PanSVR clusters nearby known SVs as a group and only keeps the top two SVs with highest number of supporting reads and the reads assigned to other nearby SVs are re-assigned to them. Herein, the SVs are clustered in a greedy manner, i.e., an SV is added to a cluster if its upstream breakpoint is within 50 bp of the downstream border of the cluster, and the cluster expends until no nearby SV can be added into it. For a cluster, PanSVR separately counts the numbers of the reads being aligned to the SVs and uses these numbers as the scores of the SVs. For the reads not in the top two clusters, each of them is re-assigned to one of the two SVs by a simple k-mer counting method. That is, PanSVR counts the numbers of identical k-mers between a read and the anchor sequences of the two SVs and re-assign the read to the SV with more identical k-mers. If the two SVs have equally high numbers, the read is randomly assigned.

The Assembly of Clustered Read and the Inference of SVs

Pan-genome augmented structure variation calling tool with read re-alignment implement an assembly for each of the clusters to produce the consensus sequence of the reads. The generated sequences are then aligned to the SV reference and PanSVR collects SV evidences from the alignment results. The method is implemented in four sub-steps as following.

Read Preprocessing

Pan-genome augmented structure variation calling tool with read re-alignment does a filtration on the reads before assembly with two rules to reduce false positives. Firstly, a proportion of reads with low scores are filtered out from the SV reference regions having extremely high read coverages. More precisely, PanSVR partitions a given reference region into 64 bp blocks and calculates the read coverages of the blocks. If a block has 1.5 times or higher read depth than average read depth, the reads having low scores in the block are discarded. Secondly, the reads are filtered by mapping quality. That is, for a given cluster, if over 80% of the reads have MAPQ = 0 for their original alignments and the scores of their improved alignments produced by PanSVR are also close to that, the read cluster is considered as an uncertain cluster and being discarded.

Assembly of Clustered Reads

Pan-genome augmented structure variation calling tool with read re-alignment uses a modified version of the assembly module of MANTA (Chen et al., 2016) to implement read assembly for all the clusters. Moreover, if a cluster belongs to a long SV region, i.e., the length of the corresponding SV is over 500 bp, the SV region is partitioned into 500 bp blocks and the assembly is separately implemented for the blocks. When the employed assembler picks up a read to extend the contig, it records at which position the read joins in the assembling contig. This information guides the realignment of the reads to the contig after assembly. Only mismatches are allowed in the realignment of reads to contig. Read coverage information on consensus sequence is calculated based on the realignment results.

Alignment of Consensus Sequence

For a consensus sequence, PanSVR detects some candidate positions in SV reference to implement local alignment at first. These candidate positions are from the mapping positions of the supporting reads in SV reference with some additional filtrations. Firstly, if all the candidate positions are out of range, the consensus sequence is discarded. Secondly, at least one read used in the generation of the consensus sequence should have a realignment score higher than that or its original alignment. After the filtration, a Needleman-Wunsch alignment is implemented for each of the candidate positions. The mismatches and indels

between the consensus sequence and local sequence in SV reference is recorded at each position of SV reference. Moreover, read depths along the consensus sequences are also stored by all the corresponding coordinates in SV reference.

SV Calling and Genotyping

Pan-genome augmented structure variation calling tool with read re-alignment infers SVs from the alignment of consensus sequences. A candidate SV other than novel insertions implied by a consensus sequence is inferred if it has high enough depth at both the two breakpoints. Meanwhile, for novel insertions, the inserted sequence should also have high depth. Moreover, the positions of breakpoints and the inserted sequences are adjusted by the variations to make a more accurate inference. After the adjustment, SVs longer than 50 bp are kept and further genotyped with the coverage information.

RESULTS

Implementation of Benchmark

To assess the ability of PanSVR, we composed an SV reference with a set of high-quality SVs at first. Mainly, the SVs are derived from PacBio CCS datasets of 16 different samples. Thirteen of them are from Human Genome Structural Variation Consortium (HGSVC) database where the datasets are phased assembly of CCS reads. There are two haplotypes of for each of the samples, and we aligned those genomes against human reference genome (version: hs37d5) by minimap2 and input the alignments into SVIM-asm (Heller and Vingron, 2019, 2020) to produce SV callsets. Moreover, we also downloaded three SV callsets of Genome in a Bottle (GIAB) Trio samples HG002, HG003, and HG004. These callsets are produced by GIAB consortium from PacBio CCS datasets using PBSV pipeline. SVs from different samples were merged by the following rule: two SVs were merged if they were of the same type and their breakpoints were within 50 bps. The merge operation was implemented by using SURVIVOR (Jeffares et al., 2017).

We benchmarked PanSVR on three real sequencing datasets produced by Illumina platforms from various samples (i.e., HG00512 and HG002) with various read lengths (i.e., 126, 148, and 250 bp). Refer to **Supplementary Table 2** for more detailed information. Two state-of-the-art short read-based SV callers, i.e., Manta and Delly, were also implemented on the same datasets for comparison. During the benchmark, leave-one-out strategy was applied for PanSVR, i.e., the SVs of the corresponding sample was removed from the known SV sets beforehand so that the constructed SV reference is blind to the benchmarked dataset. The reads were aligned against human reference hs37d5 by BWA-MEM with default settings. Manta and Delly directly detected SVs from the read alignments.

Results on Real Sequencing Datasets

The sensitivity, accuracy and F1-score of the benchmarked SV callers were assessed by using the "merge" and "genComp" commands of SURVIVOR. All the benchmarks were carried out on an Ubuntu Linux server with one AMD 3950X CPU (32 cores) and 256 GB RAM. All the SV callers were run in using 8 CPU threads. Mainly, three issues were observed from the results.

PanSVR Has Good SV Calling Yields

For all the datasets, PanSVR obviously outperformed Manta and Delly for F1-scores on both insertions and deletions (**Figure 3**). We investigated the intermediate results of PanSVR and found that the SV reference greatly helped to improve the alignment of SV-spanning reads. Although the known SV sets cannot cover all the SVs of the testing samples, the anchor sequences of the SV reference enable to rescue many reads which cannot be correctly and/or confidently aligned with the original reference. This feature largely improves the sensitivity of SV calling, especially for large insertions. For all the datasets, the numbers of insertions detected by PanSVR are nearly two times to that of Manta. Moreover, Delly showed a relatively poor ability to detect insertions, i.e., it only called a few hundreds of insertions for each sample and only a few of them were true positive. It is also worth noting that all the callers have relatively good results on deletions since short reads spanning deletions are much easier to be aligned and the SV signatures of short reads around deletion events, such as discordant read pairs and split alignments, are less complicated and more homogeneous.

As for the influence of read length on the SV calling ability, most of time, longer reads do help to achieve better SV calling results. For Delly, the F1 scores increased with the increase of read length and reach best value on the 250 bp dataset, while PanSVR and manta achieved best F1 scores on 148 bp dataset. We investigated the details of the results and found that the large numbers of low-quality bases at the tails of the 250 bp reads affected the local assembly operation of PanSVR.

PanSVR Has Good Ability to Call Long Insertions

It is a still non-trivial task for state-of-the-art short read-based callers to detect long insertions due to two issues. First, when an insertion is longer than the read length, one or two ends of a read pair around the insertion could be unmapped. Second, the length of insertion cannot be easily estimated and assembling all reads around and within an insertion is usually hard. Based on pre-built SV reference, PanSVR enable to detect long insertions with the help of SV anchor sequences that the reads can be effectively realigned to imply plenty of SV signatures. Moreover, PanSVR also has the ability to detect the SNVs and indels within the inserted strings of the sequenced sample from the realignments of the reads, so that the inserted sequences of donor samples can be correctly recovered even if they are divergent to the anchor sequences of SV reference. As showed in **Figure 4**, there are only 48 $>$ 500 bp true positive insertions in the callset of Manta, and the corresponding number for PanSVR is 917. However, we also observed that PanSVR has lowered ability to handle ALU insertions (as show in **Figure 4A** which the length distribution of the SVs detected by PanSVR has no significant peak around 300 bp). This is mainly due to that ALUs are extremely repetitive in human reference genome and the average mapping quality of the reads being aligned to ALU regions are usually close to 0. PanSVR filters out such

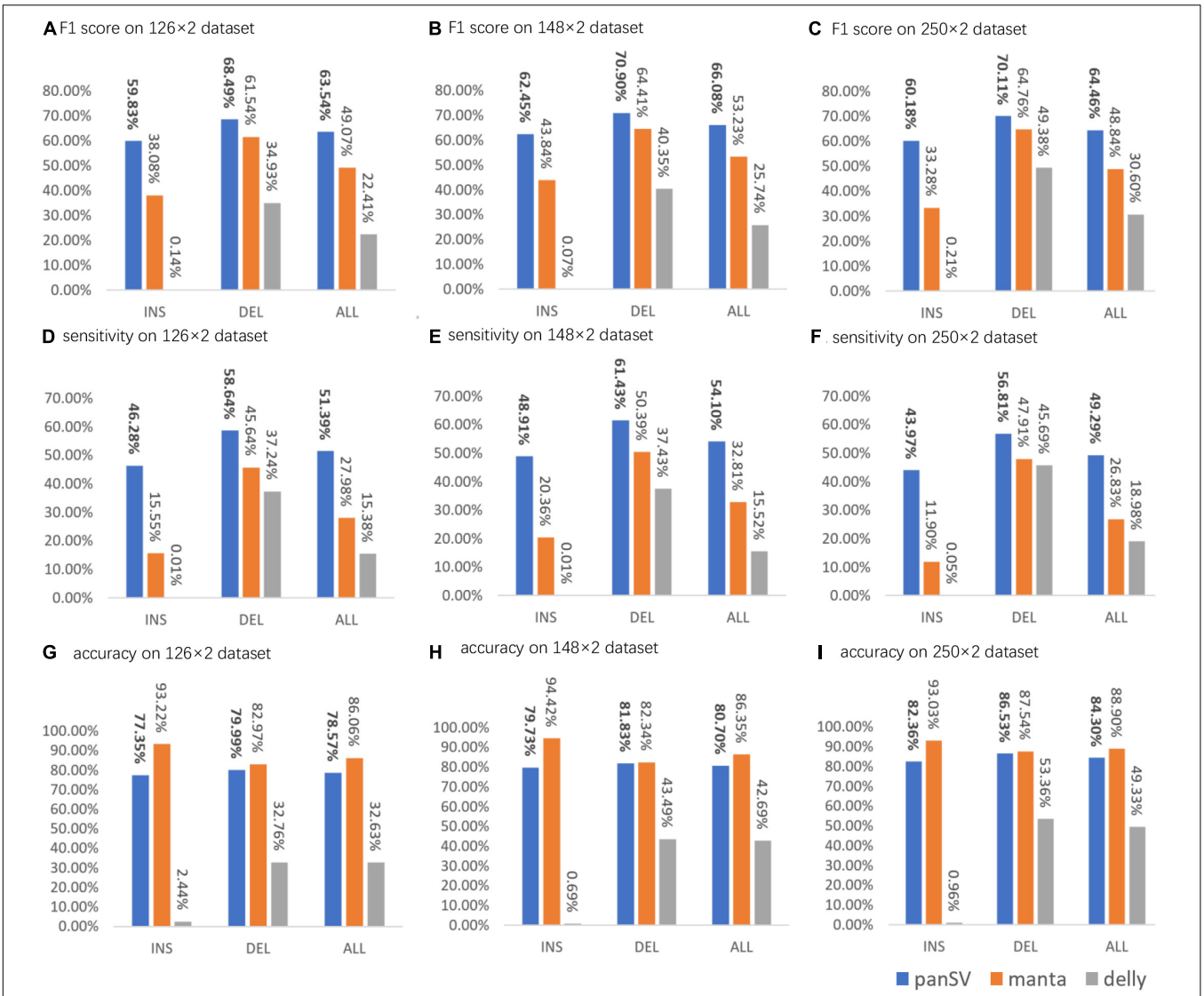

FIGURE 3 | The sensitivity, precision and F1 score of PanSVR, manta and delly on three NGS datasets. **(A,D,G)** F1 score, sensitivity and accuracy of SV calling results of PanSVR, manta and delly applied on Illumina 126 × 2 dataset; **(B,E,H)** The SV calling results of those tools on Illumina 148 × 2 dataset; **(C,F,I)** The SV calling results of those tools on Illumina 250 × 2 dataset.

regions according to the low MAPQ so that ALU insertions could be missed.

The Ability of PanSVR Could Be Complementary to State-of-the-Art SV Callers

Most of existing SV callers use chimeric alignments such as split reads, discordant read pair and large clippings as SV signatures. PanSVR does not rely on those kinds of signatures but use a different approach, so that it could produce higher-quality SV callsets by merging the results of PanSVR and other tools. We merged the results of PanSVR and Manta using SURVIVOR by various parameters. Firstly, we generated the union SV calling set of PanSVR and Manta. The SVs are treated as one when their breakpoints are distanced less than 50 bp. For all the samples and SV types, the merged SV callset achieved higher sensitivities and F1-scores than the callsets separately produced by PanSVR and Manta (**Supplementary Table 3**), although the precisions could decrease. For example, the merged callset of the 148 bp HG002 dataset called 12272 true positive SVs with 77.4% true positive rate, while PanSVR and Manta called 11540 and 6980 true positive SVs, respectively. We also tried to generate an intersection SV set from the results of the two tools. It reached more than 96% true positive rate in all three datasets, however, the F1-score slightly decreased comparing to that of Manta only (**Supplementary Table 4**).

We also assessed the speed and memory footprint of PanSVR. It takes less than 2.7 h to process all steps using 8 threads for a 60x coverage dataset. This is slower than Manta and Delly, but still affordable. This is mainly due to the realignment procedure of the approach which is more computation-intensive than that

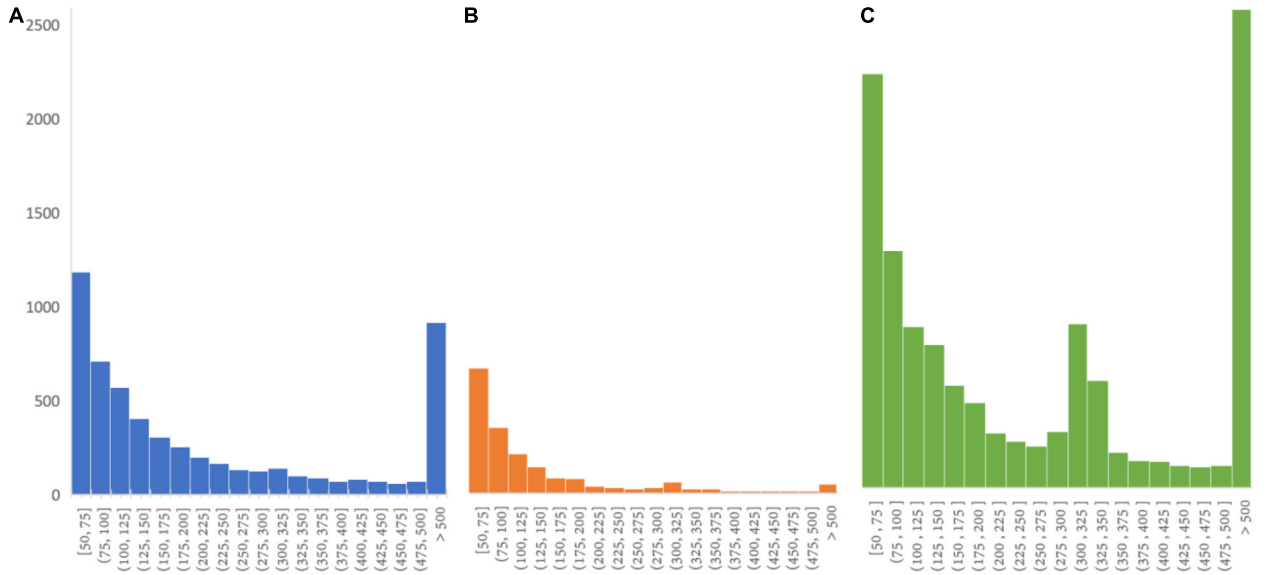

FIGURE 4 | SV length distribution of insertions from HG00512 126 × 2 sample. **(A)** Insertion length distribution of PanSVR true positive set (917 insertions longer than 500 bp); **(B)** Insertion length distribution of manta true positive set (48 insertions longer than 500 bp); **(C)** Insertion length distribution of CCS SV set (2,660 insertions longer than 500 bp).

of directly analyzing the read alignment results like most of short read-based SV callers do. Moreover, the time cost of the assembly of clustered reads is also non-neglectable. However, PanSVR still has good scalability since all the steps can be run in a parallel way. The memory footprint of the PanSVR is about 3.5 GB in the benchmark where the memory is mainly used by the RdBG-index of SV reference in the read realignment step.

DISCUSSION

Previous studies (Hickey et al., 2020; Sirén et al., 2020b) have demonstrated that ability of pan-genomes to help the alignment of short reads and SNP/INDEL calling. In this study, we introduce a pan-genome augmented read realignment and SV calling tool, PanSVR. Results on real NGS datasets demonstrate that it is feasible to use pan-genome based realignment approach to realign short reads to break through the bottleneck of short read alignment and further improve SV calling.

Mainly, we found that two main categories of SVs can be better handled with the pan-genome-based method. Firstly, the SVs in tandem repeat regions can be recused by PanSVR. This is due to that SNPs and INDELs within VNTR or STR can be used to correct short read alignments around those regions. A case is shown in **Supplementary Figure 1** that a 70 bp insertion around chr1:1913259 were successfully called by PanSVR, however, no other tool is able to detect them in the benchmark. We checked alignment results around those regions manually and found that the spanning reads can be fully mapped to that region by BWA-MEM, but with a number of mismatches and indels. The lower quality alignments affect the callers and the SVs are recognized as multiple SNP/indels. However, these reads can be re-aligned with exact matches to the SV reference by PanSVR and evidences can be collected to call the SVs confidently.

Secondly, the results indicated that pan-genome-based method greatly help to improve recall of long insertions which is surprising since additional reference information is added. It is shown that PanSVR has a nearly 20 times higher number of long insertion (>500 bp) calls than that of Manta. This is very complementary to the state-of-the-art SV calling approaches. A case is shown in **Supplementary Figure 2** that a 955 bp insertion at chr2:235423389, which cannot be called by other callers but PanSVR. The read alignments show that there are few split-read and discordant read pair signals around the SV breakpoints, so that the SVs are hard to detect, however, the realignment against the SV reference recused most of SV-spanning reads and provided homogeneous SV signatures.

The results also suggest that it is also helpful to merge the SV callsets by multiple callers to further increase sensitivity. For example, the sensitivity increased by 3.4% for the 148 bp dataset comparing to that of the original callset of PanSVR. This is consistent with previous studies (Chaisson et al., 2019) as multiple tools could be complementary to each other by various kinds of signatures and models. However, it is also worth noting that the simple union of the callsets of various tools could introduce more false positives, so that more advanced approaches for the filtration and prioritization of SV calls are still needed.

There is still a huge gap for the sensitivity of SV calling between short and long sequencing reads, although pan-genome is used. There could be caused by two issues.

Firstly, some of the SVs in donor genomes are individually specific and their breakpoints are far away from known SVs or even not related to them at all. In this situation, the pan-genome-based method cannot provide much help and the detection

of SV only depends on the alignments of the reads against original reference. A case is shown in **Supplementary Figure 3** that an 87 bp insertion in chr1:2213294 is unique for HG002 sample. It was not in the SV reference during the leave-one-out benchmark and PanSVR failed. However, with the many on-going population-scale genomics studies, it is promising to build more comprehensive SV databases. For example, a recent study (Beyter et al., 2021) has built an SV database of Iceland population with 133,886 reliably genotyped SVs and such SV databases could be available for various populations with the ubiquitous application of high-throughput sequencing technologies.

Secondly, the limited length of short reads is still a bottleneck even if under the circumstance of pan-genome. Especially, this could cause lower coverage to correct anchors in SV reference for PanSVR. A case is shown in **Supplementary Figure 4** that there is nearly no read being aligned to a 103 bp insertion in chr1:1855662. The inserted sequence is highly repetitive, i.e., ACCACCCCCCAGCTCACAGCCCACCCCCCCATCTCACCG CCCAGCCCCCCCATCTCACCAGCTGCCCCCTCCCGGGCA CACCGCCCACCCCCCCATCTCACCA. Such repeats can still not be spanned by short reads and the reads are usually mapped to other copies of the sequences with nearly perfect alignments, i.e., exactly matched without mismatch or indel. In this situation, the SV is non-trivial to be solved. Moreover, the results also indicated that PanSVR could make false positives in some cases. We checked the SVs mistakenly called by PanSVR and found that they were mainly in repeat regions. Some consensus sequences were not long enough to across the repeat region, either. Wrong alignment of them might cause wrong SV calling.

Pan-genome-based SV calling approach is promising to the comprehensive discovery of individual genomes, especially for short read datasets. With the supplement of additional SV information, it enables to produce higher-quality alignments and help to provide more evidences to make SV calls with confidence. However, there are still open problems to the use of known SVs, moreover, some of SVs can still not solved with the available SV databases. These are also important future works to us to further improve PanSVR approach. With the higher sensitivity and yield, we believe that PanSVR has the potential to many genomics studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

GL and TJ designed the method. GL implemented the method. GL and JL performed the analysis. All authors wrote the manuscript.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Grant No: 2017YFC0907503) and National Natural Science Foundation of China (Grant No: 32000467).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.731515/full#supplementary-material

REFERENCES

Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110

Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H. P., Bjornsson, E., Jonsson, H., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat. Genet.* 53, 779–786. doi: 10.1038/s41588-021-00865-4

Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1–16.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., et al. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 32, 1220–1222. doi: 10.1093/bioinformatics/btv710

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., et al. (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699.

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Khera, A. V., et al. (2019). An open resource of structural variation for medical and population genetics. *BioRxiv* 578674. doi: 10.1101/578674

Cong, P., Bai, W., Li, J., Li, N., Gai, S., Khederzadeh, S., et al. (2021). Genomic analyses of 10,376 individuals provides comprehensive map of genetic variations, structure and reference haplotypes for Chinese population. *bioRxiv* doi: 10.1101/2021.02.06.430086

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., et al. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res.* 29, 1178–1187. doi: 10.1101/gr.244939.118

Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2020). De novo assembly of 64 haplotype-resolved human genomes of diverse ancestry and integrated analysis of structural variation. *bioRxiv* doi: 10.1126/science.abf7117

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372:eabf7117.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* 36, 875–879. doi: 10.1038/nbt.4227

Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35, 2907–2915. doi: 10.1093/bioinformatics/btz041

Heller, D., and Vingron, M. (2020). SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *bioRxiv* doi: 10.1101/2020.10.27.356907

Hickey, G., Heller, D., Monlong, J., Sibbesen, J. A., Sirén, J., Eizenga, J., et al. (2020). Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.* 21, 1–17.

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8, 1–11.

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., et al. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* 21, 1–24.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie. *Nat. Methods* 9:357. doi: 10.1038/nmeth.1923

Layer, R. M., Chiang, C., Quinlan, A. R., and Hall, I. M. (2014). LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* 15, 1–19. doi: 10.1201/9781420082333.ch1

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Li, H., Feng, X., and Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 21, 1–19.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* doi: 10.6084/M9.FIGSHARE.963153.V1

Liu, B., Guo, H., Brudno, M., and Wang, Y. (2016). deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics* 32, 3224–3232.

Paten, B., Eizenga, J. M., Rosen, Y. M., Novak, A. M., Garrison, E., and Hickey, G. (2018). Superbubbles, ultrabubbles, and cacti. *J. Comput. Biol.* 25, 649–663. doi: 10.1089/cmb.2017.0251

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., et al. (2019). Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* 51, 354–362. doi: 10.1038/s41588-018-0316-4

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333–i339.

Rautiainen, M., and Marschall, T. (2020). GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biol.* 21, 1–28.

Sherman, R. M., and Salzberg, S. L. (2020). Pan-genomics in the human genome era. *Nat. Rev. Genet.* 21, 243–254. doi: 10.1038/s41576-020-0210-7

Sirén, J., Garrison, E., Novak, A. M., Paten, B., and Durbin, R. (2020a). Haplotype-aware graph indexes. *Bioinformatics* 36, 400–407.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., et al. (2020b). Genotyping common, large structural variations in 5,202 genomes using pangenomes, the Giraffe mapper, and the vg toolkit. *Biorxiv* doi: 10.1101/2020.12.04.412486

Sirén, J., Välimäki, N., and Mäkinen, V. (2011). "Indexing finite language representation of population genotypes," in *International Workshop on Algorithms in Bioinformatics*, (Saarbrücken: Springer), 270–281. doi: 10.1007/978-3-642-23038-7_23

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81.

Suzuki, H., and Kasahara, M. (2018). Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinform.* 19:33–47. doi: 10.1186/s12859-018-2014-8

The 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

The UK 10K Consortium (2015). The UK10K project identifies rare variants in health and disease. *Nature* 526:82.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J. O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* 14, 125–138. doi: 10.1038/nrg3373

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., et al. (2020). A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* 38:1357.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

METHODS
published: 24 August 2021
doi: 10.3389/fgene.2021.706952

Inferring Functional Epigenetic Modules by Integrative Analysis of Multiple Heterogeneous Networks

Zengfa Dou[1] *and Xiaoke Ma*[2]*

[1] The 20-th Research Institute, China Electronics Technology Group Corporation, Xi'an, China, [2] School of Computer Science and Technology, Xidian University, Xi'an, China

Gene expression and methylation are critical biological processes for cells, and how to integrate these heterogeneous data has been extensively investigated, which is the foundation for revealing the underlying patterns of cancers. The vast majority of the current algorithms fuse gene methylation and expression into a network, failing to fully explore the relations and heterogeneity of them. To resolve these problems, in this study we define the epigenetic modules as a gene set whose members are co-methylated and co-expressed. To address the heterogeneity of data, we construct gene co-expression and co-methylation networks, respectively. In this case, the epigenetic module is characterized as a common module in multiple networks. Then, a non-negative matrix factorization-based algorithm that jointly clusters the co-expression and co-methylation networks is proposed for discovering the epigenetic modules (called Ep-jNMF). Ep-jNMF is more accurate than the baselines on the artificial data. Moreover, Ep-jNMF identifies more biologically meaningful modules. And the modules can predict the subtypes of cancers. These results indicate that Ep-jNMF is efficient for the integration of expression and methylation data.

Keywords: DNA methylation, network biology, functional epigenetic module, non-negative matrix factorization, heterogeneous network

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Peng Gao,
Children's Hospital of Philadelphia, United States
Pu-Feng Du,
Tianjin University, China

***Correspondence:**
Xiaoke Ma
xkma@xidian.edu.cn

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

Received: *08 May 2021*
Accepted: *29 June 2021*
Published: *24 August 2021*

Citation:
Dou Z and Ma X (2021) Inferring Functional Epigenetic Modules by Integrative Analysis of Multiple Heterogeneous Networks. Front. Genet. 12:706952. doi: 10.3389/fgene.2021.706952

1. INTRODUCTION

DNA methylation modifies the cytosine base associating with cellular differentiation and cell development (Suzuki and Bird, 2008; Deaton and Bird, 2011; Teschendorff et al., 2012; Ziller et al., 2013). For example, DNA methylation regulates the expression of genes by decreasing the affinity of transcription factors (Bird and Wolffe, 1999). Furthermore, abberations of methylation directly result in oncogenesis of cancers (Varley et al., 2013). For instance, the methylation of CpG islands (CGIs) plays a critical role in renal cell cancers (Herman et al., 1994), breast cancer (Fleischer et al., 2014), and colorectal cancer (Hinoue et al., 2012).

Thus, it is promising to mine methylation patterns, such as the methylated CpG islands and epigenetic modules, because they are the foundation for revealing the mechanisms of cancers. For instance, dynamics of methylation of tissues is critical for the development of cells. The methylation patterns of genes closely associate with survival time of patients (Fleischer et al., 2014), and similarity of methylation profiles is also associated with cancer subtypes (West et al., 2013; Gavaert et al., 2015).

These efforts are insufficient to fully exploit the methylation patters because they only make use of methylation data, ignoring the regulation of methylation (Teschendorff and Relton, 2018; West et al., 2018). Since methylation directly regulates the expression of genes, it is natural to identify the epigenetic modules by integrating them. However, it is non-trivial for this issues largely due to two reasons. First, the pre-requisite of the integration of methylation and expression is the matched samples. Second, no cut-off definition of epigenetic modules is available because the regulation strategies vary. For instance, in most case, methylation in promoters negatively regulates the expression, whereas the positive regulation also exists (Varley et al., 2013).

For the first concern, the world consortia make use of the next-generation sequencing technologies to generate sample-matched data for cancers, which enables the possibility to exploit epigenetic modules. For instance, The Cancer Genome Atlas (TCGA)[1] produces genomic data for various cancers, covering mutation, transcription, methylation, etc. Furthermore, Encyclopedia of DNA Elements (ENCODE)[2] generate matched samples for cell lines and tissues.

For the second concern, even though it is intuitive to define epigenetic module for methylation profiles and networks by simply extending the traditional clustering problem, it is difficult to present a satisfied definition with heterogeneous data. The available algorithms for the integration of methylation and expression by either using a integrated network and multiple networks. Algorithms in the first class construct an integrated network, where the correlation between methylation and expression are integrate edge weight. Then, the epigenetic module in the integrated network is defined as a dense subgraph. For example, the FEM algorithm (Jiao et al., 2014) addresses this problem with the assumption that DNA methylation and expression is anti-correlated, where hot-spot and methylated modules are successfully identified. However, the recent evidence indicates that the correlation between methylation and expression could be both positive and negative (Varley et al., 2013), implying that the integrated network-based approaches are not precise enough to characterize the epigenetic modules.

To attack this issue, efforts have been devoted by using multiple networks to identify graph patterns. For example, in our previous study (Ma et al., 2014), dynamic modules are extracted from multiple networks by exploiting the temporality of cancer progression. Driver genes of cancers can be identified by exploiting the relations of various layers (Cantini et al., 2015), implying the importance and effectiveness of multiple networks. Clustering multiple networks aims to identify modules in networks, which can be achieved by extending measurement for single networks (Didier et al., 2015). These results demonstrate that multiple networks are more accurate and generalized than single networks in terms of characterizing biological patterns. In our previous study (Ma et al., 2017), the epigenetic module is a group of co-methylated and co-expressed genes in multiple networks, and then the epigenetic modules are discovered by using the M-Module algorithm (Ma et al., 2014). The success of the multiple network-based approaches demonstrates that the multiple networks model is much better than the integrated network base method.

Even though multiple network-based algorithms have been devoted to the epigenetic module discovery, many unsolved problems exit. Particularly, the quantification of modules in multiple networks is fundamental, and how to further improve performance of algorithms for epigenetic modules. In the present study, we discuss these two issues. To identify the epigenetic modules in the co-methylation and co-expression networks, the key problem is how to characterize the topological structure of modules in multiple networks. Then, we define the epigenetic module as the common module in multiple networks. To discover the functional epigenetic modules in multiple networks, a novel non-negative matrix factorization algorithm for epigenetic module (Ep-jNMF) is proposed, which jointly analyzes the gene co-expression and co-methylation networks (**Figure 1**). It first constructs the two layer networks, and extracts features using matrix factorization, where the topological structure is regularized into the objective function. Extensive experiments are performed, where Ep-jNMF achieves the best performance on the artificial networks. Moreover, it identifies more biological meaningful modules than the baselines, and some of obtained modules precisely predict the survival time of patients.

The rest of this study is organized as follows: section 2 presents the mathematical model and algorithm. The experiments and conclusion are depicted in sections 3 and 4, respectively.

2. METHODS

The model and procedure of Ep-jNMF are depicted in this section.

2.1. Notations

A network (graph) is denoted by $G = (V, E)$ with vertex set V and edge set E. Multiple network $\mathcal{G} = \{G_1, G_2, \ldots, G_M\}$ is a sequence of networks, where G_m is the m-th snapshot. In this study, the vertex set of $\mathcal{G}$ is fixed, i.e., $G_m = (V, E_m)$. The adjacent matrix of $\mathcal{G}$ is a tensor $W = (w_{ijm})_{n\times n\times M}$, where $n = |V|$ and w_{ijm} is the weight on the edge (v_i, v_j) in G_m. Actually, $W = [W_1, W_2, \ldots, W_M]$, where $W_m = (w_{ijm})_{n\times n}$ is the adjacency matrix of G_m. In this study, the attached subscript m represents the value of the variable at condition m.

Vertex degree is the sum of weights on the incident edges, i.e., $d_{im} = \sum_j w_{ijm}$. Betweenness is a typical centrality (Freeman, 1979; Brandes, 2001), which is defined as

$$betweenness_m(v) = \sum_{v_i\neq v_j, v_i\neq v, v_j\neq v} \frac{g_{ivj}}{g_{ij}},$$

where g_{ivj} and g_{ij} are the number of the shortest paths between v_i and v_j passing, and without passing v, respectively. Given a

[1]https://cancergenome.nih.gov/
[2]https://www.encodeproject.org/

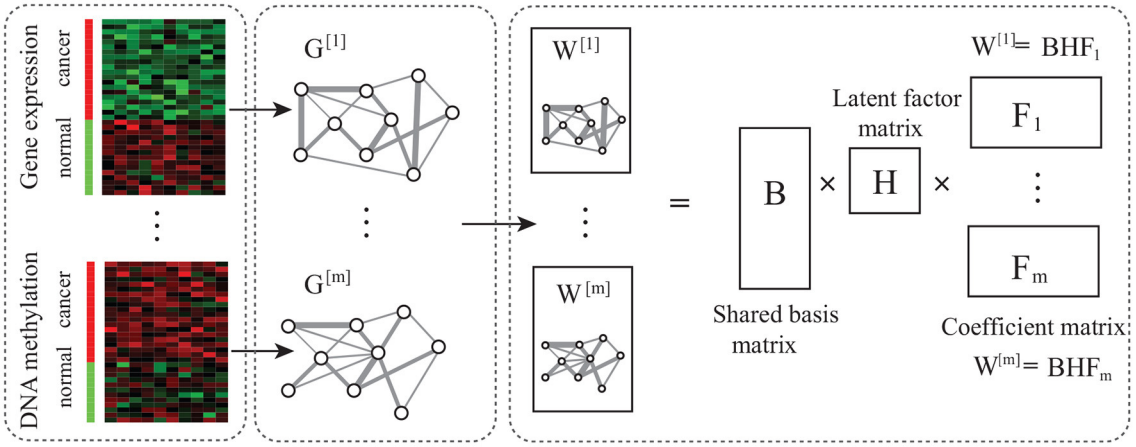

FIGURE 1 | Overview of Ep-jNMF. It is composed of network construction, matrix factorization, and module discovery, where network construction obtains the gene co-expression and co-methylation networks, and the matrix decomposition extracts features.

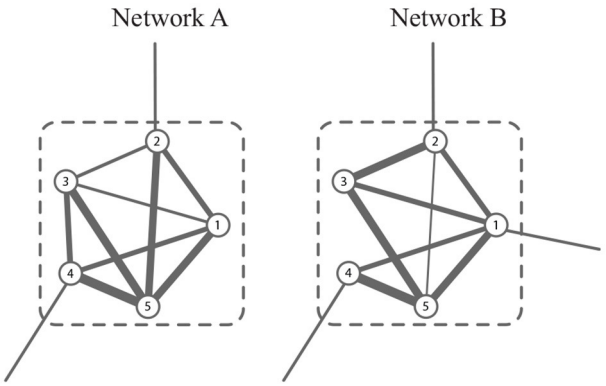

FIGURE 2 | A schematic example of common module, which is well-connected in both networks.

group of genes, denoted by C, the density of C in network G_m is defined as

$$Density_m(C) = \frac{2|E_m(C)|}{|C|(|C|-1)},$$

where $E_m(C)$ is the edge set of the subgraph induced by C in network G_m, i.e., $E_m(C) = \{(v_i, v_j)|v_i \in C, v_j \in C, (v_i, v_j) \in E_m\}$.

In G, a module is a group of vertices with more edges within it, and fewer ones outside it. In $\mathcal{G}$, the common module is a group of vertices whose connectivity is strong in all snapshots. For example, the module consisting of $\{1, 2, \ldots, 6\}$ in **Figure 2** is well-connected in both networks. In this study, we aim to obtain the common modules in the co-expression and co-methylation networks. The common module detection corresponds to a hard partitioning $\{C_1, C_2, \ldots, C_k\}$ (denoted by $\{C_l\}_{l=1}^k$) such that $C_{l_1} \cap C_{l_2} = \emptyset$ if $l_1 \neq l_2$ and $V = \sum_l C_l$, where k is the number of modules.

2.2. Mathematical Model

The quantification of connectivity of common modules in multiple networks is fundamental. Typical measurements, including the entropy function (Ma et al., 2014), modularity (Newman and Girvan, 2004), and modularity density (Li et al., 2008), are proposed. However, these strategies are inapplicable for the multiple networks. Here, we extend the modularity density D (Li et al., 2008) since it tolerates the resolution limit problem at some extent. Specifically, connectivity of module C_l in G_m is defined as

$$D_m(C_l) = \frac{1}{\sum_{v_i \in C_l} d_{im}} \big(L(C_l, C_l) - L(C_l, \overline{C}_l)\big), \tag{1}$$

where $L(C_l, C_l) = \sum_{v_i \in C_l, v_j \in C_l} w_{ijm}$ and $\overline{C}_l = V \backslash C_l$. Ideally, we maximize the connectivity of module C_l in all snapshots, i.e.,

$$\begin{cases} \max D_1(\{C_l\}), \\ \cdots \\ \max D_M(\{C_l\}). \end{cases} \tag{2}$$

However, it is difficult to reach maximal value for each network. Therefore, we transform the multi-objective function in Equation (2) into a single objective function using the geometric mean of the connectivity, i.e.,

$$D(C_l) = (\prod_m D_m(C_l))^{1/M}. \tag{3}$$

The underlying assumption is that a group of genes form a common module if and only if they are well-connected in all networks.

The partitioning $\{C_l\}_{l=1}^k$ is represented by $X_{n\times k}$ with x_{ij} =1 if $v_i \in C_j$, 0 otherwise. The overall function is the connectivity of all modules, i.e.,

$$\sum_l \max D(C_l) \tag{4}$$

$$s.t. \begin{cases} x_{ij} \in \{0,1\}, \\ \sum_{j=1}^{k} x_{ij} = 1, \\ \sum_{i=1}^{n} x_{ij} \geq 1, \end{cases}$$

where the second constraint enable the hard partitioning, and the last one ensures non-empty of modules. To avoid multi-objectives in Equation (4), we relax it as

$$\max \sum_{l} D(C_l) \tag{5}$$

$$s.t. \begin{cases} x_{ij} \in \{0,1\}, \\ \sum_{j=1}^{k} x_{ij} = 1, \\ \sum_{i=1}^{n} x_{ij} \geq 1. \end{cases}$$

2.3. The Ep-jNMF Algorithm

The algorithm consists of three components, which are introduced in turn (**Figure 1**). Networks are constructed using the Pearson correlation of gene profiles, and the PCIT package (Reverter and Chan, 2008) is adopted to remove noise.

NMF (Lee and Seung, 1999) approximates the target matrix using the product of two low-rank matrices as

$$W \approx BF \tag{6}$$

$$s.t. \begin{cases} B \geq 0, \\ F \geq 0, \end{cases}$$

where $B_{n\times k}$ and $F_{k\times n}$ are the basis and coefficient matrix, respectively, and k is the number of features. Usually, $k \ll n$ indicates that BF represents a compressed form of the original data W. Not allowing negative entries in B and F enables a non-subtractive combination of parts to form a whole. Equation (6) is solved by minimizing the approximation error as

$$e(B,F) = \|W - BF\|^2, \tag{7}$$

where $\|W\|$ is the Frobenius Norm of matrix W. Tri-factorization is more efficient than NMF (Yoo and Choi, 2010), where Equation (8) is formulated as

$$e(B,F) = \|W - BHF\|^2, \tag{8}$$

where H is the factor matrix.

For each snapshot, Ep-jNMF jointly factorizes W_m as

$$W_m \approx BHF_m. \tag{9}$$

Intuitively, we can minimize the approximation error for each snapshot as

$$\sum_{m} \min \|W_m - BHF_m\|^2 \tag{10}$$

$$s.t. \begin{cases} B \geq 0, \\ F_m \geq 0 \end{cases}$$

Algorithm 1: Ep-jNMF.

Input:
$\mathcal{G}$: Networks;
k: Number of features;

Output:
$\{C_l\}_{i=1}^{k}$: Common modules.

Procedure I: network construction

1: Constructing the gene co-expression (co-methylation) network using partial Pearson coefficient;

Procedure II: matrix decomposition

2: Fixing $F_m (1 \leq m \leq M)$ and H, update x B as equation (12);
3: Fixing B and $F_m (1 \leq m \leq M)$, update H as equation (13);
4: Fixing B and H, update $F_m (1 \leq m \leq M)$ as equation (14);
5: Keep updating the steps 3 and 4 until the termination criterion is reached;

Procedure III: common module discovery

6: Extracting modules from B;
7: **return**

However, it is difficult to reach minimization for each snapshot. Similar to Equation (5), we reformulate Equation (11)

$$\min \sum_{m} \|W_m - BF_m\|^2 \tag{11}$$

$$s.t. \begin{cases} B \geq 0, \\ F_m \geq 0. \end{cases}$$

The algorithm iteratively updates B and F_m by following the multiplicative rules (Lee and Seung, 1999), where the update rules are formulated as

$$B = B \frac{\sum_m W_m F_m^T}{B \sum_m F_m F_m^T}, \tag{12}$$

$$H = H \frac{\sum_m B^T F_m^T W_m}{B^T B F_m F_m^T}, \tag{13}$$

and

$$F_m = F_m \frac{B^T W_m}{B^T W_m B}. \tag{14}$$

Ep-jNMF (Algorithm 1) updates rules until termination is reached. For example, the approximation error threshold is set as 10^{-2}, or the maximum iteration number is 10^3. Because the initial solution is random, we repeat the procedure 50 runs with different initial solution matrices. The modules are extracted based on B, i.e., $x_{ij^*} = 1$ where $j^* = \arg max_j B_{ij}$, 0 otherwise. The Ep-jNMF algorithm involves one parameter k, which is the number of features to obtain the coefficient matrices. We select it using the instability of matrix factorization (Wu et al., 2016).

2.4. Algorithm Analysis

On the space complexity, $\mathcal{G}$ requires space $O(n^2M)$. The basis matrix requires space $O(nk)$ and the coefficient matrices need

space $O(knm)$. The space of the index matrix X is the same as the basis matrix B. In all, Ep-jNMF takes space $O(n^2m) + 2O(nk) + O(nkm) = O(n^2M)$ since $k \ll n$.

On the time complexity, for each F_m, Ep-jNMF needs time $O(rkn^2)$, where r is the number of iterations. And the running time for coefficient matrices in Ep-jNMF is $O(rkn^2M)$. Therefore, the total time complexity of Ep-jNMF is $O(rkn^2M)$.

3. EXPERIMENTS

To validate the performance of Ep-jNMF, we select sixe state-of-the-art methods for a comparison, including M-Module (Ma et al., 2014), consensus clustering (CSC) (Cantini et al., 2015), multiple-modularity method (MolTi) (Didier et al., 2015), stability NMF (sNMF) (Wu et al., 2016), FEM (Jiao et al., 2014) and spectral clustering (SPEC) (Newman, 2006a), covering single-network- and multiple-network-based approaches. The former ones are extended using the consensus strategy (Cantini et al., 2015).

3.1. Data and Criteria

The artificial networks are derived from GN benchmark (Newman, 2006b), where each snapshot consists of 4 equal size communities with 32 vertices, and the degree of vertices is fixed as 16. Parameter Z_{out} controls the noise level of networks, and Z_{out} increases from 1 to 8. By manipulating parameter Z_{out}, two types of multiple networks are generated, where in the homogeneous networks (HomoNet) the noise levels in snapshots are the same, and in heterogeneous networks (Heter-Net) it varies in different snapshots. Specifically, Z_{out} is fixed as 4 in the first snapshot, and it varies from 1 to 8 in the others. We downloaded the sample-matched gene expression and methylation profiles of breast cancer from TCGA. Specifically, the gene expression level is quantified using RPKM values and methylation level is measured by β signal, which are imputed using PCIT (Tibshirani et al., 2002).

The normalized mutual information (NMI) (Danon et al., 2005) measures the closeness of two partitioning: standard partition P^* and obtained partitioning P. NMI generates matrix N with the element N_{ij} as the size of vertices overlapped by C_i^* and C_j, which is formulated as

$$NMI(P, P^*) = \frac{-2\sum_{i=1}^{|P|}\sum_{j=1}^{|P^*|} N_{ij}\log(\frac{N_{ij}N}{N_{i.}N_{j}})}{\sum_{i=1}^{|P|} N_{i.}\log(\frac{N_{i.}}{N}) + \sum_{i=1}^{|P^*|} N_{.j}\log(\frac{N_{.j}}{N})},$$

where $|P|$ is the cardinality of P and $N_{i.} = \sum_j N_{ij}$.

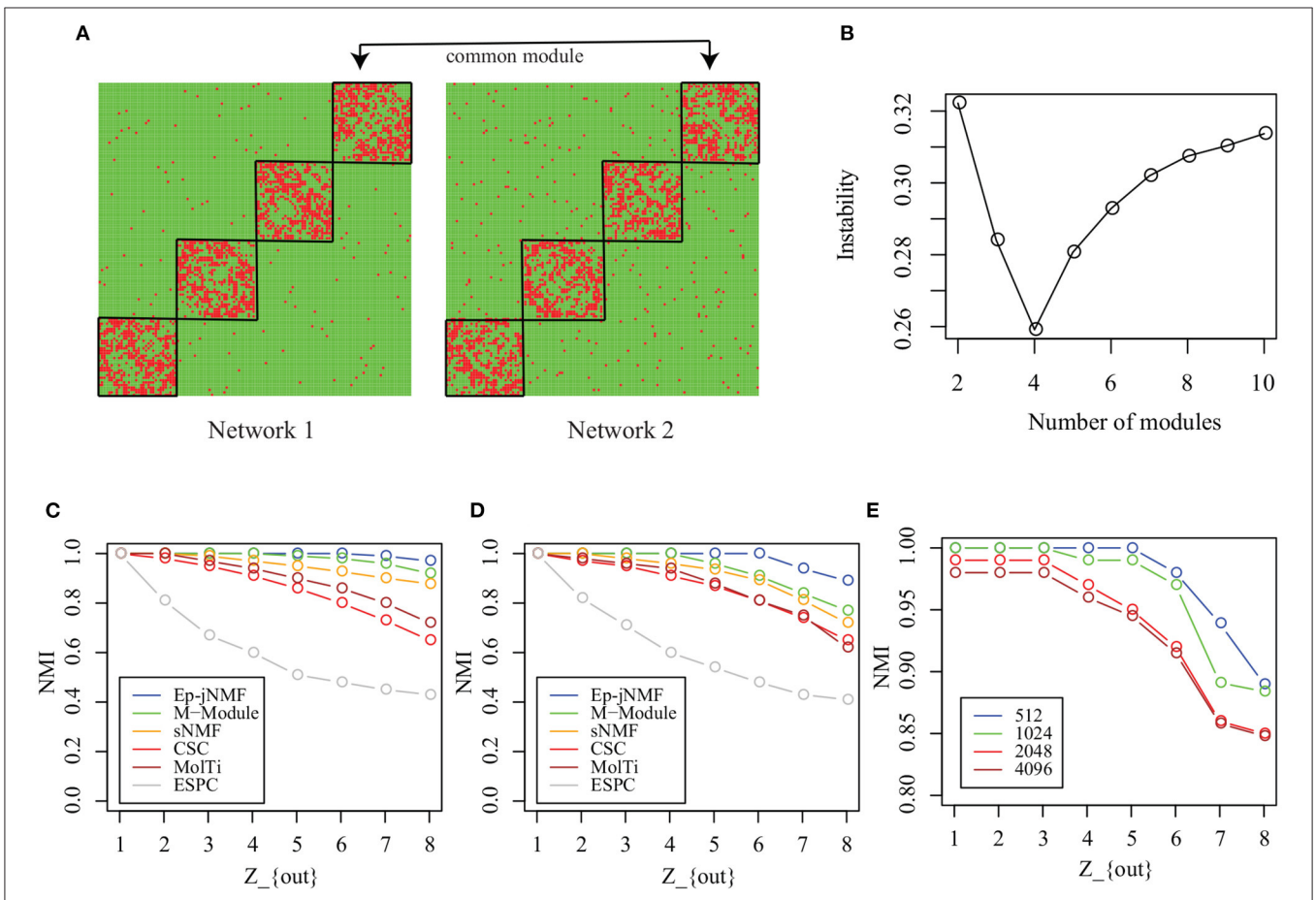

FIGURE 3 | Performance on artificial networks: **(A)** Heatmap of common modules in Homo-Net ($Z_{out} = 1$); **(B)** selection of the number of modules using instability of matrix factorization with various noise levels; **(C)** NMIs of algorithms on Homo-Net; **(D)** NMI of algorithms on Heter-Net; and **(E)** scalability analysis of Ep-jNMF.

To check whether the predicted epigenetic modules are biological meaningful, various annotation databases are selected as gold standards for the enrichment analysis, where the significance is obtained by using the hypergeometric test (corrected by Benjamini–Hochberg test) with a cutoff of 0.05.

3.2. Performance on Simulated Networks

Each simulated snapshot contains 128 vertices and 4 modules of equal size with fixed degree 16. Parameter Z_{out} controls the noise level of networks. As Z_{out} increases from 1 to 8, the module structure is obscure. In this study, we generate two types of simulated networks with $M = 2$: Homo-Net and Heter-Net. Specifically, the parameter Z_{out} of both networks of Homo-Net is the same, while the Z_{out} of one network of Heter-Net is fixed as 4 and the parameter of the other network varies from 1 to 8. **Figure 3A** is the heatmap of the Homo-Net networks with $Z_{out} = 1$, where the common modules locate at the diagonal.

First, how the Ep-jNMF algorithm selects the parameter k, i.e., the number of modules, is studied. How the instability of Ep-jNMF changes as k increases from 2 to 10 for Homo-Net

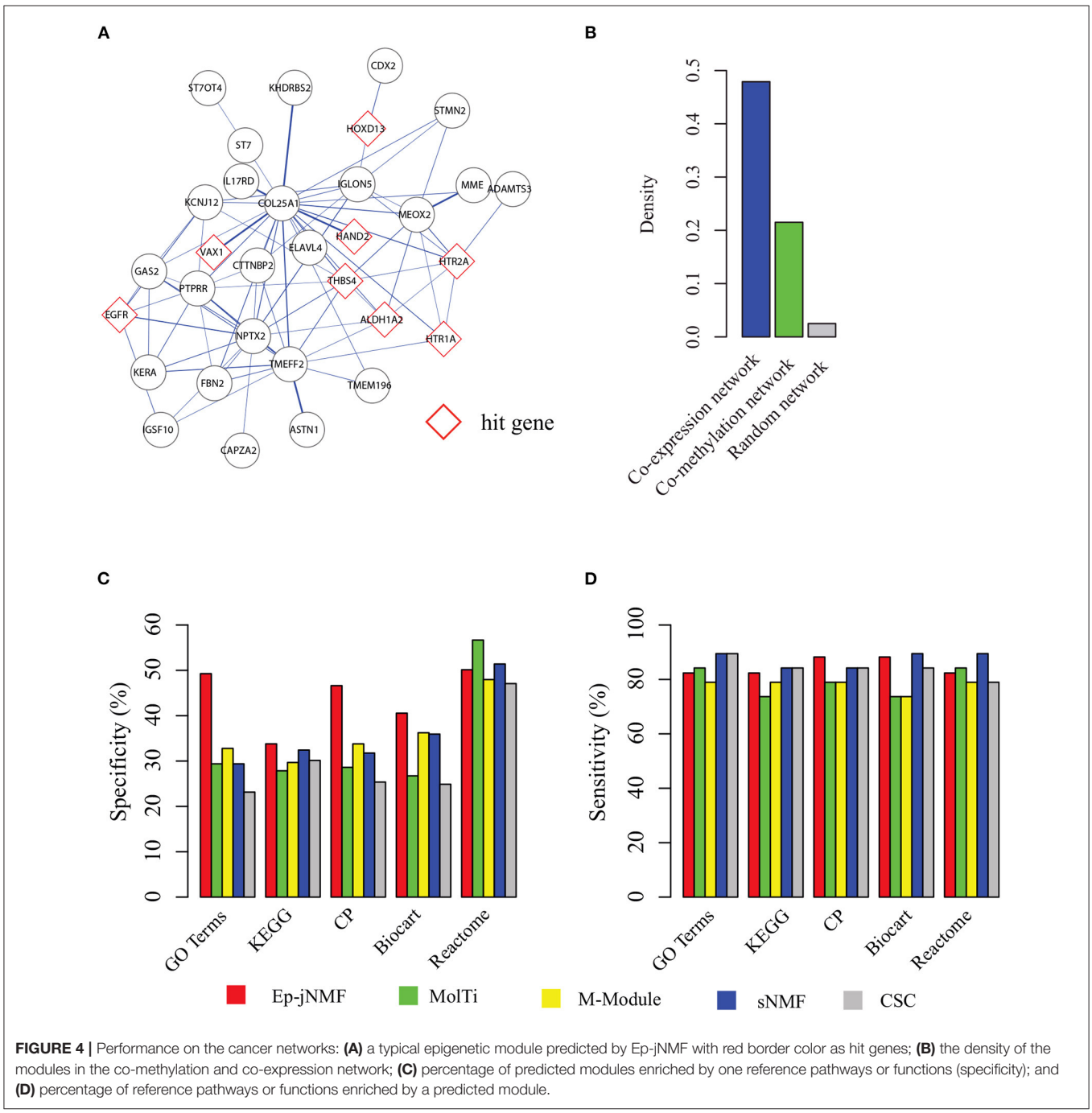

FIGURE 4 | Performance on the cancer networks: **(A)** a typical epigenetic module predicted by Ep-jNMF with red border color as hit genes; **(B)** the density of the modules in the co-methylation and co-expression network; **(C)** percentage of predicted modules enriched by one reference pathways or functions (specificity); and **(D)** percentage of reference pathways or functions enriched by a predicted module.

is shown in **Figure 3B**, where it chooses the optimal value 4 because the minimal is reached at 4. The similar pattern repeats for Heter-Net, which is not shown because of redundancy. The result demonstrates that the strategy is promising in selecting the number of modules.

Then, we compare M-Module, CSC, MolTi, sNMF, and SPEC on the simulated networks. **Figure 3C** shows the accuracy of various algorithms for Homo-Net, while **Figure 3D** shows the accuracy of various algorithms for Heter-Net. The performance of all these algorithms decreases as the parameter Z_{out} increases from 1 to 8 because the module structure is difficult to detect as Z_{out} increases. M-Module and Ep-jNMF outperform the rest of algorithms because the CSC, MolTi, and SPEC are based on the consensus clustering, which ignores the connection among multiple networks. However, M-Module and Ep-jNMF make use the multiple networks simultaneously during the module search procedure, which improves the accuracy of detecting the common modules. In all, the Ep-jNMF algorithm is better than the M-Module algorithm. More specifically, when Z_{out} is less than or equal to 5 in Homo-Net, the Ep-jNMF and M-Module algorithms have a similar performance. When Z_{out} is greater than or equal to 6, Ep-jNMF outperforms M-Module, indicating the superiority of Ep-jNMF. The similar tendency also repeats in Heter-Net (**Figure 3D**).

Finally, we investigate the accuracy of Ep-jNMF by increasing the number of vertices from 512 to 4096. The performance of Ep-jNMF is shown in **Figure 3E**, suggesting that the algorithm is robust. These results demonstrate that Ep-jNMF is promising to identify common modules in artificial networks.

3.3. Performance on Cancer Networks

For cancer networks, we select the Ep-jNMF, M-Module, MolTi, sNMF, and FEM algorithms for a comparison since they significantly outperform CSC and SPEC. The Ep-jNMF, M-Module, MolTi, sNMF, and FEM algorithms identify 17, 26, 94, 26, and 460 modules, respectively.

Figure 4A presents a functional epigenetic module obtained by Ep-jNMF with cell proliferation (p = 3.8E-4), which is critical for breast cancer metastasis (Loayza-Puch et al., 2016; Thienpont et al., 2016). Interestingly, the epigenetic module contains the HAND2 sub-module, which is validated by the biological experiments (Jones et al., 2013). The HAND2 module has been used as the benchmark for the algorithms for the methylated module (Jiao et al., 2014). Furthermore, we find that only FEM and Ep-jNMF can discover the HAND2 module, whereas the others cannot. These results imply that Ep-jNMF is effective for the identification of critical epigenetic modules. To check whether the genes within the obtained common module are well-connected in both networks, the density of the module in different snapshots is shown in **Figure 4B**. Clearly, the connectivity is strong in both snapshots because the density is 0.47 and 0.22, which is significantly higher than that in random networks. The possible reason why the module is much denser in the co-expressed network than that in the co-methylated network is that methylation is more specific than expression.

To fully validate the performance of Ep-jNMF, Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa et al., 2012), Reactome (Croft et al., 2014), Biocart (Nishimura, 2001), and Canonical pathways (Subramanian et al., 2005) are selected as reference annotation. To evaluate the performance, we first check the percentage of predicted modules that significantly enriched by at least one reference annotation, and then we calculate the percentage of the reference pathways that significantly overlaps with at least one predicted module. **Figures 4C,D** show that Ep-jNMF achieves higher specificity with comparable sensitivity, implying that the predicted modules are more meaningful in terms of the biological background.

3.4. Performance on Predicting Cancer Subtypes

Evidence proves that hub genes facilitate the prognosis of cancers (Taylor et al., 2009). Therefore, we check whether epigenetic modules also serve as biomarkers to discriminate cancer subtypes by using the methylation profiles. We select modules predicted by Ep-jNMF, FEM, sNMF, M-Module, and MolTi. Furthermore, we also include size-matched set of randomly modules to validate the performance of different features. Support vector machine is selected as classifier to calculate the percentage of patient samples that are classified correctly (accuracy). The fivefold cross-validation is used for SVM, which is shown in **Figure 5A**, indicating that modules obtained by Ep-jNMF are more discriminative than the others. Specifically, the accuracy of Ep-jNMF is 82.4%, whereas that of M-Module is 75.1% ($p = 4.9E-6$, Wilcoxon test), showing that modules in multiple networks are more accurate to capture the structure and functions of cancers. The external dataset is also performed (GSE5874), which is shown **Figure 5B**. Specifically, Ep-jNMF is also superior to the baselines (i.e., 74.6% for Ep-jNMF vs. 62.9% for M-Module, $p = 2.1E-4$, Wilcoxon test).

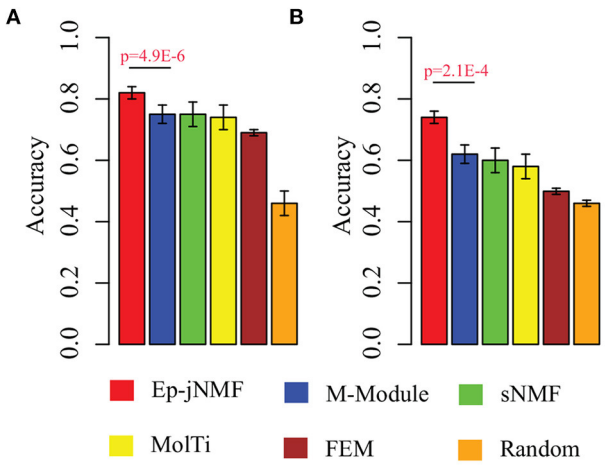

FIGURE 5 | Performance on the prediction of subtypes with various validation in terms of accuracy: **(A)** internal validation with the error bar as for the standard deviation and **(B)** external validation.

4. CONCLUSION

Epigenetic modification is a critical biological process, and mining the patterns is promising for the understanding of cancers. The advances in the next-generation sequencing technologies facilitate the generation of genomic data for cancers, which enables the integrative analysis of omic data. How to integrate gene methylation and expression data is the fundamental step for revealing the mechanisms of cancers. The traditional methods fuse them into a single network by assuming the positive and negative correlation between expression and methylation. However, these strategies are criticized for the undesirable performance since the underlying assumption is not consistent with the biological principle.

In this study, we use the multiple networks model to characterize functional epigenetic modules, which corresponds to the common modules detection in multiple networks. Finally, we present a matrix factorization algorithm for extracting the common modules from heterogeneous networks. Overall, the contributions are summarized as follows: (i) it provides a mathematical model for the functional epigenetic modules, which overcomes the limitation of the current approaches, i.e., the correlation specification between methylation and expression is not required; (ii) a joint learning method is proposed to identify the epigenetic modules in multiple networks, which avoids the structure preservation of single network-based method, which can be easily extended for other data, such as Chip-seq and mutation data; and (iii) the experiments show the superiority of Ep-jNMF.

In further research, we will investigate how to integrate heterogeneous entities, such as microRNAs, to extract the regulation programming based on multiple heterogeneous networks.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: TCGA.

AUTHOR CONTRIBUTIONS

ZD and XM designed the method and coded the algorithm. XM wrote the paper. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the open funding of Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education (Grant Nos. KFKT202009 and KFKT202010).

ACKNOWLEDGMENTS

The authors appreciate the reviewers for their suggestions.

REFERENCES

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J. M., et al.. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet*. 25, 25–29. doi: 10.1038/75556

Bird, A., and Wolffe, A. (1999). Methylation-induced repression-belts, braces, and chromatin. *Cell* 99, 451–454. doi: 10.1016/S0092-8674(00)81532-9

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Cantini, L., Medico, E., Fortunato, S., and Caselle, M. (2015). Detection of gene communities in multi-networks reveals cancer drivers. *Sci. Rep*. 5:17386. doi: 10.1038/srep17386

Croft, D., Mundo, A., Haw, R., Milacic, M., Joel, W., Wu, G., et al. (2014). The reactome module knowledgebase. *Nucleic Acids Res*. 42, D472–D477. doi: 10.1093/nar/gkt1102

Danon, L., Duch, J., Diaz-Guileram, A., and Arenas, A. (2005). Comparing community structure identification. *J. Stat. Mech. Theory Exp*. 2005:P09008. doi: 10.1088/1742-5468/2005/09/P09008

Deaton, A., and Bird, A. (2011). Cpg islands and the regulation of transcription. *Genes Dev*. 25, 1010–1022. doi: 10.1101/gad.2037511

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. *Peer J* 3:e1525. doi: 10.7717/peerj.1525

Fleischer, T., Frigessi, A., Johnson, K., Edvardsen, H., Touleimat, N., Klajic, J., et al. (2014). Genome-wide dna methylation profiles in progression to *in situ* and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol*. 15:435. doi: 10.1186/s13059-014-0435-x

Freeman, L. (1979). Centrality in social networks I: conceptual clarification. *Soc. Netw*. 1, 215–239. doi: 10.1016/0378-8733(78)90021-7

Gavaert, O., Tibshirani, R., and Plevritis, S. (2015). Pancancer analysis of DNA methylation-driven genes using methylmix. *Genome Biol*. 16:17. doi: 10.1186/s13059-014-0579-8

Herman, J., Latif, F., Weng, Y., Lerman, M., Zbar, B., Samid, D., et al. (1994). Silencing of the vhl tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl. Acad. Sci. U.S.A*. 91, 9700–9704. doi: 10.1073/pnas.91.21.9700

Hinoue, T., Weisenberger, D., Lange, C., Shen, H., Byun, H. M., Van, D. B. D., et al. (2012). Genome-scale analysis of aberrant dna methylation in colorectal cancer. *Genome Res*. 22, 271–282. doi: 10.1101/gr.117523.110

Jiao, Y., Widschwendter, M., and Teschendorff, A. (2014). A systems-level integrative framework for genome-wide dna methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 30, 2360–2366. doi: 10.1093/bioinformatics/btu316

Jones, A., Teschendorff, A., Li, Q., Hayward, J. D., Kannan, A., Mould, T., et al. (2013). Role of DNA methylation and epigenetic silencing of hand2 in endometrial cancer development. *PLoS Med*. 10:e1001551. doi: 10.1371/journal.pmed.1001551

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

Lee, D., and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. doi: 10.1038/44565

Li, Z., Zhang, S., and Wang, R. (2008). Quantative function for community detection. *Phys. Rev. E* 77:036109. doi: 10.1103/PhysRevE.77.036109

Loayza-Puch, F., Rooijers, K., Buil, L., Zijlstra, J., Vrielink, J., Lopes, R., et al. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. *Nature* 530, 490–494. doi: 10.1038/nature16982

Ma, X., Gao, L., and Tan, K. (2014). Modeling disease progression using dynamics of pathway connectivity. *Bioinformatics* 30, 2343–2350. doi: 10.1093/bioinformatics/btu298

Ma, X., Liu, Z., Zhang, Z., Huang, X., and Tang, W. (2017). Multiple network algorithm for epigenetic modules via the integration of

genome-wide DNA methylation and gene expression data. *BMC Bioinformatics* 1:18. doi: 10.1186/s12859-017-1490-6

Newman, M. (2006a). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74:036104. doi: 10.1103/PhysRevE.74.036104

Newman, M. (2006b). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Newman, M., and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* 69:026113. doi: 10.1103/PhysRevE.69.026113

Nishimura, D. (2001). Biocarta. *Biotech. Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Reverter, A., and Chan, E. (2008). Combining partial correlation and an information theory approach to the reverse engineering of gene co-expression networks. *Bioinformatics* 24, 2491–2497. doi: 10.1093/bioinformatics/btn482

Subramanian, A., Tamayo, P., Mootha, V., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Suzuki, M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat. Rev. Genet.* 9, 465–476. doi: 10.1038/nrg2341

Taylor, I., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., et al. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* 27, 199–204. doi: 10.1038/nbt.1522

Teschendorff, A., Jones, A., Fiegl, H., Sargent, A., Zhuang, J., Kitchener, H., et al. (2012). Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* 4:24. doi: 10.1186/gm323

Teschendorff, A., and Relton, C. (2018). Statistical and integrative system-level analysis of dna methylation data. *Nat. Rev. Genet.* 19, 129–147. doi: 10.1038/nrg.2017.86

Thienpont, B., Steinbacher, J., Zhao, H., D'Anna, F., Kuchnio, A., Ploumakis, A., et al. (2016). Tumour hypoxia causes dna hypermethylation by reducing tet activity. *Nature* 537, 63–68. doi: 10.1038/nature19081

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, B. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6567–6572. doi: 10.1073/pnas.082099299

Varley, K., Gertz, J., Bowling, K., Parker, S., Reddy, T. E., Pauli-Behn, F., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* 23, 555–567. doi: 10.1101/gr.147942.112

West, J., Beck, S., Wang, X., and Teschendorff, A. (2013). An integrative network algorithm identifies age-associated differential methylation interactome hotspots targeting stem-cell differentiation pathways. *Sci. Rep.* 3:1630. doi: 10.1038/srep01630

West, J., Beck, S., Wang, X., and Teschendorff, A. (2018). Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat. Rev. Clin. Oncol.* 15, 292–309. doi: 10.1038/nrclinonc.2018.30

Wu, S., Joseph, A., Hammonds, A., Celniker, S., Yu, B., and Frise, E. (2016). Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4290–4295. doi: 10.1073/pnas.1521171113

Yoo, J., and Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Inform. Process. Manage.* 46, 559–570. doi: 10.1016/j.ipm.2009.12.007

Ziller, M., Gu, H., Muller, F., Donaghey, J., Tsai, L., Kohlbacher, O., et al. (2013). Charting a dynamic dna methylation landscape of the human genome. *Nature* 500, 477–481. doi: 10.1038/nature12433

Conflict of Interest: ZD is employed by China Electronics Technology Group Corporation.

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 September 2021
doi: 10.3389/fgene.2021.728563

Identification of New Genes and Loci Associated With Bone Mineral Density Based on Mendelian Randomization

*Yijun Liu[1], Guang Jin[1], Xue Wang[2], Ying Dong[3] and Fupeng Ding[1]**

[1] Department of Orthopedics, The First Hospital of Jilin University, Changchun, China, [2] Department of Anesthesiology, The First Hospital of Jilin University, Changchun, China, [3] The Third Department of Radiotherapy, Jilin Provincial Tumor Hospital, Changchun, China

Bone mineral density (BMD) is a complex and highly hereditary trait that can lead to osteoporotic fractures. It is estimated that BMD is mainly affected by genetic factors (about 85%). BMD has been reported to be associated with both common and rare variants, and numerous loci related to BMD have been identified by genome-wide association studies (GWAS). We systematically integrated expression quantitative trait loci (eQTL) data with GWAS summary statistical data. We mainly focused on the loci, which can affect gene expression, so Summary data-based Mendelian randomization (SMR) analysis was implemented to investigate new genes and loci associated with BMD. We identified 12,477 single-nucleotide polymorphisms (SNPs) regulating 564 genes, which are associated with BMD. The genetic mechanism we detected could make a contribution in the density of BMD in individuals and play an important role in understanding the pathophysiology of cataclasis.

Keywords: BMD, GWAS, eQTL, causative gene, disease susceptibility, SMR

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Hui Ding,
University of Electronic Science and Technology of China, China
Hong Ju,
Heilongjiang Vocational College of Biology Science and Technology, China

****Correspondence:***
Fupeng Ding
liuyij@jlu.edu.cn

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

Received: *21 June 2021*
Accepted: *02 August 2021*
Published: *08 September 2021*

Citation:
Liu Y, Jin G, Wang X, Dong Y and Ding F (2021) Identification of New Genes and Loci Associated With Bone Mineral Density Based on Mendelian Randomization. Front. Genet. 12:728563. doi: 10.3389/fgene.2021.728563

INTRODUCTION

Bone mineral density (BMD), is a main risk factor for osteoporosis (OP) or systemic bone loss, which is associated with the increasing risk of fragility fracture, especially for older women (Glüer et al., 2004; Cauley et al., 2007). BMD also plays a role for causing bone fractures, including pressure fractures (Nattiv, 2000). Generally, BMD can be detected by dual-energy X-ray absorptiometry (DXA), which is a non-invasive bone densitometry method but hard to implement. Another method to measure BMD is quantitative ultrasound of the calcaneus (QUS), which is flexible, inexpensive, and easier to perform. BMDs at the spine and hip are reported to be highly heritable (Arden et al., 1996; Lee et al., 2006), which could be detected by DXA (Gonnelli et al., 2005), and are fracture risk related to fracture risk (Bauer et al., 2007).

Based on genome-wide association studies (GWAS) analysis using heel ultrasound parameters, Moayyeri et al. (2014) identified mutations at nine loci, including seven previously reported loci. GWAS, so far, have detected more than 100 genetic variants associated with BMD, including many significant loci associated with risk of fractures. In recent years, more and more BMD risk variants with low frequencies have been detected based on deep whole-genome sequencing. However, most experiment-verified variants can rarely explain approximately 5.8% of the phenotypic variance in BMD (Zheng et al., 2015). Estrada et al. (2012) identified 62 significant SNPs by performing a meta-analysis consisting of 17 BMD GWAS studies, which focused on lumbar spine or femur neck.

Kemp et al. (2017) performed a genome-wide association screen by UK Biobank and identified 307 independent SNPs located in the 203 loci. However, it remains elusive on how these genetic loci lead risk to BMD based on linkage disequilibrium phenomenon (LD) between detected SNPs and real causative mutations. In addition, due to the strict statistical significance threshold set in GWAS analysis, it is difficult to detect co-pathogenic loci in a single GWAS study. Therefore, we need to use other omics data to reveal the potential effect of these weak GWAS association signals on BMD, which may help to understand the heritability of this trait.

By these biological experiments, researchers have found several genes, which are related to BMD. Some researchers have used computational method to identify more BMD-related genes (Wu et al., 2021). Machine learning and deep learning methods have been widely used in the prediction of trait-related genetic factors (Zhuang et al., 2019; Tarwadi et al., 2020; Zhao et al., 2020a). Most of these methods predict the associations between biomolecules by feature extraction and building mathematical models (Tianyi et al., 2021; Zhao et al., 2020b, 2021a). However, these studies fail to explain the biological mechanism of results. Therefore, it is necessary to further reveal the mechanism of significant SNPs identified by GWAS (Zhao et al., 2019).

Considering the influence of LD, systematical approaches are proposed to explore the latent regulatory functions of the risk variants reported in previous GWAS studies by integrating multiple omics data (Peng and Zhao, 2020; Zhao et al., 2020c). Since gene expression is an important factor related to genetic mutations and traits, many researchers tried to reveal pathogenesis by gene expression (Zhao et al., 2021b). Researches have detected numerous expression quantitative trait loci (eQTLs) associated with BMD based on eQTL data from primary bone cell cultures (Grundberg et al., 2009; Kwan et al., 2009). Kwan et al. (2009) has found that rs136564 plays an important role in regulating the expression of a novel transcript of FAM118A, and rs136564 is also reported to be related to BMD based on GWAS analysis. Therefore, many studies focused on confirming whether an SNP can be detected by both GWAS and eQTL analysis (Farber, 2012). However, most studies focused on separately analyzing GWAS data and eQTL data rather than in an integrative way to identify disease genes (Farber and Lusis, 2008).

Mendelian randomization approach is proposed as a method of using genetic variants as instrumental variables to examine the causal influence of a modifiable exposure on diseases. Based on this assumption, we can identify the most functionally related genes to diseases. Apparently, complex traits, such as BMD, are not only derived from the effect of a single gene but also the integrated influence from complex biological networks (Schadt, 2009). In this study, we applied the Mendelian randomization (MR) method based on summary statistic data to identify novel causative genes associated with BMD. We first collected two GWAS datasets from UK Biobank [including 394,929 individuals (Zheng et al., 2015)], UK10K [including 32,965 individuals (Kim, 2018)], and blood eQTL data (Westra et al., 2013). Then SMR was implemented to investigate new genes and loci associated with BMD. As a result, we identified 12,477 SNPs regulating 564 genes, which have causal effect on BMD. Finally, we assessed the functional interactions between these genes to examine their underlying functional mechanism.

DATA AND METHODS

Data

Genome-Wide Association Studies Summary Data

The GWAS summary data were obtained from UK Biobank and UK10K project, respectively. Individuals (394,929) with genotype and phenotype data were collected from the UK Biobank. The DNA variants were filtered by MAF $>$ 0.1%. The dataset from UK10K is composed of 2,882 whole-genome sequencing (WGS data), 3,549 whole-exome sequencing (WES data), 26,543 deep imputation of genotyped samples, and 20,271 *de novo* replication genotyping. The detailed description information of GWAS datasets can be accessed from previous studies (Westra et al., 2013; Zheng et al., 2015).

Expression Quantitative Trait Loci Summary Data

It has been validated that bone metabolism is related to various types of cells such as peripheral blood monocyte cell (PBMC), B and T lymphocytes (Chalmers et al., 1981). PBMC plays an important role in studying gene expression functions related to human osteoporosis risk (Liu et al., 2005). They can also be considered as precursors of osteoclasts (Geissmann et al., 2010) and express various cytokines, which are essential in the biological process of osteoclast (Deng et al., 2011). B lymphocytes can also express biological factors associated with osteoclastogenesis and plays an important role in the immune system (Manabe et al., 2001). Recently, studies based on eQTL-mapping methods indicated that most of the disease-causative mutations actually have an influence on the expression level of nearby genes due to the phenomenon of LD (Dubois et al., 2010; Nicolae et al., 2010). Researchers have also identified that *trans*-eQTLs can reveal the downstream consequences of the variants (Fehrmann et al., 2011; Innocenti et al., 2011; Grundberg et al., 2012). In this study, we collected eQTL summary data of 5,311 samples in peripheral blood tissue, which is derived from a total of nine datasets from seven different cohorts (Westra et al., 2013).

Methods

Genome-Wide Association Studies Meta-Analysis

Since GWAS analysis focus on the effect of a single genetic variant, it ignores the interactions between different loci. However, the effect size of an SNP is different from diverse datasets. Thus, we performed a GWAS meta-analysis on two GWAS summary datasets in order to correct the effect size of multiple GWAS datasets. By assigning different weights to each SNP from different datasets, we can integrate these GWAS datasets into a more comprehensive one. There are three measurements to assess the association score between variants and the trait in GWAS dataset, β, SE, and *p*-value. β measures the estimate of a causative effect between SNP and trait, and SE indicates the standard deviation (SD) of β. The *p*-value denotes the significance level of association between SNP and the trait.

Since SE can represent the reliability of β, it can be inferred that the bigger the SE, the more inaccurate the β. Because SE is the SD of β, the weight of β can be denoted as the inverse ratio of the SE square. Thus, the weight w_i of β_i in the ith GWAS dataset can be denoted as:

$$w_i = 1/SE_i^2 \tag{1}$$

where SE_i denotes the SD of the SNP in the ith dataset.

Thus, we can integrate the effect size measurement β between different datasets, and it can be denoted as:

$$\beta = \sum_i \beta_i w_i / \sum_i w_i \tag{2}$$

In the meantime, SE after the integration of the datasets can be denoted as:

$$SE = \sqrt{1/\sum_i w_i} \tag{3}$$

Then we calculated the Z-score of SNPs based on the effect size β and SE to obtain the significance of SNPs. Z-score can be denoted as:

$$Z = \beta/SE \tag{4}$$

Then we obtained the p-value of the association after the integration of the effect of SNPs from different datasets based on the hypothesis testing of the normal distribution of the Z-score.

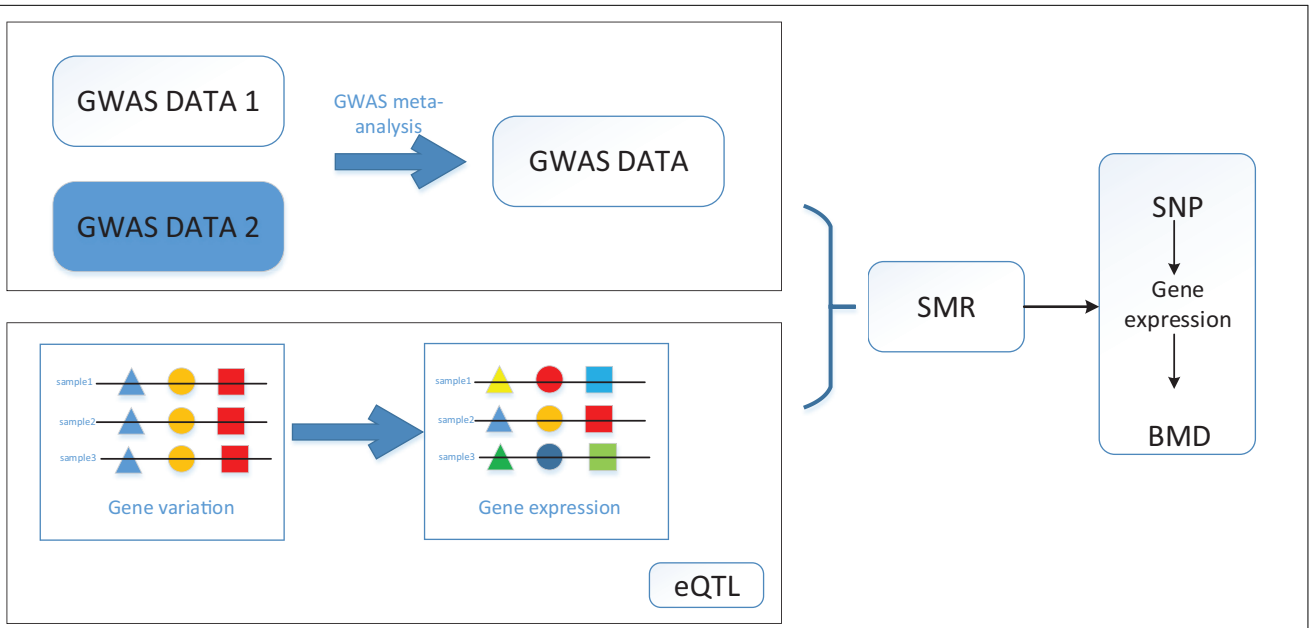

FIGURE 1 | Workflow of SMR on bone mineral density (BMD) based on genome-wide association studies (GWAS), and expression quantitative trait loci (eQTL) datasets.

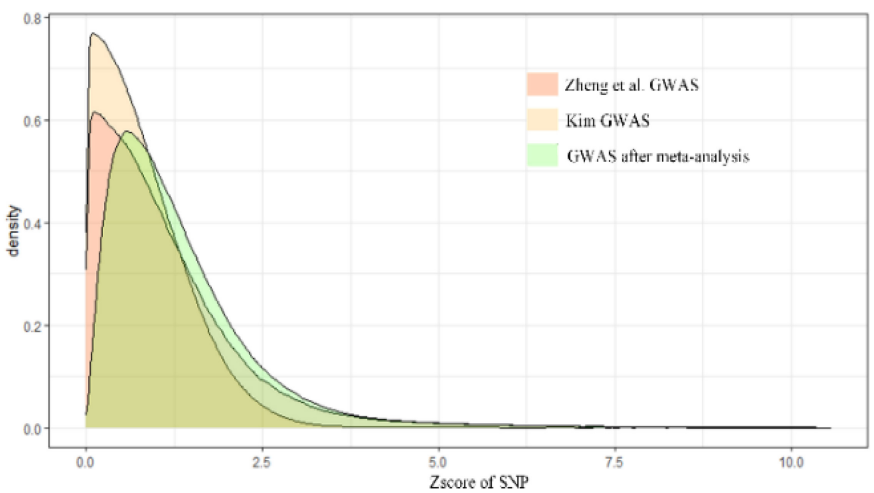

FIGURE 2 | The result of GWAS meta-analysis on BMD.

Thus, we can integrate multiple GWAS datasets by applying the above method. It can be deduced that the reliability of SNPs and SE are negatively correlated, and the weight of β is lower compared with other datasets, while the SE value is bigger. Thus, the value of β can be corrected across multiple datasets according to different weights.

Summary Data-Based Mendelian Randomization Analysis

Multiple potential and unmeasurable confounding factors may lead to huge challenges in inferring the causative relationship between genes and complex traits. However, genetic mutation is a major factor of heredity. Thus, exploring the underlying mechanism of genetic variants is important to reveal the pathologies of complex traits. Due to the linkage disequilibrium, the effect size between SNPs detected by GWAS analysis and BMD may not be accurate. Moreover, GWAS cannot fully explain the association between BMD and SNPs. Thus, the MR method is first proposed to consider a genetic variant as a factor to assess and examine for the effect size of an exposure variable on an outcome (Smith and Ebrahim, 2008). Based on the MR theory, if we use z to denote an SNP, x as the gene expression, and y as the BMD, then the association of gene expression (x) and BMD (y) can be denoted as b_{xy},

$$b_{xy} = b_{zy}/b_{zx} \tag{5}$$

where b_{zy} indicates the association between SNP and BMD, and it can be represented as the slope of z to y. b_{zx} denotes the association between SNP and gene expression, and it can be denoted as the slope of z to x. b_{zy} and b_{zx} can be obtained from two independent GWAS dataset and eQTL dataset.

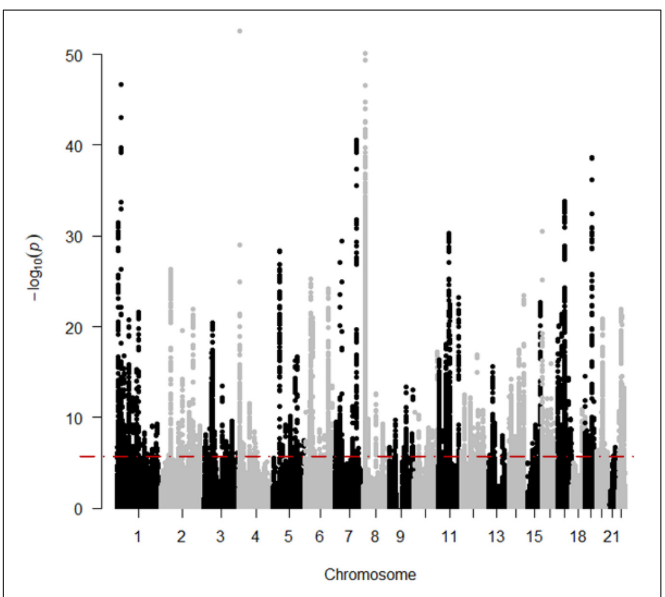

FIGURE 3 | The results of BMD-related genes based on SMR. The red line means "Significant threshold".

Then the sampling variance of the estimate value of b_{xy} can be denoted as:

$$var\left(\hat{b}_{xy}\right) = \left[var\left(y\right)\left(1 - P_{xy}^2\right)\right] \Big/ \left[n\ var\left(x\right) P_{zx}^2\right] \tag{6}$$

where n denotes the size of samples, $\hat{b}_{xy}$ denotes the estimate value of b_{xy} P_{xy}^2 indicates the proportion of variance in BMD, which is explained by gene expression, P_{zx}^2 indicates the proportion of variance in gene expression level explained by SNP. Therefore, the statistic T_{SMR} is utilized to test the significance of b_{xy}, and T_{SMR} can be represented as:

$$T_{SMR} = \hat{b}_{xy}^2 \Big/ var\left(\hat{b}_{xy}\right) \tag{7}$$

However, it is not realistic, so far, to collect genotype data and gene expression data from a very large sample size. Also, because the effect size of eQTL was unavailable, b_{zx} can be estimated from the Z-score of eQTL data as $\hat{b}_{zx}$:

$$\hat{b}_{zx} = Z_{zx} S_{zx} \tag{8}$$

where $S_{zx} = 1\sqrt{2f\left(1-f\right)\left(n + Z_{zx}^2\right)}$, f is the allele frequency, and n is the sample size. An unbiased estimate of b_{zx} could be denoted as $\hat{\varepsilon}_{zx}$. We therefore have:

$$\hat{b}_{xy} = \hat{b}_{zy} \Big/ \hat{\varepsilon}_{zx} \tag{9}$$

where $\hat{b}_{zy}$ denotes the estimate of the effect of an SNP from GWAS data for BMD, and $\hat{\varepsilon}_{zx}$ is the estimate of the effect of an SNP on the gene expression level from an eQTL data. The Delta method can be utilized to calculate the sampling variance of $\hat{b}_{xy}$ approximately (Lynch and Walsh, 1998):

$$var\left(\hat{b}_{xy}\right) \approx \frac{b_{zy}^2}{\varepsilon_{zx}^2}\left[\frac{var\left(\hat{\varepsilon}_{zx}\right)}{\varepsilon_{zx}^2} + \frac{var\left(\hat{b}_{zy}\right)}{b_{zx}^2} - \frac{2cov\left(\hat{\varepsilon}_{zx}, \hat{b}_{zy}\right)}{\varepsilon_{zx} b_{zy}}\right] \tag{10}$$

where $cov\left(\hat{\varepsilon}_{zx}, \hat{b}_{zy}\right)$ is 0 when ε_{zx} and b_{zy} are derived from independent GWAS and eQTL datasets. Because the distribution of the Z-score is known, while the distributions of ε_{zx} and b_{zy} are

TABLE 1 | Ten of the top 20 significant genes and related study.

Gene	PubMed ID
DGKQ	PMID:30048462
FDFT1	PMID:25223561
Cdc42	PMID:29314205
LRP3	PMID:27019110
TMUB2	PMID:27019110
ASB16	PMID:32269995
RERE	PMID:18597038
MS4A6A	PMID:33604283
EPDR1	PMID:32619791
SPTBN1	PMID:19801982

unknown, T_{SMR} can be approximately denoted as:

$$T_{SMR} = \frac{b_{xy}^2}{var\left(\hat{b}_{xy}\right)} \approx \frac{Z_{zy}^2 Z_{zx}^2}{Z_{zy}^2 + Z_{zx}^2} \tag{11}$$

where Z_{zy} and Z_{zx} denotes the Z-score derived from GWAS and eQTL data. Since the distribution of T_{SMR} is $x^2 = 1$, the significance of b_{xy} can be calculated by performing a χ^2-test of T_{SMR}, which is also the significance of the association between gene and BMD.

SMR Analysis for Bone Mineral Density With Expression Quantitative Trait Loci Data From Blood Tissue

We first integrated two independent GWAS datasets to obtain a more comprehensive GWAS dataset. After obtaining β and SE of each SNP based on GWAS dataset and eQTL dataset, respectively, we obtained the estimate of effect size of SNPs on BMD based on integrated GWAS summary data and estimate of effect size on gene expression based on SNPs from eQTL data. We obtained two Z-scores of the same SNP based on two datasets, Z_{GWAS} and Z_{eQTL}. The GWAS dataset provides the SNPs associated to BMD, and the eQTL dataset provides the association between these SNPs and expression level of gene. Then the SMR method is utilized to examine the effect size of SNPs on BMD excluding some irrelevant factors.

Since one single SNP can regulate multiple genes, we then identify the causative genes, which are regulated by these SNPs and are associated with BMD. We performed a Bonferroni test to filter the SNPs we obtained for the SMR method. After all, we identified 12,477 SNPs, and 564 genes regulated by these SNPs

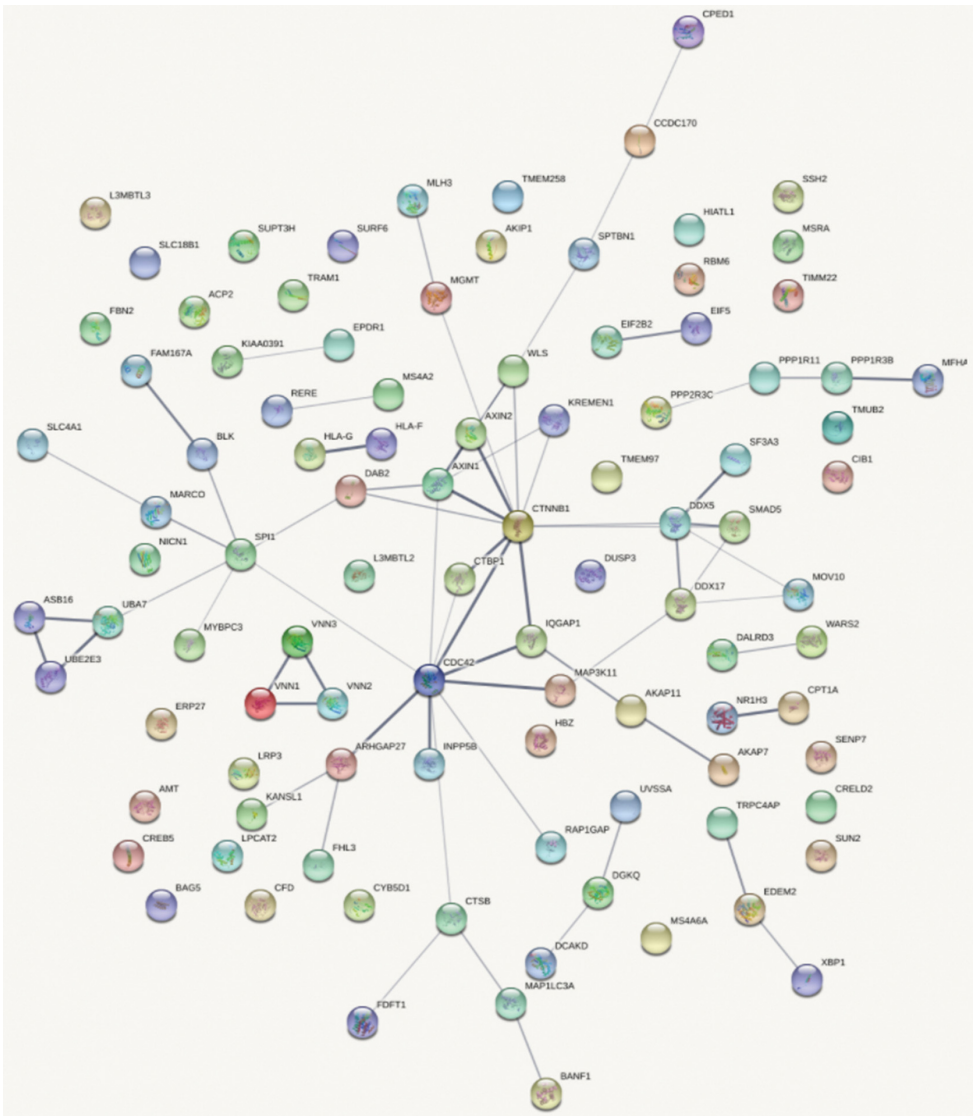

FIGURE 4 | Gene interaction network obtained from the top 100 genes.

are associated with BMD. It is clear from the result that most of the causative genes are regulated by multiple SNPs, which means detecting the disease-related genes merely depending on GWAS datasets is not reliable. The workflow is shown in **Figure 1**.

RESULTS

The result of the GWAS meta-analysis is shown in **Figure 2**. It is apparent that the original datasets from former studies are not consistent. After integration, we obtained a more precise GWAS dataset for BMD. Since there are many overlapping SNPs in the GWAS dataset and eQTL dataset, we have to filter these SNPs to find out whether the genes regulated by these SNPs are associated with BMD. Thus, the SMR method is utilized to examine latent associations between gene expression and BMD. The results of BMD-related genes based on GWAS and eQTL to test for the integrated data are shown in **Figure 3**. We identified, in total, 12,477 SNPs regulating 564 genes associated with BMD. This indicates that multiple SNPs may cooperate and effect the expression of a single gene. For example, gene FDFT1 is regulated by 451 SNPs, and most SNPs can regulate multiple genes as well, such as rs10085549, rs1073, and so on. They can regulate seven genes. **Supplementary Material** indicates the significant genes and SNPs related to BMD.

Case Study

As a result of the SMR method, we identified 12,477 significant SNPs and 564 significant genes associated with BMD. Several significant genes of the results have been reported in recent

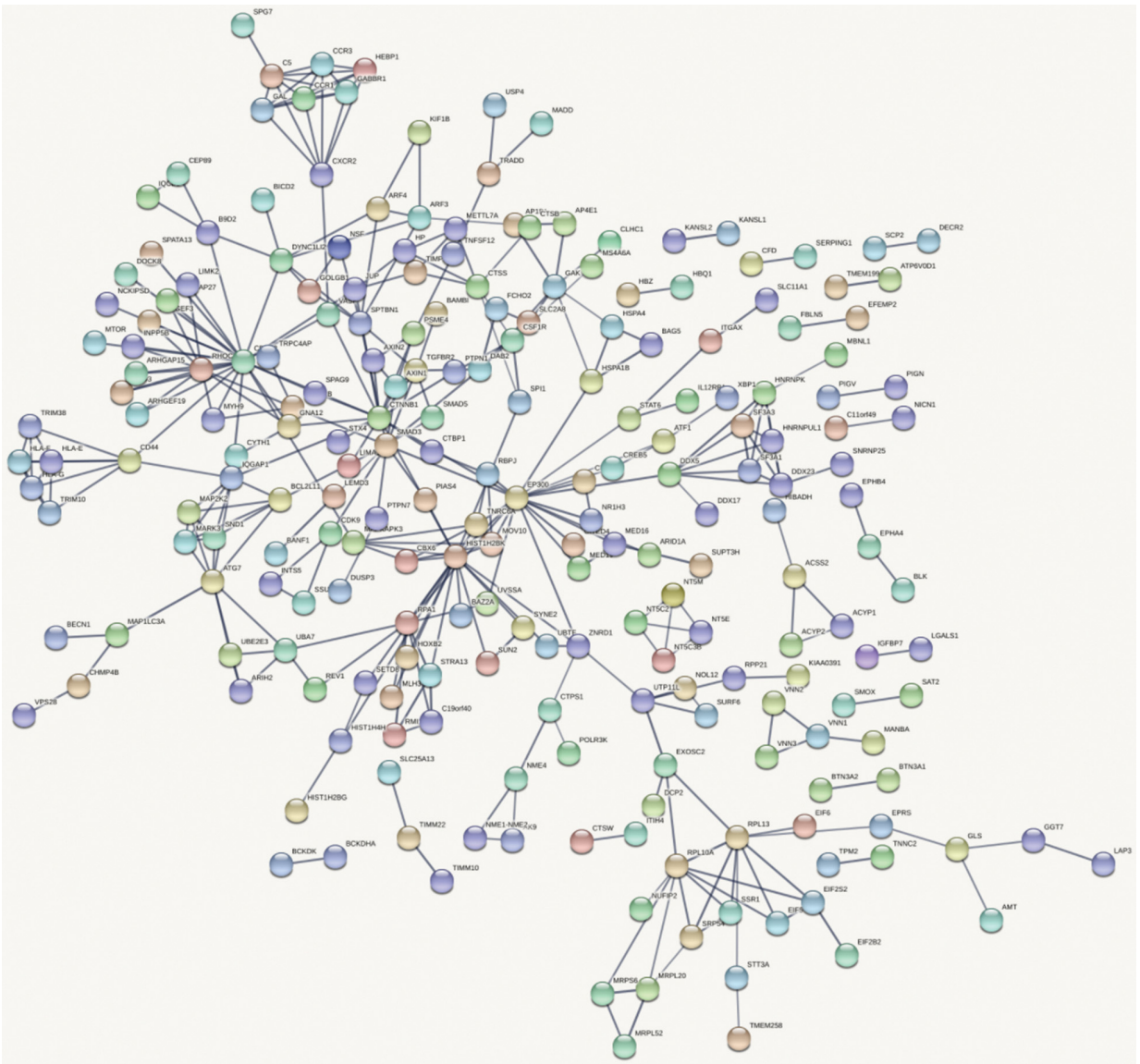

FIGURE 5 | Gene interaction network obtained from all significant genes.

studies. In the study of Kim (2008), they identified that gene DGKQ is associated with heel BMD. In the study of Wang et al. (2015) they have found the association between FDFT1 and the therapeutic response among Chinese postmenopausal women suffering from osteopenia or osteoporosis. Cdc42 is identified to be strongly related to bone deterioration in experimental osteoarthritis according to the study of Hu et al. (2018). LRP3, TMUB2 has also been reported as a risk factor for BMD of the lumbar spine (LS-BMD) (Zhu et al., 2016). RERE is reported to be a novel suspective gene associated with BMD from a group of Caucasian-origin families (Zhang et al., 2009). In total, there are 10 out of the top 20 significant genes in our results that have been reported to be related with BMD according to previous studies. **Table 1** shows these 10 genes and related GWAS studies published previously.

Gene Interaction Network Based on Bone Mineral Density

Figure 4 shows the top 100 gene interaction networks derived from the results of the SMR method on BMD. **Figure 5** shows the gene interaction network from all significant genes derived from the SMR method. Based on the top 100 gene interaction networks, Cdc42 and CTNNB1 are intensively interacted and significantly associated with BMD. It is known that the process of bone (re)modeling is based on the distinct actions of osteoclasts and osteoblasts, which are achieved by the organization of osteoclast cytoskeleton. Cdc42 belongs to the Rho GTPase subfamily, which is considered to be major regulators of cytoskeleton, and it has been reported to be a prospective therapeutic target for preventing osteoporosis (Ito et al., 2010). CTNNB1 has been reported to be related to BMD in the spine and hips (Estrada et al., 2012).

In total, we identified 12,477 SNPs and 564 genes related to BMD by the SMR method. Then we performed the case study of the identified genes to prove the effectiveness of our BMD-related gene identification method based on multiple omics data integration.

CONCLUSION

We use the SMR method to integrate omics data to identify BMD–gene associations. First, we integrated two independent GWAS data sets by adjusting the weights of SNPs to overcome that different GWAS datasets have different sample sizes. Then we reduced the impact of linkage disequilibrium and identified the impact of SNPs on BMD based on GWAS data and eQTL data. Through the Bonferroni test, we obtained 12,477 SNPs and 564 genes significantly related to BMD. Among these genes, 10 of the top 20 risk genes have been previously reported to be associated with BMD, which proves the validity of our method and the correctness of the results, but further biological experiments are needed to verify our results. Our results indicate that BMD is a highly inherited polygenic trait and is significantly associated with osteoporosis. These findings help us reveal the pathology of osteoporosis and determine the relevant pathways and therapeutic drugs.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. Written informed consent for participation was not required for this study in accordance with the National Legislation and the Institutional Requirements.

AUTHOR CONTRIBUTIONS

YL, GJ, and XW wrote the manuscript and did the experiments. FD provided ideas of this work. YL, GJ, and YD analyzed the data. All authors approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.728563/full#supplementary-material

Supplementary Table 1 | Predicted genes.

Supplementary Table 2 | Predicted SNPs.

REFERENCES

Arden, N., Baker, J., Hogg, C., Baan, K., and Spector, T. (1996). The heritability of bone mineral density, ultrasound of the calcaneus and hip axis length: a study of postmenopausal twins. *J. Bone Min. Res.* 11, 530–534. doi: 10.1002/jbmr.5650110414

Bauer, D., Ewing, S., Cauley, J., Ensrud, K., Cummings, S. R., and Orwoll, E. (2007). Quantitative ultrasound predicts hip and non-spine fracture in men: the MrOS study. *Osteop. Int.* 18, 771–777. doi: 10.1007/s00198-006-0317-5

Cauley, J. A., Hochberg, M. C., Lui, L.-Y., Palermo, L., Ensrud, K. E., Hillier, T. A., et al. (2007). Long-term risk of incident vertebral fractures. *JAMA* 298, 2761–2767. doi: 10.1001/jama.298.23.2761

Chalmers, T. C., Smith, H. Jr., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., et al. (1981). A method for assessing the quality of a randomized control trial. *Control. Clin. Trials* 2, 31–49. doi: 10.1016/0197-2456(81)90056-8

Deng, F.-Y., Lei, S.-F., Zhang, Y., Zhang, Y.-L., Zheng, Y.-P., Zhang, L.-S., et al. (2011). Peripheral blood monocyte-expressed ANXA2 gene is involved in pathogenesis of osteoporosis in humans. *Mol. Cell. Proteomics* 10:M111.011700. doi: 10.1074/mcp.M111.011700

Dubois, P. C., Trynka, G., Franke, L., Hunt, K. A., Romanos, J., Curtotti, A., et al. (2010). Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* 42, 295–302. doi: 10.1038/ng.543

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y.-H., Duncan, E. L., Ntzani, E. E., et al. (2012). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat. Genet.* 44, 491–501. doi: 10.1038/ng.2249

Farber, C. R. (2012). Systems genetics: a novel approach to dissect the genetic basis of osteoporosis,". *Curr. Osteopor. Rep.* 10, 228–235. doi: 10.1007/s11914-012-0112-5

Farber, C. R., and Lusis, A. J. (2008). Integrating global gene expression analysis and genetics. *Adv. Genet.* 60, 571–601. doi: 10.1016/s0065-2660(07)00420-8

Fehrmann, R. S., Jansen, R. C., Veldink, J. H., Westra, H.-J., Arends, D., Bonder, M. J., et al. (2011). Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* 7:e1002197. doi: 10.1371/journal.pgen.1002197

Geissmann, F., Manz, M. G., Jung, S., Sieweke, M. H., Merad, M., and Ley, K. (2010). Development of monocytes, macrophages, and dendritic cells. *Science* 327, 656–661. doi: 10.1126/science.1178331

Glüer, C. C., Eastell, R., Reid, D. M., Felsenberg, D., Roux, C., Barkmann, R., et al. (2004). Association of five quantitative ultrasound devices and bone densitometry with osteoporotic vertebral fractures in a population-based sample: the OPUS Study. *J. Bone Min. Res.* 19, 782–793. doi: 10.1359/jbmr.040304

Gonnelli, S., Cepollaro, C., Gennari, L., Montagnani, A., Caffarelli, C., Merlotti, D., et al. (2005). Quantitative ultrasound and dual-energy X-ray absorptiometry in the prediction of fragility fracture in men. *Osteop. Int.* 16, 963–968. doi: 10.1007/s00198-004-1771-6

Grundberg, E., Kwan, T., Ge, B., Lam, K. C., Koka, V., Kindmark, A., et al. (2009). Population genomics in a disease targeted primary cell model. *Genome Res.* 19, 1942–1952. doi: 10.1101/gr.095224.109

Grundberg, E., Small, K. S., Hedman, ÅK., Nica, A. C., Buil, A., Keildson, S., et al. (2012). Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* 44, 1084–1089. doi: 10.1038/ng.2394

Hu, X., Ji, X., Yang, M., Fan, S., Wang, J., Lu, M., et al. (2018). Cdc42 is essential for both articular cartilage degeneration and subchondral bone deterioration in experimental osteoarthritis. *J. Bone Min. Res.* 33, 945–958. doi: 10.1002/jbmr.3380

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., et al. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet.* 7:e1002078. doi: 10.1371/journal.pgen.1002078

Ito, Y., Teitelbaum, S. L., Zou, W., Zheng, Y., Johnson, J. F., Chappel, J., et al. (2010). Cdc42 regulates bone modeling and remodeling in mice by modulating RANKL/M-CSF signaling and osteoclast polarization. *J. Clin. Invest.* 120, 1981–1993. doi: 10.1172/JCI39650

Kemp, J. P., Morris, J. A., Medina-Gomez, C., Forgetta, V., Warrington, N. M., Youlten, S. E., et al. (2017). Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. *Nat. Genet.* 49, 1468–1475. doi: 10.1038/ng.3949

Kim, S. K. (2018). Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. *PLoS One* 13:e0200785. doi: 10.1371/journal.pone.0213962

Kwan, T., Grundberg, E., Koka, V., Ge, B., Lam, K. C., Dias, C., et al. (2009). Tissue effect on genetic control of transcript isoform variation. *PLoS Genet.* 5:e1000608. doi: 10.1371/journal.pgen.1000608

Lee, M., Czerwinski, S., Choh, A., Demerath, E., Sun, S., Chumlea, W., et al. (2006). Unique and common genetic effects between bone mineral density and calcaneal quantitative ultrasound measures: the fels longitudinal study. *Osteopor. Int.* 17, 865–871. doi: 10.1007/s00198-006-0075-4

Liu, Y.-Z., Dvornyk, V., Lu, Y., Shen, H., Lappe, J. M., Recker, R. R., et al. (2005). A novel pathophysiological mechanism for osteoporosis suggested by an in vivo gene expression study of circulating monocytes. *J. Biol. Chem.* 280, 29011–29016. doi: 10.1074/jbc.M501164200

Lynch, M., and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer.

Manabe, N., Kawaguchi, H., Chikuda, H., Miyaura, C., Inada, M., Nagai, R., et al. (2001). Connection between B lymphocyte and osteoclast differentiation pathways. *J. Immunol.* 167, 2625–2631. doi: 10.4049/jimmunol.167.5.2625

Moayyeri, A., Hsu, Y.-H., Karasik, D., Estrada, K., Xiao, S.-M., Nielson, C., et al. (2014). Genetic determinants of heel bone properties: genome-wide association meta-analysis and replication in the GEFOS/GENOMOS consortium. *Hum. Mol. Genet.* 23, 3054–3068. doi: 10.1093/hmg/ddt675

Nattiv, A. (2000). Stress fractures and bone health in track and field athletes. *J. Sci. Med. Sport* 3, 268–279. doi: 10.1016/S1440-2440(00)80036-5

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6:e1000888. doi: 10.1371/journal.pgen.1000888

Peng, J., and Zhao, T. (2020). Reduction in TOM1 expression exacerbates Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3915–3916. doi: 10.1073/pnas.1917589117

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1038/nature08454

Smith, G. D., and Ebrahim, S. (2008). "Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. in *Biosocial Surveys*," in *National Research Council (US) Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys*, eds M. Weinstein, J. W. Vaupel, and K. W. Wachter (Washington, DC: National Academies Press), 336–366.

Tarwadi, T., Jazayeri, J. A., Pambudi, S., Arbianto, A. D., Rachmawati, H., Kartasasmita, R. E., et al. (2020). In-silico molecular interaction of short synthetic lipopeptide/importin-alpha and in-vitro evaluation of transgene expression mediated by liposome-based gene carrier. *Curr. Gene Ther.* 20, 383–394. doi: 10.2174/1566523220666201005104224

Tianyi, Z., Yang, H., Valsdottir, L. R., Tianyi, Z., and Jiajie, P. (2021). Identifying drug–target interactions based on graph convolutional network and deep neural network. *Brief. Bioinform.* 22, 2141–2150. doi: 10.1093/bib/bbaa044

Wang, C., Zheng, H., He, J., Zhang, H., Yue, H., Hu, W., et al. (2015). Genetic polymorphisms in the mevalonate pathway affect the therapeutic response to alendronate treatment in postmenopausal Chinese women with low bone mineral density. *Pharmacogenom. J.* 15, 158–164. doi: 10.1038/tpj.2014.52

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243. doi: 10.1038/ng.2756

Wu, Q., Nasoz, F., Jung, J., Bhattarai, B., Han, M. V., Greenes, R. A., et al. (2021). Machine learning approaches for the prediction of bone mineral density by using genomic and phenotypic data of 5130 older men. *Sci. Rep.* 11, 4482. doi: 10.1038/s41598-021-83828-3

Zhang, H., Sol-Church, K., Rydbeck, H., Stabley, D., Spotila, L., and Devoto, M. (2009). High resolution linkage and linkage disequilibrium analyses of chromosome 1p36 SNPs identify new positional candidate genes for low bone mineral density. *Osteopor. Int.* 20, 341–346. doi: 10.1007/s00198-008-0668-1

Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* 22:bbaa212. doi: 10.1093/bib/bbaa212

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, T., Hu, Y., Zang, T., and Cheng, L. (2020c). MRTFB regulates the expression of NOMO1 in colon. *Proc. Natl. Acad. Sci.* 117, 7568–7569. doi: 10.1073/pnas.2000499117

Zhao, T., Hu, Y., Zang, T., and Wang, Y. (2019). Integrate GWAS, eQTL, and mQTL data to identify Alzheimer's disease-related genes. *Front. Genet.* 10:1021. doi: 10.3389/fgene.2019.01021

Zhao, T., Liu, J., Zeng, X., Wang, W., Li, S., Zang, T., et al. (2021a). Prediction and collection of protein–metabolite interactions. *Brief. Bioinform.* 2021:bbab014. doi: 10.1093/bib/bbab014

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., et al. (2021b). SC2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Res.* 49, D1413–D1419. doi: 10.1093/nar/gkaa838

Zheng, H. F., Forgetta, V., Hsu, Y. H., Estrada, K., Rosello-Diez, A., Leo, P. J., et al. (2015). Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. *Nature* 526, 112–117. doi: 10.1038/nature14878

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

Zhuang, H., Zhang, Y., Yang, S., Cheng, L., and Liu, S. L. (2019). A mendelian randomization study on infant length and type 2 diabetes mellitus risk. *Curr. Gene Ther.* 19, 224–231. doi: 10.2174/1566523219666190925115535

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 October 2021
doi: 10.3389/fgene.2021.762136

Identification of miRNA Signature Associated With Erectile Dysfunction in Type 2 Diabetes Mellitus by Support Vector Machine-Recursive Feature Elimination

Haibo Xu[1,2], *Baoyin Zhao*[2], *Wei Zhong*[2], *Peng Teng*[2] *and Hong Qiao*[1]*

[1]The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2]The First Hospital of Qiqihar, Qiqihar, China

Diabetic mellitus erectile dysfunction (DMED) is one of the most common complications of diabetes mellitus (DM), which seriously affects the self-esteem and quality of life of diabetics. MicroRNAs (miRNAs) are endogenous non-coding RNAs whose expression levels can affect multiple cellular processes. Many pieces of studies have demonstrated that miRNA plays a role in the occurrence and development of DMED. However, the exact mechanism of this process is unclear. Hence, we apply miRNA sequencing from blood samples of 10 DMED patients and 10 DM controls to study the mechanisms of miRNA interactions in DMED patients. Firstly, we found four characteristic miRNAs as signature by the SVM-RFE method (hsa-let-7E-5p, hsa-miR-30 days-5p, hsa-miR-199b-5p, and hsa-miR-342–3p), called DMEDSig-4. Subsequently, we correlated DMEDSig-4 with clinical factors and further verified the ability of these miRNAs to classify samples. Finally, we functionally verified the relationship between DMEDSig-4 and DMED by pathway enrichment analysis of miRNA and its target genes. In brief, our study found four key miRNAs, which may be the key influencing factors of DMED. Meanwhile, the DMEDSig-4 could help in the development of new therapies for DMED.

Keywords: micrornas, diabetes mellitus, erectile dysfunction, signature, molecular mechanisms

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Hui Ding,
University of Electronic Science and Technology of China, China
Juan Wang,
Inner Mongolia University, China

***Correspondence:**
Hong Qiao
qiaohong@hrbmu.edu.cn

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

Received: *21 August 2021*
Accepted: *22 September 2021*
Published: *11 October 2021*

Citation:
Xu H, Zhao B, Zhong W, Teng P and Qiao H (2021) Identification of miRNA Signature Associated With Erectile Dysfunction in Type 2 Diabetes Mellitus by Support Vector Machine-Recursive Feature Elimination. Front. Genet. 12:762136. doi: 10.3389/fgene.2021.762136

INTRODUCTION

Erectile dysfunction (ED) refers to the persistent or repeated failure of men to achieve and/or maintain penile erection for satisfactory sexual activity. As a common and the most neglected complication of diabetes (Zhao et al., 2020; Long et al., 2021; Yang et al., 2021), diabetic mellitus erectile dysfunction (DMED) is an important factor affecting psychological well-being, spousal relationship and family life (Malavige and Levy, 2009). The massive research indicated patient of T2MD incidence ED was significantly higher than that of the health. 75%male with diabetes is affected with ED. 66.3% is T2MD among of the data (Kouidrat et al., 2017; Cheng et al., 2018; Zagidullin et al., 2019; Zhu et al., 2021a). DMED is considered as an alternative marker for diabetes and cardiovascular disease, and is the primary feature of diabetes. Meanwhile, DMED has a multifactorial pathological process that can occur simultaneously with cardiovascular disease, neuropathy, and depression. How to effectively intervene in DMED has become an urgent problem in the global medical community.

MicroRNAs (miRNAs) are small non-coding RNAs of 19–25 nucleotides (Cheng et al., 2016; Lu and Rothenberg, 2018; Wu et al., 2019; Mo et al., 2020). They have been used as important regulators of gene expression in recent decades. Changes in their expression levels can affect multiple cellular processes and are used as molecular markers for diagnosis and follow-up (Han et al., 2021). It is widely involved in pathological processes such as cancer (Rupaimoole and Slack, 2017; Liu et al., 2021a; Lei and Shu-Lin, 2021; Sheng et al., 2021; Tang et al., 2021), DM (Vasu et al., 2019), cardiovascular events (Barwari et al., 2016), and ED (Ding et al., 2017). However, there are few related studies in DMED.

Growing evidences have indicated that miRNAs play an important role in the occurrence and development of DM and diabetic complications (Kong et al., 2011; Jiang et al., 2015). Unfortunately, the exact pathogenesis of miRNAs action on DMED remains to be largely unknown. Hence, we adopted the machine learning method (SVM-RFE), identified the characteristic miRNAs of DMED, constructed the signature of DMED and found potentially related pathways. Our work has significance for the identification of the molecular mechanism and the early prediction and diagnosis of DMED.

MATERIALS AND METHODS

Subjects

Inclusion criteria: T2DM patients admitted to Qiqihar Medical College from December 2020 to June 2021 were selected as the study subjects. Erectile dysfunction was identified by international index of erectile function -5 (IIEF-5). Diagnostic criteria for T2DM: in line with WHO diabetes diagnostic criteria in 1999; Diagnostic and grading criteria for ED: 1) Regular sexual partner and normal sexual life; 2) History of ED for more than 6 months; 3) Erectile dysfunction was assessed according to IIEF-5, and the total score ≤21 was divided into ED; A score of five to seven was severe, 8–11 was moderate, 12–21 was mild, and ≥22 was no ED. Healthy married men with erectile dysfunction aged 30–70 years with a course of 2–10 years were included in the study.

Exclusion criteria: 1) The informed consent is not signed or the medical records are incomplete; 2) Incomplete research data; 3) Type 1 diabetes mellitus (T1DM), adult latent autoimmune diabetes mellitus (LADA), acute complications of diabetes mellitus; 4) Hypogonadism, thyroid disease, adrenal disease, pituitary disease, etc.; 5) Pelvic and urinary genital malformations, inflammation, tumor, trauma, surgical history; 6) Serious blood system, cardiovascular, liver, kidney disease or other disease affecting sexual activity; 7) Spinal cord injury; 8) Smoking, alcohol, drug abuse and masturbation history; 9) A history of drug abuse; (10) Receiving ED treatment or drugs that affect ED; Such as immunosuppressants, glucocorticoids, diuretics, receptor blockers, antioxidants, etc.; 11) History of mental illness, and the ED caused by anxiety, depression and other psychological factors was excluded according to the SAS standard score <50 and SDS standard score <53.

According to the above criteria, a total of 20 male T2DM patients were included and divided into DMED group (10 cases) and DM group (10 cases). The study was conducted in accordance with the Declaration of Helsinki,and with approval from The first hospital of Qiqihar ethics committee for clinical trials (2020-KY-007–01).Written informed consent was obtained from all the participants.

Clinical Data of Patients

General information such as name, age, sex, height, weight, marital history, personal history, infection history, surgery and trauma history were collected. IIEF-5 scores, SAS scale and SDS scale were obtained using a questionnaire. Body mass index (BMI) was calculated by formula: BMI = Body weight/height^2 (kg/m^2). The remaining indicators were determined by clinical tests.

Collection of Serum Samples

Clinical serum samples were collected by utilizing residual specimens from patients undergoing routine medical care. Each blood sample is 4–6 ml. Samples were then centrifuged and the supernatant was stored at −80°C in the centrifuge tube.

Sample Sequencing

MiRNA sequencing was performed on a total of 20 cases, 10 cases in each group with no statistical difference in age, course of disease and BMI.

The Method of Sample Detecting

Use Agilent 2100 Bioanalyzer to test sample integrity and concentration, and NanoDrop to Inorganic ions orpolycarbonate contamination. This step aimed to provide a reference for library construction and later analysis.

Library Construction

Filter Small RNA: Use the 200ng-1 μg of RNA sample, then separate RNA segment of different size by PAGE gel, select 18–30 nt (14–30 ssRNA Ladder Marker, TAKARA) stripe and recycle; Adaptor ligation: Prepare connection 3′adaptor system (Reaction condition:70°C for 2min; 25°C for 2 h); Secondly add RT-Primer, (Reaction condition: 65°C for 15 min; ramp to 4°C at a rate of 0.3°C/s); Thirdly add 5′adaptor mix system (Reaction condition: 70°C for 2 min; 25°C for 1 h).

RT PCR: Prepare First Strand Master Mix and Super Script II (Invitrogen) reverse transcription (Reaction condition: 42°C for 1 h; 70°C for 15 min); Several rounds of PCR amplification with PCR Primer Cocktail and PCR Master Mix were performed to enrich the cDNA fragments (Reaction condition: 95°C for 3 min; 15–18 cycles of (98°C for 20 s, 56°C for 15 s, 72°C for 15 s); 72°C for 10 min; 4°C hold); Purify PCR products: Then the PCR products were purified with PAGE gel, dissolve the recycled products in EB solution.

Circularization

The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA (ssCir DNA) were formatted as the final library.

Library Quality Control

Library was validating on the Agilent Technologies 2100 bioanalyzer.

Sequencing

The library was amplified with phi29 to make DNA nanoball (DNB) which have more than 300 copies of one molecular. The DNBs were loaded into the patterned nanoarray and single end 50 bases reads were generated in the way of combinatorial Probe-Anchor Synthesis (cPAS).

Feature Selection of Diabetic Mellitus Erectile Dysfunction Based on Support Vector Machine-Recursive Feature Elimination

Support Vector Machine-Recursive Feature Elimination (SVM-RFE) is a sequence backward selection algorithm based on the maximum interval principle of Support Vector Machine (SVM) (Guyon et al., 2002; Tang et al., 2018; Cheng et al., 2019; Yang et al., 2020; Liu et al., 2021b; Joshi et al., 2021).

The counts data based on miRNA sequencing were combined with the improved SVM-REF method proposed by Kai-Bo Duan et al. (Duan et al., 2005) to select miRNAs. Due to the randomness in Kai-Bo Duan's method, in order to obtain a model with relatively small error, the whole process was repeated for 1,000 times. The model with the smallest error was selected as the final model. If there were multiple models with the same minimum error, the model with a large number of miRNAs was selected. The characteristic miRNAs, called DMEDSig-4, were finally screened.

Differentially Expressed

According to the identified DMEDSig-4, we combined the miRNA read counts data calculated by the sequencing process. In view of the negative binomial distribution of counts data, we used the R package DESeq2 to calculate the differences between the DM groups and DEMD groups (Love et al., 2014).

Identification of miRNA Target Genes

The miRNA-targeted mRNAs of DMEDSig-4 were pooled using the online bioinformatics analysis tool (EncoRI). Firstly, we searched EncoRI database (http://starbase.sysu.edu.cn/) (Li et al., 2014), which included seven databases (microT, miRanda, miRmap, PITA, RNA22, PicTar and TargetScan). Then, we entered characteristic miRNA of DMEDSig-4, set CLIP-Data≥5, Program-Number≥5, and Degradome-Data≥1 to obtain miRNA target genes.

Analysis of Gene Function Enrichment Regulatory Network

The R package FGNet allows functional enrichment analysis (FEA) to be performed on a list of genes or expression sets and the results to be converted into a network (Aibar et al., 2015; Azimi et al., 2021). The network can provide an overview of the biological function of genes/terms, and allows easy seeing of links between genes, the overlap between clusters, etc. We selected the annotation tool topGO for functional annotation of target genes. GO was used to describe gene functions along with three aspects: biological process (BP), cellular component (CC) and molecular function (MF). The $p < 0.01$ was considered significant.

TABLE 1 | Comparisons of clinical data. BMI, body mass index; IIEF-5, international index of erectile function 5; SAS, self-rating anxiety scale; SDS, self-rating depression scale; FPG, fasting plasma glucose; HbA1c, glycated hemoglobin; TC, total cholesterol; TG, triglyceride; TT, testosterone; TSH, thyroid stimulating hormone; Scr, serum creatinine; Urea, carbamide; ALT, alanine aminotransferase; AST, aspartate aminotransferase.

Parameter	Group (DMED)	Group (DM)	*p* Value
Age, years	48.7 ± 5.03	46.6 ± 6.19	0.416
Diabetes duration, years	4 (2–8)	2 (2–5)	0.147
BMI, kg/m^2	26.19 ± 3.08	26.71 ± 3.39	0.725
IIEF-5 score	12.30 ± 4.85	23.40 ± 1.07	<0.0001***
SAS score	41.7 ± 4.9	40 ± 6.5	0.517
SDS score	48.5 (46–51)	50 (50–51)	0.244
FPG, mmol/l	12.61 (8.29–14.11)	9.46 (8.31–10.02)	0.364
HbA1c, %	9.75 ± 2.63	8.56 ± 1.59	0.236
TC, mmol/l	5.54 ± 1.15	4.87 ± 1.33	0.243
TG, mmol/l	2.4 ± 1.41	2.99 ± 2.09	0.468
TT, ng/ml	4.08 ± 1.11	3.87 ± 1.01	0.658
TSH, uIU/ml	2.21 ± 0.69	1.75 ± 0.95	0.225
Scr, umol/L	70.8 ± 11.03	63 ± 6.25	0.067
UREA, mmol/L	5.88 ± 1.24	6.03 ± 1.43	0.813
ALT	18.8 (16.5–21.4)	25.15 (19.8–34.5)	0.059
AST	16.18 ± 3.92	21 ± 8.93	0.136

RESULTS

Statistical Analysis of Clinical Data

Firstly, we collected 20 high-quality samples (10 DMED and 10 DM) from 60 patients according to the inclusion criteria. Then, we collected and collated the clinical information of these 20 samples. Subsequently, statistical analysis was performed for the DMED group and the DM group, including age, diabetes duration, BMI, fasting plasma glucose, glycated hemoglobin, total cholesterol, triglyceride, testosterone, thyroid stimulating hormone, serum creatinine, carbamide, alanine aminotransferase, aspartate aminotransferase. Meanwhile, we also conducted a questionnaire survey on these 20 patients, and obtained IIEF-5 score, self-rating anxiety scale (SAS) score and self-rating depression scale (SDS) score. In this project, a total of 20 samples were tested using DNBSEQ platform. The average ratio of sample to genome was 78.22%. A total of 1,044 small RNAs were detected.

SAS9.4 international standard statistical programming software was used for statistical analysis. Measurement data processing normal distribution and variance homogeneity were measured, The measurement data conforming to normal distribution are expressed by mean ± standard deviation,comparison between two groups with sample t tests,non normal distribution adopt median and IQR to express, makes the non-parametric test. $p < 0.05$ could be

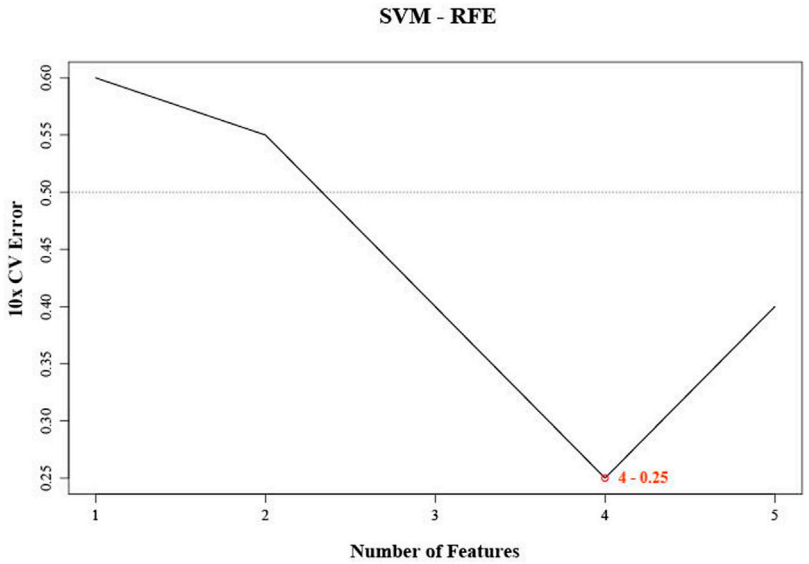

FIGURE 1 | SVM-RFE were used for feature selection. The point highlighted indicates the lowest error rate, and the corresponding miRNA at this point are the best signature selected by SVM-RFE.

considered statistically significant (**Table 1**). Finally, according to the results, there were no statistically significant differences with regard to age, diabetes duration and BMI ($p > 0.05$). In addition, there was no significant difference in other statistical indicators ($p > 0.05$) except IIEF-5 ($p < 0.05$). This showed that we chose samples to avoid the influence of other factors as much as possible.

Construction of microRNAs Signature Associated With Diabetic Mellitus Erectile Dysfunction

Many pieces of evidence have shown that miRNAs have diagnostic and predictive value for DMED in mice (Li et al., 2017; Cong et al., 2020; Huo et al., 2020). However, few correlational researches of DMED have been conducted on the human body. In order to further explore the mechanism of miRNAs regulating DMED in the human body, we selected 10 DM groups and 10 DMED groups for miRNA sequencing based on the inclusion criteria. The miRNA counts data of 20 samples were obtained. Firstly, miRNA counts data were filtered to delete miRNAs with counts of 0 in most samples. The selected miRNAs should have counts of non-0 in at least 18 samples. Due to the limited sample size of miRNA expression profiles in DM groups and DMED groups, we adopted the machine learning method of SVM-RFE to screen characteristic miRNAs (Sanz et al., 2018).

Firstly, the miRNA counts data contained all miRNAs that will be imported. Secondly, the algorithm used SVM model training samples to calculate the weight of each miRNA. Subsequently, we ranked miRNAs according to their weights and deleted the bottom-ranked miRNAs from the subset (Lin et al., 2017). Meanwhile, the remaining miRNAs were used to train the model again for the next iteration. Finally, the required number of miRNAs were selected. The later the miRNA was removed from the subset, the more significant the miRNAs were. We employed SVM-RFE machine learning method. Firstly, SVM-RFE method can be used for linear/nonlinear classification as well as regression with low generalization error rate. That is to say, he has a good learning ability, and the results of learning have a good extension. Secondly, SVM-RFE method can solve the problem of machine learning in the case of small samples, solve the problem of high dimension, and avoid the problem of neural network structure selection and local minimum point. Finally, SVM-RFE method is the best off-the-shelf classifier and can get a low error rate. Meanwhile, SVM-RFE method can make good classification decisions for data points outside the training set.

Since SVM-REF was more sensitive to feature changes, the ranking of features was different each time. For the robustness of feature selection, we refer to the method of Kai-Bo Duan (Duan et al., 2005). We used ten-fold cross-validation here by adding resampling to each iteration to stabilize the ranking (Zhu et al., 2021b). After 1,000 cycles of the algorithm (**Supplementary Table S1**), four characteristic miRNAs (hsa-let-7E-5p, hsa-miR-30 days-5p, hsa-miR-199b-5p, and hsa-miR-342–3p) were obtained according to the lowest error rate of 0.25 in ten-fold cross-validation (**Figure 1**). We found that the error rate decreased significantly from the first to the fourth feature number, and then increased significantly from the fifth. Obviously, the feature number of four had the best differentiation between DMED groups and DM groups. These four miRNAs (hsa-let-7E-5p, hsa-miR-30 days-5p, hsa-miR-199b-5p, and hsa-miR-342–3p) obtained by machine learning methods had the best classification performance in the DMED and DM groups. Therefore, we referred to this predictive

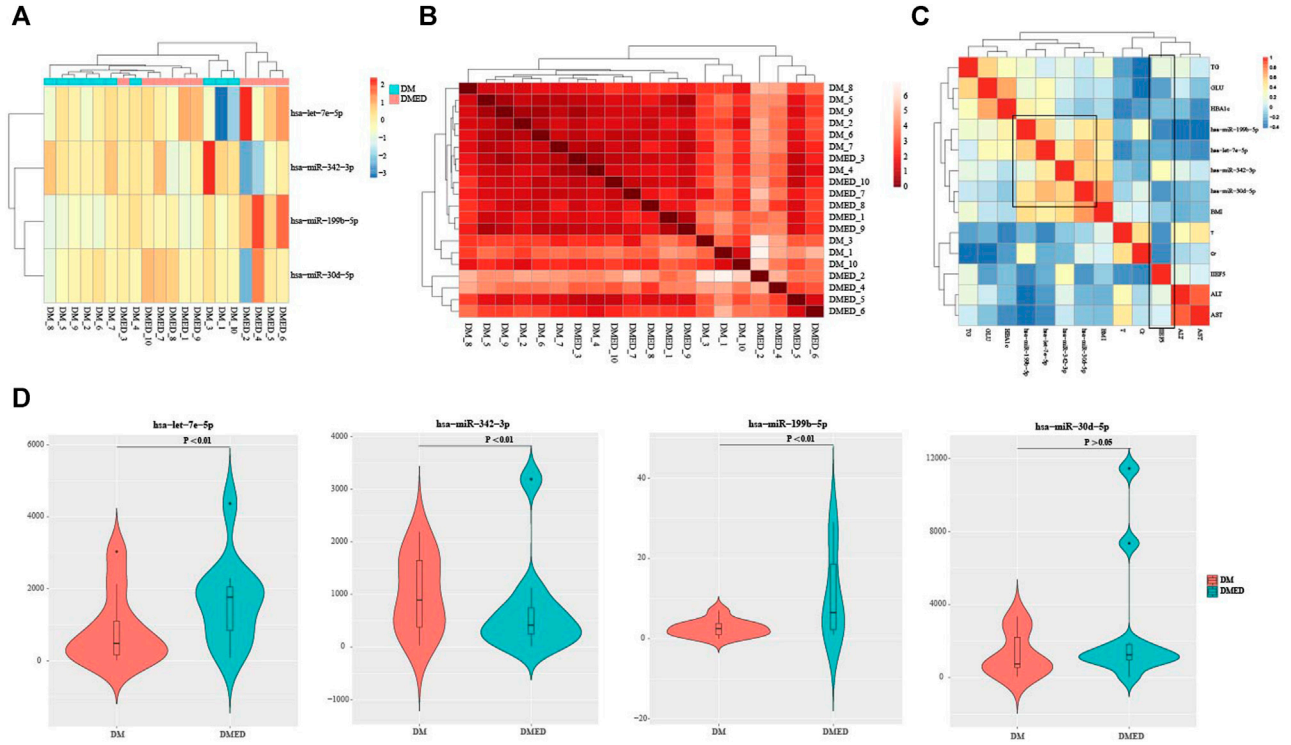

FIGURE 2 | The verification of DMEDSig-4 classification ability and the association of clinical information. **(A)** Clustering heat map of DMEDSig-4 in DM and DMED groups. Red: Up-regulated expression; Blue: Down-regulated expression. **(B)** Heat map of similarity between DM groups and DMED groups. **(C)** Correlation analysis of DMEDSig-4 and clinical factors. BMI, body mass index, IIEF5, international index of erectile function 5; HbA1c, glycated hemoglobin; TT, testosterone; GLU, glucose; Cr, creatinine; TG, triglyceride; ALT, Alanine transaminase; AST, aspartic acid transaminase. **(D)** Violin diagram of differential expression of DMEDSig-4 in DM and DMED samples.

signature as DMEDSig-4. In the following results, the performance of DMEDSig-4 was verified and analyzed.

Classification Capability Verification and Clinical Association Analysis of Diabetic Mellitus Erectile DysfunctionSig-4

In order to investigate the expression patterns of four characteristic miRNAs predicted by DMEDSig-4, we first conducted hierarchical clustering according to counts data of these four miRNAs and plotted the clustering heat map of DMEDSig-4 in 20 samples (**Figure 2A**). The results of hierarchical clustering analysis showed that the 20 samples were clustered into three clusters by DMEDSig-4. The two clusters on the right were completely composed of samples from the DM or DMED groups. Although the cluster on the left was a mixture of the two types of samples, except for DMED_3, DM and DMED were significantly clustered into the same cluster. This phenomenon indicated that all samples were characterized by the DMEDSig-4 expression pattern. In addition, we also calculated the similarity between samples according to Euclidean distance (**Figure 2B**). The results showed that the Euclidean distance between some samples was very small, indicating these samples had a high similarity. For example, samples DM_8, DM_5, DM_9, DM_2, DM_6 and DM_7. Samples DMED_10, DMED_7, DMED_8, DMED_1 and DMED_9.

In order to further explore the correlation between DMEDSig-4 and clinical factors, we calculated the Spearman correlation between miRNAs and clinical indicators of all samples (**Figure 2C**, **Supplementary Table S2**). It could be observed that the four miRNAs in DMEDSig-4 were positively correlated with each other, indicating that there might be a mechanism of co-operative regulation of DMEDSig-4. In general, the diagnosis of ED depends on the history of disease and the IIEF-5 (Rosen et al., 1997; Hatzichristou et al., 2002). It could be observed that IIEF-5 was negatively correlated with DMEDSig-4. Meanwhile, IIEF-5 was negatively correlated with testosterone, glycated globin, creatinine, fasting blood glucose and other indicators. Previous studies had verified that testosterone was a protective factor in DMED (Diaz-Arjonilla et al., 2009), and the negative correlation between testosterone and indicators confirmed the correctness of IIEF-5 data.

In order to verify the classification ability of the four characteristic miRNAs in DMEDSig-4 for samples, we calculated the significant difference of miRNA expression between the DM groups and the DMED groups. The results showed that the differences of hsa-let-7e-5p, hsa-miR-199b-5p and hsa-miR-342–3p were significant ($p < 0.01$), while the effect of hsa-miR-30 days-5p was not significant ($p > 0.05$) (**Figure 2D**).

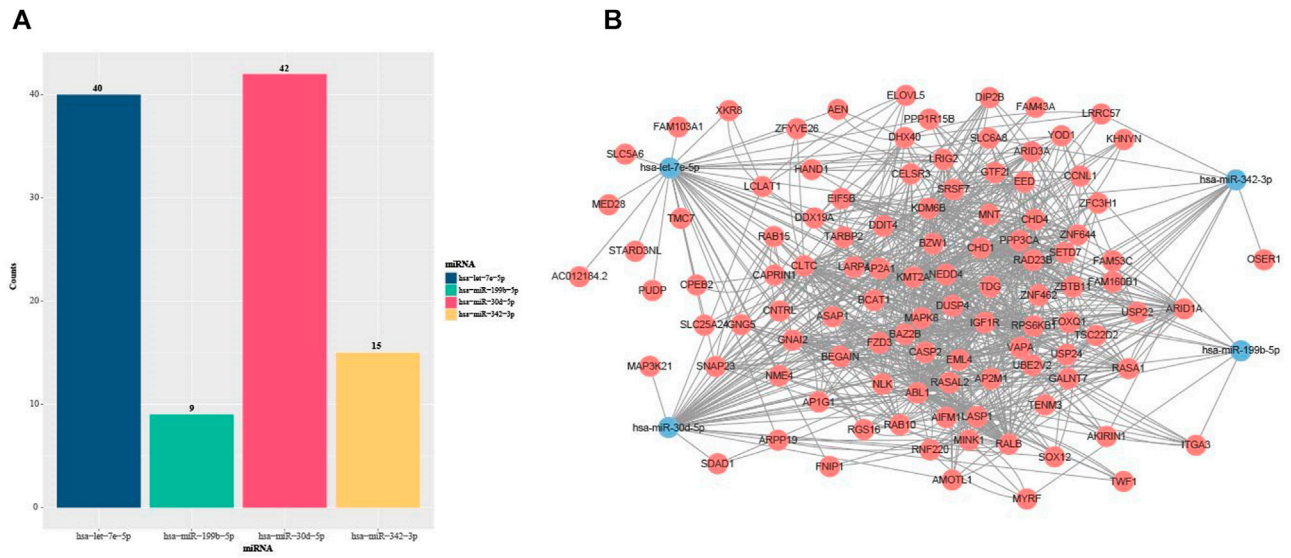

FIGURE 3 | Statistics of miRNA target miRNAs in DMEDSig-4. **(A)** Bar chart of DEMDSig-4 target genes **(B)** miRNA-mRNA interaction network diagram of DEMDSig-4. Red: target mRNAs; Blue: miRNAs of DMEDSig-4.

These results indicated that the miRNAs we identified have certain characteristics between the DM and DMED groups, regardless of the individual miRNA level or the overall expression pattern, and there were also certain correlations with clinical indicators. These showed that DMEDSig-4 had a certain role in helping to identify DMED.

Identification of Target microRNAs of Diabetic Mellitus Erectile DysfunctionSig-4

Many pieces of evidence demonstrated that miRNAs were capable of regulating various biological and pathological processes via inhibiting target mRNA translation or promoting mRNA degradation (Fabian et al., 2010; Fang et al., 2019; Riaz and Li, 2019; Wang et al., 2020). Meanwhile, the miRNAs could act as signatures of disease, strong indicators of prognosis or potential therapeutic targets (Zhu et al., 2011; Huang, 2018; Norsworthy et al., 2020). In order to further study the effects of the DMEDSig-4 target mRNAs on ED, we used the online bioinformatics database ENCORI (http://starbase.sysu.edu.cn/) to identify the target mRNAs. According to the parameters set, the target genes corresponding to each miRNA were obtained. The results of predicting targeted mRNAs on four miRNAs showed that there were 40 targeted mRNAs for hsa-let-7e-5p, 42 targeted mRNAs for hsa-miR-30 days-5p, nine targeted mRNAs for hsa-miR-199b-5p, and 15 targeted mRNAs hsa-miR-342–3p (**Figure 3A**).

In order to observe the relationship between miRNAs and target mRNAs, we constructed the miRNA-mRNA interaction network diagram based on DMEDSig-4 and target mRNAs through STRING database (https://string-db.org/) (**Figure 3B**). The network consisted of four miRNA nodes, 105 mRNA nodes and 870 edges. Among them, 764 edges were the relationship between target mRNAs, and 106 edges were the relationship between miRNA and target mRNAs. There was only one intersection among miRNA target mRNAs, indicating that these miRNAs did not tend to jointly regulate a target mRNA, and the identified target mRNAs had a close interaction relationship, suggesting that these mRNAs might act together on the same pathway.

Functional Annotation of microRNAs and Its Target microRNAs

In order to investigate the function of DEMDSig-4 targeted mRNAs, we used the topGO (Yang et al., 2021) annotation tool in R package FGNet for functional annotation of DEMDSig-4. We aimed to the functional analysis of each target mRNA searching further to verify the characteristic function of DEMDSig-4. The resulting network represented the links and associations between clusters of mRNAs and enriched terms. We annotated the biological process (BP), cellular component (CC) and molecular function (MF) of the target mRNAs. A total of 255 clusters and descriptions, we provided in the form of supplementary files (**Supplementary Table S3**). Here, we focused on the biological process of the target mRNAs. Due to the large number of biological processes and the complexity of the network, we manually selected representative biological processes for demonstration, including most of the mRNAs targeted for DMEDSig-4 (**Figure 4A**). The biological process includes: Regulation of nitrogen metabolic Process non-canonical Wnt signaling Pathway Neuron Projection Development “Wnt Signaling Pathway, Planar Cell Polarity Pathway” Regulation of Neuron projection Development, And “Density Lipoprotein receptor particle metabolic Process”. We found that the subnetwork was divided into two broad functional categories, including metabolic function and neural

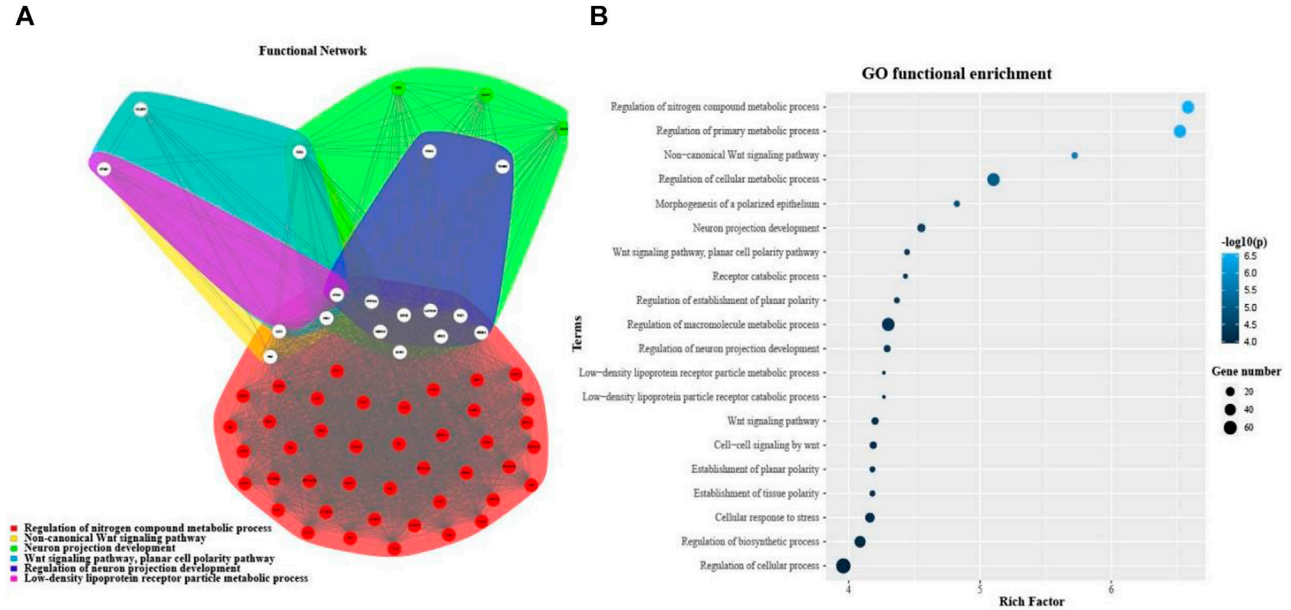

FIGURE 4 | Functional annotation of DMEDSig-4. **(A)** Functional Gene Network (FGN) of clusters defined by topGO for DMEDSig-4 Targeted Genes. Edges: link genes annotated under the same enrichment terms. Node colors: functional clusters containing genes. White nodes: hub genes expressed in multiple clusters. **(B)** Bubble chart of function enrichment results of DMEDSig-4.

function, which were closely related to the pathological mechanism of ED (Müller and Mulhall, 2006; Shamloul and Ghanem, 2013).

Although FGNet provided a broad overview of the biological effects of human-specific genetic alterations by clustering functional terms within clusters and establishing relationships between such clusters, it lacked the detail that could be obtained by analyzing each functional category individually (Bitar et al., 2019; Zhang et al., 2021). Therefore, we turned to gene ontology (GO) analysis. The top 20 salient biological processes were functionally annotated in terms of *p* values (**Figure 4B**). The results demonstrated most of the signaling pathways were associated with DMED at the molecular and cellular levels, which could provide important information for revealing the most significant biological functions of DMED. We found that the biological processes of DMEDSig-4 were mainly divided into cell metabolism, neural signal transmission and planar cell polarity pathway. For example, we queried that the biological process "regulation of nitrogen Compound metabolic process" was associated with endothelial dysfunction (Andersson, 2003; Yuyun et al., 2018; Cyr et al., 2020). Endothelial dysfunction was recognized as a mainstay in the pathophysiology of the disease (Castela and Costa, 2016). As for the "WNT signaling pathway", studies had demonstrated that the Wnt family contributed to the pathogenesis of diabetes-induced erectile dysfunction (Shin et al., 2014; Ghatak et al., 2017). ED was also involved in the regulation of the metabolic and nervous systems (Burnett, 2005; Ryan and Gajraj, 2012; Mitidieri et al., 2020). For example, GO terms included "regulation of primary metabolic process" "regulation of cellular metabolic process" and "cellular response to stress". Literature verification showed that the above pathways were related to the pathological mechanism of DMED (Matfin et al., 2005; Zsoldos et al., 2019).

Finally, we annotated the characteristic miRNAs in DMEDSig-4 to prove the mechanism relationship between DMEDSig-4 and DMED. The expression of hsa-miR-342–3p helped to identify patients with cardiovascular disease (Seleem et al., 2019; Ray et al., 2020). Meanwhile, the expression level of this miRNA was significantly increased in diabetic nephropathy (Eissa et al., 2016; Jiang et al., 2020). Importantly, hsa-miR-342–3p was differentially expressed in obese children with and without endothelial dysfunction (Khalyfa et al., 2016), which was one of the important factors causing DMED. Meanwhile, hsa-miR-199b-5p, hsa-miR-30 days-5p and hsa-let-7e-5p were all related to diabetic kidney damage and cardiovascular diseases (Jia et al., 2016; Fedorko et al., 2017; Sun et al., 2018), which were all risk factors for ED.

DISCUSSION

Erectile dysfunction (ED) is a common and often overlooked complication of diabetes that can wreak havoc on men both physically and mentally. Studies have shown that type 2 diabetes mellitus (T2DM) is widely associated with ED and is a risk factor for ED. Interestingly, several studies have demonstrated that miRNAs are involved in the pathogenesis of ED. For example, Rama Natarajan et al. explored the role of miRNAs in the pathology of diabetic complications and also discussed the potential use of miRNAs as novel diagnostic and therapeutic targets for diabetic complications (Natarajan et al., 2012; Wang et al., 2019). Wang et al. found that upregulation of miR-320 was

associated with impaired angiogenesis in diabetes (Wang et al., 2009). Yan Wen et al. found that miR-205 may contribute to the pathogenesis of DMED via down-regulation of androgen receptor expressions (Wen et al., 2019). Although a large number of studies have proved the regulatory relationship of miRNAs on ED, there is still a lack of research on the relationship between miRNAs and ED in the context of T2DM. This study aimed to further explore ED signature associated with diabetes by analyzing miRNA expression data in patients with DM. This signature may play a certain role in the diagnosis and treatment of DMED. This study is not only a preliminary attempt on miRNA signature of DMED, but also may serve as the basis for subsequent studies.

In terms of data, we collected a large amount of clinical information and conducted preliminary screening to select patients (disease history/clinical information) as similar as possible and to eliminate the interference of other factors to the greatest extent possible. Finally, we selected 10 DM patients and 10 DMED patients as the final study subjects. However, due to the limited time and cost, our patient cohort is still relatively small, and there is no additional data set verification, so the identified DMEDSig-4 may not have good universality, which will be our further research direction. In this study, we used a machine learning approach (SVM-RFE) to identify potential miRNA features in a sample of DM and DMED patients. First, four optimal feature miRNAs (hsa-miR-342–3p, hsa-miR-199b-5p, hsa-miR-30 days-5p and hsa-let-7e-5p) were identified after 1,000 cycles of the algorithm, called DMEDSig-4. They all had a good classification effect, and there might be a potential mechanism of co-regulation. Subsequently, after associating with clinical factors, we found that DMEDSig-4 was positively correlated with each other and negatively correlated with IIEF-5. Then, we searched for the miRNAs targeted by DMDESig-4 and constructed a miRNA-mRNA interaction network. The results showed that the network consisted of four miRNA nodes, 105 mRNA nodes and 870 edges. Meanwhile, there was only one intersection between the targeted miRNAs of miRNA, indicating that these miRNAs did not tend to jointly regulate a target mRNA. Importantly, the identified target mRNAs had a close interaction relationship, suggesting that these mRNAs might act together on the same pathway. This might play an enlightening role in the subsequent studies of miRNA on DMED. Finally, we searched for the miRNAs targeted by DMDESig-4 and performed functional enrichment. The results showed that the DMDESig-4 pathways were closely related to DM and ED, which might contribute to the pathogenesis of ED. The literature review has shown that DMEDSig-4 was associated with cardiovascular disease, diabetic nephropathy and liver injury, which were all potential risk factors for ED (Hu et al., 2021). Clinical ED patients are often accompanied by cardiovascular disease, kidney and liver damage and other symptoms.

We hope that the characterization of miRNAs will contribute to a comprehensive understanding of their pathways in DMED and improve therapeutic strategies for patients with DMED. We hope that the identification of DMEDSig-4 will contribute to a comprehensive understanding of its pathway mechanism in DMED and improve therapeutic strategies for patients with DMED.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182053.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Qiqihar First Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

HQ designed the experiment, HX wrote the manuscript and conducted the experiment, HX, BZ, WZ, and PT contributed the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.762136/full#supplementary-material

REFERENCES

Aibar, S., Fontanillo, C., Droste, C., and De Las Rivas, J. (2015). Functional Gene Networks: R/Bioc Package to Generate and Analyse Gene Networks Derived from Functional Enrichment and Clustering. *Bioinformatics* 31 (10), 1686–1688. doi:10.1093/bioinformatics/btu864

Andersson, K. E. (2003). Erectile Physiological and Pathophysiological Pathways Involved in Erectile Dysfunction. *J. Urol.* 170 (2 Pt 2), S6–S13. doi:10.1097/01.ju.0000075362.08363.a4

Azimi, M., Totonchi, M., Rahimi, M., Firouzi, J., Sahranavard, P., Emami Razavi, A., et al. (2021). An Integrated Analysis to Predict micro-RNAs Targeting Both Stemness and Metastasis in Human Gastric Cancer. *J. Gastroenterol. Hepatol.* 36 (2), 436–445. doi:10.1111/jgh.15176

Barwari, T., Joshi, A., and Mayr, M. (2016). MicroRNAs in Cardiovascular Disease. *J. Am. Coll. Cardiol.* 68 (23), 2577–2584. doi:10.1016/j.jacc.2016.09.945

Bitar, M., Kuiper, S., O'Brien, E. A., and Barry, G. (2019). Genes with Human-specific Features Are Primarily Involved with Brain, Immune and Metabolic Evolution. *BMC Bioinformatics* 20 (Suppl. 9), 406. doi:10.1186/s12859-019-2886-2

Burnett, A. L. (2005). Metabolic Syndrome, Endothelial Dysfunction, and Erectile Dysfunction: Association and Management. *Curr. Urol. Rep.* 6 (6), 470–475. doi:10.1007/s11934-005-0043-0

Castela, Â., and Costa, C. (2016). Molecular Mechanisms Associated with Diabetic Endothelial-Erectile Dysfunction. *Nat. Rev. Urol.* 13 (5), 266–274. doi:10.1038/nrurol.2016.23

Cheng, L., Shi, H., Wang, Z., Hu, Y., Yang, H., Zhou, C., et al. (2016). IntNetLncSim: an Integrative Network Analysis Method to Infer Human lncRNA Functional Similarity. *Oncotarget* 7 (30), 47864–47874. doi:10.18632/oncotarget.10012

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

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018). Exposing the Causal Effect of C-Reactive Protein on the Risk of Type 2 Diabetes Mellitus: A Mendelian Randomization Study. *Front. Genet.* 9, 657. doi:10.3389/fgene.2018.00657

Cong, R., Wang, Y., Wang, Y., Zhang, Q., Zhou, X., Ji, C., et al. (2020). Comprehensive Analysis of lncRNA Expression Pattern and lncRNA-miRNA-mRNA Network in a Rat Model with Cavernous Nerve Injury Erectile Dysfunction. *J. Sex. Med.* 17 (9), 1603–1617. doi:10.1016/j.jsxm.2020.05.008

Cyr, A. R., Huckaby, L. V., Shiva, S. S., and Zuckerbraun, B. S. (2020). Nitric Oxide and Endothelial Dysfunction. *Crit. Care Clin.* 36 (2), 307–321. doi:10.1016/j.ccc.2019.12.009

Diaz-Arjonilla, M., Schwarcz, M., Swerdloff, R. S., and Wang, C. (2009). Obesity, Low Testosterone Levels and Erectile Dysfunction. *Int. J. Impot. Res.* 21 (2), 89–98. doi:10.1038/ijir.2008.42

Ding, J., Tang, Y., Tang, Z., Zhang, X., and Wang, G. (2017). A Variant in the Precursor of MicroRNA-146a Is Responsible for Development of Erectile Dysfunction in Patients with Chronic Prostatitis via Targeting NOS1. *Med. Sci. Monit.* 23, 929–937. doi:10.12659/msm.898406

Duan, K.-B., Rajapakse, J. C., Wang, H., and Azuaje, F. (2005). Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data. *IEEE Trans.on Nanobioscience* 4 (3), 228–234. doi:10.1109/tnb.2005.853657

Eissa, S., Matboli, M., and Bekhet, M. M. (2016). Clinical Verification of a Novel Urinary microRNA Panal: 133b, -342 and -30 as Biomarkers for Diabetic Nephropathy Identified by Bioinformatics Analysis. *Biomed. Pharmacother.* 83, 92–99. doi:10.1016/j.biopha.2016.06.018

Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA Translation and Stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379. doi:10.1146/annurev-biochem-060308-103103

Fang, S., Pan, J., Zhou, C., Tian, H., He, J., Shen, W., et al. (2019). Circular RNAs Serve as Novel Biomarkers and Therapeutic Targets in Cancers. *Cgt* 19 (2), 125–133. doi:10.2174/1566523218666181109142756

Fedorko, M., Juracek, J., Stanik, M., Svoboda, M., Poprach, A., Buchler, T., et al. (2017). Detection of Let-7 miRNAs in Urine Supernatant as Potential Diagnostic Approach in Non-metastatic clear-cell Renal Cell Carcinoma. *Biochem. Med. (Zagreb)* 27 (2), 411–417. doi:10.11613/bm.2017.043

Ghatak, K., Yin, G. N., Choi, M.-J., Limanjaya, A., Minh, N. N., Ock, J., et al. (2017). Dickkopf2 Rescues Erectile Function by Enhancing Penile Neurovascular Regeneration in a Mouse Model of Cavernous Nerve Injury. *Sci. Rep.* 7 (1), 17819. doi:10.1038/s41598-017-17862-5

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learn.* 46 (1), 389–422. doi:10.1023/A:1012487302797

Han, W., Lu, D., Wang, C., Cui, M., and Lu, K. (2021). Identification of Key mRNAs, miRNAs, and mRNA-miRNA Network Involved in Papillary Thyroid Carcinoma. *Cbio.* 16 (1), 146–153. doi:10.2174/1574893615999200608125427

Hatzichristou, D., Hatzimouratidis, K., Bekas, M., Apostolidis, A., Tzortzis, V., and Yannakoyorgos, K. (2002). Diagnostic Steps in the Evaluation of Patients with Erectile Dysfunction. *J. Urol.* 168 (2), 615–620. doi:10.1097/00005392-200208000-00044

Hu, Y., Qiu, S., and Cheng, L. (2021). Integration of Multiple-Omics Data to Analyze the Population-specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* 2021, 7036592. doi:10.1155/2021/7036592

Huang, Y. (2018). The Novel Regulatory Role of lncRNA-miRNA-mRNA axis in Cardiovascular Diseases. *J. Cel. Mol. Med.* 22 (12), 5768–5775. doi:10.1111/jcmm.13866

Huo, W., Li, H., Zhang, Y., and Li, H. (2020). Epigenetic Silencing of microRNA-874-3p Implicates in Erectile Dysfunction in Diabetic Rats by Activating the Nupr1/Chop-mediated Pathway. *FASEB j.* 34 (1), 1695–1709. doi:10.1096/fj.201902086r

Jia, K., Shi, P., Han, X., Chen, T., Tang, H., and Wang, J. (2016). Diagnostic Value of miR-30d-5p and miR-125b-5p in Acute Myocardial Infarction. *Mol. Med. Rep.* 14 (1), 184–194. doi:10.3892/mmr.2016.5246

Jiang, X., Luo, Y., Zhao, S., Chen, Q., Jiang, C., Dai, Y., et al. (2015). Clinical Significance and Expression of microRNA in Diabetic Patients with Erectile Dysfunction. *Exp. Ther. Med.* 10 (1), 213–218. doi:10.3892/etm.2015.2443

Jiang, Z. H., Tang, Y. Z., Song, H. N., Yang, M., Li, B., and Ni, C. L. (2020). miRNA-342 S-uppresses R-enal I-nterstitial F-ibrosis in D-iabetic N-ephropathy by T-argeting SOX6. *Int. J. Mol. Med.* 45 (1), 45–52. doi:10.3892/ijmm.2019.4388

Joshi, P., Vedhanayagam, M., and Ramesh, R. (2021). An Ensembled SVM Based Approach for Predicting Adverse Drug Reactions. *Cbio.* 16 (3), 422–432. doi:10.2174/1574893615999200707141420

Khalyfa, A., Kheirandish-Gozal, L., Bhattacharjee, R., Khalyfa, A. A., and Gozal, D. (2016). Circulating microRNAs as Potential Biomarkers of Endothelial Dysfunction in Obese Children. *Chest* 149 (3), 786–800. doi:10.1378/chest.15-0799

Kong, L., Zhu, J., Han, W., Jiang, X., Xu, M., Zhao, Y., et al. (2011). Significance of Serum microRNAs in Pre-diabetes and Newly Diagnosed Type 2 Diabetes: a Clinical Study. *Acta Diabetol.* 48 (1), 61–69. doi:10.1007/s00592-010-0226-0

Kouidrat, Y., Pizzol, D., Cosco, T., Thompson, T., Carnaghi, M., Bertoldo, A., et al. (2017). High Prevalence of Erectile Dysfunction in Diabetes: a Systematic Review and Meta-Analysis of 145 Studies. *Diabet. Med.* 34 (9), 1185–1192. doi:10.1111/dme.13403

Lei, T., and Shu-Lin, W. (2021). Exploring miRNA Sponge Networks of Breast Cancer by Combining miRNA-disease-lncRNA and miRNA-target Networks. *Curr. Bioinformatics* 16 (3), 385–394. doi:10.2174/1574893615999200711171

Li, D.-S., Feng, L., Luo, L.-H., Duan, Z.-F., Li, X.-L., Yin, C.-H., et al. (2017). The Effect of microRNA-328 Antagomir on Erectile Dysfunction in Streptozotocin-Induced Diabetic Rats. *Biomed. Pharmacother.* 92, 888–895. doi:10.1016/j.biopha.2017.05.071

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucleic Acids Res.* 42 (Database issue), D92–D97. doi:10.1093/nar/gkt1248

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., and Wei, H. (2017). Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics. *Molecules* 23 (1), 52. doi:10.3390/molecules23010052

Liu, D., Huang, Y., Nie, W., Zhang, J., and Deng, L. (2021). SMALF: miRNA-Disease Associations Prediction Based on Stacked Autoencoder and XGBoost. *BMC Bioinformatics* 22 (1), 219. doi:10.1186/s12859-021-04135-2

Liu, S., Tang, H., Liu, H., and Wang, J. (2021). Multi-label Learning for the Diagnosis of Cancer and Identification of Novel Biomarkers with High-Throughput Omics. *Cbio.* 16 (2), 261–273. doi:10.2174/1574893615999200623130416

Long, J., Yang, H., Yang, Z., Jia, Q., Liu, L., Kong, L., et al. (2021). Integrated Biomarker Profiling of the Metabolome Associated with Impaired Fasting Glucose and Type 2 Diabetes Mellitus in Large-Scale Chinese Patients. *Clin. Transl. Med.* 11 (6), e432. doi:10.1002/ctm2.432

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Lu, T. X., and Rothenberg, M. E. (2018). MicroRNA. *J. Allergy Clin. Immunol.* 141 (4), 1202–1207. doi:10.1016/j.jaci.2017.08.034

Malavige, L. S., and Levy, J. C. (2009). Erectile Dysfunction in Diabetes Mellitus. *J. Sex. Med.* 6 (5), 1232–1247. doi:10.1111/j.1743-6109.2008.01168.x

Matfin, G., Jawa, A., and Fonseca, V. A. (2005). Erectile Dysfunction: Interrelationship with the Metabolic Syndrome. *Curr. Diab. Rep.* 5 (1), 64–69. doi:10.1007/s11892-005-0070-8

Mitidieri, E., Cirino, G., d'Emmanuele di Villa Bianca, R., and Sorrentino, R. (2020). Pharmacology and Perspectives in Erectile Dysfunction in Man. *Pharmacol. Ther.* 208, 107493. doi:10.1016/j.pharmthera.2020.107493

Mo, F., Luo, Y., Fan, D., Zeng, H., Zhao, Y., Luo, M., et al. (2020). Integrated Analysis of mRNA-Seq and miRNA-Seq to Identify C-MYC, YAP1 and miR-3960 as Major Players in the Anticancer Effects of Caffeic Acid Phenethyl Ester in Human Small Cell Lung Cancer Cell Line. *Cgt* 20 (1), 15–24. doi:10.2174/1566523220666200523165159

Müller, A., and Mulhall, J. P. (2006). Cardiovascular Disease, Metabolic Syndrome and Erectile Dysfunction. *Curr. Opin. Urol.* 16 (6), 435–443. doi:10.1097/01.mou.0000250284.83108.a6

Natarajan, R., Putta, S., and Kato, M. (2012). MicroRNAs and Diabetic Complications. *J. Cardiovasc. Trans. Res.* 5 (4), 413–422. doi:10.1007/s12265-012-9368-5

Norsworthy, P. J., Thompson, A. G. B., Mok, T. H., Guntoro, F., Dabin, L. C., Nihat, A., et al. (2020). A Blood miRNA Signature Associates with Sporadic Creutzfeldt-Jakob Disease Diagnosis. *Nat. Commun.* 11 (1), 3960. doi:10.1038/s41467-020-17655-x

Ray, S. L., Coulson, D. J., Yeoh, M. L. Y., Tamara, A., Latief, J. S., Bakhashab, S., et al. (2020). The Role of miR-342 in Vascular Health. Study in Subclinical Cardiovascular Disease in Mononuclear Cells, Plasma, Inflammatory Cytokines and PANX2. *Int. J. Mol. Sci.* 21 (19), 7217. doi:10.3390/ijms21197217

Riaz, F., and Li, D. (2019). Non-coding RNA Associated Competitive Endogenous RNA Regulatory Network: Novel Therapeutic Approach in Liver Fibrosis. *Cgt* 19 (5), 305–317. doi:10.2174/1566523219666191107113046

Rosen, R. C., Riley, A., Wagner, G., Osterloh, I. H., Kirkpatrick, J., and Mishra, A. (1997). The International Index of Erectile Function (IIEF): a Multidimensional Scale for Assessment of Erectile Dysfunction. *Urology* 49 (6), 822–830. doi:10.1016/s0090-4295(97)00238-0

Rupaimoole, R., and Slack, F. J. (2017). MicroRNA Therapeutics: towards a new era for the Management of Cancer and Other Diseases. *Nat. Rev. Drug Discov.* 16 (3), 203–222. doi:10.1038/nrd.2016.246

Ryan, J. G., and Gajraj, J. (2012). Erectile Dysfunction and its Association with Metabolic Syndrome and Endothelial Function Among Patients with Type 2 Diabetes Mellitus. *J. Diabetes its Complications* 26 (2), 141–147. doi:10.1016/j.jdiacomp.2011.12.001

Sanz, H., Valim, C., Vegas, E., Oller, J. M., and Reverter, F. (2018). SVM-RFE: Selection and Visualization of the Most Relevant Features through Non-linear Kernels. *BMC Bioinformatics* 19 (1), 432. doi:10.1186/s12859-018-2451-4

Seleem, M., Shabayek, M., and Ewida, H. A. (2019). MicroRNAs 342 and 450 Together with NOX-4 Activity and Their Association with Coronary Artery Disease in Diabetes. *Diabetes Metab. Res. Rev.* 35 (5), e3130. doi:10.1002/dmrr.3130

Shamloul, R., and Ghanem, H. (2013). Erectile Dysfunction. *The Lancet* 381 (9861), 153–165. doi:10.1016/s0140-6736(12)60520-0

Sheng, Y., Jiang, Y., Yang, Y., Li, X., Qiu, J., Wu, J., et al. (2021). CNA2Subpathway: Identification of Dysregulated Subpathway Driven by Copy Number Alterations in Cancer. *Brief Bioinform.* 22 (5), bbaa413. doi:10.1093/bib/bbaa413

Shin, S. H., Kim, W. J., Choi, M. J., Park, J.-M., Jin, H.-R., Yin, G. N., et al. (2014). Aberrant Expression of Wnt Family Contributes to the Pathogenesis of Diabetes-Induced Erectile Dysfunction. *Andrology* 2 (1), 107–116. doi:10.1111/j.2047-2927.2013.00162.x

Sun, Z., Ma, Y., Chen, F., Wang, S., Chen, B., and Shi, J. (2018). miR-133b and miR-199b Knockdown Attenuate TGF-B1-Induced Epithelial to Mesenchymal Transition and Renal Fibrosis by Targeting SIRT1 in Diabetic Nephropathy. *Eur. J. Pharmacol.* 837, 96–104. doi:10.1016/j.ejphar.2018.08.022

Tang, H., Zhao, Y.-W., Zou, P., Zhang, C.-M., Chen, R., Huang, P., et al. (2018). HBPred: a Tool to Identify Growth Hormone-Binding Proteins. *Int. J. Biol. Sci.* 14 (8), 957–964. doi:10.7150/ijbs.24174

Tang, M., Liu, C., Liu, D., Liu, J., Liu, J., and Deng, L. (2021). PMDFI: Predicting miRNA-Disease Associations Based on High-Order Feature Interaction. *Front. Genet.* 12, 656107. doi:10.3389/fgene.2021.656107

Vasu, S., Kumano, K., Darden, C. M., Rahman, I., Lawrence, M. C., and Naziruddin, B. (2019). MicroRNA Signatures as Future Biomarkers for Diagnosis of Diabetes States. *Cells* 8 (12), 1533. doi:10.3390/cells8121533

Wang, J.-y., Yang, Y., Ma, Y., Wang, F., Xue, A., Zhu, J., et al. (2020). Potential Regulatory Role of lncRNA-miRNA-mRNA axis in Osteosarcoma. *Biomed. Pharmacother.* 121, 109627. doi:10.1016/j.biopha.2019.109627

Wang, L., Xuan, Z., Zhou, S., Kuang, L., and Pei, T. (2019). A Novel Model for Predicting LncRNA-Disease Associations Based on the LncRNA-MiRNA-Disease Interactive Network. *Cbio.* 14 (3), 269–278. doi:10.2174/1574893613666180703105258

Wang, X., Qian, R., Zhang, W., Chen, S., Jin, H., and Hu, R. (2009). MicroRNA-320 Expression in Myocardial Microvascular Endothelial Cells and its Relationship with Insulin-like Growth Factor-1 in Type 2 Diabetic Rats. *Clin. Exp. Pharmacol. Physiol.* 36 (2), 181–188. doi:10.1111/j.1440-1681.2008.05057.x

Wen, Y., Liu, G., Zhang, Y., and Li, H. (2019). MicroRNA-205 Is Associated with Diabetes Mellitus-induced Erectile Dysfunction via Down-regulating the Androgen Receptor. *J. Cel. Mol. Med.* 23 (5), 3257–3270. doi:10.1111/jcmm.14212

Wu, Y., Lu, X., Shen, B., and Zeng, Y. (2019). The Therapeutic Potential and Role of miRNA, lncRNA, and circRNA in Osteoarthritis. *Cgt* 19 (4), 255–263. doi:10.2174/1566523219666190716092203

Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al. (2021). Risk Prediction of Diabetes: Big Data Mining with Fusion of Multifarious Physical Examination Indicators. *Inf. Fusion* 75, 140–149. doi:10.1016/j.inffus.2021.02.015

Yang, X.-F., Zhou, Y.-K., Zhang, L., Gao, Y., and Du, P.-F. (2020). Predicting LncRNA Subcellular Localization Using Unbalanced Pseudo-k Nucleotide Compositions. *Cbio.* 15 (6), 554–562. doi:10.2174/1574893614666190902151038

Yuyun, M. F., Ng, L. L., and Ng, G. A. (2018). Endothelial Dysfunction, Endothelial Nitric Oxide Bioavailability, Tetrahydrobiopterin, and 5-methyltetrahydrofolate in Cardiovascular Disease. Where Are We with Therapy? *Microvasc. Res.* 119, 7–12. doi:10.1016/j.mvr.2018.03.012

Zagidullin, B., Aldahdooh, J., Zheng, S., Wang, W., Wang, Y., Saad, J., et al. (2019). DrugComb: an Integrative Cancer Drug Combination Data portal. *Nucleic Acids Res.* 47 (W1), W43–W51. doi:10.1093/nar/gkz337

Zhang, Z., Zhang, S., Li, X., Zhao, Z., Chen, C., Zhang, J., et al. (2021). Reference Genome and Annotation Updates lead to Contradictory Prognostic Predictions in Gene Expression Signatures: a Case Study of Resected Stage I Lung Adenocarcinoma. *Brief Bioinform.* 22 (3), bbaa081. doi:10.1093/bib/bbaa081

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a Novel Deep Learning Method for Prioritizing lncRNA Target Genes. *Bioinformatics* 36 (16), 4466–4472. doi:10.1093/bioinformatics/btaa428

Zhu, M., Yi, M., Kim, C. H., Deng, C., Li, Y., Medina, D., et al. (2011). Integrated miRNA and mRNA Expression Profiling of Mouse Mammary Tumor Models Identifies miRNA Signatures Associated with Mammary Tumor Lineage. *Genome Biol.* 12 (8), R77. doi:10.1186/gb-2011-12-8-r77

Zhu, Q., Fan, Y., and Pan, X. (2021). Fusing Multiple Biological Networks to Effectively Predict miRNA-Disease Associations. *Cbio.* 16 (3), 371–384. doi:10.2174/1574893615999200715165335

Zhu, Z., Han, X., and Cheng, L. (2021). Identification of Gene Signature Associated with Type 2 Diabetes Mellitus by Integrating Mutation and Expression Data. *Curr. Gene Ther.* doi:10.2174/1566523221666210707140839

Zsoldos, M., Pajor, A., and Pusztafalvi, H. (2019). A Szexuális Funkciózavar És a Metabolikus Szindróma Kapcsolata. *Orvosi. Hetilap.* 160 (3), 98–103. doi:10.1556/650.2019.31235

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fgene.2021.765400

The Transcriptome Characteristics of Severe Asthma From the Prospect of Co-Expressed Gene Modules

Bin Li[1,2,3†], Wen-Xuan Sun[1†], Wan-Ying Zhang[1,3], Ye Zheng[1], Lu Qiao[1], Yue-Ming Hu[1], Wei-Qiang Li[1], Di Liu[1], Bing Leng[1], Jia-Ren Liu[1,3], Xiao-Feng Jiang[1] and Yan Zhang[2]**

[1]Department of Clinical Laboratory, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China, [2]School of Life Science and Technology, Computational Biology Research Center, Harbin Institute of Technology, Harbin, China, [3]Heilongjiang Longwei Precision Medical Laboratory Center, Harbin, China

Rationale: Severe asthma is a heterogeneous disease with multiple molecular mechanisms. Gene expression studies of asthmatic bronchial epithelial cells have provided biological insights and underscored possible pathological mechanisms; however, the molecular basis in severe asthma is still poorly understood.

Objective: The objective of this study was to identify the features of asthma and uncover the molecular basis of severe asthma in distinct molecular phenotype.

Methods: The k-means clustering and differentially expressed genes (DEGs) were performed in 129 asthma individuals in the Severe Asthma Research Program. The DEG profiles were analyzed by weighted gene co-expression network analysis (WGCNA), and the expression value of each gene module in each individual was annotated by gene set variation analysis (GSVA).

Results: Expression analysis defined five stable asthma subtype (AS): 1) Phagocytosis-Th2, 2) Normal-like, 3) Neutrophils, 4) Mucin-Th2, and 5) Interferon-Th1 and 15 co-expressed gene modules. "Phagocytosis-Th2" enriched for receptor-mediated endocytosis, upregulation of Toll-like receptor signal, and myeloid leukocyte activation. "Normal-like" is most similar to normal samples. "Mucin-Th2" preferentially expressed genes involved in O-glycan biosynthesis and unfolded protein response. "Interferon-Th1" displayed upregulation of genes that regulate networks involved in cell cycle, IFN gamma response, and CD8 TCR. The dysregulation of neural signal, REDOX, apoptosis, and O-glycan process were related to the severity of asthma. In non-TH2 subtype (Neutrophils and Interferon-Th1) with severe asthma individuals, the neural signals and IL26-related co-expression module were dysregulated more significantly compared to that in non-severe asthma. These data infer differences in the molecular evolution of asthma subtypes and identify opportunities for therapeutic development.

Abbreviations: Normal: Healthy controls, AS: asthma subtype, BAl: bronchoalveolar lavage, FeNO: fractional exhaled nitric oxide, ppb: parts-per-billion, Up: the number of upregulated genes in the module, Down: the number of downregulated genes in the module, NonS: the mean value of module in non-severe asthma samples. meanS: the mean value of module in severe asthma samples.

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Chuan-Le Xiao,
Sun Yat-sen University, China
Guoqing Liu,
Inner Mongolia University of Science and Technology, China

***Correspondence:**
Xiao-Feng Jiang
jiangxiaofeng@hrbmu.edu.cn
Yan Zhang
zhangtyo@hit.edu.cn

†These authors share Co-first authorship

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

Received: *27 August 2021*
Accepted: *29 September 2021*
Published: *25 October 2021*

Citation:
Li B, Sun W-X, Zhang W-Y, Zheng Y, Qiao L, Hu Y-M, Li W-Q, Liu D, Leng B, Liu J-R, Jiang X-F and Zhang Y (2021) The Transcriptome Characteristics of Severe Asthma From the Prospect of Co-Expressed Gene Modules. Front. Genet. 12:765400. doi: 10.3389/fgene.2021.765400

Conclusions: Asthma is a heterogeneous disease. The co-expression analysis provides new insights into the biological mechanisms related to its phenotypes and the severity.

Keywords: Phagocytosis-Th2, normal-like, neutrophils, mucin-Th2, Interferon-Th1

INTRODUCTION

Asthma is a chronic disorder, characterized by airway hyper-responsiveness (AHR) and remodeling with variable degrees of eosinophilic and neutrophilic inflammation resulting in significant morbidity and mortality (Wilson et al., 2006; Kim et al., 2010). It affects about 5% of the population (Global et al., 2017). According to the clinical characteristics, it is mainly divided into the acute and the non-acute asthma, which is further divided into mild, moderate, and severe asthma individual. About 5–10% of the patients do not respond well to standard treatment and have a poor prognosis (Higgins, 2003). The bronchial epithelial cells act as a physical barrier in airway immunity and as central modulators of inflammatory response (Hamilton et al., 2001). Environmental stimuli promote epithelial cell synthesis and secretion by a variety of mediators, such as cytokines, chemokines, reactive oxygen species, lipid, and peptide mediators and eventually involved in recruiting leukocytes, mucus secretion, vascular permeability, bronchoconstriction, and airway hyper-responsiveness (Whitsett, 2018).

Gene expression and genetic variation studies both indicate that asthma is a polygenic and heterogeneous disease with multiple molecular roots (Langfelder and Horvath, 2008; Belsky et al., 2013). Based on the gene profiles in bronchial epithelial cells associated with fractional exhaled nitric oxide (FeNO), Modena et al. identified five phenotypes of asthma. The results showed that a large number of individuals were severe asthma in each subtype (Modena et al., 2014). However, the typical characteristics of phenotype and the features related to severe asthma in phenotype were also unclear. Therefore, revealing these characteristics in each molecular subtype could be valuable for individualized treatment of severe asthma.

In recent years, the WGCNA (Langfelder and Horvath, 2008) (weighted gene co-expression network analysis) is a new system biology approach that can be used to identify co-expression gene sets that largely represent the typical biological characteristics in complex disease. Therefore, in this study, we used the co-expressed gene modules (GMs) that were used to uncover the typical features in subtype and severe asthma.

MATERIALS AND METHODS

Study Population and Data Processing

As part of SARP (Severe Asthma Research Program), bronchial brushing samples and matching demographic data were obtained from 155 participants (129 asthmatics and 26 healthy subjects) from 2009 to 2011. Gene expression of the SARP and external cohorts are available online (GEO database; http://www.ncbi.\.

K-Means and Limma Analysis

According to the transcription of bronchial epithelial cells, the k-means method integrated in the ConsensusClusterPlus (Wilkerson and Hayes, 2010) package was adopted to identify stable subtypes, and the stability was evaluated by iterating for 1,000 times at a sub-sampling rate of 0.95. A total of 4,650 (MAD: median absolute deviation >0.5) genes were used as input. Starting from k = 4-5, a significant improvement in clustering stability can be observed, but it has no effect on $k > 5$ (**Figure 1C**). They were termed by asthma subtype. After the establishment of these ASs, the Bayesian method in Limma package was used to select the differentially expressed genes (DEGs) between each ASs and the normal. The cutoff setting: FC $\geq$ log (1.5), FDR $\leq$ 0.05.

To determine the inflammatory Th2 group, K-means on 155 subjects was performed based on microarray expression profiles of three Th2 marker genes (Woodruff et al., 2007) (periostin: POSTN, channel regulator 1: CLCA1, and serpin peptidase inhibitor clade B member 2: SERPINB2). and three main clusters were identified. They were named Th-H (Th2-high), Th-M (Th2-moderate), and Th-L (Th2-low).

WGCNA Co-expressed Analysis

Using the default parameter setting and the 2,664 DEGs selected in ASs, the WGCNA was performed. This method clusters genes into modules using a topological overlap measure (TOM) (Langfelder et al., 2008). The TOM was a highly robust interconnection measurement method that essentially provided a measure of the connection strength between two adjacent genes and all other genes in a network. Genes were clustered using 1-TOM as the distance measure and GMs were defined as branches of the resulting cluster tree using a dynamic branch-cutting algorithm. Based on the dysregulated direction of each gene in asthma, the genes in each co-expressed module were split into 2 GMs.

Gene Set Variation Analysis

Gene Set Variation Analysis (GSVA) was performed using the R package "GSVA" (Hänzelmann et al., 2013) (function gsva - arguments: method = "gsva", mx. diff = TRUE). GSVA implements a non-parametric unsupervised method of gene set enrichment that allowed an assessment of the relative enrichment of a selected pathway across the sample space. The output of GSVA was a gene set by sample matrix of GSVA enrichment scores that were approximately normally distributed. GSVA enrichment scores were generated for each gene set using the normalized gene expression data.

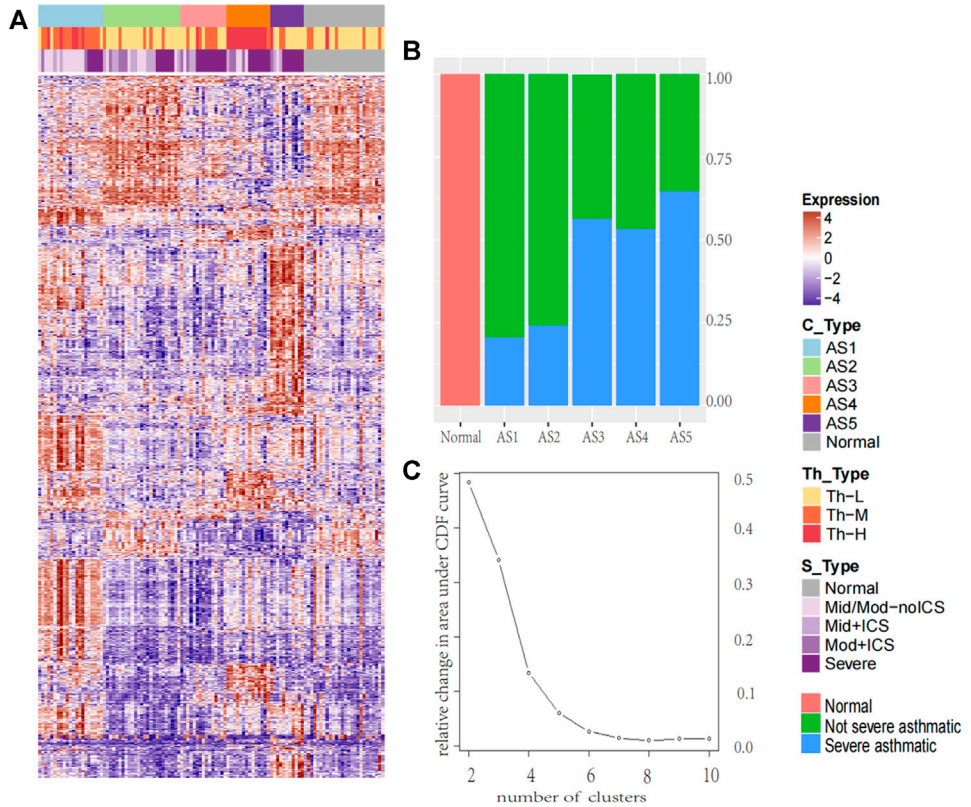

FIGURE 1 | The heatmap of molecular subtype in asthma and the percentage of severe asthma in each cluster. The heatmap of molecular subtype in asthma and the percentage of severe asthma in each cluster **(A)** The k-means identified five stable asthma subtypes (AS1–5), the heatmap was derived by 2,664 DEGs selected in each subtype. To define the Th2 subtype, the unsupervised hierarchical clustering was performed based on the microarray expression levels of periostin (POSTN), channel regulator 1 (CLCA1), and serpin peptidase inhibitor, clade B, member 2 (SERPINB2). We named these subtypes: (1) Th-H, High; (2) Th-M, Moderate; (3) Th-L, Low. According to the severity of the disease, the samples were divided into five groups: Normal for normal group; Mid/Mod-noICS for mild and moderate without ICS treatment; Mid + ICS for mild and ICS treatment group; Mod + ICS for moderate and ICS treatment group; and Severe for severe asthma group **(B)** The percentage of severe asthma in each cluster (red, normal samples, green: non-severe asthma individuals; blue: severe asthma individuals).

Pathway Enrichment Analysis

Using the clusterProfiler (Yu et al., 2012) package, the functional enrichment analysis was performed for the up- and downregulated GMs. Significance cutoff was defined as FDR <0.05 for multiple testing.

Correlation Analysis

The correlation analysis was performed by Pearson. The GSVA score in each gene set represented its overall expression in individual. The C2 (curated gene sets) dataset was downloaded from Molecular Signatures Database (MSigDB), which is a collection of annotated gene sets for pathway analysis (http://software.broadinstitute.org/gsea/msigdb).

RESULTS

K-Means and Differentially Expressed Gene Analysis

The k-means was performed, and five stable asthma subtypes were obtained (**Figure 1A**). These subtypes were named as follows: 1) Phagocytosis-Th2, 2) Normal-like, 3) Neutrophils-Type, 4) Mucin-Th2, and 5) Interferon-Th1 based on the differential expression modules and their related biological clinical characteristics. These five subtypes were associated with specific clinical characteristics (**Table 1**). "Interleukin-Th2" is the youngest group (mean age = 29) with an elevated FeNO (43 ppb). "Normal-like" has the highest Juniper AQLQ (mean 5, p = 4.2E-12). "Neutrophils-Type" has the highest levels of neutrophils in blood (mean, 60, p = 0.02) and BAL (mean, 4, p = 0.02) and the lowest total cells count in BAL (mean, 2.3, p = 3.9E-05). "Mucin-Th2" has the highest FeNO overall (mean, 46, p = 2.5E-05) and the greatest reversibility (mean, 21, p = 2.5E-05). The "Interferon-Th1" has the highest lymphocytes (mean, 14, p = 0.01) in BAL. In addition, three subtypes ("Neutrophils-Type", "Mucin-Th2", and "Interferon-Th1") had more percentage of severe asthma individuals (chi-square, p < 0.001) (**Figure 1B**).

The Co-expression Features of AS

A total of 2,664 DEGs were detected among these five ASs compared to the normal samples. Using 2,664 DEG as input, a

TABLE 1 | Summary of clinical characteristics of the SARP cohort in AS.

	Normal	AS1	AS2	AS3	AS4	AS5	*p*-value
Inflammatory cells in blood							
Total WBC	5.5 ± 1.3	6.1 ± 1.8	6.4 ± 1.8	7 ± 2.7	6.8 ± 2.9	7 ± 2.9	0.0007
Neutrophils, %	54 ± 6.3	52.8 ± 10	61 ± 8.7	62 ± 15.4	62 ± 12.9	62 ± 13.9	0.02
Basophils, %	1 ± 0.5	0.6 ± 0.5	0.6 ± 0.5	0.2 ± 0.5	1 ± 0.5	1 ± 0.5	0.73
Eosinophils, %	2 ± 1.1	4 ± 2.6	2 ± 1.2	3 ± 4.4	5 ± 2.7	2.5 ± 3.3	0.0009
Lymphocytes, %	33 ± 5.6	34 ± 8.2	31 ± 9.1	26 ± 11.1	29 ± 11.2	26 ± 10.9	0.004
Monocytes, %	8 ± 2.5	8 ± 1.8	7 ± 2	8 ± 4.8	6 ± 1.4	7 ± 2	0.001
Inflammatory cells in BAL							
BAL Total cells	6.1 ± 3.5	8 ± 8.8	7 ± 8.8	2.3 ± 2	4.3 ± 3.6	4.7 ± 8.9	3.9E-05
BAL macrophages,%	86.4 ± 9	91 ± 5.8	90 ± 6.4	85 ± 11.8	89 ± 18.6	81 ± 12.6	0.002
BAL lymphocytes,%	9 ± 7.5	5.8 ± 4.8	7.8 ± 6	9.3 ± 7.2	7 ± 9.5	14 ± 8.3	0.01
BAL eosinophils,%	0.2 ± 1.9	0.4 ± 1.6	0.4 ± 0.7	0.7 ± 1.8	1 ± 2.6	0.3 ± 1.1	0.01
BAL Neutrophils,%	2 ± 4.2	1.3 ± 2.7	1.5 ± 1.8	4 ± 8.2	2 ± 12	2.7 ± 8.7	0.02
Inflammatory cells in sputum							
Total cells, millions	2.1 ± 3.7	1.6 ± 3.8	2 ± 1.4	1.9 ± 5.4	1.4 ± 1.2	1.1 ± 3.6	0.46
Total WBC, millions	1.1 ± 2.2	0.8 ± 3.5	1.3 ± 1	1.3 ± 4.8	1.1 ± 1	0.7 ± 3.6	0.45
Viability of WBCs,%	67 ± 16.6	54 ± 25.9	68 ± 25.7	74 ± 12.2	69 ± 26.9	67 ± 27.8	0.54
Macrophages, %	28 ± 15.1	42 ± 23.8	46 ± 25.9	33 ± 16.1	42 ± 18	42 ± 30.6	0.46
Bronchial epithelial cells,%	3 ± 6	3 ± 9.5	2.5 ± 6.9	2.5 ± 10	5 ± 10.7	1 ± 1.6	0.11
Eosinophils, %	0.9 ± 6.7	0.7 ± 2.9	0.5 ± 10	2.2 ± 15	7.2 ± 11	2.7 ± 1.4	0.40
Lymphocytes, %	1.1 ± 1.4	1.7 ± 3.2	1.4 ± 2.5	1.8 ± 2.2	1 ± 1.5	1.1 ± 1.2	0.82
Pulmonary function and other characteristics							
Baseline FEV1, % predicted	94.5 ± 9	83 ± 16.4	81 ± 24.5	67 ± 25.3	59 ± 19.4	67 ± 18.3	9.9E-08
Baseline FVC, % predicted	96 ± 10.9	93 ± 13.3	85 ± 20.6	81 ± 22.2	69.4 ± 20	89 ± 17.7	0.0009
Maximum FEV1 reversal, %	5.3 ± 3.6	13 ± 16.2	8.7 ± 24	15 ± 38.4	21 ± 27.7	15 ± 13.1	2.5E-05
Juniper AQLQ	7 ± 0.1	4.4 ± 1.2	5 ± 1.3	4.9 ± 1.4	3.8 ± 1.3	3.9 ± 1.1	4.2E-12
Age, years	28 ± 11.9	29 ± 10.7	43 ± 12.2	48 ± 13.7	42 ± 11.1	35 ± 15.5	0.005
Age when first diagnosed	NA	6 ± 10.7	12 ± 14	10 ± 20	13 ± 8.9	9 ± 14.5	0.001
Body mass index	24 ± 5.2	29 ± 5.6	28 ± 6.3	31 ± 6.3	33 ± 10.2	27 ± 6.3	0.052
Number_of_positive_skin_reactions	1.5 ± 3.2	4 ± 3	2 ± 3.2	4 ± 4.1	5 ± 3.6	3 ± 1.9	0.02
FeNO, ppb	21 ± 50.9	43 ± 30.9	17 ± 14.9	43 ± 34.6	46 ± 66.6	22 ± 18.3	2.5E-05

total of 15 coordinately expressed GMs representing distinct biological processes were obtained. In these 15 GMs with only upregulated genes, 10 discriminated these five asthma clusters (**Figure 2**). In the validation dataset, compared to the normal group, the overall expression level of these modules was consistent with the trend in this study.

Phagocytosis-Th2 Subtype: AS1

Two core gene programs (GM2 and GM3) characterized "Phagocytosis-Th2", which included gene networks involved in leukocyte migration (5.3%, FDR = 9.15E-05), osteoclast differentiation (4.5%, FDR = 0.006), receptor-mediated endocytosis (5.7%, FDR = 7.03E-04, COLEC12, MSR1, CD163), and antigen processing and presentation *via* MHC class II (6%, p = 1.42E-13, HLA-DMA (Gao et al., 2020), HLA-DRB5, HLA-DMB, HLA-DRB4, HLA-DPB1, HLA-DRA, HLA-DRB3, HLA-DOA, HLA-DQA2 (Lasky-Su et al., 2012), HLA-DRB1, HLA-DPA1). The clinical characteristics showed that the eosinophils (mean: 4, p = 0.0009) were abnormally increased in blood. FeNo (mean: 43, p = 2.5E-05) levels were higher, while the lymphocyte counts (mean, 5.8, p = 0.01) in BAL were lower compared to normal samples. The overall expression of GM2 in this cluster was significantly related to phagocytosis category, which was the reason why it was termed as "Phagocytosis-Th2".

Normal-like Subtype: AS2

The Normal-like subtype was the most similar to the normal group, with the mildest clinical symptoms and the fewest differential expressed genes. The top three upregulated genes were PHACTR3 (Itoh et al., 2014), SLCO1B3, and GNMT, which were related to the response of glucocorticoids.

Neutrophils-type Subtype: AS3

The GM10 was a typical feature of Neutrophils-Type. This module only contained eight upregulated genes, including SLCO1B3, PHACTR3, TPO, and FKBP5 (Binder, 2009), which were also related to the glucocorticoids and severity. The TPO was a marker that related to the severity of asthma (Voraphani et al., 2014).

Mucin-Th2 Subtype: AS4

Three core gene programs (GM6, GM9, and GM14) characterized "Mucin-Th2", which included gene networks involved in O-linked glycosylation (2%, p = 0.01,

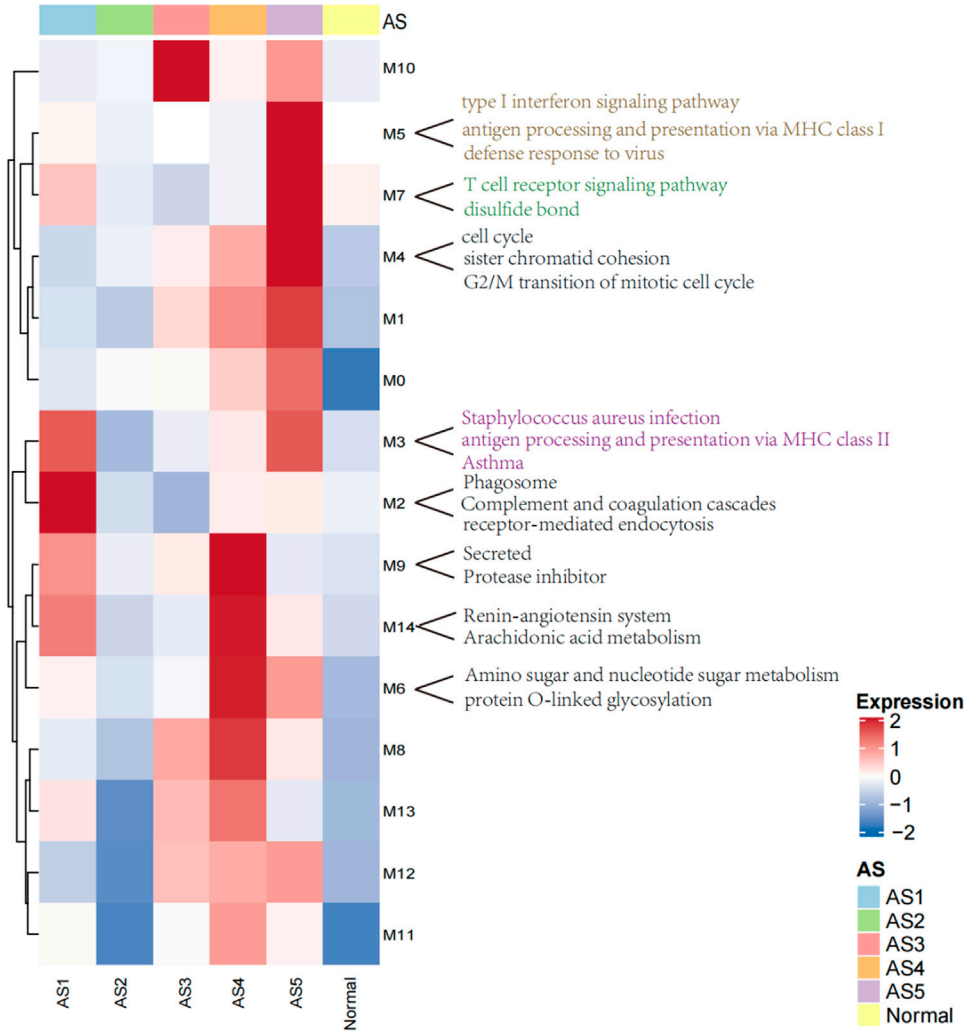

FIGURE 2 | The heatmap of modules with upregulated genes in ASs. The overall expression is represented by red and blue; the red indicates high expressed in cluster, and the blue indicates low expressed in cluster. M0–M14 represents 15 co-expressed modules with upregulated genes.

i.e., GALNT7, PGM3 (Zhang et al., 2019), and GALNT10), amino acid biosynthetic process (3%, p = 0.08, i.e., FOLH1, FOLH1B, and PYCR1), and negative regulation of endopeptidase activity (5.7%, i.e., SERPINA11, SERPINB10, SERPINB2 (Sánchez-Ovando et al., 2020), AHSG, WFIKKN2, and FETUB (Diao et al., 2016)). Individualized functional analysis showed that about 90% of individuals in "Mucin-Th2" significantly enriched Mucin type O-Glycan biosynthesis pathway. This cluster has the highest expression of Th2 marker genes (CLCA1, POSTN, and SERPINB2) and has the highest FeNo overall (median = 46 ppb, p = 2.5E-05) and eosinophils in blood, BAL, and sputum (**Table 1**). Although Mucin-Th2 was the typical Th2, no inflammation and immune-related modules (GM2, GM3, GM5, and GM7) were found overexpressed in this cluster.

Interferon-Th1 Subtype: AS5

Three core gene programs (GM4, GM5, and GM7) characterized Interferon-Th1, which included gene networks involved in cell division (23%, p = 2.39E-29, i.e., ERCC6L, CDCA2, and CDCA3), type I interferon response (15%, p = 5.84E-30, i.e., IFITM3, IFITM1, and IFITM2), antigen processing and presentation *via* MHC class I (6%, p = 6.53E-09, i.e., HLA-H, HLA-B, HLA-C, HLA-A, HLA-F, B2M, HLA-G, and HLA-E), and T-cell activation (8.9%, p = 7.17E-12, i.e., ITK, ZAP70, TNFSF14, CD8B, CD8A, and CD48). The overall expression of GM5 was significantly related to interferon response, which was the reason we termed this type "Interferon-Th1". This subtype was a typical non-Th2 subtype with normal FeNo (mean:22) and eosinophils in peripheral blood, BAL, and sputum.

The Characteristics Related to Severity in Asthma

In GMs with upregulated genes, 9 GMs (0, 1, 3, 4, 6, 8, 10, 12, and 14) were positively correlated to the severity and positive association with the use of ICS and OCS. They were also the

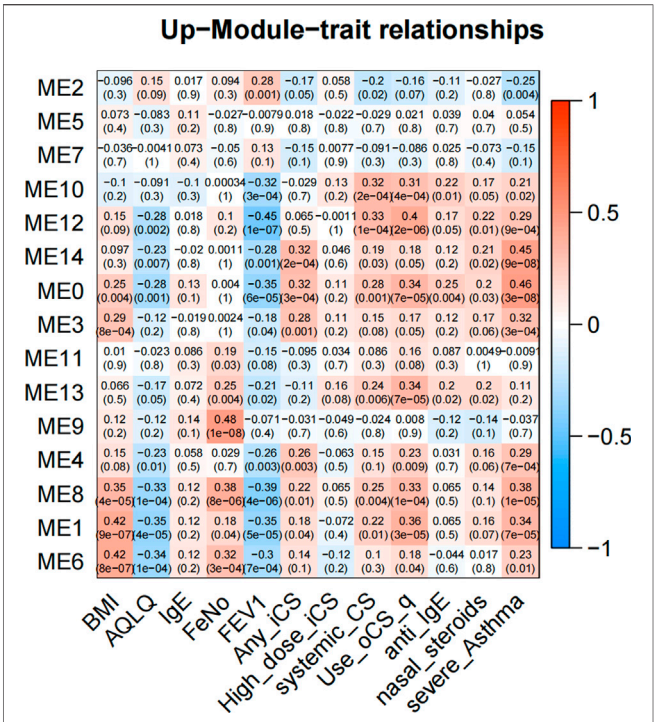

FIGURE 3 | The heatmap of correlation between co-expressed modules with upregulated genes and clinical characteristics. The clinical characteristics include BMI, AQLQ, IgE, fractional exhaled nitric oxide (FeNO), FEV1% predicted, use of inhaled (ICS) and oral corticosteroids (OCS), high dose use of ICS, systemic use CS, anti_IgE treatment, nasal steroids, and the severity of asthma. Positive correlations are red, and negative correlations are blue.

most negatively correlated with FEV1% predicted and AQLQ (**Figure 3**). These genes encode proteins related to calcium ion transmembrane transport (6.34%, p = 0.003, i.e., P2RY12, LOXHD1, CACNB4, and PKDREJ), apoptotic process (7.26%, p = 0.006, i.e., MTFP1, C8ORF4, PTPRH, and LGALS7B), O-glycan processing (4.46%, p = 1.33E-06, i.e., GALNT14, MUC1, and MUC2), and amino acid biosynthetic process (3%, p = 0.003, i.e., FOLH1, FOLH1B, and PYCR1). In 3 GMs (0, 1, 8), this expression increased with each step of disease severity: healthy control (Normal) < mild-to-moderate asthma not treated with ICS (mild-mod-noICS) < mild-to-moderate asthma treated with ICS (mild-mod-ICS) < severe asthma (severe) (**Figure 5**).

In GMs with downregulated genes, 8 GMs (1, 2, 4, 6, 7, 8, 12, and 14) were negatively correlated to the severity and negative association with use of ICS and OCS. They were also the most positively correlated with FEV1% predicted and AQLQ (**Figure 4**). These genes encode proteins related to cell adhesion (5.24%, p = 0.002, i.e., CD164, COL16A1, PRKCE, and KIAA1462), innate immune response (10%, p = 4.94E-04, i.e., C1QA, MARCO, and SAA1), potassium and sodium ion transmembrane transport (3.7%, p = 5.20E-4, i.e., KCNB1, KCNA1, SLC20A2, and SCN11A), lipoprotein metabolic process (5.62%, p = 2.05E-5, i.e., LRP1, APOC2, APOC1, LPL, and APOE), cellular oxidant detoxification (5.8%, p = 0.01, i.e., GSTM2, GPX3, and CYGB), and neuron signal (10%, p = 1.77E-4, i.e., TUBB2B, SPOCK1, and NRCA). In 4 GMs (2, 4, 7, and 8), this expression decreased with each step of disease severity: healthy control (Normal) < mild-to-moderate asthma not treated with ICS (mild-mod-noICS) < mild-to-moderate asthma treated with ICS (mild-mod-ICS) < severe asthma (SA) (**Figure 5**).

The comparison between mild-mod-ICS and mild-mod-noICS showed that ICS/OCS significantly reduced the expression of 3 GMs (2, 7, and 13) with upregulated genes, while increasing the expression of GM4 with upregulated modules.

The Severe Characteristics in AS

A total of 8 GMs (0-up, 1-up, 8-up, 11-up, 13-up, 8-down, 10-down, 12-down) were significantly different between the severe and non-severe group in specific phenotypes (**Table 2**), and 6 of them were shown in **Figure 6**. Compared to that in normal and non-severe samples, the GM0-up related to calcium ion *trans*-membrane transport (6.3%, p = 0.004, i.e., P2RY12, LOXHD1 and CACNB4) and negative regulation of endopeptidase activity (4.7%, p = 0.04, i.e., SERPINB3 and SERPINB4) was abnormally high expressed in severe asthma across all subtypes.

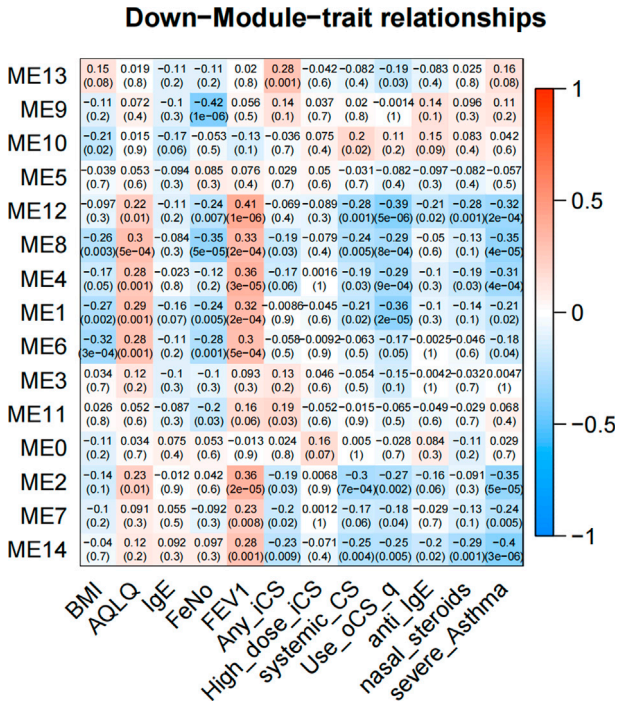

FIGURE 4 | The heatmap of correlation between co-expressed modules with downregulated genes and clinical characteristics. The clinical characteristics include BMI, AQLQ, IgE, fractional exhaled nitric oxide (FeNO), FEV1% predicted, use of inhaled (ICS) and oral corticosteroids (OCS), high dose use of ICS, systemic use CS, anti-IgE treatment, nasal steroids, and the severity of asthma. Positive correlations are red, and negative correlations are blue.

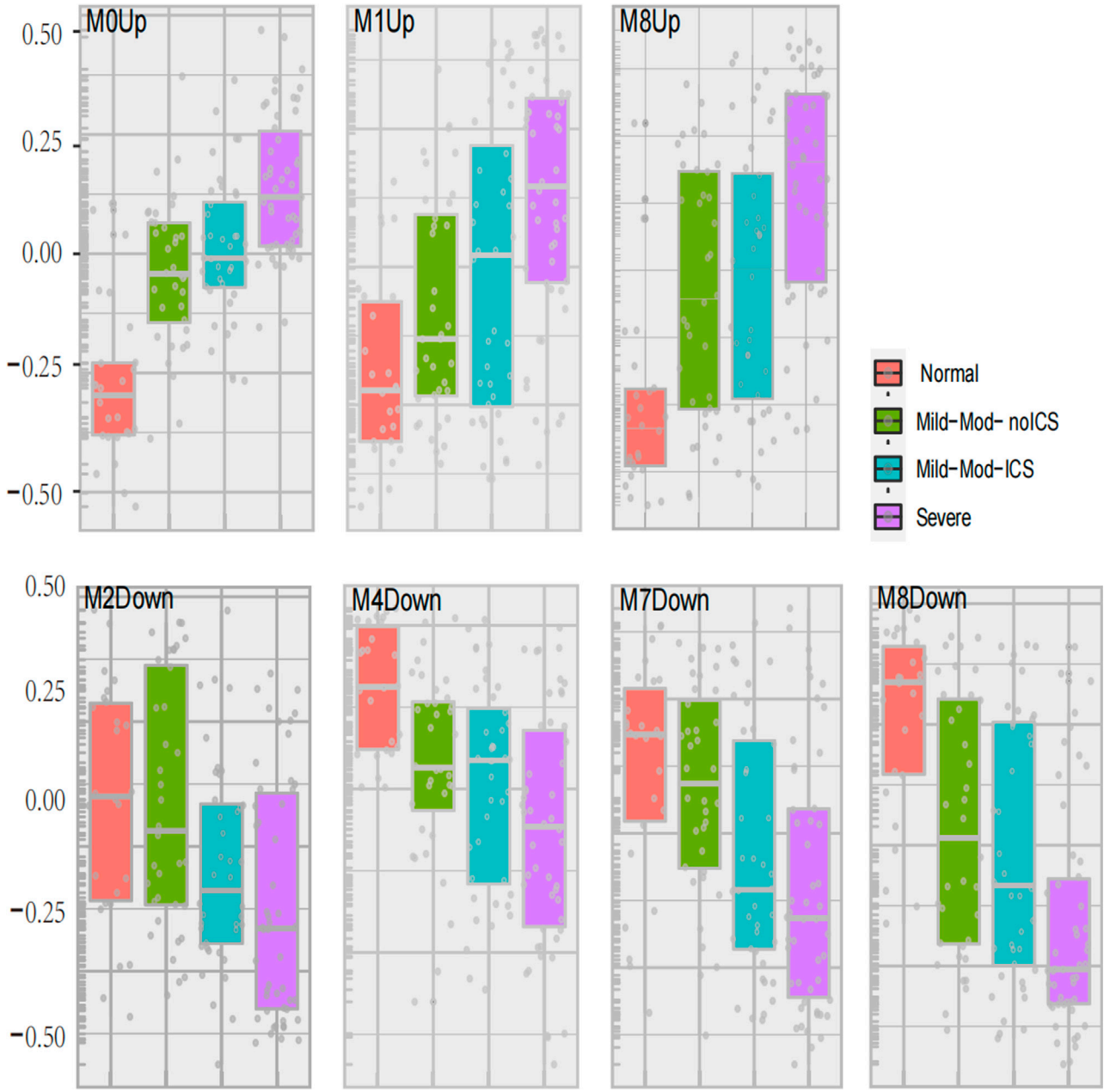

FIGURE 5 | Seven modules expression related to asthma severity. The geometric means were measured according to asthma severity. These classes included healthy controls (Normal), mild-to-moderate asthma not on inhaled corticosteroids (Mild-Mod-noICS), mild-to-moderate asthma on inhaled corticosteroids (Mild-Mod-ICS), and severe asthma (SA).

In "Phagocytosis-Th2", the GM10-down related to nitrogen compound metabolic process (2.06%, p = 0.02, i.e., VNN1 and VNN3) was differentially expressed between severe and non-severe groups. In "Mucin-Th2", the GM1-up related to O-glycan processing (4.7%, p = 1.33E-06, i.e., GALNT14, MUC1, and MUC2), GM8-up related to apoptotic process (17%, p = 0.01, i.e., LGALS7B, MAL, and SGK1), and GM8-down related to immune response (7.1%, p = 0.01, i.e., C3, CXCL6, IL6, and SUSD2) were significantly different between severe and non-severe groups. The GM1-up and GM8-up were much higher expressed while the GM8-down was lower expressed in severe asthma in the "Mucin-Th2" phenotype.

In "Neutrophils-Type" and "Interferon-Th1" (non-Th2) phenotypes, two specific modules (GM13-up and GM12-down) show different expression between severe asthma and non-severe asthma. The GM13-up related to interleukin-26 (IL26, PIK3R5, and LRRC2) was much higher expressed in severe individuals while the GM12-down module related to the nervous system (10%, p = 1.77E-

TABLE 2 | The differential expression of nine co-expression modules between severe and non-severe individuals in specific subtypes.

AS	Module	meanNonS	meanS	*p*-value
AS1	GM10-Down	−0.159	−0.52	0.02
AS1	GM0-Up	−0.036	0.157	0.01
AS3	GM0-Up	-0.031	0.148	0.008
AS3	GM11-Up	−0.237	0.252	0.016
AS3	GM13-Up	−0.203	0.367	0.005
AS4	GM8-Down	−0.436	−0.551	0.037
AS4	GM1-Up	0.13	0.434	0.018
AS4	GM8-Up	0.298	0.554	0.007
AS5	GM12-Down	0.154	−0.478	8.82E-06
AS5	GM13-Up	−0.314	0.191	0.016

04, i.e., TUBB2B, SPOCK1, and ASCL1) was much lower expressed in severe asthma individuals. The similar results were shown in the validation data (**Figure 7**).

DISCUSSION

Asthma is a heterogeneous disease with multiple immune and non-immune mechanisms (Ito et al., 2004; McKinley et al., 2008). This study shows that the transcriptome of bronchial epithelial cells was related to ASs and the severity.

Using cluster analysis, five stable ASs were obtained. Multiple clinical features had significant differences among these ASs. These results were similar to that of previous studies (Langfelder and Horvath, 2008). Using the expression of 2,664 DEG profiles as input, the WGCNA co-expression analysis was performed and the module was split based on the dysregulated direction of each gene; 30 GMs were obtained, 15 of which contain only upregulated genes and the other 15 contain downregulated genes (**Table 3**). These split modules were essential for the description of the typical characteristics in ASs. In each co-expressed module, there was a negative correlation between the two gene sets with up- and

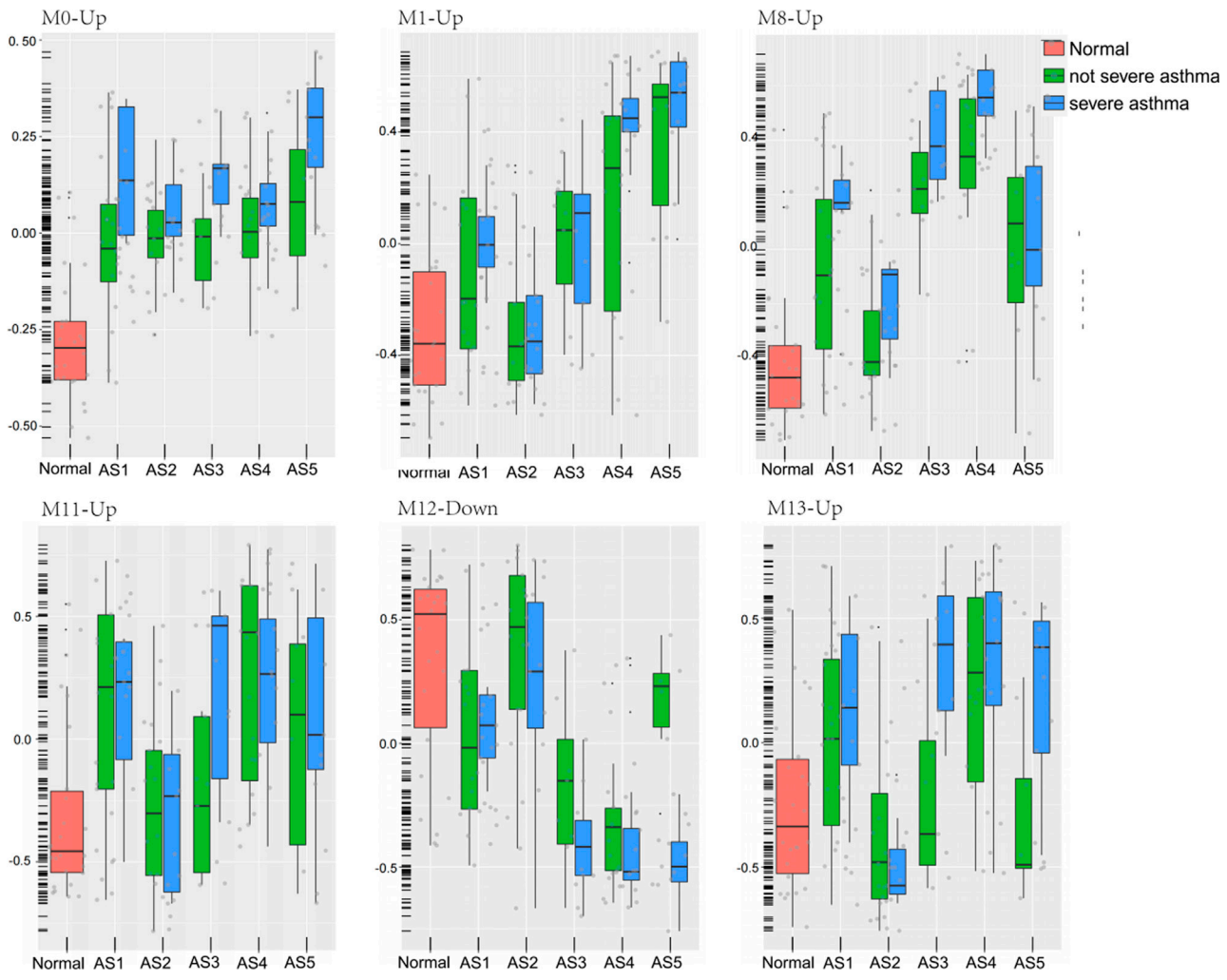

FIGURE 6 | Six modules expression in each phenotype between severe and non-severe groups. The geometric means were measured according to each cluster. These classes included healthy controls (Normal), five asthma phenotypes (AS1–5), non-severe asthma, and severe asthma (SA).

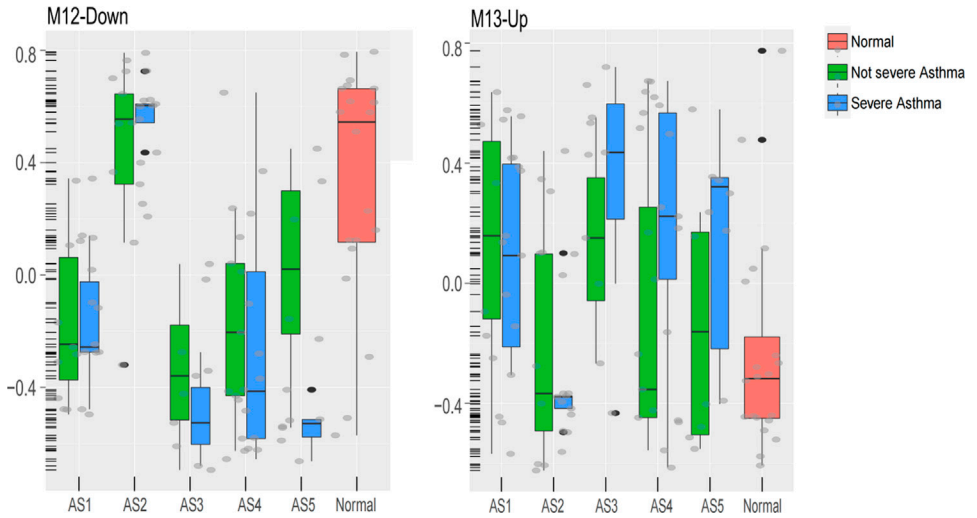

FIGURE 7 | The expression of M13-Up and M12-Down modules in clusters between severe and non-severe groups.

TABLE 3 | The number of upregulated and downregulated genes in co-expressed modules.

	Up	Down
Module0	66	86
Module1	185	354
Module2	210	42
Module3	190	85
Module4	194	28
Module5	154	58
Module6	108	56
Module7	118	24
Module8	32	109
Module9	103	25
Module10	8	117
Module11	72	4
Module12	3	50
Module13	43	2
Module14	27	11

downregulated genes, respectively. For example, in GM9 with upregulated genes, multiple Th2-related marker genes (CLCA1, POSTN, etc.) were included, while in GM9 with downregulated genes, the MUC5B (Zhang et al., 2019), SLC28A3, and CSGALNACT1 genes were included and closely related to the reduction of airway defense response (Ridley and Thornton, 2018; Rojas et al., 2019). The MUC5B plays a key role in airway defense. The lack of MUC5B leads to lung inflammation, impaired immune balance, and chronic infection mediated by a variety of bacteria (Roy et al., 2014). These two aspects (up and down features) might be two effective strategies for individualized treatment in asthma.

Among these 15 modules containing upregulated genes, 4 GMs (2, 3, 5, and 7) were typical immune-related modules and had obvious subtype distribution characteristics. The GM2 and GM3 were typical characteristics of "Phagocytosis-Th2" while the GM5 and GM7 were the typical features of "Interferon-Th1". The function enrichment analysis showed that the GM2 and GM3 were mainly related to the receptor-mediated endocytosis and antigen presentation *via* MHC-II. The GM5 and GM7 were highly expressed in "Interferon-Th1" and mainly related to type I interferon response and T-cell toxicity. Although the "Mucin-Th2" was a typical Th2 phenotype, the four immune-related modules mentioned above were not significantly increased in this cluster, while the increased glycosylation of O-type glycans, amino sugar and nucleoside sugar metabolism, proteolysis, and unfolded protein reaction were highly expressed in this cluster and positively correlated to the expression of Th2 markers. It suggested that these biological functions could be coordinated with the Th2 signals and related to the physio-pathological mechanisms in "Mucin-Th2". The increased mucus in airway was a typical clinical feature of Th2 asthma (Dunican et al., 2018; Lambrecht et al., 2019). Studies had shown that the galectin-10 (Galectin-10 and Gal10) crystal structure in airway mucus stimulates the immune system and induces the changes in airway inflammation and mucus secretion (Nyenhuis et al., 2019; Persson et al., 2019). Clearing the crystal structure can effectively relieve airway inflammation and asthma symptoms.

Correlation analysis showed that the apoptotic process and O-glycan processing were positively correlated to asthma severity, while the cell adhesion, innate immune response, potassium and sodium ion transmembrane transport, cellular oxidant detoxification, and neuron signal were negatively correlated to asthma severity. Correcting the imbalance of oxidation and anti-oxidation in the lung may be an important method to relieve asthma symptoms in clinic. GSH is the most important antioxidant in lung tissues (Brigelius-Flohé and Maiorino, 2013). GSH can inhibit a variety of pathogen replication and survival, and increasing the GSH can effectively improve the body's ability to resist foreign microorganisms (Fitzpatrick et al., 2012). The inhibition of the activity of the CYP450 pathway could destroy the phagocytosis of macrophage and reduces the clearance efficiency of inflammatory stimuli (Bystrom et al., 2013).

In "Phagocytosis-Th2", the GM10-down was differentially expressed between severe and non-severe groups. In "Mucin-Th2", the O-glycan processing (GM1-Up), apoptotic process (GM8-Up), and oxidation–reduction process (GM8-down) were significantly different between severe and non-severe groups. The oxygen free radical increase in bronchoalveolar lavage fluid and peripheral blood associated with the severity of disease (Mossberg et al., 2009; Sangiuolo et al., 2015), especially in a typical Th2 phenotype (Huang et al., 2019).

In "Neutrophils-Type" and "Interferon-Th1" (non-Th2) phenotype, interleukin-26 (IL26)-related function was upregulated and related to the severity of asthma. IL-26 is a member of IL-10 cytokine family, is abundant in human airways, and induces the production of pro-inflammatory cytokines (Louhaichi et al., 2020). Stimulation of cultured $CD4^+$ T cells with monocyte by recombining IL-26 promoted the generation of RORγ $Th17^+$ cells, inducing the production of IL-17A, IL-1β, IL-6, and TNF-α (Louhaichi et al., 2020). Therefore, IL-26 could appear as a novel pro-inflammatory cytokine, produced in airways, and may be a promising target to treat inflammatory asthma.

Although we clustered the transcriptome of bronchial epithelial cells and revealed the typical features in five stable subtypes, the heterogeneity in asthma is much higher than that in subtypes. The characteristics in individuals were more likely a mixture of typical features in multiple subtypes. For asthma, distinguishing subtypes was only a powerful method for uncovering the heterogeneity of complex diseases. Therefore, the individualized analysis based on phenotypes in asthma was a powerful tool for individualized diagnosis and treatment.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article.

AUTHOR CONTRIBUTIONS

Acquisition of data: BL accessed to the transcriptom data online; Conception and design: YZ, X-FJ, BL; Analysis and interpretation: BL, W-YZ, W-XS, YZ, LQ, Y-MH, W-QL, DL; Wrote the article: BL, W-XS; Approved and edited the article: All authors approved the article.

FUNDING

This study is funded by the National Natureal science foundation number 81171657, No,30371364.

ACKNOWLEDGMENTS

Thanks to Sally E. Wenzel in the Department of Environmental and Occupational Health, the Director of the University of Pittsburgh Asthma Institute, and the SARP investigators and patients for providing the clinical data. Thanks to Professor Jiang Meng from school of Computer Science of Harbin Institute of Technology for his selfless help in article analysis and writing.

REFERENCES

Belsky, D. W., Sears, M. R., Hancox, R. J., Harrington, H., Houts, R., Moffitt, T. E., et al. (2013). Polygenic Risk and the Development and Course of Asthma: an Analysis of Data from a Four-Decade Longitudinal Study. *Lancet Respir. Med.* 1 (6), 453–461. doi:10.1016/s2213-2600(13)70101-2

Binder, E. B. (2009). The Role of FKBP5, a Co-chaperone of the Glucocorticoid Receptor in the Pathogenesis and Therapy of Affective and Anxiety Disorders. *Psychoneuroendocrinology* 34 (Suppl. 1), S186–S195. doi:10.1016/j.psyneuen.2009.05.021

Brigelius-Flohé, R., and Maiorino, M. (2013). Glutathione Peroxidases. *Biochim. Biophys. Acta (Bba) - Gen. Subjects* 1830 (5), 3289–3303. doi:10.1016/j.bbagen.2012.11.020

Bystrom, J., Thomson, S. J., Johansson, J., Edin, M. L., Zeldin, D. C., Gilroy, D. W., et al. (2013). Inducible CYP2J2 and its Product 11,12-EET Promotes Bacterial Phagocytosis: a Role for CYP2J2 Deficiency in the Pathogenesis of Crohn's Disease? *PLoS One* 8 (9), e75107. doi:10.1371/journal.pone.0075107

Diao, W.-q., Shen, N., Du, Y.-p., Liu, B.-b., Sun, X.-y., Xu, M., et al. (2016). Fetuin-B (FETUB): a Plasma Biomarker Candidate Related to the Severity of Lung Function in COPD. *Sci. Rep.* 6, 30045. doi:10.1038/srep30045

Dunican, E. M., Elicker, B. M., Gierada, D. S., Nagle, S. K., Schiebler, M. L., Newell, J. D., et al. (2018). Mucus Plugs in Patients with Asthma Linked to Eosinophilia and Airflow Obstruction. *J. Clin. Invest.* 128 (3), 997–1009. doi:10.1172/jci95693

Fitzpatrick, A. M., Jones, D. P., and Brown, L. A. S. (2012). Glutathione Redox Control of Asthma: from Molecular Mechanisms to Therapeutic Opportunities. *Antioxid. Redox Signaling* 17 (2), 375–408. doi:10.1089/ars.2011.4198

Gao, J., Wu, M., Wang, F., Jiang, L., Tian, R., Zhu, X., et al. (2020). CD74, a Novel Predictor for Bronchopulmonary Dysplasia in Preterm Infants. *Medicine (Baltimore)* 99 (48), e23477. doi:10.1097/md.0000000000023477

Global, Regional, and National Deaths, Prevalence, Disability-Adjusted Life Years, and Years Lived with Disability for Chronic Obstructive Pulmonary Disease and Asthma, 1990-2015: a Systematic Analysis for the Global Burden of Disease Study 2015. *Lancet Respir. Med.* 2017;5(9):691–706.doi:10.1016/S2213-2600(17)30293-X

Hamilton, L. M., Davies, D. E., Wilson, S. J., Kimber, I., Dearman, R. J., and Holgate, S. T. (2001). The Bronchial Epithelium in Asthma-Mmuch More Than a Passive Barrier. *Monaldi Arch. Chest Dis.* 56 (1), 48–54.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data. *BMC Bioinformatics* 14, 7. doi:10.1186/1471-2105-14-7

Higgins, J. C. (2003). The 'crashing asthmatic.'. *Am. Fam. Physician* 67 (5), 997–1004. doi:10.1080/02724634.2003.10010569

Huang, W.-C., Liu, C.-Y., Shen, S.-C., Chen, L.-C., Yeh, K.-W., Liu, S.-H., et al. (2019). Protective Effects of Licochalcone A Improve Airway Hyper-Responsiveness and Oxidative Stress in a Mouse Model of Asthma. *Cells* 8 (6), 617. doi:10.3390/cells8060617

Ito, K., Hanazawa, T., Tomita, K., Barnes, P. J., and Adcock, I. M. (2004). Oxidative Stress Reduces Histone Deacetylase 2 Activity and Enhances IL-8 Gene Expression: Role of Tyrosine Nitration. *Biochem. Biophysical Res. Commun.* 315 (1), 240–245. doi:10.1016/j.bbrc.2004.01.046

Itoh, A., Uchiyama, A., Taniguchi, S., and Sagara, J. (2014). Phactr3/scapinin, a Member of Protein Phosphatase 1 and Actin Regulator (Phactr) Family, Interacts with the Plasma Membrane via Basic and Hydrophobic Residues in the N-Terminus. *PLoS One* 9 (11), e113289. doi:10.1371/journal.pone.0113289

Kim, H. Y., DeKruyff, R. H., and Umetsu, D. T. (2010). The many Paths to Asthma: Phenotype Shaped by Innate and Adaptive Immunity. *Nat. Immunol.* 11 (7), 577–584. doi:10.1038/ni.1892

Lambrecht, B. N., Hammad, H., and Fahy, J. V. (2019). The Cytokines of Asthma. *Immunity* 50 (4), 975–991. doi:10.1016/j.immuni.2019.03.018

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining Clusters from a Hierarchical Cluster Tree: the Dynamic Tree Cut Package for R. *Bioinformatics* 24 (5), 719–720. doi:10.1093/bioinformatics/btm563

Lasky-Su, J., Himes, B. E., Raby, B. A., Klanderman, B. J., Sylvia, J. S., Lange, C., et al. (2012). HLA-DQ Strikes Again: Genome-wide Association Study Further confirmsHLA-DQin the Diagnosis of Asthma Among Adults. *Clin. Exp. Allergy* 42 (12), 1724–1733. doi:10.1111/cea.12000

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The Sva Package for Removing Batch Effects and Other Unwanted Variation in High-Throughput Experiments. *Bioinformatics* 28 (6), 882–883. doi:10.1093/bioinformatics/bts034

Louhaichi, S., Mlika, M., Hamdi, B., Hamzaoui, K., and Hamzaoui, A. (2020). Sputum IL-26 Is Overexpressed in Severe Asthma and Induces Proinflammatory Cytokine Production and Th17 Cell Generation: A Case-Control Study of Women. *Jaa* Vol. 13, 95–107. doi:10.2147/jaa.s229522

McKinley, L., Alcorn, J. F., Peterson, A., DuPont, R. B., Kapadia, S., Logar, A., et al. (2008). TH17 Cells Mediate Steroid-Resistant Airway Inflammation and Airway Hyperresponsiveness in Mice. *J. Immunol.* 181 (6), 4089–4097. doi:10.4049/jimmunol.181.6.4089

Modena, B. D., Tedrow, J. R., Milosevic, J., Bleecker, E. R., Meyers, D. A., Wu, W., et al. (2014). Gene Expression in Relation to Exhaled Nitric Oxide Identifies Novel Asthma Phenotypes with Unique Biomolecular Pathways. *Am. J. Respir. Crit. Care Med.* 190 (12), 1363–1372. doi:10.1164/rccm.201406-1099oc

Mossberg, N., Movitz, C., Hellstrand, K., Bergström, T., Nilsson, S., and Andersen, O. (2009). Oxygen Radical Production in Leukocytes and Disease Severity in Multiple Sclerosis. *J. Neuroimmunol* 213 (1-2), 131–134. doi:10.1016/j.jneuroim.2009.05.013

Nyenhuis, S. M., Alumkal, P., Du, J., Maybruck, B. T., Vinicky, M., and Ackerman, S. J. (2019). Charcot-Leyden crystal Protein/galectin-10 Is a Surrogate Biomarker of Eosinophilic Airway Inflammation in Asthma. *Biomarkers Med.* 13 (9), 715–724. doi:10.2217/bmm-2018-0280

Persson, E. K., Verstraete, K., Heyndrickx, I., Gevaert, E., Aegerter, H., Percier, J.-M., et al. (2019). Protein Crystallization Promotes Type 2 Immunity and Is Reversible by Antibody Treatment. *Science* 364 (6442), eaaw4295. doi:10.1126/science.aaw4295

Ridley, C., and Thornton, D. J. (2018). Mucins: the Frontline Defence of the Lung. *Biochem. Soc. Trans.* 46 (5), 1099–1106. doi:10.1042/bst20170402

Rojas, D. A., Iturra, P. A., Méndez, A., Ponce, C. A., Bustamante, R., Gallo, M., et al. (2019). Increase in Secreted Airway Mucins and Partial Muc5b STAT6/FoxA2 Regulation during Pneumocystis Primary Infection. *Sci. Rep.* 9 (1), 2078. doi:10.1038/s41598-019-39079-4

Roy, M. G., Livraghi-Butrico, A., Fletcher, A. A., McElwee, M. M., Evans, S. E., Boerner, R. M., et al. (2014). Muc5b Is Required for Airway Defence. *Nature* 505 (7483), 412–416. doi:10.1038/nature12807

Sánchez-Ovando, S., Baines, K. J., Barker, D., Wark, P. A., and Simpson, J. L. (2020). Six Gene and TH2 Signature Expression in Endobronchial Biopsies of Participants with Asthma. *Immun. Inflamm. Dis.* 8 (1), 40–49. doi:10.1002/iid3.282

Sangiuolo, F., Puxeddu, E., Pezzuto, G., Cavalli, F., Longo, G., Comandini, A., et al. (2015). HFE Gene Variants and Iron-Induced Oxygen Radical Generation in Idiopathic Pulmonary Fibrosis. *Eur. Respir. J.* 45 (2), 483–490. doi:10.1183/09031936.00104814

Voraphani, N., Gladwin, M. T., Contreras, A. U., Kaminski, N., Tedrow, J. R., Milosevic, J., et al. (2014). An Airway Epithelial iNOS-DUOX2-Thyroid Peroxidase Metabolome Drives Th1/Th2 Nitrative Stress in Human Severe Asthma. *Mucosal Immunol.* 7 (5), 1175–1185. doi:10.1038/mi.2014.6

Whitsett, J. A. (2018). Airway Epithelial Differentiation and Mucociliary Clearance. *Ann. Am. Thorac. Soc.* 15 (Suppl. 3), S143–S148. doi:10.1513/AnnalsATS.201802-128AW

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* 26 (12), 1572–1573. doi:10.1093/bioinformatics/btq170

Wilson, D. H., Adams, R. J., Ruffin, R. E., Tucker, G., Taylor, A. W., and Appleton, S. (2006). Trends in Asthma Prevalence and Population Changes in South Australia, 1990-2003. *Med. J. Aust.* 184 (5), 226–229. doi:10.5694/j.1326-5377.2006.tb00207.x

Woodruff, P. G., Boushey, H. A., Dolganov, G. M., Barker, C. S., Yang, Y. H., Donnelly, S., et al. (2007). Genome-wide Profiling Identifies Epithelial Cell Genes Associated with Asthma and with Treatment Response to Corticosteroids. *Proc. Natl. Acad. Sci.* 104 (40), 15858–15863. doi:10.1073/pnas.0707413104

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, Q., Wang, Y., Qu, D., Yu, J., and Yang, J. (2019). The Possible Pathogenesis of Idiopathic Pulmonary Fibrosis Considering MUC5B. *Biomed. Res. Int.* 2019, 9712464. doi:10.1155/2019/9712464

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 November 2021
doi: 10.3389/fgene.2021.782005

Integrative Analysis for Elucidating Transcriptomics Landscapes of Systemic Lupus Erythematosus

*Haihong Zhang[1], Yanli Wang[1], Jinghui Feng[2], Shuya Wang[1], Yan Wang[1], Weisi Kong[1] and Zhiyi Zhang[1]**

[1]Department of Rheumatology and Immunology, The First Affiliated Hospital of Harbin Medical University, Harbin, China, [2]Department of Gerontology, The First Affiliated Hospital of Harbin Medical University, Harbin, China

Systemic lupus erythematosus (SLE) is a complex and heterogeneous autoimmune disease that the immune system attacks healthy cells and tissues. SLE is difficult to get a correct and timely diagnosis, which makes its morbidity and mortality rate very high. The pathogenesis of SLE remains to be elucidated. To clarify the potential pathogenic mechanism of SLE, we performed an integrated analysis of two RNA-seq datasets of SLE. Differential expression analysis revealed that there were 4,713 and 2,473 differentially expressed genes, respectively, most of which were up-regulated. After integrating differentially expressed genes, we identified 790 common differentially expressed genes (DEGs). Gene functional enrichment analysis was performed and found that common differentially expressed genes were significantly enriched in some important immune-related biological processes and pathways. Our analysis provides new insights into a better understanding of the pathogenic mechanisms and potential candidate markers for systemic lupus erythematosus.

Keywords: systemic lupus erythematosus, differential expression analysis, gene functional enrichment analysis, RNA-seq, protein-protein interaction

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Chen Qingfeng,
Guangxi University, China
Shihua Zhang,
Wuhan University of Science and Technology, China

****Correspondence:***
Zhiyi Zhang
zhangzhiyi2014@163.com

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

Received: *23 September 2021*
Accepted: *20 October 2021*
Published: *04 November 2021*

Citation:
Zhang H, Wang Y, Feng J, Wang S, Wang Y, Kong W and Zhang Z (2021) Integrative Analysis for Elucidating Transcriptomics Landscapes of Systemic Lupus Erythematosus. Front. Genet. 12:782005. doi: 10.3389/fgene.2021.782005

INTRODUCTION

Systemic lupus erythematosus is a chronic autoimmune disease (Beccastrini et al., 2013; Davies et al., 2021). Its clinical manifestations are heterogeneous and involve one or more organs such as skin, kidney, joints, and nervous system (Von Feldt, 1995; Adinolfi et al., 2016; Ronco et al., 2021). The latest data from the US Lupus Registry and published studies around the world can more accurately estimate the incidence and prevalence of SLE. It is estimated that the incidence of 23.2 cases per 100,000 people in North America is the highest in the world (Tsokos, 2011; Rees et al., 2017). SLE is a heterogeneous rheumatic systemic disease with extremely diverse clinical manifestations and diverse pathogenesis (Wu et al., 2021). In addition, it is one of the most varied diseases in its epidemiology and etiology, with different types of immune dysfunction (Oku and Atsumi, 2018). SLE patients' immune system activation is characterized by exaggerated B cells and T cells responses (Tsokos, 2011). The health-related quality of life of SLE patients is significantly impaired (Di Battista et al., 2018). To obtain a better diagnosis and treatment method, it is necessary to explore the pathogenesis of SLE.

Since the successful application of high-throughput technology, it has been widely used in almost all biological research fields (Hess et al., 2020). With the development of high-throughput technology (Hess et al., 2020), biological research has been transformed from a single gene level to a full

transcriptome level, which has greatly advanced many research fields in biology (Wang et al., 2009; McDermaid et al., 2019). Cheng. et al. based on the genome-wide expression data of peripheral blood mononuclear cells (PBMC) of SLE patients found a novel marker of SLE (Cheng et al., 2021). Jiang. et al. discovered a new type of lncRNA that plays an important role in the pathogenesis of SLE based on the whole transcriptome data of PBMC of SLE patients (Jiang et al., 2021). However, these studies were only conducted on a single dataset, and there was heterogeneity between different datasets. Therefore, through a comprehensive analysis of multiple datasets, more robust results will be obtained.

In this study, we conducted a systematic analysis of two gene expression datasets of SLE. First, differential expression analysis was performed to obtain differentially expressed genes (DEGs) in each dataset. To obtain robust results, we intersected the DEGs s of the two datasets. We found that 790 genes were differentially expressed in both datasets. Finally, gene function enrichment analysis showed that common DEGs were enriched in immune-related biological pathways. Overall, our research provided new insight into the molecular mechanism of SLE.

MATERIALS AND METHODS

Datasets

"Systemic Lupus Erythematosus" and "RNA-seq" were used as the keywords for searching the GEO database. The gene expression datasets of PBMC from freshly isolated healthy controls and SLE patients were downloaded from the GEO database (GSE162828 and GSE169080), the platforms used were GPL24676, and GPL20795. GSE162828 included 10 samples of peripheral blood mononuclear cells and was divided into the SLE group (5 samples) and healthy controls group (5 samples). GSE169080 included seven samples of peripheral blood mononuclear cells and was divided into SLE group (4 samples) and healthy controls group (3 samples) (Clough and Barrett, 2016; Cheng et al., 2021; Jiang et al., 2021).

Data Pre-processing and Identification of Differential Expressed Genes

R package DESeq2 (1.26.0) was used for the analysis of the original datasets (Love et al., 2014). $|\log FC| > 1$ and p. adj <0.05 were defined as the cutoff values for further analysis of DEGs. Volcano and heatmap were constructed by R package ggplot2. Venn plot (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to draw the intersection of two databases.

Analyzing of DEGs on Protein-Protein Interaction Network

Protein-protein interaction (PPI) network analysis helps to study the molecular mechanism of diseases from a systematic perspective and discover new drug targets (Wu et al., 2019). STRING (https://string-db.org/) is a database covering more than 5,000 organisms with known and predicted protein-protein interactions, providing direct (physical) and indirect (functional) associations (Szklarczyk et al., 2017). We used String (https://string-db.org/) to generate biological networks for proteins, and the results were analyzed by Cytoscape (Shannon et al., 2003; Szklarczyk et al., 2017).

Gene Functional Enrichment Analysis

Gene Ontology (GO) is an ontology widely used in the field of bioinformatics, which covers three aspects of biology: biological process (BP), cellular component (CC), and molecular function (MF) (Thomas, 2017). Kyoto Encyclopedia of Genes and Genomes (KEGG) is a biological system advanced function and utility database based on molecular-level information from genome sequencing and other high-throughput experimental technologies (Kanehisa et al., 2017). In this study, R package clusterProfiler was used to perform GO functional annotation and KEGG pathway enrichment analysis for DEGs (Yu et al., 2012).

RESULTS

Differentially Expressed Genes Between SLE Patients and Healthy Controls

To obtain abnormally expressed genes in SLE patients, we separately analyzed the differential expression of two GEO datasets (GSE162828 and GSE169080). As shown in **Figure 1A**, there were 4,713 DEGs, including 2,717 up-regulated and 1,996 down-regulated in the GSE162828 dataset. In the GSE169080 dataset, there were 2,473 DEGs, including 1,552 up-regulated and 921 down-regulated (**Figure 1B**). In both datasets, the number of up-regulated DEGs was more than the number of down-regulated DEGs (**Figure 1C**). In the GSE162828 dataset, the up-regulated DEGs accounted for 56.7% of all DEGs. At the same time, the up-regulated DEGs accounted for 62.8% of all DEGs in the GSE169080 dataset. The trends in the two datasets were roughly the same.

In addition, the heatmap showed that DEGs can group samples by sample type, namely SLE patients (SLE) and healthy controls (Norm) (**Figures 1D,E**). These genes were highly concordant within groups. The expression level of these genes between SLE patients and healthy controls exhibited a large difference in both databases.

Identification of Common Differentially Expressed Genes by Integrated Analysis

Due to the heterogeneity between different datasets, the analysis results of different datasets may have certain differences (Ying et al., 2020). The gene expression in different samples may be different (Bao et al., 2021). To avoid this problem, integrating multiple datasets and a large number of samples help obtain more

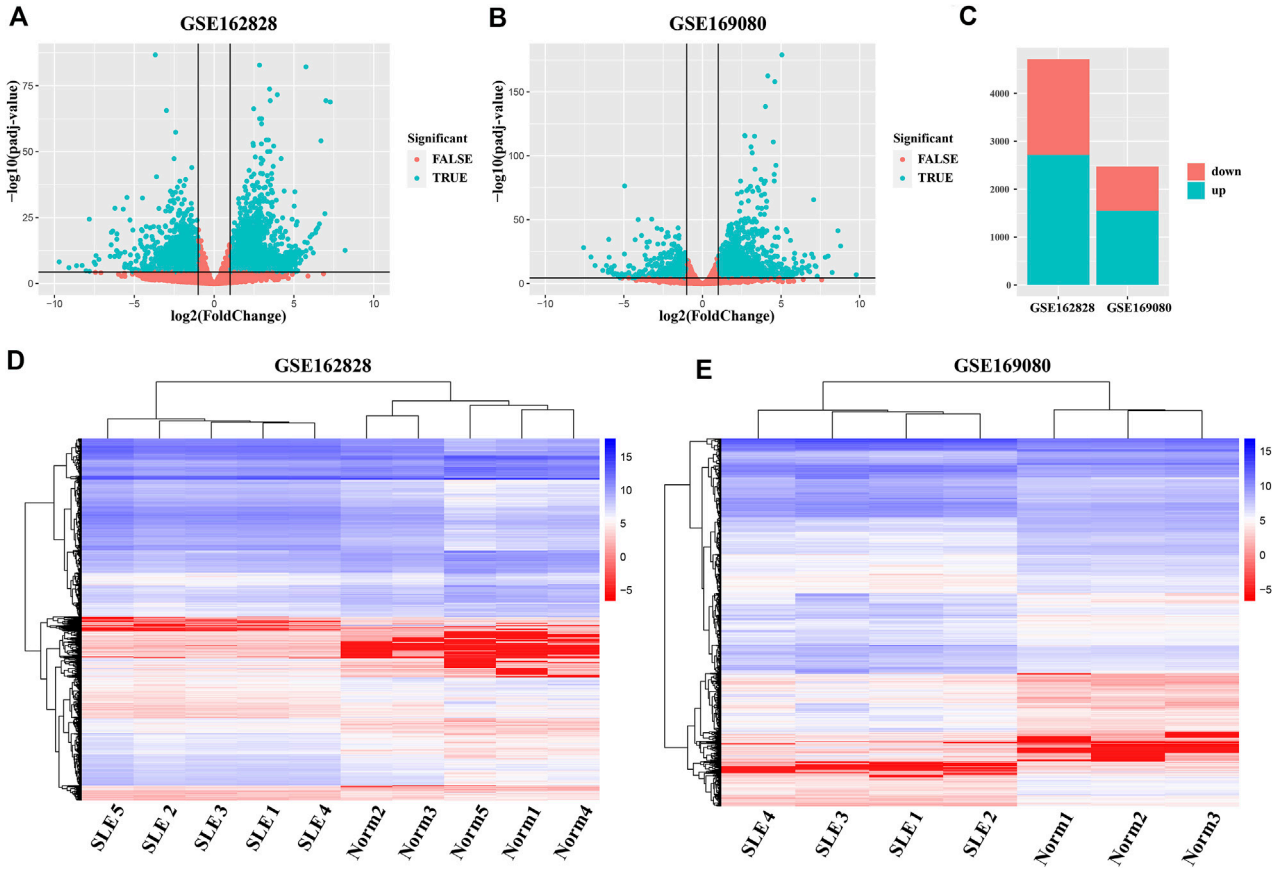

FIGURE 1 | Analysis of differentially expressed genes (DEGs) between SLE patients and healthy controls. **(A,B)** The volcano plots exhibited the differentially expressed genes in SLE patients groups compared to healthy controls groups. Each dot in the figure represented one gene. The green dots indicated the differentially expressed genes, while the red dots denoted no significant difference. **(C)** Barplot showed the number of DEGs whose expression levels were up-regulated (green) and down-regulated (red) in the two datasets. **(D, E)** Hierarchy Clustering Analysis. Repeated samples are clustered together, indicating the repeatability of samples and the differences between samples.

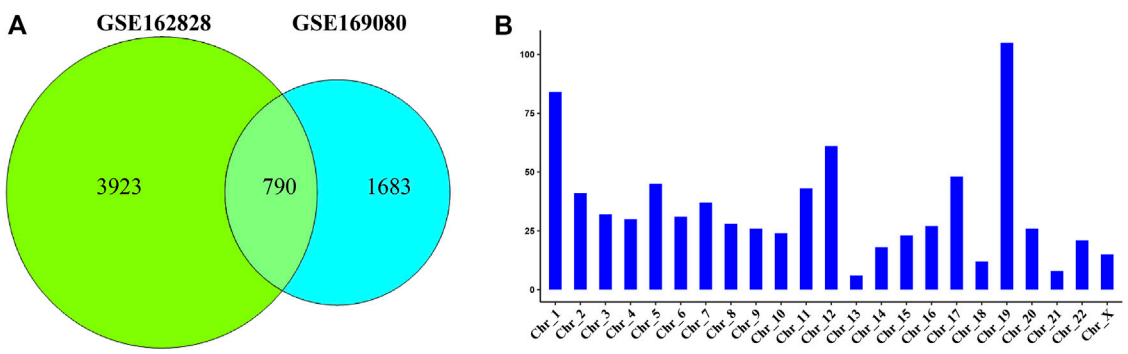

FIGURE 2 | Common differentially expressed genes. **(A)** Venn plot showed the intersection of DEGs in two datasets **(B)** The histogram showed the distribution of DEGs on chromosomes.

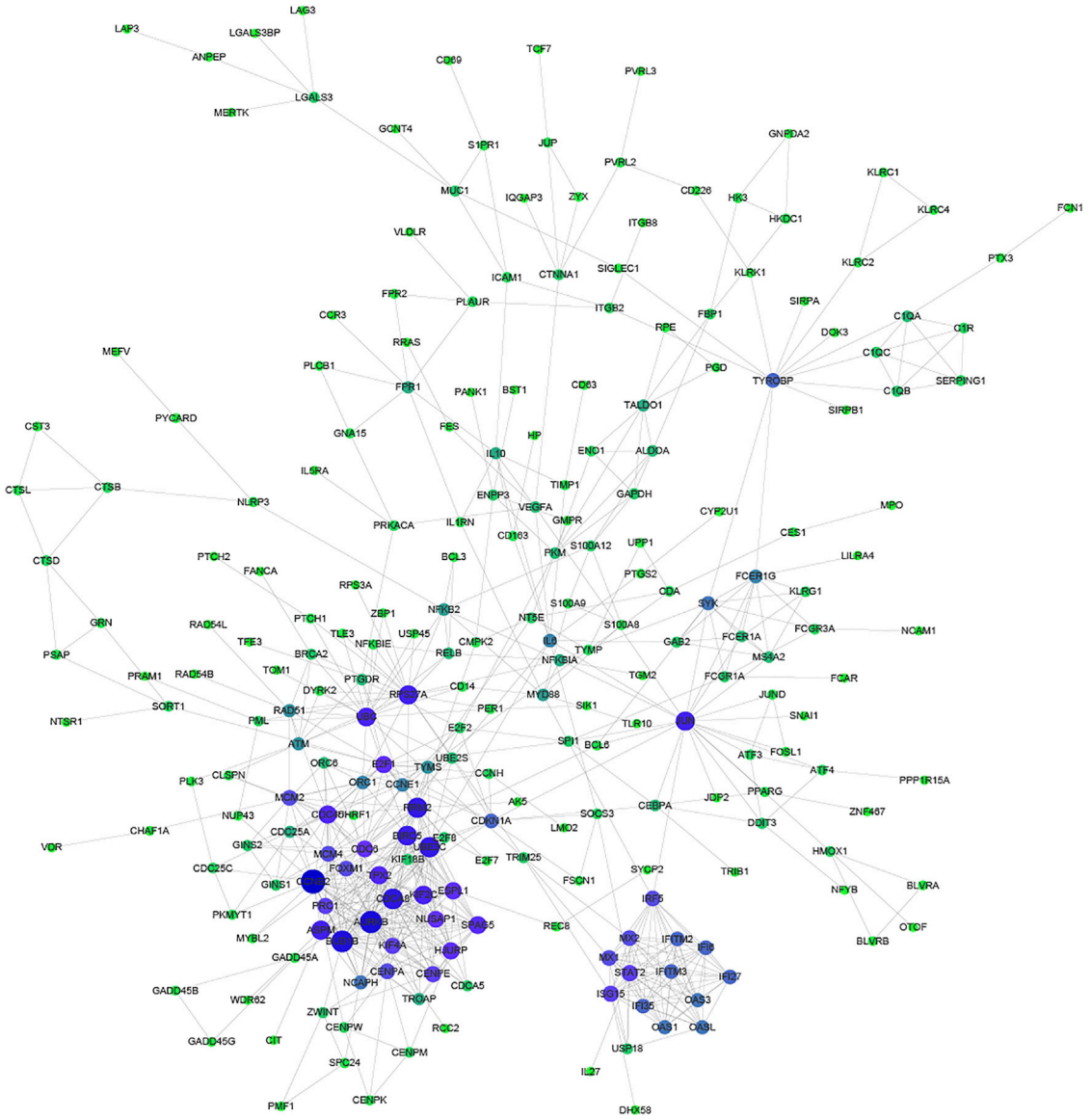

FIGURE 3 | Protein-Protein Interaction network of common DEGs. The size and color of the node depending on the degree, the larger the degree, the larger the node.

solid results (Kou et al., 2020). In this study, we integrated DEGs from two datasets to obtain common DEGs.

The Venn diagram showed that 790 DEGs were shared between the two datasets (**Figure 2A**). They accounted for 16.8% (GSE162828) and 31.9% (GSE169080) of the two datasets, respectively. There were 3,923 DEGs only in the GSE162828 dataset, and 1,683 DEGs only in GSE169080 dataset. This may be caused by different sequencing technologies and sample heterogeneity.

We defined these 790 DEGs as common DEGs. To further explore the distribution of common DEGs on the chromosomes, we had made statistics on the chromosomal locations of these genes. As shown in **Figure 2B**, we found that these genes were distributed on every chromosome. Most of these genes were distributed on chromosome 19. On the contrary, they were only 6 DEGs on chromosome 13.

Analysis of Common Differentially Expressed Genes on Protein-Protein Interaction Network

Proteins usually perform biological functions in concert. It has been shown that there is a close relationship between Protein-Protein Interaction (PPI) and the biological

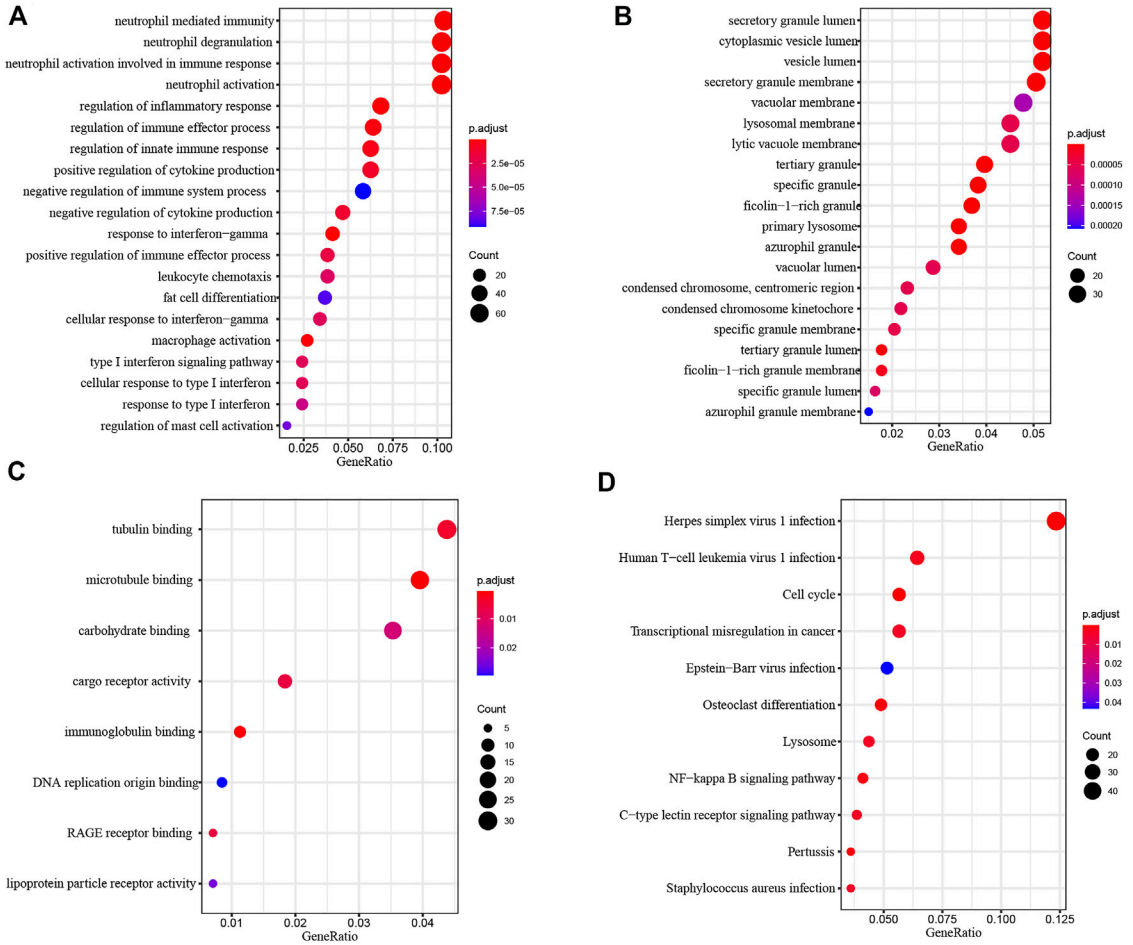

FIGURE 4 | Functional Enrichment Analysis of common DEGs. **(A)** Biological process analysis of common DEGs. **(B)** Cellular component analysis of common DEGs. **(C)** Molecular function analysis of common DEGs. **(D)** KEGG analysis of common DEGs.

functions of gene/protein clusters (Li H. et al., 2019). To further analyze the correlation between common DEGs, STRING and Cytoscape were used to construct the PPI network (**Figure 3**). Part of common DEGs was predicted to have a strong association with other genes. The size and color of the node depending on the degree, the larger the degree, the larger the node.

Especially, *CCNB2*, *CDCA8*, *AURKB*, *BUB1B*, *RRM2*, *BIRC5*, and *UBE2C* had the largest degree. *CCNB2* is an essential component of the cell cycle regulatory machinery (Takashima et al., 2014; Li R. et al., 2019). *CDCA8* is an essential regulator of mitosis and cell division (Zhang et al., 2020). *AURKB* participates in the regulation of alignment and segregation of chromosomes during mitosis and meiosis through association with microtubules (Ahmed et al., 2021). *BUB1B* encodes a kinase involved in the spindle checkpoint function (Zhang et al., 2021). *RRM2* encodes one of two non-identical subunits for ribonucleotide reductase (Mazzu et al., 2020). *BIRC5* encodes negative regulatory proteins that prevent apoptotic cell death (Adamopoulos et al., 2021). *UBE2C* is required for the destruction of mitotic cyclins and cell cycle progression (Jin et al., 2020).

Functional Enrichment Analysis of Common Differentially Expressed Genes

To investigate the biological function of common DEGs, we used clusterProfiler to perform Functional enrichment analysis. Biological Process (BP) enrichment showed that the common DEGs were enriched in neutrophil mediated immunity, neutrophil degranulation, neutrophil activation involved in immune response, neutrophil activation and regulation of inflammatory response (**Figure 4A**). Cellular Component (CC) enrichment showed that the common DEGs were mainly enriched in secretory granule lumen, cytoplasmic vesicle lumen, vesicle lumen, secretory granule membrane and vacuolar membrane (**Figure 4B**). Molecular Function (MF) enrichment showed that the common DEGs were significantly

enriched in tubulin binding, microtubule binding, carbohydrate binding, cargo receptor activity and immunoglobulin binding (**Figure 4C**).

KEGG pathway analysis provided a potential functional cluster of common DEGs, indicating that the common DEGs were clustered in Herpes simplex virus one infection, Human T−cell leukemia virus one infection, Cell cycle, Transcriptional misregulation in cancer and Epstein−Barr virus infection (**Figure 4D**).

DISCUSSION

SLE is a multi-system autoimmune inflammation that can affect multiple organs and cause extensive and severe clinical manifestations (Wu et al., 2021). The current understanding of the pathogenesis of SLE is not comprehensive. The key driving factors involved in the occurrence and development of SLE remain to be determined. In this study, we provided new insights into the transcriptome of SLE based on RNA-seq data.

The results showed that compared with the normal healthy control groups, a large number of genes in SLE patients were abnormally expressed. Through integrated analysis, we found that there were 790 shared DEGs in the two databases. The results indicated that these common DEGs may lead to the occurrence and development of SLE. Previous studies had shown that lncRNA and circRNA are important factors leading to the occurrence of SLE (Cheng et al., 2021; Jiang et al., 2021). We found that the differential expression of these common DEGs might play an important role in this process.

Through further analysis, we found that the DEGs tended to up-regulated in the two datasets. Through protein-protein interaction network analysis of commonly dysregulated genes, we found that there was a strong correlation between these genes. These PPI networks may have affected the occurrence and development of SLE. Pathway enrichment results showed that common DEGs were significantly enriched in immune-related pathways such as neutrophil mediated immunity, neutrophil degranulation, neutrophil activation involved in the immune response.

In summary, we integrated and analyzed high-throughput sequencing RNA-seq datasets to uncover potential molecular mechanisms of SLE. Our findings provide new clues for possible targeted therapy of SLE. Further studies on the functions of those common DEGs hoped to better understand SLE by integrating more data.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/GSE162828 and GSE169080

AUTHOR CONTRIBUTIONS

ZZ designed the experiments. HZ obtained the data from GEO. YW, JF, SW, YW, and WK analyzed the data. HZ and ZZ wrote the manuscript. All authors read and approved the manuscript.

REFERENCES

Adamopoulos, P. G., Tsiakanikas, P., Adam, E. E., and Scorilas, A. (2021). Unraveling Novel Survivin mRNA Transcripts in Cancer Cells Using an In-House Developed Targeted High-Throughput Sequencing Approach. *Genomics* 113 (1 Pt 2), 573–581. doi:10.1016/j.ygeno.2020.09.053

Adinolfi, A., Valentini, E., Calabresi, E., Tesei, G., Signorini, V., Barsotti, S., et al. (2016). One Year in Review 2016: Systemic Lupus Erythematosus. *Clin. Exp. Rheumatol.* 34 (4), 569–574.

Ahmed, A., Shamsi, A., Mohammad, T., Hasan, G. M., Islam, A., and Hassan, M. I. (2021). Aurora B Kinase: a Potential Drug Target for Cancer Therapy. *J. Cancer Res. Clin. Oncol.* 147 (8), 2187–2198. doi:10.1007/s00432-021-03669-5

Bao, X., Shi, R., Zhao, T., Wang, Y., Anastasov, N., Rosemann, M., et al. (2021). Integrated Analysis of Single-Cell RNA-Seq and Bulk RNA-Seq Unravels Tumour Heterogeneity Plus M2-like Tumour-Associated Macrophage Infiltration and Aggressiveness in TNBC. *Cancer Immunol. Immunother.* 70 (1), 189–202. doi:10.1007/s00262-020-02669-7

Beccastrini, E., D'Elios, M. M., Emmi, G., Silvestri, E., Squatrito, D., Prisco, D., et al. (2013). Systemic Lupus Erythematosus: Immunopathogenesis and Novel Therapeutic Targets. *Int. J. Immunopathol Pharmacol.* 26 (3), 585–596. doi:10.1177/039463201302600302

Cheng, Q., Chen, M., Chen, X., Chen, X., Jiang, H., Wu, H., et al. (2021). Novel Long Non-coding RNA Expression Profile of Peripheral Blood Mononuclear Cells Reveals Potential Biomarkers and Regulatory Mechanisms in Systemic Lupus Erythematosus. *Front. Cel Dev. Biol.* 9, 639321. doi:10.3389/fcell.2021.639321

Clough, E., and Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods Mol. Biol.* 1418, 93–110. doi:10.1007/978-1-4939-3578-9_5

Davies, K., Dures, E., and Ng, W.-F. (2021). Fatigue in Inflammatory Rheumatic Diseases: Current Knowledge and Areas for Future Research. *Nat. Rev. Rheumatol.* 17, 651–664. doi:10.1038/s41584-021-00692-1

Di Battista, M., Marcucci, E., Elefante, E., Tripoli, A., Governato, G., Zucchi, D., et al. (2018). One Year in Review 2018: Systemic Lupus Erythematosus. *Clin. Exp. Rheumatol.* 36 (5), 763–777.

Hess, J. F., Kohl, T. A., Kotrová, M., Rönsch, K., Paprotka, T., Mohr, V., et al. (2020). Library Preparation for Next Generation Sequencing: A Review of Automation Strategies. *Biotechnol. Adv.* 41, 107537. doi:10.1016/j.biotechadv.2020.107537

Jiang, Z., Zhong, Z., Miao, Q., Zhang, Y., Ni, B., Zhang, M., et al. (2021). circPTPN22 as a Novel Biomarker and ceRNA in Peripheral Blood Mononuclear Cells of Rheumatoid Arthritis. *Mol. Med. Rep.* 24 (2), 617. doi:10.3892/mmr.2021.12256

Jin, Z., Zhao, X., Cui, L., Xu, X., Zhao, Y., Younai, F., et al. (2020). UBE2C Promotes the Progression of Head and Neck Squamous Cell Carcinoma. *Biochem. Biophysical Res. Commun.* 523 (2), 389–397. doi:10.1016/j.bbrc.2019.12.064

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45 (D1), D353–D361. doi:10.1093/nar/gkw1092

Kou, N., Zhou, W., He, Y., Ying, X., Chai, S., Fei, T., et al. (2020). A Mendelian Randomization Analysis to Expose the Causal Effect of IL-18 on Osteoporosis Based on Genome-wide Association Study Data. *Front. Bioeng. Biotechnol.* 8, 201. doi:10.3389/fbioe.2020.00201

Li, H., Long, J., Xie, F., Kang, K., Shi, Y., Xu, W., et al. (2019a). Transcriptomic Analysis and Identification of Prognostic Biomarkers in Cholangiocarcinoma. *Oncol. Rep.* 42 (5), 1833–1842. doi:10.3892/or.2019.7318

Li, R., Jiang, X., Zhang, Y., Wang, S., Chen, X., Yu, X., et al. (2019b). Cyclin B2 Overexpression in Human Hepatocellular Carcinoma Is Associated with Poor Prognosis. *Arch. Med. Res.* 50 (1), 10–17. doi:10.1016/j.arcmed.2019.03.003

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Mazzu, Y. Z., Armenia, J., Nandakumar, S., Chakraborty, G., Yoshikawa, Y., Jehane, L. E., et al. (2020). Ribonucleotide Reductase Small Subunit M2 Is a Master Driver of Aggressive Prostate Cancer. *Mol. Oncol.* 14 (8), 1881–1897. doi:10.1002/1878-0261.12706

McDermaid, A., Monier, B., Zhao, J., Liu, B., and Ma, Q. (2019). Interpretation of Differential Gene Expression Results of RNA-Seq Data: Review and Integration. *Brief Bioinform* 20 (6), 2044–2054. doi:10.1093/bib/bby067

Oku, K., and Atsumi, T. (2018). Systemic Lupus Erythematosus: Nothing Stale Her Infinite Variety. *Mod. Rheumatol.* 28 (5), 758–765. doi:10.1080/14397595.2018.1494239

Rees, F., Doherty, M., Grainge, M. J., Lanyon, P., and Zhang, W. (2017). The Worldwide Incidence and Prevalence of Systemic Lupus Erythematosus: a Systematic Review of Epidemiological Studies. *Rheumatology (Oxford)* 56 (11), 1945–1961. doi:10.1093/rheumatology/kex260

Ronco, P., Beck, L., Debiec, H., Fervenza, F. C., Hou, F. F., Jha, V., et al. (2021). Membranous Nephropathy. *Nat. Rev. Dis. Primers* 7 (1), 69. doi:10.1038/s41572-021-00303-z

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). 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

Takashima, S., Saito, H., Takahashi, N., Imai, K., Kudo, S., Atari, M., et al. (2014). Strong Expression of Cyclin B2 mRNA Correlates with a Poor Prognosis in Patients with Non-small Cell Lung Cancer. *Tumor Biol.* 35 (5), 4257–4265. doi:10.1007/s13277-013-1556-7

Thomas, P. D. (2017). The Gene Ontology and the Meaning of Biological Function. *Methods Mol. Biol.* 1446, 15–24. doi:10.1007/978-1-4939-3743-1_2

Tsokos, G. C. (2011). Systemic Lupus Erythematosus. *N. Engl. J. Med.* 365 (22), 2110–2121. doi:10.1056/NEJMra1100359

Von Feldt, J. M. (1995). Systemic Lupus Erythematosus. *Postgrad. Med.* 97 (483), 7986–7994. doi:10.1080/00325481.1995.11945982

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* 10 (1), 57–63. doi:10.1038/nrg2484

Wu, D., Ai, L., Sun, Y., Yang, B., Chen, S., Wang, Q., et al. (2021). Role of NLRP3 Inflammasome in Lupus Nephritis and Therapeutic Targeting by Phytochemicals. *Front. Pharmacol.* 12, 621300. doi:10.3389/fphar.2021.621300

Wu, H.-T., Chen, W.-T., Li, G.-W., Shen, J.-X., Ye, Q.-Q., Zhang, M.-L., et al. (2019). Analysis of the Differentially Expressed Genes Induced by Cisplatin Resistance in Oral Squamous Cell Carcinomas and Their Interaction. *Front. Genet.* 10, 1328. doi:10.3389/fgene.2019.01328

Ying, X., Jin, X., Wang, P., He, Y., Zhang, H., Ren, X., et al. (2020). Integrative Analysis for Elucidating Transcriptomics Landscapes of Glucocorticoid-Induced Osteoporosis. *Front. Cel Dev. Biol.* 8, 252. doi:10.3389/fcell.2020.00252

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, C., Zhao, L., Leng, L., Zhou, Q., Zhang, S., Gong, F., et al. (2020). CDCA8 Regulates Meiotic Spindle Assembly and Chromosome Segregation during Human Oocyte Meiosis. *Gene* 741, 144495. doi:10.1016/j.gene.2020.144495

Zhang, P. C., Zhong, X. R., Zheng, H., Li, L., Chen, F., Shen, M. J., et al. (2021). Prognostic Values of Spindle Checkpoint Protein BUB1B in Triple Negative Breast Cancer. *Zhonghua Bing Li Xue Za Zhi* 50 (6), 645–649. doi:10.3760/cma.j.cn112151-20210131-00108

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 November 2021
doi: 10.3389/fgene.2021.763915

Construction of sRNA Regulatory Network for *Magnaporthe oryzae* Infecting Rice Based on Multi-Omics Data

Enshuang Zhao[1], *Hao Zhang*[1,2]*, *Xueqing Li*[2], *Tianheng Zhao*[2] *and Hengyi Zhao*[2]

[1]College of Software, Jilin University, Changchun, China, [2]College of Computer Science and Technology, Jilin University, Changchun, China

Studies have shown that fungi cause plant diseases through cross-species RNA interference mechanism (RNAi) and secreted protein infection mechanism. The small RNAs (sRNAs) of *Magnaporthe oryzae* use the RNAi mechanism of rice to realize the infection process, and different effector proteins can increase the autotoxicity by inhibiting pathogen-associated molecular patterns triggered immunity (PTI) to achieve the purpose of infection. However, the coordination of sRNAs and proteins in the process of *M. oryzae* infecting rice is still poorly understood. Therefore, the combination of transcriptomics and proteomics to study the mechanism of *M. oryzae* infecting rice has important theoretical significance and practical value for controlling rice diseases and improving rice yields. In this paper, we used the high-throughput data of various omics before and after the *M. oryzae* infecting rice to screen differentially expressed genes and sRNAs and predict protein interaction pairs based on the interolog and the domain-domain methods. We were then used to construct a prediction model of the *M. oryzae*-rice interaction proteins according to the obtained proteins in the proteomic network. Finally, for the differentially expressed genes, differentially expressed sRNAs, the corresponding mRNAs of rice and *M. oryzae*, and the interacting protein molecules, the *M. oryzae*-rice sRNA regulatory network was built and analyzed, the core nodes were selected. The functional enrichment analysis was conducted to explore the potential effect pathways and the critical infection factors of *M. oryzae* sRNAs and proteins were mined and analyzed. The results showed that 22 sRNAs of *M. oryzae*, 77 secretory proteins of *M. oryzae* were used as effect factors to participate in the infection process of *M. oryzae*. And many significantly enriched GO modules were discovered, which were related to the infection mechanism of *M. oryzae*.

Keywords: ***Magnaporthe oryzae*, rice, multi-omics, sRNA, protein, machine learning**

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Shiwei Sun,
Institute of Computing Technology, (CAS), China
Hongmin Cai,
South China University of Technology, China

***Correspondence:**
Hao Zhang
zhangh@jlu.edu.cn

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

Received: *24 August 2021*
Accepted: *28 September 2021*
Published: *12 November 2021*

Citation:
Zhao E, Zhang H, Li X, Zhao T and Zhao H (2021) Construction of sRNA Regulatory Network for Magnaporthe oryzae *Infecting Rice Based on Multi-Omics Data. Front. Genet. 12:763915. doi: 10.3389/fgene.2021.763915*

INTRODUCTION

Rice is an important crop, providing a portion of staple food for more than half of the world's population (Ruiz-Sánchez et al., 2010). However, rice blast is the most severe disease of rice, caused by *Magnaporthe oryzae*, which seriously affects crop stability and sustainability around the world (Imam et al., 2015). Therefore, research on how to control rice blast is widespread.

Although *M. oryzae* is a model fungus for the study of plant-fungal diseases, current studies have shown that the long-term control performance of rice blast by using rice fungicides in the field or selecting rice varieties resistant to *M. oryzae* is still unstable (Deng and Naqvi, 2019). Therefore, people have done a lot of research on *M. oryzae* infecting rice and achieved some research results. However, the interaction mechanism between fungi and plants is very complicated, and it is currently challenging to analyze the molecular interaction mechanism only by biological experiments (Li et al., 2017; Nelson et al., 2018). Therefore, researchers began to use biocomputing methods to assist and guide biological experiments based on the emergence of many omics data related to fungus-plant interactions, such as genomics, transcriptomics, proteomics and metabolomics multi-omics data to reveal interactions between biomolecules and explore key factors in biological processes.

For exploring the key biomolecules in the process of fungus-plant interactions small RNAs (sRNAs) were first studied in depth. sRNAs refer to those that do not encode proteins in the organism and are mostly 18nt-40nt in length (Mueth et al., 2015). The common mechanism of action of sRNAs is RNA interference (RNAi). The effector complex RISC is added to one of the sRNA strands to achieve the purpose of inhibiting protein biosynthesis (Majumdar et al., 2017). Researchers have found that using the host plant's RNAi mechanism by pathogenic sRNAs to achieve the infection process may be ubiquitous in the fungus-plant interaction mechanism (Weiberg et al., 2013; Cai et al., 2018).

In addition, fungi as eukaryotes, their secreted proteins are transported across the membrane by endocytosis and exocytosis (Pompa et al., 2017; Riquelme et al., 2018). Secreted proteins are proteins produced by the nucleus, processed and transported through the endoplasmic reticulum and Golgi apparatus, and secreted outside of cells or other cells. They play key biological regulatory roles, such as hormones, antibodies, and enzymes (Faso et al., 2009). In addition, studies have found that pathogens invade hosts through secreted proteins to achieve an attack on the hosts' immune effect. For example, when soil-borne pathogenic fungi invade plants, they secrete an effector protein (Verticillium dahliae polysaccharide deacetylase, VdPDA1), which deacetylates chitin oligosaccharides produced by plants to resist infection by pathogens, thus reducing or inactivating the immune system of plants, to achieve the purpose of infection (Cui et al., 2020).

However, at present, the research on the mechanism of fungus-plant interaction is still in its infancy (Kim et al., 2016; Larsen et al., 2016; Großkinsky et al., 2018; Wang et al., 2019). In addition to genomics research combining plant disease resistance genes and sRNA for analysis (Zhang et al., 2016; Raman et al., 2017), other omics analysis is still based on single omics analysis, and some sRNAs (Zhang et al., 2019; Chang et al., 2020), proteins (Solomon and Oliver, 2001; Grenville-Briggs et al., 2005; Grohmann and Bronte, 2010; McGaha et al., 2012; Yang et al., 2012), metabolites (Parker et al., 2009) have been identified. In fungi infecting plants, how sRNA and protein molecules are involved in the regulation is still unknown. Therefore, based on differentially expressed genes, differentially expressed sRNAs and protein interaction pairs in the process of *M. oryzae* infecting rice, this study proposed a new method to analyze the multi-omics data of *M. oryzae* infecting rice and constructed a multi-omics data integration-based *M. oryzae*-rice interaction network. It also wholly presented the interaction relationship between the markers of various omics in the process of *M. oryzae* infecting rice and revealed the key nodes that play a regulatory role in *M. oryzae* infection in rice. This paper found a possible solution for studying the mechanism of *M. oryzae* infecting rice and provided research ideas for preventing and controlling rice and other food crops.

DATA AND METHODS

Firstly, the genomic, transcriptome, and proteome data were analyzed to establish the *M. oryzae*-rice sRNA interaction network and *M. oryzae*-rice protein interaction network. Then, the sRNA and protein interaction networks of *M. oryzae* and rice were analyzed. Finally, the PPI interaction networks and GO functional enrichment modules of *M. oryzae* and rice were excavated, respectively, and the key factors of multiple omics joint regulations and the biological processes involved were explored. The design roadmap for this work is shown in **Figure 1**.

Data Source

Regarding the genome and transcriptome, this paper used the gene chip expression data of rice before and after *M. oryzae* infection with rice at 72 h, sRNA data of *M. oryzae* cultured on a complete medium for 16 h, the mixed sRNA data of the rice infected by *M. oryzae* for 72 h (Raman et al., 2013), the gene expression data of rice after 48 h of culture, the gene expression data of rice after 48 h of infection by *M. oryzae* (Chujo et al., 2013), and the mRNA data of rice. These are all from the NCBI database. Regarding the proteome, high-throughput protein data of mode hosts, mode pathogens, rice and *M. oryzae* were obtained from HPIDB, NCBI and Uniport databases. We first obtained the protein IDs of *M. oryzae* and rice from the NCBI and Uniport databases. Because different databases have different identifiers for the same protein, the obtained protein IDs must be converted uniformly. Here, the protein IDs of the Uniport database were selected as the unified protein ID identifiers, and the high-throughput data of these proteins were obtained after the protein IDs were converted.

Data Preprocessing

The Acquisition of Differentially Expressed Genes in Rice

The commonly used R software packages for the gene chip probe level data processing include affy, affyPLM, affycomp, gcrma, etc. In this step, the affy package was used to analyze the rice gene differential expression. Firstly, the background noise of the gene chips was denoised by the MAS method. Then, in order to eliminate the influence of signal strength and other factors between different chips, the linear normalization method was used for chip data. Next, the expression amount of the gene

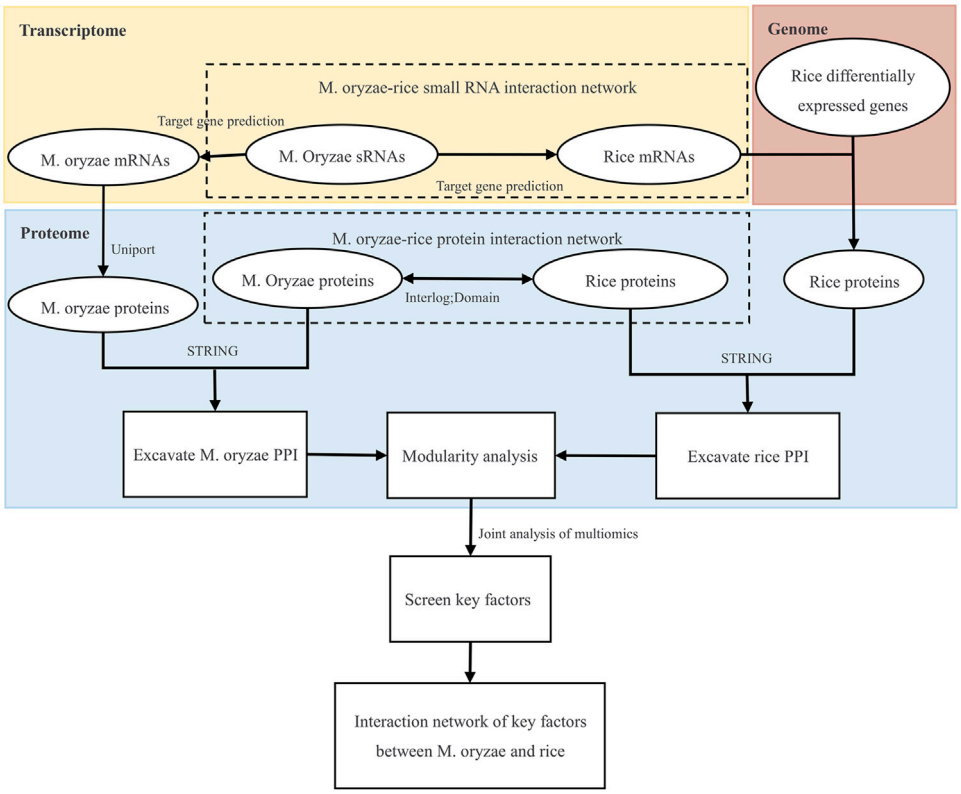

FIGURE 1 | Overall design route.

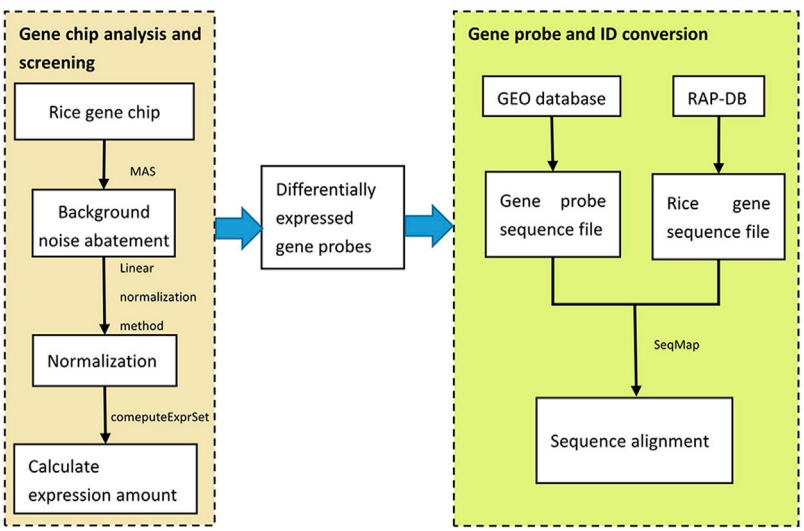

FIGURE 2 | Algorithms for the analysis of differentially expressed genes.

probes was calculated by the hybridization signal of the probeset using the function computeExprSet in the affy package.

Then, the sequence number of the probes used by the gene chip was retrieved from the GEO database and the probe sequences were downloaded. Then, the whole rice genome sequences were downloaded from RAP-DB, and the sequence alignment between the gene probe sequences and the whole rice genome sequences was performed by using the SeqMap sequence

alignment tool to find the rice gene IDs corresponding to the gene probes. Finally, by extracting the matched rice gene IDs, the conversion from gene probes to gene IDs was completed, and 1.5-fold differentially expressed rice genes were screened out, totaling 1,368. This process is shown in **Figure 2**.

Differentially Expressed sRNAs Screening of *M. oryzae*

First, to remove the adapters and get the correct sRNA sequences, the cutadapt tool was used to remove the sRNA data adapters. Next, genome matching was performed on the sRNA data of *M. oryzae* after removing the adapters to remove the sRNA data that were not of *M. oryzae* from the data. The specific operation was to perform the mapping operation on the mixed sRNA data of *M. oryzae* after removing the adapters and match it to the genome of *M. oryzae* to obtain the pure sRNA data of *M. oryzae*. The genome matching tools used in this section were bowtie and samtools.

Since there are several sRNA sequences of different lengths in FASTQ files, it is necessary to control the length of these sRNA sequences. According to the available length of plant sRNAs, we selected the sRNA sequences of *M. oryzae* from 18nt to 25nt, and suggested that these sRNAs could be used to predict the target genes of rice. File A containing sRNA sequence, sequence length and sequence expression amount of *M. oryzae* was obtained from the *M. oryzae* sRNA data after length control and without genome mapping. Then, the file after genome mapping was extracted, and each sRNA sequence of *M. oryzae* was extracted into file B. Finally, the two files were matched. After matching each sRNA sequence in file A, *M. oryzae* sRNA data appearing in file B was output.

In this paper, the 3/4 quantile normalization method was used to normalize the sRNA expression amount data before and after the infection of *M. oryzae*. The specific method was to rank the sRNA expression amount of *M. oryzae* from high to low and find the *M. oryzae* sRNA ranked in 3/4. Then, this expression amount was taken as the baseline of the lower expression level, and the expression amounts of other samples were converted to multiples of this expression amount. Finally, the data of *M. oryzae* differentially expressed sRNA after normalized treatment were statistically analyzed, and the expression amount and expression rate was used for screening. The following formula calculated the expression rate:

$$Growth\_Rate = \frac{count_{after} - count_{before}}{count_{before}}$$

It was found that there were 4933 new sRNA data after infection, and the expression amount was sorted, and the top 146 sRNA data with the highest expression amount were selected. The data of 6,100 sRNA species before and after the infection of *M. oryzae* were screened by two conditions: expression amount and expression rate. A total of 220 species *M. oryzae* sRNAs were screened out by selecting sRNAs whose differential expression amount increment was more significant than or equal to 9 and differential expression increase rate was more significant than or equal to 2. Similarly, the sRNA data of *M. oryzae* with differential expression amount increment less than −116.5 were selected, and there were 257 kinds of sRNA data. The differential expression amount increment and expression amount increase rate of sRNAs above were all greater than the corresponding mean values of increase or decrease. Because the sRNA differential expression increase rate ranged from 0 to 1, and the change rate was meager, only the increment of differential expression amount was used to screen the decreased expression sRNA of *M. oryzae*. The distribution map of *M. oryzae* differentially expressed sRNAs is shown in **Figure 3**.

Preprocessing of Protein Data

Blast sequence alignment was performed on the protein amino acid sequences of downloaded rice and *M. oryzae*. Proteins with sequence similarity more significant than 95 were removed as repeated proteins to eliminate the error in the same protein sequencing by different sequencing platforms and avoid duplicating the same protein that was considered to be caused by two different proteins.

Prediction of *M. oryzae*-Rice sRNA Interaction Pairs

Using the bioinformatics method to accurately and rapidly predict the target genes of miRNA can provide clues for studying the function of miRNA. Using target gene prediction software to predict miRNA target genes is more efficient and faster than experimental biological methods. There are many standard target gene tools, including TargetScan, miRcode, miRDB, RNA22, and tapir, the target gene prediction tool used in this paper. Before the prediction, T was converted to U in the sRNA data and such sequence files were converted to FASTA files. After the sequence base conversion of FASTA files, the tapir tool can be used to predict the target genes of the sRNA sequence files of *M. oryzae*.

First, the FASTA CDS files of *M. oryzae* and rice were downloaded, and the FASTA files of the differentially expressed *M. oryzae* sRNAs were obtained. Then, when the tapir tool was used for target gene prediction, the matching score was set as 0.5 and the free energy ratio was set as 0.7. After target gene matching, Python script was applied to process the prediction results, and the final target gene prediction result file was obtained.

In this section, 366 kinds of differential expression amount up-regulated and newly added of *M. oryzae* sRNAs were targeted to rice mRNAs. A total of 1,857 rice mRNAs were obtained. After gene IDs matching and deduplication of these mRNAs, 1,121 rice gene IDs were obtained. In the same way, 257 kinds of *M. oryzae* sRNAs with down-regulated differential expression amounts were targeted to *M. oryzae*, and 664 *M. oryzae* mRNAs and 264 *M. oryzae* genes were obtained.

Prediction of *M. oryzae*-Rice Protein Interaction Pairs

Sequence-Based

The protein interaction prediction method based on sequence features (interolog method) is based on the principle that homologous proteins have similar functional and structural characteristics (Thanasomboon et al., 2017). Here, the

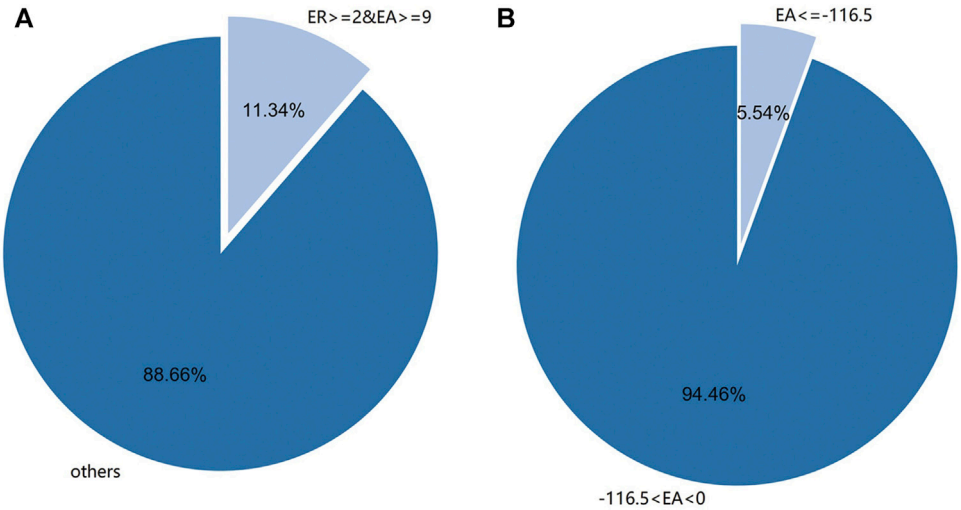

FIGURE 3 | (A) The distribution of differential expression rate (ER) of sRNAs of *M. oryzae* was greater than or equal to 2 and the expression amount (EA) was more significant than or equal to 9. **(B)** The distribution of differential EA of sRNAs of *M. oryzae* was less than or equal to −116.5.

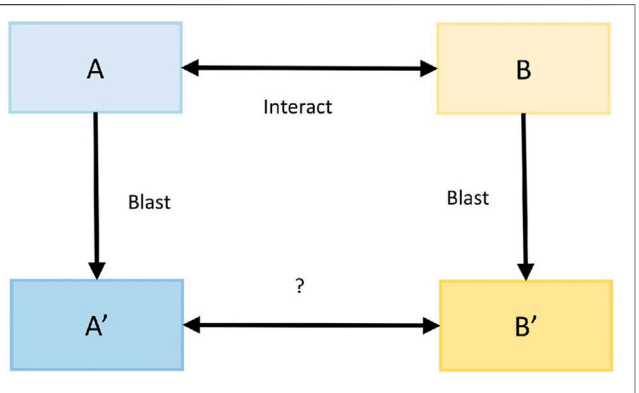

FIGURE 4 | Interolog method.

interspecific interolog method predicted the protein interaction relationship between *M. oryzae* and rice. First, the confirmed interaction mode host and mode pathogen protein sequences were recorded as A and B, while the protein sequences of rice and *M. oryzae* were recorded as A′and B′. Then, for each protein amino acid sequence in A′, sequence alignment was carried out with the protein amino acid sequences in A, and the accuracy was obtained. Similarly, file B also followed this step. Finally, the accuracy of the interaction relationship pairs between A′and B′ was calculated by interacting with the proteins in A and B. The process is shown in **Figure 4**.

In this process, the interolog method was used to screen the protein interaction pairs between rice and *M. oryzae*, and the threshold was set as E-value less than or equal to 1E-5 and similarity greater than or equal to 30. Then, the model pathogen and mode host protein pairs corresponding to *M. oryzae* and rice proteins were matched, and the protein pair files of *M. oryzae* and rice were obtained based on the interolog method.

Domain-Based

The available domain-based protein interaction prediction method (domain-domain interaction method) is based on the principle that interacting protein pairs may have the exact functional domains (Lee et al., 2006). For example, for the confirmed interactions between mode host protein A and mode pathogen protein B, if rice protein A′and *M. oryzae* protein B′ have the same interaction functional domains as protein A and protein B, then rice protein A′and *M. oryzae* protein B′ interact. The process is shown in **Figure 5**.

In this process, functional domains were obtained from the protein amino acid sequences of mode hosts, mode pathogens, rice and *M. oryzae* through the Pfam database. E-value was selected as 1E-5 and the coincidence rate was selected as 90%. TSV files containing protein IDs, protein functional domains and E-values were obtained. Then, protein domain files were extracted and sorted to obtain protein interaction relationship pairs based on functional domains.

Prediction of Secreted Protein of *M. oryzae*

By combining the interolog method and domain-domain method, 83664 pairs of protein interactions were obtained following the two methods. However, not all *M. oryzae* proteins can be transported across the membrane, it is necessary to do the secreted protein identification of the above *M. oryzae* proteins and screen out the *M. oryzae*-rice protein interaction network that *M. oryzae* proteins were secreted proteins.

In this paper, the secreted proteins of *M. oryzae* were predicted on TMHMM. The FASTA files of 323 *M. oryzae* protein amino acid sequences were obtained through the Uniport database and imported into the TMHMM website to obtain their secreted proteins' predicted results. When the expected number of amino acids in the transmembrane helix of a protein is greater than or equal to 18, or when the transmembrane helix number of N-the best predicted is greater than or equal to 1, the protein can be considered a secreted protein. Therefore, protein IDs with

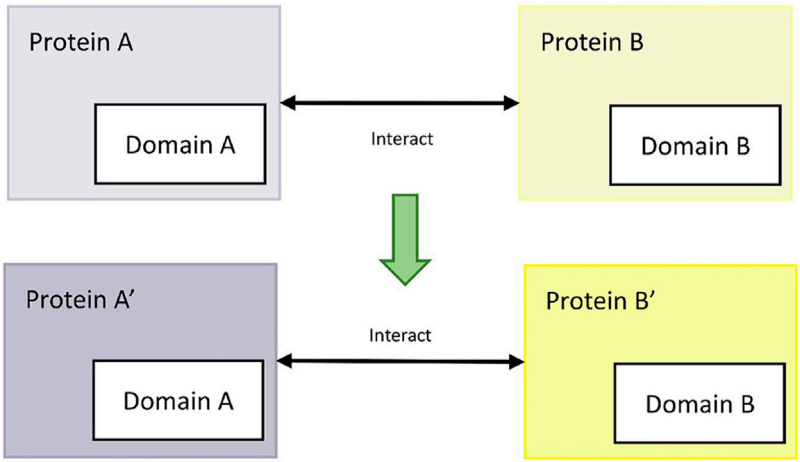

FIGURE 5 | Domain-domain method.

parameters ExpAA greater than or equal to 18 or PredHel greater than or equal to 1 were extracted. The obtained *M. oryzae* secreted proteins were matched and screened with the previous 83664 *M. oryzae*-rice protein interaction pairs, and finally, 7352 *M. oryzae*-rice protein interaction pairs were obtained.

Construction of a Prediction Model for Cross-Species Regulatory Protein Pairs Between *M. oryzae* and Rice

Acquisition and Processing of Positive and Negative Samples

This paper established a prediction model for *M. oryzae* and rice interaction protein pairs and obtained protein interaction pairs through the sequence and functional structure prediction in experiments. The 7352 data of the effective interaction pairs of *M. oryzae* and rice obtained above were used as positive samples. The negative samples were randomly selected from other *M. oryzae* rice protein interaction pairs except the positive samples that the ratio of positive and negative samples was 1:1.

For the protein features of *M. oryzae* and rice, the proteins' amino acid sequences and functional domains were used as the feature data. In addition, functional domain texts were preprocessed before training, including unifying special symbols, spaces, upper and lower case letters of each functional domain and removing stop words to achieve standardized processing of data samples.

Construction of Protein Interaction Pair Prediction Model Based on textRNN

Recurrent Neural Network (RNN) is mainly used in sequence prediction, character generation, emotion recognition, man-machine dialogue, etc. RNN is a kind of recursive neural network that takes sequence data as input, recurses in the sequence's evolution direction, and connects all nodes in a chain. The sequence information determines the task of the event itself, which requires previous knowledge and current information to determine the output result jointly. As a result, textRNN can more effectively address the problem of contextual semantic relevance. Considering that the protein's amino acid sequence and functional domains belonged to short texts, which have contextual semantic relevance characteristics, this paper used textRNN to construct the protein interaction pair binary classification model.

A multi-layer RNN network needs to be established in the construction of RNN model. The dropout layer was added after each RNN kernel function, and the amino acid sequences after the *M. oryzae* and rice protein interaction pair segmentation were used as the input variable of the RNN model. The first hidden layer activated this input. Then the successive activations were performed layer by layer to get the output. Each hidden layer had its own weight and bias. Parameters such as the classification results, accuracy and loss function of the output protein interaction pairs were output by the output layer. The optimal RNN protein interaction model was obtained by adjusting learning_rate, dropout_keep_prob and total iteration cycles according to the learning curve and confusion matrix. Finally, different evaluation indexes were applied to evaluate and verify the model. The accuracy of protein interaction pairs predicted by the interolog method and domain-domain method in this paper was proved.

Analysis of Regulatory Network Between *M. oryzae* and Rice

In order to analyze the obtained sRNA and protein interaction network of *M. oryzae*-rice, and the network diagram of *M. oryzae*-rice protein interaction was too significant. Therefore, the PPI networks of *M. oryzae* and rice jointly regulated by various omics were explored, respectively. First, the PPI network of *M. oryzae* was mined based on the proteins regulated by *M. oryzae* differentially expressed sRNAs and *M. oryzae* proteins in the *M. oryzae*-rice protein interaction network. And the PPI network of rice was mined based on the proteins regulated by rice differentially expressed genes and rice proteins in the *M. oryzae*-rice protein interaction network. Then the PPI networks of *M. oryzae* and rice were analyzed for GO pathway enrichment, and

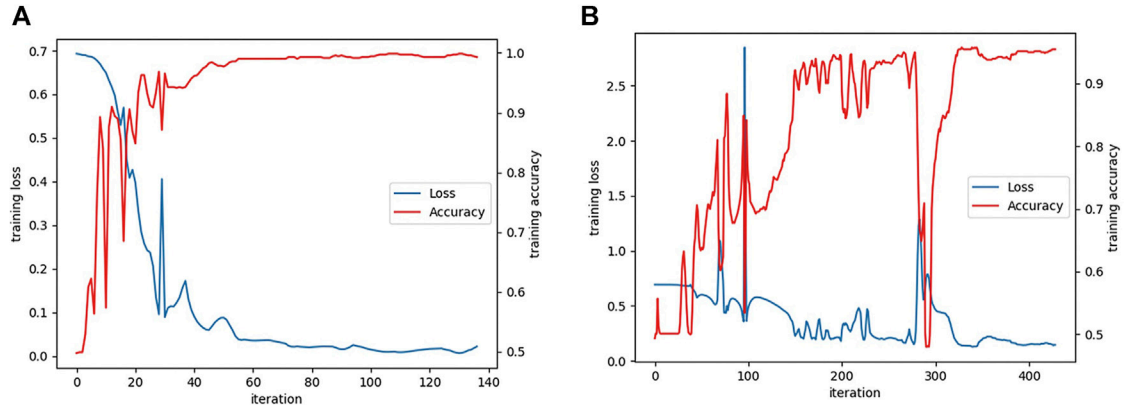

FIGURE 6 | (A) The learning curves of textRNN model on functional domain and **(B)** amino acid sequence characteristics.

TABLE 1 | Evaluation indexes of textRNN model on the functional domain.

	Precision	Recall	F1-score	Support
0	0.99	0.99	0.99	420
1	0.99	0.99	0.99	419
Accuracy			0.99	839
Macro avg	0.99	0.99	0.99	839
Weightrd avg	0.99	0.99	0.99	839

TABLE 2 | Evaluation indexes of textRNN model on the amino acid sequence.

	Precision	Recall	F1-score	Support
0	0.97	0.98	0.98	399
1	0.98	0.97	0.97	398
Accuracy			0.97	797
Macro avg	0.97	0.97	0.97	797
Weightrd avg	0.97	0.97	0.97	797

the modules were separated. Finally, by analyzing the isolated *M. oryzae* and rice protein networks, the main modules' biological functions and KEGG enrichment pathways were described. The key nodes of *M. oryzae*-rice and their interaction networks were mined by using multi-omics network data to explore the molecular mechanism of *M. oryzae* and rice interaction.

RESULTS

Prediction Model Results of the Interspecies Regulatory Protein Pairs Between *M. oryzae* and Rice

The learning curves of the textRNN model on the functional domain and amino acid sequence are shown in **Figure 6**.

The model was evaluated according to the precision, recall and F1 indexes, and the accuracy indexes of the TEXTRNN model in the functional domain and amino acid sequence are shown in **Table 1** and **Table 2**, respectively.

When textRNN model with the functional domain as feature data was tested, the testAcc of textRNN model was 98.81%, testLoss was 0.029, and the confusion matrix was: $\begin{bmatrix}414 & 6\\ 4 & 415\end{bmatrix}$.

When textRNN model with protein amino acid sequence as feature data was tested, the testAcc of textRNN model was 97.49%, testLoss was 0.086, and the confusion matrix was: $\begin{bmatrix}391 & 8\\ 12 & 386\end{bmatrix}$.

Therefore, the textRNN model can be used to predict the *M. oryzae*-rice protein interaction pairs, and the prediction model performed well in this paper. Furthermore, the prediction of protein interaction pairs in plants infected by other fungi can also refer to this model.

Analysis Results of *M. oryzae*-Rice Transcriptome and Proteome Networks

After target prediction of the 623 kinds of *M. oryzae* sRNAs, 1,857 *M. oryzae*-rice sRNA interaction pairs and 664 *M. oryzae* internal sRNA interaction pairs were obtained. By digging positive and negative regulatory factors, 1,166 *M. oryzae* genes, 1,121 rice genes, 1,173 *M. oryzae* proteins and 1,677 rice proteins were found to be involved in the biological process of *M. oryzae* infection to rice. In addition, the transcriptome network of *M. oryzae* and rice was visualized by the Cytoscape tool. There were 20 sRNA-mRNA interaction clusters with two or more sRNAs involved in regulation. The network diagram was shown in **Supplementary Figure S1**.

Based on the 7,352 *M. oryzae* and rice protein interaction pairs obtained previously, the *M. oryzae*-rice protein interaction network diagram was drawn with a total of 11 rice protein interaction clusters. The network diagram was shown in **Supplementary Figure S2**.

A total of 593 kinds of *M. oryzae* sRNAs and 581 kinds of *M. oryzae* secreted proteins directly involved in the two interaction mechanisms were excavated through the *M. oryzae*-rice sRNA interaction network and protein interaction network, and they were put into the STRING database for GO pathway enrichment

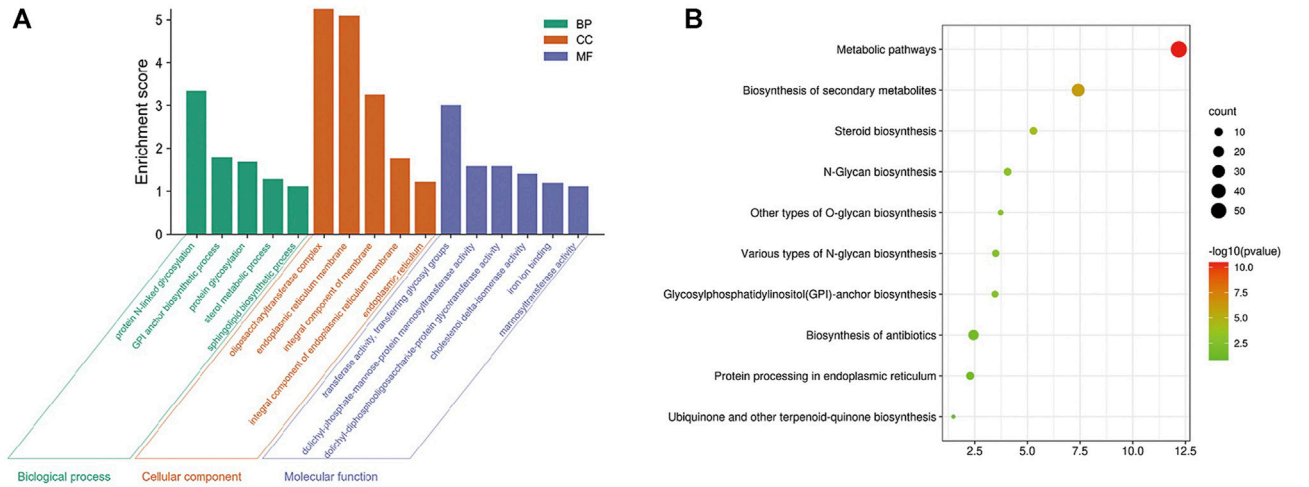

FIGURE 7 | (A) Go enrichment module diagram of *M. oryzae*. The first three significantly enriched Go-terms of the biological process modules are protein N-linked glycosylation, GPI anchor biosynthesis process and protein glycosylation. The first three significantly enriched GO-terms of the cell component modules are oligosaccharyltransferase complex, endoplasmic reticulum membrane, and integral component of the membrane. The first three significantly enriched GO-terms of molecular function modules are transferase activity, transferring glycosyl groups, dolichyl-phosphate-mannose-protein mannosyltransferase activity, and dolichyl-diphosphooligosaccharide-protein glycotransferase activity. **(B)** KEGG enrichment bubble diagram of *M. oryzae*. There are 10 significant enrichment pathways in KEGG enrichment pathways, which are mainly related to various glycan organisms, anchor organisms, steroid organisms, biosynthesis of secondary metabolites, protein processing and metabolic pathways.

analysis and KEGG enrichment analysis. First, the *p*-value was set as 1E-16, and the GO enrichment results and KEGG enrichment results were derived. The PPI network diagram of *M. oryzae* was too large to be shown in this paper and was shown in **Supplementary Figure S3**. Next, GO enrichment analysis (**Figure 7A**) and KEGG pathway enrichment analysis (**Figure 7B**) were carried out on PPI interaction network diagram of *M. oryzae*. It can be seen that most of these enrichment pathways were involved in the biological processes of sRNA synthesis, protein synthesis and transport in *M. oryzae*. To some extent, the above conclusions proved that *M. oryzae* could complete the infection process of rice through sRNAs and secreted proteins.

The obtained rice differentially expressed genes, rice proteins regulated by mRNAs and rice proteins in the protein interaction network of *M. oryzae* and rice were analyzed by GO enrichment and KEGG enrichment. However, there were too many rice-related protein nodes. Firstly, the protein nodes obtained by three ways were mined through the STRING database for their PPI. Then the rice protein interaction pairs obtained were imported into Cytoscape to obtain the rice protein interaction network. Finally, the rice protein interaction network was divided into modules and the largest five rice modules were screened out. The GO enrichment pathways (**Figure 8A**) and KEGG pathways (**Figure 8B**) of each module were excavated, respectively.

Modularity Analysis Results of *M. oryzae* and Rice Regulatory Networks (Cluster 1–10)

In this paper, Clusterviz, a Cytoscape plug-in, was used to segment the protein interaction network between *M. oryzae* and rice into modules, and the FAG-EC algorithm was selected to intercept only the subnet modules with more than six nodes. Next, the segmentation subnet modules were sorted by complexity, and GO function enrichment analysis was carried out for each module. The largest five subnets with significant function enrichment analysis were selected for subsequent analysis and named Cluster 1–10. Then, each subnet's GO functional modules and KEGG enrichment pathways were mined to explore their biological processes.

Cytoscape calculated the network topology attributes, and its plug-in NetworkAnalyzer was used to calculate the degree and betweenness of nodes in each subnet. Betweenness is a measure of the centrality of a node in the network. In some sense, it measures the influence of a node on information spread through the network. The following formula calculates betweenness:

$$C_b(n) = \sum_{s \neq n \neq t} (\delta_{st}(n)/\delta_{st})$$

Where s and t are genes different from n in the network, δ_{st} represents the shortest path from s to t, and δ_{st} (n) represents the shortest path from s to t and through n.

The nodes of each subnet were sorted according to betweenness, and the top 6 nodes with the highest betweenness in each subnet were obtained, which were regarded as the central nodes of the subnet and marked in the subnet interaction diagram.

After the segmentation module of the regulatory network of *M. oryzae*, five largest significant functional enrichment subnets were selected, which were the *M. oryzae* helicase activity and protein synthesis module (Cluster 1), *M. oryzae* DNA repair-related module (Cluster 2), *M. oryzae* RNA transport and molecular transport-related module (Cluster 3), *M. oryzae* gene expression and mRNA processing-related module (Cluster 4) and *M. oryzae*

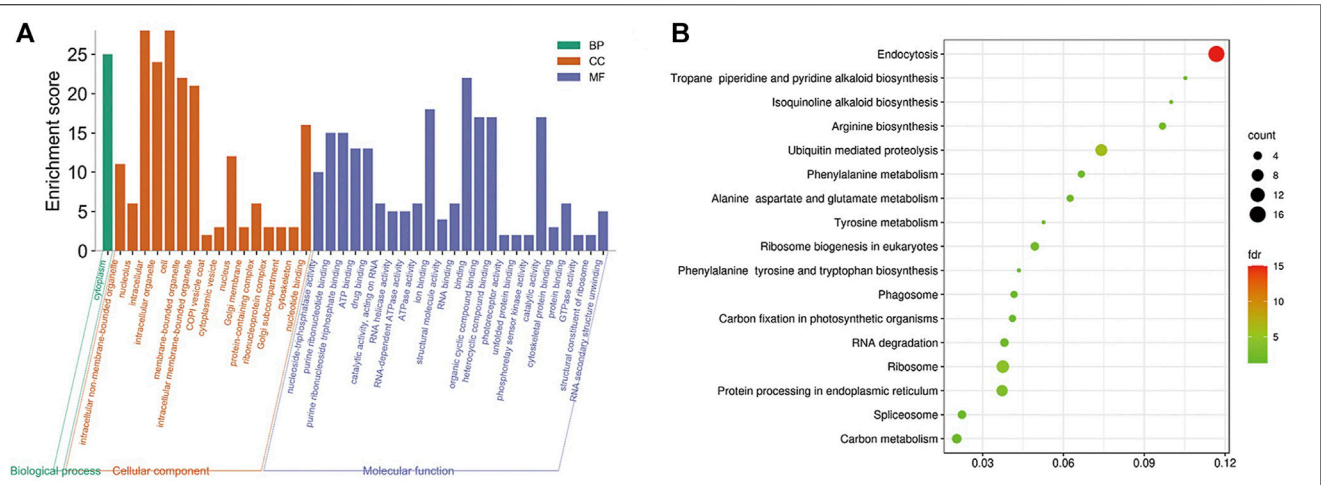

FIGURE 8 | (A) Five modules significantly enriched GO functional enrichment module diagram and **(B)** KEGG enrichment pathway bubble diagram of rice.

biosynthetic pathway-related subnet (Cluster 5). Cluster 1 was closely related to a series of protein synthesis processes and helicase activity (**Supplementary Figure S4**). The KEGG enrichment pathways of Cluster 2 mainly included nucleotide excision repair pathway, homologous recombination and mismatch repair pathway, etc (**Supplementary Figure S5**). The KEGG enrichment pathways of Cluster 3 mainly involved RNA transport, MAPK signaling pathway-yeast and endocytosis pathway (**Supplementary Figure S6**). The GO items of Cluster 4 were mainly involved in RNA transcription, translation and protein synthesis. The KEGG enrichment pathways of Cluster 4 mainly included basic transcription factor enrichment pathway, RNA polymerase enrichment pathway, pyrimidine metabolism enrichment pathway, purine metabolism enrichment pathway, nucleotide excision repair enrichment pathway, ribosome biogenesis in eukaryotes enrichment pathway and metabolic pathway enrichment pathway (**Supplementary Figure S7**). The KEGG enrichment pathways of Cluster 5 mainly included steroid biosynthesis, antibiotic biosynthesis, secondary metabolite biosynthesis, terpenoid skeleton biosynthesis and metabolic pathway (**Supplementary Figure S8**).

After the segmentation module of the regulatory network of rice, five largest significant functional enrichment subnets were selected, which were the rice protein binding functional module (Cluster 6), rice GTP and nucleoside triphosphatase-related module (Cluster 7), rice gene expression, transport and metabolism-related module (Cluster 8), rice protein synthesis module (Cluster 9) and rice gene expression and defense response regulation module in rice (Cluster 10). Cluster 6 was significantly enriched in the unfolded protein binding function module (**Supplementary Figure S9**). Cluster 7 was significantly enriched in GTPase activity, GTP binding and nucleoside-triphosphatase activity (**Supplementary Figure S10**). The go terms of Cluster 8 were related to regulation of gene expression, transport pathway of biomolecules, and rice metabolic pathways. These GO functional modules showed that the infection process of *M. oryzae* affected the gene expression and metabolism of rice (**Supplementary Figure S11**). Cluster 9 was significantly enriched in nucleus, ribosome, ribonucleoprotein complex, cytoplasm, cell, translation and structural constituent of ribosome. Most of these GO modules were related to the protein synthesis process (**Supplementary Figure 12**). The go terms of Cluster 10 were related to regulation of gene expression, protein synthesis, and rice defense module. These GO functional modules showed that the infection process of *M. oryzae* affected the differential gene expression in rice (**Supplementary Figure S13**).

PPI Network Analysis and Screening Results of Main Regulatory Factors of *M. oryzae* and Rice

After the 366 sRNAs up-regulated during the *M. oryzae* infecting rice process to predict the target genes of rice mRNAs, 1,857 rice mRNAs were obtained, pointing to 1,121 rice genes. After the 257 sRNAs were down-regulated during the *M. oryzae* infecting rice process to predict the target genes of *M. oryzae* mRNAs, 664 *M. oryzae* mRNAs were obtained, and 264 *M. oryzae* genes were involved in regulation. The 664 kinds of *M. oryzae* mRNAs were input into the Uniport database to obtain 2,644 protein IDs corresponding to these mRNAs. According to GO, the obtained protein IDs were matched with their interacting protein IDs to expand the proteins involved in regulation by *M. oryzae*. These expanded proteins also used TMHMM to predict secreted proteins, and 337 *M. oryzae* proteins were obtained. Then 601 protein IDs, which were involved in the transboundary regulation of the secreted proteins of *M. oryzae*, were matched with the *M. oryzae*-rice protein interaction pair network to obtain the *M. oryzae* and rice sRNA-protein interaction network (**Figure 9**).

Analysis Results of the Core Nodes of the Interaction Network Between *M. oryzae* and Rice

The *M. oryzae* infecting rice interaction network diagram and the rice response of *M. oryzae* infection network diagram obtained

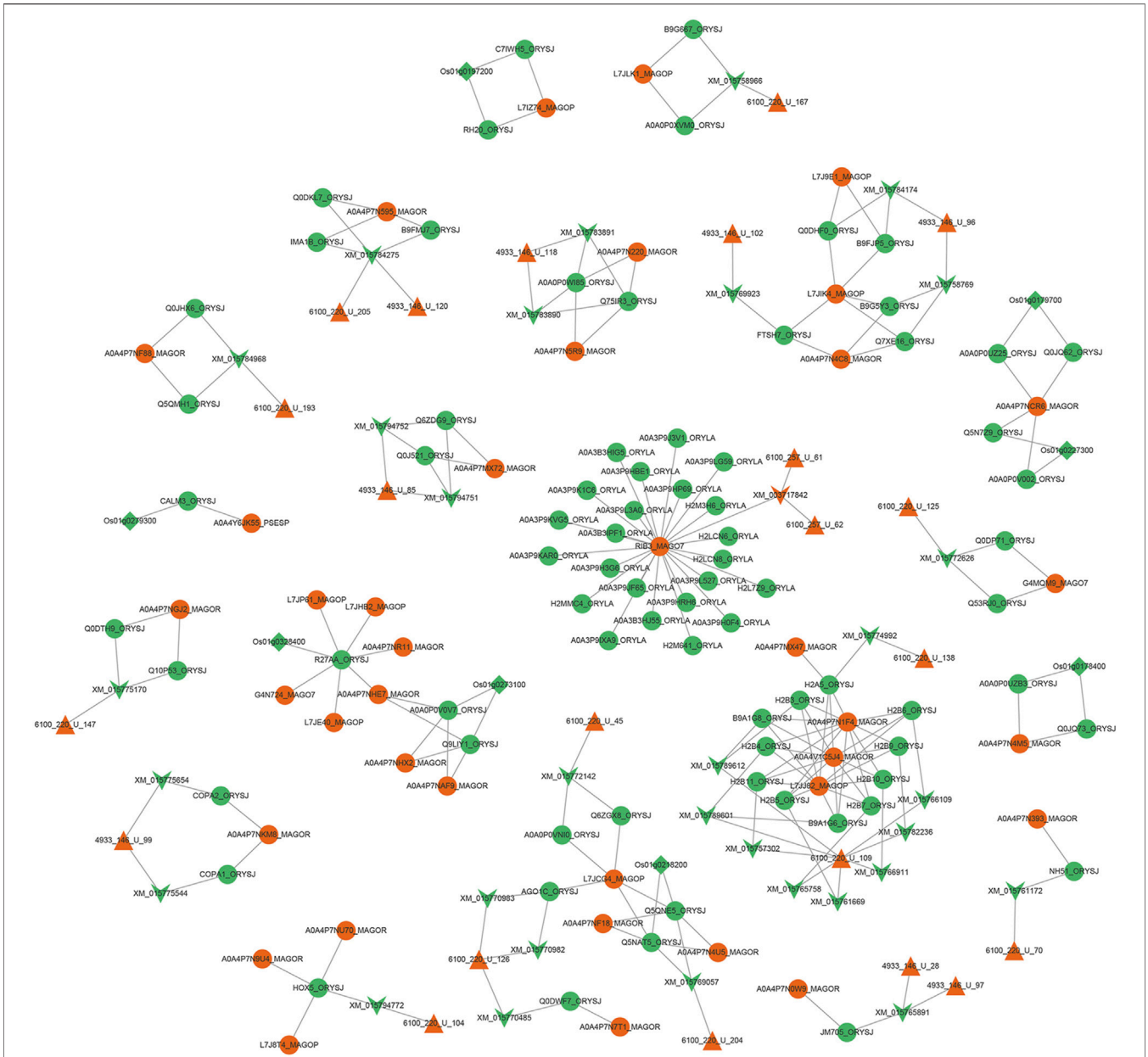

FIGURE 9 | Interaction network diagram between *M. oryzae* and rice main regulatory factors. The red regular triangles are the sRNA nodes of *M. oryzae*, the green inverted triangles are the mRNA nodes of rice, the red inverted triangle is the mRNA node of *M. oryzae*, the green circles are the protein nodes of rice, the red circles are the protein nodes of *M. oryzae*, and the green diamonds are the gene nodes of rice. According to the screening of degree and betweenness, the key protein nodes can be found as RIB3_MAGO7, L7JCG4_MAGOP, A0A4P7NCR6_MAGOR, L7JIK4_MAGOP, HOX5_ORYSJ and R27AA_ORYSJ. The key mRNA nodes can be found as XM_015784275 and XM_015765891. The key sRNA nodes can be found as 4933_146_U_99 and 6100_220_U_126. The key genetic nodes can be found as Os01g0178400 and Os01g0197200.

above were combined to find the biomolecules that play a role in them. However, the large number of these biomolecules was not conducive to our further analysis of *M. oryzae* and rice interaction mechanism, so core node mining was needed. In this study, the biomolecules involved in the infection of rice by *M. oryzae* were extracted by multi-omics joint analysis, including 8 rice differentially expressed genes, 31 rice mRNAs, 77 rice proteins, 22 *M. oryzae* sRNAs, 1 *M. oryzae* mRNA, and 38 *M. oryzae* proteins (**Supplementary Table S1**).

22 differentially expressed sRNAs were found, including 12 up-regulated sRNA data of *M. oryzae*, 8 newly increased sRNA data of *M. oryzae*, and 2 down-regulated sRNA data of *M. oryzae*. 20 up-regulated and newly added sRNA data were used to infect rice by targeting rice mRNAs for rice RNA

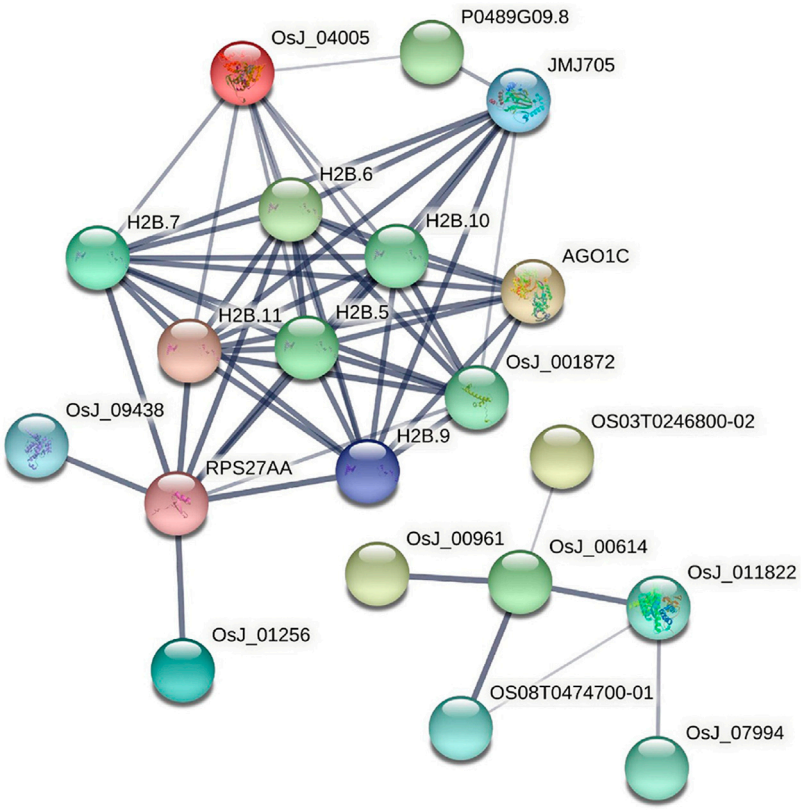

FIGURE 10 | PPI network diagram of rice core nodes.

silencing. And 2 down-regulated sRNAs of *M. oryzae* may increase some proteins in *M. oryzae* to achieve the purpose of invading rice by secreted proteins.

77 rice core proteins were imported into the STRING database, 32 influential rice gene nodes were obtained, and GO function enrichment analysis and KEGG pathway enrichment analysis was conducted. There were 19 interacting gene nodes. The PPI interaction network diagram of rice core nodes is shown in **Figure 10**.

The enrichment analysis of the GO pathway of rice core protein nodes found that the significantly enriched GO functions in rice were distributed in three aspects. One was gene expression-related modules, including negative regulation of gene expression, gene expression regulation, epigenetic regulation, gene silencing, and gene expression. The second was protein molecular synthesis and transport-related modules, including protein complex, protein heterodimerization activity, nucleic acid binding, protein binding, organic circular compound binding, heterocyclic compound binding, DNA binding, intracellular protein transport. The third was metabolism-related modules, including protein metabolism process, macromolecular metabolism process, proteolysis, nitrogen compound metabolism process, cellular macromolecular decomposition process, regulation of nitrogen compound metabolism process, regulation of primary metabolic process, primary metabolic process, etc. According to the KEGG pathways enrichment analysis of rice core nodes, the significantly enriched KEGG pathways were protein processing and endocytosis in the endoplasmic reticulum. These GO functional modules with significant enrichment of rice key proteins were basically consistent with the GO functions of the main modules of the rice regulatory network, which verified the accuracy of the rice core proteins mined through multi-omics joint analysis.

GO functional modules of the *M. oryzae* infecting rice mechanism and rice core nodes for the combined analysis found that the GOs were significantly enriched in the gene expression regulation module, protein synthesis and transport module, and metabolism module. The significant enrichment of gene expression modules indicated that *M. oryzae* silenced rice genes through RNA silencing mechanism to achieve the purpose of infecting rice. In addition, the protein synthesis and transport module showed that *M. oryzae* infected rice by secreted proteins. The module included protein synthesis, nucleic acid binding, protein binding, organic cyclic compound binding, heterocyclic compound binding and transport. These results indicated that *M. oryzae* invaded rice by secreted proteins which combined with some proteins or biomolecules in rice to affect the defense mechanism of rice, thus realizing the infection process. Based on the analysis of KEGG metabolic pathway in rice, it was found that these key proteins in rice affected the metabolic mechanism of rice. After *M. oryzae* infected rice by sRNAs and secreted

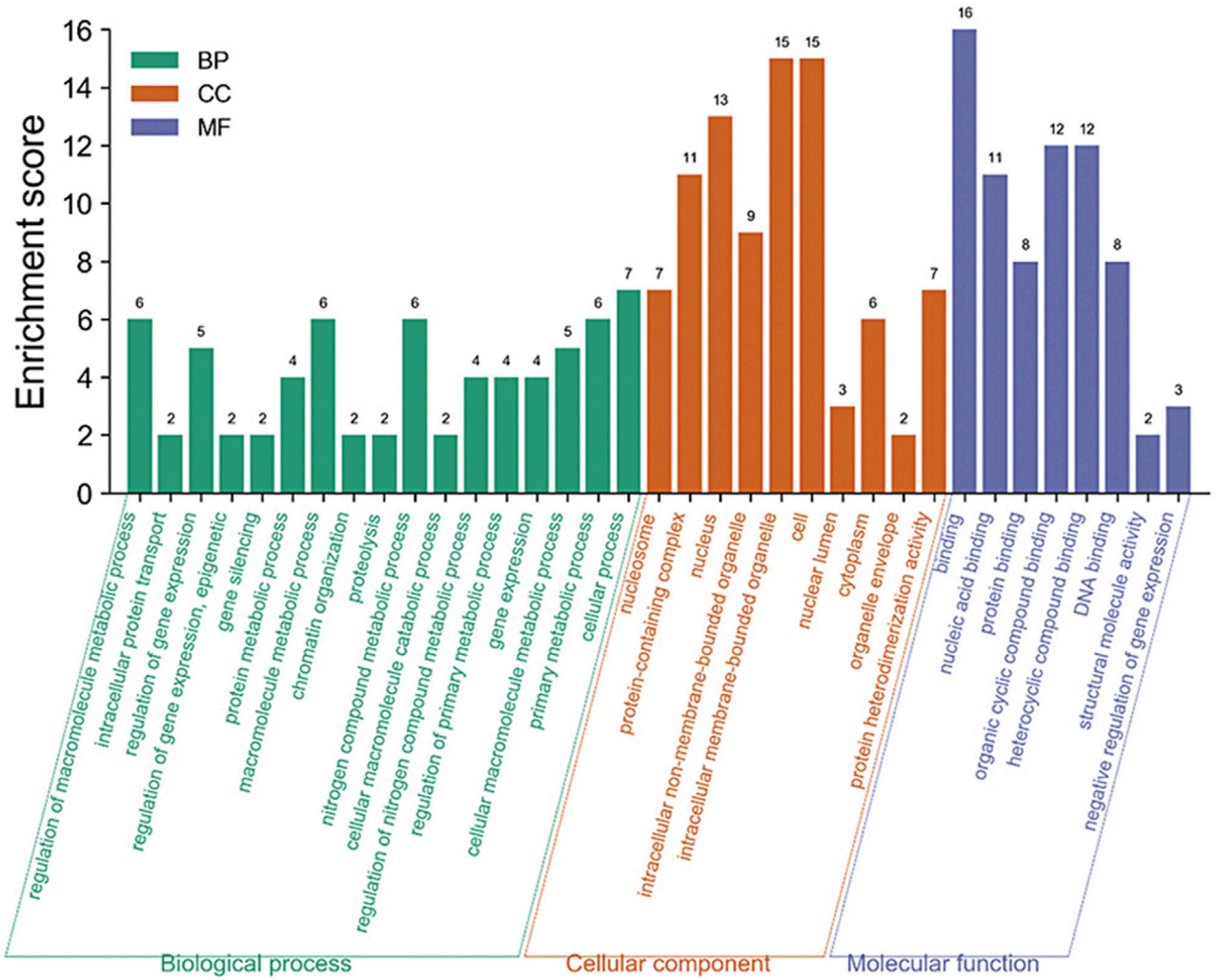

FIGURE 11 | GO pathway enrichment analysis diagram of rice core proteins.

proteins, the rice metabolism was affected, including nitrogen compound metabolism, protein metabolism, other biological macromolecules metabolism, etc. The table of these GO enrichment modules was shown in **Supplementary Table S2**. The GO enrichment function diagram of rice core proteins is shown in **Figure 11**.

DISCUSSION

In this study, a variety of omics data of *M. oryzae* and rice were used to excavate the interaction network between *M. oryzae* and rice to explore the mechanism of *M. oryzae* infection on rice to mine the key nodes involved in the interaction process. The data of each omics used in this paper included sRNA data before and after *M. oryzae* infecting rice, *M. oryzae* mRNA data, *M. oryzae* protein data, *M. oryzae* gene expression data before and after *M. oryzae* infecting rice, rice mRNA data, rice protein data, and protein data of mode host-mode fungus. First, each omics data was screened separately to mine differentially expressed rice gene data, *M. oryzae*-rice sRNA interaction pairs, and *M. oryzae*-rice protein interaction pairs. Then, the interaction network of each omics was analyzed longitudinally to construct the regulatory network of *M. oryzae*-rice multi-omics interaction and explore its biological process.

In genomics, a total of 1,368 1.5-fold differentially expressed rice genes were extracted by screening the gene expression data of rice before and after the infection of *M. oryzae*. In transcriptomics, this study analyzed the sRNA data of *M. oryzae* before and after infection with rice and obtained 366 kinds of up-regulated and newly added sRNAs of *M. oryzae*, which all had the possibility of interacting with host rice, that is, to infect rice by RNA silencing mechanism. In addition, for the 257 species of *M. oryzae* sRNAs whose expression levels were reduced during the infection process, it may be through the regulation of the protein expression in *M. oryzae*, through the secreted protein into the rice to achieve the purpose of infection. Therefore, according to the two infection mechanisms of *M.*

oryzae, the 623 kinds of *M. oryzae* sRNAs screened were analyzed. Furthermore, through the method of target gene prediction, 1,857 sRNA interaction pairs of *M. oryzae*-rice and 664 sRNA interaction pairs of *M. oryzae* were found.

In proteomics, some studies have proved that the secreted proteins of the pathogen can enter the host body and interact with the host proteins to interfere with the protein expression of the host. However, it is not clear which protein molecules are involved in the infection process of *M. oryzae* to affect the defense and growth of rice in the existing studies. In this paper, the protein interaction pairs between mode pathogens and mode hosts that experiments have verified were collected and used as the prediction template. Firstly, the interolog method based on homology was used to predict the protein interaction pairs between *M. oryzae* and rice. Next, the domain-domain method was used to make the second prediction of the protein interaction pairs predicted by the interolog method. Then TMHMM secreted protein prediction tool was used to screen the secreted proteins of *M. oryzae*. In the screening of the final protein interaction pairs, the three prediction methods should be met simultaneously, and 7,352 protein interaction pairs of *M. oryzae*-rice were obtained.

In this study, a total of 8 rice differentially expressed genes, 31 rice mRNAs, 77 rice proteins, 22 *M. oryzae* sRNAs, 1 *M. oryzae* mRNA and 38 *M. oryzae* proteins were identified as the core nodes of the *M. oryzae* and rice multi-omics interaction network by high-throughput data analysis, combined with joint analysis of *M. oryzae* and rice multi-omics data, which involved significantly enriched GO modules. Most of them were related to gene expression, molecular protein synthesis, molecular transport and metabolism, that is, the infection mechanism of *M. oryzae*. However, all the experiments in this paper were based on the premise that sRNA and protein interaction mechanisms exist between *M. oryzae* and rice. The accuracy of this experiment still needs to be further verified. In addition, due to the mutual regulation between plants and pathogens, some host sRNAs and secreted proteins can enter the fungi during the infection process to resist infection. However, this paper only studied the infection mechanism of *M. oryzae* and neglected the analysis of the defense mechanism of rice. Moreover, significant enrichment of biomolecular transport modules was found in the GO function enrichment analysis of key factors of *M. oryzae* in this study, but it is not clear which rice biomolecules are involved in the defense mechanism. And although there are some insufficient, this paper for the *M. oryzae* infecting rice joint analysis of multi-omics data, which provided a specific data basis for further study of the mechanism of *M. oryzae*-rice interaction, made some specific contributions to the prevention of diseases and insect pests in rice and provided a new train of thought and theoretical basis for the fungus-plant interactions mechanism research.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110088; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025; https://rapdb.dna.affrc.go.jp/download/irgsp1.html; https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX214117; https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRX214123; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM973470; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM973471; https://www.ncbi.nlm.nih.gov/Traces/wgs/AACU03?val=AACU03.1; https://www.ncbi.nlm.nih.gov/Traces/wgs/AACU03?val=LVCG01.1; https://hpidb.igbb.msstate.edu/about.html#stats.

AUTHOR CONTRIBUTIONS

Conceptualization, HZ; methodology, HZ, EZ, and XL; software, EZ and XL; validation, EZ; formal analysis, EZ; investigation, XL; resources, XL; data curation, EZ; writing—original draft preparation, EZ and HZ; writing—review and editing, EZ, XL, and HZ; visualization, XL; supervision, HZ; project administration, HZ; funding acquisition, HZ. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by the National Natural Science Foundation of China (Grant No. 62072210).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.763915/full#supplementary-material

REFERENCES

Cai, Q., He, B., Kogel, K.-H., and Jin, H. (2018). Cross-kingdom RNA Trafficking and Environmental RNAi - Nature's Blueprint for Modern Crop protection Strategies. *Curr. Opin. Microbiol.* 46, 58–64. doi:10.1016/j.mib.2018.02.003

Chang, H., Zhang, H., Qin, g. M., Zhang, T., Zhang, T., Liu, Y., et al. (2020). Identification of Novel Phytophthora Infestans Small RNAs Involved in Potato Late Blight Reveals Potential Cross-Kingdom Regulation to Facilitate Oomycete Infection. *Int. J. Data Min. Bioinformatics* 23 (2), 119–141. doi:10.1504/IJDMB.2020.107379

Chujo, T., Miyamoto, K., Shimogawa, T., Shimizu, T., Otake, Y., Yokotani, N., et al. (2013). OsWRKY28, a PAMP-Responsive Transrepressor, Negatively Regulates Innate Immune Responses in rice against rice Blast Fungus. *Plant Mol. Biol.* 82, 23–37. doi:10.1007/s11103-013-0032-5

Cui, C., Wang, J.-J., Zhao, J.-H., Fang, Y.-Y., He, X.-F., Guo, H.-S., et al. (2020). A Brassica miRNA Regulates Plant Growth and Immunity through Distinct Modes of Action. *Mol. Plant* 13 (2), 231–245. doi:10.1016/j.molp.2019.11.010

Deng, Y. Z., and Naqvi, N. I. (2019). Metabolic Basis of Pathogenesis and Host Adaptation in Rice Blast. *Annu. Rev. Microbiol.* 73, 601–619. doi:10.1146/annurev-micro-020518-115810

Faso, C., Chen, Y.-N., Tamura, K., Held, M., Zemelis, S., Marti, L., et al. (2009). A Missense Mutation in the Arabidopsis COPII Coat Protein Sec24A Induces the

Formation of Clusters of the Endoplasmic Reticulum and Golgi Apparatus. *Plant Cell* 21 (11), 3655–3671. doi:10.1105/tpc.109.068262

Grenville-Briggs, L. J., Avrova, A. O., Bruce, C. R., Williams, A., Whisson, S. C., Birch, P. R. J., et al. (2005). Elevated Amino Acid Biosynthesis in Phytophthora Infestans during Appressorium Formation and Potato Infection. *Fungal Genet. Biol.* 42 (3), 244–256. doi:10.1016/j.fgb.2004.11.009

Grohmann, U., and Bronte, V. (2010). Control of Immune Response by Amino Acid Metabolism. *Immunol. Rev.* 236, 243–264. doi:10.1111/j.1600-065X.2010.00915.x

Großkinsky, D. K., Syaifullah, S. J., and Roitsch, T. (2018). Integration of Multi-Omics Techniques and Physiological Phenotyping within a Holistic Phenomics Approach to Study Senescence in Model and Crop Plants. *J. Exp. Bot.* 69 (4), 825–844. doi:10.1093/jxb/erx333

Imam, J., Alam, S., Mandal, N. P., Maiti, D., Variar, M., and Shukla, P. (2015). Molecular Diversity and Mating Type Distribution of the Rice Blast Pathogen Magnaporthe Oryzae in North-East and Eastern India. *Indian J. Microbiol.* 55 (1), 108–113. doi:10.1007/s12088-014-0504-6

Kim, J., Woo, H. R., and Nam, H. G. (2016). Toward Systems Understanding of Leaf Senescence: An Integrated Multi-Omics Perspective on Leaf Senescence Research. *Mol. Plant* 9 (6), 813–825. doi:10.1016/j.molp.2016.04.017

Larsen, P. E., Sreedasyam, A., Trivedi, G., Desai, S., Dai, Y., Cseke, L. J., et al. (2016). Multi-Omics Approach Identifies Molecular Mechanisms of Plant-Fungus Mycorrhizal Interaction. *Front. Plant Sci.* 6, 1061. doi:10.3389/fpls.2015.01061

Lee, H., Deng, M., Sun, F., and Chen, T. (2006). An Integrated Approach to the Prediction of Domain-Domain Interactions. *BMC Bioinformatics* 7, 269. doi:10.1186/1471-2105-7-269

Li, W., Zhu, Z., Chern, M., Yin, J., Yang, C., Ran, L., et al. (2017). A Natural Allele of a Transcription Factor in Rice Confers Broad-Spectrum Blast Resistance. *Cell* 170 (1), 114–126. doi:10.1016/j.cell.2017.06.008

Majumdar, R., Rajasekaran, K., and Cary, J. W. (2017). RNA Interference (RNAi) as a Potential Tool for Control of Mycotoxin Contamination in Crop Plants: Concepts and Considerations. *Front. Plant Sci.* 8, 200. doi:10.3389/fpls.2017.00200

McGaha, T. L., Huang, L., Lemos, H., Metz, R., Mautino, M., Prendergast, G. C., et al. (2012). Amino Acid Catabolism: a Pivotal Regulator of Innate and Adaptive Immunity. *Immunol. Rev.* 249 (1), 135–157. doi:10.1111/j.1600-065X.2012.01149.x

Mueth, N. A., Ramachandran, S. R., and Hulbert, S. H. (2015). Small RNAs from the Wheat Stripe Rust Fungus (Puccinia Striiformis f.Sp. Tritici). *BMC Genomics* 16 (1), 718. doi:10.1186/s12864-015-1895-4

Nelson, R., Wiesner-Hanks, T., Wisser, R., and Balint-Kurti, P. (2018). Navigating Complexity to Breed Disease-Resistant Crops. *Nat. Rev. Genet.* 19 (1), 21–33. doi:10.1038/nrg.2017.82

Parker, D., Beckmann, M., Zubair, H., Enot, D. P., Caracuel-Rios, Z., Overy, D. P., et al. (2009). Metabolomic Analysis Reveals a Common Pattern of Metabolic Re-programming during Invasion of Three Host Plant Species byMagnaporthe Grisea. *Plant J.* 59 (5), 723–737. doi:10.1111/j.1365-313X.2009.03912.x

Pompa, A., De Marchis, F., Pallotta, M. T., Benitez-Alfonso, Y., Jones, A., Schipper, K., et al. (2017). Unconventional Transport Routes of Soluble and Membrane Proteins and Their Role in Developmental Biology. *Int. J. Mol. Sci.* 18 (4), 703. doi:10.3390/ijms18040703

Raman, V., Simon, S. A., Demirci, F., Nakano, M., Meyers, B. C., and Donofrio, N. M. (2017). Small RNA Functions Are Required for Growth and Development of Magnaporthe Oryzae. *MPMI* 30 (7), 517–530. doi:10.1094/MPMI-11-16-0236-R

Raman, V., Simon, S. A., Romag, A., Demirci, F., Mathioni, S. M., Zhai, J., et al. (2013). Physiological Stressors and Invasive Plant Infections Alter the Small RNA Transcriptome of the rice Blast Fungus, Magnaporthe Oryzae. *BMC Genomics* 14, 326. doi:10.1186/1471-2164-14-326

Riquelme, M., Aguirre, J., Bartnicki-García, S., Braus, G. H., Feldbrügge, M., Fleig, U., et al. (2018). Fungal Morphogenesis, from the Polarized Growth of Hyphae to Complex Reproduction and Infection Structures. *Microbiol. Mol. Biol. Rev.* 82 (2), e00068–17. doi:10.1128/MMBR.00068-17

Ruiz-Sánchez, M., Aroca, R., Muñoz, Y., Polón, R., and Ruiz-Lozano, J. M. (2010). The Arbuscular Mycorrhizal Symbiosis Enhances the Photosynthetic Efficiency and the Antioxidative Response of rice Plants Subjected to Drought Stress. *J. Plant Physiol.* 167 (11), 862–869. doi:10.1016/j.jplph.2010.01.018

Solomon, P. S., and Oliver, R. P. (2001). The Nitrogen Content of the Tomato Leaf Apoplast Increases during Infection by Cladosporium Fulvum. *Planta* 213 (2), 241–249. doi:10.1007/s004250000500

Thanasomboon, R., Kalapanulak, S., Netrphan, S., and Saithong, T. (2017). Prediction of Cassava Protein Interactome Based on Interolog Method. *Sci. Rep.* 7 (1), 17206. doi:10.1038/s41598-017-17633-2

Wang, X., Zhang, R., Shi, Z., Zhang, Y., Sun, X., Ji, Y., et al. (2019). Multi-omics Analysis of the Development and Fracture Resistance for maize Internode. *Sci. Rep.* 9 (1), 8183. doi:10.1038/s41598-019-44690-6

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., et al. (2013). Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways. *Science* 342 (6154), 118–123. doi:10.1126/science.1239705

Yang, F., Jensen, J. D., Svensson, B., Jørgensen, H. J. L., Collinge, D. B., and Finnie, C. (2012). Secretomics Identifies Fusarium Graminearum Proteins Involved in the Interaction with Barley and Wheat. *Mol. Plant Pathol.* 13 (5), 445–453. doi:10.1111/j.1364-3703.2011.00759.x

Zhang, H., Liu, S., Chang, H., Zhan, M., Qin, Q.-M., Zhang, B., et al. (2019). Mining Magnaporthe Oryzae sRNAs with Potential Transboundary Regulation of Rice Genes Associated with Growth and Defense through Expression Profile Analysis of the Pathogen-Infected Rice. *Front. Genet.* 10, 296. doi:10.3389/fgene.2019.00296

Zhang, Y., Xia, R., Kuang, H., and Meyers, B. C. (2016). The Diversification of PlantNBS-LRRDefense Genes Directs the Evolution of MicroRNAs that Target Them. *Mol. Biol. Evol.* 33 (10), 2692–2705. doi:10.1093/molbev/msw154

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 November 2021
doi: 10.3389/fgene.2021.779186

Graph Embedding Based Novel Gene Discovery Associated With Diabetes Mellitus

Jianzong Du[1†], *Dongdong Lin*[1†], *Ruan Yuan*[1], *Xiaopei Chen*[1], *Xiaoli Liu*[1*] *and Jing Yan*[1,2*]

[1]*Zhejiang Hospital, Hangzhou, China,* [2]*Zhejiang Provincial Key Lab of Geriatrics, Zhejiang Hospital, Hangzhou, China*

Diabetes mellitus is a group of complex metabolic disorders which has affected hundreds of millions of patients world-widely. The underlying pathogenesis of various types of diabetes is still unclear, which hinders the way of developing more efficient therapies. Although many genes have been found associated with diabetes mellitus, more novel genes are still needed to be discovered towards a complete picture of the underlying mechanism. With the development of complex molecular networks, network-based disease-gene prediction methods have been widely proposed. However, most existing methods are based on the hypothesis of guilt-by-association and often handcraft node features based on local topological structures. Advances in graph embedding techniques have enabled automatically global feature extraction from molecular networks. Inspired by the successful applications of cutting-edge graph embedding methods on complex diseases, we proposed a computational framework to investigate novel genes associated with diabetes mellitus. There are three main steps in the framework: network feature extraction based on graph embedding methods; feature denoising and regeneration using stacked autoencoder; and disease-gene prediction based on machine learning classifiers. We compared the performance by using different graph embedding methods and machine learning classifiers and designed the best workflow for predicting genes associated with diabetes mellitus. Functional enrichment analysis based on Human Phenotype Ontology (HPO), KEGG, and GO biological process and publication search further evaluated the predicted novel genes.

Keywords: diabetes mellitus, graph embedding, novel gene discovery, molecular network, disease gene prediction

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Sheng Yang,
Nanjing Medical University, China
Zhen Tian,
Zhengzhou University, China

***Correspondence:**
Xiaoli Liu
liuxiaoli1010@126.com
Jing Yan
zjicu@vip.163.com

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

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

Received: *18 September 2021*
Accepted: *20 October 2021*
Published: *25 November 2021*

Citation:
Du J, Lin D, Yuan R, Chen X, Liu X and Yan J (2021) Graph Embedding Based Novel Gene Discovery Associated With Diabetes Mellitus. Front. Genet. 12:779186. doi: 10.3389/fgene.2021.779186

INTRODUCTION

Diabetes mellitus is a chronic disease where the blood sugar in patients is abnormally elevated because of the underproductive pancreas or the ineffective response toward insulin (Kharroubi and Darwish, 2015). According to the global diabetes map (ninth edition) published by the International Diabetes Federation (IDF) in 2019 (Cho et al., 2018), the number of diabetic patients worldwide is increasing, with an average global growth rate of 51%. There are currently 463 million diabetic patients. According to the growing trend, there will be 700 million diabetic patients worldwide by 2045 (Cho et al., 2018). Diabetes mellitus and its multiple complications have largely increased the risk of mortality, blindness, and kidney failure of patients, and posed a heavy burden on human society. It is urgent to investigate the disease mechanisms and find more effective cures.

There are different types of diabetes: type 1 diabetes (T1D), type 2 diabetes (T2D), gestational diabetes and other types (Geerlings and Hoepelman, 1999; Kharroubi and Darwish, 2015). For

different types of diabetes, the causes and risk factors vary. Type 1diabetes is an autoimmune disease, where the insulin-producing cells in the pancreas are attacked by the immune system of patients. The pathogenesis of type 1 diabetes is still unclear, but some researchers think it is caused by a combination of genetic and environmental factors. The genome-wide association studies (GWAS) have identified over 60 susceptibility loci for T1D (Systematic evaluation of genes and genetic variants associated with Type 1 diabetes susceptibility). And post-GWAS functional analyses (Shabalin, 2012; Westra et al., 2013; Fagny et al., 2017; Wang et al., 2019a; van der Wijst et al., 2020) such as expression quantitative trait loci (eQTL) analysis have been performed to infer the underlying causal genes (Nyaga et al., 2018). Cells become resistant to insulin in type 2 diabetes, resulting in higher demand for insulin. However, the dysfunction of pancreatic β cells decreases secretion of insulin, leading to evaluated blood sugar levels in patients. The pathogenesis of T2D is also unclear, but the genetic studies of T2D provided novel susceptibility loci and candidate genes. Similarly, the mechanisms of other types of diabetes are also not clear. It is urgent to discover genes associated with diabetes mellitus to find therapeutic targets and improve diagnoses (Kharroubi and Darwish, 2015).

There have been intense efforts to predict genes associated with complex diseases in recent years (Ghiassian et al., 2015; Peng et al., 2017; Agrawal et al., 2018; Cheng et al., 2019; Wang et al., 2020). GWASs can directly reveal the associations between genome variants and diseases (Zhu et al., 2016a; Zhu et al., 2016b; Visscher et al., 2017; Gallagher and Chen-Plotkin, 2018; Visscher and Goddard, 2019). However, most GWAS SNPs locate in non-coding regions, i.e., intronic or inter-genetic regions, leading to a limited discovery of disease genes. Functional analysis, such as eQTL analysis (Wang et al., 2021a; Wang et al., 2021b), can further translate GWAS signals to functional genes through measuring the regulation pattern between genomic variations (genotypes) and transcriptome variations (gene expression level). These statistical methods have achieved tremendous success in discovering disease-associated genes. And these discoveries have also been recorded in biological databases such as DisGeNet (Piñero et al., 2015; Piñero et al., 2016; Piñero et al., 2020). However, these methods mostly are based on simple "gene-disease" associations and ignore the underlying functional collaborations among genes.

With the development of molecular networks, such as protein-protein interaction (PPI) networks and gene regulatory networks, it is feasible to investigate disease genes based on gene networks (Peng et al., 2021a). Under the hypothesis of guilt–by–association (GBA), the novel disease-associated genes can be predicted by measuring the neighborhood structures of known disease genes. In recent years, there have been many network-based methods emerging as powerful tools for disease-gene prediction (Wang et al., 2019b; Wang et al., 2019c; Yang et al., 2019). The task of disease-gene prediction can be considered as a classification problem in machine learning. There are two types of classification in disease-gene prediction based on the types of entity the methods aim to predict. One is node classification, where genes in the gene network can be separated into two groups: known disease-genes and unlabeled genes, and the prediction methods aim to give a rank to unlabeled genes based on the prediction model. Top-ranked genes will be predicted as novel disease genes. Methods such as PRINCE (Vanunu et al., 2010), VAVIEN (Erten et al., 2011), and N2A-SVM (Peng et al., 2019a) belong to this category. The other type of classification in disease-gene prediction is edge classification, also called link prediction. In this category, genes and diseases both exist in the network as nodes, which comprise a heterogeneous graph. The prediction methods learn features from known disease-gene edges and predict novel disease-gene links. The feature of a disease-gene link is combined from a pair of node features. Methods such as RWRH (Li and Patra, 2010) and RWPCN (Yang et al., 2011) belong to this category.

From the aspect of features extracted from the network, the disease-gene prediction methods can be separated into handcrafted feature-based methods and automatic feature representation-based methods. In the first category, methods engineered features for nodes in biological networks, such as using node degree, graphlet degree, common neighbors, shortest path length meta-paths, etc. However, methods relying on direct neighborhood counting can only capture the local network structure while ignoring the global structure. To overcome this issue, Xu et al. proposed a method by integrating multiple topological features to predict disease genes (Xu and Li, 2006). In their methods, they expanded the neighbors of a seed by considering 2-hop neighbors. Besides the network topological structure, some methods integrated more biological data as features. DERanking (Nitsch et al., 2010) incorporated differential expression in features. BRIDGE (Chen et al., 2013) integrated multiple data sources besides the PPI network, such as gene expression, gene ontology (GO), and the KEGG database. DiGI (Tran et al., 2020) used gene co-expression network, functional pathways, PPI network, and other cofunction networks in feature engineering. Although these methods based on handcrafted features have achieved tremendous success in multiple fields, there needs a lot of domain knowledge and it may also introduce biases with manually engineered features.

In recent years, graph embedding learning methods emerged as powerful tools for extracting the latent features from networks. Graph embedding is also known as graph representation learning, aiming at mapping large and sparse graph data into low-dimensional dense feature vectors. There are matrix factorization-based graph embedding methods [such as IMC (Natarajan and Dhillon, 2014) and PCFM (Zeng et al., 2017)], and also methods based on skip-gram based neuron networks [such as LINE (Tang et al., 2015), DeepWalk (Perozzi et al., 2014), and Node2Vec (Grover and Leskovec, 2016)], and graph neuron networks [such as graph convolutional network (Wu et al., 2020)]. These techniques have been widely used in bioinformatics applications such as the discovery of antibiotics (Stokes et al., 2020), disease genes (Peng et al., 2021b), disease modules (Wang et al., 2020), drug targets (Peng et al., 2021c), drug side-effects (Han et al., 2021), RNA-targets (Peng et al., 2019b), molecular network edges (Perozzi et al., 2014; Ribeiro et al., 2017; Peng et al., 2021d), etc. However, there has been a lack of research on discovering genes associated with diabetes mellitus using cutting-edge graph-embedding techniques. In this study,

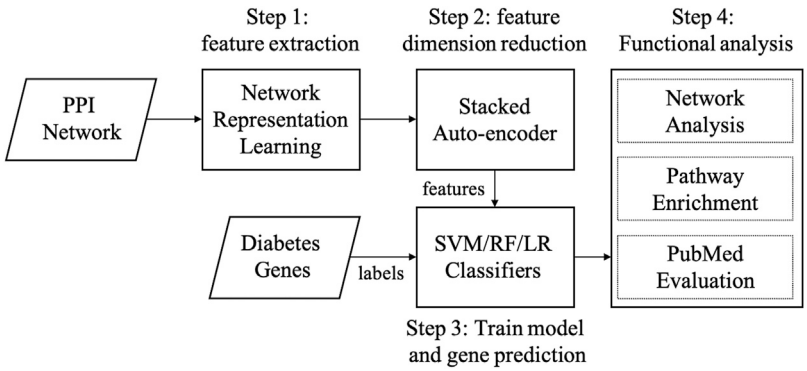

FIGURE 1 | Workflow of our method. Abbreviations: SVM: supporting vector machine, RF: random forest, LR: logistic regression.

we designed a computational framework based on graph embedding approaches to discover novel genes associated with diabetes mellitus without distinction between diabetes types. We first extracted gene features from a PPI network. During this phase, we compared three cutting-edge graph embedding methods, i.e., LINE (Tang et al., 2015), DeepWalk (Perozzi et al., 2014), and Node2Vec (Grover and Leskovec, 2016). Next, we applied a stacked auto-encoder to further process the node embeddings into lower-dimensional space. Finally, we used widely-used machine learning classifiers for the task of gene prediction. In the experiments, we evaluated the performance of our model by using five-fold cross-validation, and we also compared the performance using various graph embedding methods, hyper-parameters, and machine learning classifiers.

METHODOLOGY

There are three main steps in our graph embedding based diabetes-gene prediction model: 1) we used three cutting graph embedding methods, i.e., LINE, DeepWalk, and Node2Vec, to extract node features from a PPI network; 2) A three-layer stacked autoencoder was applied to further reduce feature dimension and automatic feature extraction; 3) disease gene prediction using support vector machine (SVM) (Chang and Lin, 2011), and other two widely-used classifiers (random forest and logistic regression) were compared. Four metrics (AUPRC, AUROC, ACC, and F1 score) were used to measure the performance in five-fold cross-validation. Functional enrichment and network analysis were applied for evaluation. The workflow of our method is shown in **Figure 1**.

Extract Features From PPI Network Based on Graph Embedding

To extract the latent feature from PPI network, we adopt three cutting-edge graph embedding methods: Node2vec, DeepWalk, and LINE, and compared their performance in the task of predicting genes associated with diabetes mellitus. DeepWalk draws on the idea of the word2vec algorithm. Word2vec is a commonly used word embedding method in natural language learning (NLP). It describes the co-occurrence relationship between words and words through the sentence sequence in the corpus and then learns the vector representation of words based on skip-gram neuronal network model. The DeepWalk algorithm is similar to word2vec and uses the co-occurrence relationship between nodes in the graph to learn the vector representation of nodes. DeepWalk uses random walk to sample paths with fixed lengths. The paths are consisted of randomly visited nodes and are similar to sentences in NLP. And then word2vec is used to learn the co-occurrence relationship of nodes based on skip-gram neuronal network model. The weights on the hidden layer of skip-gram model will be the latent features.

Node2vec is a graph embedding method improved based on DeepWalk. The novel part of Node2vec is that it uses a biased random walk process to generate random paths. The hyperparameters p and q are used to control the directions of random walk in consonance with breadth-first search (BFS) or depth-first search (DFS) in the PPI network. Parameter p determines the process of revisiting the nodes within random walk and q affects the possibility of capturing local or global nodes. Compared to DeepWalk, Node2vec provides more various elements, and particularly, if the value of p and q both equal 1, these two algorithms are the same.

LINE is also a method based on the assumption of neighborhood similarity, except that LINE uses BFS to construct neighborhoods while DeepWalk uses DFS to construct neighborhoods. LINE also takes into account the first-order and second-order similarities between nodes and can be applied to various types of networks and large-scale networks. However, some vertices have few adjacent points, which leads to insufficient learning of embedding vectors and insufficient use of high-level information.

Feature Regeneration and Reduction Using Stacked Autoencoder

Autoencoder is an unsupervised artificial neural network that can automatically extract latent features from data. Autoencoder has been successfully applied in many applications, such as speech recognition, self-driving cars, human gesture detection, etc. The

autoencoder structure is composed of three parts: the input layer, the hidden layer, and the output layer, which correspond to the encoder, bottleneck and decoder respectively. Among them, the encoder is responsible for selecting key features from the data, and the decoder is responsible for recreating the original data using key components. Since the number of hidden layer nodes is less than the number of input nodes, the autoencoder can reduce the data dimension by retaining only the features needed to reconstruct the data. The autoencoder is also a feed-forward network, which can be trained using the same procedure as the feed-forward network. Although Autoencoder has the same input and output, it also has a certain degree of loss, so autoencoder is also called lossy compression technology.

Since there are complicated relationships within the elements in some data sets, only one autoencoder cannot meet the requirements. To reduce the dimensionality of the input features, a single autoencoder may not be able to complete it. In response to this situation, the stacked autoencoder was proposed. As the name suggests, stacked autoencoders are multiple autoencoders stacked on top of each other. The specific process of the stacked autoencoder method is described as follows: First, given the initial input, train the first-layer autoencoder in an unsupervised way to reduce the reconstruction error to the set value. Second, take the output of the hidden layer of the first autoencoder as the input of the second autoencoder, and use the same method as above to train the autoencoder. Third, repeat the second step until all autoencoders are initialized. Finally, use the weights of the hidden layer of the last stacked autoencoder as the final features.

Machine Learning Classifiers Used for Disease Gene Prediction

After the process of network representation learning and feature denoising, we apply classification methods for the final prediction task. Three widely-used machine learning algorithms were used for predicting genes associated with diabetes mellitus: support vector machine (SVM), Logistic regression, and Random Forest. Logistic regression models the relationship between predictor variables and a categorical response variable. Given feature vector $\boldsymbol{x}$ and the label $y \in \{0, 1\}$ of each sample, the logistic regression models feature $\boldsymbol{x}$ and the probability of y by **Eq. 1**, where $\boldsymbol{w}$ represents weights and b represents bias. This equation means the log odds of prediction $y = 1$ equals linear regression of input feature $\boldsymbol{x}$. The parameters $\boldsymbol{w}$ and b can be estimated by maximum likelihood estimation.

$$\boldsymbol{w}^T\boldsymbol{x} + b = ln\frac{p(y=1|\boldsymbol{x})}{1-p(y=1|\boldsymbol{x})} \text{ i.e., } p(y=1|\boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{w}^T\boldsymbol{x}+b)}} \tag{1}$$

Random Forest is an integrated algorithm composed of decision trees, which achieves excellent performance in many applications. Decision tree is a supervised learning algorithm based on "if-then-else" rules. When we perform the classification task, the input samples are classified by each decision tree separately. And each decision tree will get its own classification result. Those decision trees form the random forest, and it will ensemble all prediction results, and output the label with the most consistent evidence.

Support vector machines (SVM) is a binary classification model. Its basic model is a linear classifier featured with the largest interval between two classes in the feature space. Kernel techniques can be applied to SVM, which makes it a non-linear classifier. The learning strategy of SVM is to maximize the interval, which can be formalized as a problem of solving convex quadratic programming. As shown in **Eq. 2**, the SVM model is to construct the hyperplane ($\boldsymbol{\omega}$ is the variable coefficient, γ is the constant), so that the labels of the samples can be divided correctly.

$$\boldsymbol{\omega}\boldsymbol{x}^T + \gamma = 0 \tag{2}$$

Metrics for Evaluating Prediction Performance

In the task of binary classification, samples in the test set can be separated into four classes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). And the sample size of the test set (N) equals to the sum of TP, TN, FP, and FN. Based on these measures, we used four metrics to evaluate the prediction performance: accuracy (ACC), area under the receiver operating characteristic curve (AUROC), area under the precision and recall curve (AUPRC) and F1 score. The accuracy is defined as the ratio of number of correctly predicted samples (TP + TN) and the sample size of the test set (N). However, ACC is not robust in study with unbalanced samples, which means there is only a small number of positive/negative sample. The other three metrics can solve this problem to some extent. The PR curve is defined based on precision and recall which are defined in **Eqs 3**, **4**, respectively. The precision and recall are on y and x-axis respectively. Since there are N possible thresholds of prediction probability, there would be N points, i.e., (precision, recall) on the PR curve.

$$precision = \frac{TP}{TP+FP} \tag{3}$$

$$recall = \frac{TP}{TP+FN} \tag{4}$$

Similarly, the ROC is defined based on true positive rate (TPR) and false positive rate (FPR), which are defined in **Eqs 5**, **6** respectively. In ROC, the TPR and FPR are on y and x-axis respectively. F1 score is a combination of precision and recall, which is defined in **Eq. 7**.

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

$$FPR = \frac{FP}{TN+FP} \tag{6}$$

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

The area under ROC and PRC (AUROC and AUPRC) are widely used to compare the performance of different classifiers. Given a

series of points $\{(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)\}$ on the ROC or PRC curve, the area under the curve (AUC) can be approximately computed by **Eq. 8**.

$$AUC = \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i) \cdot (y_i + y_{i+1}) \tag{8}$$

RESULTS AND DISCUSSION

Datasets

We first downloaded the diabetes mellitus associated genes from DisGeNet database (as of June 2021, UMLS CUI: C0011849). 2,803 genes were recorded in this database, and each gene was assigned with a gene-disease association (GDA) score, indicating the levels of evidence. The GDA score takes into account the number and type of sources (level of curation, organisms), and the number of publications supporting the association. After filtering GDA score with threshold set to 0.1, there were 476 genes left that were used for model training in the downstream prediction.

The protein-protein interaction network was obtained from Menche et al.'s work (Menche et al., 2015). This PPI network consists of multiple sources of protein interactions, such as regulatory interactions, yeast two-hybrid high-throughput interactions, literature curated databases, metabolic enzyme-coupled interactions, protein-protein complexes, etc. By combining those interactions, we obtained this PPI network of 13,460 proteins and 141,296 interactions.

Network Representation Learning Using DeepWalk, LINE, and Node2vec

We extracted the node features of the PPI network using the technique of network representation learning or graph embedding, which maps the topological features of nodes in the network into the embedding space. To choose a proper method, three cutting-edge network representation learning methods were used for feature extraction. And we compared their performance using five-fold cross-validation. To balance the sample size of positive samples and negative samples, we randomly selected the same number of nodes not labeled as disease genes as negative samples.

We run these methods on the PPI network and generate features with 512 dimensions. Then the features were further processed by a stacked autoencoder with three levels, which will reduce noises and generate latent features. The 512-dimensional features were converted to 64-dimensional features using this autoencoder. And SVM was used for final classification using the same setting parameters.

Figure 2 shows the average AUROC, AUPRC, F1 score, and accuracy (ACC) values of three methods achieved in this experiment. We can see that Node2vec achieves the best performance under all metrics. And DeepWalk is the second-best method. This is easy to understand because Node2vec improves DeepWalk by a biased random-walk strategy (see details in Methods).

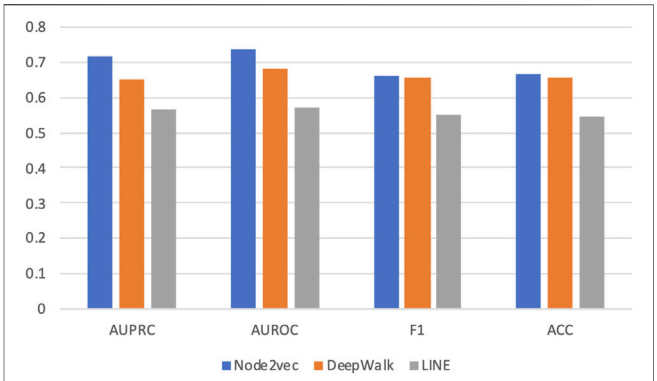

FIGURE 2 | Prediction performance in five-fold cross validation based on three graph embedding methods. Three different graph embedding methods are compared: DeepWalk, LINE, and Node2vec. Four metrics are used for performance evaluation: AUROC, AUPRC, F1 score, and accuracy (ACC).

Feature Dimension Affects Prediction Performance

As a non-end-to-end model, our framework first generates features of network nodes and then predicts disease-associated genes based on SVM. All of the three network-representation-learning methods mentioned above are based on a skip-gram neuron network model, where the dimension of output features equals the number of neurons in the hidden layer of skip-gram neuron network. To explore the impact of feature dimensions on our predicting framework, we compared the performance of the representation learning methods with various dimensional features extracted from the PPI network. Those features were all converted to 64-dimensional features using the stacked autoencoder described above, followed by the SVM classifier under the same settings (RBF kernel and other settings in default).

Based on five-fold cross-validation, we got the results shown in **Figure 3**. The four sub-panels in **Figure 3** represent the prediction performance on diabetes genes using different feature dimensions (i.e., 64, 128, 256, and 512 feature dimensions) generated by three network representation learning methods. The average AUROC, AUPRC, F1 score, and ACC values were compared.

When the feature dimension equals 64, Node2vec achieved the best performance in ACC, F1 score, and AUROC. And LINE achieved the best performance in AUPRC and the second-best performance on ACC and F1 score. While as the feature dimension increased to 128 and 256, the DeepWalk achieved the best performance, and Node2vec achieved the second-best rank. However, The Node2vec achieved the maximum AUROC (0.74) and AUPRC (0.72) scores with 512 feature dimensions compared with other methods in various feature dimensions. In summary, the feature dimension and network representation learning method both affect the prediction performance in a

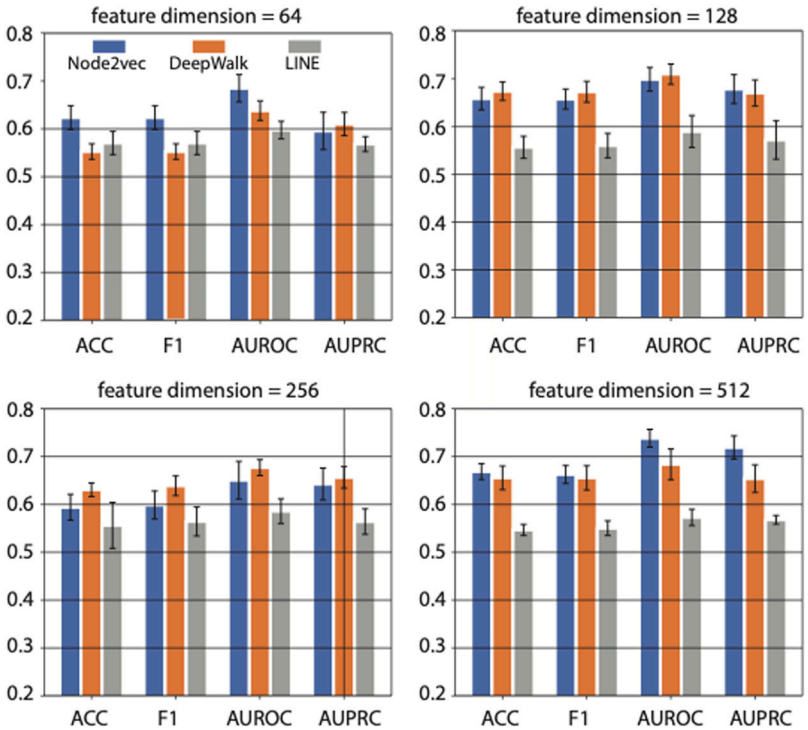

FIGURE 3 | The effects of feature dimension on prediction performance. Four feature dimensions (i.e., 64, 128, 256, and 512) generated by graph embedding methods are used for comparison. Three different graph embedding methods are also compared.

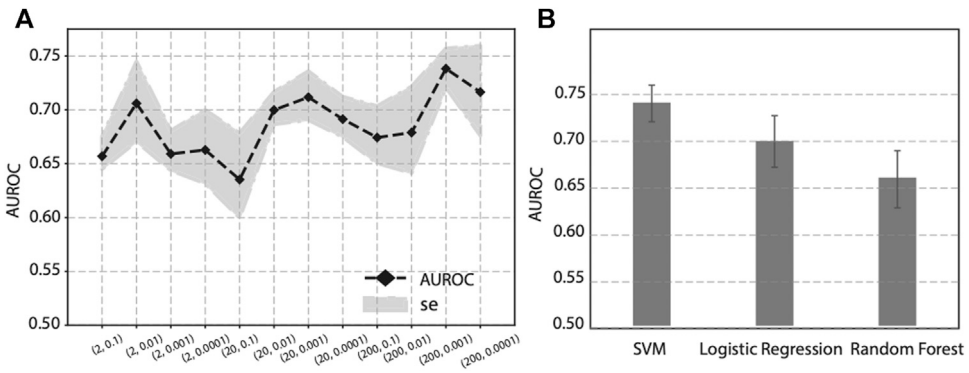

FIGURE 4 | Effect on prediction performance by hyper-parameters in Node2vec and different machine learning classifiers. **(A)** Prediction performance under various *p* and *q* values in Node2vec. **(B)** Prediction performance of SVM, Logistic regression and Random Forest in five-fold cross validation.

task-dependent way. In our case, i.e., predicting genes associated with diabetes mellitus, we choose Node2vec as the method of feature learning from PPI network, and output 512-dimensional features in downstream analysis.

Exploring the Effect of Hyper-Parameters in Node2vec and Different Classifiers

As previous publications have pointed out, the hyper-parameter *p* and *q*, in Node2vec have potential influence to feature learning and downstream analysis. To optimize the two parameters, we performed a grid search on *p* and *q*, and calculated the corresponding performance. Since p controls the random walk to visit new nodes or visited nodes, we set *p* in a larger manner to encourage the random walk to visit new nodes, and we choose $p \in (2, 20, 200)$. And q controls the random walk towards a BFS or DFS graph search. To let the random walk be biased to a DFS search, we set $q \in (0.1, 0.01, 0.001, 0.0001)$. The performance of various *p* and q values is shown in **Figure 4A**. It seems there is not a linear relationship between (*p*, *q*) values and the performance.

TABLE 1 | Top 15 genes predicted associated with diabetes mellitus.

Gene id	Gene name	Gene description	Score
331	BIRC4	X-linked inhibitor of apoptosis	0.78
7098	TLR3	Toll like receptor 3	0.77
55905	ZNF313	Ring finger protein 114	0.76
8915	BCL10	BCL10 immune signaling adaptor	0.76
3654	IRAK1	Interleukin 1 receptor associated kinase 1	0.75
3659	IRF1	Interferon regulatory factor 1	0.75
84270	CARD19	Caspase recruitment domain family member 19	0.75
64320	RNF25	Ring finger protein 25	0.75
340061	TMEM173	Stimulator of interferon response CGAMP interactor 1	0.74
59307	SIGIRR	Single ig and TIR domain containing	0.74
9451	EIF2AK3	Eukaryotic translation initiation factor 2 alpha kinase 3	0.74
5608	MAP2K6	Mitogen-activated protein kinase 6	0.73
51135	IRAK4	Interleukin 1 receptor associated kinase 4	0.73
220885	RPSAP15	Ribosomal protein SA pseudogene 15	0.73
9344	TAOK2	TAO kinase 2	0.73

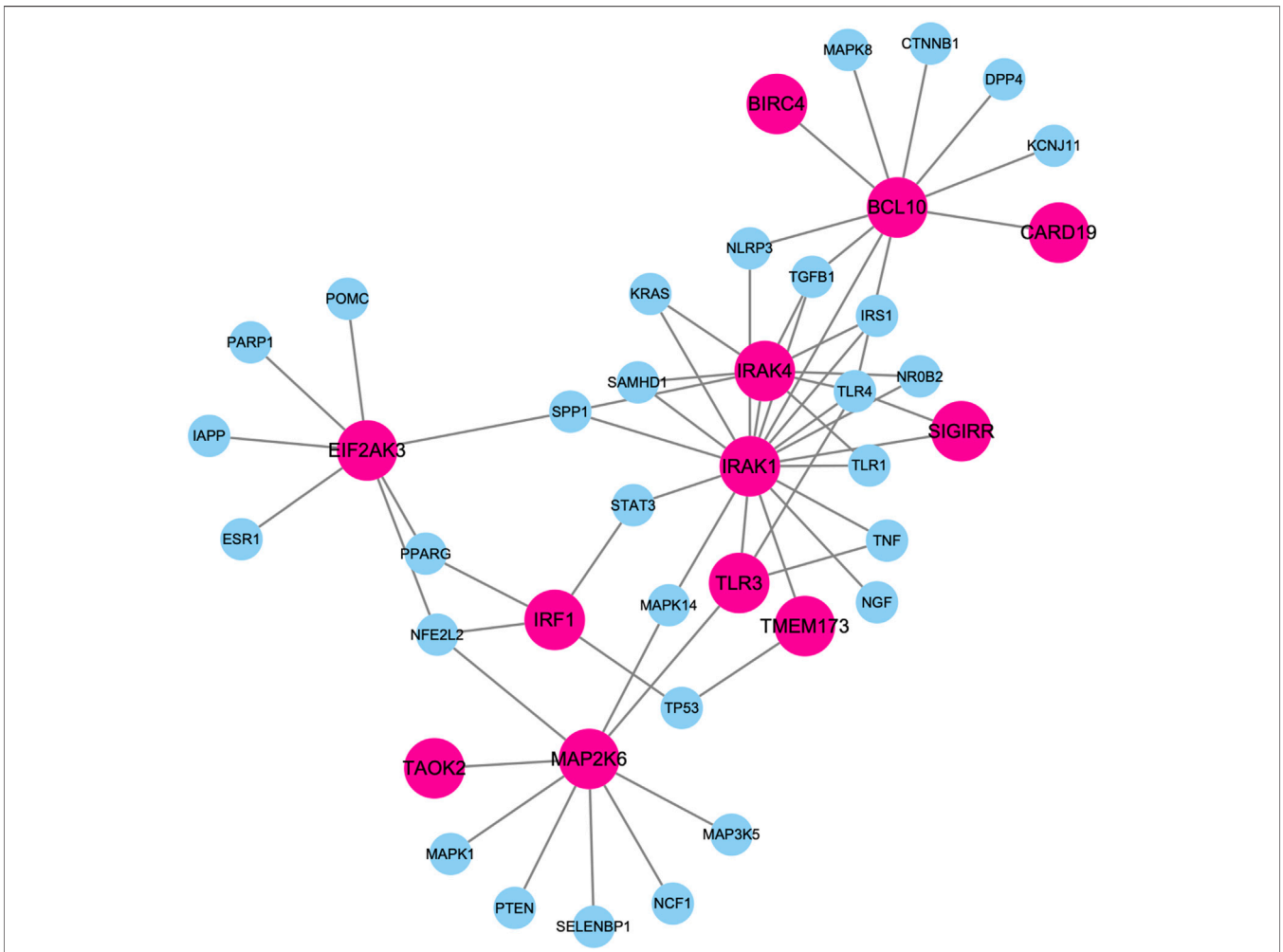

FIGURE 5 | Largest component of PPI subnetwork among these top-predicted genes and known genes associated with diabetes mellitus. Nodes in pink represent top predicted genes. Nodes in blue represent know diabetes genes.

As we can see, when $p = 200$ and $q = 0.001$, it achieves the best performance (AUROC = 0.74) on this specific task, i.e., prediction genes associated with diabetes mellitus. Since the best combination of (p, q) values varies from study to study, it is recommended to perform a grid search to find the best hyperparameters.

To evaluate the effect of different classifiers, we compared SVM with two other widely-used classifiers: Logistic regression and Random Forest. Using the same features obtained from Node2vec followed by a stacked autoencoder, we compared the prediction performance of SVM, Logistic regression, and Random Forest in five-fold cross-validation. The results are shown in **Figure 4B**, where we can see SVM achieves the best performance than Logistic regression and Random Forest. Based on this analysis, our prediction model will use SVM as classifier to predict genes associated with diabetes mellitus.

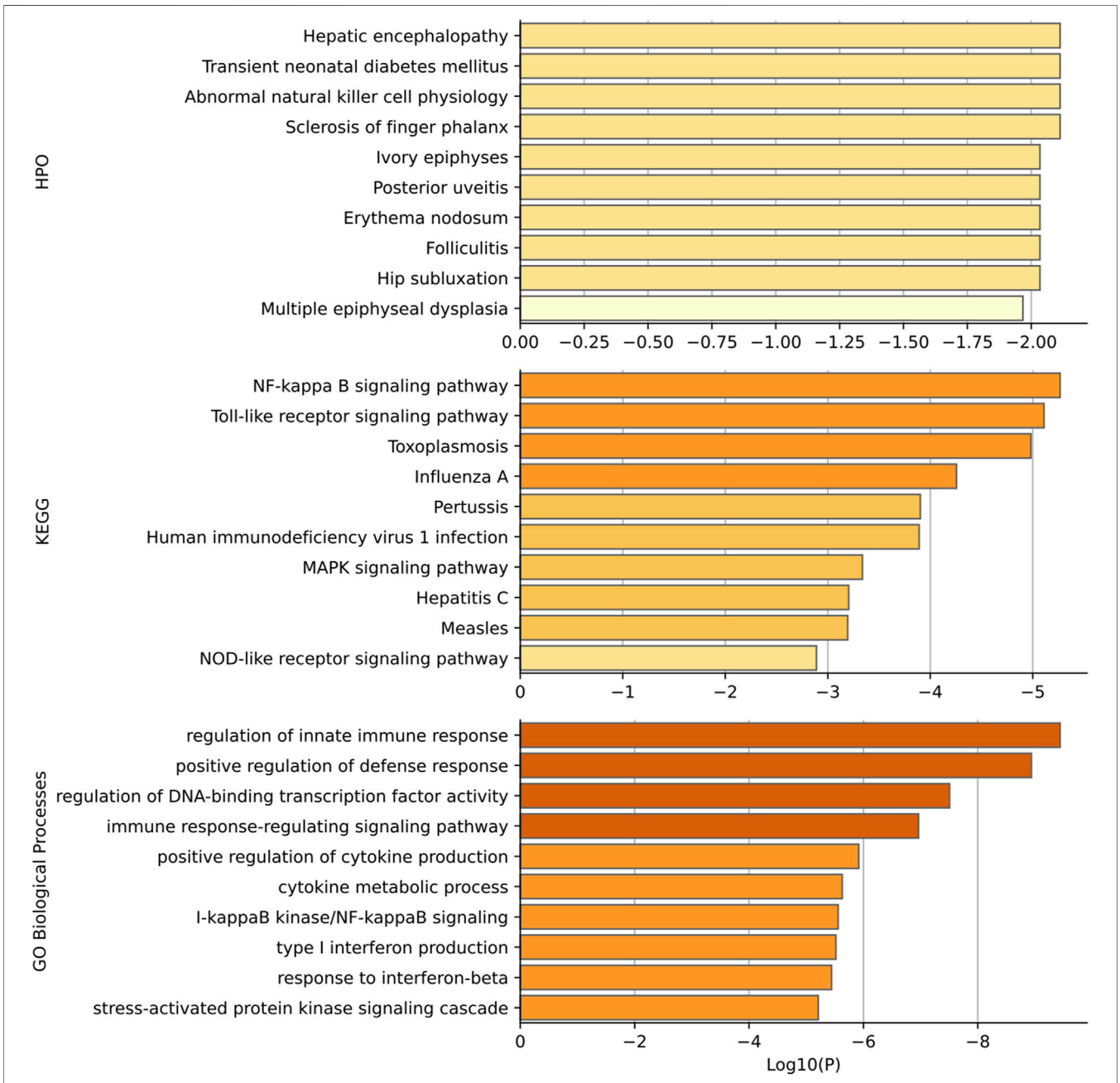

FIGURE 6 | Functional enrichment results based on HPO, KEGG, and GO. p-values are shown in log scale and only top 10 terms are shown in each category.

Top Genes Predicted to Be Associated With Diabetes Mellitus

To discover novel genes associated with diabetes mellitus, we predicted all unlabeled genes in the PPI network using the final trained model. The model uses Node2vec (with p = 200 and q = 0.001) to extract node features in 512-dimension followed by a three-layer autoencoder to compress the feature to 64-dimension, and SVM is applied to predict the possibility of unlabeled genes to be a diabetes gene. The SVM model was trained using all the 476 genes labels as disease-related. Then all the unlabeled genes were predicted by SVM. We ranked the gene predicted by our methods and listed the top 15 genes in **Table 1**. The size of the top 15 genes is artificially set.

Researchers have delineated the relevance of some predicted genes to diabetes mellitus. Zhou et al. (2017), evaluated the gene-environment interactions and haplotype associations and extrapolated the pathogenic role of genetic variants in the TLR3-TRIF-TRAF3-INF-β in causing type 2 diabetes mellitus. Al Dubayee et al. (2021), examined the increased expression of BCL10 and reduced expression of caspase-7 from peripheral blood mononuclear cells of diabetic individuals during the apoptosis in insulin resistance, which reveals close relationship between BCL10 gene and diabetes mellitus. Maikel et al. (Colli et al., 2018), utilized immunofluorescence to discern the positive correlation between expression of PDL1 and IRF1, based on the fact that PDL1 expression is elevated in insulin-containing islets of individuals with type 1 diabetes, IRF1 and Diabetes Mellitus show a high probability of interaction.

Figure 5 shows the largest component of PPI subnetwork among these top-predicted genes and known genes associated with diabetes mellitus. Those predicted genes are closely connected with known diabetes genes in the database. For example, IRAK1 and IRAK4 have the highest degrees connecting both known genes and predicted genes. It has been shown that deletion of IRAK1 improves glucose tolerance by elevating insulin sensitivity (Sun et al., 2017). IRAK4 inhibitors can block MyD88 dependent signaling, which contributes to the pathogenesis of type I diabetes (Sabnis, 2021).

Functional Enrichment Analysis of the Predicted Genes

Gene set enrichment analysis has been performed for the top 15 genes predicted to be related to diabetes mellitus. Gene functional categories in Human Phenotype Ontology (HPO), KEGG, and GO biological process were used for over-representation analysis using WebGestaltR (Liao et al., 20192019). The top enrichment terms are shown in **Figure 6**. Our predicted genes have shown over-representation in genes of the HPO term "transient neonatal diabetes mellitus" with suggestive p-value < 0.01. The top HPO term enriched was "hepatic encephalopathy," and it has been shown that diabetes mellitus plays a role in hepatic encephalopathy by releasing and enhancing the inflammatory cytokines (Ampuero et al., 2013). In KEGG enrichment results, the term "NF-kappa B signaling pathway" achieves the best significance with p-value < $5*10^{-5}$. Romeo et al. (2002) has shown that diabetes and high glucose can induce the activation of nuclear factor-kB (NF-kappa B), which regulates a proapoptotic program in retinal pericytes. The second term is "Toll-like receptor signaling pathway diabetes" with enrichment p-value < $5*10^{-5}$. Dasu and Martin (2014) has shown the increased toll-like receptors (TLRs) expression and activation contribute to the hyper inflammation in human diabetic wounds. The third enriched term is "toxoplasmosis". There have been findings that patients with toxoplasmosis are more susceptible to be diabetics than those without toxoplasmosis, suggesting a role of toxoplasmosis in diabetes mellitus (Shirbazou et al., 2013). Most enriched terms in GO are related with the immune response. And it has been well established that patients with diabetes mellitus have more susceptibility to infections (Berbudi et al., 2020). The high blood glucose levels, as well as the inflammatory mediators produced by adipocytes and macrophages, can result in the immune response (Geerlings and Hoepelman, 1999).

CONCLUSION

Diabetes mellitus has widely affected the population in the world, without knowing the underlying mechanism. Discovering genes associated with diabetes will pave the way for developing novel efficient therapies. In this work, we designed a computational framework for diabetes gene prediction based on graph embedding techniques. This framework consists of three main steps: network feature extraction based on graph embedding methods; feature denoising and regeneration using stacked autoencoder; and disease-gene prediction based on machine learning classifiers. By comparing with different graph embedding methods and widely-used machine learning classifiers, we proved the efficiency and accuracy of our method. By applying this method to diabetes gene discovery, we found novel genes that have been reported in publications with clear association evidence but not recorded in the database. Through functional enrichment analysis based on Human Phenotype Ontology (HPO), KEGG, and GO biological process, we found the top predicted genes are enriched in multiple terms that have been proved to have a role in diabetes mellitus. Our computational method may also benefit gene discoveries for other complex diseases.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

JD and RY conceived the study, DL and XC designed and performed the experiments, XL wrote and revised the manuscript, JY supervised the study.

ACKNOWLEDGMENTS

We thank the reviewers for their suggestive questions. We thank the funding support from Wanshu. We also deeply thank Prof. Tao Wang from Northwestern Polytechnical University for providing the guidance and in-house pipelines. Besides, we would like to thank the contributors of databases, software used in our manuscript.

REFERENCES

Agrawal, M., Zitnik, M., and Leskovec, J. (2018). Large-scale Analysis of Disease Pathways in the Human Interactome. *PSB* 23, 111–122. doi:10.1142/9789813235533_0011:

Al Dubayee, M., Alshahrani, A., Aljada, D., Zahra, M., Alotaibi, A., Ababtain, I., et al. (2021). Gene Expression Profiling of Apoptotic Proteins in Circulating Peripheral Blood Mononuclear Cells in Type II Diabetes Mellitus and Modulation by Metformin. *Dmso* 14, 1129–1139. doi:10.2147/dmso.s300048

Ampuero, J., Ranchal, I., del Mar Díaz-Herrero, M., del Campo, J. A., Bautista, J. D., and Romero-Gómez, M. (2013). Role of Diabetes Mellitus on Hepatic Encephalopathy. *Metab. Brain Dis.* 28, 277–279. doi:10.1007/s11011-012-9354-2

Berbudi, A., Rahmadika, N., Tjahjadi, A. I., and Ruslami, R. (2020). Type 2 Diabetes and its Impact on the Immune System. *Cdr* 16, 442–449. 10 Data Availability Statement Publicly available datasets were analyzed in this study. doi:10.2174/1573399815666191024085838

Chang, C.-C., and Lin, C.-J. (2011). Libsvm. *ACM Trans. Intell. Syst. Technol.* 2, 1–27. doi:10.1145/1961189.1961199

Chen, Y., Wu, X., and Jiang, R. (2013). Integrating Human Omics Data to Prioritize Candidate Genes. *BMC Med. Genomics* 6, 57–12. doi:10.1186/1755-8794-6-57

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. Acids* 18, 590–604. doi:10.1016/j.omtn.2019.09.019

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., et al. (2018). IDF Diabetes Atlas: Global Estimates of Diabetes Prevalence for 2017 and Projections for 2045. *Diabetes Res. Clin. Pract.* 138, 271–281. doi:10.1016/j.diabres.2018.02.023

Colli, M. L., Hill, J. L. E., Marroquí, L., Chaffey, J., Dos Santos, R. S., Leete, P., et al. (2018). PDL1 Is Expressed in the Islets of People with Type 1 Diabetes and Is Up-Regulated by Interferons-α And-γ via IRF1 Induction. *EBioMedicine* 36, 367–375. doi:10.1016/j.ebiom.2018.09.040

Dasu, M. R., and Martin, S. J. (2014). Toll-like Receptor Expression and Signaling in Human Diabetic Wounds. *Wjd* 5, 219. doi:10.4239/wjd.v5.i2.219

Erten, S., Bebek, G., and Koyutürk, M. (2011). Vavien: an Algorithm for Prioritizing Candidate Disease Genes Based on Topological Similarity of Proteins in Interaction Networks. *J. Comput. Biol.* 18, 1561–1574. doi:10.1089/cmb.2011.0154

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C.-Y., Lopes-Ramos, C. M., et al. (2017). Exploring Regulation in Tissues with eQTL Networks. *Proc. Natl. Acad. Sci. USA* 114, E7841–E7850. doi:10.1073/pnas.1707375114

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS Era: from Association to Function. *Am. J. Hum. Genet.* 102, 717–730. doi:10.1016/j.ajhg.2018.04.002

Geerlings, S. E., and Hoepelman, A. I. M. (1999). Immune Dysfunction in Patients with Diabetes Mellitus (DM). *FEMS Immunol. \\& Med. Microbiol.* 26, 259–265. doi:10.1111/j.1574-695x.1999.tb01397.x

Ghiassian, S. D., Menche, J., and Barabási, A. L. (2015). A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *Plos Comput. Biol.* 11, e1004120. doi:10.1371/journal.pcbi.1004120

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable Feature Learning for Networks," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 855–864.

Han, X., Kong, Q., Liu, C., Cheng, L., and Han, J. (2021). SubtypeDrug: a Software Package for Prioritization of Candidate Cancer Subtype-specific Drugs. *Bioinformatics* 37, 2491–2493. doi:10.1093/bioinformatics/btab011

Kharroubi, A. T., and Darwish, H. M. (2015). Diabetes Mellitus: The Epidemic of the century. *Wjd* 6, 850. doi:10.4239/wjd.v6.i6.850

Li, Y., and Patra, J. C. (2010). Genome-wide Inferring Gene-Phenotype Relationship by Walking on the Heterogeneous Network. *Bioinformatics* 26, 1219–1224. doi:10.1093/bioinformatics/btq108

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and ZhangWebGestalt, B. (20192019). WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi:10.1093/nar/gkz401

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering Disease-Disease Relationships through the Incomplete Interactome. *Science* 347, 1257601. doi:10.1126/science.1257601

Natarajan, N., and Dhillon, I. S. (2014). Inductive Matrix Completion for Predicting Gene-Disease Associations. *Bioinformatics* 30, i60–i68. doi:10.1093/bioinformatics/btu269

Nitsch, D., Gonçalves, J. P., Ojeda, F., De Moor, B., and Moreau, Y. (2010). Candidate Gene Prioritization by Network Analysis of Differential Expression Using Machine Learning Approaches. *BMC Bioinformatics* 11, 1–16. doi:10.1186/1471-2105-11-460

Nyaga, D. M., Vickers, M. H., Jefferies, C., Perry, J. K., and O'Sullivan, J. M. (2018). Type 1 Diabetes Mellitus-Associated Genetic Variants Contribute to Overlapping Immune Regulatory Networks. *Front. Genet.* 9, 535. doi:10.3389/fgene.2018.00535

Peng, J., Guan, J., Hui, W., and Shang, X. (2021). A Novel Subnetwork Representation Learning Method for Uncovering Disease-Disease Relationships. *Methods* 192, 77–84. doi:10.1016/j.ymeth.2020.09.002

Peng, J., Guan, J., and Shang, X. (2019). Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder. *Front. Genet.* 10, 226. doi:10.3389/fgene.2019.00226

Peng, J., Han, L., and Shang, X. (2021). A Novel Method for Predicting Cell Abundance Based on Single-Cell RNA-Seq Data. *BMC Bioinformatics* 22, 1–15. doi:10.1186/s12859-021-04187-4

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A Learning-Based Framework for miRNA-Disease Association Identification Using Neural Networks. *Bioinformatics* 35, 4364–4371. doi:10.1093/bioinformatics/btz254

Peng, J., Lu, J., Shang, X., and Chen, J. (2017). Identifying Consistent Disease Subnetworks Using Dnet. *Methods* 131, 104–110. doi:10.1016/j.ymeth.2017.07.024

Peng, J., Wang, Y., Guan, J., Li, J., Han, R., Hao, J., et al. (2021). An End-To-End Heterogeneous Graph Representation Learning-Based Framework for Drug-Ttarget Interaction Prediction. *Brief. Bioinform.* 22, bbaa430. doi:10.1093/bib/bbaa430

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2021). Integrating Multi-Network Topology for Gene Function Prediction Using Deep Neural Networks. *Brief. Bioinform.* 22, 2096–2105. doi:10.1093/bib/bbaa036

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online Learning of Social Representations," in Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 701710.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET Knowledge Platform for Disease Genomics: 2019 Update. *Nucleic Acids Res.* 48, D845–D855. doi:10.1093/nar/gkz1021

Piñero, J., Bravo, A., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., et al. (2016). DisGeNET: a Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants. *Nucleic Acids Res.* 45, D833–D839. doi:10.1093/nar/gkw943

Piñero, J., Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., et al. (2015). DisGeNET: a Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. *Database* 2015. doi:10.1093/database/bav028

Ribeiro, L. F. R., Saverese, P. H. P., and Figueiredo, D. R. struc2vec. (2017). "Learning Node Representations from Structural Identity," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 385–394.

Romeo, G., Liu, W.-H., Asnaghi, V., Kern, T. S., and Lorenzi, M. (2002). Activation of Nuclear Factor- B Induced by Diabetes and High Glucose Regulates a Proapoptotic Program in Retinal Pericytes. *Diabetes* 51, 2241–2248. doi:10.2337/diabetes.51.7.2241

Sabnis, R. W. (2021). *Thienopyridinyl and Thiazolopyridinyl Compounds as IRAK4 Inhibitors*.

Shabalin, A. A. (2012). Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Shirbazou, S., Delpisheh, A., Mokhetari, R., and Tavakoli, G. (2013). Serologic Detection of Anti Toxoplasma Gondii Infection in Diabetic Patients. *Iran. Red Crescent Med. J.* 15, 701–703. doi:10.5812/ircmj.5303

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell* 180, 688–702. doi:10.1016/j.cell.2020.01.021

Sun, X.-J., Kim, S. P., Zhang, D., Sun, H., Cao, Q., Lu, X., et al. (2017). Deletion of Interleukin 1 Receptor-Associated Kinase 1 (Irak1) Improves Glucose Tolerance Primarily by Increasing Insulin Sensitivity in Skeletal Muscle. *J. Biol. Chem.* 292, 12339–12350. doi:10.1074/jbc.m117.779108

Tang, J., et al. (2015). "Line: Large-Scale Information Network Embedding," in Proceedings of the 24th international conference on world wide web, 1067–1077.

Tran, V. D., Sperduti, A., Backofen, R., and Costa, F. (2020). Heterogeneous Networks Integration for Disease-Gene Prioritization with Node Kernels. *Bioinformatics* 36, 2649–2656. doi:10.1093/bioinformatics/btaa008

van der Wijst, M., de Vries, D. H., Groot, H. E., Trynka, G., Hon, C. C., Bonder, M. J., et al. (2020). The Single-Cell eQTLGen Consortium. *Elife* 9. doi:10.7554/eLife.52155

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, e1000641. doi:10.1371/journal.pcbi.1000641

Visscher, P. M., and Goddard, M. E. (2019). From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* 211, 1125–1130. doi:10.1534/genetics.118.301594

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* 101, 5–22. doi:10.1016/j.ajhg.2017.06.005

Wang, T., Liu, Y., Ruan, J., Dong, X., Wang, Y., and Peng, J. (2021). A Pipeline for RNA-Seq Based eQTL Analysis with Automated Quality Control Procedures. *BMC Bioinformatics* 22, 403–418. doi:10.1186/s12859-021-04307-0

Wang, T., Peng, Q., Liu, B., Liu, X., Liu, Y., Peng, J., et al. (2019). eQTLMAPT: Fast and Accurate eQTL Mediation Analysis with Efficient Permutation Testing Approaches. *Front. Genet.* 10, 1309. doi:10.3389/fgene.2019.01309

Wang, T., Hua, Y., Xu, Z., and Yu, J. S. (2021). Enhancing Discoveries of Molecular QTL Studies with Small Sample Size Using Summary Statistic Imputation. *Brief. Bioinform.* 20, bbab370. doi:10.1093/bib/bbab370

Wang, T., Peng, J., Peng, Q., Wang, Y., and Chen, J. (2019). FSM: Fast and Scalable Network Motif Discovery for Exploring Higher-Order Network Organizations. *Methods* 173, 83–93. doi:10.1016/j.ymeth.2019.07.008

Wang, T., Peng, Q., Liu, B., Liu, Y., and Wang, Y. (2020). Disease Module Identification Based on Representation Learning of Complex Networks Integrated from GWAS, eQTL Summaries, and Human Interactome. *Front. Bioeng. Biotechnol.* 8, 418. doi:10.3389/fbioe.2020.00418

Wang, T., Ruan, J., Yin, Q., Dong, X., and Wang, Y. (2019). "An Automated Quality Control Pipeline for eQTL Analysis with RNA-Seq Data," in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1780–1786. doi:10.1109/bibm47256.2019.8983006

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic Identification of Trans eQTLs as Putative Drivers of Known Disease Associations. *Nat. Genet.* 45, 1238–1243. doi:10.1038/ng.2756

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Networks Learn. Syst.* 32, 4–24.

Xu, J., and Li, Y. (2006). Discovering Disease-Genes by Topological Features in Human Protein-Protein Interaction Network. *Bioinformatics* 22, 2800–2805. doi:10.1093/bioinformatics/btl467

Yang, P., Li, X., Wu, M., Kwoh, C.-K., and Ng, S.-K. (2011). Inferring Gene-Phenotype Associations via Global Protein Complex Network Propagation. *PLoS One* 6, e21502. doi:10.1371/journal.pone.0021502

Yang, W., Han, J., Ma, J., Feng, Y., Hou, Q., Wang, Z., et al. (2019). Prediction of Key Gene Function in Spinal Muscular Atrophy Using Guilt by Association Method Based on Network and Gene Ontology. *Exp. Ther. Med.* 17, 2561–2566. doi:10.3892/etm.2019.7216

Zeng, X., Ding, N., Rodríguez-Patón, A., and Zou, Q. (2017). Probability-based Collaborative Filtering Model for Predicting Gene-Disease Associations. *BMC Med. Genomics* 10, 76–53. doi:10.1186/s12920-017-0313-y

Zhou, Z., Zeng, C., Nie, L., Huang, S., Guo, C., Xiao, D., et al. (2017). The Effects of TLR3, TRIF and TRAF3 SNPs and Interactions with Environmental Factors on Type 2 Diabetes Mellitus and Vascular Complications in a Han Chinese Population. *Gene* 626, 41–47. doi:10.1016/j.gene.2017.05.011

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., and Powell, J. E. (2016). Analysis Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* 48, 481–487. doi:10.1038/ng.3538

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., et al. (2016). Integration of Summary Data from GWAS and eQTL Studies Predicts Complex Trait Gene Targets. *Nat. Genet.* 48, 481–487. doi:10.1038/ng.3538

Conflict of Interest: 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 December 2021
doi: 10.3389/fgene.2021.758131

Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment

Jiaogen Zhou[1†], *Wei Xiong*[2†], *Yang Wang*[3] *and Jihong Guan*[3*]

[1]*Jiangsu Provincial Engineering Research Center for Intelligent Monitoring and Ecological Management of Pond and Reservoir Water Environment, Huaiyin Normal University, Huian, China,* [2]*Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai, China,* [3]*Department of Computer Science and Technology, Tongji University, Shanghai, China*

Over the past decades, massive amounts of protein-protein interaction (PPI) data have been accumulated due to the advancement of high-throughput technologies, and but data quality issues (noise or incompleteness) of PPI have been still affecting protein function prediction accuracy based on PPI networks. Although two main strategies of *network reconstruction* and *edge enrichment* have been reported on the effectiveness of boosting the prediction performance in numerous literature studies, there still lack comparative studies of the performance differences between *network reconstruction* and *edge enrichment*. Inspired by the question, this study first uses three protein similarity metrics (local, global and sequence) for network reconstruction and edge enrichment in PPI networks, and then evaluates the performance differences of network reconstruction, edge enrichment and the original networks on two real PPI datasets. The experimental results demonstrate that edge enrichment work better than both network reconstruction and original networks. Moreover, for the edge enrichment of PPI networks, the sequence similarity outperformes both local and global similarity. In summary, our study can help biologists select suitable pre-processing schemes and achieve better protein function prediction for PPI networks.

Keywords: edge enrichment, network reconstruction, protein-protein interaction networks, protein function prediction, protein sequence annotation

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Cheng Liang,
Shandong Normal University, China
Yongjun Tang,
Central South University, China

****Correspondence:***
Jihong Guan
jhguan@tongji.edu.cn

†These authors have contributed equally to this work

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

Received: *13 August 2021*
Accepted: *11 November 2021*
Published: *14 December 2021*

Citation:
Zhou J, Xiong W, Wang Y and Guan J (2021) Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment. Front. Genet. 12:758131. doi: 10.3389/fgene.2021.758131

1 INTRODUCTION

Over the past decades, massive amounts of un-annotated protein sequence data have been accumulated with the advancement of high-throughput biological technologies. Due to high costs and time-consummation of experimental determining protein function annotation, the proportion of annotated proteins has been still relatively low (Sharan et al., 2007; Barrell et al., 2009). The increasing efforts have been made to predict protein functions.

As the best-known and early method of protein function prediction, homology-based prediction method indeed gave rise to a series of protein function prediction methods based on protein sequence or structural similarity (Sleator and Walsh, 2010). At the same time, the emerging of available protein databases, such as FATCAT (Ye and Godzik, 2004), PAST (Täubig et al., 2006) and PROCAT (Wallace et al., 1996), has further helped to improve the effectiveness of protein prediction. However, the low sequence similarity scores often occur when comparing target protein sequences with source protein sequences (Ofran et al., 2005), and thus this significantly reduces the effective application of homology-based prediction methods.

With the increasing amounts of the measured protein-protein interaction (PPI) data, more and more protein function prediction methods based on PPI networks are proposed and generally outperform the above homology-based prediction methods. In PPI networks, proteins and protein-protein interactions are represented by nodes and edges, respectively (Sharan et al., 2007; Chen et al., 2020; Wu et al., 2020; Waiho et al., 2021). Up to now, numerous algorithms have been used in protein function prediction based on PPI networks, such as edge-betweenness clustering (Dunn et al., 2005), Graphlet-based edge clustering (Solava et al., 2012), clique percolation (Adamcsek et al., 2006), GRAAL (Kuchaiev et al., 2010), hybrid-property based method (Hu et al., 2011), and IsoRank (Singh et al., 2008). Moreover, advanced machine learning and deep learning techniques have also been used for protein function prediction, including collective classification (Xiong et al., 2013; Wu et al., 2014), active learning (Xiong et al., 2014), DeepInteract (Sunil et al., 2017), ConvsPPIS (Zhu et al., 2020), PhosIDN (Yang et al., 2021) and WinBinVec (Abdollahi et al., 2021), etc.

The above methods mainly use existing PPI data. However, current PPI data mainly generated by high-throughput or TAP-MS techniques (Berggard et al., 2007), are often in presence of noise and incompleteness, and this unavoidably causes adverse effects on the prediction performance. Two main methods of *network reconstruction* and *edge enrichment* are proposed to effectively boost the prediction performance. Different strategies are used for network reconstruction or edge enrichment. For example, Bogdanov and Singh (2010) presented a network reconstruction approach by extracting functional neighborhood features using random walk with restart. Chua et al. (2007) used weighting strategies to enrich PPI networks, and adopted a local prediction method to predict the functions of un-annotated proteins. Xiong et al. (2013) applied collective classification to PPI networks with enriched edges to predict protein functions.

Although the above two types of approaches achieve promising performance improvements, there still lack comparative studies of the performance differences between network reconstruction and edge enrichment. We do not still know which one is better in performance, or specifically, which one should be applied for different situations. Inspired by the question, we conducte a comprehensive comparison of two network transformation of network reconstruction and edge enrichment for boosting the performance of PPI network-based protein functional annotation. Concretely, we first use three different protein similarity metrics for network reconstruction and edge enrichment of PPI networks, and then evaluate the performance differences between the two transformed networks (network reconstruction and edge enrichment) and original networks on two real PPI datasets. The results of experiments demonstrate that edge enrichment work better than both network reconstruction and original networks. Moreover, for the edge enrichment of PPI networks, the sequence similarity outperformes both local and global similarity. More detailed work will be presented in later sections.

2 MATERIALS AND METHODS

2.1 Similarity Metrics

As we point out above, the noise and incompleteness of PPI network data adversely affects the performance of protein functional annotation. Network reconstruction and edge enrichment are major approaches to improve PPI data quality. In this work, we carry out comparison study on these two approaches by reconstructing and enriching original networks using various protein similarity metrics, including sequence similarity, local similarity and global similarity. In what follows, we describe and discuss these similarity measures in detail.

2.1.1 Protein Sequence Similarity

BLAST method (Altschul et al., 1997) is used to measure the similarity between any two proteins in this study. The similarity of a given protein V_x with other proteins is defined as

$$S(V_x) = [S_{x,1}, S_{x,2}, \ldots, S_{x,i}, \ldots, S_{x,n}] \tag{1}$$

where $S_{x,i}$ is the similarity score between the pair of proteins V_x and V_i. Due to ignoring self-similarity, $S_{x,i} = 0$ is set when $x = i$.

2.1.2 Local Similarity Indices

We consider three kinds of local similarity indices, including *Common Neighbors* (CN), *Jaccard Index* and Functional Similarity (FS).

Common Neighbors. Given nodes u and v, their neighboring sets are N_u and N_v, respectively. The CN is defined as the neighborhood overlap of the nodes (Newman, 2001). The more identical neighbors two nodes have, the higher the CN value is. The measure of CN is as follows:

$$S_{CN}(u, v) = |N_u \cap N_v| \tag{2}$$

Jaccard Index. Given nodes u and v and their corresponding neighboring sets of N_u and N_v, Jaccard index is used to measure the similarity between the N_u and N_v sets, and it is calculated as:

$$S_{Jaccard}(u, v) = \frac{|N_u \cap N_v|}{|N_u \cup N_v|} \tag{3}$$

Functional Similarity (FS). For a PPI network, FS index was first used to measure the similarity of any pair of proteins (Chua et al., 2006), and it is defined as follows:

$$S_{FS}(u, v) = \frac{2|N_u \cap N_v|}{|N_u - N_v| + 2|N_u \cap N_v| + \lambda_{u,v}} \times \frac{2|N_u \cap N_v|}{|N_v - N_u| + 2|N_u \cap N_v| + \lambda_{v,u}} \tag{4}$$

where $\lambda_{u,v} = \max(0, n_{avg} - (|N_u - N_v|) + |N_u \cap N_v|))$, and by using the $\lambda_{u,v}$ factor, similarity weights between protein pairs are penalized when their common neighbors are too few. n_{avg} is the average number of close neighbors that each node has in the network. In a weighted PPI network, the labeled weights of edges mean interaction confidences between pairs of proteins. Thus, we can modify the FS index to take into account the confidence of each interaction. The extended FS index for weighted PPI networks, named FS.R, is defined as follows:

$$S_{FS.R}(u,v) = \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w} r_{v,w}}{\sum_{w\in N_u} r_{u,w} + \sum_{w\in(N_u\cap N_v)} r_{u,w} r_{v,w} + \lambda_{u,v}} \times \frac{2\sum_{w\in(N_u\cap N_v)} r_{u,w} r_{v,w}}{\sum_{w\in N_v} r_{v,w} + \sum_{w\in(N_u\cap N_v)} r_{u,w} r_{v,w} + \lambda_{v,u}}. \quad (5)$$

2.1.3 Global Similarity Indices

Two global similarity indices are considered in this paper, they are Katz index and random walk with restart.

Katz Index. This index is proposed by Lü and Zhou (2011). It sums the set of paths directly and deals with the paths by length so that the shorter paths get more weights. Formally,

$$S_{Katz}(u,v) = \sum_{L=1}^{\infty} \beta^L \cdot \left|paths_{uv}^{<L>}\right| = \beta A_{uv} + \beta^2 (A^2)_{uv} + \beta^3 (A^3)_{uv} + \ldots \quad (6)$$

where $paths_{uv}^{<L>}$ is the set of the paths, which connect the nodes of u and v with a path length of L. The parameter of β controls the path weights.

Random Walk with Restart (RWR). Tong et al. (2008) used RWR index to measure the relevance score between node j and node i in a PPI network. Given the adjacency matrix $W_{n,n}$ of a PPI network, a random walker transmits from the starting node i to one of its neighbors at random with probability c, and returns to the node i with the probability $1 - c$. Finally, the walker will stay stably at node j with probability $R_{i,j}$. The steady-state probability $R_{i,j}$ is defined as RWR index. We have

$$\overrightarrow{R_i} = c\tilde{W}^T \overrightarrow{R_i} + (1-c)\overrightarrow{e_i} \quad (7)$$

where $\overrightarrow{e_i}$ is the starting vector, the ith element is 1 and the other elements are 0. $\tilde{W}$ is a weighted matrix. For an unweighted network, $\tilde{W}_{ij} = 1/m$ (where m is the number of neighbors that node i has) if i and j are connected, and $\tilde{W}_{ij} = 0$ otherwise. For a weighted network,

$$\begin{cases} \tilde{W}_{ij} = W_{ij} / \sum_{j=1}^{n} W_{ij}, & if\ i\ and\ j\ are\ connected. \\ \tilde{W}_{ij} = 0, & otherwise. \end{cases} \quad (8)$$

2.2 Network Reconstruction and Edge Enrichment

Network reconstruction is carried out as follows: First, the similarity scores between protein pairs in the original PPI network are calculated according to the above similarity indexes. Next, some interactions are selected to reconstruct the PPI network based on the similarity scores. As in Liben-Nowell and Kleinberg (2007), an appropriate score threshold is used such that the number of protein pairs with higher scores than the threshold is as same as possible to the interaction number of the original network. Then, a new network is formed by using the protein pairs with higher scores over the threshold. However, this approach may lead to absence of some proteins in the new network. Alternatively, for any node N_i in the original network, we first remove all its interactions. We find the top k neighbors most similar to the node N_i. Then, the k edges from the node N_i to its top k neighbors are created, and their similarity scores are used as edge weights in the new network. Thus, we have

$$S(N_i)_k = [S_{i,1}, S_{i,2}, \ldots, S_{i,k}]. \quad (9)$$

Edge enrichment is also performed in two steps as in network reconstruction, the only difference is that all interactions in the original network are preserved. An enriched network has two types of edges: *explicit edges* (old edges) and similarity-inferred edges (new edges). Here, there are two questions to be addressed: One is how to combine the edge weights with different semantics, and another is how many edges are added for each protein, that is, how to optimize the parameter k (see **Eq. 9**). The questions will be discussed in the following sections.

2.3 Protein Function Prediction Approaches

In this study, protein function predictions on two real PPI datasets are performed using two different approaches.The first one is majority method, which is a local neighbor counting approach (Schwikowski et al., 2000). The second is a global protein function prediction approach, which is common called *collective classification* (Xiong et al., 2013). Details of this approach are presented in the following subsections.

2.4 Gibbs Sampling Based Collective Classification

Gibbs sampling (GS) includes two main processes of *bootstrapping* and *iterative classification* (Sen et al., 2008). The pseudo-code is illustrated below.

ALGORITHM 1 | Gibbs sampling

```
for each query protein V_x do
    compute the initial a_x^→ using 𝒳 ∩ 𝒩_x^w (original network), 𝒳 ∩ 𝒩_x^s (reconstructed network), 𝒳 ∩ 𝒩_x^w
    and 𝒳 ∩ 𝒩_x^s (enriched network).
end for
for each query protein V_x do
    update a_x^→ using current assignments to 𝒩_x^w (original network), 𝒩_x^s (reconstructed network), 𝒩_x^w
    and 𝒩_x^s (enriched network).
end for
for each query protein V_x do
    update a_x^→ using current assignments to 𝒩_x^w (original network), 𝒩_x^s (reconstructed network), 𝒩_x^w
    and 𝒩_x^s (enriched network).
    create b_xi^→ to record the m-rank result
end for
for each query protein V_x do
    calculate the final result c_x^→ based on matrix M_x
end for
```

2.4.1 Bootstrapping

The closer the proteins to each other, the more similar their functions become in a PPI network. For an unannotated protein, its probability distribution is estimated using a weighted voting method. In the original or reconstructed network, there is only one kind of annotated neighbors to vote. An unannotated protein V_x has the corresponding explicit neighbors of N_x or k similarity-inferred neighbors. For the above neighbor sets, we have their edge weights as follows:

$$\mathcal{N}_x^w = [w_{x1}, w_{x2}, \ldots, w_{xi}, \ldots, w_{xN_x}] \quad \mathcal{N}_x^s = [S_{x,1}, S_{x,2}, \ldots, S_{x,i}, \ldots, S_{x,k}] \quad (10)$$

The probability of V_x having the jth function F_j (V_xF_j) is calculated as follows:

$$P_x^j = \frac{1}{Z_x^w}\sum_{i=1}^{N_x} w_{x,i} f_{i,j} \quad P_x^j = \frac{1}{Z_x^s}\sum_{i=1}^{k} S_{x,i} f_{i,j} \tag{11}$$

where Z_x^w and Z_x^s are the normalizers:

$$Z_x^w = \sum_{j=1}^{m}\sum_{i=1}^{N_x} w_{x,i} f_{i,j} \quad Z_x^s = \sum_{j=1}^{m}\sum_{i=1}^{k} S_{x,i} f_{i,j} \tag{12}$$

However, in the enriched network, there are both old (explicit) and new (similarity-inferred) neighbors which need to be voted. So, the parameter $\lambda \in (0, 1)$ is used to combine the two types of different neighbors. Given a query protein V_x, the V_xF_j probability is calculated as follows:

$$P_x^j = \lambda \frac{1}{Z_x^w}\sum_{i=1}^{N_x} w_{x,i} f_{i,j} + (1-\lambda)\frac{1}{Z_x^s}\sum_{i=1}^{k} S_{x,i} f_{i,j} \tag{13}$$

A higher P_x^j value indicates a higher probability that protein V_x is more likely to have jth function F_j. The V_xF_j probability distribution is represented as:

$$\overrightarrow{a_x} = [P_x^1, P_x^2, \ldots, P_x^m] \tag{14}$$

2.4.2 Iterative Classification

Iterative classification has two main steps of burn-in and sampling. In burn-in period, iteration number is fixed, and $\overrightarrow{a_x}$ is updated in each iteration. In sampling period, we update $\overrightarrow{a_x}$ in each iteration, and also count how many times the jth function F_j for protein V_x are sampled. Considering each protein with one or more functions, therefore, we define the most likely function of the protein V_x as follow:

$$b_x^j = argmax_{j\in[1,m]} P_x^j \tag{15}$$

where b_x^j represents the jth most likely function of the protein V_x, that is the jth-rank result. We further use $\overrightarrow{b_{xi}}$ vector to record all ranking results in the ith iteration.

$$\overrightarrow{b_{xi}} = [b_{xi}^1, b_{xi}^2, \ldots, b_{xi}^m]. \tag{16}$$

The matrix M_x with s rows and m columns is produced after running the predetermined s number of iterations.

$$M_x = [\overrightarrow{b_{x1}}, \overrightarrow{b_{x2}}, \ldots, \overrightarrow{b_{xs}}]^T. \tag{17}$$

Finally, we obtain the required m-dimensional vector $\overrightarrow{c_x}$ for query protein V_x:

$$\overrightarrow{c_x} = [c_x^1, c_x^2, \ldots, c_x^m]. \tag{18}$$

where c_x^1 is the first ranked prediction in the ith column of M_x.

3 RESULTS AND DISCUSSION

3.1 Data Preprocessing and Experimental Workflow

The two PPI datasets of A and B are used in our study. The datasets A and B are downloaded from the databases of BioGRID (Stark et al., 2011) and STRING (Szklarczyk et al., 2011), respectively. The datasets A and B are annotated as in Ashburner et al. (2000). The datasets in this study are based on Gene Ontology (GO) annotation. GO annotations consist of three basic namespaces: molecular function, biological process and cellular component. We construct one protein interaction network for each GO namespace using only physical interactions.Therefore, there are totally six PPI networks (three for *S.cerevisiae* and the other three for *M.musculus*) in Dataset A. For Dataset B, we construct two PPI networks (one for *S.cerevisiae* and another for *M.musculus*).More detailed information was listed in the supplementary material (**Supplementary Table S1**).

The comparison of the function prediction performance on the reconstructed and enriched networks with that on the original networks is first performed using the cross validation of leave-one-out method (LOOM). LOOM takes each protein in turn as a query protein, and carries out function prediction with the remaining proteins in the network. As the bootstrapping in *Gibbs sampling* based collective classification does not result in updating of the query protein, therefore we use the *majority* method to predict protein functions in LOOM cross validation. Then, the annotated protein proportion is changed from 10% to 90%, and the average performance of 10 experiments is reported for each of all proportions. The *majority* method is not suitable in this setting because it is a local neighbor counting approach and

TABLE 1 | Comparison of performance differences between similarity indices (Dataset A: *M.musculus*).

Indices	Molecular function			Biological process				Cellular component		
	1st rank	2nd rank	3rd rank	1st rank	2nd rank	3rd rank	4th rank	1st rank	2nd rank	3rd rank
Origin	0.28	0.12	0.10	0.39	0.23	0.13	0.09	1.63	0.45	0.24
CN	0.21	0.09	0.07	0.27	0.22	0.14	0.09	1.44	0.47	0.16
Jaccard	0.30	**0.16**	0.11	**0.49**	**0.30**	0.129	0.11	1.94	0.56	0.25
FS	**0.33**	0.15	**0.15**	0.47	0.28	**0.16**	**0.12**	**2.13**	**0.61**	**0.27**
CN+	0.27	0.14	0.10	0.37	0.26	0.14	0.11	1.70	0.54	0.21
Jaccard+	0.35	0.16	0.12	**0.54**	**0.34**	0.15	0.12	2.03	0.62	0.27
FS+	**0.38**	**0.16**	**0.15**	0.52	0.30	**0.16**	**0.14**	**2.23**	**0.69**	**0.29**
Katz	0.29	0.13	0.12	0.45	0.23	**0.17**	0.11	1.70	0.54	0.28
RWR	**0.32**	**0.15**	**0.13**	**0.49**	**0.26**	0.16	**0.12**	**2.23**	**0.61**	**0.30**
Katz+	0.31	**0.16**	0.14	0.47	0.26	**0.19**	**0.14**	2.13	0.59	0.27
RWR+	**0.35**	0.15	**0.16**	**0.52**	**0.28**	0.17	0.12	**2.45**	**0.64**	**0.33**

TABLE 2 | Comparison of performance differences between similarity indices (Dataset B).

Indices	*S.cerevisiae*			*M.musculus*		
	1st rank	2nd rank	3rd rank	1st rank	2nd rank	3rd rank
Origin	2.23	0.75	0.49	1.94	1.49	0.82
CN	1.50	0.54	0.29	1.28	0.69	0.43
Jaccard	1.55	0.62	0.39	1.51	1.13	**0.79**
FS	**1.70**	**0.64**	**0.41**	**1.56**	**1.22**	0.75
CN+	1.85	0.67	0.43	1.63	1.27	0.72
Jaccard+	1.95	0.65	**0.49**	1.78	1.33	0.78
FS+	**2.13**	**0.72**	0.47	**1.92**	**1.51**	**0.81**
Katz	1.63	0.62	0.41	1.70	**1.33**	0.75
RWR	**1.78**	**0.67**	**0.43**	**1.78**	1.27	**0.79**
Katz+	1.86	0.64	0.47	1.86	**1.51**	0.79
RWR+	**2.23**	**0.75**	**0.52**	**2.03**	1.49	**0.85**

TABLE 3 | The influence of the parameter of *k* (*M.musculus* in Dataset A).

Indices	Molecular function			Biological process				Cellular component		
	1st rank	2nd rank	3rd rank	1st rank	2nd rank	3rd rank	4th rank	1st rank	2nd rank	3rd rank
Origin	0.28	0.12	0.10	0.39	0.23	0.13	0.09	1.63	0.45	0.24
BLAST 1	0.34	0.19	0.09	0.30	0.16	0.08	0.04	0.79	0.34	0.13
BLAST 5	0.43	0.26	0.13	**0.45**	**0.22**	0.12	0.08	**0.98**	**0.38**	0.18
BLAST 10	**0.45**	**0.27**	0.11	0.43	0.18	**0.13**	0.09	0.96	0.35	0.17
BLAST 15	0.41	0.21	**0.13**	0.42	0.20	0.11	**0.09**	0.92	0.33	**0.19**
BLAST+1	0.39	0.24	0.15	0.47	0.26	0.15	0.12	1.71	0.49	0.27
BLAST+5	0.47	**0.29**	**0.23**	**0.56**	0.30	**0.18**	**0.14**	**2.02**	**0.67**	0.32
BLAST+10	**0.49**	0.24	0.21	0.54	0.32	0.14	0.11	1.94	0.58	0.29
BLAST+15	0.46	0.27	0.20	0.49	**0.34**	0.15	0.12	1.86	0.62	**0.33**
FS 10	0.30	0.14	0.12	0.42	0.24	0.14	0.09	1.71	0.54	0.23
FS 30	0.33	0.15	**0.15**	**0.47**	0.28	0.16	0.12	**2.13**	0.61	0.27
FS 50	**0.35**	0.16	0.17	0.46	**0.30**	**0.18**	0.14	2.04	**0.64**	**0.28**
FS 100	0.32	**0.18**	0.12	0.45	0.27	0.17	**0.15**	1.95	0.57	0.26
FS+10	0.32	0.14	0.12	0.26	0.14	0.14	0.10	1.95	0.58	0.25
FS+30	0.39	0.16	0.15	0.52	**0.30**	**0.16**	0.14	**2.23**	**0.70**	**0.30**
FS+50	**0.41**	0.15	0.11	**0.54**	0.24	0.14	**0.14**	2.21	0.67	0.25
FS+100	0.38	**0.18**	**0.16**	0.50	0.27	0.16	0.13	2.07	0.64	0.27
RWR 10	0.25	0.13	0.11	0.41	0.21	0.14	0.09	1.86	0.54	0.24
RWR 30	**0.32**	0.15	0.13	**0.49**	**0.26**	0.16	0.11	2.23	**0.61**	**0.30**
RWR 50	0.31	**0.16**	0.11	0.47	0.21	**0.17**	0.12	**2.33**	0.58	0.27
RWR 100	0.29	0.15	**0.15**	0.44	0.22	0.16	**0.14**	2.12	0.55	0.30
RWR+10	0.29	0.14	0.13	0.47	0.23	0.15	0.12	2.13	0.57	0.29
RWR+30	**0.35**	0.15	**0.16**	**0.52**	**0.28**	**0.18**	0.12	**2.45**	**0.64**	0.33
RWR+50	0.34	**0.16**	0.15	0.49	0.28	0.14	**0.13**	2.36	0.62	0.32
RWR+100	0.31	0.15	0.15	0.46	0.25	0.16	0.12	2.23	0.58	**0.36**

does not work well in sparsely-labeled network. Thus, the *Gibbs sampling* based collective classification is used to predict protein functions. The main hardware configuration of an Inter dual-core processor (3 GHz) and 16GB RAM, with a Linux operating system, and Python 3.0 is as the programming environment for running the algorithms.

Finally, as in Bogdanov and Singh (2010), the ratio of the number of *true positive* (TP) predictions to the number of *false positive*predictions (FP) is produced in the cross validation, i.e. TP/FP is used to assess prediction accuracy of PPI networks. We define the overall *i*th rank *true positive* (TP) as the number of proteins whose *i*th rank predicted function c_x^i is one of the true functions of protein V_x, and the overall *i*th rank *false positive* (FP) as the number of proteins whose *i*th rank predicted function c_x^i is not one of the true functions of protein V_x.

3.2 Similarity Index Selection and the Effect of the Parameters *k* and λ

In this study, in addition to sequence similarity, the PPI networks are reconstructed and enriched by using three local similarity indices (CN, Jaccard and FS)and two global similarity indices (Katz and RWR). In order to choose the best ones for the following experiments, the performance differences between the five similarity indices are evaluated over the two datasets of A and B. The experimental results over the dataset A are presented in

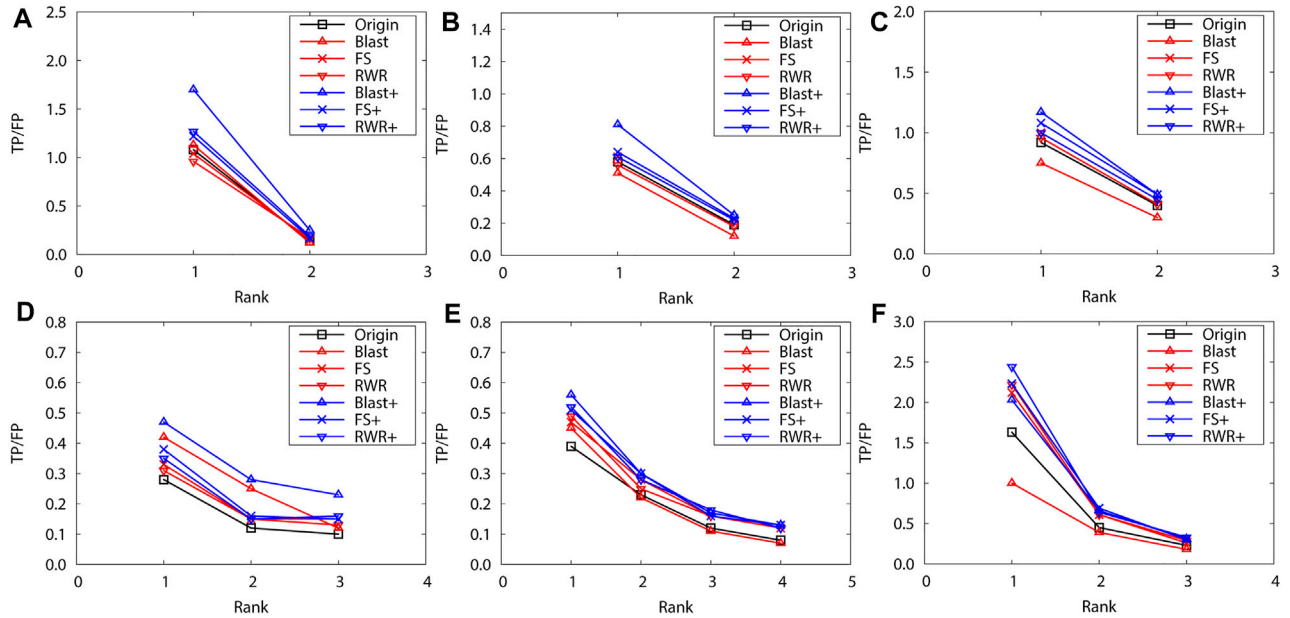

FIGURE 1 | The performance evaluation by leave-one-out validation over the PPI networks (Dataset A: *S.cerevisiae* and *M.musculus*). Here, the sub figures in the horizontal and vertical directions represent the experimental results for the PPI networks of different data sets and function types, respectively. Horizontally, the top three subplots represent ones on *S.cerevisiae*, and the bottom for ones on *M.musculus*. **(A)** and **(D)** Molecular function, **(B)** and **(E)** Biological process, **(C)** and **(F)** Cellular component.

TABLE 4 | The effect of the parameter *k* (Dataset B).

Indices	*S.cerevisiae*			*M.musculus*		
	1st rank	**2nd rank**	**3rd rank**	**1st rank**	**2nd rank**	**3rd rank**
Origin	2.23	0.75	0.49	1.94	1.49	0.82
BLAST 1	0.96	0.37	0.17	1.28	0.59	0.35
BLAST 5	1.18	**0.43**	**0.28**	1.63	**0.75**	0.45
BLAST 10	**1.21**	0.39	0.24	**1.70**	0.72	0.41
BLAST 15	1.15	0.42	0.26	1.56	0.70	**0.50**
BLAST+1	2.11	0.64	0.45	2.15	1.51	0.82
BLAST+5	**2.83**	**0.82**	0.64	**2.45**	**1.63**	**0.87**
BLAST+10	2.57	0.75	**0.65**	2.33	1.57	0.85
BLAST+15	2.40	0.69	0.62	2.28	1.49	0.76
FS 10	1.53	0.55	0.38	1.33	1.06	0.68
FS 30	1.72	**0.64**	**0.41**	1.56	**1.22**	0.75
FS 50	**1.75**	0.57	0.38	1.64	1.19	**0.79**
FS 100	1.63	0.61	0.37	**1.68**	1.18	0.73
FS+10	1.93	0.65	0.40	1.85	0.42	0.79
FS+30	**2.13**	**0.72**	0.47	**1.92**	**1.51**	**0.81**
FS+50	2.05	0.70	0.44	1.83	1.40	0.76
FS+100	1.90	0.67	**0.49**	1.92	1.45	0.78
RWR 10	1.50	0.57	0.36	1.57	1.08	0.69
RWR 30	**1.78**	**0.67**	0.43	**1.78**	1.27	0.79
RWR 50	1.72	0.63	0.40	1.69	**1.31**	0.74
RWR 100	1.70	0.61	**0.45**	1.64	1.17	**0.82**
RWR+10	2.00	0.70	0.46	1.88	1.40	0.69
RWR+30	**2.23**	0.75	**0.52**	**2.03**	**1.49**	**0.85**
RWR+50	2.11	0.72	0.49	1.94	1.43	0.82
RWR+100	1.94	**0.81**	0.48	1.82	1.45	0.75

TABLE 5 | The influence of the parameter λ (*M.musculus* in Dataset A).

Indices	Molecular function			Biological process				Cellular component		
	1st rank	2nd rank	3rd rank	1st rank	2nd rank	3rd rank	4th rank	1st rank	2nd rank	3rd rank
Origin	0.28	0.12	0.10	0.39	0.23	0.13	0.09	1.63	0.45	0.24
BLAST+0.1	0.38	0.21	0.11	0.43	0.24	0.12	0.13	1.52	0.41	0.20
BLAST+0.3	0.40	0.25	0.18	0.49	0.29	**0.17**	0.11	1.65	0.54	0.25
BLAST+0.5	0.44	**0.30**	0.16	0.537	0.27	0.15	**0.15**	1.85	0.62	0.28
BLAST+0.7	**0.47**	0.29	**0.23**	**0.56**	**0.32**	0.16	0.14	**2.02**	**0.67**	**0.34**
BLAST+0.9	0.33	0.16	0.15	0.42	0.23	0.13	0.10	1.76	0.55	0.27
FS+0.1	0.31	0.14	0.12	0.49	0.24	0.15	0.13	1.94	0.59	0.27
FS+0.3	0.35	0.16	**0.16**	**0.53**	**0.33**	0.13	0.10	1.86	0.68	0.25
FS+0.5	0.37	0.15	0.13	0.49	0.31	**0.18**	0.11	2.04	0.63	0.28
FS+0.7	**0.39**	**0.16**	0.15	0.52	0.30	0.16	**0.14**	**2.23**	**0.70**	**0.30**
FS+0.9	0.30	0.13	0.10	0.42	0.222	0.14	0.11	1.86	0.57	0.26
RWR+0.1	0.30	0.13	0.12	0.47	0.29	0.16	0.12	2.12	0.59	0.28
RWR+0.3	0.33	0.15	0.14	0.50	**0.31**	0.11	0.08	2.22	0.64	0.32
RWR+0.5	0.35	0.13	**0.17**	0.50	0.24	0.16	0.10	2.32	**0.74**	0.30
RWR+0.7	**0.37**	**0.17**	0.14	**0.52**	0.28	**0.18**	**0.12**	**2.45**	0.64	**0.33**
RWR+0.9	0.30	0.13	0.10	0.43	0.26	0.14	0.10	1.94	0.57	0.27

TABLE 6 | The influence of the parameter λ (Dataset B).

Indices	*S.cerevisiae*			*M.musculus*		
	1st rank	2nd rank	3rd rank	1st rank	2nd rank	3rd rank
Origin	2.23	0.75	0.49	1.94	1.49	0.82
BLAST+0.1	1.56	0.63	0.41	1.76	1.28	0.72
BLAST+0.3	1.89	0.70	0.58	2.01	1.44	0.78
BLAST+0.5	2.56	0.75	**0.66**	2.34	1.37	0.82
BLAST+0.7	**2.83**	**0.82**	0.64	**2.45**	**1.63**	**0.87**
BLAST+0.9	2.36	0.79	0.56	2.12	1.51	0.85
FS+0.1	1.86	0.66	0.42	1.70	1.33	0.74
FS+0.3	1.93	0.64	0.45	1.86	1.38	0.81
FS+0.5	2.06	0.69	0.43	**2.02**	1.44	**0.87**
FS+0.7	**2.13**	0.72	**0.47**	1.92	1.51	0.84
FS+0.9	1.99	**0.75**	0.41	1.88	**1.62**	0.79
RWR+0.1	1.82	0.62	0.42	1.65	1.38	0.77
RWR+0.3	1.94	0.65	0.48	1.83	1.44	**0.83**
RWR+0.5	2.02	0.71	**0.54**	1.95	1.46	0.73
RWR+0.7	**2.23**	**0.75**	0.52	**2.03**	**1.49**	0.82
RWR+0.9	2.12	0.69	0.47	1.92	1.43	0.77

Supplementary Table S3 and **Table 1**, and ones over the Dataset B listed in **Table 2**. Using FS as the local similarity index and RWR as the global similarity index generally achieve the best performance. Hence, FS and RWR are selected as the local similarity index and global similarity index, respectively in the following experiments.

The effect of two parameters on the performance of network reconstruction and edge enrichment are also examined in our study. The first one is the number of similarity-inferred edges k. The prediction performance on the Datasets of A and B is listed in **Supplementary Table S4**, **Table 3**, and **Table 4**, with the varying values of k. For both the datasets A and B, experimental results show that BLAST roughly achieves the best performance by setting $k = 5$. When the values of $k = \{10, 30, 50, 100\}$ are used for FS and RWR, using $k = 30$ or $k = 50$ generally works best in most cases, and the overall performance is relatively robust for the reconstructed or enriched networks. Hence, in the following experiments, the parameter value of k is used as 5, 30, 30 for BLAST, FS and RWR, respectively.

The second parameter λ dominates the tradeoff between explicit edges and similarity-inferred edges. Further, the effect of the parameter λ is evaluated on the prediction performance when it varies from 0.1 to 0.9. The results on the Dataset A are listed in Supplementary material (see **Supplementary Table S5**) and **Table 5**, and ones on the Dataset B in **Table 6**, respectively. Generally, the λ value has a relatively small impact on prediction accuracy, unless it is too large or too small. In the following experiments, the λ value is set uniformly at 0.7.

3.3 Performance Evaluation on Dataset A

The performance comparison of reconstructed and enriched networks with that of the original networks is first carried out by

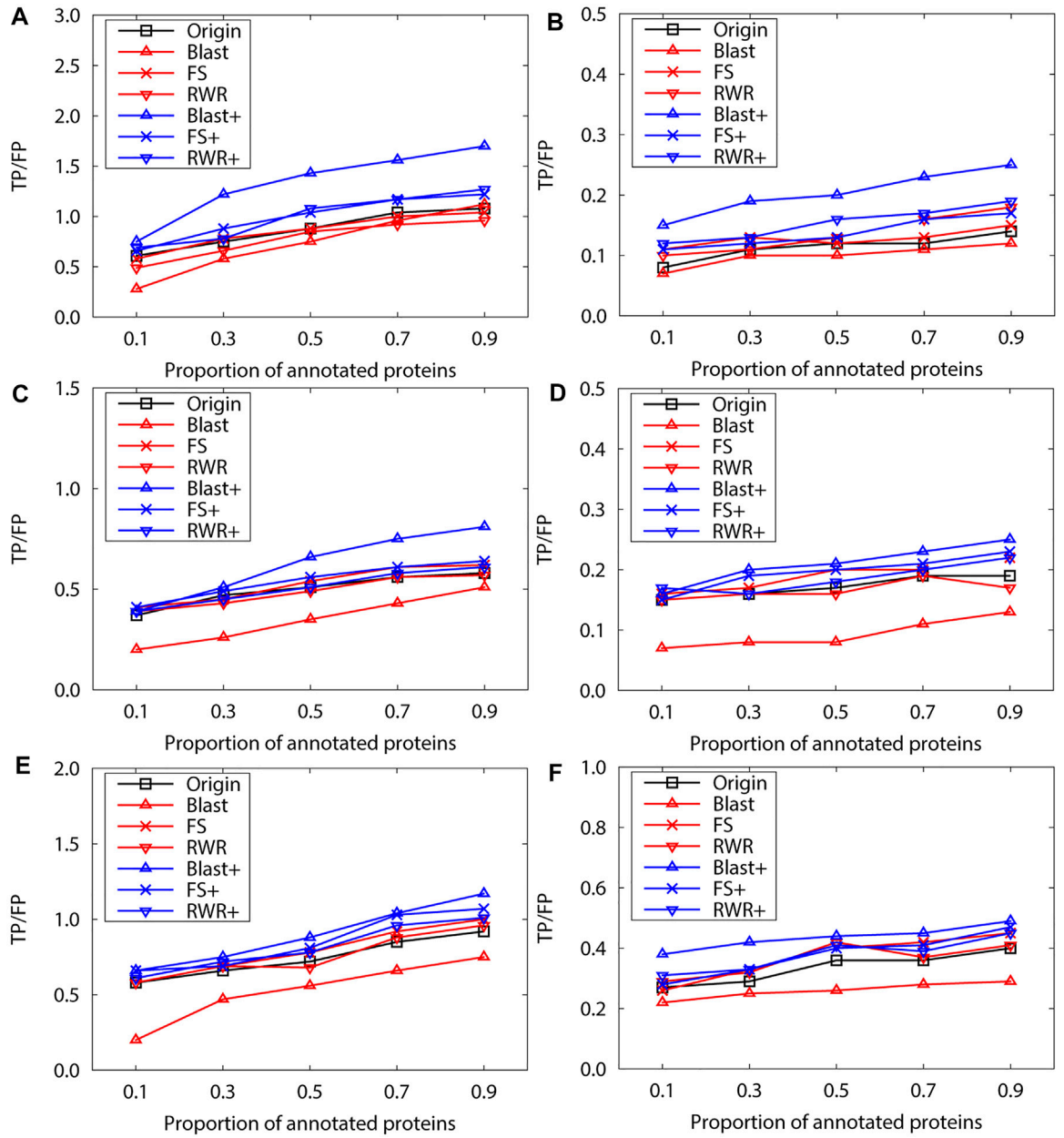

FIGURE 2 | The performance evaluation over the sparsely-labeled networks (Dataset A: *S.cerevisiae*). Here, the sub figures in the horizontal and vertical directions represent the experimental results for the PPI networks of different function types and rank predicted functions, respectively. Horizontally, the top two subplots represent ones of molecular function, the middle for ones of biological process, and the bottom for ones of cellular component **(A)**, **(C)** and **(E)** first rank predicted function, **(B)**, **(D)** and **(F)** second rank predicted function.

leave-one-out validation. The top protein function prediction is selected according to the average number of useful functions per protein in the PPI networks. Therefore, only the top 2 predictions are performed on the PPI networks of *S.cerevisiae* in the Dataset A, and the top 3 or 4 predictions are examined for *M.musculus* in Dataset A.

Obviously, edge enrichment gains more accurate predictions than network reconstruction and original networks, due to the combination of explicit and implicit (similarity-inferred) edges (**Figure 1**). The results clearly indicate that edge enrichment indeed gains better prediction performance by adding similarity-inferred edges to PPI networks. BLAST-enriched networks always worke best, while BLAST-reconstructed networks always work worst. This is because BLAST-inferred edges are based on protein sequence information that is short in the original networks. The useful information in the original network greatly increases by adding BLAST-inferred edges, and consequently boosts prediction accuracy. However, in the reconstructed networks, the original PPI edges are put aside first, BLAST-reconstructed networks contain only protein

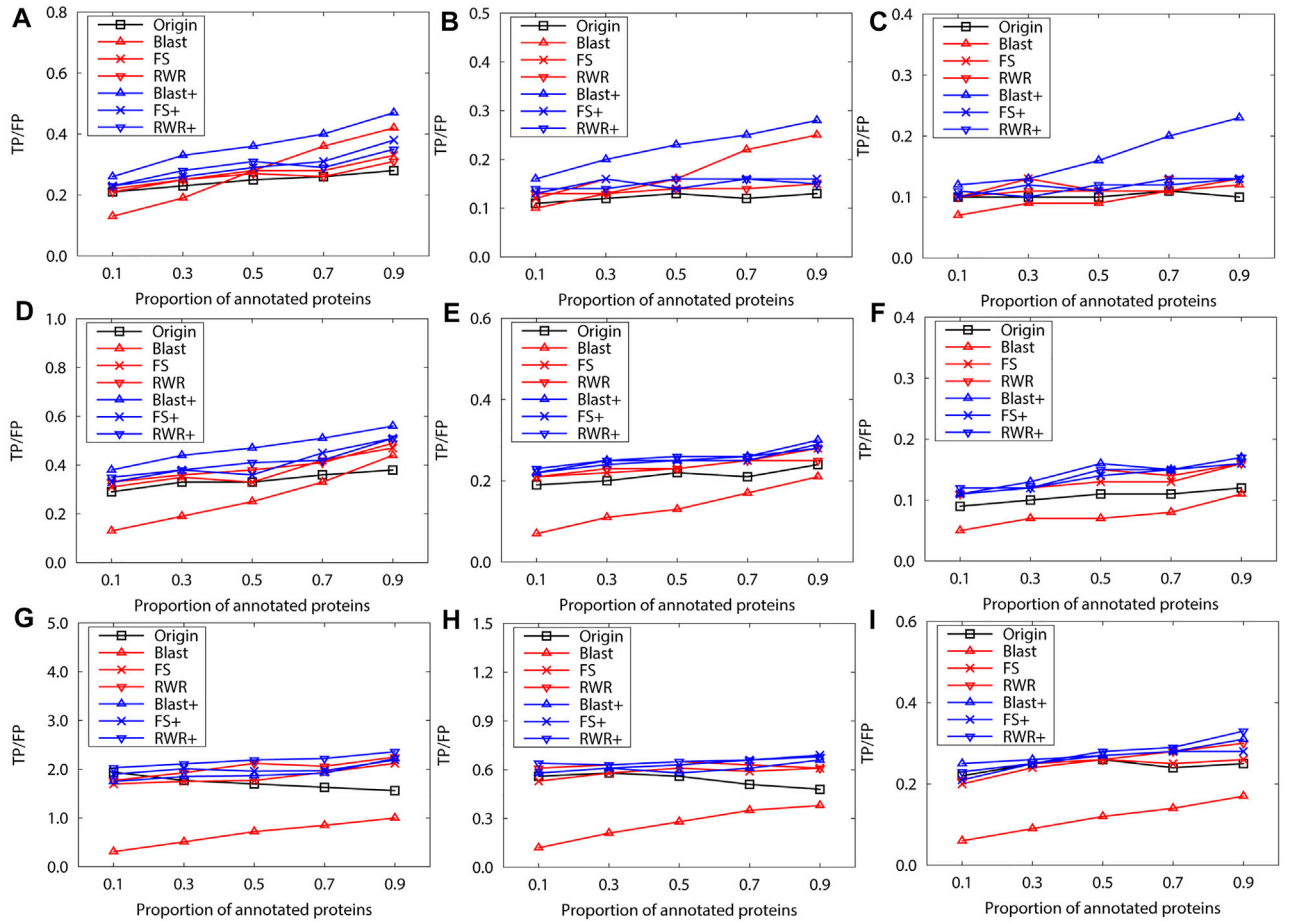

FIGURE 3 | The performance evaluation over the sparsely-labeled networks (Dataset A: *M.musculus*). Here, the sub figures in the horizontal and vertical directions represent the experimental results for the PPI networks of different function types and rank predicted functions, respectively. Horizontally, the top three subplots represent ones on the PPI networks of molecular function, the middle for ones of biological process, and the bottom for ones of cellular component **(A)**, **(D)** and **(G)** first rank predicted function, **(B)**, **(E)** and **(H)** second rank predicted function, **(C)**, **(F)** and **(I)** third rank predicted function.

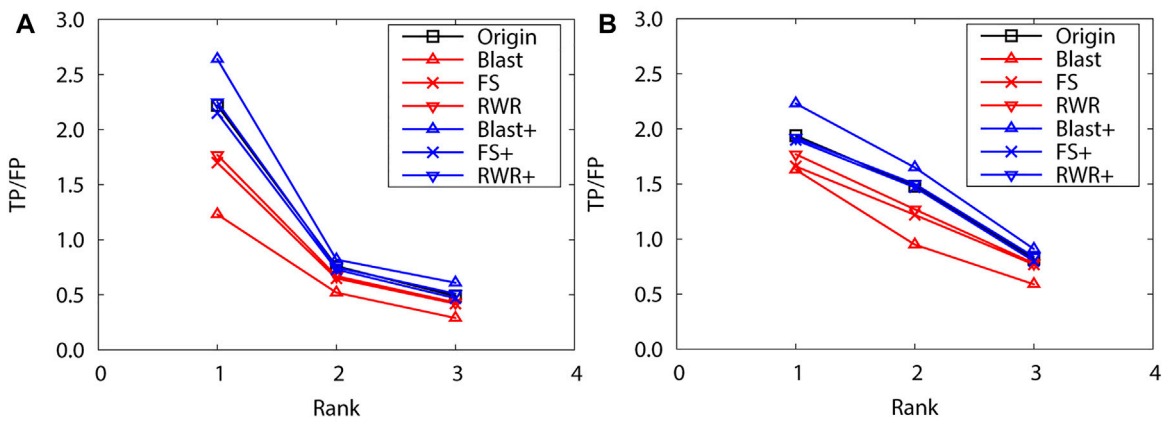

FIGURE 4 | The performance evaluation by leave-one-out validation over the PPI networks (Dataset B: *S.cerevisiae* and *M.musculus*) **(A)** *S.cerevisiae* **(B)** *M.musculus*.

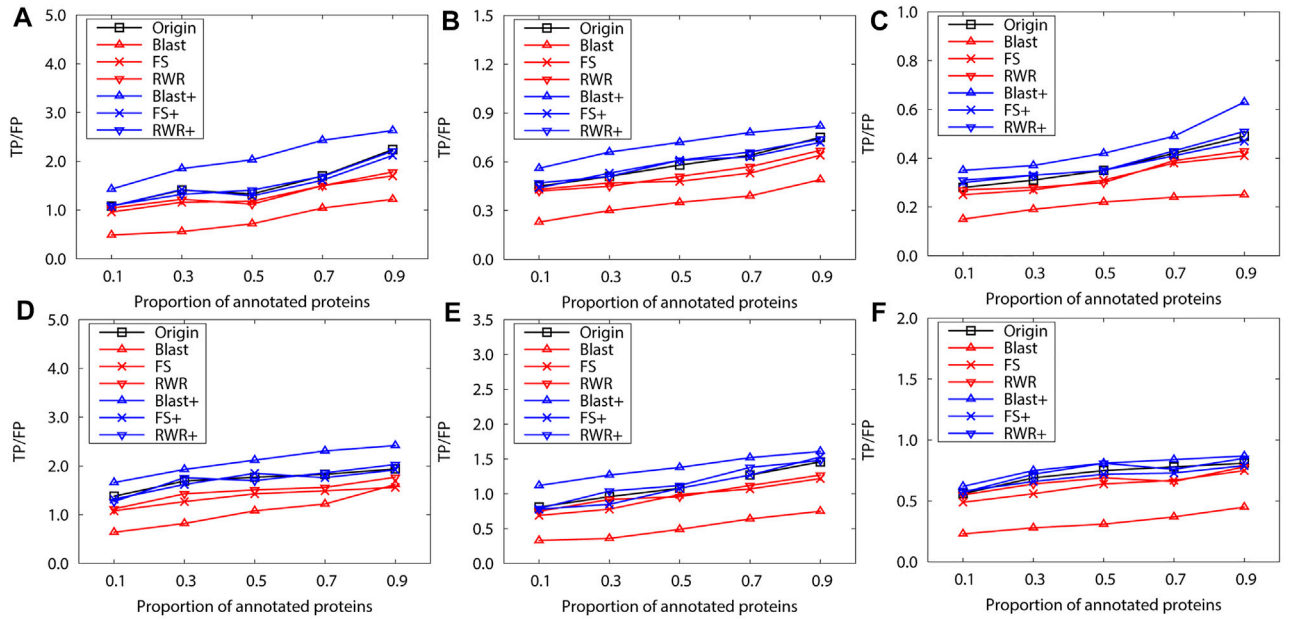

FIGURE 5 | The performance evaluation over the sparsely-labeled networks (Dataset B: *S.cerevisiae* and *M.musculus*). Here, the sub figures in the horizontal and vertical directions represent the experimental results for different data types and rank predicted functions, respectively. Horizontally, the top three subplots represent ones over the dataset of *S.cerevisiae*, and the bottom for ones of *M.musculus*. **(A)** and **(D)** first rank predicted function, **(B)** and **(E)** second rank predicted function, **(C)** and **(F)** third rank predicted function.

sequence information, and thus performe worst. The experimental results also validate that FS-reconstructed networks and RWR-reconstructed networks work better than the original networks in most cases. This is because the reconstructed networks filter out noisy or spurious interactions in the original PPI networks.

We further evaluate prediction accuracy of these three kinds of networks by using Gibbs Sampling in sparse-labeled PPI networks. Concretely, in PPI networks, the annotated protein proportion is changed from 0.1 to 0.9, and the remaining protein functions are predicted. For each proportion of the annotated proteins, the average prediction accuracy of running 10 experiments is presented on the PPI networks of *S.cerevisiae* (**Figure 2**)and *M.musculus* (**Figure 3**), respectively. The enrichment gains more accurate predictions than network reconstruction and original networks. The BLAST-enriched networks always work the best, while the BLAST-reconstructed networks always perform the worst. As expected, the experimental results also validate that FS-reconstructed networks and RWR-reconstructed networks generally performe better than the original networks. As the annotated protein proportion in the original networks increases, the prediction performance gets better for most networks, especially for the 1-st rank function. However, the prediction performance of the original network slightly declines as its annotated protein proportion increases (**Figure 3G, H**).

3.4 Performance Evaluation on Dataset B

As above, the performance of reconstructed and enriched networks is first compared with that of the original networks by leave-one-out validation. Here, the top 3 protein function predictions are considered for both PPI networks of *S. cerevisiae* and *M. musculus*. As expected, edge enrichment gaines higher accurate predictions than network reconstruction and original networks. Moreover, BLAST-enriched networks get best, while the BLAST-reconstructed networks always work worst (**Figure 4**). The reasons are the same as for the dataset A.

Next, we evaluate the prediction performance of these networks in sparse-labeled conditions with the collective classification method. Similarly, the average prediction performance is generated over running 10 experiments, with the annotated-protein proportion varying from 0.1 to 0.9. Generally, the experimental results present a similar trend to the above for the dataset A (**Figure 5**). However, FS-reconstructed networks and RWR-reconstructed networks do not outperform the original networks, due to the quality properties of the dataset itself. This is mainly because many informative interactions are deleted and the prediction performance is impaired when reconstructing the networks based on similarity.

To validate this point, 10% and 50% interactions of the original network of the dataset B are randomly selected to construct two sparse networks. The leave-one-out validation is then performed over the two sparse networks. The selection process have two steps: First, a random weight is assigned to each edge of the original network, and a minimum spanning tree is constructed on the new network. The randomness of the minimum spanning tree (*MST*) is ensured by the random weights, and *MST* ensures the connectivity of the sparse network. Second, the *MST* is expanded by adding a number of edges, which are randomly selected from the original network (but not

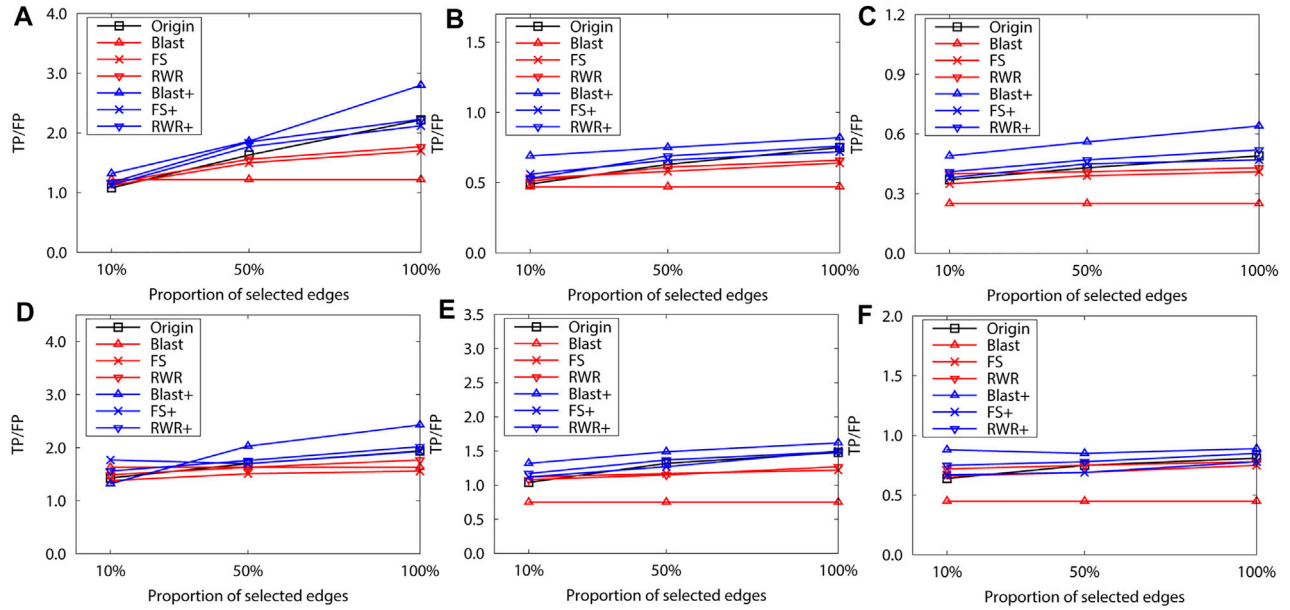

FIGURE 6 | The performance evaluation by leave-one-out validation over the PPI networks (Dataset B: *S.cerevisiae* and *M.musculus*). Here, the sub figures in the horizontal and vertical directions represent the experimental results for different data types and rank predicted functions, respectively. Horizontally, the top three subplots represent ones over the dataset of *S.cerevisiae*, and the bottom for ones of *M.musculus* **(A)** and **(D)** first rank predicted function, **(B)** and **(E)** second rank predicted function, **(C)** and **(F)** third rank predicted function.

already on the *MST*). Hence, the number of edges in the sparse network is equal to 10% or 50% of edges in the original network. The sparse network preserves the basic topological properties of the original network.

The final experimental results also confirm the above-mentioned phenomenon. For example, in **Figure 6**, the FS-reconstructed networks and the RWR-reconstructed networks work better than the original networks when the networks are very sparse (e.g. 10%). However, as the networks become denser, the FS-reconstructed networks and the RWR-reconstructed networks get worse than the original networks.

4 CONCLUSION

The systematic comparison of two network transformation approaches (network reconstruction and edge enrichment) is performed using three different protein similarity metrics (sequence similarity, local and global similarity). In summary, edge enrichment performs better than network reconstruction and original networks, while network reconstruction is more effective on relatively small and incomplete PPI networks. The edge enrichment of PPI networks based on sequence similarity outperforms those based on both local and global similarity. As the PPI networks become more and more complete, the effectiveness of both edge enrichment and network reconstruction will decrease or relatively decrease.

Research efforts will be further expanded in future, which include: 1) how the removal of noisy edges and addition of informative edges affect the prediction performance; 2) a combining approach that combines the best properties of all these indices is developed since the similarity indices considered here have different properties and performances.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: (1) Datasets A: BioGRID, https://downloads.thebiogrid.org/BioGRID. (2) Datasets A: STRING, https://string-db.org.

AUTHOR CONTRIBUTIONS

JZ, JG, WX and JG, JH designed and performed the experiments. JZ, JG, WX and YW analyzed the data. The manuscript was written by JZ, JG and WX and approved by all authors.

FUNDING

This work was supported by National Natural Science Foundation of China (NSFC) under Grants Nos. 41877009, 61772367, 62172300, U1936205 and by the Fundamental Research Funds for the Central Universities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.758131/full#supplementary-material

REFERENCES

Abdollahi, S., Lin, P.-C., and Chiang, J.-H. (2021). Winbinvec: Cancer-Associated Protein-Protein Interaction Extraction and Identification of 20 Various Cancer Types and Metastasis Using Different Deep Learning Models. *IEEE J. Biomed. Health Inform.* 25, 4052–4063. doi:10.1109/JBHI.2021.3093441

Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). CFinder: Locating Cliques and Overlapping Modules in Biological Networks. *Bioinformatics* 22, 1021–1023. doi:10.1093/bioinformatics/btl039

Altschul, S., Madden, T., 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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA Database in 2009--an Integrated Gene Ontology Annotation Resource. *Nucleic Acids Res.* 37, D396–D403. doi:10.1093/nar/gkn803

Berggård, T., Linse, S., and James, P. (2007). Methods for the Detection and Analysis of Protein-Protein Interactions. *Proteomics* 7, 2833–2842. doi:10.1002/pmic.200700131

Bogdanov, P., and Singh, A. K. (2010). Molecular Function Prediction Using Neighborhood Features. *Ieee/acm Trans. Comput. Biol. Bioinf.* 7, 208–217. doi:10.1109/TCBB.2009.81

Chen, Y., Wang, W., Liu, J., Feng, J., and Gong, X. (2020). Protein Interface Complementarity and Gene Duplication Improve Link Prediction of Protein-Protein Interaction Network. *Front. Genet.* 11, 291. doi:10.3389/fgene.2020.00291

Chua, H. N., Sung, W.-K., and Wong, L. (2007). An Efficient Strategy for Extensive Integration of Diverse Biological Data for Protein Function Prediction. *Bioinformatics* 23, 3364–3373. doi:10.1093/bioinformatics/btm520

Chua, H. N., Sung, W.-K., and Wong, L. (2006). Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions. *Bioinformatics* 22, 1623–1630. doi:10.1093/bioinformatics/btl145

Dunn, R., Dudbridge, F., and Sanderson, C. M. (2005). The Use of Edge-Betweenness Clustering to Investigate Biological Function in Protein Interaction Networks. *BMC Bioinformatics* 6, 39. doi:10.1186/1471-2105-6-39

Hu, L., Huang, T., Shi, X., Lu, W.-C., Cai, Y.-D., and Chou, K.-C. (2011). Predicting Functions of Proteins in Mouse Based on Weighted Protein-Protein Interaction Network and Protein Hybrid Properties. *PLOS ONE* 6, e14556. doi:10.1371/journal.pone.0014556

Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W., and Pržulj, N. (2010). Topological Network Alignment Uncovers Biological Function and Phylogeny. *J. R. Soc. Interf.* 7, 1341–1354. doi:10.1098/rsif.2010.0063

Liben-Nowell, D., and Kleinberg, J. (2007). The Link-Prediction Problem for Social Networks. *J. Am. Soc. Inf. Sci.* 58, 1019–1031. doi:10.1002/asi.20591

Lü, L., and Zhou, T. (2011). Link Prediction in Complex Networks: A Survey. *Physica A: Stat. Mech. its Appl.* 390, 1150–1170. doi:10.1016/j.physa.2010.11.027

Newman, M. E. J. (2001). Clustering and Preferential Attachment in Growing Networks. *Phys. Rev. E* 64, 025102. doi:10.1103/PhysRevE.64.025102

Ofran, Y., Punta, M., Schneider, R., and Rost, B. (2005). Beyond Annotation Transfer by Homology: Novel Protein-Function Prediction Methods to Assist Drug Discovery. *Drug Discov. Today* 10, 1475–1482. doi:10.1016/S1359-6446(05)03621-4

Patel, S., Tripathi, R., Kumari, V., and Varadwaj, P. (2017). Deepinteract: Deep Neural Network Based Protein-Protein Interaction Prediction Tool. *Cbio* 12, 551–557. doi:10.2174/1574893611666160815150746

Schwikowski, B., Uetz, P., and Fields, S. (2000). A Network of Protein-Protein Interactions in Yeast. *Nat. Biotechnol.* 18, 1257–1261. doi:10.1038/82360

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective Classification in Network Data. *AIMag* 29, 93–106. doi:10.1609/aimag.v29i3.2157

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based Prediction of Protein Function. *Mol. Syst. Biol.* 3, 88. doi:10.1038/msb4100129

Singh, R., Xu, J., and Berger, B. (2008). Global Alignment of Multiple Protein Interaction Networks with Application to Functional Orthology Detection. *Proc. Natl. Acad. Sci.* 105, 12763–12768. doi:10.1073/pnas.0806627105

Sleator, R. D., and Walsh, P. (2010). An Overview of In Silico Protein Function Prediction. *Arch. Microbiol.* 192, 151–155. doi:10.1007/s00203-010-0549-9

Solava, R. W., Michaels, R. P., and Milenković, T. (2012). Graphlet-based Edge Clustering Reveals Pathogen-Interacting Proteins. *Bioinformatics* 28, i480–i486. doi:10.1093/bioinformatics/bts376

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., et al. (2011). The Biogrid Interaction Database: 2011 Update. *Nucleic Acids Res.* 39, D698–D704. doi:10.1093/nar/gkq1116

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The String Database in 2011: Functional Interaction Networks of Proteins, Globally Integrated and Scored. *Nucleic Acids Res.* 39, D561–D568. doi:10.1093/nar/gkq973

Täubig, H., Buchner, A., and Griebsch, J. (2006). Past: Fast Structure-Based Searching in the PDB. *Nucleic Acids Res.* 34, W20–W23. doi:10.1093/nar/gkl273

Tong, H., Faloutsos, C., and Pan, J.-Y. (2008). Random Walk with Restart: Fast Solutions and Applications. *Knowl Inf. Syst.* 14, 327–346. doi:10.1007/s10115-007-0094-2

Waiho, K., Afiqah-Aleng, N., Iryani, M. T. M., and Fazhan, H. (2021). Protein-protein Interaction Network: an Emerging Tool for Understanding Fish Disease in Aquaculture. *Rev. Aquacult.* 13, 156–177. doi:10.1111/raq.12468

Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1996). Derivation of 3D Coordinate Templates for Searching Structural Databases: Application to Ser-His-Asp Catalytic Triads in the Serine Proteinases and Lipases. *Protein Sci.* 5, 1001–1013. doi:10.1002/pro.5560050603

Wu, Q., Ye, Y., Ng, M. K., Ho, S.-S., and Shi, R. (2014). Collective Prediction of Protein Functions from Protein-Protein Interaction Networks. *BMC Bioinformatics* 15, S9. doi:10.1186/1471-2105-15-S2-S9

Wu, Z., Liao, Q., and Liu, B. (2020). A Comprehensive Review and Evaluation of Computational Methods for Identifying Protein Complexes from Protein-Protein Interaction Networks. *Brief. Bioinform.* 21, 1531–1548. doi:10.1093/bib/bbz085

Xiong, W., Liu, H., Guan, J., and Zhou, S. (2013). Protein Function Prediction by Collective Classification with Explicit and Implicit Edges in Protein-Protein Interaction Networks. *BMC Bioinformatics* 14, S4. doi:10.1186/1471-2105-14-S12-S4

Xiong, W., Xie, L., Zhou, S., and Guan, J. (2014). Active Learning for Protein Function Prediction in Protein-Protein Interaction Networks. *Neurocomputing* 145, 44–52. doi:10.1016/j.neucom.2014.05.075

Yang, H., Wang, M., Liu, X., Zhao, X.-M., and Li, A. (2021). PhosIDN: an Integrated Deep Neural Network for Improving Protein Phosphorylation Site Prediction by Combining Sequence and Protein-Protein Interaction Information. *Bioinformatics*, btab551. doi:10.1093/bioinformatics/btab551

Ye, Y., and Godzik, A. (2004). FATCAT: a Web Server for Flexible Structure Comparison and Structure Similarity Searching. *Nucleic Acids Res.* 32, W582–W585. doi:10.1093/nar/gkh430

Zhu, H., Du, X., and Yao, Y. (2020). Convsppis: Identifying Protein-Protein Interaction Sites by an Ensemble Convolutional Neural Network with Feature Graph. *Cbio* 15, 368–378. doi:10.2174/1574893614666191105155713

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 | Frontiers in Genetics

ORIGINAL RESEARCH
published: 08 April 2022
doi: 10.3389/fgene.2022.801902

Structural Genomic Analysis of SARS-CoV-2 and Other Coronaviruses

Qiong Zhang[1,2,3], *Huai-Lan Guo*[3,4], *Jing Wang*[3,4], *Yao Zhang*[3,4], *Ping-Ji Deng*[3,4]* and *Fei-Feng Li*[3,4]*

[1]School of Pharmaceutical Sciences, Hubei University of Medicine, Shiyan, China, [2]Hubei Key Laboratory of Wudang Local Chinese Medicine Research, Hubei University of Medicine, Shiyan, China, [3]Hubei Biomedical Detection Sharing Platform in Water Source Area of South to North Water Diversion Project, Hubei University of Medicine, Shiyan, China, [4]School of Public Health, Hubei University of Medicine, Shiyan, China

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is the causative agent of the coronavirus disease 2019 (COVID-19) pandemic. In this study, we conducted a comparative analysis of the structural genes of SARS-CoV-2 and other CoVs. We found that the sequence of the E gene was the most evolutionarily conserved across 200 SARS-CoV-2 isolates. The E gene and M gene sequences of SARS-CoV-2 and NC014470 CoV were closely related and fell within the same branch of a phylogenetic tree. The absolute diversity of E gene and M gene sequences of SARS-CoV-2 isolates was similar to that of common CoVs (C-CoVs) infecting other organisms. The absolute diversity of the M gene sequence of the KJ481931 CoV that can infect humans was similar to that of SARS-CoV-2 and C-CoVs infecting other organisms. The M gene sequence of KJ481931 CoV (infecting humans), SARS-CoV-2 and NC014470 CoV (infecting other organisms) were closely related, falling within the same branch of a phylogenetic tree. Patterns of variation and evolutionary characteristics of the N gene and S gene were very similar. These data may be of value for understanding the origins and intermediate hosts of SARS-CoV-2.

Keywords: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), common coronaviruses (C-CoVs), structural gene, evolution, intermediate hosts

OPEN ACCESS

Edited by:
Lei Deng,
Central South University, China

Reviewed by:
Seyed Reza Mohebbi,
Shahid Beheshti University of Medical Sciences, Iran
Yan Yousheng,
Capital Medical University, China

***Correspondence:**
Fei-Feng Li
20200510@hbmu.edu.cn
Ping-Ji Deng
dengpj@hbmu.edu.cn

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

Received: *26 October 2021*
Accepted: *01 March 2022*
Published: *08 April 2022*

Citation:
Zhang Q, Guo H-L, Wang J, Zhang Y, Deng P-J and Li F-F (2022) Structural Genomic Analysis of SARS-CoV-2 and Other Coronaviruses. Front. Genet. 13:801902. doi: 10.3389/fgene.2022.801902

INTRODUCTION

The coronaviruses (CoVs) are a large family of viruses that infect many organisms, including humans (Ma et al., 2020). The primary symptoms resulting from CoV infection are respiratory diseases and severe acute respiratory syndrome (Ashour et al., 2020). CoVs are enveloped viruses with a positive sense single stranded RNA genome. CoVs were first discovered in patients with the common cold in 1966 (Tyrrell and Bynoe 1966; Velavan and Meyer 2020).

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) belongs to the *Betacoronavirus* genus and the *Sarbecovirus* subgenus (Ceraolo and Giorgi 2020; Li F et al., 2020). Infection by SARS-CoV-2 results in a syndrome called coronavirus disease 2019 (COVID-19); the virus has caused a global pandemic, resulting in large numbers of illnesses and deaths [(An update on the epidemiological characteristics of novel coronavirus pneumoniaCOVID-19) 2020]. The main features of COVID-19 are high transmissibility and high mortality [Lai et al., 2020, (An update on the epidemiological characteristics of novel coronavirus pneumoniaCOVID-19) 2020]. Since the first patient with COVID-19 was identified (Lai, Shih, Ko, Tang and Hsueh 2020), more than 68 million additional cases have been confirmed globally with over 1.5 million deaths.

Many organisms have been considered as potential intermediate hosts of SARS-CoV-2 [Guo et al., 2020; Jiang and Shi 2020, (An update on the epidemiological characteristics of novel coronavirus pneumoniaCOVID-19) 2020; Zhang et al., 2020c; Zhou et al., 2020]. In a previous study, we concluded that SARS-CoV-2 may have evolved from a distant common ancestor of other common CoVs (C-CoVs), and may have persisted in an unidentified primary host for a long period (Li X et al., 2020). However, the origins and the intermediate hosts of SARS-CoV-2 remain unclear.

The SARS-CoV-2 genome is about 30 kb in size, making it one of the largest known viral RNA genomes. The genome contains four structural genes: S, E, M and N (Comas-Garcia 2019; Khailany et al., 2020). The "crown-like" appearance of SARS-CoV-2 results from the presence of the spike (S) glycoprotein (encoded by the S gene) on the surface of the virus (Jacofsky et al., 2020). The S protein binds to angiotensin-converting enzyme-2 (ACE2) and mediates fusion of the viral envelope with host cells (Lu et al., 2020). The other major SARS-CoV-2 envelope protein is the transmembrane (M) glycoprotein (encoded by the M gene) (Jacofsky et al., 2020). The main functions of the M protein are viral envelope formation and virion assembly (Ujike and Taguchi 2015; Jacofsky et al., 2020). The SARS-CoV-2 capsid and genomic RNA are linked by the basic (N) phosphoprotein (encoded by the N gene) (Khailany et al., 2020; Mousavizadeh and Ghasemi 2020). The other structural protein is the envelope (E) protein (encoded by the E gene), which is involved in virion assembly, release, and viral pathogenesis (Schoeman and Fielding 2019). The sequences of SARS-CoV-2 structural genes or proteins may contain information on the origins and intermediate hosts of the virus, which may be useful for vaccine development.

In this study, we analyzed the sequences of the structural genes of SARS-CoV-2 and C-CoVs that infect humans and other organisms. We aimed to understand variation and evolutionary characteristics of SARS-CoV-2 structural gene sequences.

MATERIALS AND METHODS

Materials

We obtained structural gene sequences from 200 SARS-CoV-2 isolates, 126 C-CoVs that infect humans, and 53 C-CoVs that infect other organisms from the NCBI database (https://www.ncbi.nlm.nih.gov/sars-cov-2/).

Analysis of Variation in SARS-CoV-2 Structural Gene Sequences

To analyze variation in the structural gene sequences of 200 SARS-CoV-2 isolates, we carried out multiple sequence alignments using Vector NTI software (Li et al., 2016). We analyzed the influence of mutations in structural gene sequences on the functions of structural proteins using DNAMAN software. We used MEGA-X software (Gorbalenya et al., 2020) to analyze the evolutionary features of SARS-CoV-2 structural gene sequences.

Comparative Analysis of Structural Genes in SARS-CoV-2 and Other CoVs

We chose SARS-CoV-2 structural genes that showed sequence variation or evolutionary relatedness to C-CoVs for further analysis (**Table 1**). Using Vector NTI software and MEGA-X software (Kumar et al., 2018), we conducted a comparative sequence analysis of the structural gene sequences of SARS-CoV-2, C-CoVs that infect humans, and C-CoVs that infect other organisms.

RESULTS

Genomic Analysis of SARS-CoV-2 Structural Gene Sequences

The four structural genes encoded in the SARS-CoV-2 genome are E (228 nt), M (669 nt), N (908 nt), and S (3,822 nt). As shown in **Figure 1**, the similarities and absolute diversities of SARS-CoV-2 structural gene sequences were very high (**Figure 1** A,B).

Two SARS-CoV-2 isolates had two single nucleotide polymorphisms (SNPs) within the E gene (**Figure 1** C,D and **Table 1**), nine isolates had three variations (one mutation and two SNPs) within the M gene (**Figure 1** C,D and **Table 1**), 28 isolates had 22 variations (13 mutations and nine SNPs) within the N gene (**Figure 1**, C–T and **Table 1**) and 89 isolates had 25 variations (16 mutations and nine SNPs) within the S gene (**Figure 1**, C–T and **Table 1**).

The variance rates (VRs) of structural genes among the 200 SARS-CoV-2 isolates were 1% (E), 4.5% (M), 14% (N) and 44.5% (S) (**Table 1**). The gene size variance rates (GSVRs) of the four genes were 0.44/10,000 (E), 0.67/10,000 (M), 1.54/10,000 (N) and 1.16/10,000 (S) (**Table 1**). The sequence of the E gene was the most highly conserved across the 200 SARS-CoV-2 isolates.

Influence of Mutations in SARS-CoV-2 Structural Genes on the Features of Structural Proteins

We identified 30 mutations within the structural genes of 200 SARS-CoV-2 isolates. Subsequently, we analyzed the influence of these mutations on the features of structural proteins. As shown in **Supplementary Figure S1**, the Val70→Ile substitution in the M gene of the MT263397 isolate had little effect on the transmembrane segment of the M protein.

In the N gene, six mutations affected N protein hydrophobicity, three mutations affected protein hydrophilicity, 10 mutations affected protein secondary structure, and four mutations affected the transmembrane segment (**Supplementary Figure S2**).

One mutation in the S gene affected S protein hydrophobicity, one mutation affected protein hydrophilicity, and three mutations affected protein secondary structure (**Supplementary Figure S3**).

In general, mutations in the N gene of SARS-CoV-2 isolates occurred between amino acid residues 200 to 300 and had large

TABLE 1 | Analysis of structural gene sequences of 200 severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates.

Genes	Size (nt)	Variations[1]	Variance rate[2] (%)	Gene size variance rate[3]	SNPs	Mutations	For further analysis
E gene	228	2	1	0.44/10,000	MT263389, MT259248	—	MT263389, MT259248, ***MT263410***[6]
M gene	669	9	4.5	0.67/10,000	MT259252, MT263384, MT263410, MT263389, MT263443, MT263388, MT263422, MT263447	MT263397	MT263410, MT263389, MT263397, ***MT263074***[6]
N gene	908	28	14	1.54/10,000	MT263398, ***MT256917***[4], ***MT256918***[4], MT259270, MT263430, MT259267, MT263421, MT263451, MT258382, MT263435, MT263458, MT263395, MT259237	MT259237, MT259269, MT259274, MT263429, ***MT256917***[4], ***MT256918***[4], MT258379, MT259250, MT259263, MT263402, MT263074, MT263386, MT263410, MT263411, MT256924, MT263422, LC534419	MT263410, MT263074, MT263422, MT259237, MT259269, MT256917, MT263386, MT263411, MT258382, MT263398, MT259274, MT259270, MT263429, MT259267, MT263421, MT256924, LC534419, MT263435, MT263395, ***MT263389***[6]
S gene	3,822	89	44.5	1.16/10,000	MT259262, MT263410, MT259257, MT263441, MT263469, MT263386, MT259287, MT263074, MT259269, MT259227	MT263414, MT263460, MT263384, MT259249, ***MT263466(2)***[5], MT259236, MT259276, MT263403, MT263412, MT263418, MT259262, MT259282, MT259253, MT262915, MT263457, MT263443, MT263393, MT263420, MT263385, MT263387, MT251973, MT251976, MT251979, MT258378, MT258379, MT258380, MT258382, MT258383, MT259235, MT259239, MT259240, MT259243, MT259244, MT259246, MT259248, MT259249, MT259250, MT259251, MT259256, MT259258, MT259260, MT259261, MT259263, MT259264, MT259265, MT259273, MT259277, MT263431, MT263436, MT259278, MT259281, MT259286, MT263074, MT263390, MT263391, MT263392, MT263394, MT263402, MT263406, MT263408, MT263411, MT263413, MT263415, MT263417, MT263426, MT263428, MT263432, MT263433, MT263437, MT263438, MT263439, MT263442, MT263445, MT263446, MT263459, MT263465, MT263467, MT263468	MT263410, MT263074-3, MT263466, MT263384, MT263443, MT259269, MT263386, MT259249, MT263414, MT259262, MT259257, MT259236, MT259282, MT263441, MT262915, MT259287, MT251973, MT263393, MT263385, MT259253, MT263457, MT263420, MT259227, ***MT263389***[6]

Notes: [1]Variations include single nucleotide polymorphisms (SNPs) and mutations.
[2]Variance rate= (variations/200) × 100%.
[3]Gene size variance rate= (variations/200/gene size) × 10,000/10,000.
[4]There were two variations in the MT256917 and MT256918 CoVs, respectively.
[5]There were two mutations in the MT263466 CoV.
[6]No variation controls for further analysis of structural genes.

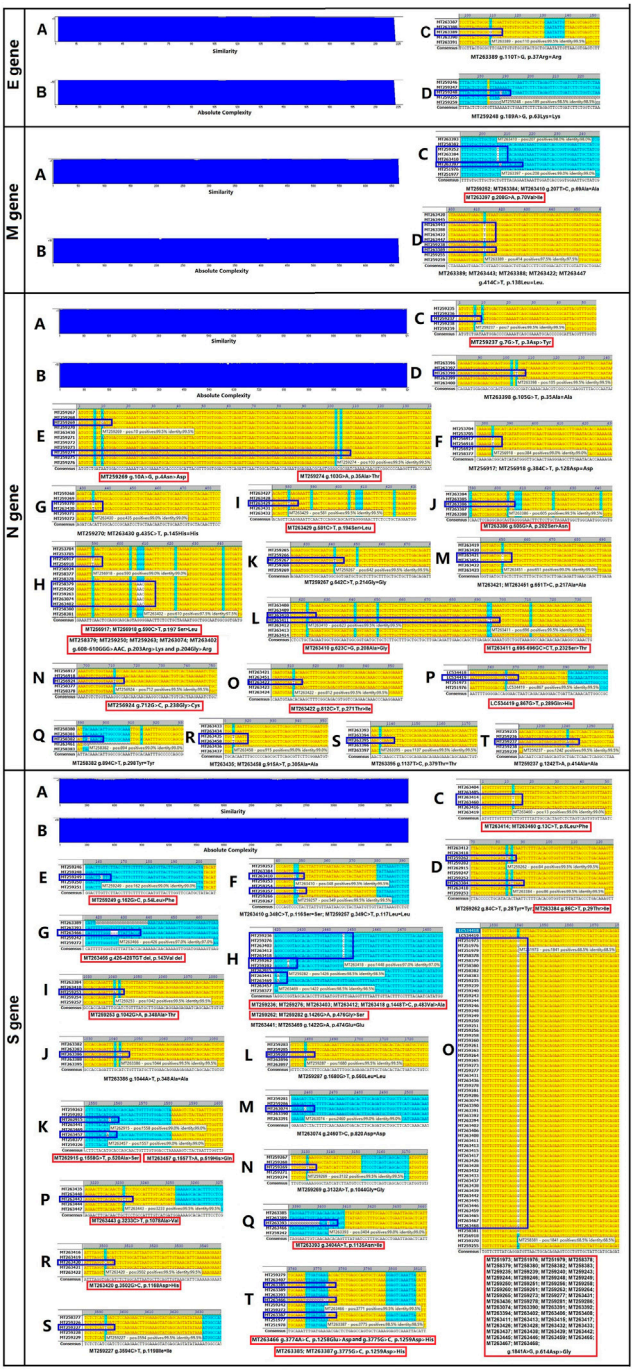

FIGURE 1 | Absolute diversity and variations in the structural genes of 200 severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates. The similarity and absolute diversity in structural genes sequences were very high. Two SARS-CoV-2 isolates had two single nucleotide polymorphisms (SNPs) within the E gene, nine isolates had three variations (one mutation and two SNPs) within the M gene, 28 strains had 22 variations (13 mutations and nine SNPs) within the N gene, and 89 strains had 25 variations (16 mutations and nine SNPs) within the S gene.

impacts on the function of the protein (**Figure 1** and **Supplementary Figure S2**).

Phylogenetic Analysis of SARS-CoV-2 Structural Gene Sequences

Next, we analyzed the evolutionary characteristics of the structural genes of SARS-CoV-2 isolates. As shown in **Figure 2**, the SARS-CoV-2 structural genes showing increased variation also showed distinct evolutionary features. The sequence of the E gene was the most evolutionarily conserved across the 200 SARS-CoV-2 isolates (**Figure 2**). We selected the sequences of structural genes that showed variation and evolutionary relatedness with C-CoVs for further analysis (**Table 1**).

Comparative Analysis of Structural Gene Sequences of SARS-CoV-2 and C-CoVs That Infect Humans

To understand the relationships between the structural genes of SARS-CoV-2 and C-CoVs that also infect humans, we carried out a comparative sequence analysis of selected structural gene sequences from SARS-CoV-2 (**Table 1**) and C-CoVs that infect humans. As shown in **Figure 3**, the E gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between KJ481931 and MG011357 (**Figure 3A**). In terms of their E gene sequences, SARS-CoV-2 and KJ481931 were the most closely related evolutionarily (**Figure 3A**), and the absolute diversities of the E gene sequences of these two CoVs was similar (**Figure 3B**).

The M gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between KJ48193 and a group of other CoVs (KP209309, KY581691, KY581689, KY581686, KP209307, KP209313, and KP209306). The M gene sequences of SARS-CoV-2 and KJ481931 were the most closely related evolutionarily (**Figure 3C**), and the absolute diversities of the M gene sequences of these two CoVs was similar (**Figure 3D**).

The N gene and S gene sequences of SARS-CoV-2 isolates were evolutionarily distinct (**Figure 3E** and **Figure 3G**). The absolute diversities of N gene sequences in SARS-CoV-2 isolates differed from those of all other C-CoVs (**Figure 3F**). However, the S gene sequences of SARS-CoV-2 isolates and KJ481931 were the most closely related evolutionarily (**Figure 3G**), and the absolute diversities of the S gene sequences of these two CoVs were similar (**Figure 3H**).

Comparative Analysis of Structural Gene Sequences of SARS-CoV-2 and C-CoVs That Infect Other Organisms

To understand the relationships between the structural genes of SARS-CoV-2 and C-CoVs that infect other organisms, we carried out a comparative sequence analysis of selected structural gene sequences from SARS-CoV-2 (**Table 1**) and C-CoVs that infect

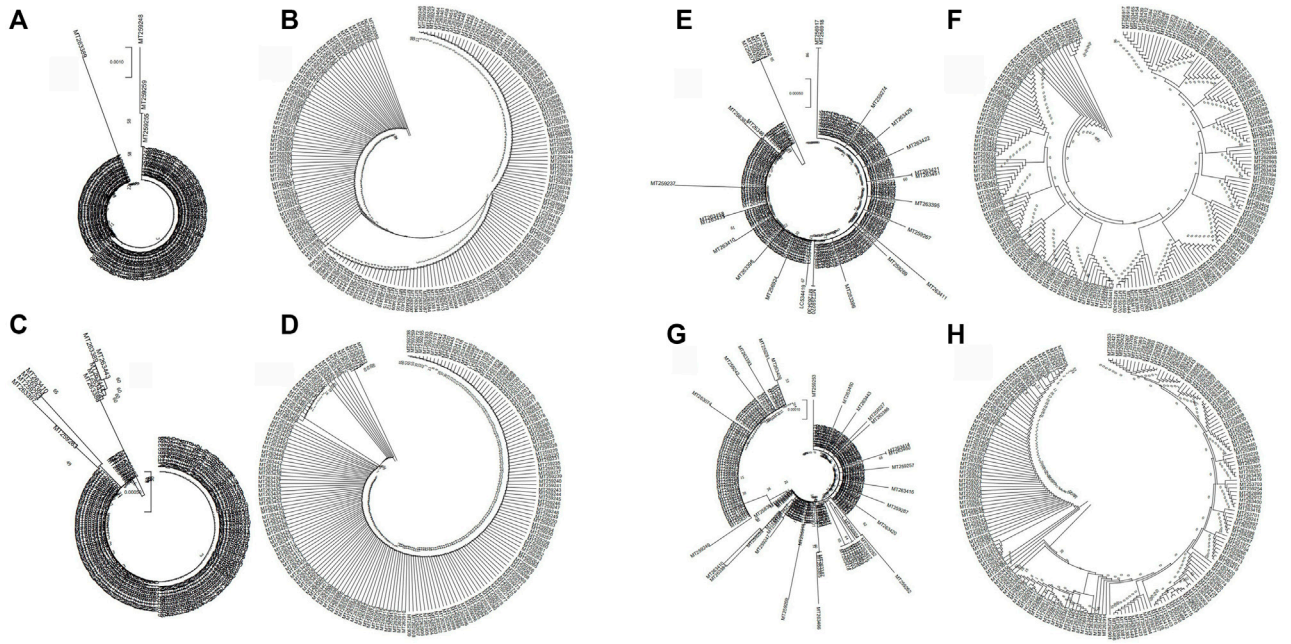

FIGURE 2 | Evolutionary characteristics of the structural genes of 200 severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates. **(A,B)**: The two isolates with single nucleotide polymorphisms (SNPs) in the E gene were evolutionary distinct. **(C,D)**: The nine isolates with variations in the M gene were evolutionary distinct. **(E,F)**: The 28 isolates with variations in the N gene were evolutionarily distinct. **(G,H)**: The 88 strains with variations in the S gene evolutionarily distinct.

other organisms. As shown in **Figure 4**, the E gene sequences of SARS-CoV-2 isolates were most closely evolutionarily related to NC014470, DQ415914, NC026011, NC006213, JN874559, and U00735l; NC014470 was also located within the same branch of a phylogenetic tree as SARS-CoV-2 isolates (**Figure 4A**). The absolute diversities of E gene sequences from NC014470, DQ415914, NC026011, NC006213, JN874559, and U00735 were similar to those of E gene sequences from SARS-CoV-2 isolates (**Figure 4B**).

The M gene sequences of SARS-CoV-2 isolates were most closely related to NC014470, EF065513 and NC030886 (**Figure 4C**). The absolute diversities of M gene sequences from NC014470, EF065513 and NC030886 were similar to those of M gene sequences from SARS-CoV-2 isolates (**Figure 4D**).

In terms of N gene and S gene sequences, SARS-CoV-2 was most closely evolutionarily related to NC014470; these two CoVs formed a separate clade in a phylogenetic tree (**Figure 4E** and **Figure 4G**). The absolute diversity of N gene sequences from SARS-CoV-2 isolates was similar to that of the N gene sequence of NC014470 (**Figure 4F**). However, the absolute diversity of the S gene sequence from NC014470 was more similar to those of the S gene sequences of other C-CoVs (**Figure 4H**).

Comparative Analysis of Structural Gene Sequences of SARS-CoV-2 and C-CoVs That Infect Humans and Other Organisms

We next wanted to analyze the evolutionary relationships among the structural genes of SARS-CoV-2 and C-CoVs that infect humans and other organisms. We performed a comparative sequence analysis of the structural genes from SARS-CoV-2 isolates (**Table 1**) and those from C-CoVs (**Table 2**). As shown in **Figure 5**, the E gene sequences of SARS-CoV-2 isolates and C-CoVs could be grouped into three clades (CI, CII and CIII) (**Figures 5A,B**). In terms of their E gene sequences, SARS-CoV-2 isolates were most closely related to NC014470; these two CoVs represented evolutionary intermediates in the phylogenetic tree between C-CoVs that infect humans and those that infect other organisms (**Figures 5A,B**). The absolute diversity of E gene sequences of SARS-CoV-2 isolates was most similar to that of the E gene sequences of C-CoVs that infect other organisms (**Figure 5C**).

The M gene sequences of SARS-CoV-2 isolates and C-CoVs could be also grouped into three clades (CI, CII and CIII) (**Figures 5D,E**). The M gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between NC014470 (infecting other organisms) and KJ481931 (infecting humans); SARS-CoV2 isolates grouped closely together in a same branch of the phylogenetic tree (**Figures 5D,E**). The absolute diversity of the M gene sequences of SARS-CoV-2 isolates was more similar to those of the M gene sequences of C-CoVs that infect other organisms (**Figure 5F**). However, the absolute diversity of the M gene sequence of KJ481931 (infecting humans) was more similar to that of M gene sequences from SARS-CoV-2 isolates and C-CoVs that infect other organisms (**Figure 5F**).

The N gene sequences of SARS-CoV-2 isolates were closely related and grouped together within the same branch of a phylogenetic tree (**Figures 5G,H**). The N gene sequence of NC014470 was an evolutionary intermediate between SARS-CoV-2 isolates and C-CoVs that infect humans (**Figures**

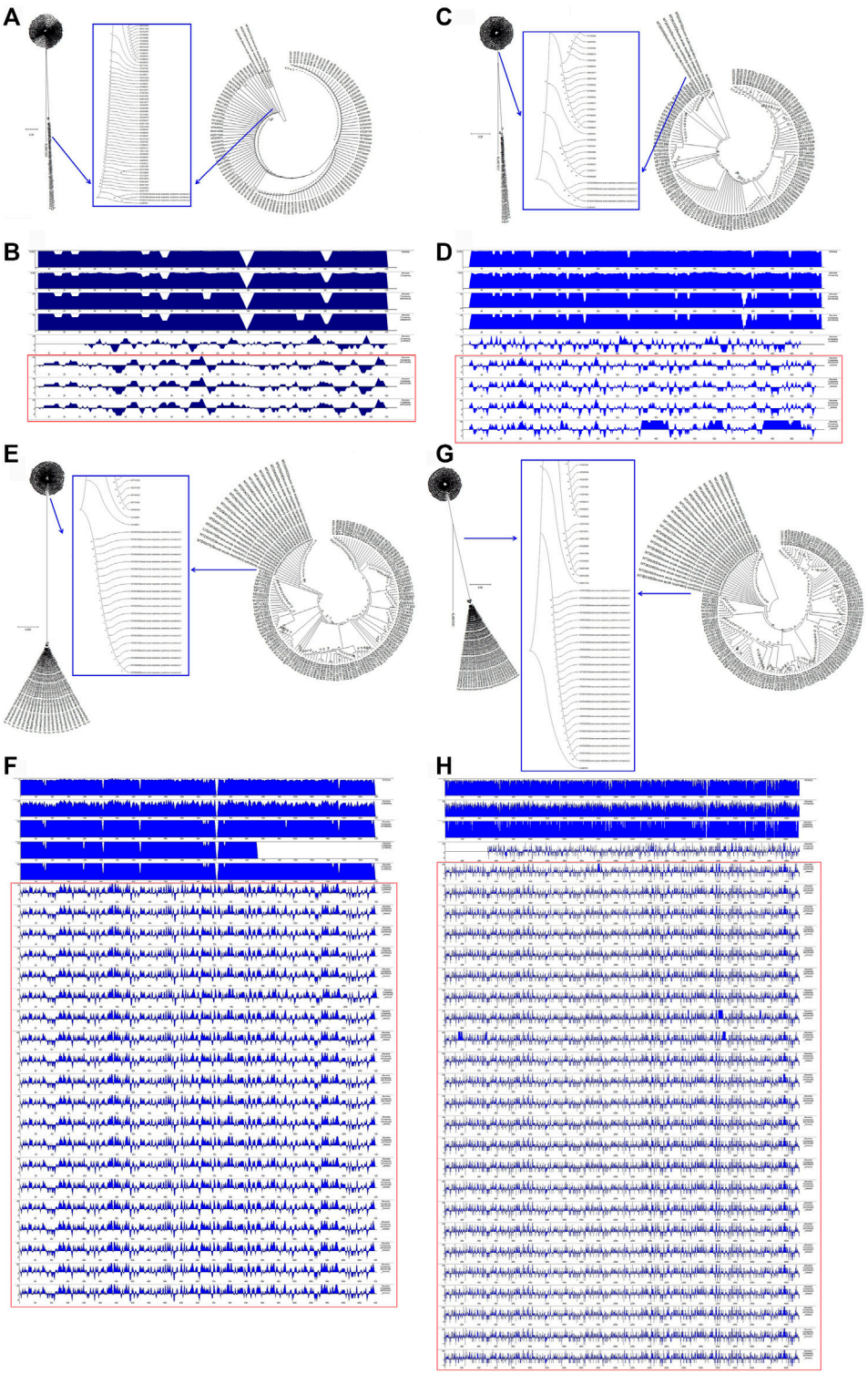

FIGURE 3 | Evolutionary characteristics and absolute diversity of structural genes in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates and common coronaviruses (C-CoVs) that infect humans. **(A)**: The E gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between KJ481931 and MG011357. **(C)**: The M gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between KJ48193 and a group of C-CoVs (KP209309, KY581691, KY581689, KY581686, KP209307, KP209313, and KP209306). **(B,D)**: The absolute diversities of the E and M gene sequences within the KJ481931 C-CoV were similar to those of the E and M gene sequences of SARS-CoV-2 isolates. **(E,G)**: The N and S gene sequences of SARS-CoV-2 isolates were evolutionarily distinct. **(F,H)**: The absolute diversities of the N and S gene sequences of SARS-CoV-2 isolates differed from those of all C-CoVs that infect humans.

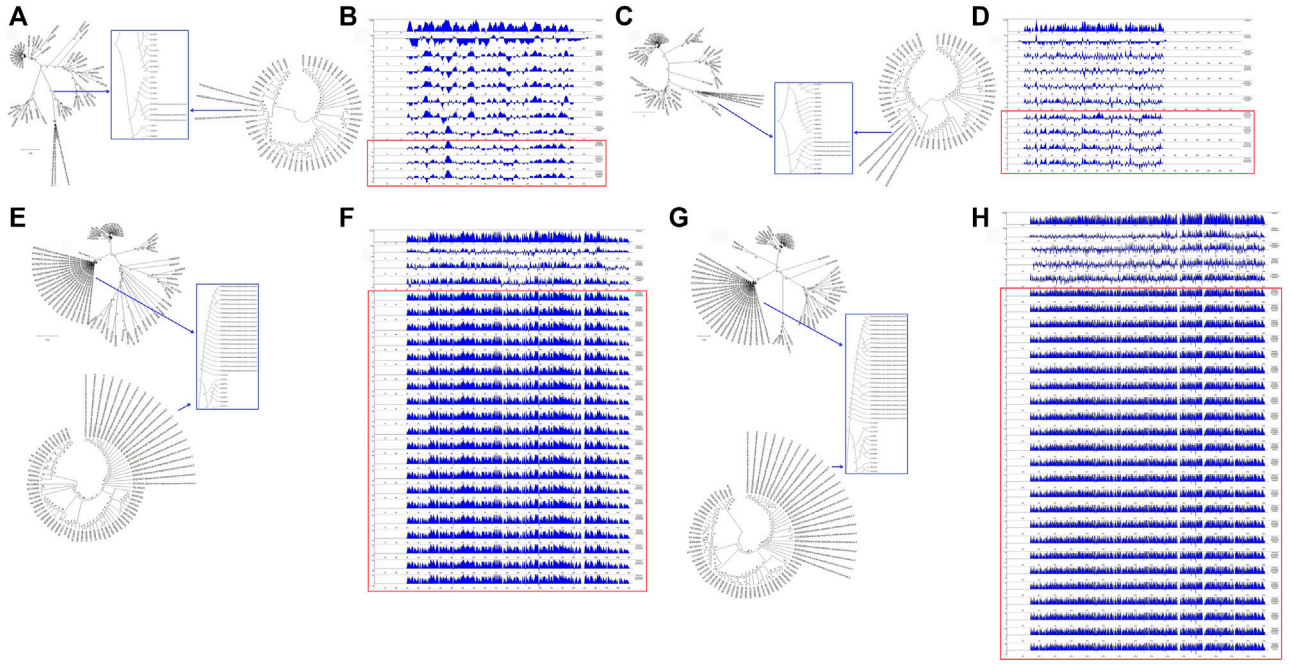

FIGURE 4 | Evolutionary characteristics and absolute diversities of structural genes in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates and common coronaviruses (C-CoVs) that infect other organisms. **(A,C)**: The E and M gene sequences of SARS-CoV-2 isolates were evolutionarily intermediates between C-CoVs that infect other organisms. **(A)**: In terms of E gene sequences, SARS-CoV-2 isolates were most closely related to the C-CoVs NC014470, DQ415914, NC026011, NC006213, JN874559 and U00735. **(B)**: The absolute diversities of the E gene sequences of the C-CoVs NC014470, DQ415914, NC026011, NC006213, JN874559, and U00735 were similar to those of the E gene sequences of SARS-CoV-2 isolates. **(C)**: In terms of M gene sequences, SARS-CoV-2 isolates were most closely evolutionarily related to the C-CoVs NC014470, EF065513 and NC030886. **(D)**: The absolute diversities of the M gene sequences of the C-CoVs NC014470, EF065513 and NC030886 were similar to those of M gene sequences of SARS-CoV-2 isolates. **(E,G)**: In terms of N and S gene sequences, SARS-CoV-2 isolates were most closely evolutionarily related to the C-CoV NC014470, forming a separate clade. **(F)**: The absolute diversity of N gene sequences of SARS-CoV-2 isolates was similar to that of the C-CoV NC014470. **(H)**: The absolute diversity of the S gene sequence of the C-CoV NC014470 was similar to those of other C-CoVs.

TABLE 2 | Analysis of structural gene sequences of common coronaviruses (C-CoVs) evolutionarily related to severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2).

Genes	C-CoVs infecting humans	C-CoVs other organisms
E gene	KJ481931, MG011357	NC014470, DQ415914, NC026011, NC006213, JN874559, U00735
M gene	KJ481931, KP209309, KY581691, KY581689, KY581686, KP209307, KP209313, KP209306	NC014470, EF065513, NC030886
N gene	KJ156911, KJ156905	NC014470
S gene	KJ481931, MG011344	NC014470

5G,H). The absolute diversity of the N gene sequences of SARS-CoV-2 isolates differed from the absolute diversity of the N gene sequences of C-CoVs (**Figure 5I**).

The evolutionary features and absolute diversities of the S gene sequences of SARS-CoV-2 isolates and C-CoVs that infect other organisms or humans were very similar to those of the N gene sequences (**Figures 5J–L**).

DISCUSSION

Genetic information determines the functions and characteristics of biological factors and organisms. Gene annotation and evolutionary analysis are important steps in interpreting sequence information (Khailany et al., 2020). In this work, we profiled variations in the structural gene sequences of SARS-CoV-2 isolates. We analyzed the evolutionary characteristics and absolute diversities of structural gene sequences of SARS-CoV-2 isolates and C-CoVs that infect humans and other organisms.

CoVs are positive-single-stranded RNA viruses. The major symptoms caused by CoV infection are respiratory tract infections. SARS-CoV, Middle East Respiratory Syndrome (MERS)-CoV and SARS-CoV-2 are three highly contagious and deadly CoVs that have caused outbreaks in humans (Singh Tomar and Arkin 2020). The genomes of SARS-CoV and SARS-CoV-2 share approximately 80% identity, but are

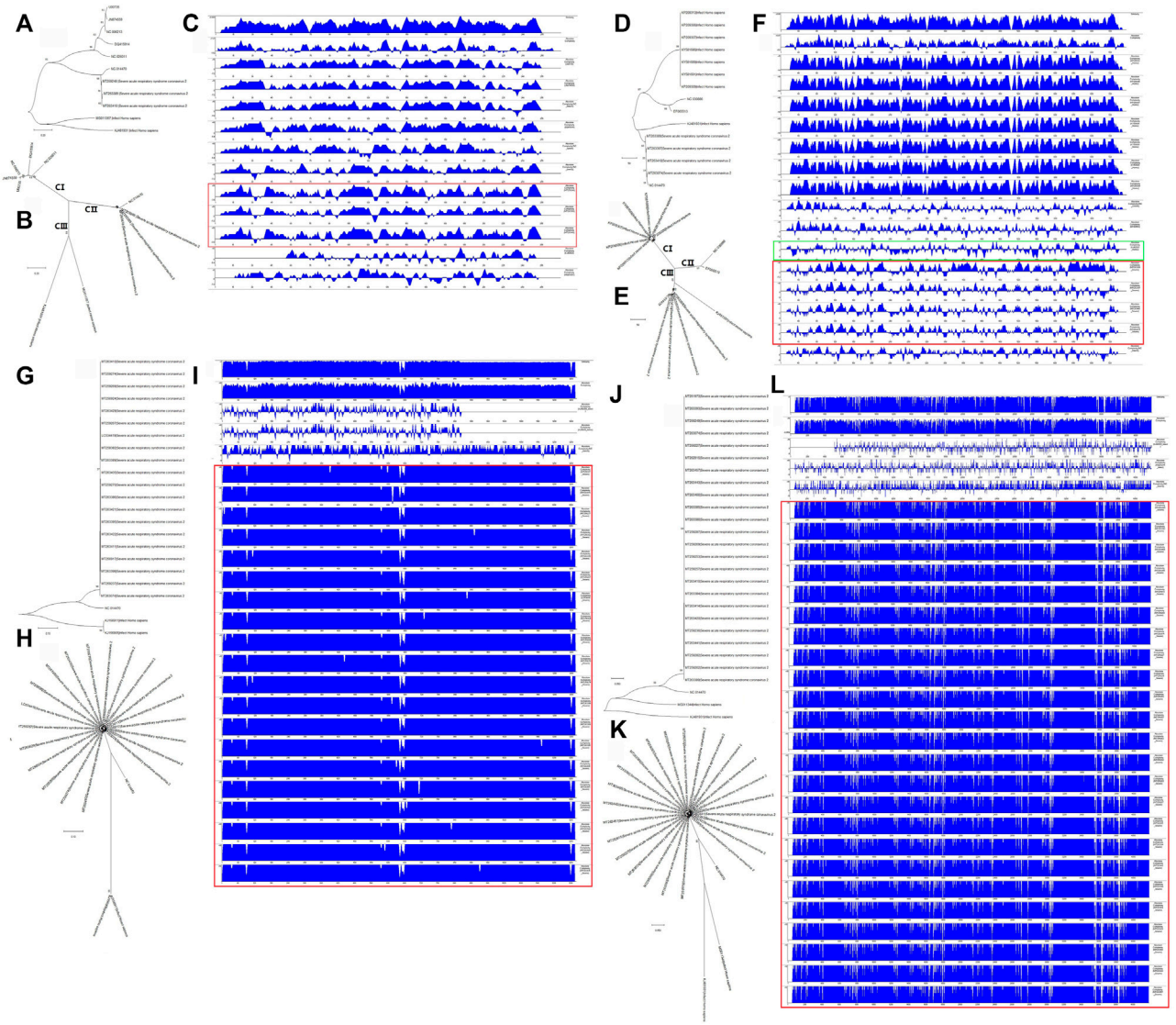

FIGURE 5 | Evolutionary characteristics and absolute diversities of structural genes of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates and common coronaviruses (C-CoVs). **(A,B,D,E)**: The E and M gene sequences of SARS-CoV-2 isolates and common CoVs could be grouped into three clades (CI, CII and CIII). A, B: The E gene sequences of SARS-CoV-2 isolates and C-CoV NC014470 were evolutionary intermediates between C-CoVs that infect humans and other organisms. **(C)**: The absolute diversities of E gene sequences of SARS-CoV-2 isolates were more similar to those of C-CoVs that infect other organisms. **(D,E)**: The M gene sequences of SARS-CoV-2 isolates, NC014470 (infecting other organisms) and KJ481931 (infecting humans) were closely related and grouped together in the same branch of a phylogenetic tree. The M gene sequences of SARS-CoV-2 isolates were evolutionary intermediates between those of NC014470 (infecting other organisms) and KJ481931 (infecting humans). **(F)**: The absolute diversities of M gene sequences of SARS-CoV-2 isolates were more similar to those of C-CoVs infecting other organisms. **(F)** (Green box): The absolute diversity of M gene sequences of KJ481931 (infecting humans) was more similar to those of M gene sequences from SARS-CoV-2 isolates and C-CoVs that infect other organisms. **(G,H,J,K)**: The N and S gene sequences of SARS-CoV-2 strains grouped closely together on the same branch of an evolutionary tree. **(G,H,J,K)**: The N and S gene sequences of NC014470 were located between those of SARS-CoV-2 isolates and C-CoVs that infect humans. **(I,L)**: The absolute diversities of N and S gene sequences of SARS-CoV-2 isolates were unlike those of C-CoVs.

distinct from those of other C-CoVs that infect humans (Lu et al., 2020, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2 2020).

The SARS-CoV-2 genome including four structural genes encoding structural proteins: E, M, S and N (Comas-Garcia 2019). The functions of the E protein include assembly, release, and pathogenesis of CoVs (Schoeman and Fielding 2019). Important features of the E gene and protein are their small size and the high hydrophobicity of the E protein. Those features suggests that the E protein may act as a viroporin, and that CoVs lacking the E protein may be less virulent. The E protein many serve as a vaccine candidate (Fett et al., 2013; Regla-Nava et al., 2015). In this work, using the genome sequences of 200 SARS-CoV-2 isolates, we found that only two isolates had SNPs within the E gene. The sequence of the E

gene was the most highly conserved across the 200 SARS-CoV-2 isolates.

The genomes of many CoVs contain an E gene, including SARS-CoV (Torres et al., 2006; Parthasarathy et al., 2008), MERS-CoV (Surya et al., 2015), human CoV 229E (Wilson et al., 2006), and SARS-CoV-2. In terms of their E gene sequences, we found the SARS-CoV-2 was most closely evolutionarily related to NC014470 [a C-CoV that infects bats (Drexler et al., 2010)]; these two CoVs were evolutionary intermediates between C-CoVs that infect humans and those that infect other organisms. The absolute diversity of the E gene sequences of SARS-CoV-2 isolates was more similar to that of E gene sequences from C-CoVs that infect other organisms.

The genetic and evolutionary features of M gene sequences within the 200 SARS-CoV-2 isolates were very similar to those of E gene sequences. As a major envelope protein, the M protein is responsible for viral envelope formation and virion assembly (Ujike and Taguchi 2015; Jacofsky et al., 2020). Here, we found that nine of 200 isolates showed variations (one mutation and eight SNPs) in the M gene. The VR and GSVR of the M gene were slightly higher than those of the E gene. However, the M protein is a major envelope protein (Ujike and Taguchi 2015), and the mutation (Val70→Ile) in the M gene of MT263397 had little impact on the transmembrane segment of the M protein. The M gene and protein is another good candidate for SARS-CoV-2 vaccine development.

The evolutionary features of M gene sequences were very interesting. The M gene sequences of SARS-CoV-2 isolates were evolutionary intermediate between those of NC014470 (infecting other organisms) and KJ481931 [infecting humans; (Marthaler et al., 2014)]; the M gene sequences of these CoVs were grouped closely together within the same branch of a phylogenetic tree. The absolute diversity of the M gene sequence from KJ481931 was more similar to that of M gene sequences from SARS-CoV-2 isolates and to those of M gene sequences of C-CoVs that infect other organisms.

During CoV infection, the N protein and viral RNA enter host cells together, where they are involved in viral assembly, release and genome replication (Narayanan et al., 2003). In the early stages of infection, antibodies against the N protein are highly specific (Shi et al., 2003; Leung et al., 2004; Tan et al., 2004). In this study, we found that 28 of 200 SARS-CoV-2 isolates showed a total of 22 variations within the N gene. Mutations mainly occurred between amino acid residues 200 to 300 and had a large impact on N protein function.

The genetic and evolutionary features of N and S structural genes within the 200 SARS-CoV-2 isolates were very similar. The VRs of N and S genes were 14 and 44.5%, respectively. However, the S gene sequence is longer than the N gene sequence (Khailany et al., 2020). The GSVR of the S gene was 1.16/10,000, lower than that of the N gene (1.54/10,000). We identified 58 isolates bearing the same variation (Asp614→Gly), but mutations in the S gene had little effect on protein function. The N gene sequence was less conserved than the S gene sequence.

The main function of the S protein is to mediate CoV entry into host cells (Tortorici and Veesler 2019). Among the four structural proteins, the S protein is the largest (Khailany et al., 2020). In the S protein, SARS-CoV-2 and SARS-CoV share 76% amino acid identity (de Groot 2006; Zhang et al., 2020a). Entry of SARS-CoV-2 into host cells can be prevented by antibodies raised against SARS-CoV (Hoffmann et al., 2020). The S protein of SARS-CoV-2 shared 93 and 97% amino acid identity with Bat CoV RaTG13 and Pangolin-CoV, respectively (Zhang et al., 2020b; Special Expert Group for Control of the Epidemic of Novel Coronavirus Pneumonia of the Chinese Preventive Medicine Association, 2020; Zhou et al., 2020). These results strongly suggest potential intermediate hosts based on conservation of the S protein. However, in our study we found that S gene sequences of SARS-CoV-2 isolates were evolutionarily independent in a phylogenetic tree, with a relatively large evolutionary distance separating the S genes of SARS-CoV-2 and C-CoVs. The absolute diversity of S gene sequences within SARS-CoV-2 isolates was also unlike those of S genes sequences from all the other C-CoVs.

CONCLUSION

On the basis of these results, we conclude that the E and M structural genes of SARS-CoV-2 and the NC014470 and KJ481931 CoVs are important for understanding the origins and intermediate hosts of SARS-CoV-2.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

Study conception and design: F-FL, QZ and P-JD. Data collection and analysis: F-FL, QZ, H-LG, JW, YZ, and P-JD. Funding: F-FL; drafting/revision of the manuscript: all authors.

FUNDING

This work was supported by grants from the Cultivating Project for Young Scholars at Hubei University of Medicine (No. 2020QDJZR025 to F-FL), the National Natural Science Foundation of China (No. 81372998 to H-LG) and the Special Emergency Research Project on COVID-19 at Hubei University of Medicine (No. 2020XGFYZR04 to JW).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.801902/full#supplementary-material

Supplementary Figure S1 | Influence of a mutation in the M gene (g.208G→A, p.70Val→Ile) of the MT263397 coronavirus on protein structure and function. The mutation had little effect on the transmembrane segment of the M protein.

Supplementary Figure S2 | Influence of mutations in the N genes of 200 severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates on protein structure and function. Six mutations (MT259237, p.3Asp→Tyr; MT263429, p.194Ser→Leu; MT256917 and MT256918, p.197Ser→Leu; MT263410, p.208Ala→Gly; MT256924, p.238Gly→Cys; and MT263422, p.271Thr→Ile) had effects on protein hydrophobicity. Three mutations (MT259237, p.3Asp→Tyr; MT263429, p.194Ser→Leu; MT256917 and MT256918, p.197Ser→Leu) had effects on protein hydrophilicity. Ten mutations (MT259237, p.3Asp→Tyr; MT259274, p.35Ala→Thr; MT263429, p.194Ser→Leu; MT256917 and MT256918, p.197Ser>→Leu; MT263386, p.202Ser→Asn; MT258379, MT259250, MT259263, MT263074, and MT263402, p.203Arg→Lys and p.204Gly→Arg; MT263411, p.232Ser→Thr; MT256924, p.238Gly→Cys; MT263422, p.271Thr→Ile; and LC534419, p.289Gln→His) had effects on protein secondary structure. Four mutations (MT259237, p.3Asp→Tyr; MT263386, p.202Ser→Asn; MT258379, MT259250, MT259263, MT263074, and MT263402, p.203Arg→Lys and p.204Gly→Arg; and MT263422, p.271Thr→Ile) had effects on protein transmembrane segments.

Supplementary Figure S3 | Influence of mutations in the S genes of 200 severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) isolates on protein structure and function. One mutation in the S gene affected protein hydrophobicity (MT263384, p.29Thr→Ile) and one mutation affected hydrophilicity (MT251973 and MT251976, p.614Asp→Gly). Three mutations (MT259253, p.348Ala→Thr; MT263466, p.1258Glu→Asp and p.1259Asp→His; MT263385, MT263387, p.1259Asp→His) in the S gene had effects on protein secondary structure.

REFERENCES

Ashour, H. M., Elkhatib, W. F., Rahman, M. M., and Elshabrawy, H. A. (2020). Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. *Pathogens* 9, 9. doi:10.3390/pathogens9030186

Ceraolo, C., and Giorgi, F. M. (2020). Genomic Variance of the 2019-nCoV Coronavirus. *J. Med. Virol.* 92, 522–528. doi:10.1002/jmv.25700

Comas-Garcia, M. (2019). Packaging of Genomic RNA in Positive-Sense Single-Stranded RNA Viruses: A Complex Story. *Viruses* 11, 11. doi:10.3390/v11030253

de Groot, R. J. (2006). Structure, Function and Evolution of the Hemagglutinin-Esterase Proteins of corona- and Toroviruses. *Glycoconj J.* 23, 59–72. doi:10.1007/s10719-006-5438-8

Drexler, J. F., Gloza-Rausch, F., Glende, J., Corman, V. M., Muth, D., Goettsche, M., et al. (2010). Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-dependent RNA Polymerase Gene Sequences. *J. Virol.* 84, 11336–11349. doi:10.1128/jvi.00650-10

Fett, C., DeDiego, M. L., Regla-Nava, J. A., Enjuanes, L., and Perlman, S. (2013). Complete protection against Severe Acute Respiratory Syndrome Coronavirus-Mediated Lethal Respiratory Disease in Aged Mice by Immunization with a Mouse-Adapted Virus Lacking E Protein. *J. Virol.* 87, 6551–6559. doi:10.1128/jvi.00087-13

Gorbalenya, A. E., Baker, S., Baric, R., and de Groot, R. J. (2020). The Species Severe Acute Respiratory Syndrome-Related Coronavirus: Classifying 2019-nCoV and Naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536–544. doi:10.1038/s41564-020-0695-z

Guo, W. L., Jiang, Q., Ye, F., Li, S. Q., Hong, C., Chen, L. Y., et al. (2020). Effect of Throat Washings on Detection of 2019 Novel Coronavirus. *Clin. Infect. Dis.* 71, 1980–1981. doi:10.1093/cid/ciaa416

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271–280. e278. doi:10.1016/j.cell.2020.02.052

Jacofsky, D., Jacofsky, E. M., and Jacofsky, M. (2020). Understanding Antibody Testing for COVID-19. *The J. Arthroplasty* 35, S74–S81. doi:10.1016/j.arth.2020.04.055

Jiang, S., and Shi, Z. L. (2020). The First Disease X Is Caused by a Highly Transmissible Acute Respiratory Syndrome Coronavirus. *Virol. Sin* 35, 263–265. doi:10.1007/s12250-020-00206-5

Khailany, R. A., Safdar, M., and Ozaslan, M. (2020). Genomic Characterization of a Novel SARS-CoV-2. *Gene Rep.* 19, 100682. doi:10.1016/j.genrep.2020.100682

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096

Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J., and Hsueh, P.-R. (2020). Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus Disease-2019 (COVID-19): The Epidemic and the Challenges. *Int. J. Antimicrob. Agents* 55, 105924. doi:10.1016/j.ijantimicag.2020.105924

Leung, D. T. M., Tam, F. C. H., Ma, C. H., Chan, P. K. S., Cheung, J. L. K., Niu, H., et al. (2004). Antibody Response of Patients with Severe Acute Respiratory Syndrome (SARS) Targets the Viral Nucleocapsid. *J. Infect. Dis.* 190, 379–386. doi:10.1086/422040

Li F, F.-F., Zhang, Q., Wang, G.-Y., and Liu, S.-L. (2020). Comparative Analysis of SARS-CoV-2 and its Receptor ACE2 with Evolutionarily Related Coronaviruses. *Aging* 12, 20938–20945. doi:10.18632/aging.104024

Li, F. F., Yan, P., Zhao, Z. X., Liu, Z., Song, D. W., Zhao, X. W., et al. (2016). Polymorphisms in the CHIT1 Gene: Associations with Colorectal Cancer. *Oncotarget* 7 (7), 39572–39581. doi:10.18632/oncotarget.9138

Li X, X., Zai, J., Zhao, Q., Nie, Q., Li, Y., Foley, B. T., et al. (2020). Evolutionary History, Potential Intermediate Animal Host, and Cross-Species Analyses of SARS-CoV-2. *J. Med. Virol.* 92 (6), 602–611. doi:10.1002/jmv.25731

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. (2020). Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding. *Lancet* 395 (395), 565–574. doi:10.1016/S0140-6736(20)30251-8

Ma, D., Chen, C.-B., Jhanji, V., Xu, C., Yuan, X.-L., Liang, J.-J., et al. (2020). Expression of SARS-CoV-2 Receptor ACE2 and TMPRSS2 in Human Primary Conjunctival and Pterygium Cell Lines and in Mouse Cornea. *Eye* 34, 1212–1219. doi:10.1038/s41433-020-0939-4

Marthaler, D., Jiang, Y., Collins, J., and Rossow, K. (2014). Complete Genome Sequence of Strain SDCV/USA/Illinois121/2014, a Porcine Deltacoronavirus from the United States. *Genome Announc* 2, 2. doi:10.1128/genomeA.00218-14

Mousavizadeh, L., and Ghasemi, S. (2020). Genotype and Phenotype of COVID-19: Their Roles in Pathogenesis. *J. Microbiol. Immunol. Infect.* 54 (2), 159–163. doi:10.1016/j.jmii.2020.03.022

Narayanan, K., Chen, C.-J., Maeda, J., and Makino, S. (2003). Nucleocapsid-independent Specific Viral RNA Packaging via Viral Envelope Protein and Viral RNA Signal. *J. Virol.* 77, 2922–2927. doi:10.1128/jvi.77.5.2922-2927.2003

Parthasarathy, K., Ng, L., Lin, X., Liu, D. X., Pervushin, K., Gong, X., et al. (2008). Structural Flexibility of the Pentameric SARS Coronavirus Envelope Protein Ion Channel. *Biophys. J.* 95 (95), L39–L41. doi:10.1529/biophysj.108.133041

Regla-Nava, J. A., Nieto-Torres, J. L., Jimenez-Guardeño, J. M., Fernandez-Delgado, R., Fett, C., Castaño-Rodríguez, C., et al. (2015). Severe Acute Respiratory Syndrome Coronaviruses with Mutations in the E Protein Are Attenuated and Promising Vaccine Candidates. *J. Virol.* 89, 3870–3887. doi:10.1128/jvi.03566-14

Schoeman, D., and Fielding, B. C. (2019). Coronavirus Envelope Protein: Current Knowledge. *Virol. J.* 16, 1669. doi:10.1186/s12985-019-1182-0

Shi, Y., Yi, Y., Li, P., Kuang, T., Li, L., Dong, M., et al. (2003). Diagnosis of Severe Acute Respiratory Syndrome (SARS) by Detection of SARS Coronavirus Nucleocapsid Antibodies in an Antigen-Capturing Enzyme-Linked Immunosorbent Assay. *J. Clin. Microbiol.* 41, 5781–5782. doi:10.1128/jcm.41.12.5781-5782.2003

Singh Tomar, P. P., and Arkin, I. T. (2020). SARS-CoV-2 E Protein Is a Potential Ion Channel that Can Be Inhibited by Gliclazide and Memantine. *Biochem. Biophys. Res. Commun.* 530, 10–14. doi:10.1016/j.bbrc.2020.05.206

Special Expert Group for Control of the Epidemic of Novel Coronavirus Pneumonia of the Chinese Preventive Medicine Association (2020). An Update on the Epidemiological Characteristics of Novel Coronavirus

pneumoniaCOVID-19. *Zhonghua Liu Xing Bing Xue Za Zhi.* 41, 139–144. doi:10.3760/cma.j.issn.0254-6450.2020.02.002

Surya, W., Li, Y., Verdia-Baguena, C., Aguilella, V. M., and Torres, J. (2015). MERS Coronavirus Envelope Protein Has a Single Transmembrane Domain that Forms Pentameric Ion Channels. *Virus. Res.* 201, 61–66. doi:10.1016/j.virusres.2015.02.023

Tan, Y. J., Goh, P. Y., Fielding, B. C., Shen, S., Chou, C. F., Fu, J. L., et al. (2004). Profiles of Antibody Responses against Severe Acute Respiratory Syndrome Coronavirus Recombinant Proteins and Their Potential Use as Diagnostic Markers. *Clin. Diagn. Lab. Immunol. Mar.* 11, 362–371. doi:10.1128/cdli.11.2.362-371.2004

Torres, J., Parthasarathy, K., Lin, X., Saravanan, R., Kukol, A., and Liu, D. X. (2006). Model of a Putative Pore: the Pentameric Alpha-Helical Bundle of SARS Coronavirus E Protein in Lipid Bilayers. *Biophys. J.* 91, 938–947. doi:10.1529/biophysj.105.080119

Tortorici, M. A., and Veesler, D. (2019). Structural Insights into Coronavirus Entry. *Adv. Virus. Res.* 105, 93–116. doi:10.1016/bs.aivir.2019.08.002

Tyrrell, D. A., and Bynoe, M. L. (1966). Cultivation of Viruses from a High Proportion of Patients with Colds. *Lancet* 1, 76–77. doi:10.1016/s0140-6736(66)92364-6

Ujike, M., and Taguchi, F. (2015). Incorporation of Spike and Membrane Glycoproteins into Coronavirus Virions. *Viruses. Apr* 3 (7), 1700–1725. doi:10.3390/v7041700

Velavan, T. P., and Meyer, C. G. (2020). The COVID-19 Epidemic. *Trop. Med. Int. Health Mar.* 25, 278–280. doi:10.1111/tmi.13383

Wilson, L., Gage, P., and Ewart, G. (2006). Hexamethylene Amiloride Blocks E Protein Ion Channels and Inhibits Coronavirus Replication. *Virol. Sep* 30353, 294–306. doi:10.1016/j.virol.2006.05.028

Zhang, T., Wu, Q. F., and Zhang, Z. G. (2020c). *Pangolin Homology Associated with 2019-nCoV*. bioRxiv. doi:10.1101/2020.02.19.950253

Zhang, T., Wu, Q., and Zhang, Z. (2020a). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* 30, 1346–1351. e1342. doi:10.1016/j.cub.2020.03.022

Zhang, T., Wu, Q., and Zhang, Z. (2020b). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* 30, 1578. doi:10.1016/j.cub.2020.03.022

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nat. Mar* 579, 270–273. doi:10.1038/s41586-020-2012-7

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership