COMPUTATIONAL METHODS IN PREDICTING COMPLEX DISEASE ASSOCIATED GENES AND ENVIRONMENTAL FACTORS

EDITED BY: Yudong Cai, Jialiang Yang, Tao Huang and Minxian Wallace Wang
PUBLISHED IN: Frontiers in Genetics, Frontiers in Neuroscience and Frontiers in Physiology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-875-5
DOI 10.3389/978-2-88966-875-5

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

COMPUTATIONAL METHODS IN PREDICTING COMPLEX DISEASE ASSOCIATED GENES AND ENVIRONMENTAL FACTORS

Topic Editors:
Yudong Cai, Shanghai University, China
Jialiang Yang, Geneis (Beijing) Co. Ltd, China
Tao Huang, Shanghai Institute of Nutrition and Health (CAS), China
Minxian Wallace Wang, Broad Institute, United States

Citation: Cai, Y., Yang, J., Huang, T., Wang, M. W., eds. (2021). Computational Methods in Predicting Complex Disease Associated Genes and Environmental Factors. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88966-875-5

Table of Contents

frontiers in Genetics

EDITORIAL
published: 21 April 2021
doi: 10.3389/fgene.2021.679651

Editorial: Computational Methods in Predicting Complex Disease Associated Genes and Environmental Factors

Yudong Cai[1], *Jialiang Yang*[2], *Tao Huang*[3]* *and Minxian Wallace Wang*[4]

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Geneis (Beijing) Co. Ltd., Beijing, China, [3] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai, China, [4] Medical and Population Genetics Program, Broad Institute of MIT and Harvard, Cambridge, MA, United States

Keywords: computational method, complex disease, environmental factor, disease gene, genetic factor, epigenetic factor

Editorial on the Research Topic

Computational Methods in Predicting Complex Disease Associated Genes and Environmental Factors

With the advances of sequencing and experimental techniques, the molecular mechanisms of human Mendelian diseases have been more or less elucidated. However, there are also many complex diseases whose disease/pathology development involves the interaction of large numbers of biomolecules across multi-molecular levels including DNA, RNA, proteins, and methylation, as well as the impact of environmental and human lifestyle factors. The understanding of such diseases is one of the biggest challenges in modern biology and medical sciences. The progress in this field will shed light on complex disease pathology, prevention, prognosis, diagnosis, and treatment in a personalized manner.

In recent years, large amounts of data from human genome sequencing, metagenome sequencing, and information about the impact of environmental and lifestyle factors on complex diseases have been produced, collected, and stored in large scale databases such as the National Alzheimer's Coordinating Center (NACC) database, the database of Genotypes and Phenotypes (dbGaP) and UK Biobank. The large amount of data poses a big challenge, as well as a great opportunity, to reveal the secrets behind complex diseases using machine learning, statistics, and bioinformatics tools along with validation through experimental work. In fact, many computational studies have already been performed within this research area; however, most are focused on disease-associated factors at a single-molecular level, such as genetic factors, epigenetic factors, environmental factors, and so on. A more systematic study on the interactions among these factors, alongside experimental validation, might present a comprehensive view on the disease pathogenicity and thus may hold the key to truly understanding complex diseases.

In this special issue, there are 18 studies of complex diseases.

Li et al. compared the gene expression profiles between patients with heart failure ($n = 177$) and without heart failure ($n = 136$) using multiple feature selection strategies and identified 38 HF signature genes. Their results can facilitate the early detection of heart failure and can reveal its molecular mechanisms.

Liang et al. proposed a novel antiviral Drug Repositioning method based on minimizing Matrix Nuclear Norm (DRMNN). Experiments have shown that DRMNN is better than other algorithms in predicting which drugs are effective against influenza A virus. Within the 10 drugs most likely to be effective against H3N2 viruses, six drugs are reported to have some effect on the viruses.

OPEN ACCESS

Edited and reviewed by:
James J. Cai,
Texas A&M University, United States

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

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

Received: *12 March 2021*
Accepted: *26 March 2021*
Published: *21 April 2021*

Citation:
Cai Y, Yang J, Huang T and Wang MW (2021) Editorial: Computational Methods in Predicting Complex Disease Associated Genes and Environmental Factors. Front. Genet. 12:679651. doi: 10.3389/fgene.2021.679651

Liu et al. recruited 20 patients undergoing cardiac surgery (10 with paroxysmal atrial fibrillation and 10 with persistent atrial fibrillation) and 10 healthy subjects. With proteomic analysis, they identified the differentially expressed proteins and investigated their roles in Atrial fibrillation (AF).

Li et al. developed a set of computational approaches integrating multiple machine-learning algorithms, including Monte Carlo feature selection (MCFS), incremental feature selection (IFS), and support vector machine (SVM), to identify gene expression characteristics on different phases of Myocardial infarction (MI). The functional enrichment analyses followed by protein-protein interaction analysis identified several hub genes (IL1R1, TLR2, and TLR4) which may be new diagnostic molecules for MI.

Su et al. described a convolutional neural network called F-S-Net that fused the information from multimodal medical images and used the semantic information contained within these images for glioma segmentation. F-S-Net was found to achieve a dice coefficient of 0.9052 and Jaccard similarity of 0.8280, outperforming several previous segmentation methods.

Wang et al. screened the genes associated with neuropathic pain (NP) using differential analysis along with random walk with restart (RWR). They discovered eight hub genes that were closely related to NP occurrence and development, which may help to provide potent theoretical basis for NP treatment.

Hao et al. integrated four gene expression datasets which collectively included 65 nasal polyp samples from Chronic rhinosinusitis with nasal polyps (CRSwNP) patients and 54 nasal mucosal samples from healthy controls. They identified 76 co-differentially expressed genes (co-DEGs, including 45 upregulated and 31 downregulated) in CRSwNP patients compared with the healthy controls. Protein-protein interaction (PPI) network analysis and real-time quantitative PCR (RT-qPCR) showed that seven genes might be crucial in CRSwNP pathogenesis.

Guan et al. constructed cell type-specific predictive models for autism spectrum disorder (ASD) based on individual genes and gene sets, respectively, to screen cell type-specific ASD-associated genes and gene sets. They found that the functions of genes with predictive power for ASD were different and the top important genes were distinct across different cells, highlighting the cell-type heterogeneity of ASD.

Zhu et al. proposed and compared 10 protein–protein interaction (PPI)-based computational methods to study the connections between diabetes and 254 diseases. They found that a method called DIconnectivity_eDMN performed the best in the sense that it inferred a disease rank (according to its relation with diabetes) most consistent with that by literature mining.

Zhang et al. analyzed the blood gene expression profiles of 73 Caucasian women with high and low bone mineral density (BMD). The WGCNA yielded three gene modules, including 26 lncRNAs and 55 mRNAs as hub genes in the blue module, 36 lncRNAs and 31 mRNAs as hub genes in the turquoise module, and 56 mRNAs and 30 lncRNAs as hub genes in the brown module. The mRNAs and lncRNAs identified in this WGCNA could be novel clinical targets in the diagnosis and management of osteoporosis.

Sun et al. proposed a mathematical model based on matrix decomposition, named MFMDA, to identify potential miRNA–disease associations by integrating known miRNA and disease-related data, similarities between miRNAs and between diseases. While most predicted miRNAs were confirmed by external databases of experimental literature, they also identified a few novel disease-related miRNAs for further experimental validation.

Wang et al. identify the key modules and hub genes related to the annulus fibrosus in intervertebral disc degeneration (IDD) through: (1) constructing a weighted gene co-expression network; (2) identifying key modules and hub genes; (3) verifying the relationships of key modules and hub genes with IDD; and (4) confirming the expression pattern of hub genes in clinical samples. They generated a comprehensive overview of the gene networks underlying annulus fibrosus in intervertebral disc degeneration.

Wang et al. proposed a new method called Matrix completion algorithm based on q-kernel information (QIMCMDA) for miRNA-disease association prediction. Its performance was significantly better than other commonly used technologies. QIMCMDA may become an excellent supplement in the field of biomedical research in the future.

Liu et al. proposed a novel network inference algorithm using Random Walk with Restart (RWRNET) that combined local and global topology relationships. The proposed method was compared with several state-of-the-art methods on the basis of six benchmark datasets and the results demonstrated the effectiveness of the proposed method.

An et al. evaluated the pharmacological effects of novel peptide drugs (P-ONE and P-TWO) at the small RNA (sRNA) level using an allergic rhinitis (AR) model. They found that sRNA target genes had a specific enrichment pattern and may contribute to the effects of the novel peptides.

Gao et al. identified orphan genes in balanced and unbalanced *Arabidopsis thaliana* gene datasets. They compared several ensemble models and found that SMOTE-ENN-XGBoost model, which combined over-sampling and under-sampling algorithms with XGBoost, achieved higher predictive accuracy than the other balanced algorithms with XGBoost models. Thus, SMOTE-ENN-XGBoost provided a theoretical basis for developing evaluation criteria for identifying orphan genes in unbalanced and biological datasets.

Wei et al. developed a machine learning method to classify colon and rectal cancer into three immune subtypes named High-Immunity Subtype, Medium-Immunity Subtype, and Low-Immunity Subtype, respectively. A prognostic signature of six genes (CERCAM, CD37, CALB2, MEOX2, RASGRP2, and PCOLCE2) was identified by the multivariable COX analysis, which was further used to develop an accurate model to predict the prognosis of colon and rectal cancer patients.

Chidambaran et al. recruited 171 adolescents (14.5 ± 1.8 years, 75.4% female) undergoing spine fusion and tested ranked deciles of 1,336 prioritized genes for increased representation of variants associated with chronic postsurgical pain (CPSP). Penalized regression (LASSO) selected 20 variants for calculating weighted polygenic risk scores (PRS). Systems

biology guided PRS improved predictive accuracy of CPSP risk in a pediatric cohort.

In recent years, there are more and more studies of complex diseases using computational methods on multi omics data. By integrating genetic factors, epigenetic factors, environmental factors, and so on, the underlying mechanisms of complex diseases may be revealed and we may find the cures.

AUTHOR CONTRIBUTIONS

TH wrote the editorial and all authors have approved the submission.

Conflict of Interest: JY is the Vice President of Geneis (Beijing) Co. Ltd.

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

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 20 May 2020
doi: 10.3389/fnins.2020.00488

Identification of the Hub Genes Related to Nerve Injury-Induced Neuropathic Pain

Kai Wang[1†], Duan Yi[1†], Zhuoyin Yu[2], Bin Zhu[1], Shuiqing Li[1] and Xiaoguang Liu[3]*

[1] Department of Pain Medicine Center, Peking University Third Hospital, Beijing, China, [2] Department of Anesthesiology, Peking University Third Hospital, Beijing, China, [3] Department of Orthopedic, Peking University Third Hospital, Beijing, China

Background: The reactivity enhancement of pain sensitive neurons in the nervous system is a feature of the pathogenesis for neuropathic pain (NP), yet the underlying mechanisms need to be fully understood. In this study, we made an attempt to clarify the NP-related hub genes and signaling pathways so as to provide effective diagnostic and therapeutic methods toward NP.

Methods: Microarray expression profile GSE30691 including the mRNA-seq data of the spared nerve injury (SNI)-induced NP rats was accessed from the GEO database. Then, genes associated with NP development were screened using differential analysis along with random walk with restart (RWR). GO annotation and KEGG pathway analyses were performed to explore the biological functions and signaling pathways where the genes were activated. Afterward, protein-protein interaction (PPI) analysis and GO analysis were conducted to further identify the hub genes which showed an intimate correlation with NP development.

Results: Totally 94 genes associated with NP development were screened by differential analysis and RWR analysis, and they were observed to be predominantly enriched in hormone secretion and transport, cAMP signaling pathway and other NP occurrence associated functions and pathways. Thereafter, the 94 genes were subjected to PPI analysis to find the genes much more associated with NP and a functional module composed of 48 genes were obtained. 8 hub genes including C3, C1qb, Ccl2, Cxcl13, Timp1, Fcgr2b, Gal, and Lyz2 were eventually identified after further association and functional enrichment analyses, and the expression of these 8 genes were all higher in SNI rats by comparison with those in Sham rats.

Conclusion: Based on the data collected from GEO database, this study discovered 8 hub genes that were closely related to NP occurrence and development, which help to provide potent theoretical basis for NP treatment.

Keywords: neuropathic pain, nerve injury model, bioinformatics analysis, hub gene, functional association network

OPEN ACCESS

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

Reviewed by:
Jinhai Wang,
First Affiliated Hospital, School of Medicine, Zhejiang University, China
Wentao Dai,
Shanghai Center for Bioinformation Technology, China

****Correspondence:***
Xiaoguang Liu
lxg_pku1@163.com

†These authors share first authorship

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Neuroscience

Received: *03 March 2020*
Accepted: *20 April 2020*
Published: *20 May 2020*

Citation:
Wang K, Yi D, Yu Z, Zhu B, Li S and Liu X (2020) Identification of the Hub Genes Related to Nerve Injury-Induced Neuropathic Pain. Front. Neurosci. 14:488. doi: 10.3389/fnins.2020.00488

HIGHLIGHT

- 94 genes closely related to neuropathic pain occurrence are identified using differential analysis and random walk with restart.
- 8 hub genes that are implicated with neuropathic pain regulation are verified by means of protein association analysis along with GO annotation and KEGG pathway analyses.

INTRODUCTION

Pain is a survival mechanism that can act as a warning sign of ongoing or impending tissue damage. In evolutionary terms, the activation of high threshold mechanical nociceptors or other types of specialized nociceptor plays a protective role and can serve as a warning system for dangerous stimuli (Cohen and Mao, 2014). Neuropathic pain (NP) is a kind of chronic pain induced by the injury or dysfunction of the central or peripheral nervous system (Jensen et al., 2011; Finnerup et al., 2016; Watson and Sandroni, 2016). Smith et al. (2007) discovered that compared with the nociceptive pain, NP produced a more negative impact on the life quality. However, the specific mechanisms underlying NP remain elusive and there is still a lack of the effective therapeutic methods. Therefore, it is urgent to further clarify the underlying mechanisms toward NP and exploit the relevant genes and signaling pathways, so as to provide theoretical basis and new ideas for future treatment.

The pathogenesis of NP is complex. The current discovery has shown that NP is not only involved in the excitability of transmitting pain sensitive neurons, but also related to central and peripheral sensitization (von Hehn et al., 2012; Meacham et al., 2017). Central pain, a subtype of NP (like spinal cord injury-induced pain), manifests as a series of symptoms and signs that are developed after the injury of the central nervous system, such as nerve pain caused by headache, abdominal pain, etc. (Cohen and Mao, 2014). In addition, other than the inducement of inflammatory response in some local tissues, peripheral nerve injury or tissue damage can also cause alterations of the inflammatory-related cytokines in the central nervous system, such as the elevation of interleukin-1β (IL-1β), IL-6, tumor necrosis factor-α (TNF-α), chemokines and neurotrophic factors (Rubio and Sanz-Rodriguez, 2007; Wei et al., 2012; Matsuo et al., 2014; Sato et al., 2014; Gerard et al., 2015). Zhang et al. (2013) reported that inhibiting CXCL1-CXCL2 signal might be used as a novel therapeutic method for NP treatment. Moreover, Xiong et al. (2016) found that M1-type small glial cells could produce a large number of pro-inflammatory factors, resulting in the aggravation of nerve injury and consequently leading to the neurological dysfunction. Hence, further investigating the molecular mechanisms underlying NP and clarifying the effective targets are significant for the application of pain medication in clinical targeted therapies.

Bioinformatics can provide tools for analysis of large amounts of information, like the microarray technique, which has been widely applied in high-throughput gene expression detection (Schena et al., 1995; Allison et al., 2006) and can be reliably used for the identification of novel targets for clinical diagnosis and treatment (Chen et al., 2015). This study aimed to discuss the molecular mechanisms of nerve injury-induced NP, and in turn identify the hub genes and signaling pathways associated with NP pathogenesis. Due to the certain difficulties and the risk of experimenting on human being, we adopted animal models to study the NP pathogenesis. In our study, microarray GSE30691 including the mRNA-seq data of the spared nerve injury (SNI)-induced NP rats was downloaded from the GEO database. Multiple bioinformatics methods were adopted here for screening the genes and pathways which were associated with NP occurrence. In the meantime, hub genes intimately relevant to NP development were identified. Our findings would provide new thoughts for exploration of genes and biological pathways that are involved in nerve injury-induced NP.

MATERIALS AND METHODS

Data Source

The mRNA expression microarray GSE30691 was downloaded from the Gene Expression Omnibus (GEO) database[1]. The dataset was composed of the L4-5 dorsal root ganglion (DRG) segments from the rats at 0, 3, 7, 21, and 40 days after SNI and from the rats at 3, 7, and 21 days after a sham operation. Three independent experiments were performed in each period.

Differential Analysis

Statistical software R (version 3.3.2)[2] and packages of Bioconductor[3] were applied for analysis of the differentially expressed genes (DEGs). Differential analysis was performed on the genes from the SNI and Sham rats in 3, 7, and 21-day three time periods using the "limma" package (Smyth, 2011), with $|\log FC| > 0.585$ and FDR < 0.05 used as the screening threshold.

Random Walk With Restart (RWR) for Screening NP-Related Genes

In order to make the analysis more reliable, a network which can execute on the protein-protein interaction (PPI) network was designed. RWR is a classic ranking algorithm which is originated from the random walk. With the aid of RWR analysis, the global structure information of the network can be explored, which is helpful to estimate the proximity between two nodes (Zhang et al., 2018; Fan et al., 2019; Valdeolivas et al., 2019).

The PPI network we had constructed was denoted as a graph G = (V, E) comprising of a set of genes V and a set of interactions E. The graph could be characterized by an n × n adjacency matrix A:

$$A_{[i,j]'} = \frac{A_{[i,j]}}{\sum_{k=1}^{n} A_{[i,j]}} \quad (1)$$

where n refers to the total number of the nodes. A_([i,j]) = 1 if node i and node j are interacted, and 0 otherwise. In the RWR algorithm, each node in the network was conferred a restart

[1]https://www.ncbi.nlm.nih.gov/geo/

[2]https://www.r-project.org/

[3]http://www.bioconductor.org/

probability and all probabilities constituted a vector which was defined as Pt:

$$P\_(t+1) = (1-r)A^{\wedge\prime} P\_t + rP\_0 \quad (2)$$

where A is the column-wise normalized adjacency matrix A. Pt is the previous state probabilities at time t. r is the restart probability. P0 is the initial state probabilities, a column vector with 1/m for the m seed genes (NP-related genes identified in L4-5 DRG segments from SNL cohort) and 0 for other genes on the network.

The iteration process was repeated until the difference between two vectors was smaller than 1×10^{-5}. New NP-related genes were subsequently identified and Venn diagram was plotted to obtain RWR genes.

GO Annotation and KEGG Pathway Analyses

As we had screened the genes associated with NP using the RWR analysis, R package "clusterProfiler" was used to perform GO annotation and KEGG pathway analyses, with the critical value of $p < 0.05$ and $q < 0.05$ (Yu et al., 2012). Afterward, the enrichment results were visualized with the aid of R package "enrichplot" so as to further analyze the biological functions and pathways by which the genes affected NP.

PPI Network Construction

The NP-associated genes we identified were projected onto a PPI network for functional association analysis (confidence > 0.400) using the STRING database[4]. Thereafter, the Cytoscape plugin "MCODE" was applied to find the functional module, while "ClueGO" and "CluePedia" were used for enrichment analysis toward the genes in the module.

RESULTS

Identification of NP-Associated Genes

To find the genes that were tightly correlated with NP, differential analysis was run for the genes in the microarray GSE30691. The results revealed that a total of 51, 99, and 63 DEGs were identified from the SNI group versus the Sham group at 3, 7, and 21 days after SNI, respectively (**Figure 1A**). Thereafter, the DEGs were projected onto a PPI network, and the DEGs of each time period were regarded as seed genes for follow-up RWR analysis. Eventually, a total of 95, 98, and 97 NP-associated genes were screened in three periods, respectively, and the 94 common genes identified using a Venn diagram were considered to be closely correlated with NP (**Figure 1B**).

[4]https://string-db.org/cgi/input.pl

GO and KEGG Analyses on the NP-Associated Genes

As abovementioned, 94 genes were identified to be closely related to NP. In order to investigate the role of these genes in NP, GO annotation and KEGG pathway analyses were conducted. As revealed in **Figure 2A**, the most significantly activated biological functions of these genes were hormone secretion and transport, potassium ion transport, humoral immune response and negative regulation of immune system process, etc. While the most noteworthy enriched signaling pathways were complement and coagulation cascade, neuroactive ligand-receptor interaction, cAMP signaling pathway and ECM-receptor interaction, etc. (**Figure 2B**). All of these functions and pathways have been proven to show an intimate correlation with NP development, which supports our result that the 94 genes we identified are significantly associated with NP.

PPI Network Analysis and Identification of Hub Genes

To gain more insight into the role of these 94 genes in NP development and find the hub genes which were significantly implicated in, a PPI network based on these 94 genes was established on STRING database for functional association analysis and sequentially visualized on Cytoscape. The plugin "MCODE" was used to find functional modules and eventually a module consisting of 48 genes with the highest score was obtained (**Figure 3A**). After that, biological functions where the 48 genes were most activated were explored by means of GO annotation. It turned out that the genes were predominantly enriched in some NP development associated functions, including

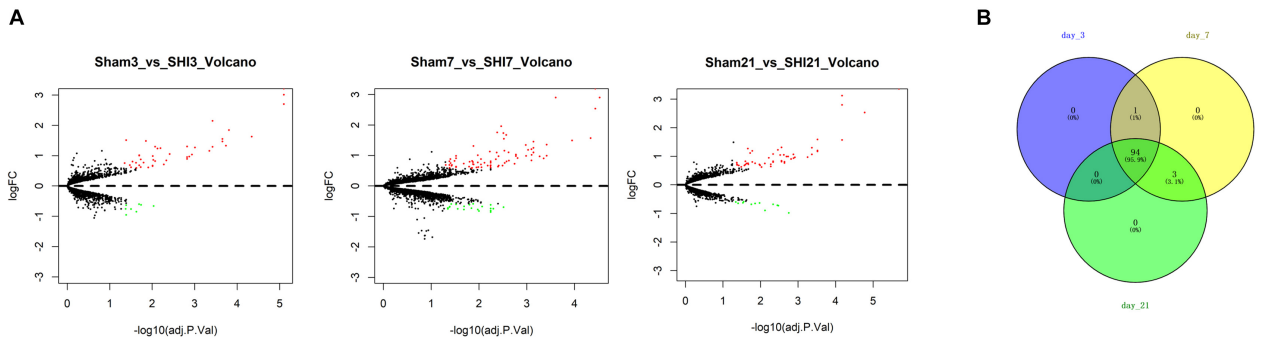

FIGURE 1 | Identification of NP-associated genes. **(A)** Volcano Plots of the DEGs in the SNI group versus the Sham group at 3, 7, and 21 days after surgery, respectively; **(B)** Venn diagram showed the common NP-associated genes from the 3 time periods.

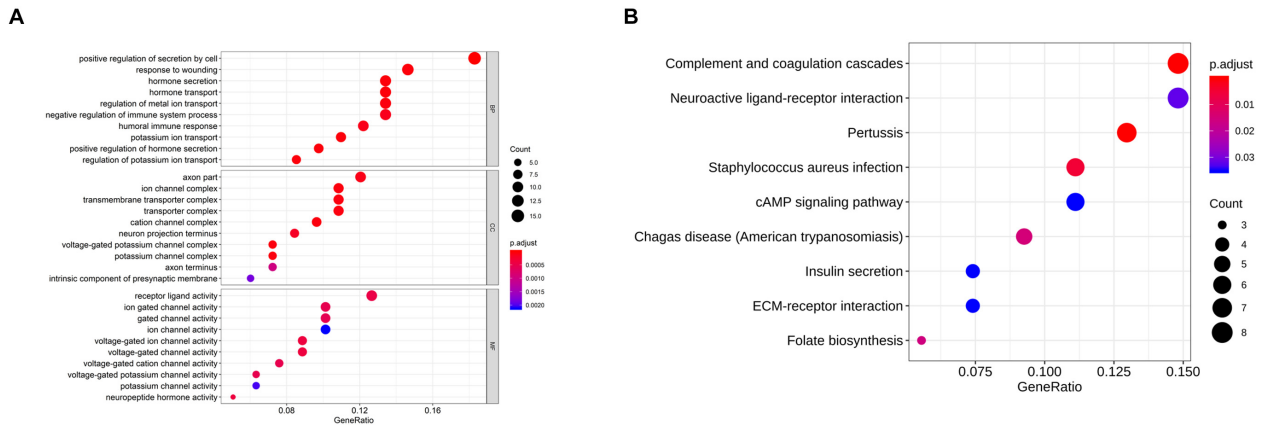

FIGURE 2 | GO and KEGG analyses on the NP-associated genes. **(A)** The most enriched GO terms of the 94 NP-associated genes; **(B)** the most activated KEGG pathways of the 94 NP-associated genes.

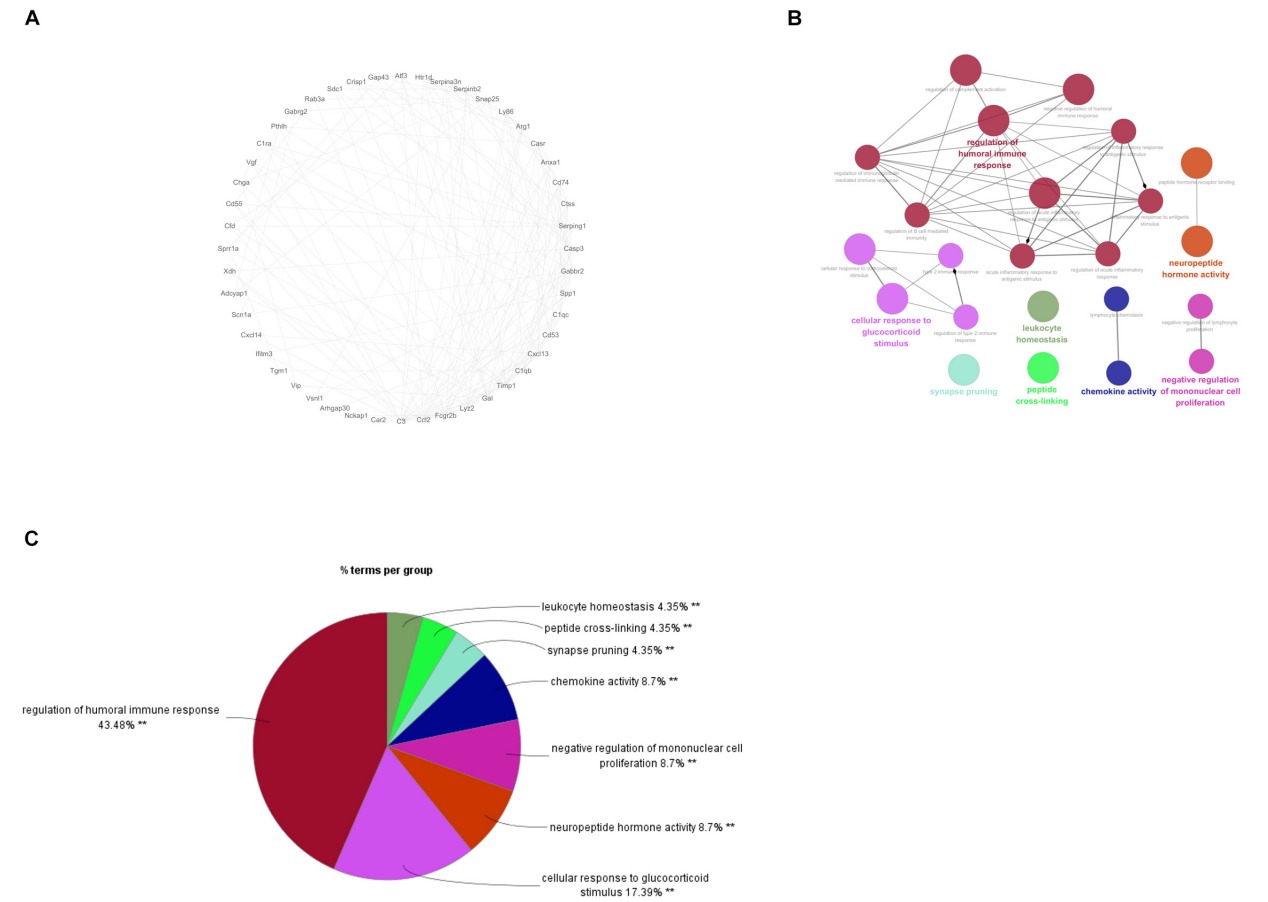

FIGURE 3 | PPI Network Analysis and identification of hub genes. **(A)** PPI network for the 48 genes involved in the functional module of a highest score; **(B)**: the most enriched GO terms for the 48 genes; C: Proportional graph of the enriched GO terms of the 48 genes.

regulation of humoral immune response, cellular response to glucocorticoid stimulus, neuropeptide hormone activity, negative regulation of mononuclear cell proliferation and chemokine activity (**Figures 3B,C**).

As the 48 genes in the module were found to be enriched in NP-associated functions, further PPI network analysis was conducted to identify the hub genes that were most relevant to NP occurrence and development. The connectivity degree of

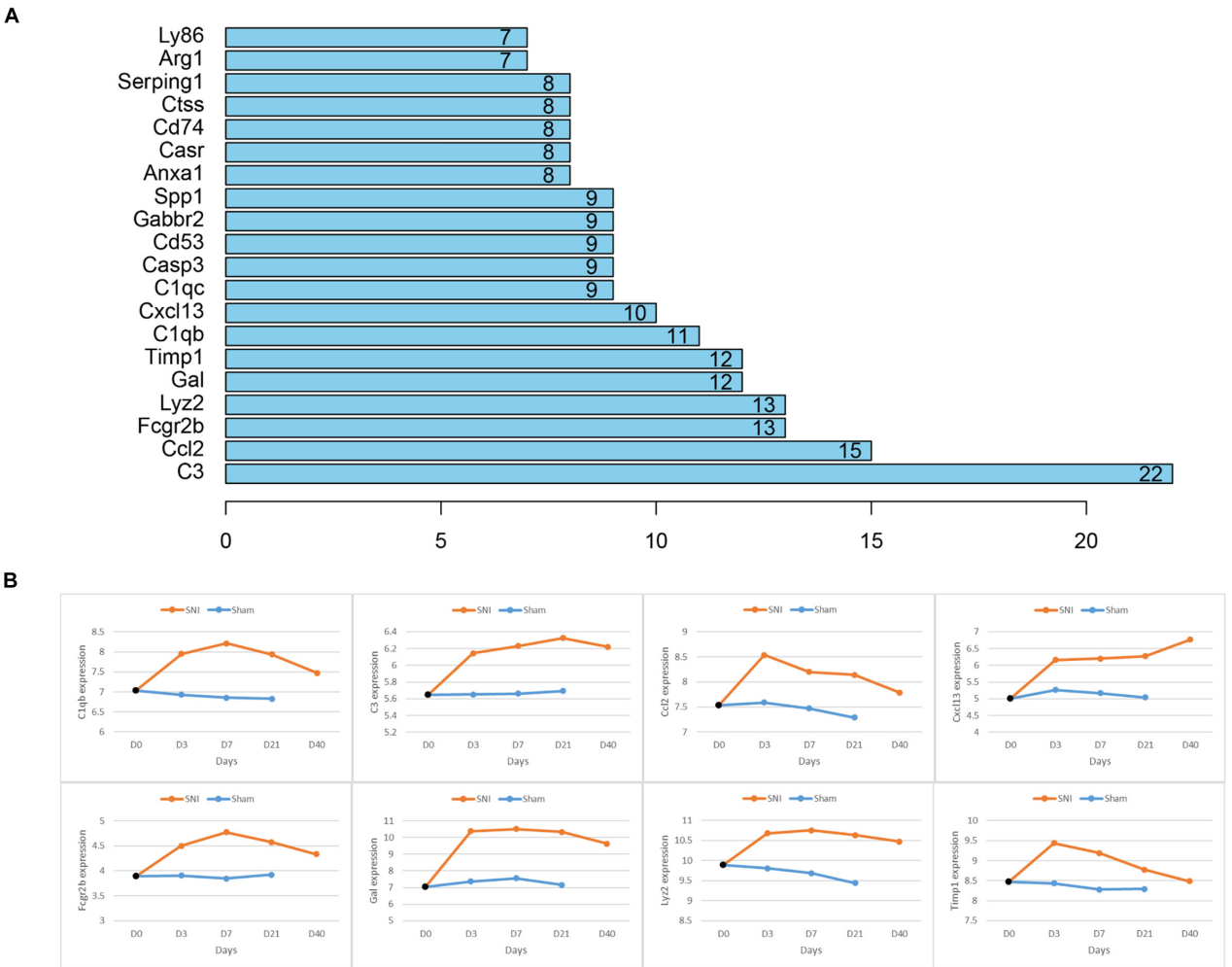

FIGURE 4 | Hub genes associated with NP and the relative expression of each hub gene in different time periods between SNI and Sham rats. **(A)** Node number of the hub genes in the PPI network; **(B)** the relative expression of each hub gene with a degree higher than 10 in different time periods between Sham and SNI rats.

each gene was calculated and it turned out that 8 genes (C3, C1qb, Ccl2, Cxcl13, Timp1, Fcgr2b, Gal, Lyz2) which had a degree higher than 10 were identified and here were regarded as the hub genes significantly associated with NP development (**Figure 4A**). For further verification, we detected the expression of the 8 genes in different time periods between SNI and Sham rats and found that all these 8 genes exhibited a much higher expression in SNI rats in comparison with those in Sham rats in the same period (**Figure 4B**). In view of these results, the 8 hub genes were confirmed to be most associated with NP development.

DISCUSSION

NP is a complex chronic pain with elusive mechanisms currently (Kerstman et al., 2013; Gilron et al., 2015). It has been reported that NP is commonly associated with paresthesia, hyperesthesia, paralgesia and hyperalgesia (Bouhassira and Attal, 2016). In addition, some changes in the whole nervous system are also implicated with NP, such as the ectopic action potential, the generation of the new synaptic circuitry and the neuro-immune interaction (Taylor, 2001; Zhuo et al., 2011). Therefore, it is a necessity to extend our knowledge on the NP pathogenesis, which is of great significance on setting of the treatment strategies for responsive NP prevention and efficacy improvement.

It has been revealed that NP is always accompanied by the alteration of genes on the sensory pathways (von Hehn et al., 2012). In the present study, we adopted the microarray technique to identify the NP-related DEGs and the activated signaling pathways from the SNI rat models. Microarray technique is a tool able to quantify the expression levels of thousands of genes across the biological samples simultaneously, and it can provide the complex regulations among genes based on the expression data of the whole genome, which helps us find better targets for NP treatment (Gao et al., 2018). During the whole analysis process, some factors like the sample attributes, processing tools, handling methods and results screening all made some effects on the final results. To make the results more reliable, we used

multiple analytical methods, such as differential analysis, RWR, GO annotation, and KEGG pathway analyses. More specifically, mRNA expression data from the SNI and Sham rats 3, 7, and 21 days after operation were obtained from microarray GSE30691 through the GEO database. Subsequently, DEGs in the three time periods were screened and projected onto a PPI network, after which the DEGs in each period were taken as seed genes for RWR analysis. Eventually, 94 common genes were identified and considered to be associated with NP development. Our findings lay a foundation for future investigation of the molecular mechanisms underlying NP occurrence and development.

After identification of the NP-associated genes, we performed enrichment analysis and found that these genes were predominantly enriched in some biological functions like hormone secretion and transport, potassium ion transport, humoral immune response, negative regulation of immune system process, while these functions have been proven to be involved in NP occurrence and development (Jaggi et al., 2015; Jang et al., 2018; Wang et al., 2018). Additionally, cAMP signaling pathway and ECM-receptor interaction are two signaling pathways that have been confirmed to be implicated with multiple functions in regulation of NP (Zhou et al., 2017; Yan et al., 2018; Yan et al., 2019), and our study observed that the genes we identified were activated in these two pathways as well. Given the findings above, the specific role of these genes in NP development requires further exploration.

Despite the genes and pathways associated with NP development we found, 8 hub genes (C3, C1qb, Ccl2, Cxcl13, Timp1, Fcgr2b, Gal, Lyz2) that were responsible for NP development regulation were identified and some of them have been reported to present an intimate correlation with NP development. Levin ME et al. conducted the microarray analysis on the data from the SNI-induced NP rats, and the results demonstrated that multiple complement factors like C1 inhibitor, C1q α, β, and γ, C1r, C1s, C2, C3, C4, and C7 were all up-regulated, and rats with less complement C3 in plasma (cobra venom factor-treated) had relative attenuated pain behaviors (Levin et al., 2008). This study found that C3 was remarkably increased in SNI rats and exhibited a rising trend within 0–21 days. As for Timp1, Gal and C1qb, researchers discovered that Gal and C1qb could be used as potential biomarkers for NP occurrence (Buckley et al., 2018; Yang et al., 2018). Besides, a study on CXCL13 made by Jiang et al. (2016) revealed that CXCL13 could make an effect on NP development via targeting CXCR5. These genes were all observed to be significantly highly expressed in SNI rats in our study. Moreover, CCL2 has been verified to play a vital role in NP development (Zhao et al., 2017), yet there has been no study performed to investigate the role of Fcgr2b and Lyz2 in NP. Overall, our identification of the 8 hub genes further confirms their significance in NP development.

In conclusion, we found 94 NP-associated genes and corresponding enriched biological functions and signaling pathways by means of multiple bioinformatics approaches. Furthermore, 8 hub genes that were implicated with NP development regulation were identified. Our findings lay a foundation for future exploration of the molecular mechanisms underlying NP development and help to find potential targets for NP diagnosis and treatment.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this study were downloaded and accessed from the Gene Expression Omnibus (GEO) database: https://www.ncbi.nlm.nih.gov/geo/, with accession no: GSE30691.

AUTHOR CONTRIBUTIONS

KW, DY, ZY, BZ, SL, and XL: study design, wrote the paper, and revised the manuscript and gave the final approval of the version. KW, DY, and ZY: literature search. BZ, SL, and XL: acquired the data.

FUNDING

This study was supported by Beijing Municipal Natural Science Foundation (Grant No. 7192226), China Central Health Scientific Research Project (Grant No. W2017BJ53), and National Natural Science Foundation of China (Grant No. 81972103).

REFERENCES

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7, 55–65. doi: 10.1038/nrg1749

Bouhassira, D., and Attal, N. (2016). Translational neuropathic pain research: a clinical perspective. *Neuroscience* 338, 27–35. doi: 10.1016/j.neuroscience.2016.03.029

Buckley, D. A., Jennings, E. M., Burke, N. N., Roche, M., McInerney, V., Wren, J. D., et al. (2018). The development of translational biomarkers as a tool for improving the understanding, diagnosis and treatment of chronic neuropathic pain. *Mol. Neurobiol.* 55, 2420–2430. doi: 10.1007/s12035-017-0492-8

Chen, G., Fang, X., and Yu, M. (2015). Regulation of gene expression in rats with spinal cord injury based on microarray data. *Mol. Med. Rep.* 12, 2465–2472. doi: 10.3892/mmr.2015.3670

Cohen, S. P., and Mao, J. (2014). Neuropathic pain: mechanisms and their clinical implications. *BMJ* 348:f7656. doi: 10.1136/bmj.f7656

Fan, X. N., Zhang, S. W., Zhang, S. Y., Zhu, K., and Lu, S. (2019). Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. *BMC Bioinformatics* 20:87. doi: 10.1186/s12859-019-2675-y

Finnerup, N. B., Haroutounian, S., Kamerman, P., Baron, R., Bennett, D. L., Bouhassira, D., et al. (2016). Neuropathic pain: an updated grading system for research and clinical practice. *Pain* 157, 1599–1606. doi: 10.1097/j.pain.0000000000000492

Gao, Y., Sun, N., Wang, L., Wu, Y., Ma, L., Hong, J., et al. (2018). Bioinformatics analysis identifies p53 as a candidate prognostic biomarker for neuropathic pain. *Front. Genet.* 9:320. doi: 10.3389/fgene.2018.00320

Gerard, E., Spengler, R. N., Bonoiu, A. C., Mahajan, S. D., Davidson, B. A., Ding, H., et al. (2015). Chronic constriction injury-induced nociception is relieved

by nanomedicine-mediated decrease of rat hippocampal tumor necrosis factor. *Pain* 156, 1320–1333. doi: 10.1097/j.pain.0000000000000181

Gilron, I., Baron, R., and Jensen, T. (2015). Neuropathic pain: principles of diagnosis and treatment. *Mayo Clin. Proc.* 90, 532–545. doi: 10.1016/j.mayocp.2015.01.018

Jaggi, A. S., Kaur, A., Bali, A., and Singh, N. (2015). Expanding spectrum of sodium potassium chloride co-transporters in the pathophysiology of diseases. *Curr. Neuropharmacol.* 13, 369–388. doi: 10.2174/1570159x13666150205130359

Jang, J. H., Park, J. Y., Oh, J. Y., Bae, S. J., Jang, H., Jeon, S., et al. (2018). Novel analgesic effects of melanin-concentrating hormone on persistent neuropathic and inflammatory pain in mice. *Sci. Rep.* 8:707. doi: 10.1038/s41598-018-19145-z

Jensen, T. S., Baron, R., Haanpaa, M., Kalso, E., Loeser, J. D., Rice, A. S., et al. (2011). A new definition of neuropathic pain. *Pain* 152, 2204–2205. doi: 10.1016/j.pain.2011.06.017

Jiang, B. C., Cao, D. L., Zhang, X., Zhang, Z. J., He, L. N., Li, C. H., et al. (2016). CXCL13 drives spinal astrocyte activation and neuropathic pain via CXCR5. *J. Clin. Invest.* 126, 745–761. doi: 10.1172/JCI81950

Kerstman, E., Ahn, S., Battu, S., Tariq, S., and Grabois, M. (2013). Neuropathic pain. *Handb. Clin. Neurol.* 110, 175–187. doi: 10.1016/B978-0-444-52901-5.00015-0

Levin, M. E., Jin, J. G., Ji, R. R., Tong, J., Pomonis, J. D., Lavery, D. J., et al. (2008). Complement activation in the peripheral nervous system following the spinal nerve ligation model of neuropathic pain. *Pain* 137, 182–201. doi: 10.1016/j.pain.2007.11.005

Matsuo, H., Uchida, K., Nakajima, H., Guerrero, A. R., Watanabe, S., Takeura, N., et al. (2014). Early transcutaneous electrical nerve stimulation reduces hyperalgesia and decreases activation of spinal glial cells in mice with neuropathic pain. *Pain* 155, 1888–1901. doi: 10.1016/j.pain.2014.06.022

Meacham, K., Shepherd, A., Mohapatra, D. P., and Haroutounian, S. (2017). Neuropathic pain: central vs. peripheral mechanisms. *Curr. Pain Headache Rep.* 21:28. doi: 10.1007/s11916-017-0629-5

Rubio, N., and Sanz-Rodriguez, F. (2007). Induction of the CXCL1 (KC) chemokine in mouse astrocytes by infection with the murine encephalomyelitis virus of Theiler. *Virology* 358, 98–108. doi: 10.1016/j.virol.2006.08.003

Sato, K. L., Johanek, L. M., Sanada, L. S., and Sluka, K. A. (2014). Spinal cord stimulation reduces mechanical hyperalgesia and glial cell activation in animals with neuropathic pain. *Anesth. Analg.* 118, 464–472. doi: 10.1213/ANE.0000000000000047

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470. doi: 10.1126/science.270.5235.467

Smith, B. H., Torrance, N., Bennett, M. I., and Lee, A. J. (2007). Health and quality of life associated with chronic pain of predominantly neuropathic origin in the community. *Clin. J. Pain* 23, 143–149. doi: 10.1097/01.ajp.0000210956.31997.89

Smyth, G. K. (2011). *limma: Linear Models for Microarray Data*. New York, NY: Springer. 397–420.

Taylor, B. K. (2001). Pathophysiologic mechanisms of neuropathic pain. *Curr. Pain Headache Rep.* 5, 151–161. doi: 10.1007/s11916-001-0083-1

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 35, 497–505. doi: 10.1093/bioinformatics/bty637

von Hehn, C. A., Baron, R., and Woolf, C. J. (2012). Deconstructing the neuropathic pain phenotype to reveal neural mechanisms. *Neuron* 73, 638–652. doi: 10.1016/j.neuron.2012.02.008

Wang, S. L., Zhao, Z. K., Sun, J. F., Sun, Y. T., Pang, X. Q., Zeng, Z. W., et al. (2018). Review of Anemone raddeana Rhizome and its pharmacological effects. *Chin. J. Integr. Med.* 24, 72–79. doi: 10.1007/s11655-017-2901-2

Watson, J. C., and Sandroni, P. (2016). Central neuropathic pain syndromes. *Mayo Clin. Proc.* 91, 372–385. doi: 10.1016/j.mayocp.2016.01.017

Wei, X. H., Yang, T., Wu, Q., Xin, W. J., Wu, J. L., Wang, Y. Q., et al. (2012). Peri-sciatic administration of recombinant rat IL-1beta induces mechanical allodynia by activation of src-family kinases in spinal microglia in rats. *Exp. Neurol.* 234, 389–397. doi: 10.1016/j.expneurol.2012.01.001

Xiong, X. Y., Liu, L., and Yang, Q. W. (2016). Functions and mechanisms of microglia/macrophages in neuroinflammation and neurogenesis after stroke. *Prog. Neurobiol.* 142, 23–44. doi: 10.1016/j.pneurobio.2016.05.001

Yan, L. P., Qian, C. X., Ma, C., and Wang, L. L. (2018). [Effect of Electroacupuncture of "Weizhong" (BL 40) and "Huantiao" (GB 30) on cAMP-PKA-CREB signaling of spinal cord in rats with neuropathic pain]. *Zhen Ci Yan Jiu* 43, 788–792. doi: 10.13702/j.1000-0607.180250

Yan, X. T., Xu, Y., Cheng, X. L., He, X. H., Wang, Y., Zheng, W. Z., et al. (2019). SP1, MYC, CTNNB1, CREB1, JUN genes as potential therapy targets for neuropathic pain of brain. *J. Cell Physiol.* 234, 6688–6695. doi: 10.1002/jcp.27413

Yang, J. A., He, J. M., Lu, J. M., and Jie, L. J. (2018). Jun, Gal, Cd74, and C1qb as potential indicator for neuropathic pain. *J. Cell Biochem.* 119, 4792–4798. doi: 10.1002/jcb.26673

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, J., Suo, Y., Liu, M., and Xu, X. (2018). Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein-protein interaction network. *Biochim. Biophys. Acta Mol. Basis Dis.* 1864, 2369–2375. doi: 10.1016/j.bbadis.2017.11.017

Zhang, Z. J., Cao, D. L., Zhang, X., Ji, R. R., and Gao, Y. J. (2013). Chemokine contribution to neuropathic pain: respective induction of CXCL1 and CXCR2 in spinal cord astrocytes and neurons. *Pain* 154, 2185–2197. doi: 10.1016/j.pain.2013.07.002

Zhao, H., Alam, A., Chen, Q., Ma, A. E., Pal, A., Eguchi, S., et al. (2017). The role of microglia in the pathobiology of neuropathic pain development: what do we know? *Br. J. Anaesth.* 118, 504–516. doi: 10.1093/bja/aex006

Zhou, J., Xiong, Q., Chen, H., Yang, C., and Fan, Y. (2017). Identification of the spinal expression profile of non-coding rnas involved in neuropathic pain following spared nerve injury by sequence analysis. *Front. Mol. Neurosci.* 10:91. doi: 10.3389/fnmol.2017.00091

Zhuo, M., Wu, G., and Wu, L. J. (2011). Neuronal and microglial mechanisms of neuropathic pain. *Mol. Brain* 4:31. doi: 10.1186/1756-6606-4-31

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 Physiology

ORIGINAL RESEARCH
published: 03 June 2020
doi: 10.3389/fphys.2020.00483

Identification of Post-myocardial Infarction Blood Expression Signatures Using Multiple Feature Selection Strategies

Ming Li[1†], Fuli Chen[2†], Yaling Zhang[3], Yan Xiong[2], Qiyong Li[2] and Hui Huang[2*]*

[1] Department of Cardiology, Eastern Hospital, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China, [2] Department of Cardiology, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China, [3] Department of Nephrology, Eastern Hospital, Sichuan Academy of Medical Sciences & Sichuan Provincial People's Hospital, Chengdu, China

Myocardial infarction (MI) is a type of serious heart attack in which the blood flow to the heart is suddenly interrupted, resulting in injury to the heart muscles due to a lack of oxygen supply. Although clinical diagnosis methods can be used to identify the occurrence of MI, using the changes of molecular markers or characteristic molecules in blood to characterize the early phase and later trend of MI will help us choose a more reasonable treatment plan. Previously, comparative transcriptome studies focused on finding differentially expressed genes between MI patients and healthy people. However, signature molecules altered in different phases of MI have not been well excavated. We developed a set of computational approaches integrating multiple machine learning algorithms, including Monte Carlo feature selection (MCFS), incremental feature selection (IFS), and support vector machine (SVM), to identify gene expression characteristics on different phases of MI. 134 genes were determined to serve as features for building optimal SVM classifiers to distinguish acute MI and post-MI. Subsequently, functional enrichment analyses followed by protein-protein interaction analysis on 134 genes identified several hub genes (IL1R1, TLR2, and TLR4) associated with progression of MI, which can be used as new diagnostic molecules for MI.

Keywords: myocardial infarction, Monte Carlo feature selection, incremental feature selection, support vector machine, gene

OPEN ACCESS

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

Reviewed by:
Xiaogang Guo,
Zhejiang University, China
Yun Li,
University of Pennsylvania, United States

***Correspondence:**
Qiyong Li
lqycdu@163.com
Hui Huang
huangtong315143020@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *06 March 2020*
Accepted: *20 April 2020*
Published: *03 June 2020*

Citation:
Li M, Chen F, Zhang Y, Xiong Y, Li Q and Huang H (2020) Identification of Post-myocardial Infarction Blood Expression Signatures Using Multiple Feature Selection Strategies. Front. Physiol. 11:483. doi: 10.3389/fphys.2020.00483

INTRODUCTION

Myocardial infarction (MI), one of the most common cardiac diseases, has been a serious threat to human health worldwide for a long period. According to the third universal definition of MI, it is the condition of myocardial necrosis in a clinical setting consistent with myocardial ischemia (Bax et al., 2012). MI occurs when the blood flow is impaired and the cardiomyocyte is injured due to the lack of oxygen supply (Lu et al., 2015). Patients with coronary atherosclerosis have a high risk of developing a MI when inflammation takes place in the vascular wall (Thygesen et al., 2007). Usually a more serious event is termed as acute myocardial infarction (AMI). The symptoms of MI include chest pain, shortness of breath, abnormal heart beating, and fatigue (Kosuge et al., 2006). Smoking and dyslipidemia are thought to be important risk factors for MI, which is correlated with

the increasing mortality rate in China (Critchley et al., 2004). Approximately three million cases of MI are diagnosed every year and the annual incidence rate is about 600 cases per 100,000 people (Rogers et al., 2008; Nascimento et al., 2019). The average mortality of MI is approximately 27% according to statistics (White and Chew, 2008), making it a major cause of death in the world.

After the onset of MI, many pathological processes occur, such as the death of myocardial cells, and will develop into different conditions depending on the status of the patient. MI can be classified pathologically as acute, healing, or healed, which is roughly correlated with the disease duration. Acute MI describes a severe event usually accompanied by activated inflammation at early onset. Then it progresses to healing, which can be characterized by the presence of mononuclear cells and fibroblasts and the absence of polymorphonuclear leukocytes. The entire process reaching the healed state of MI takes about several months when cellular infiltration fades away and scar tissue appears (Thygesen et al., 2007). The different phases after onset reflect distinct pathological conditions. So, a better understanding of the phases will contribute to the treatment of MI and improve the outcomes of patients.

Early and rapid diagnosis is important for the decision of treatment and improvement of survival. There are several methods for the evaluation of MI including electrocardiography (ECG) and cardiac markers. The ECG has a high specificity of 90% for MI but a poor sensitivity of 20% (Zimetbaum and Josephson, 2003). Serum biomarkers of myocardial necrosis, such as cardiac troponin (I or T), which can specifically reflect myocardial injury, show high clinical sensitivity and can improve the diagnostic accuracy (Jaffe et al., 2000). Levels of MB isoforms of creatine (CK-MB) also exhibit the ability to identify MI as an increased CK-MB value is associated with myocarditis and electrical cardioversion (Members et al., 2007). Although the traditional clinical approach has shown excellent performance for diagnosing MI, an increasing number of studies have proven that molecular markers, like the transcription profile in serum, are capable of reflecting detailed pathological conditions and subsequent progress of MI, which will help to determine the optimal treatment.

Owing to the great development in RNA-seq technology, many novel genes are found to play crucial roles in various diseases. It has been reported that the specific expression pattern of certain genes is relevant to the pathological condition of MI. For examples, H-FABP, which is involved in myocardial fatty-acid metabolism, is rapidly released into the cytosol in early MI and can act as an early marker (Glatz et al., 1988). B-type Natriuretic Peptide (BNP) is secreted by the ventricles in response to the tension of cardiomyocytes and leads to the reduction of blood pressure, making it a prognostic marker after MI (De Lemos et al., 2001). Growth Differentiation Factor-15 (*GDF15*) is specifically expressed in the heart when ischemia or reperfusion happened, and increasing *GDF15* indicates a higher risk of death in MI patients (Wollert et al., 2007). Besides, non-coding RNAs are also found to be involved in the pathogenesis of MI. Circulating miR-208a, which is only detected in AMI patients, is thought to be the novel potential biomarker for early diagnosis with higher sensitivity and specificity (Wang et al., 2010). Given that the progress of MI involves numerous complex biological processes and pathways, the overall transcriptome analysis will contribute to revealing a more detailed molecular mechanism and an easier way to locate the key genes related to pathogenesis of MI.

In this study, we utilized bioinformatics methods to explore the key gene networks associated with MI from the vast transcriptomic data. Previous studies which aimed to find the biomarker for MI put the focus on separated genes but ignored the linkage among them. With the application of bioinformatics, we can study the complex expression network consisting of multiple genes with less time consumed and a higher efficiency. Transcriptomic data was obtained from the published paper which performed whole blood RNA profiling at different time points in cohort with MI (Vanhaverbeke et al., 2019). In order to identify the key biomarkers for distinguishing different pathological extents, we manually divided all patients into three categories based on the duration of MI. These three different groups roughly reflect distinct pathological conditions. Next, we constructed an optimal support vector machine (SVM) model with the application of a feature selection method called Monte Carlo Feature Selection (MCFS) (Chen et al., 2018a, 2019a,b,d, 2020; Pan et al., 2018, 2019a,b; Wang et al., 2018; Jiang et al., 2019; Li et al., 2019) and incremental feature selection (IFS) (Chen et al., 2018b, 2019d; Lei et al., 2018; Li and Huang, 2018; Sieber et al., 2018; Zhang et al., 2018; Wang and Huang, 2019; Yan et al., 2019). 134 optimal genes were selected which show specific expression patterns during varied phases of MI and can distinguish different categories with a highly accuracy. The functional enrichment analysis suggested the important biological processes and pathways related to the progress of MI and corresponding hub genes were identified by gene network analysis. The selected genes in the current study can serve as novel biomarkers for different phases of MI and contribute to revealing the pathological mechanism of MI.

MATERIALS AND METHODS

Dataset

The blood gene expression profiles of 166 samples which incorporate three phases of MI (D0: acute MI, D30: 30-days post-MI, and Y1: 1-year post-MI) were downloaded with the gene expression omnibus (GEO) under accession number of GSE123342 (Vanhaverbeke et al., 2019). There were 65 D0, 64 D30, and 37 Y1 samples. There were 70,523 probes in Affymetrix Human Transcriptome Array 2.0 corresponding to 30,905 genes. The probes for the same gene were averaged and the data was quantile normalized (Bolstad et al., 2003). We wanted to find the genes with changed expression patterns in post-MI.

Monte Carlo Feature Selection (MCFS)

Monte Carlo feature selection has been a widely used method for feature selection (Chen et al., 2018a, 2019a,b,d, 2020; Pan et al., 2018, 2019a,b; Wang et al., 2018; Jiang et al., 2019; Li et al., 2019). It was originally developed by Draminski et al. (2008).

It randomly constructed many tree classifiers of the sub datasets from the original dataset and assigned the importance to a feature based on how much it participated in the tree classifiers. The java software dmLab[1] with default parameters (Draminski et al., 2008) was used to apply the Monte-Carlo feature selection method.

To be more specific, the original dataset was divided into s subsets of m features (m<<d, where d is the total number of features, i.e., 30,905 genes in this study). Then, for each subset, t trees were constructed. Therefore, a total of s·t classification trees were constructed. At last, the relative importance (RI) of each feature was estimated as follows:

$$RI_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{no.\ in\ n_g(\tau)}{no.\ in\ \tau} \right)^v \quad (1)$$

where $IG(n_g(\tau))$ was the information gain (IG) of node $n_g(\tau)$, $(no.\ in\ n_g(\tau))$ was the number of samples in node $n_g(\tau)$, $(no.\ in\ \tau)$ was the number of samples in tree τ, *wAcc* was the weighted accuracy over all samples, and *u* and *v* were two regular factors which were set as default.

After running MCFS, all features can be ranked based on their RI. The higher the RI, the more important a feature was.

Incremental Feature Selection (IFS)

With MCFS, all features were ranked. But we still did not know how many genes we should choose. Ideally, we wanted the number of selected genes to be small but their classification performance to be great. To find the balance and the optimal signature, we adopted IFS (Chen et al., 2018b, 2019d; Lei et al., 2018; Li and Huang, 2018; Sieber et al., 2018; Zhang et al., 2018; Wang and Huang, 2019; Yan et al., 2019). During IFS, a serial of feature sets $F = [f_1, f_2, \ldots, f_N]$ were constructed. N ranged from 1 to 1000. For each feature set, we constructed corresponding support vector machine (SVM) classifiers using the R function svm with default parameters in package e1071[2] and evaluated the performance using leave-one-out cross validation (LOOCV). Therefore, we can get a serial of LOOCV accuracies which corresponded to different feature sets with various numbers of features. With the help of the IFS curve, we can balance the model complexity and classification performance. If the number of features was too small, the performance would be bad. If the number of features was too large, too much noise would be introduced and the performance would decrease. The optimal selection would be achieved when the number of features was small and the accuracy was high.

Functional Enrichment Analysis

The biological functions of the optimal MI signature genes were analyzed using hypergeometric enrichment analysis (Shi et al., 2018a,b). The significance of the signature genes onto Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Gene Ontology (GO) biological process (BP), molecular function (MF), and cell component (CC) were represented with hypergeometric *p* values.

[1]http://www.ipipan.eu/staff/m.draminski/mcfs.html

[2]https://cran.r-project.org/web/packages/e1071/index.html

RESULTS

Feature Ranking Based on MCFS Method

In this study, we exploited newly published gene expression profiles of patients with MI (Vanhaverbeke et al., 2019). Each patient was represented by 30,905 gene expression features. We integrated expression profiles of all patients into one matrix for quantile normalization followed by applying the MCFS method for ranking analysis. Each feature was assessed by estimating the relative importance (RI) value. After evaluating all features, we generated a feature list *F* in descending order of RI values of features. The ranked features with RI values were provided in **Supplementary Table S1**.

Establishing Classifier Using SVM With IFS

According to the feature list obtained by the MCFS algorism, the IFS method was employed to identify optimal feature sets which could train the best performance for SVM. To save computing time, we established the series of feature subsets (F_1, F_2, F_3, . . ., F_{1000}) based on the top 1 to 1000 genes in *F*. For each feature set, we established a classifier by SVM algorithm and estimated optimal parameters through Leave-One-Out Cross-Validation (LOOCV). The LOOCV accuracies on multiple feature subsets were shown in **Figure 1**, from which we can see that the accuracy reached a plateau area when the top 134 features were used for

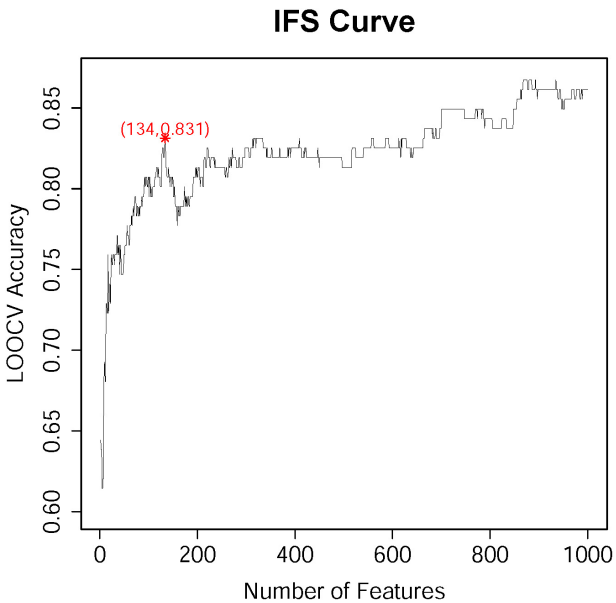

FIGURE 1 | The IFS-curve obtained by IFS method. The *X*-axis represents the number of features participating in the classification. The *Y*-axis represents the LOOCV accuracy produced by SVM. The accuracy reached 0.831 when the top 134 features were used. When even more features were added, the accuracy did not increase too much. It reached the plateau area. Therefore, to balance the number of features and the accuracy, 134 features were selected.

TABLE 1 | The confusion matrix of the predicted results using the 134 features.

	Predicted D0*	Predicted D30**	Predicted Y1***
Actual D0*	47	12	6
Actual D30**	2	58	4
Actual Y1***	1	3	33

, acute MI; **, 30-days post-MI; *, 1-year post-MI.*

building the classifier. The 134 optimal features were listed in **Supplementary Table S2**. The confusion matrix of the predicted results using the 134 features was shown in **Table 1**. It can be seen that all three classifiers had a great performance.

Cluster Analysis With Optimal Features

In order to confirm the performance of identified optimal features/genes representing different phases of samples, we performed cluster analysis on expression profiles of 134 optimal genes in 166 samples which incorporate three phases of MI (D0: acute MI, D30: 30-days post-MI, and Y1: 1-year post-MI). We used a heatmap to visualize the expression of such optimal genes among three groups of samples (**Figure 2**). The cluster tree illustrated that most samples belonging to the same phase can be clustered together and different phases were classified into different branches. In addition, these optimal genes were also classified into three clusters which correspond to high expression in three phases. The largest gene cluster with 90 genes was highly expressed in D0, the cluster with 16 genes had a high expression of D30, and the cluster with 28 genes was highly expressed in Y1.

The expression levels of genes like KLHL8, HCLS1, MOB3A, IL17RA, ETF1, ZFAS1, CRK, MXD1, UBXN2B, FCAR, and EXTL3 decreased in post-MI while the expression levels of genes like DCK and RNU4-7P increased in post-MI. We plotted the boxplots of several representative genes in **Figure 3**. For

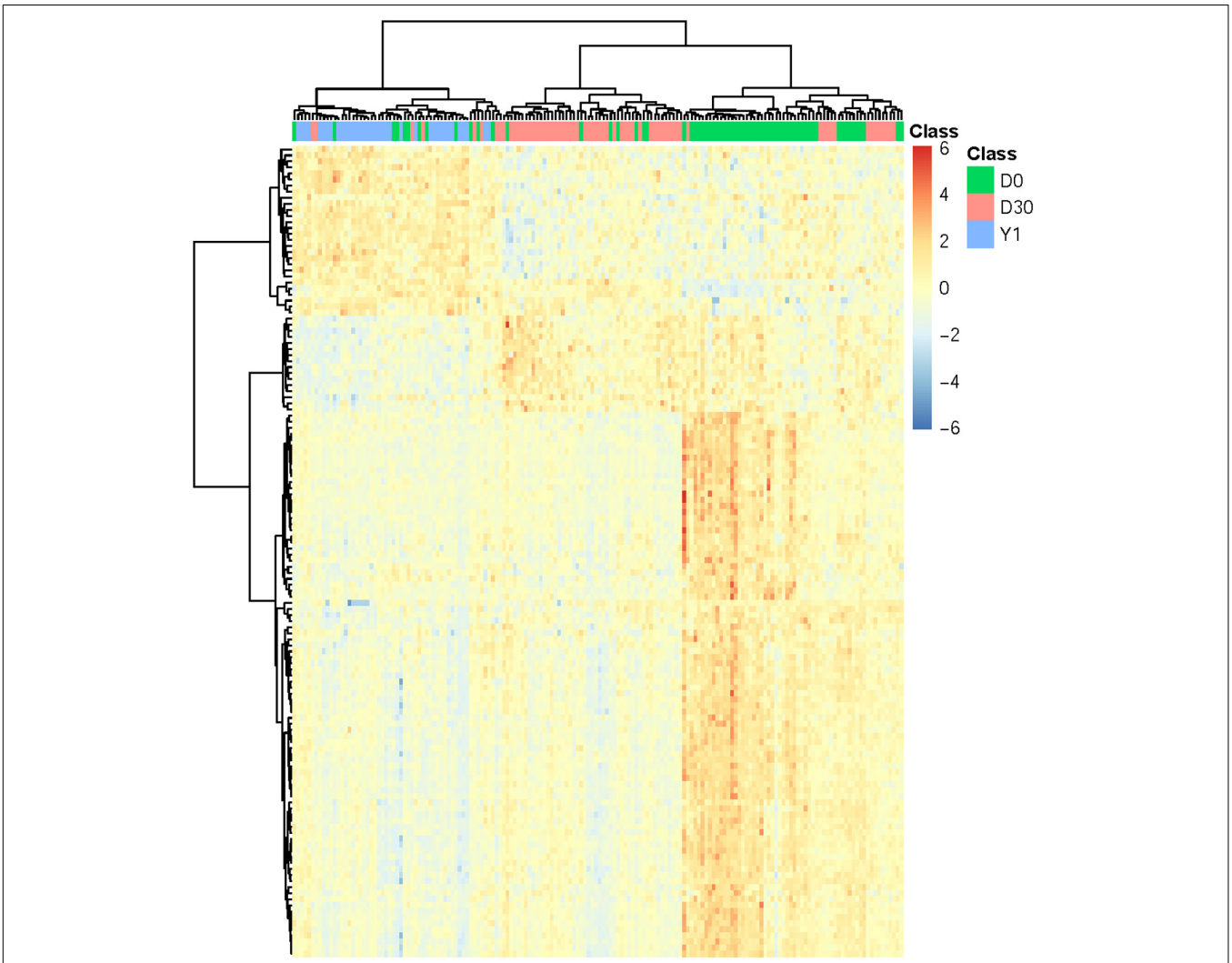

FIGURE 2 | Heatmap of all MI samples on the top 134 genes. The columns refer to samples and the rows refer to genes. Different phases of samples were colored by green (D0 represents acute MI), red (D30 represents 30-days post-MI), and blue (Y1 represents 1-year post-MI), respectively. It can be seen that the samples from different time points had different expression patterns. For each time point, there was a corresponding cluster with highly expressed genes at this time point.

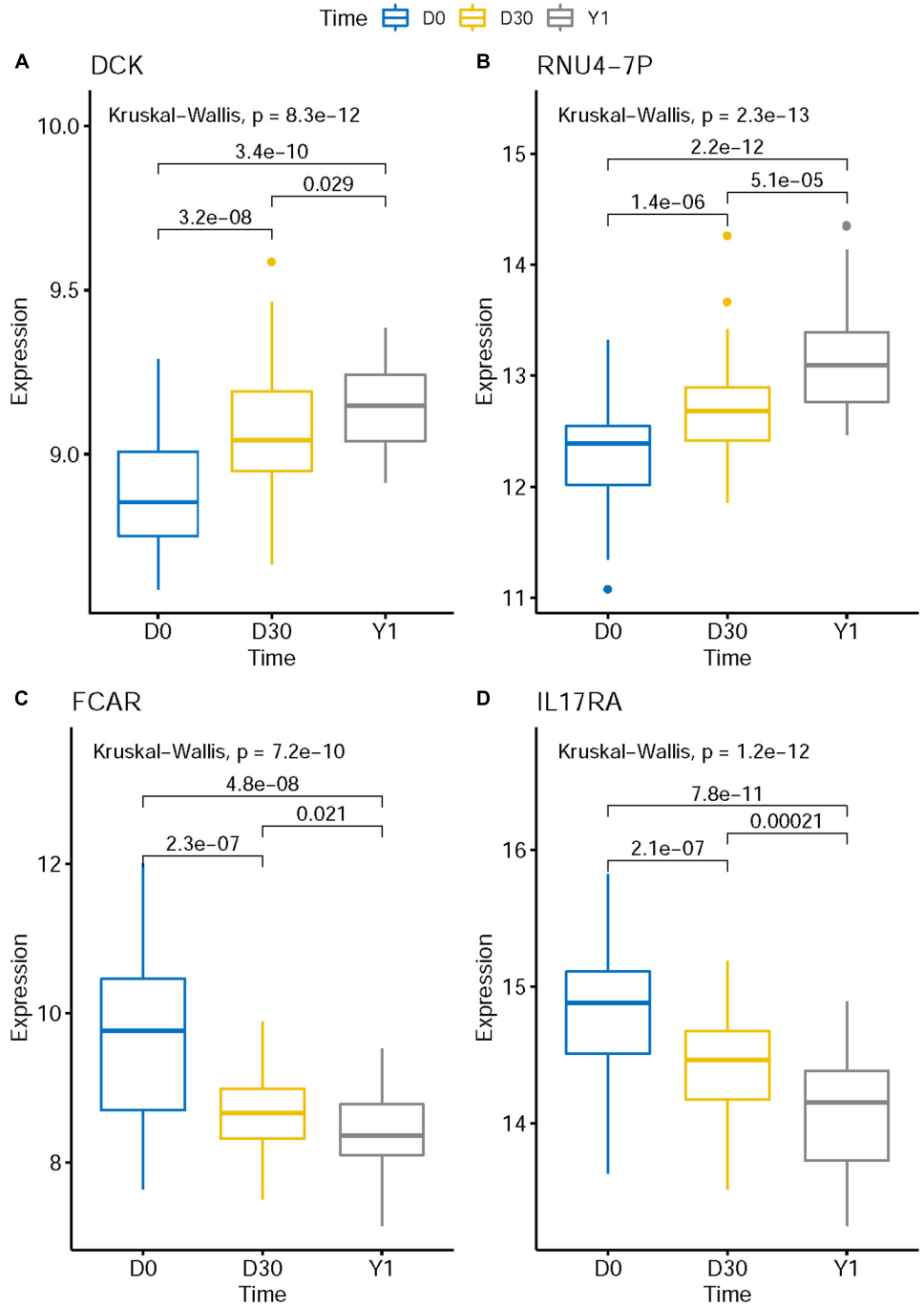

FIGURE 3 | The boxplots of representative post-MI expression patterns. The expression level of genes like DCK **(A)** and RNU4-7P **(B)** increased in post-MI while the expression levels of genes like FCAR **(C)** and IL17RA **(D)** decreased in post-MI. These expression patterns may reveal the mechanisms of MI.

example, in **Figure 3C**, the expression levels of FCAR on D0 was significantly higher than on D30 and the expression levels on D30 was significantly higher than on Y1. There was a consistent post-MI trend of FCAR. These expression patterns may reveal the mechanisms of MI. FCAR is a member of the immunoglobulin superfamily and encodes a receptor for the Fc region of IgA. The cell surface receptors for immunoglobulin, such as the protein of FCAR, can activate many inflammatory processes involved in atherosclerosis and coronary artery disease (Daëron, 1997; Gavasso et al., 2005). The variation in FCAR which causes an amino acid alteration was found to increase the risk of MI and coronary heart disease, indicating the potential functional role of FCAR in the development of cardiovascular disease (Iakoubova et al., 2006, 2008).

Functional Enrichment Analysis on Optimal Features

We next performed functional enrichment analysis on these 134 optimal features/genes. A hypergeometric distribution test was applied to calculate *p* value to determine the significantly enriched entries. Firstly, we performed Gene Ontology enrichment analysis on the gene set. In biological progress aspect, the top 3 GO terms were GO: 0044264, GO: 0046903, and GO: 0005976, which correspond to cellular polysaccharide metabolic process, secretion, and polysaccharide metabolic process, respectively (**Supplementary Table S3**). The top GO term of cellular component was GO: 0005964, corresponding to phosphorylase kinase complex (**Supplementary Table S4**). The most significantly enriched GO term of molecular function was GO: 0004908, which was annotated to interleukin-1 receptor activity (**Supplementary Table S5**). Secondly, KEGG enrichment analysis was applied to discover the signaling pathways involved in these optimal genes. In this part, we found the insulin signaling pathway (hsa04910) was the top enriched KEGG pathway (**Supplementary Table S6**).

Analysis of Gene Interaction Networks

To investigate the correlation of optimal genes, we applied gene interaction analysis on 134 features/genes to construct gene interaction networks. Proteins encoded by such classes of genes were input into a STRING database (Szklarczyk et al., 2018), mining interaction relationship. Although part of the genes showed no association with other genes, we found an interaction network consisting of dozens of genes and predicted three hub genes, including IL1R1, TLR2, and TLR4 (**Figure 4**), which may interact with each other to play a non-negligible role in the progression of MI.

IL1R1, TLR2, and TLR4 showed promising associations with MI. It was reported that the knockout of IL1R1 caused a reduction of leukocyte production after MI, leading to a decreased inflammation with better outcome (Sager et al., 2015). In another mice study, the up-regulated IL1R1 at 7 days post-MI prolonged the inflammation by suppressing neutrophil apoptosis (Iyer et al., 2015).

TLR2 plays a fundamental role in the activation of innate immunity (Binder et al., 2002). There are usually high levels of cytokines that result in inflammation in MI patients; TLR2 served as a key receptor to activate the corresponding pathways (Pagano et al., 2012). The experimental data indicated that circulatory TLR2 is relevant to different manifestations of myocardial I/R injury (Arslan et al., 2010). And the inhibition of TLR2 has beneficial effects on I/R injury in a murine model of MI (Arslan et al., 2009). TLR2 is the key receptor which can induce the inflammation after MI, therefore many MI-related genes show close interactions with TLR2.

TLR4 regulates the cytokines after cardiac damage (Arslan et al., 2010). Activation of TLR4 was related to myocytic inflammatory reaction in MI patients 14 days after onset, suggesting that TLR4 signaling plays a role in the progress after MI (Satoh et al., 2006).

DISCUSSION

Optimal Genes Associated With Classification of MI

Using the feature selection, 134 genes were extracted and exhibited an excellent performance in our prediction model of SVM, suggesting that these genes may participate in the progression of MI. Here, we took some of the selected genes as examples to give a detailed discussion to validate the relevance of a given gene in distinguishing different pathological phases of MI. Through a literature review, several experimental evidences or analysis results have been found to confirm the reliability of our prediction.

DLGAP1-AS1

The top ranked feature identified by our computational analysis turned out to be DLGAP1-AS1, an RNA gene which is affiliated with the lncRNA class. A recent publication has reported that high expression of lncRNA DLGAP1-AS1 was detected in rats with acute ischemia-reperfusion (I/R) injury. And decreased DLGAP1-AS1 can alleviate vascular endothelial cell injury via PI3K pathway (Shen et al., 2020). The cause of I/R injury is mainly attributed to the reperfusion of the MI area, and vascular endothelial cells are the key defense with the occurrence of I/R injury (Carden and Granger, 2000; Causey et al., 2012). So, it came to the inference that down-regulated DLGAP1-AS1 serves as the protective regulator to mediate vascular endothelial cells in preventing I/R injury after the MI. This builds relevance for the alteration in DLGAP1-AS1 expression in the progression of MI. Besides that, gene DLGAP1 showed significant differential expression in Flk-1 knockout mice under the treatment of heart perfusion (Thirunavukkarasu et al., 2008). Flk-1 is one of the most important receptors that trigger cardioprotective signals and plays a crucial role in I/R injury (Shalaby et al., 1995; Addya et al., 2005), as DLGAP1-AS1 can target DLGAP1 and regulate its expression. This finding provided further support to suggest DLGAP1-AS1 was closely related to the progression of MI.

PYGL

The following ranked gene was Glycogen Phosphorylase L (PYGL), which encodes a homodimeric protein that is involved in galactose metabolism (Tomihira et al., 2004). Early research has mentioned the application of glycogen phosphorylase in the diagnosis of myocardial ischemic injury and infarction (Krause et al., 1996; Mair, 1998). Recently, PYGL was reported to display an up-regulated expression in an acute MI cohort compared to normal controls (Zhang et al., 2017). Another study has demonstrated that up-regulated PYGL may induce the RIP1-dependent necrosis after I/R injury, implying that PYGL is associated with the subsequent progress after AMI and I/R injury (Oerlemans et al., 2012). This evidence proves our prediction results were reasonable.

MEGF9

MEGF9 was also identified as an important gene related to the classification of MI. MEGF9 is a protein coding gene and is associated with Fiedler's Myocarditis disease. Some studies have

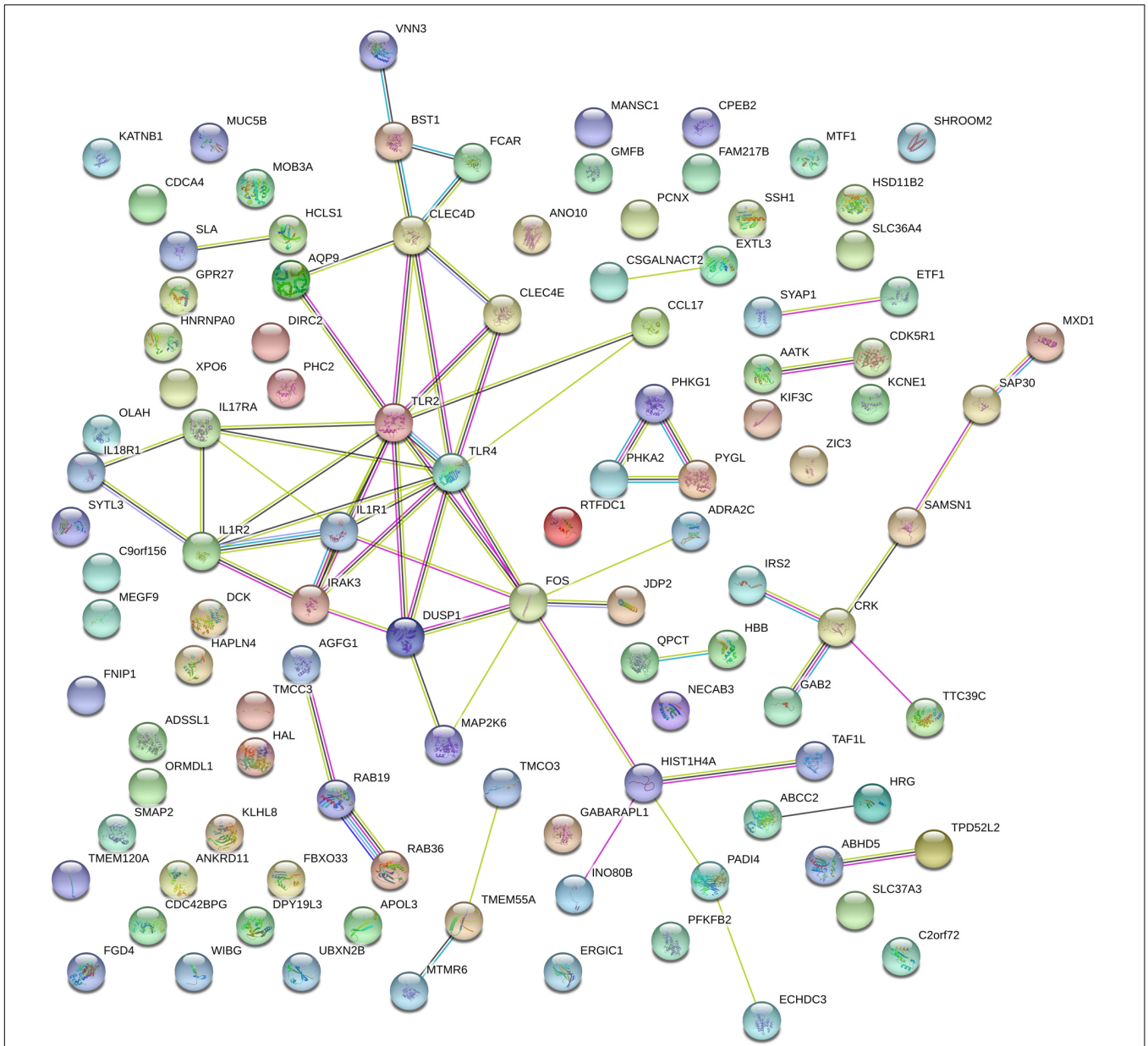

FIGURE 4 | Gene networks containing IL1R1, TLR2, TLR4, and other related genes. These gene interactions were extracted from the protein-protein interaction network reported in a STRING database and plotted by the online drawing tool of STRING. IL1R1, TLR2, and TLR4 were located in the center positions and were hub genes.

observed differentially expressed MEGF9 and identified it as the key gene involved in AMI and MI (Cheng et al., 2017; Qiu and Liu, 2019). As demonstrated by genome-wide linkage analysis in autosomal dominant congenital heart defects, the risk region in chromosome 9q was found and MEGF9 turned out to be one of the candidate genes in this position. However, no mutations were found in this gene through the sequence analysis, suggesting that MEGF9 may play its role by post-transcriptional regulation instead of at the genome level (Van De Meerakker et al., 2011). Hence, the specific expression pattern could be a signature for diagnosing MI and even distinguishing different phases of MI.

PHC2

Next, another gene called PHC2, which is associated with the metabolism of proteins, was selected by our computational analysis. PHC2 was reported as one of the differentially expressed genes in patients with MI compared to controls by bioinformatics screening (Wu et al., 2018). Another study also confirmed the key role of PHC2 in the pathogenesis of MI through protein-protein interaction network analysis (Qiu and Liu, 2019). These results implied that PHC2 may act as a hub gene which can mediate some other genes' interaction and regulate downstream pathways, and then influence the progress of MI. Our analysis highlighted the

importance of PHC2, pointing out that this specific gene may be applied as a marker for the prediction of recurrent MI.

Through literature review and reasonable inference, the selected genes mentioned above were all found to play crucial roles in the progress of MI and show the discriminative ability to indicate the pathological degree of disease. It validated the reliability of our prediction model. Considering the length limitation of the article, we can't give extended descriptions of all 134 selected genes. We believed that these 134 selected genes were meaningful during the development of MI and its subsequent progression, and they will contribute to the research of molecular mechanism and provide benefits for the therapy of disease.

Gene Ontology Enrichment Analysis

Given that the selected 134 genes were deemed as important features for the classification of different phases of MI, we performed GO and KEGG functional enrichment analysis to explore the key biological processes or pathways during the progress of disease. As shown in **Supplementary Tables S3–S6**, we analyzed the enriched GO terms and KEGG pathways which showed statistical significance. A detailed discussion was given about the linkage between certain functional sets and MI.

Based on the enrichment results of 134 selected genes, we found some GO biological process terms with high scores turned out to be involved in the polysaccharide metabolic process, including GO: 0044264 and GO: 0005976. As early as 1965, scientists have noticed the important role of glucose load in MI (Cohen and Shafrir, 1965). Recent studies reported that certain polysaccharide compounds can affect myocardial injury via regulating the inflammation response (Li et al., 2011; Lim et al., 2016). As demonstrated by experiments on rat, the polysaccharide extract from Momordica charantia down-regulated the expression of NF-kappaB and ameliorated oxidative stress and inflammation, which caused a cardioprotective effect against MI (Raish, 2017). Polysaccharide metabolism plays an important role during the progression of MI, so the biological processes related to polysaccharide metabolism are meaningful and can be used to indicate the progression of disease based on its specific pattern.

Apart from GO terms that belong to biological processes, we found these 134 genes are also enriched in a cellular components term GO: 0005964 with the highest probability. GO: 0005964 refers to phosphorylase kinase complex. For cardiomyocytes, the storage of glycogen is important during the emergency situation. Increasing Ca^{2+} concentration in cytosol can induce glycogenolysis by the activation of phosphorylase kinase, which can alleviate myocardial damage during MI or cardiac surgery (Raish, 2017). In fact, some phosphorylases have been applied in the diagnosis of myocardial ischemic injury and infarction since the serum level of phosphorylase showed a signature with the diseases (Krause et al., 1996). It is reasonable for the MI-related genes to be enriched in such GO term that would mean the phosphorylase play a crucial role during the progression of MI.

The most enriched GO terms of molecular function turned out to be interleukin-1 (IL-1)-related functions including GO: 0004908 and GO: 0019966, which represent IL-1 receptor activity and IL-1 binding, respectively. An interleukin-1 receptor gene ST2 was increased in the serum after MI, suggesting that this gene may participate in innate immunity during myocardial injury (Weinberg et al., 2002). What's more, ST2 was reported to be able to predict the clinical outcome in AMI due to its role in cardiac pathophysiology (Shimpo et al., 2004). Many publications have observed the elevated serum level of IL-1 receptor in patients with AMI (Shibata et al., 1997; Balbay et al., 2001). These findings proved the important role of IL-1 in the progression of MI, and confirmed the relation between selected genes and MI.

KEGG Pathways Enrichment Analysis

The KEGG pathways enrichment analysis provided various pathway results. Among these, the highest enriched pathway turned out to be hsa04910, which is an insulin signaling pathway. Increased insulin can promote the metabolism of glucose to maintain the balance of blood glucose. The connection between abnormal insulin signaling and heart disease has already been reported, in that diabetes mellitus significantly increased the risk of ischemic heart disease (Miettinen et al., 1998). Insulin can protect cardiomyocytes from apoptosis through activating downstream pathways such as PI3K and Akt (Yao et al., 2014). It was reported that impaired insulin signaling will cause the dysfunction of mitochondria after MI due to the reduced glucose transport and oxygen content (Sena et al., 2009). Thus, the insulin signaling pathway is important during the progression of MI and influences the pathological degree of disease.

CONCLUSION

Taken together, the gene features yielded by our model showed strong relevance to the pathological progression of MI, suggesting their discriminative ability in the classification of different phases of disease. This validated the reliability of our machine learning model and proved that it can be used as a novel approach to predict the status of MI patients. Our work will contribute to the precise diagnosis and help to decide on the optimal treatment for each patient with MI. In addition, the genes identified by our analysis provided new understanding about the pathogenesis of MI and established a solid foundation for future research.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the GSE123342.

AUTHOR CONTRIBUTIONS

HH and QL contributed to the conception and design. HH, ML, and FC contributed to the development of methodology. All authors contributed to analysis and interpretation of data, writing, review, and/or revision of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2020.00483/full#supplementary-material

TABLE S1 | The ranked features with RI values.

TABLE S2 | The 134 optimal features for building the classifier.

TABLE S3 | Gene ontology enrichment analysis for biological progress aspect.

TABLE S4 | Gene ontology enrichment analysis for cellular component.

TABLE S5 | Gene ontology enrichment analysis for molecular function.

TABLE S6 | KEGG enrichment analysis for discovering the signaling pathways involved in these optimal genes.

REFERENCES

Addya, S., Shiroto, K., Turoczi, T., Zhan, L., Kaga, S., Fukuda, S., et al. (2005). Ischemic preconditioning-mediated cardioprotection is disrupted in heterozygous Flt-1 (VEGFR-1) knockout mice. *J. Mol. Cell Cardiol.* 38, 345–351. doi: 10.1016/j.yjmcc.2004.11.033

Arslan, F., Keogh, B., McGuirk, P., and Parker, A. E. (2010). TLR2 and TLR4 in ischemia reperfusion injury. *Mediators Inflamm.* 2010:704202. doi: 10.1155/2010/704202

Arslan, F., Smeets, M., O'Neill, L., Keogh, B., McGuirk, P., Timmers, L., et al. (2009). Myocardial ischemia/reperfusion injury is mediated by leukocytic TLR2 and reduced by systemic administration of a novel anti-TLR2 antibody. *Eur. Heart J.* 121:317. doi: 10.1161/CIRCULATIONAHA.109.880187

Balbay, Y., Tikiz, H., Baptiste, R., Ayaz, S., Şaşmaz, H., and Korkmaz, ŞJ. A. (2001). Circulating interleukin-1 beta, interleukin-6, tumor necrosis factor-alpha, and soluble ICAM-1 in patients with chronic stable angina and myocardial infarction. *Angiology* 52, 109–114. doi: 10.1177/000331970105200204

Bax, J. J., Baumgartner, H., Ceconi, C., Dean, V., Fagard, R., Funck-Brentano, C., et al. (2012). Third universal definition of myocardial infarction. *Eur. Heart J.* 60:1581.

Binder, C. J., Chang, M.-K., Shaw, P. X., Miller, Y. I., Hartvigsen, K., Dewan, A., et al. (2002). Innate and acquired immunity in atherogenesis. *Nat. Med.* 8, 1218–1226. doi: 10.1038/nm1102-1218

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Carden, D. L., and Granger, D. N. (2000). Pathophysiology of ischaemia–reperfusion injury. *J. Pathol.* 190, 255–266. doi: 10.1002/(sici)1096-9896(200002)190:3<255::aid-path526>3.0.co;2-6

Causey, M. W., Salgar, S., Singh, N., Martin, M., and Stallings, J. D. (2012). Valproic acid reversed pathologic endothelial cell gene expression profile associated with ischemia–reperfusion injury in a swine hemorrhagic shock model. *J. Vasc. Surg.* 55, 1096–1103.

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

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

Chen, L., Pan, X., Guo, W., Gan, Z., Zhang, Y.-H., Niu, Z., et al. (2020). Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* 112, 2524–2534. doi: 10.1016/j.ygeno.2020.02.004

Chen, L., Pan, X., Zeng, T., Zhang, Y., Huang, T., and Cai, Y. (2019a). Identifying essential signature genes and expression rules associated with distinctive development stages of early embryonic cells. *IEEE Access.* 7, 128570–128578. doi: 10.1109/ACCESS.2019.2939556

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

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

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

Cheng, M., An, S., and Li, J. (2017). Identifying key genes associated with acute myocardial infarction. *Medicine* 96:e7741. doi: 10.1097/MD.0000000000007741

Cohen, A. M., and Shafrir, E. J. D. (1965). Carbohydrate metabolism in myocardial infarction: behavior of blood glucose and free fatty acids after glucose loading. *Diabetes* 14, 84–86. doi: 10.2337/diab.14.2.84

Critchley, J., Liu, J., Zhao, D., Wei, W., and Capewell, S. J. C. (2004). Explaining the increase in coronary heart disease mortality in Beijing between 1984 and 1999. *Circulation* 110, 1236–1244. doi: 10.1161/01.cir.0000140668.91896.ae

Daëron, M. (1997). Fc receptor biology. *Annu. Rev. Immunol.* 15, 203–234.

De Lemos, J. A., Morrow, D. A., Bentley, J. H., Omland, T., Sabatine, M. S., McCabe, C. H., et al. (2001). The prognostic value of B-type natriuretic peptide in patients with acute coronary syndromes. *N. Engl. J. Med.* 345, 1014–1021.

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

Gavasso, S., Nygård, O., Pedersen, E. R., Aarseth, J. H., Bleie, Ø, Myhr, K.-M., et al. (2005). Fcγ receptor IIIA polymorphism as a risk-factor for coronary artery disease. *Atherosclerosis* 180, 277–282. doi: 10.1016/j.atherosclerosis.2004.12.011

Glatz, J., Van Bilsen, M., Paulussen, R., Veerkamp, J., Van der Vusse, G., Reneman, R., et al. (1988). Release of fatty acid-binding protein from isolated rat heart subjected to ischemia and reperfusion or to the calcium paradox. *Biochim. Biophys. Acta* 961, 148–152. doi: 10.1016/0005-2760(88)90141-5

Iakoubova, O. A., Tong, C. H., Chokkalingam, A. P., Rowland, C. M., Kirchgessner, T. G., Louie, J. Z., et al. (2006). Asp92Asn polymorphism in the myeloid IgA Fc receptor is associated with myocardial infarction in two disparate populations: CARE and WOSCOPS. *Arterioscler. Thromb. Vasc. Biol.* 26, 2763–2768. doi: 10.1161/01.atv.0000247248.76409.8b

Iakoubova, O. A., Tong, C. H., Rowland, C. M., Kirchgessner, T. G., Young, B. A., Arellano, A. R., et al. (2008). Association of the Trp719Arg polymorphism in kinesin-like protein 6 with myocardial infarction and coronary heart disease in 2 prospective trials: the CARE and WOSCOPS trials. *J. Am. Coll. Cardiol.* 51, 435–443. doi: 10.1016/j.jacc.2007.05.057

Iyer, R. P., Patterson, N. L., Zouein, F. A., Ma, Y., Dive, V., de Castro Brás, L. E., et al. (2015). Early matrix metalloproteinase-12 inhibition worsens post-myocardial infarction cardiac dysfunction by delaying inflammation resolution. *Int. J. Cardiol.* 185, 198–208. doi: 10.1016/j.ijcard.2015.03.054

Jaffe, A. S., Ravkilde, J., Roberts, R., Naslund, U., Apple, F. S., Galvani, M., et al. (2000). It's time for a change to a troponin standard. *Am. Heart. Assoc.* 102, 1216–1220. doi: 10.1161/01.cir.102.11.1216

Jiang, Y., Pan, X., Zhang, Y., Huang, T., and Gao, Y. (2019). Gene expression difference between primary and metastatic renal cell carcinoma using patient-derived xenografts. *IEEE Access.* 7, 142586–142594. doi: 10.1109/ACCESS.2019.2944132

Kosuge, M., Kimura, K., Ishikawa, T., Ebina, T., Hibi, K., Tsukahara, K., et al. (2006). Differences between men and women in terms of clinical features of ST-segment elevation acute myocardial infarction. *Circ. J.* 70, 222–226. doi: 10.1253/circj.70.222

Krause, E.-G., Rabitzsch, G., Noll, F., Mair, J., and Puschendorf, B. (1996). Glycogen phosphorylase isoenzyme BB in diagnosis of myocardial ischaemic injury and infarction. *Mol. Cell Biochem.* 160, 289–295. doi: 10.1007/978-1-4613-1279-6_37

Lei, C., ShaoPeng, W., Yu-Hang, Z., Lai, W., XianLing, X., Tao, H., et al. (2018). Prediction of Nitrated Tyrosine residues in protein sequences by extreme learning machine and feature selection methods. *Comb. Chem. High Throughput Screen.* 21, 393–402. doi: 10.2174/1386207321666180531091619

Li, C., Gao, Y., Xing, Y., Zhu, H., Shen, J., Tian, J. J. F., et al. (2011). Fucoidan, a sulfated polysaccharide from brown algae, against myocardial ischemia–reperfusion injury in rats via regulating the inflammation response. *Food Chem. Toxicol.* 49, 2090–2095. doi: 10.1016/j.fct.2011.05.022

Li, J., and Huang, T. (2018). Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies. *Biochim. Biophys. Acta* 1864(6 Pt B), 2241–2246. doi: 10.1016/j.bbadis.2017.10.036

Li, J., Lu, L., Zhang, Y.-H., Xu, Y., Liu, M., Feng, K., et al. (2019). Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine. *Cancer Gene Ther.* 27, 56–69. doi: 10.1038/s41417-019-0105-y

Lim, S. H., Kim, Y., Yun, K. N., Kim, J. Y., Jang, J.-H., Han, M.-J., et al. (2016). Plant-based foods containing cell wall polysaccharides rich in specific active monosaccharides protect against myocardial injury in rat myocardial infarction models. *Sci. Rep.* 6, 1–15. doi: 10.1038/srep38728

Lu, L., Liu, M., Sun, R., Zheng, Y., and Zhang, P. (2015). Myocardial infarction: symptoms and treatments. *Cell Biochem. Biophys.* 72, 865–867. doi: 10.1007/s12013-015-0553-4

Mair, J. (1998). Glycogen phosphorylase isoenzyme BB to diagnose ischaemic myocardial damage. *Clin. Chim. Acta* 272, 79–86. doi: 10.1016/s0009-8981(97)00254-4

Members, N. W. G., Morrow, D. A., Cannon, C. P., Jesse, R. L., Newby, L. K., Ravkilde, J., et al. (2007). National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines: clinical characteristics and utilization of biochemical markers in acute coronary syndromes. *Circulation* 53, 552–574. doi: 10.1373/clinchem.2006.084194

Miettinen, H., Lehto, S., Salomaa, V., Mähönen, M., Niemelä, M., Haffner, S. M., et al. (1998). Impact of diabetes on mortality after the first myocardial infarction. *Diabetes Care* 21, 69–75. doi: 10.2337/diacare.21.1.69

Nascimento, B. R., Brant, L. C. C., Marino, B. C., Passaglia, L. G., and Ribeiro, A. L. P. (2019). Implementing myocardial infarction systems of care in low/middle-income countries. *Heart* 105, 20–26. doi: 10.1136/heartjnl-2018-313398

Oerlemans, M. I., Liu, J., Arslan, F., den Ouden, K., van Middelaar, B. J., Doevendans, P. A., et al. (2012). Inhibition of RIP1-dependent necrosis prevents adverse cardiac remodeling after myocardial ischemia–reperfusion in vivo. *Basic Res. Cardiol.* 107:270.

Pagano, S., Satta, N., Werling, D., Offord, V., De Moerloose, P., Charbonney, E., et al. (2012). Anti-apolipoprotein A-1 IgG in patients with myocardial infarction promotes inflammation through TLR2/CD14 complex. *J. Intern. Med.* 272, 344–357. doi: 10.1111/j.1365-2796.2012.02530.x

Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019a). Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms. *Int. J. Mol. Sci.* 20:2185. doi: 10.3390/ijms20092185

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

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

Qiu, L., and Liu, X. (2019). Identification of key genes involved in myocardial infarction. *Eur. J. Med. Res.* 24:22. doi: 10.1186/s40001-019-0381-x

Raish, M. (2017). *Momordica charantia* polysaccharides ameliorate oxidative stress, hyperlipidemia, inflammation, and apoptosis during myocardial infarction by inhibiting the NF-κB signaling pathway. *Int. J. Biol. Macromol.* 97, 544–551. doi: 10.1016/j.ijbiomac.2017.01.074

Rogers, W. J., Frederick, P. D., Stoehr, E., Canto, J. G., Ornato, J. P., Gibson, C. M., et al. (2008). Trends in presenting characteristics and hospital mortality among patients with ST elevation and non-ST elevation myocardial infarction in the National Registry of Myocardial Infarction from 1990 to 2006. *Am. Heart J.* 156, 1026–1034. doi: 10.1016/j.ahj.2008.07.030

Sager, H. B., Heidt, T., Hulsmans, M., Dutta, P., Courties, G., Sebas, M., et al. (2015). Targeting interleukin-1β reduces leukocyte production after acute myocardial infarction. *Circulation* 132, 1880–1890. doi: 10.1126/scitranslmed.aaf1435

Satoh, M., Shimoda, Y., Maesawa, C., Akatsu, T., Ishikawa, Y., Minami, Y., et al. (2006). Activated toll-like receptor 4 in monocytes is associated with heart failure after acute myocardial infarction. *Int. J. Cardiol.* 109, 226–234. doi: 10.1016/j.ijcard.2005.06.023

Sena, S., Hu, P., Zhang, D., Wang, X., Wayment, B., Olsen, C., et al. (2009). Impaired insulin signaling accelerates cardiac mitochondrial dysfunction after myocardial infarction. *J. Mol. Cell Cardiol.* 46, 910–918. doi: 10.1016/j.yjmcc.2009.02.014

Shalaby, F., Rossant, J., Yamaguchi, T. P., Gertsenstein, M., Wu, X.-F., Breitman, M. L., et al. (1995). Failure of blood-island formation and vasculogenesis in Flk-1-deficient mice. *Nature* 376, 62–66. doi: 10.1038/376062a0

Shen, G.-H., Song, Y., Yao, Y., Sun, Q.-F., Jing, B., Wu, J., et al. (2020). Downregulation of DLGAP1-Antisense RNA 1 Alleviates Vascular Endothelial Cell Injury Via Activation of the Phosphoinositide 3-kinase/Akt Pathway Results from an Acute Limb Ischemia Rat Model. *Eur. J. Vasc. Endovasc. Surg.* 59, 98–107. doi: 10.1016/j.ejvs.2019.06.032

Shi, X., Cheng, L., Jiao, X., Chen, B., Li, Z., Liang, Y., et al. (2018a). Rare copy number variants identify novel genes in Sporadic total anomalous pulmonary vein connection. *Front. Genet.* 9:559. doi: 10.3389/fgene.2018.00559

Shi, X., Huang, T., Wang, J., Liang, Y., Gu, C., Xu, Y., et al. (2018b). Next-generation sequencing identifies novel genes with rare variants in total anomalous pulmonary venous connection. *EBiomedicine* 38, 217–227. doi: 10.1016/j.ebiom.2018.11.008

Shibata, M., Endo, S., Inada, K., Kuriki, S., Harada, M., Takino, T., et al. (1997). Elevated plasma levels of interleukin-1 receptor antagonist and interleukin-10 in patients with acute myocardial infarction. *J. Interferon Cytokine Res.* 17, 145–150. doi: 10.1089/jir.1997.17.145

Shimpo, M., Morrow, D. A., Weinberg, E. O., Sabatine, M. S., Murphy, S. A., Antman, E. M., et al. (2004). Serum levels of the interleukin-1 receptor family member ST2 predict mortality and clinical outcome in acute myocardial infarction. *Circulation* 109, 2186–2190. doi: 10.1161/01.cir.0000127958.21003.5a

Sieber, P., Schafer, A., Lieberherr, R., Le Goff, F., Stritt, M., Welford, R. W. D., et al. (2018). Novel high-throughput myofibroblast assays identify agonists with therapeutic potential in pulmonary fibrosis that act via EP2 and EP4 receptors. *PLoS One* 13:e0207872. doi: 10.1371/journal.pone.0207872

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Thirunavukkarasu, M., Addya, S., Juhasz, B., Pant, R., Zhan, L., Surrey, S., et al. (2008). Heterozygous disruption of Flk-1 receptor leads to myocardial ischaemia reperfusion injury in mice: application of affymetrix gene chip analysis. *J. Cell Mol. Med.* 12, 1284–1302. doi: 10.1111/j.1582-4934.2008.00269.x

Thygesen, K., Alpert, J. S., and White, H. D. (2007). Universal definition of myocardial infarction. *Eur. Heart J.* 50, 2173–2195.

Tomihira, M., Kawasaki, E., Nakajima, H., Imamura, Y., Sato, Y., Sata, M., et al. (2004). Intermittent and recurrent hepatomegaly due to glycogen storage in a patient with type 1 diabetes: genetic analysis of the liver glycogen phosphorylase gene (PYGL). *Diabetes Res. Clin. Pract.* 65, 175–182. doi: 10.1016/j.diabres.2003.12.004

Van De Meerakker, J. B., Van Engelen, K., Mathijssen, I. B., dit Deprez, R. H. L., Lam, J., Wilde, A. A., et al. (2011). A novel autosomal dominant condition consisting of congenital heart defects and low atrial rhythm maps to chromosome 9q. *Eur. J. Hum. Genet.* 19, 820–826. doi: 10.1038/ejhg.2011.33

Vanhaverbeke, M., Vausort, M., Veltman, D., Zhang, L., Wu, M., Laenen, G., et al. (2019). Peripheral Blood RNA Levels of QSOX1 and PLBD1 are new independent predictors of left ventricular Dysfunction After acute myocardial infarction. *Circulation* 12:e002656. doi: 10.1161/CIRCGEN.119.002656

Wang, D., Li, J. R., Zhang, Y. H., Chen, L., Huang, T., and Cai, Y. D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 9:155. doi: 10.3390/genes9030155

Wang, G.-K., Zhu, J.-Q., Zhang, J.-T., Li, Q., Li, Y., He, J., et al. (2010). Circulating microRNA: a novel potential biomarker for early diagnosis of acute myocardial infarction in humans. *Eur. Heart J.* 31, 659–666. doi: 10.1093/eurheartj/ehq013

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

Weinberg, E. O., Shimpo, M., De Keulenaer, G. W., MacGillivray, C., Tominaga, S.-I., Solomon, S. D., et al. (2002). Expression and regulation of ST2, an interleukin-1 receptor family member, in cardiomyocytes and myocardial infarction. *Circulation* 106, 2961–2966. doi: 10.1161/01.cir.0000038705.69871.d9

White, H. D., and Chew, D. P. (2008). Acute myocardial infarction. *Etiology* 372, 570–584.

Wollert, K. C., Kempf, T., Peter, T., Olofsson, S., James, S., Johnston, N., et al. (2007). Prognostic value of growth-differentiation factor-15 in patients with non-ST-elevation acute coronary syndrome. *Circulation* 115:962. doi: 10.1161/circulationaha.106.650846

Wu, K., Zhao, Q., Li, Z., Li, N., Xiao, Q., Li, X., et al. (2018). Bioinformatic screening for key mi RNA s and genes associated with myocardial infarction. *FEBS Open Bio* 8, 897–913. doi: 10.1002/2211-5463.12423

Yan, X., Yu-Hang, Z., JiaRui, L., Xiaoyong, P., Tao, H., and Yu-Dong, C. (2019). New computational tool based on machine-learning algorithms for the identification of rhinovirus infection-related genes. *Comb. Chem. High Throughput Screen.* 22, 1–1. doi: 10.2174/1386207322666191129114741

Yao, H., Han, X., and Han, X. (2014). The cardioprotection of the insulin-mediated PI3K/Akt/mTOR signaling pathway. *Am. J. Cardiovasc. Drugs* 14, 433–442. doi: 10.1007/s40256-014-0089-9

Zhang, S., Liu, W., Liu, X., Qi, J., and Deng, C. J. M. (2017). Biomarkers identification for acute myocardial infarction detection via weighted gene co-expression network analysis. *Medicine* 96:e8375. doi: 10.1097/MD.0000000000008375

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

Zimetbaum, P. J., and Josephson, M. E. (2003). Use of the electrocardiogram in acute myocardial infarction. *N. Engl. J. Med.* 348, 933–940. doi: 10.1056/nejmra022700

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 July 2020
doi: 10.3389/fgene.2020.00740

Computational Identification of 29 Colon and Rectal Cancer-Associated Signatures and Their Applications in Constructing Cancer Classification and Prognostic Models

Ran Wei[1†], Hengchang Liu[1†], Chunxiang Li[2], Xu Guan[1], Zhixun Zhao[1], Chenxi Ma[1], Xishan Wang[1] and Zheng Jiang[1]*

[1] Department of Colorectal Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China, [2] Department of Thoracic Surgery, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co., Ltd., China

Reviewed by:
Tianbao Li,
University of Texas Health Science Center at San Antonio, United States
Ju Xiang,
Changsha Medical University, China

***Correspondence:**
Zheng Jiang
071106237@fudan.edu.cn

[†]These authors have contributed equally to this work

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

Received: *12 May 2020*
Accepted: *22 June 2020*
Published: *22 July 2020*

Citation:
Wei R, Liu H, Li C, Guan X, Zhao Z, Ma C, Wang X and Jiang Z (2020) Computational Identification of 29 Colon and Rectal Cancer-Associated Signatures and Their Applications in Constructing Cancer Classification and Prognostic Models. Front. Genet. 11:740. doi: 10.3389/fgene.2020.00740

Systematic classification of colon and rectal cancer-associated signatures is critical for the classification and prognosis of cancer patients. In this study, we identified a panel of 29 colon and rectal cancer-associated signatures from bioinformatics analyses on both TCGA and GEO datasets. Based on the signatures, we developed a machine learning method to classify colon and rectal cancer into three immune subtypes named High-Immunity Subtype, Medium-Immunity Subtype, and Low-Immunity Subtype, respectively. Reconfirmed by different datasets, this classification was associated with the tumor mutational burden (TMB) and many cancer-associated pathways. Compared to Medium-Immunity and Low-Immunity, patients with High-Immunity Subtype have a greater immune cell infiltration and better survival prognosis. In addition, a prognostic signature of six differentially−expressed and survival-associated genes among the three cancer subtypes (*CERCAM*, *CD37*, *CALB2*, *MEOX2*, *RASGRP2*, and *PCOLCE2*) was identified by the multivariable COX analysis, which was further used to develop an accurate model to predict the prognosis of colon and rectal cancer patients.

Keywords: Colon and rectal cancer, signature, prognosis, immunogenomic profiling, machine learning

INTRODUCTION

Colon and rectal tumors are among the most lethal and common malignancies after lung and prostate cancer (Sanoff et al., 2007; Wilkinson et al., 2010; Bray et al., 2018). It has been estimated that 53,990 new cases would be diagnosed in 2019 in the United States alone (Yothers et al., 2013). Distant metastasis is the main factor affecting the overall survival (OS) of patients with colon cancer, and prevention can reduce its incidence (Sanoff et al., 2007; Bray et al., 2018). Nevertheless, mortality remains high in case of advanced disease (Wilkinson et al., 2010). In patients with locally advanced or distantly metastatic colon cancer, conventional treatments are often insufficient to achieve a curative effect (Pagès et al., 2018; Wang Y. et al., 2018). Consequently, early detection

and monitoring of the development of colon cancer using sensitive biomarkers could increase the proportion of patients diagnosed before the onset of aggressive disease.

Immunotherapy is a significant part of precision medicine in oncotherapy, enhancing the ability of the host immune system to fight advanced cancer types (Becht et al., 2016; Gutting et al., 2019). In recent decades, cutting-edge immunotherapies offered the promise of alternative treatment methods for many types of cancer (Sharma and Allison, 2015; Palucka and Coussens, 2016). Recent studies indicate that inhibiting immune checkpoint receptors expressed on T cells can boost the elimination of colon cancer cells *in vivo*. Furthermore, programmed cell death protein 1 (PD-1) and cytotoxic T lymphocyte associated antigen-4 (CTLA-4) have been proved as effective targets for the treatment of patients with immunogenic tumors, especially in mismatch repair-deficient colon cancer and melanoma (Brahmer et al., 2012; Sasidharan Nair et al., 2018). Some studies also showed that MSI tumor classification may be a predictive biomarker for PD-1 inhibition because of its association with increased expression of PD-1 and other immune-checkpoint molecules (Basile et al., 2017; Chouhan and Sammour, 2018). At the same time, multiple studies investigated tumor immunology in colon cancer (Kather and Halama, 2019). The colon is not only one of the most significant digestive organs, but also contains the largest accumulation of immune cells in the body, which regulate this very large immune barrier (Fletcher et al., 2018). Some studies have showed that ulcerative colitis, which is partly considered an autoimmune disease, can promote the development of colon cancer, but the underlying signaling mechanism needs further research (Bopanna et al., 2017; Lopez et al., 2018). Due to the abundant immune cells in the colon cancer microenvironment (Fridman et al., 2012), the type, density, and location of diverse immune cells is a promising resource for predicting the clinical outcomes. In addition, the evaluation of the extent of tumor-infiltration by T-lymphocytes, macrophages and mast cells could be considered as a significant biomarker for TNM staging and prognosis (Yang et al., 2017; Han et al., 2018). Indeed, the density of T-lymphocytes and mast cells should be treated as a widely available prognostic biomarker in colon and rectal cancer, which is related to their functions in immune suppression, inflammation, and tumor development (Marech et al., 2014; Lv et al., 2019). The cancer microenvironment also commonly consists of stromal cells originating from the mesenchyma, which can regulate immune cell trafficking and activation to influence the prognosis of different cancer types and disease stages (Greten et al., 2004; Koliaraki et al., 2015). In order to promote the development of effective immunotherapy strategies, it is important to investigate the immunomodulatory role of the immune and stromal compartments of tumors. By combining different immunotherapeutic methods with other therapeutic approaches, and paying attention to the association between immunotherapy response and the tumor mutation burden (TMB), it is possible to significantly improve the efficacy of cancer therapy.

In this study, we used the "Cell type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT)" algorithm, which employs support vector regression and has already been employed for immune score model construction in several cancer types (Newman et al., 2015; Zeng et al., 2018). Furthermore, we classified both rectal and colon cancer into three distinct subtypes: High-Immunity Subtype, Medium-Immunity Subtype, and Low-Immunity Subtype using immunogenomic profiling based on "Estimation of Stromal and Immune cells in Malignant Tumors using Expression data (ESTIMATE)" (Yoshihara et al., 2013; Vincent et al., 2015). We employed CIBERSORT and ESTIMATE to evaluate the proportions of immune cells and subtype-specific molecular features in samples from 870 colon and rectal cancer patients and 70 normal controls based on gene expression profiles available in public databases. This investigation aimed to assess the potential clinical utility of differentially expressed genes form distinct subtypes for prognostic stratification and their potential as biomarkers for targeted colon and rectal cancer therapy. Additionally, we explored underlying functional signaling mechanisms via bioinformatic analyses. The results of this study lay a great promise and foundation for subsequent in-depth immune-related studies for the precision treatment of colon and rectal cancer.

MATERIALS AND METHODS

RNA-Sequencing Data and Bioinformatics Analysis

Transcriptomic RNA-sequencing data of colon cancer and rectal cancer patients were obtained from The Cancer Genome Atlas (TCGA)[1], which contained data from a colon adenocarcinoma (COAD, n = 467) cohort and rectal adenocarcinoma (READ, n = 172) tissues. The exclusion criteria were normal COAD and READ samples and an OS of <30 days. Besides level 3 HTSeq-FPKM data were transformed into TPM (transcripts per million reads) for the following analyses. The TPM data for 430 patients with COAD were employed for further analyses. Gene expression datasets of colon cancer and rectal cancer patients obtained using an GPL570 platform were searched against the gene expression omnibus (GEO)[2]. The raw CEL files of matching microarray data were processed using the robust multichip average algorithm (Irizarry et al., 2003). Then, microarray presets could be mapped to gene symbols according to the platform annotation file and normalized employing a robust multi-array averaging method using the "affy" and "simpleaffy" packages (Irizarry et al., 2003).

Implementation of Single-Sample Gene Set Enrichment Analysis (ssGSEA)

The R package gsva was used for quantitative ssGSEA of infiltrating immune cell types. The gene signatures of immune cell populations could be applied to individual colon and rectal cancer samples with the ssGSEA (Barbie et al., 2009; Bindea et al., 2013). The enrichment levels of 29 immune signatures which are related to innate immunity [CD56 bright natural killer (NK) cells, NK cells, CD56dim NK cells, plasmacytoid dendritic cells

[1]https://cancergenome.nih.gov/

[2]https://www.ncbi.nlm.nih.gov/geo/

(DCs), activated DCs, immature DCs, neutrophils, eosinophils, monocytes, mast cells, and macrophages] and adaptive immunity (activated B cells, immature B cells, activated $CD4^+$ T cells, effector memory $CD4^+$ T cells, central memory $CD4^+$ T cells, central memory $CD8^+$ T cells, effector memory $CD8^+$ T cells, activated $CD8^+$ T cells, T follicular helper cells, NK T cells, Tγδ, Th1, Th2, Th17, and Treg), were quantified in each sample based on the ssGSEA score. Finally, hierarchical clustering of colon and rectal cancer was conducted on the basis of ssGSEA scores of the 29 immune signatures.

Evaluation of Immune Cell Infiltration Levels, Tumor Purity, and Stromal Content in Colon and Rectal Cancer

Estimation of Stromal and Immune cells in Malignant Tumors using Expression data was employed to analyze the stromal content (stromal score), tumor purity, and immune cell infiltration level (immune score) for colon and rectal cancer sample (He et al., 2018).

Comparison of the Proportions of Immune Cell Subsets Between Colon and Rectal Cancer Subtypes

The transcriptomic RNA-sequencing data with standard annotation were uploaded to the CIBERSORT web portal[3], and the algorithm was run employing the LM22 signature with 1000 permutations (Newman et al., 2015). The inferred fractions of immune cell populations produced by CIBERSORT were considered accurate if the CIBERSORT output had a $p < 0.05$ (Ali et al., 2016), and were considered eligible for further analysis. The final CIBERSORT output estimates were normalized for each sample to add up to one, enabling their direct interpretation as cell fractions for comparison across different datasets and immune cell types. For parts of each immune cell type, the optimal cut-off value was made as the point with the most important split (log-rank test) (Budczies et al., 2012).

Identification of Differentially Expressed Genes

The statistical software R (version3.5.2) and the Bioconductor linear model package for microarray data "limma"[4] were used to identify the differentially expressed genes (DEGs) between the High-Immunity Subtype and Low-Immunity Subtype (FDR < 0.05) colon and rectal cancer tissues in TCGA (He et al., 2018). DEGs were defined by a p-value < 0.05 and $|\log_2 \text{FoldChange}| > 1$. For genes corresponding to multiple probe sets, the average data of the multiple probes were used as the gene expression values (Wei et al., 2018). The values of genes over 20% of the total samples were eliminated (Qin et al., 2012). After pre-processing the data, the Wilcoxon signed rank test was used to select significant DEGs using the "limma" package in Bioconductor (He et al., 2018).

[3]http://cibersort.stanford.edu/

[4]http://www.bioconductor.org/

Identification of Colon Cancer Subtype-Specific Gene Ontology and Networks

The step-by-step method of the weighted gene co-expression network analysis (WGCNA) in R was employed to identify the gene modules (gene ontology) and construct the module and network that were significantly related with the genes highly correlated with immune cell infiltration based on gene co-expression analysis (He et al., 2018). The adjacency matrix and the topological overlap matrix (TOM) was used to calculate according to the corresponding soft threshold, and the corresponding dissimilarities between each gene were calculated. We employed the dynamic tree cut method, and the branches of the hierarchical cluster tree would be cut to identify modules.

Gene Set Enrichment Analysis

H: Hallmark gene sets; C2: curated gene sets [including Kyoto Encyclopedia of Genes and Genomes (KEGG)]; C5: Gene Ontology (GO) gene sets; C7: immunologic signatures gene sets v6.2 collections were downloaded from Molecular Signatures Database as the target gene sets with which GSEA performed using the software gsea-3.0. The whole transcriptome of all tumor samples was used for GSEA, and only gene sets with NOM $p < 0.05$ and FDR $q < 0.05$ were considered as significant.

Survival Analysis

The R package clusterProfiler was employed to conduct gene functional enrichment analyses to identify biological themes among gene clusters (Yu et al., 2012). The R package survival receiver operating characteristic (ROC) was used to calculate the AUC of the survival ROC curves to validate the performance of the prognostic signature (Sun, 2017; Lin et al., 2019). Kaplan–Meier curves were plotted to verify the statistical relationship between genes and the OS of the high-risk group and low-risk groups from the TCGA datasets with the log-rank tests. Using multivariate Cox proportional hazard regression to identify prognostic clinicopathologic factors for OS in colon and rectal cancer patients. They were utilized to verify the differences of survival between the patients in the two different risk groups. The six-gene signature and nomogram were developed from the final (forward and backward elimination methods) Cox model to predict the OS of colon and rectal cancer patients. Besides the performance of the prediction model was validated internally and externally by bootstrap method. Bootstrap-corrected OS rates were calculated by averaging the Kaplan-Meier estimates based on 2000 bootstrap samples.

RESULTS

Immunogenomic Profiling Identifies Three Colon and Rectal Cancer Subtypes

Figure 1 shows a schematic representation of the process for selecting colon and rectal cancer samples. A total of 735 patients with complete overall survival information were

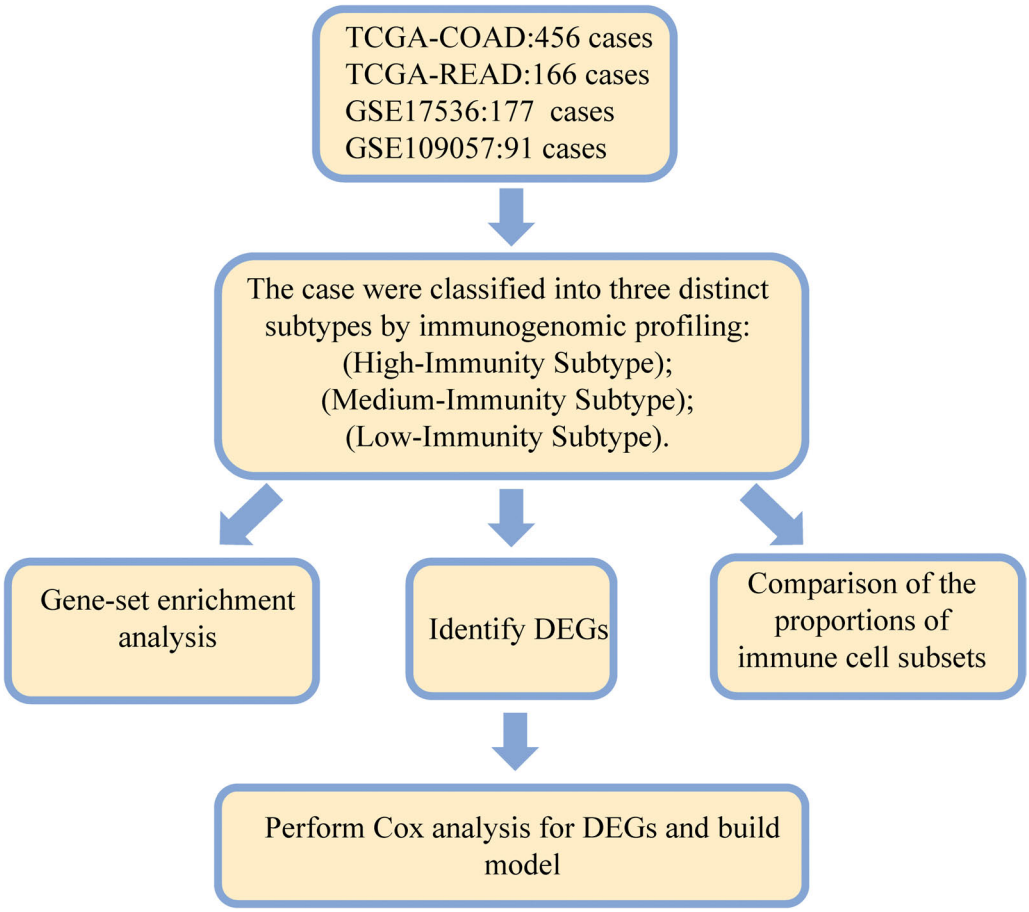

FIGURE 1 | Flow diagram of the analysis procedure: data collection, preprocessing, and analysis. DEGs, differentially expressed genes; ROC, receiver operating characteristic; TCGA, The Cancer genome atlas; GEO, Gene Expression Omnibus.

included from TCGA, GSE17536, and GSE109057. A total of 29 immune-associated gene sets, representing diverse immune cell types, functions, and pathways, were analyzed in the datasets via the ssGSEA scores (Yoshihara et al., 2013; Vincent et al., 2015) to quantify the enrichment levels of immune cells, pathways or functions in the colon and rectal cancer samples. The ssGSEA scores of the 29 gene sets from these microarray datasets were then used to conduct hierarchical clustering, which revealed three types of colon and rectal cancer (**Figure 2**). We classified the three clusters as: High-Immunity Subtype, Medium-Immunity Subtype, and Low-Immunity Subtype, and the immune scores were higher in High-Immunity Subtype and lower in Low-Immunity Subtype (Yoshihara et al., 2013; Vincent et al., 2015). In addition, we found that tumor purity and stromal score of the three colon and rectal cancer subtypes had opposite trends (**Figure 2**).

Composition of Immune Cells in Three Colon and Rectal Cancer Subtypes

The result showed that the High-Immunity Subtype have significantly higher immune scores than Low-Immunity Subtype in colon and rectal cancer (**Figures 3A,B**). Mann–Whitney U test. $^{**}p < 0.01$; $^{***}p < 0.001$; $p \geq 0.05$, not significant. Besides the levels of TMB were similar with immune scores in High-Immunity Subtype and Low-Immunity Subtype, which showed that the TMB was associated with different Immunity types **Figures 3C,D**. Kruskal–Wallis rank sum test. $^{**}p < 0.01$; $^{***}p < 0.001$; $p \geq 0.05$, not significant. Owing to the significant value of 29 immune-associated gene, we tended to establish a comprehensive exploration of these genes' molecular characteristics. The result of genetic alterations testing showed that Missense Mutation was commonly occurring type of mutation (**Figure 3E**). Besides We embarked on the immune cell constitution in colon and rectal cancer tissues versus normal colon tissues in **Figure 4B**. From the results, the fractions of M1 macrophages, activated $CD4^+$ memory T cells, M1 macrophages, activated NK cells, and neutrophils were consistently higher in the High-Immunity Subtype than in the Low-Immunity Subtype in colon cancer. The fractions of activated $CD8^+$ memory T cells, B cells and Plasma cells were consistently higher in the High-Immunity Subtype than in the Low-Immunity Subtype in rectal cancer. And a summary of the immune cell composition in tumor cases showed that macrophages M1, macrophages M2, mast cells,

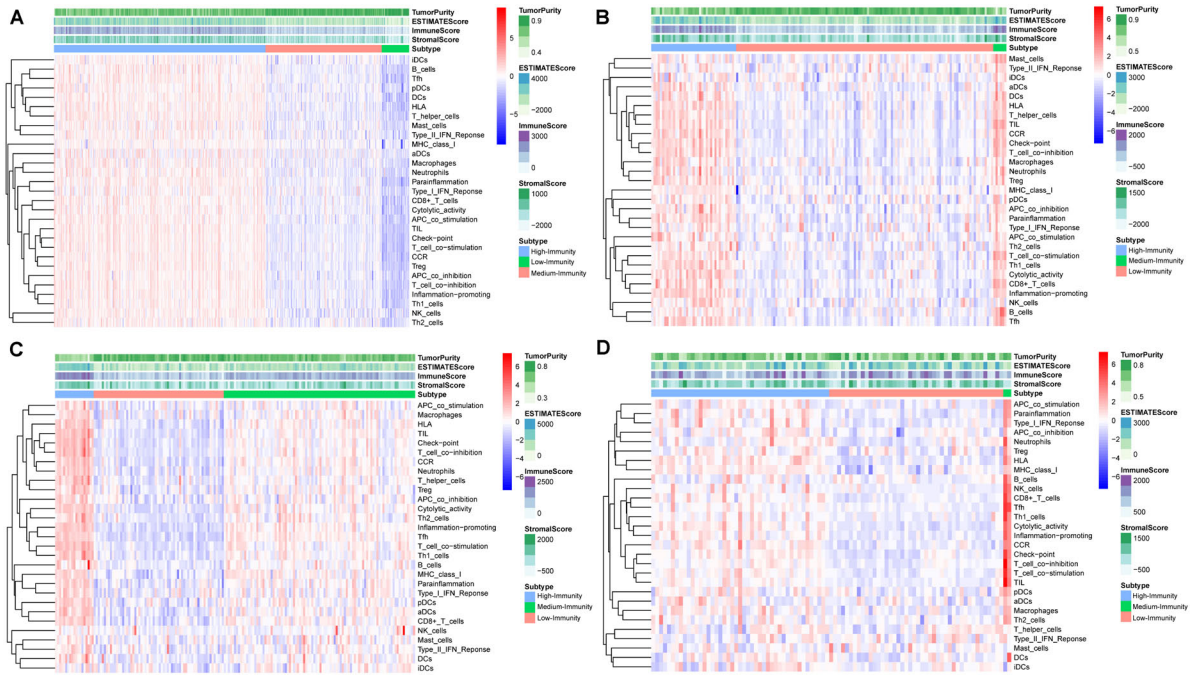

FIGURE 2 | Hierarchical clustering of colon and rectal cancer yields three stable subtypes in four different datasets named High-Immunity Subtype, Medium-Immunity Subtype and Low-Immunity Subtype. Tumor purity, Stromal score, and Immune score were evaluated by ESTIMATE. **(A)** The colon cancer patients in TCGA-COAD database. **(B)** The rectal cancer patients in TCGA-READ database. **(C)** The colon cancer patients in GSE17536 database. **(D)** The colon cancer patients in GSE109057 database.

T cells and neutrophils were most common immune cell fractions in colon and rectal cancer in **Figures 4A,D**. Besides different types of immune cells affect each other's fractions, macrophages M0 in high fractions may decrease the fractions of activated $CD8^+$ memory T cells (**Figures 4C,E**).

Expression of Genes on Immune Cells Showed Significantly Higher Expression Levels in High-Immunity Subtype

We found that the expression of most HLA genes were significantly higher in High-Immunity Subtype than in Low-Immunity Subtype (Kruskal-Wallis test, $P < 0.001$) (**Figure 5**). Besides the expression of various immune cell subpopulation marker genes (Yoshihara et al., 2013) were the highest in High-Immunity Subtype and the lowest in Low-Immunity Subtype, such as CD8A ($CD8^+$ T cells), TNFSF14 (APC co stimulation), CD79A (B cells), CD28 (Tumor Infiltrating Lymphocyte), and CD28 (T cell co-stimulation) in colon and rectal cancer (**Figure 5**). ANOVA test. $P < 0.01$; $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

Identification of Subtype-Specific Pathways, and Gene Ontology of Colon and Rectal Cancer

We employed the GSEA to indent the KEGG pathways and gene ontology enriched in High-Immunity Subtype and Low-Immunity Subtype (**Figures 6A,B**). Notably, the positive regulation of humoral immune response, up-regulation of mast cell activation associated with immune response, regulation of T-helper 1,2 cell differentiation and establishment of T cell polarity. Besides the pathways on Immunity moderation were highly increased in High-Immunity Subtype and included antigen processing and presentation pathways, NF-kappa B signaling, p53 signaling pathway, VEGF signaling pathway, Hippo signaling pathway, PI3K-Akt and mTOR signaling pathway and MAPK signaling pathway, which proved that the immune activity was promoted in High-Immunity Subtype. And some previous study proved that the promotion of PI3K-Akt and MAPK cascades positively associated with the elevated of various immune pathways (Sun, 2017). Besides the immune scores were related with colon cancer Stage. The immune scores of Stage IV was lower than Stage I. Based on the selection criteria after preprocessing the raw data, we identified the DEGs of High-Immunity Subtype and Low-Immunity Subtype in TCGA-COAD and TCGA-READ. 2378 DEGs between High-Immunity Subtype and Low-Immunity Subtype colon cancer were identified in TCGA-COAD dataset. The DEGs were analyzed for co-expression network analysis with employing the WGCNA package, and finally, a total of 18 modules were identified. The ME in the brown, yellow, red and pink modules showed significantly higher association with cancer progression than other modules. And more, the four modules with cancer development was identified as the clinically significant module, which was selected for further analysis (**Figures 6C,E–F**). Kaplan–Meier curves for OS based on three colon cancer

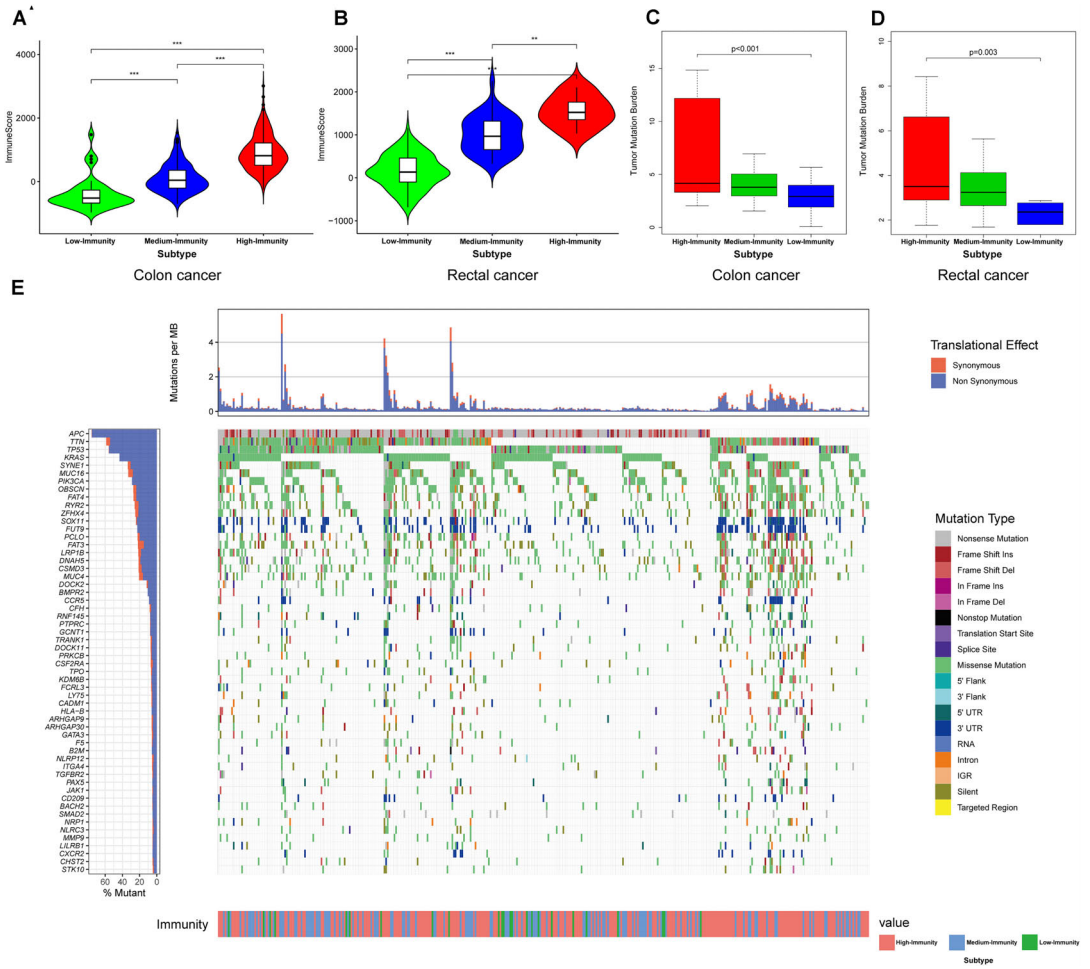

FIGURE 3 | Three Immunity subtypes show differential phenotypes. **(A,B)** Comparison of the immune cell infiltration levels between three colorectal cancer subtypes. Mann–Whitney *U* test. **(C,D)** Comparison of the TMB levels between three colorectal cancer subtypes. Kruskal–Wallis rank sum test. $^{**}p < 0.01$; $^{***}p < 0.001$; $p \geq 0.05$, not significant. **(E)** Mutation landscape of immune-related genes.

immune subtypes. The High-Immunity Subtype had the best survival, whereas other classes were associated with poor outcome (**Figure 6D**). Log-rank test, $p = 0.008$. We screened hub DEGs with excellent biomarker potential to evaluate prognosis between three immunity types in colon cancer. A forest plot of expression profiles based on multivariate Cox regression analysis revealed that this immune-based prognostic index could be a significant tool for the assessment of colon cancer prognosis (**Figure 6G**). And the expression of six DEGs were higher in High-Immunity Subtype than in Low-Immunity Subtype in colon cancer (**Figure 6H**). ANOVA test $^{**}P < 0.01$; $^{***}P < 0.001$.

Correlation of Immune Cells Proportion With Six-Gene Signature Expression

To further confirm the correlation of six-gene signature expression with the immune microenvironment, and 22 kinds of immune cell profiles in COAD samples were constructed. The results from the difference and correlation analyses showed that lots kinds of immune cells were correlated with the expression of six-gene signature (**Figure 7** and **Supplementary Figures S1**, **S2**). Among them, T cells and Macrophages positively correlated with *CALB2, CD37, CERCAM, MEOX2, RASGRP2, PCOLCE2* expression. The blue line in each plot was fitted linear model indicating the proportion tropism of the immune cell along with six-gene signature expression, and Pearson coefficient was used for the correlation test. These results further supported that the levels of six-gene signature expression affected the immune activity.

Prognostic Value of Overlapped DEGs Between High-Immunity Subtype and Low-Immunity Subtype in Colon and Rectal Cancer

According to the multivariate Cox regression analysis, we established a prognostic signature to divide the colon cancer and rectal cancer patients into two groups with discrete clinical outcomes with regards to OS (**Figure 8**). The prognostic index

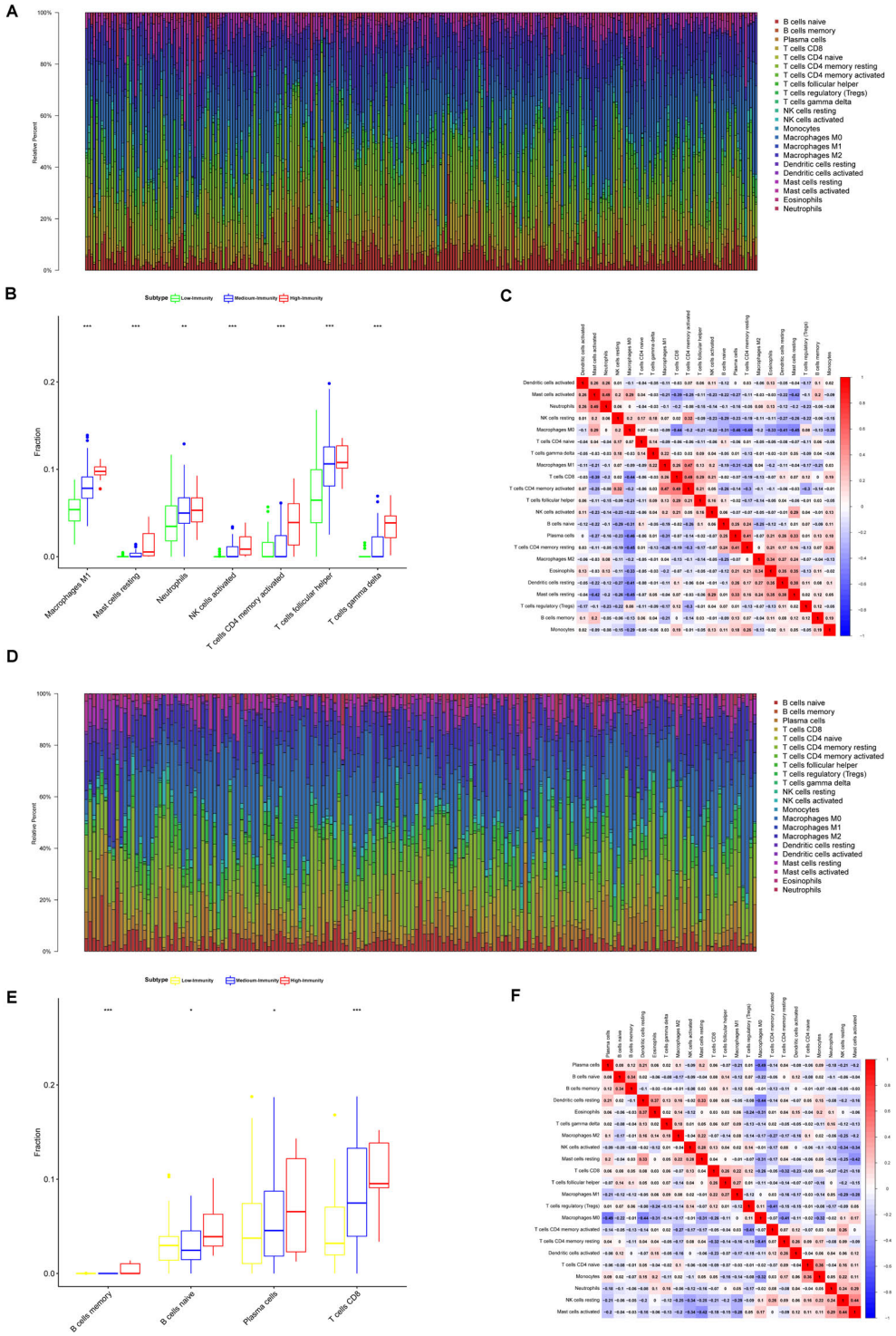

FIGURE 4 | Composition of immune cells in three colon and rectal cancer subtypes and correlation analysis. **(A)** Barplot showing the fractions of 22 immune cells of colon cancer patients in TCGA-COAD database. Column names of plot were sample ID. **(B)** Comparison of the proportions of immune cell subsets between colon cancer subtypes in TCGA-COAD. ANOVA test, *P* values are shown. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $p \geq 0.05$, not significant. **(C)** Heatmap showing the correlation between immune cells of colon cancer cases in TCGA-COAD database. The shade of each tiny color box represented corresponding correlation value between two cells. **(D)** Barplot showing the fractions of 22 immune cells of colon cancer patients in TCGA-READ database. Column names of plot were sample ID. **(E)** Comparison of the proportions of immune cell subsets between rectal cancer subtypes in TCGA-READ database. ANOVA test, *P* values are shown. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $p \geq 0.05$, not significant. **(F)** Heatmap showing the correlation between immune cells of colon cancer cases in TCGA-READ database. The shade of each tiny color box represented corresponding correlation value between two cells.

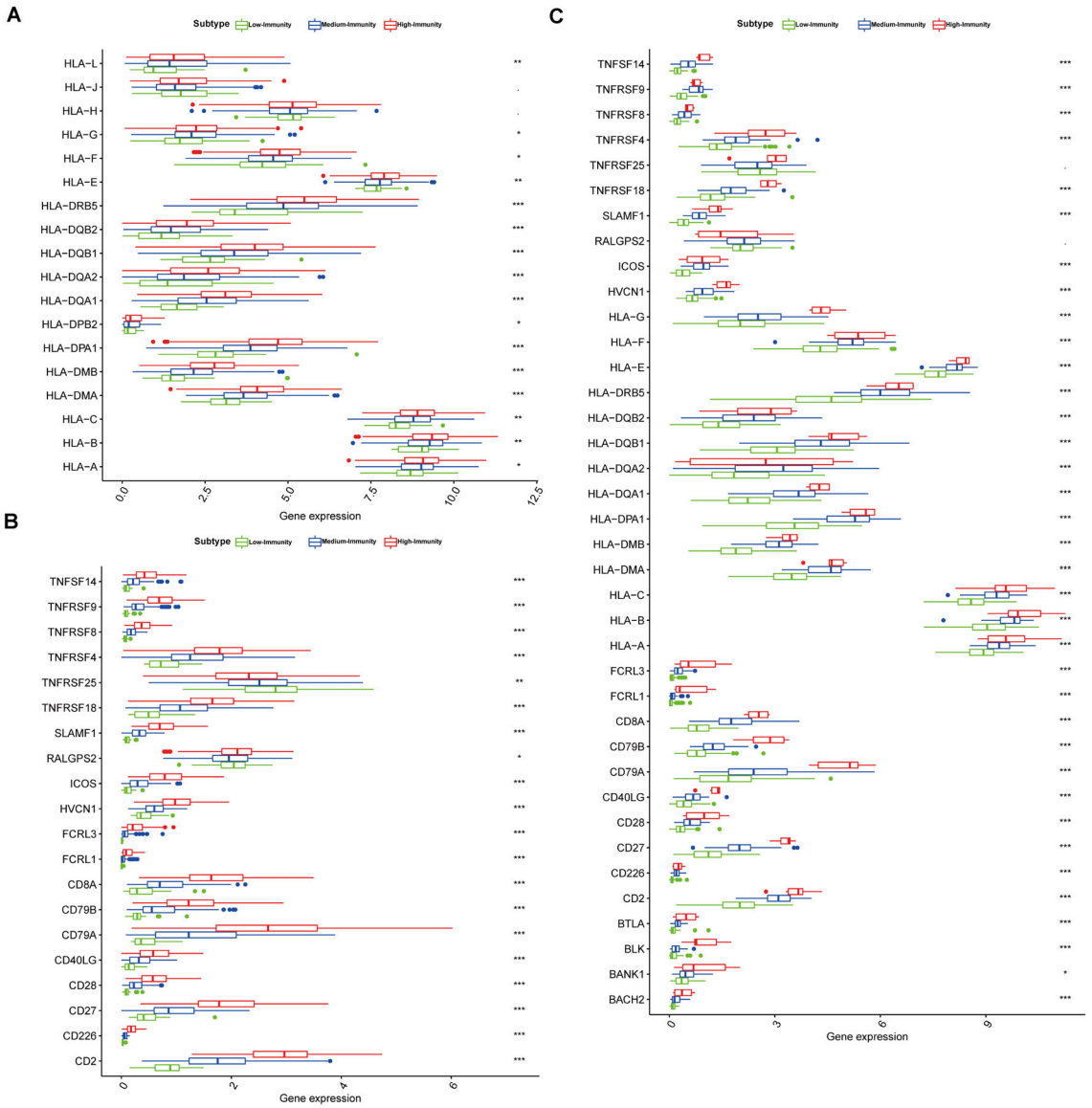

FIGURE 5 | Three colon and rectal cancer subtypes show differential phenotypes. **(A)** Comparison of the expression levels of HLA genes between colon cancer subtypes in TCGA-COAD database. ANOVA test. **(B)** Comparison of the expression levels of genes on immune cells between colon cancer subtypes in TCGA-COAD database. **(C)** Comparison of the expression levels of genes on immune cells between rectal cancer subtypes in TCGA-READ database. ANOVA test $P < 0.01$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

formula for colon cancer was as follows: Risk scores = [Status of *CERCAM* × (0.3314)] + [Status of *CD37* × (−0.5627)] + [Status of *CALB2* × 0.2474] + [Status of *MEOX2* × (−0.5889)] + [Status of *RASGRP2* × (1.0606)] + [Status of *PCOLCE2* × (0.6738)]. This prognostic index based on the immune subtypes could be a valuable tool for distinguishing among colon and rectal cancer patients on the base of potential discrete clinical outcomes. We calculated the risk scores of hub genes and divided the patients into a high-risk group and a low-risk group on the basis of the median risk score in colon and rectal cancer. The correlation of gene expression and survival status is shown in **Figures 8A–D**. The results of survival analysis proved that the OS of the high-risk group was significantly lower than that of the low-risk colon cancer patients (log-rank test, $p < 0.001$). The area under the ROC curve was 0.731 in colon cancer, which indicated a moderate power of the prognostic signature based on DEGs between the high-immunity subtype and low-immunity subtype in survival monitoring. The predictive power of this index for the OS of colon and rectal cancer patients was investigated in the validation cohort (**Figures 8E–H**). The results of this prognostic index suggested a significant difference between the high-risk group and low-risk group with regard to the OS of rectal cancer patients in the validation cohort (log-rank test, $p < 0.05$). A nomogram for predicting the

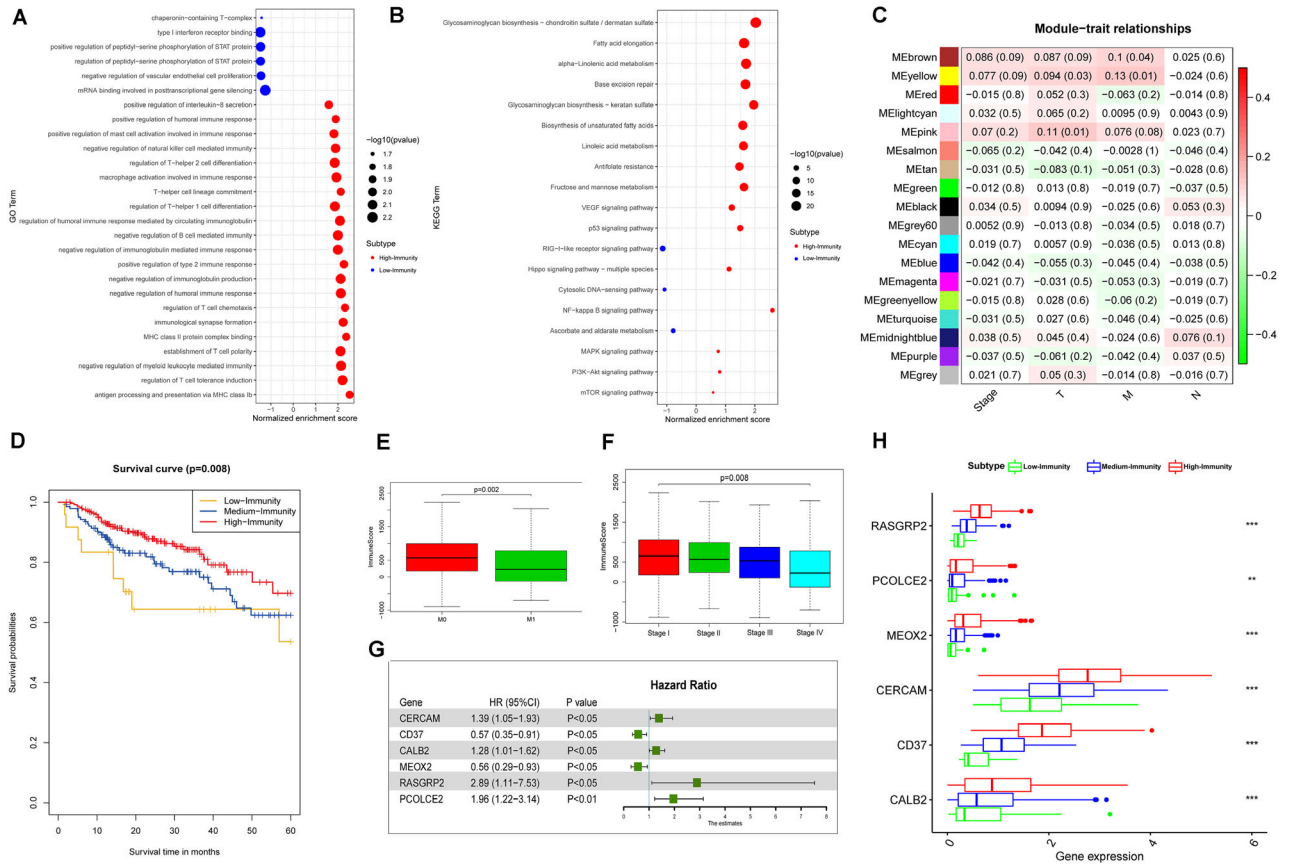

FIGURE 6 | Identification of colon and rectal cancer subtype-specific pathways, gene ontology, DEGs and networks. **(A)** Gene ontology enriched in High-Immunity Subtype and Low-Immunity Subtype of colon cancer in TCGA-COAD database. **(B)** KEGG pathways enriched in High-Immunity Subtype and Low-Immunity Subtype of colon cancer in TCGA-COAD database. **(C)** Heatmap of the correlation between MEs and clinical traits of colon cancer. **(D)** Kaplan–Meier curves for OS based on three colon cancer immune subtypes. The High-Immunity Subtype had the best survival, whereas other classes were associated with poor outcome (log-rank test, $p = 0.008$). **(E,F)** Comparison of the immune cell infiltration levels between Stage and M-Stage in colon cancer. Wilcoxon rank sum or Kruskal–Wallis rank sum test served as the statistical significance test. $^{**}p < 0.01$; $^{***}p < 0.001$. **(G)** Forest plot of hazard ratios showing the prognostic values of five hub genes with multivariable Cox analyses. **(H)** Comparison of the expression levels of five hub genes between colon cancer subtypes in TCGA-COAD database. ANOVA test. $^{**}P < 0.01$, $^{***}P < 0.001$.

3- and 5-year OS was established based on the independent variables (**Figure 9A**). The age, Stage-T, Stage-M, Immunity Type and Six-gene model were further included in the nomogram. A weighted total score calculated from these factors was applied to predict the 3- and 5-year OS of the colon cancer patients. The nomogram cohort was divided into 4 equal groups for validation. The error bars represent the 95% CIs of these estimates. A closer distance between two curves suggests higher accuracy (**Figure 9B**).

The Six-Gene Signature Had Potential to Be Indicators of Immune Microenvironment Modulation

Given the levels of prognostic index risk were negatively correlated with the survival, GSEA was employed in the high-risk and the low-risk groups compared with the median level of risk scores. As shown in **Figure 10A** and **Supplementary Table S2**, for GO collection defined by MSigDB the genes in high-risk group were mainly enriched in immune-related activities, such as the regulation of cytokine, 2 type response and mast cell mediated immunity. For KEGG collection defined by MSigDB, multiple immune functional singling pathways genes sets were enriched in the high-risk group (**Figure 10B** and **Supplementary Table S2**). For HALLMARK collection defined by MSigDB, the genes were enriched in tumor progression-related pathways, including angiogenesis, apoptosis, IL6-JAK and P53 singling pathway (**Figure 10C** and **Supplementary Table S2**). For the immunologic gene sets collection defined by MSigDB, multiple immune functional gene sets were enriched in the high-risk group (**Figure 10D** and **Supplementary Table S2**).

DISCUSSION

Although the significance of classification based on immune signatures in tumor immunotherapy has been established, the

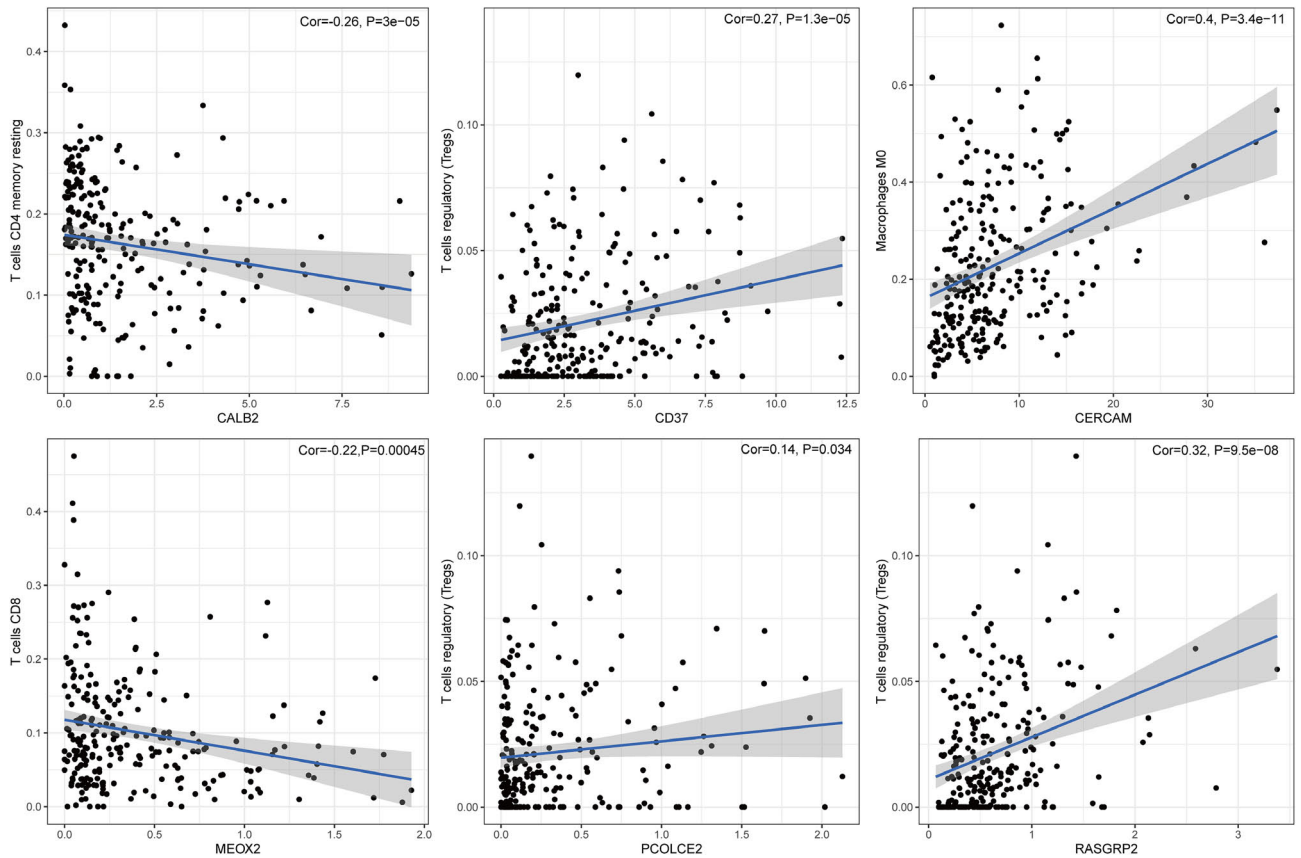

FIGURE 7 | The correlation between immune cell proportion and the expression of six signature genes. The scatter plot showed the correlation of immune cell proportion with the six-gene signature expression ($p < 0.05$). The blue line in each plot was fitted by a linear model indicating the proportion tropism of the immune cell along with six-gene signature expression, and the Pearson coefficient was used for the correlation test.

functions and clinical significance of hub genes have not been explored in colon and rectal cancer. This genome-wide profiling study identified and classified DEGs in colon and rectal cancer, which promotes our understanding of their clinical significance and illuminates potential molecular characteristics. The results show that colon and rectal cancer could be classified into three stable subtypes, a High-Immunity Subtype, Medium-Immunity Subtype, and Low-Immunity Subtype, which were reproducible and predictable. The High-Immunity colon and rectal cancer subtype was enriched in immune response activating and regulating cancer-associated pathways, including the Toll-like receptor signaling pathway, B cell receptor signaling pathway, PI3K-Akt signaling pathway, and NF-κB signaling pathway. Notably, the NF-κB signaling pathway is associated with immune signatures in colon cancer, and it plays a significant role in mediating tumor immunity (Sun, 2017). Moreover, it has a significant negative correlation with the proliferation and differentiation of immune cells as well as the synthesis of immunoglobulins (Su et al., 2017; Wang and Xia, 2018). Additionally, the PI3K-Akt signaling pathway can affect the production of cytokines by T cells and participate in immunosuppression, while mTOR plays a significant role in regulating cell proliferation and protein synthesis, which makes it a promising target for cancer treatment (Lucas et al., 2016; Zheng et al., 2018). The immune signature of the Immunity Low colon cancer subtype was decreased, but enriched in type I interferon receptor binding and serine phosphorylation of STAT protein, which is associates with the regulation of oncogene transcription in tumor apoptosis, proliferation, and angiogenesis (Yu et al., 2009; Li et al., 2017; Zhang et al., 2018). These results indicate the existence of potential positive or negative associations between activation of signaling pathways and immunity in colon and rectal cancer.

The immune context plays a significant role in tumorigenesis and progression, and these insights could influence tumor immunotyping and clinical treatment (Dumauthioz et al., 2018; Locy et al., 2018). Our results showed that the High-Immunity Subtype had stronger immune cell infiltration and anti-tumor immune activity, such as high levels of macrophages, B cells and cytotoxic T cells. Many studies attempted to assess the density of $CD8^+$ and $CD3^+$ lymphocytes in the tumor proper via IHC staining, but the obtained data could not comprehensively reflect the immune cell infiltration and anti-tumor immune activities (Qin et al., 2013; Wong et al., 2018). CIBERSORT was employed to evaluate the proportions of 22 immune cell subsets in colon and rectal cancer, which indicated that $CD8^+$ T cells,

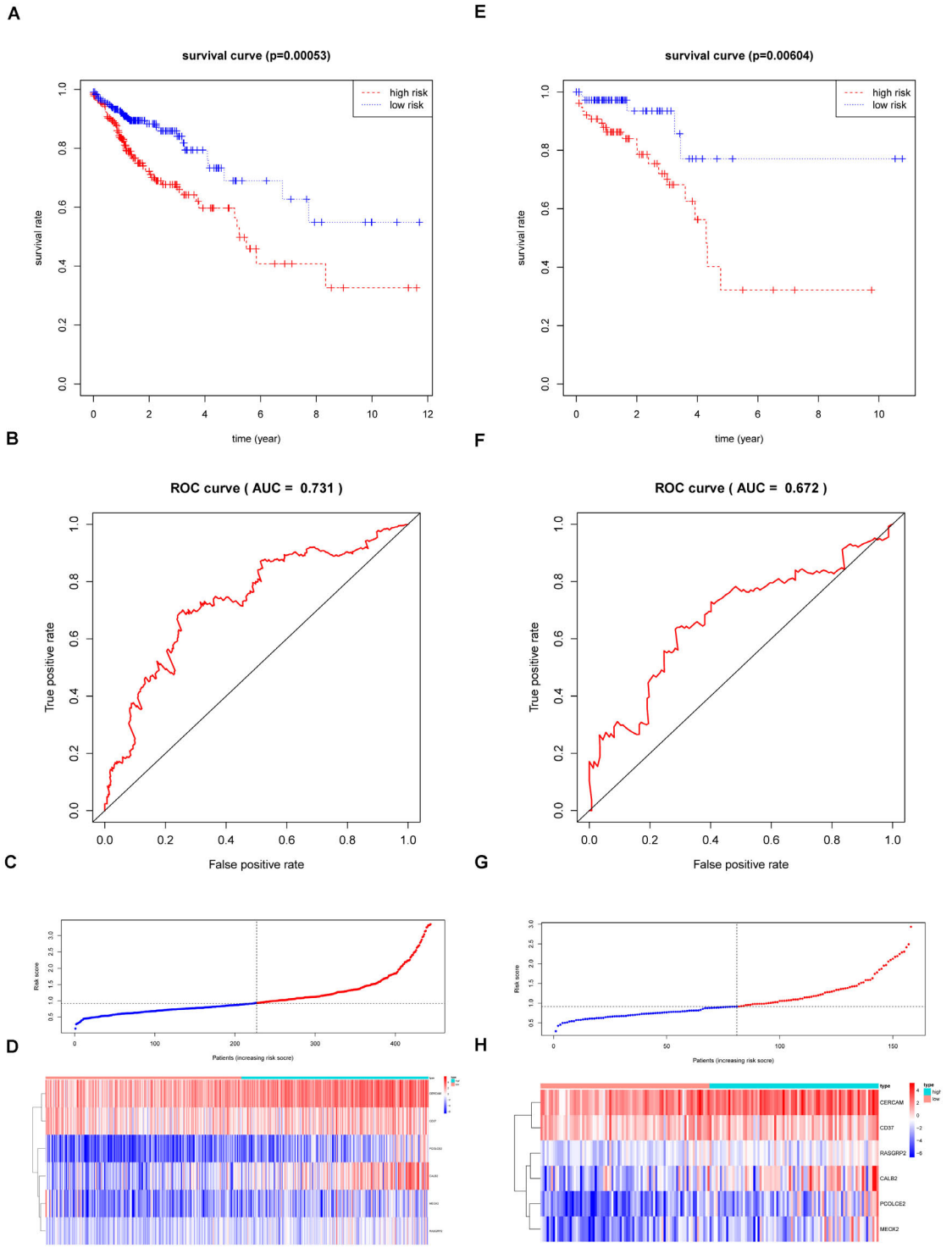

FIGURE 8 | The Survival analysis and prognostic performance of the six-gene signature of colon and rectal cancer. **(A)** The Kaplan–Meier test of the risk score for the overall survival of colon cancer between high-risk and low-risk patients in TCGA-COAD database (log-rank test, $p < 0.001$); **(B)** The prognostic value of the risk score showed by the time-dependent receiver operating characteristic (ROC) curve for predicting the 5 years overall survival. in TCGA-COAD database; **(C)** Risk score curve of the six-gene signature of colon cancer in TCGA-COAD database; **(D)** Heatmap showed the expression of six genes by risk score of colon cancer in TCGA-COAD database; **(E)** The Kaplan–Meier test of the risk score for the overall survival of rectal cancer between high-risk and low-risk patients in TCGA-READ database (log-rank test, $p < 0.001$); **(F)** The prognostic value of the risk score showed by the time-dependent receiver operating characteristic (ROC) curve for predicting the 5 years overall survival in TCGA-COAD database; **(G)** Risk score curve of the six-gene signature of rectal cancer in TCGA-READ database; **(H)** Heatmap showed the expression of six genes by risk score of rectal cancer in TCGA-READ database.

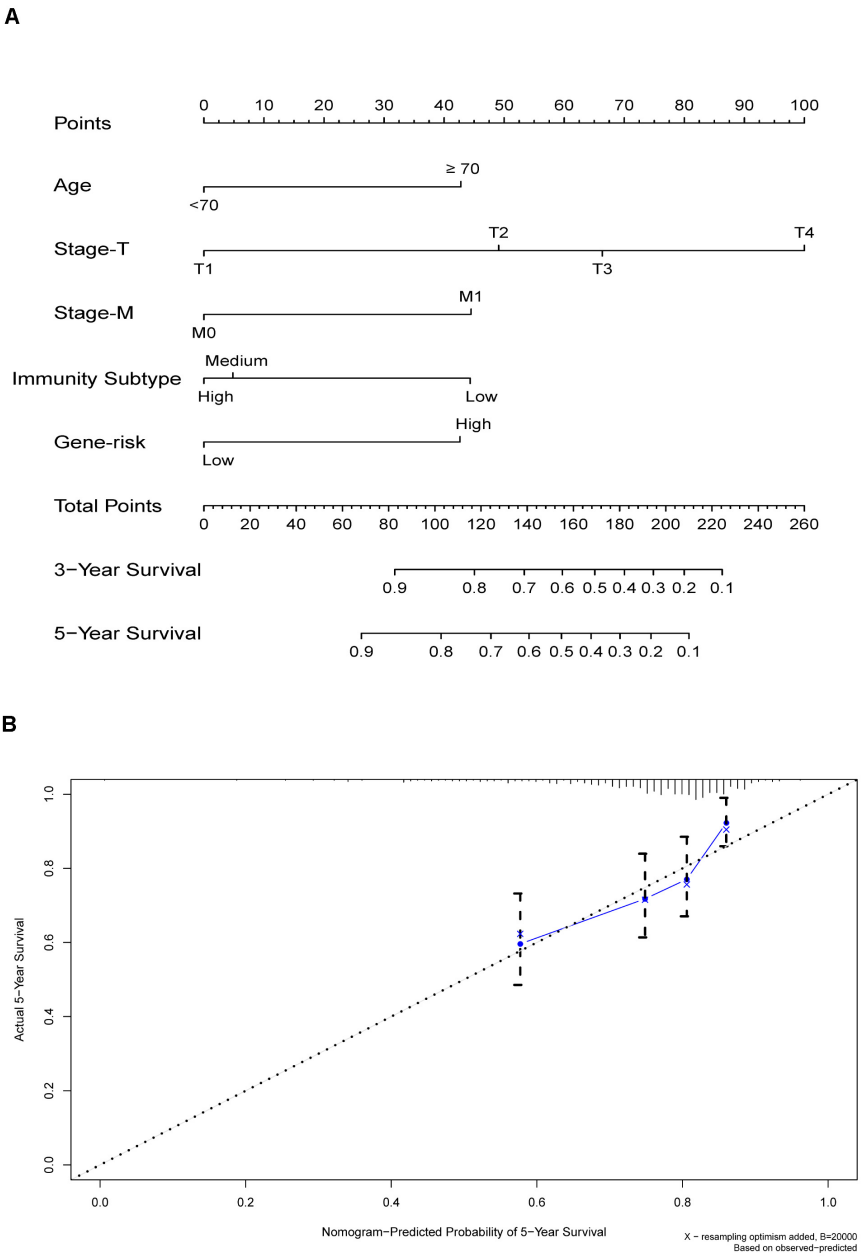

FIGURE 9 | (A) Nomogram for predicting 3- and 5-year OS in colon cancer. To calculate probability of OS, first determine the value for each factor by drawing a vertical line from that factor to the points scale. "Points" is a scoring scale for each factor, and "total points" is a scale for total score. Then sum all of the individual values and draw a vertical line from the total points scale to the 3-, and 5-year OS probability lines to obtain OS estimates. **(B)** The nomogram cohort was divided into four equal groups for validation. The gray line represents the perfect match between the actual (*y*-axis) and nomogram-predicted (*x*-axis) survival probabilities. Black circles represent nomogram-predicted probabilities for each group, and X's represent the bootstrap-corrected estimates. Error bars represent the 95% CIs of these estimates. A closer distance between two curves suggests higher accuracy.

M2 macrophages, M1 macrophages, M0 macrophages and mast cells were present4 in higher numbers in the High-Immunity Subtype than in the Low-Immunity Subtype, which confirmed the elevated anti-tumor immune activity in the High-Immunity Subtype. Macrophages represent the first line of defense against foreign pathogens, recognizing a wide range of endogenous and exogenous ligands via important effectors in innate immunity (Duluc et al., 2009; Rhee, 2016). However, M2 macrophages can release pro-angiogenic molecules and growth factors that promote cancer development, as well as inhibit the antitumor immunity of T cells and NK cells (Pollard, 2004; Lewis and Pollard, 2006; Sica et al., 2006), which is in agreement with the findings of this study. In addition, somatic mutations in tumor DNA could give rise to neoantigens recognizable and targetable

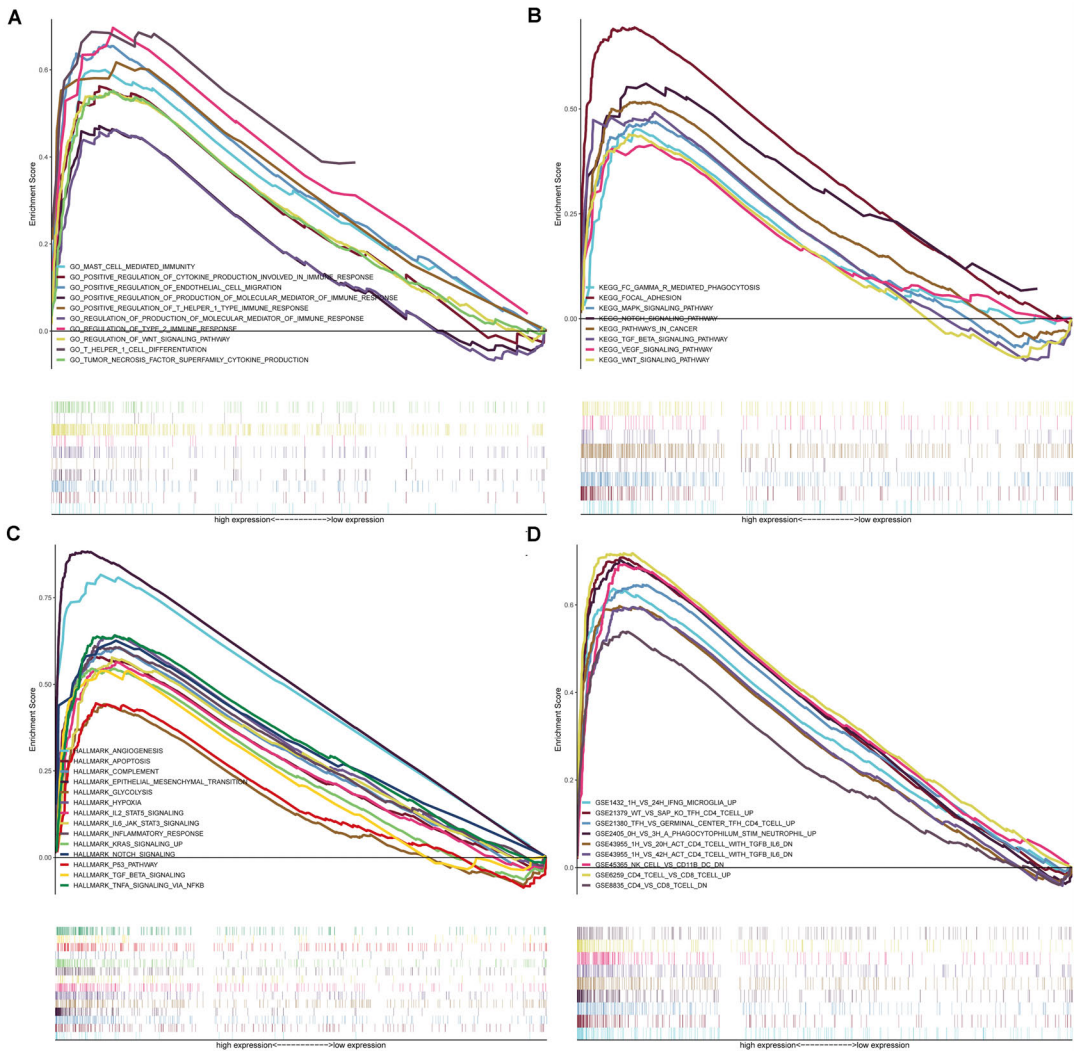

FIGURE 10 | GSEA for samples with high-risk and low-risk based on the prognostic index of six-gene signature. **(A)** The enriched gene sets in GO collection by the high-risk sample. Each line representing one particular gene set with unique color, and up-regulated genes located in the left approaching the origin of the coordinates, by contrast the down-regulated lay on the right of *x*-axis. Only gene sets with NOM $p < 0.05$ and FDR $q < 0.05$ were considered significant. And only several leading gene sets were displayed in the plot. **(B)** The enriched gene sets in KEGG by samples with high-risk sample. And only several leading gene sets were displayed in the plot. **(C)** Enriched gene sets in HALLMARK collection by samples of high-risk sample. Only several leading gene sets are shown in plot. **(D)** Enriched gene sets in C7 collection, the immunologic gene sets, by samples of high-risk sample. Only several leading gene sets are shown in plot.

by the immune system with major histocompatibility complex (MHC) (Wong et al., 2018). As a measure of somatic mutations in cancer cells, TMB is useful in estimating tumor neoantigenic load (Rhee, 2016), and thus critical for the identification of patients likely to respond to immune checkpoint blockade (Wong et al., 2018). In this study, the level of TMB is significantly higher in High-Immunity Subtype than in Medium-Immunity Subtype and Low-Immunity Subtype, confirming the relationship between TMB and immunity.

To investigate the molecular mechanisms and clinical value of potential targets, we established an immune-based prognostic index to develop a convenient and reliable protocol for monitoring the immune status and clinical outcomes in colon and rectal cancer patients. The index is based on the fractions of six genes identified among the differentially expressed genes from the stable High-Immunity and Low-Immunity subtypes, all of which were up-regulated in the High-Immunity Subtype. However, the potential molecular mechanisms of these genes remain poorly understood. Few studies on the function and mechanism of *CERCAM* in colon and rectal cancer have been published. *CD37* belongs to the tetraspanin SUPERFAMILY that is widely expressed and forms complexes with other tetraspanins and MHC class II on mature B cells (Xu-Monette et al., 2016). Some studies indicated that *CD37* may be associated with various different cellular processes, including migration, adhesion, proliferation of lymphocytes and survival, and it is significant for interactions between T- and B-cells as well as for immunoglobulin G/immunoglobulin A production

(van Spriel et al., 2004, 2009; van Spriel, 2011; Beckwith et al., 2015). Cells with high expression of *CALB2* (Calbindin-2) were derived from primary colon tumors, and it could be a diagnostic marker for malignant mesotheliomas (Chu et al., 2005; Blum et al., 2018). Furthermore, *CALB2* could be a modifier of 5-fluorouracil sensitivity, promoting cell death in colorectal cancer cells through activation of the intrinsic apoptotic pathway following treatment with this chemotherapy agent (Stevenson et al., 2011). The Mesenchyma *MEOX2* (Homeobox 2) was previously shown to be related to malignant progression and clinical prognosis in lung cancer, hepatocellular carcinoma, laryngeal carcinoma and gliomas (Tachon and Masliantsev, 2019). Furthermore, *MEOX2* was also found to regulate the migration and proliferation of endothelial cells with NF-kB downregulation (Patel et al., 2005). Additionally, *MEOX2* promoter sequences have been treated as part of a test for cancer-specific DNA methylation cluster markers in colorectal cancer, and it may regulate the resistance to chemotherapeutics such as cisplatinum (De Carvalho et al., 2012; Ávila-Moreno et al., 2014). *RASGRP2* (RAS guanyl releasing protein 2), is a guanine-nucleotide-exchange-factor that can activate small GTPases, such as Ras and Rap (Irizarry et al., 2003). Additionally, *RASGRP2* was identified as a high-avidity target antigen for $CD4^+$ T cells, and its expression is thought to be upregulated by HLA-DR to activate and propagate autoreactive $CD4^+$ T cells (Jelcic et al., 2018). Moreover, *RASGRP2* is related to immune-mediated thrombosis and thrombocytopenia, and mediates platelet and T-cell adhesion with integrin-independent neutrophil chemotaxis via integrin-mediated activation of Rap1 (Cifuni et al., 2008; Carbo et al., 2010). Some studies have also shown that *RASGRP2* could promote the migratory, invasive and proliferative capacity *in vitro*, as well as confer chemoresistance in prostate cancer, metastatic melanoma, and colon cancer (Yang et al., 2008; Wang et al., 2017; Wang L.X. et al., 2018). The upregulation of *PCOLCE2* expression leads to enhanced extracellular matrix organization, which has in turn promotes cancer cells adhesion, and may be employed to predict tumors with a propensity for developing metastasis in lung cancer, gynecological cancers or rectal cancer (Thutkawkorapin et al., 2016; Adhikary et al., 2017; Lim et al., 2017; Zhang and Wang, 2019). Furthermore, we established the nomogram to predict the survival more accurately for colon cancer patients with visualization results, which can further improve the compliance and therapeutic effect of patients. For example, a 70-year-old (43 points in the model) colon cancer patients with T3 stage (65 points), M0 stage (0 points), High-Immunity Subtype (0 points) with high-risk (42 points) has a total of 150 points, resulting in the estimated 3-, 5-year OS of about 65.0 and 55%. The 3- and 5-year OS of patients with High-Immunity Subtype were both remarkably improved combined with low six-gene signature risk.

There are a number of limitations to this study. For example, we screened the genes by identifying overlapping DEGs from different stable immune subtypes. Although these genes were able to identify the stable immune subtypes of colon and rectal cancer and their prognostic powers was validated in this study, the results are based on RNA-sequencing data, lacking functional validation of the target genes. This should be addressed in future studies. Furthermore, only limited data were used for performance evaluation and it is necessary to collect more datasets for a more comprehensive evaluation. Because of the lack of *in vitro* or *in vivo* experiments, the reliability of the analysis of molecular mechanism could be limited. And some prospective study could be carried out to validate the findings of this retrospective study. Functional experiments for the validation of the identified DEGs and corresponding downstream signaling pathways are needed to therapeutic targets and reveal novel diagnostic for colon and rectal cancer. Although the multivariate Cox proportional hazards regression analysis was employed widely to identify key factors involved in the establishment of a prognostic model, several machine learning algorithms might achieve better prediction results, such as Decision Tree, Naïve Bayes, and Random Forest. We will test these algorithms in the future. In the future, many questions remain to be solved on cancer immune therapy, including the correlation between immunogenomics, proteomics, and metabolomics, which can be used to understand the immunological changes in rectal and colon cancer. We hope that our systematic analysis will be of great help in promoting risk stratification, therapeutic decision-making in patients with colon and rectal cancer.

CONCLUSION

This study demonstrates the utility of colon and rectal cancer immune subtypes based on immune signatures in the diagnosis, treatment evaluation, and prognosis. The proposed DEGs models could assist in formulating more efficient therapeutic strategies for improving the personalized management of colon and rectal cancer patients.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ZJ, RW, and HL made substantial contributions to the conception and designed the study. ZJ supervised the acquisition of the data. CL, XG, and RW participated in the data analysis and statistical analysis. RW, HL, CL, XG, ZZ, CM, and XW contributed to interpretation of the results. ZZ, CM, and HL performed the revision of manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Program Project for Precision Medicine in National Research and Development Plan of China (2018YFC1315000), Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2018PT32012), and CAMS Innovation Fund for Medical Sciences (CIFMS) (2019-I2M-2-002 and 2017-I2M-1-006).

ACKNOWLEDGMENTS

The authors acknowledge the great efforts of the TCGA project and GEO in the creation of the database.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00740/full#supplementary-material

REFERENCES

Adhikary, T., Wortmann, A., Finkernagel, F., Lieber, S., Nist, A., Stiewe, T., et al. (2017). Interferon signaling in ascites-associated macrophages is linked to a favorable clinical outcome in a subgroup of ovarian carcinoma patients. *BMC Genomics* 18:243. doi: 10.1186/s12864-017-3630-9

Ali, H. R., Chlon, L., Pharoah, P. D., and Markowetz, F. (2016). Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. *PLoS Med.* 13:e1002194. doi: 10.1371/journal.pmed.1002194

Ávila-Moreno, F., Armas-López, L., Álvarez-Moran, A. M., López-Bujanda, Z., Ortiz-Quintero, B., Hidalgo-Miranda, A., et al. (2014). Overexpression of MEOX2 and TWIST1 is associated with H3K27me3 levels and determines lung cancer chemoresistance and prognosis. *PLoS One* 9:e114104. doi: 10.1371/journal.pone.0114104

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 108–112. doi: 10.1038/nature08460

Basile, D., Garattini, S. K., Bonotto, M., Ongaro, E., Casagrande, M., Cattaneo, M., et al. (2017). Immunotherapy for colorectal cancer: where are we heading? *Exp. Opin. Biol. Ther.* 17, 709–721. doi: 10.1080/14712598.2017.1315405

Becht, E., de Reyniès, A., Giraldo, N. A., Pilati, C., Buttard, B., Lacroix, L., et al. (2016). Immune and stromal classification of colorectal cancer is associated with molecular subtypes and relevant for precision immunotherapy. *Clin. Cancer Res.* 22, 4057–4066. doi: 10.1158/1078-0432.ccr-15-2879

Beckwith, K. A., Byrd, J. C., and Muthusamy, N. (2015). Tetraspanins as therapeutic targets in hematological malignancy: a concise review. *Front. Physiol.* 6:91. doi: 10.3389/fphys.2015.00091

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39, 782–795. doi: 10.1016/j.immuni.2013.10.003

Blum, W., Pecze, L., Rodriguez, J. W., Steinauer, M., and Schwaller, B. (2018). Regulation of calretinin in malignant mesothelioma is mediated by septin 7 binding to the CALB2 promoter. *BMC Cancer* 18:475. doi: 10.1186/s12885-018-4385-7

Bopanna, S., Ananthakrishnan, A. N., Kedia, S., Yajnik, V., and Ahuja, V. (2017). Risk of colorectal cancer in Asian patients with ulcerative colitis: a systematic review and meta-analysis. *Lancet Gastroenterol. Hepatol.* 2, 269–276. doi: 10.1016/s2468-1253(17)30004-3

Brahmer, J. R., Tykodi, S. S., Chow, L. Q., Hwu, W. J., Topalian, S. L., Hwu, P., et al. (2012). Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. *N. Engl. J. Med.* 366, 2455–2465. doi: 10.1056/NEJMoa1200694

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Budczies, J., Klauschen, F., Sinn, B. V., Győrffy, B., Schmitt, W. D., Darb-Esfahani, S., et al. (2012). Cutoff finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLoS One* 7:e51862. doi: 10.1371/journal.pone.0051862

Carbo, C., Duerschmied, D., Goerge, T., Hattori, H., Sakai, J., Cifuni, S. M., et al. (2010). Integrin-independent role of CalDAG-GEFI in neutrophil chemotaxis. *J. Leukoc. Biol.* 88, 313–319. doi: 10.1189/jlb.0110049

Chouhan, H., and Sammour, T. (2018). The interaction between BRAF mutation and microsatellite instability (MSI) status in determining survival outcomes after adjuvant 5FU based chemotherapy in stage III colon cancer. *J. Surg. Oncol.* 118, 1311–1317. doi: 10.1002/jso.25275

Chu, A. Y., Litzky, L. A., Pasha, T. L., Acs, G., and Zhang, P. J. (2005). Utility of D2-40, a novel mesothelial marker, in the diagnosis of malignant mesothelioma. *Mod. Pathol.* 18, 105–110. doi: 10.1038/modpathol.3800259

Cifuni, S. M., Wagner, D. D., and Bergmeier, W. (2008). CalDAG-GEFI and protein kinase C represent alternative pathways leading to activation of integrin alphaIIbbeta3 in platelets. *Blood* 112, 1696–1703. doi: 10.1182/blood-2008-02-139733

De Carvalho, D. D., Sharma, S., You, J. S., Su, S. F., Taberlay, P. C., Kelly, T. K., et al. (2012). DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell* 21, 655–667. doi: 10.1016/j.ccr.2012.03.045

Duluc, D., Corvaisier, M., Blanchard, S., Catala, L., Descamps, P., Gamelin, E., et al. (2009). Interferon-gamma reverses the immunosuppressive and protumoral properties and prevents the generation of human tumor-associated macrophages. *Int. J. Cancer* 125, 367–373. doi: 10.1002/ijc.24401

Dumauthioz, N., Labiano, S., and Romero, P. (2018). Tumor resident memory T Cells: new players in immune surveillance and therapy. *Front. Immunol.* 9:2076. doi: 10.3389/fimmu.2018.02076

Fletcher, R., Wang, Y. J., Schoen, R. E., Finn, O. J., Yu, J., and Zhang, L. (2018). Colorectal cancer prevention: immune modulation taking the stage. *Biochim. Biophys. Acta Rev. Cancer* 1869, 138–148. doi: 10.1016/j.bbcan.2017.12.002

Fridman, W. H., Pagès, F., Sautès-Fridman, C., and Galon, J. (2012). The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* 12, 298–306. doi: 10.1038/nrc3245

Greten, F. R., Eckmann, L., Greten, T. F., Park, J. M., Li, Z. W., Egan, L. J., et al. (2004). IKKbeta links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell* 118, 285–296. doi: 10.1016/j.cell.2004.07.013

Gutting, T., Burgermeister, E., Härtel, N., and Ebert, M. P. (2019). Checkpoints and beyond – immunotherapy in colorectal cancer. *Semin. Cancer Biol.* 55, 78–89. doi: 10.1016/j.semcancer.2018.04.003

Han, J., Chen, M., Wang, Y., Gong, B., Zhuang, T., Liang, L., et al. (2018). Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Sci. Rep.* 8:9912. doi: 10.1038/s41598-018-28299-9

He, Y., Jiang, Z., Chen, C., and Wang, X. (2018). Classification of triple-negative breast cancers based on immunogenomic profiling. *J. Exp. Clin. Cancer Res.* 37:327. doi: 10.1186/s13046-018-1002-1

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4, 249–264. doi: 10.1093/biostatistics/4.2.249

Jelcic, I., Al Nimer, F., Wang, J., Lentsch, V., Planas, R., Jelcic, I., et al. (2018). Memory B Cells activate brain-homing, autoreactive CD4(+) T Cells in multiple sclerosis. *Cell* 175, 85–100.e23. doi: 10.1016/j.cell.2018.08.011

Kather, J. N., and Halama, N. (2019). Harnessing the innate immune system and local immunological microenvironment to treat colorectal cancer. *Br. J. Cancer* 120, 871–882. doi: 10.1038/s41416-019-0441-6

Koliaraki, V., Pasparakis, M., and Kollias, G. (2015). IKKβ in intestinal mesenchymal cells promotes initiation of colitis-associated cancer. *J. Exp. Med.* 212, 2235–2251. doi: 10.1084/jem.20150542

Lewis, C. E., and Pollard, J. W. (2006). Distinct role of macrophages in different tumor microenvironments. *Cancer Res.* 66, 605–612. doi: 10.1158/0008-5472.can-05-4005

Li, H. B., Tong, J., Zhu, S., Batista, P. J., Duffy, E. E., Zhao, J., et al. (2017). m(6)A mRNA methylation controls T cell homeostasis by targeting the IL-7/STAT5/SOCS pathways. *Nature* 548, 338–342. doi: 10.1038/nature23450

Lim, S. B., Tan, S. J., Lim, W. T., and Lim, C. T. (2017). An extracellular matrix-related prognostic and predictive indicator for early-stage non-small cell lung cancer. *Nat. Commun.* 8:1734. doi: 10.1038/s41467-017-01430-6

Lin, P., Guo, Y. N., Shi, L., Li, X. J., Yang, H., He, Y., et al. (2019). Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. *Aging* 11, 480–500. doi: 10.18632/aging.101754

Locy, H., de Mey, S., de Mey, W., De Ridder, M., Thielemans, K., and Maenhout, S. K. (2018). Immunomodulation of the tumor microenvironment: turn foe into friend. *Front. Immunol.* 9:2909. doi: 10.3389/fimmu.2018.02909

Lopez, A., Pouillon, L., Beaugerie, L., Danese, S., and Peyrin-Biroulet, L. (2018). Colorectal cancer prevention in patients with ulcerative colitis. *Best Pract. Res. Clin. Gastroenterol.* 3, 103–109. doi: 10.1016/j.bpg.2018.05.010

Lucas, C. L., Chandra, A., Nejentsev, S., Condliffe, A. M., and Okkenhaug, K. (2016). PI3Kδ and primary immunodeficiencies. *Nat. Rev. Immunol.* 16, 702–714. doi: 10.1038/nri.2016.93

Lv, Y., Zhao, Y., Wang, X., Chen, N., Mao, F., Teng, Y., et al. (2019). Increased intratumoral mast cells foster immune suppression and gastric cancer progression through TNF-α-PD-L1 pathway. *J. Immunother. Cancer.* 7:54. doi: 10.1186/s40425-019-0530-3

Marech, I., Ammendola, M., Gadaleta, C., Zizzo, N., Oakley, C., Gadaleta, C. D., et al. (2014). Possible biological and translational significance of mast cells density in colorectal cancer. *World J. Gastroenterol.* 20, 8910–8920. doi: 10.3748/wjg.v20.i27.8910

Newman, A. M., Liu, C. L., and Green, M. R. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

Pagès, F., Mlecnik, B., Marliot, F., Bindea, G., Ou, F. S., Bifulco, C., et al. (2018). International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* 391, 2128–2139. doi: 10.1016/s0140-6736(18)30789-x

Palucka, A. K., and Coussens, L. M. (2016). The basis of oncoimmunology. *Cell* 164, 1233–1247. doi: 10.1016/j.cell.2016.01.049

Patel, S., Leal, A. D., and Gorski, D. H. (2005). The homeobox gene Gax inhibits angiogenesis through inhibition of nuclear factor-kappaB-dependent endothelial cell gene expression. *Cancer Res.* 65, 1414–1424. doi: 10.1158/0008-5472.can-04-3431

Pollard, J. W. (2004). Tumour-educated macrophages promote tumour progression and metastasis. *Nat. Rev. Cancer* 4, 71–78. doi: 10.1038/nrc1256

Qin, L., Wang, W. Z., Liu, H. R., Xu, W. B., Qin, M. W., Zhang, Z. H., et al. (2013). CD4+ and CD8+ T lymphocytes in lung tissue of NSIP: correlation with T lymphocytes in BALF. *Respir. Med.* 107, 120–127. doi: 10.1016/j.rmed.2012.09.021

Qin, S., Kim, J., Arafat, D., and Gibson, G. (2012). Effect of normalization on statistical and biological interpretation of gene expression profiles. *Front. Genet.* 3:160. doi: 10.3389/fgene.2012.00160

Rhee, I. (2016). Diverse macrophages polarization in tumor microenvironment. *Arch. Pharm. Res.* 39, 1588–1596. doi: 10.1007/s12272-016-0820-y

Sanoff, H. K., Bleiberg, H., and Goldberg, R. M. (2007). Managing older patients with colorectal cancer. *J. Clin. Oncol.* 25, 1891–1897. doi: 10.1200/jco.2006.10.1220

Sasidharan Nair, V., Toor, S. M., Taha, R. Z., Shaath, H., and Elkord, E. (2018). DNA methylation and repressive histones in the promoters of PD-1, CTLA-4, TIM-3, LAG-3, TIGIT, PD-L1, and galectin-9 genes in human colorectal cancer. *Clin. Epigenet.* 10:104. doi: 10.1186/s13148-018-0539-3

Sharma, P., and Allison, J. P. (2015). The future of immune checkpoint therapy. *Science* 348, 56–61. doi: 10.1126/science.aaa8172

Sica, A., Schioppa, T., Mantovani, A., and Allavena, P. (2006). Tumour-associated macrophages are a distinct M2 polarised population promoting tumour progression: potential targets of anti-cancer therapy. *Eur. J. Cancer* 42, 717–727. doi: 10.1016/j.ejca.2006.01.003

Stevenson, L., Allen, W. L., Proutski, I., Stewart, G., Johnston, L., McCloskey, K., et al. (2011). Calbindin 2 (CALB2) regulates 5-fluorouracil sensitivity in colorectal cancer by modulating the intrinsic apoptotic pathway. *PLoS One* 6:e20276. doi: 10.1371/journal.pone.0020276

Su, P., Liu, X., Pang, Y., Liu, C., Li, R., Zhang, Q., et al. (2017). The archaic roles of the lamprey NF-κB (lj-NF-κB) in innate immune responses. *Mol. Immunol.* 92, 21–27. doi: 10.1016/j.molimm.2017.10.002

Sun, S. C. (2017). The non-canonical NF-κB pathway in immunity and inflammation. *Nat. Rev. Immunol.* 17, 545–558. doi: 10.1038/nri.2017.52

Tachon, G., and Masliantsev, K. (2019). Prognostic significance of MEOX2 in gliomas. *Mod. Pathol.* 32, 774–786. doi: 10.1038/s41379-018-0192-6

Thutkawkorapin, J., Picelli, S., Kontham, V., Liu, T., Nilsson, D., and Lindblom, A. (2016). Exome sequencing in one family with gastric- and rectal cancer. *BMC GEneti.* 17:41. doi: 10.1186/s12863-016-0351-z

van Spriel, A. B. (2011). Tetraspanins in the humoral immune response. *Biochem. Soc. Trans.* 39, 512–517. doi: 10.1042/bst0390512

van Spriel, A. B., Puls, K. L., Sofi, M., Pouniotis, D., Hochrein, H., Orinska, Z., et al. (2004). A regulatory role for CD37 in T cell proliferation. *J. Immunol.* 172, 2953–2961. doi: 10.4049/jimmunol.172.5.2953

van Spriel, A. B., Sofi, M., Gartlan, K. H., van der Schaaf, A., Verschueren, I., Torensma, R., et al. (2009). The tetraspanin protein CD37 regulates IgA responses and anti-fungal immunity. *PLoS Pathog.* 5:e1000338. doi: 10.1371/journal.ppat.1000338

Vincent, K. M., Findlay, S. D., and Postovit, L. M. (2015). Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. *Breast Cancer Res.* 17:114. doi: 10.1186/s13058-015-0613-0

Wang, B. D., Ceniccola, K., Hwang, S., Andrawis, R., Horvath, A., Freedman, J. A., et al. (2017). Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer. *Nat. Commun.* 8:15921. doi: 10.1038/ncomms15921

Wang, F., and Xia, Q. (2018). Back to homeostasis: negative regulation of NF-κB immune signaling in insects. *Dev. Compar. Immunol.* 87, 216–223. doi: 10.1016/j.dci.2018.06.007

Wang, L. X., Li, Y., and Chen, G. Z. (2018). Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. *PLoS One* 13:e0190447. doi: 10.1371/journal.pone.0190447

Wang, Y., Lin, H. C., Huang, M. Y., Shao, Q., Wang, Z. Q., Wang, F. H., et al. (2018). The immunoscore system predicts prognosis after liver metastasectomy in colorectal cancer liver metastases. *Cancer Immunol. Immunother.* 67, 435–444. doi: 10.1007/s00262-017-2094-8

Wei, H., Li, J., Xie, M., Lei, R., and Hu, B. (2018). Comprehensive analysis of metastasis-related genes reveals a gene signature predicting the survival of colon cancer patients. *PeerJ* 6:e5433. doi: 10.7717/peerj.5433

Wilkinson, N. W., Yothers, G., Lopa, S., Costantino, J. P., Petrelli, N. J., and Wolmark, N. (2010). Long-term survival results of surgery alone versus surgery plus 5-fluorouracil and leucovorin for stage II and stage III colon cancer: pooled analysis of NSABP C-01 through C-05. A baseline from which to compare modern adjuvant trials. *Ann. Surg. Oncol.* 17, 959–966. doi: 10.1245/s10434-009-0881-y

Wong, Y. N. S., Joshi, K., Khetrapal, P., Ismail, M., Reading, J. L., Sunderland, M. W., et al. (2018). Urine-derived lymphocytes as a non-invasive measure of the bladder tumor immune microenvironment. *J. Exp. Med.* 215, 2748–2759. doi: 10.1084/jem.20181003

Xu-Monette, Z. Y., Li, L., Byrd, J. C., Jabbar, K. J., Manyam, G. C., Maria de Winde, C., et al. (2016). Assessment of CD37 B-cell antigen and cell of origin significantly improves risk prediction in diffuse large B-cell lymphoma. *Blood* 128, 3083–3100. doi: 10.1182/blood-2016-05-715094

Yang, H., Zhang, X., Cai, X. Y., Wen, D. Y., Ye, Z. H., Liang, L., et al. (2017). From big data to diagnosis and prognosis: gene expression signatures in liver hepatocellular carcinoma. *PeerJ.* 5:e3089. doi: 10.7717/peerj.3089

Yang, L., Zhang, L., Wu, Q., and Boyd, D. D. (2008). Unbiased screening for transcriptional targets of ZKSCAN3 identifies integrin beta 4 and vascular endothelial growth factor as downstream targets. *J. Biol. Chem.* 283, 35295–35304. doi: 10.1074/jbc.M806965200

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

Yothers, G., O'Connell, M. J., Lee, M., Lopatin, M., Clark-Langone, K. M., Millward, C., et al. (2013). Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *J. Clin. Oncol.* 31, 4512–4519. doi: 10.1200/jco.2012.47.3116

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* 16, 284–287. doi: 10.1089/omi.2011.0118

Yu, H., Pardoll, D., and Jove, R. (2009). STATs in cancer inflammation and immunity: a leading role for STAT3. *Nat. Rev. Cancer* 9, 798–809. doi: 10.1038/nrc2734

Zeng, D., Zhou, R., Yu, Y., Luo, Y., Zhang, J., Sun, H., et al. (2018). Gene expression profiles for a prognostic immunoscore in gastric cancer. *Br. J. Surg.* 105, 1338–1348. doi: 10.1002/bjs.10871

Zhang, H., Watanabe, R., Berry, G. J., Tian, L., Goronzy, J. J., and Weyand, C. M. (2018). Inhibition of JAK-STAT signaling suppresses pathogenic immune responses in medium and large vessel vasculitis. *Circulation* 137, 1934–1948. doi: 10.1161/circulationaha.117.030423

Zhang, X., and Wang, Y. (2019). Identification of hub genes and key pathways associated with the progression of gynecological cancer. *Oncol. Lett.* 18, 6516–6524. doi: 10.3892/ol.2019.11004

Zheng, W., O'Hear, C. E., Alli, R., Basham, J. H., Abdelsamed, H. A., Palmer, L. E., et al. (2018). PI3K orchestration of the in vivo persistence of chimeric antigen receptor-modified T cells. *Leukemia* 32, 1157–1167. doi: 10.1038/s41375-017-0008-6

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 September 2020
doi: 10.3389/fgene.2020.560812

Pharmacological Effects of Novel Peptide Drugs on Allergic Rhinitis at the Small Ribonucleic Acids Level

Li-Feng An[1†], Zhan-Dong Li[2,3†], Lin Li[1], Hao Li[2] and Jian Yu[2,3]*

[1] Department of Otorhinolaryngology Head and Neck Surgery, China-Japan Union Hospital of Jilin University, Changchun, China, [2] College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [3] Measurement Biotechnique Research Center, Jilin Engineering Normal University, Changchun, China

Using an allergic rhinitis (AR) model, we evaluated the pharmacological effects of novel peptide drugs (P-ONE and P-TWO) at the small RNA (sRNA) level. Using high-throughput sequencing, we assessed the sRNA expression profile of the negative control, AR antagonist (positive control), P-ONE, and P-TWO groups. By functional clustering and Gene Ontology and KEGG pathway analyses, we found that sRNA target genes have a specific enrichment pattern and may contribute to the effects of the novel peptides. Small RNA sequencing confirmed the biological foundations of novel and traditional AR treatments and suggested unique pharmacological effects. Our findings will facilitate evaluation of the pathogenesis of AR and of the pharmacological mechanisms of novel peptide drugs.

Keywords: allergic rhinitis, peptide drugs, small RNAs, high-throughput sequencing, gene ontology, KEGG pathway

OPEN ACCESS

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

Reviewed by:
Yu-Hang Zhang,
Channing Division of Network Medicine, Brigham and Women's Hospital, United States
Hu Zhou,
Chinese Academy of Sciences, China

****Correspondence:***
Lin Li
lilin01@jlu.edu.cn

[†] These authors have contributed equally to this work

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

Received: *10 May 2020*
Accepted: *24 August 2020*
Published: *11 September 2020*

Citation:
An L-F, Li Z-D, Li L, Li H and Yu J (2020) Pharmacological Effects of Novel Peptide Drugs on Allergic Rhinitis at the Small Ribonucleic Acids Level. Front. Genet. 11:560812. doi: 10.3389/fgene.2020.560812

INTRODUCTION

Allergic rhinitis (AR) (Maoz-Segal et al., 2019) is defined as inflammation of the nasal mucosa induced by an allergic reaction; it is also known as anaphylaxis (allergy) (Turner et al., 2019). Based on the clinical symptoms, AR can be classified into four groups based on its persistence and severity (Settipane and Charnock, 2016; Jung et al., 2020). For instance, AR of > 4-week duration is classified as persistent and AR with only mild symptoms as mild. These classifications can be combined; for instance, mild persistent AR (Settipane and Charnock, 2016). According to an independent survey, one in five people in Australia (Smith et al., 2017; Price et al., 2018) and one in three in the United States (Han et al., 2016) suffer or have suffered from AR, typically accompanied by asthma and allergic complications (Hill et al., 2016). Similar frequencies have been reported in other countries (Cardell et al., 2016; Bousquet et al., 2018). Therefore, AR is an important threat to human health globally.

Initially, the pathogenesis of AR was evaluated based on its pathological characteristics, such as inflammation and bacterial infection (Ledford and Lockey, 2016). However, these symptoms are similar to those of other diseases such as infectious rhinitis (the common cold) (Meltzer et al., 2000), indicating that phenotypic features cannot explain the differential susceptibility among populations. Next-generation sequencing enables genomic and transcriptomic analysis of disease. Polymorphisms of genes such as FcγRIIIa (Zeyrek et al., 2008) and those encoding

histamine-metabolizing enzymes (García-Martín et al., 2007) are reported to be functionally related to the onset of such diseases. Moreover, the expression of some microRNAs (miRNAs; e.g., miR-370, miR-539, and miR-299) is altered during AR pathogenesis (Specjalski et al., 2016).

Histamine H4 receptor, a member of the G protein-coupled receptor superfamily, is a core regulatory factor of AR (Takahashi et al., 2009; Broide, 2010; Shiraishi et al., 2013). During the pathogenesis of AR, H4 receptor is upregulated, triggering immune over-activation (Lundberg et al., 2011; Walter et al., 2011) and remodeling of the inflammatory microenvironment by modulation of IL-6 and INF-γ expression (Peng et al., 2019). Among the drugs targeting the core pathogenic processes of AR, many target the H4 receptor. Indeed, two vaccines developed based on the immunological epitopes of the H4 receptor were effective in animal models (Wang et al., 2018). Such vaccines trigger an immune response against abnormally expressed H4 receptor. Th2 cells and IgE have been demonstrated to contribute to AR pathogenesis and the efficacy of H4 receptor-based therapeutics (Wang et al., 2018; Peng et al., 2019); however, the underlying biological mechanisms are unclear.

At the methodological level, in this study, we focused on two major biological/bioinformatics techniques: (1) establishment of guinea pig allergic rhinitis model; (2) analyses on small RNA sequencing data. For the establishment of guinea pig allergic rhinitis, researchers from multiple countries have developed and modified various methods to establish stable, reproducible and comparable models to mimic the pathogenesis of allergic rhinitis in human beings. In 2006, researchers from University of British Columbia (Canada) summarized the general workflow for the establishment of stable guinea pig allergic rhinitis model using ovalbumin (Al Suleimani et al., 2007), which is one of the most common allergens for guinea pigs' respiratory tracts. In the next decades, the detailed techniques have been gradually modified but the major establishment procedure remains stable, implying the stability and efficiency of such ovalbumin based methods. Apart from that, another major methodological challenges for our study turns out to be the comparable small RNA sequencing analyses. With the development of computational methods, a general workflow for small RNA sequencing analyses has already been established including small RNA clustering, novel small RNA discovery, miRNA target prediction, differential expression of small RNA, evolutionary analysis, and functional analysis (Baran-Gale et al., 2015; Fuchs et al., 2015; Buschmann et al., 2016). In this study, we applied the latest workflow/software for small RNA identification and annotation [miRDeep2 (Friedländer et al., 2008) and RIPmiR (Breakfield et al., 2012)], revealing a robust small RNA profiling results for further analyses and summarization.

MicroRNAs are important in the pathogenesis of AR. In this study, miRNA profiling and a guinea pig model of AR enabled identification of the therapeutic mechanisms of two epitopes of the H4 receptor. All in all, based on the microRNA profiling techniques and blood samples from guinea pig model of allergic rhinitis (AR), we focused the underlying therapeutic mechanisms of two reported epitopes against H4 receptor for allergic rhinitis treatment at the microRNA level, trying to reveal their potential pharmacological mechanisms by targeting H4 receptors.

MATERIALS AND METHODS

Reagents and Instruments

The following reagents were used: TRIzol (Invitrogen, 15596018), DEPC water (Ambion, AM9915), chloroform, isopropanol, and isoamyl alcohol (Xilong Chemistry). The following instruments were used: cryogenic centrifuge (Eppendorf), vortex oscillator (Qilinbeier), and TissueLyser II (Qiagen).

Animal Models

We used 38-week-old male guinea pigs (Changchun Biological Products Research Institute Co., Ltd.; SCXK (Ji) 2016-0008) to establish a model of AR.

Model Establishment

Following widely reported rhinitis guinea pig model establishment protocol (Narita et al., 1998), we established a guinea pig model of AR using ovalbumin (OVA). OVA causes less irritation and fewer side effects than toluene diisocyanate but is prone to degeneration or coagulation and so must be made fresh immediately before use.

Small Peptide Screening

We first purified anti-HRH4 monoclonal IgG to a high purity (95%) for phage peptide library screening. Using HR4 antibody as the antigen, we screened out two peptides with high affinity for the monoclonal HRH4 IgG; these were named P-ONE (FNKWMDCLSVTH) and P-TWO (TFKFTLSYRQVH) and have been patented (Patent 1: "Vaccine based on mimicking human histamine receptor 4 (HR4) epitope and construction method thereof", Application No.: 201510382851.1, Publication No.: 105017385B and Patent 2: "Using a phage antibody library to screen human histamine receptor 4 (HR4) epitope mimetic peptides and a vaccine construction method", Application No.: 201510382781.X, Publication No.: 105037499B).

Preparation of Vaccines

The peptide and CTB were dissolved in physiological saline, and the same volume of liposome Lipofect was added such that each 200 μL contained 100 μg of peptide and 5 μg of CTB. The mixture was stored overnight at 4°C and on the following day was brought to room temperature.

(1) Add 1 mL of saline to the antagonist to make a 25 mg/mL solution.
(2) Add 50 μL of normal saline to CTB to make a 10 mg/mL solution.
(3) The antagonist requires a total of 200 μL of nasal drops and is formulated as follows: Antagonist (peptide, 100 μg) 4 μL, CTB (5 μg) 0.5 μL, normal saline 95.5 μL, and liposomes 100 μL.

(4) P-ONE 20.7 mg, P-TWO 20.5 mg, and control vaccine 20.5 mg. Normal saline (NS) is added to P-ONE, P-TWO, and control vaccine as shown in **Table 1**.

Evaluation of Animal Model of AR

There is no uniform standard for the evaluation of AR models. Instead, such models are evaluated based on their ability to repeatedly trigger an allergic reaction.

Symptoms of AR, combined with changes in animal behavior and characteristic pathomorphological changes, were assessed to evaluate the model. After stimulation, animals with AR exhibit symptoms such as sneezing, scratching the nose, scratching the face, and a running nose. Most prior studies adopted the symptom score of Zhao (1993). We tested the model by evaluating sniffing, sneezing, and nasal discharge. During the evaluation, the superimposed quantitative score was applied to indicate the success of modeling. Symptom score is tested and calculated at the last time. After stimulation, each animal was observed for 30 min. The scoring criteria were as follows:

(1) Nasal itching: 1 point for one or two instances of light nose blowing, 2 points for moderate scratching of the nose/face, and 3 points for violent scratching of the nose/face.
(2) Sneeze: 1 point for 1–3, 2 points for 4–10, and 3 points for $\geq$ 11.
(3) Clearing the nose: 1 point for nostril flow, 2 points for the front nostril, and 3 points for the runny surface.

The three symptom scores were summed and a total score of $\geq$ 5 was considered a success. This experiment is based on observation records and combined with related behavioral indicators, verifying the success of the model.

RNA Sampling

After the last dose on day 85, behavioral indicators were evaluated. We extracted RNA from blood samples for miRNA sequencing using the RNeasy Plus Micro and Mini Kits (Qiagen).

We fragmented and digested tissue samples by two methods. The first method is a lapping machine-based method. An appropriate amount of tissue sample was placed in a numbered grinding and crushing tube and 1.5 mL of TRIzol lysate was added. The mixture was ground in a TissueLyser II grinder for 30 s, and allowed to stand for 5 min. The second method was performed using liquid nitrogen. TRIzol lysate (1.5 mL) was transferred into a 2 mL EP tube. An appropriate amount of tissue sample was ground into powder in liquid nitrogen, transferred to the lysate, and allowed to stand flat for 5 min. Next, the disrupted tissue samples were centrifuged at 4°C and 12,000 g for 5 min. The supernatant was transferred to an EP tube containing 300 μL of chloroform: isoamyl alcohol (24: 1), mixed by inverting and shaking vigorously, and centrifuged at 12,000 *g* at 4°C for 8 min. If the middle layer was thick and the water phase turbid, extraction was repeated using the same volume of chloroform: isoamyl alcohol (24: 1).

The supernatant was transferred to a centrifuge tube containing 600 μL of isopropanol. Do not suck into the middle layer (micro-tissue or micro-cell sample, add 2 μL of 5 mg/mL glycogen-assisted precipitation during precipitation), mixed by inversion, and placed at −20°C for $\geq$ 2 h. Next, the sample was centrifuged at 17,500 *g* for 25 min at 4°C, the supernatant was discarded, and the pellet was washed with 0.9 mL of 75% ethanol and invert the suspended pellet. The sample was centrifuged at 4°C for 3 min at 17,500 *g* (depending on the precipitation), washed with 75% ethanol, and centrifuged at 17,500 *g* for 3 min at 4°C. The supernatant was discarded, residual liquid was removed after brief centrifugation, and allowed to dry for 3–5 min. Finally, the pellet was dissolved in 30–200 μL DEPC or RNase-free water.

Small RNA Library Construction

We used the Agilent 2100 Bioanalyzer to evaluate sample integrity and concentration, and NanoDrop to detect inorganic ions or polycarbonate contamination.

To construct an RNA library, 0.2–1 μg of RNA was subjected to electrophoresis, 18–30 nt bands were selected (14–30 ssRNA Ladder Marker, TaKaRa) stripe and recycle. Next, we prepared a connection 3′ adaptor system at 70°C for 2 min and 25°C for 2 h and added RT primer at 65°C for 15 min followed by a ramp to 4°C at 0.3°C/s. Finally, we added the 5′ adaptor mix system at 70°C for 2 min and 25°C for 1 h. For reverse transcriptase-polymerase chain reaction (RT-PCR), we used First-Strand Master Mix and Super Script II (Invitrogen) and performed reverse transcription at 42°C for 1 h and 70°C for 15 min. Next, several rounds of PCR amplification using a PCR Primer Cocktail and Master Mix were performed at 95°C for 3 min; followed by 15–18 cycles of 98°C for 20 s, 56°C for 15 s, and 72°C for 15 s; followed by 72°C for 10 min; and a hold at 4°C. The PCR products were purified by electrophoresis and dissolved in EB.

The double-stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single-stranded circular DNA (ssCir DNA) was used as the final library. The library was validating using an Agilent Technologies 2100 Bioanalyzer. The library was amplified with phi29 to generate a DNA nanoball (DNB), which harbored > 300 copies of one molecule. The DNBs were loaded into the patterned nanoarray and single-end 50-base reads were generated by combinatorial probe-anchor synthesis (cPAS).

TABLE 1 | Vaccine preparation for different experimental and control groups.

Groups	Peptide (uL)	CTB (uL)	Physiological saline (uL)	Liposome (uL)
Negative Control	NA	NA	75	75
Positive Control	3.659 (JNJ77777120)	0.375	71	75
P-ONE	3.623 (P-ONE)	0.375	71	75
P-TWO	3.659 (P-TWO)	0.375	71	75

Small RNA Sequencing and Analysis

Data Filtering

The impurities in raw data include 5′ primer contaminants, no-insert tags, oversized insertion tags, low-quality tags, poly-A tags, small tags, and tags lacking a 3′ primer. Generally, an adaptor contaminant is caused by low sample quality or adaptor or sample concentration. The higher the adapter proportion, the greater the contamination. Low-quality tags are those with > 4 bases and a quality of < 10 or those with > 6 bases and a quality of < 13.

The above contaminant tags were removed, and the length distribution of clean tags was analyzed to evaluate sample composition. Small RNAs (sRNAs) are typically 18–30 nt in length (miRNAs, 21 or 22 nt; small interfering RNAs [siRNAs], 24 nt; and PIWI-interacting RNAs [piRNAs], 30 nt). The data were processed by removing tags of low quality, with 5′ primer contaminants, lacking a 3′ primer, without insertions, with poly-A, and of < 18 nt. The length distribution of the clean tags was summarized. After filtering, the remaining clean tags were stored in FASTQ format (Cock et al., 2009).

Reads Mapping

In general, the higher the alignment ratio, the closer the genetic relationship between the sample and the reference species. A low rate may be due to low similarity with the reference genome or to contaminants. Bowtie (Langmead et al., 2009) was used to map clean reads to the reference genome and to other sRNA databases. Please note that for Rfam we used cmsearch (Nawrocki and Eddy, 2013) with the default parameters.

Small RNA Classification

When annotating, some sRNA tags may be mapped to more than one category. To ensure that each sRNA was mapped to only one category, we used the priority miRNA > piRNA > small nucleolar RNA [snoRNA] > Rfam > other small RNA.

Small RNA Prediction

We used miRDeep2 (Friedländer et al., 2008) (for animals) and RIPmiR (Breakfield et al., 2012) (for plants) to predict novel miRNAs by exploring the characteristic hairpin structure of miRNA precursors. Piano (Wang et al., 2014), which is based on the support vector machine (SVM) (Scholkopf and Smola, 2001) algorithm and transposon interaction information, was used to predict piRNAs. The SVM classifier can be used in a wide range of species including human, mouse, rat, fruit fly, and insects. siRNA is a 22–24 nt double-strand RNA, one strand of which is 2 nt longer than the other. Due to this structural feature, we aligned tags to identify sRNAs meeting that criterion. Such tags were regarded as siRNA candidates.

Small RNA Expression

The sRNA expression level was calculated by the transcripts per million kilobases (TPM) method (John et al., 2004), which eliminates the influence of sequencing discrepancy. The data can be used for comparing gene expression between samples. To calculate the TPM the following formula was used:

$$TPM = \frac{C * 10^6}{N} \quad (1)$$

Target Prediction

To identify targets we used RNAhybrid (Krüger and Rehmsmeier, 2006), miRanda (John et al., 2004), or TargetScan (Maziere and Enright, 2007; Agarwal et al., 2015) for animal, and psRobot (Wu et al., 2012) or TargetFinder (Fahlgren and Carrington, 2010) for plants. The default parameters were as shown in **Table 2**.

Screening of DESs (Differential Expressed Sequences)

RNA sequencing could be modeled as a random sampling process, in which each read is sampled independently and uniformly from every possible nucleotide in the sample (Jiang and Wong, 2009). Under this assumption the number of reads from a gene (or transcript isoform) follows a binomial distribution (and can be approximated by the Poisson distribution).

Using the statistical model described above, DEGseq (Wang et al., 2009) proposes a novel method based on the MA-plot, a statistical analysis tool used to detect and visualize intensity-dependent ratios of microarray data (Yang et al., 2002). Let C1 and C2 denote the counts of reads mapped to a specific gene obtained from two samples, with Ci∼binomial (ni, pi), i = 1,2, where ni denotes the total number of mapped reads and pi the probability of a read coming from that gene. We define M = log2C1 - log2C2, and A = (log2C1 + log2C2 2). It can be shown that under the random sampling assumption the conditional distribution of M given that A = a (a is an observation of A) follows an approximately normal distribution. For each gene on the MA plot, we perform the hypothesis test H0: p1 = p2 versus H1: p1 ≠ p2. A P-value is assigned based on the conditional normal distribution.

The P-values calculated for each gene are adjusted to Q-values for multiple testing corrections by two strategies (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). To improve accuracy, we defined a differentially expressed gene (DEG) as a fold-change of ≥ 2 and Q-value of ≤ 0.001. RNA-seq experiments have low technical background noise and the Poisson model fits the data well. In such cases, the technical replicates can be pooled to increase the sequencing depth and detect subtle changes in gene expression. Otherwise, a method that estimates noise by comparing the replicates is recommended.

Screening of DESs (Poisson Distribution)

Based on a prior report (Audic and Claverie, 1997), BGI (Beijing Genomics Institute) developed an algorithm to identify DEGs

TABLE 2 | Default parameter for target prediction.

Methods	Parameter
miRanda	-en -20 -strict
RNAhybrid	-b 100 -c -f 2,8 -m 100000 -v 3 -u 3 -e -20 -p 1 -s 3utr_human
psRobot	-gl 17 -p 8 -gn 1
TargetFinder	-c 4

between two samples. If x is defined as the number of reads from sRNA A, x yields the Poisson distribution:

$$p(x) = \frac{e^{\lambda}\lambda^{x}}{x!} \text{ (}\lambda\text{ is the real transcripts of the gene)} \quad (2)$$

$$2\sum_{i=0}^{i=y} p(i|x) \quad (3)$$

Or

$$2 * \left(1 - \sum_{i=0}^{i=y} p(i|x)\right)(if \sum_{i=0}^{i=y} p(i|x) > 0.5 \quad (4)$$

$$p(y|x) = \left(\frac{N_2}{N_1}\right) y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x|y|1)}} \quad (5)$$

In the equation above, the *P*-value of the differential gene expression test is corrected by the Bonferroni method (Atkinson, 2002). DES analysis is then performed on the sample; however, this generates thousands of hypotheses simultaneously (only if gene x is differentially expressed between the two groups); therefore, correction for false positive (type I errors) and false negative (type II) errors is performed by the false discovery rate (FDR) method (Benjamini and Yekutieli, 2001). In the next step, it is assumed that we have selected R DEGs among which S genes show differential expression, and the V genes are false positives. The error ratio (Q) is as follows: Q = V R. The user sets a cutoff value for Q (e.g., BGI sets a default cutoff of 5%), and the FDR is preset to < 0.05. To assess the significance of differences in gene expression, an FDR of ≤ 0.001 and an absolute Log2Ratio value of ≥ 1 are set as the default thresholds. More stringent criteria, such as a smaller FDR and larger fold-change value, can also be used to identify DEGs.

Next, we performed multiple hypothesis tests for the *P*-value of the differential gene expression test and determine the *P*-value field by controlling the FDR result. The conditions were set in advance so that the FDR cannot exceed 0.05. We also calculated the gene expression level (FPKM value) to assess differences in gene expression between samples. The smaller the FDR value, the greater the difference multiple, indicating a significant difference in expression. Genes with an FDR ≤ 0.001 and multiples of more than two-fold were regarded as differentially expressed.

Hierarchical Clustering Analysis

We performed hierarchical clustering of differentially expressed miRNAs using R package "pheatmap" (Kolde and Kolde, 2015). For more than two groups, hierarchical clustering of the intersection was performed, followed by union DESs.

Gene Ontology Enrichment Analysis

Gene Ontology (GO) (The Gene Ontology Consortium, 2016) is an international standard gene functional classification system. It offers a dynamically updated and controlled vocabulary, as well as a defined concept to comprehensively describe properties of genes and their products. GO has three ontologies: molecular function, cellular component, and biological process. The basic unit of GO is the GO term; each term belongs to a type of ontology.

GO enrichment analysis finds all GO terms that are significantly enriched in a list of DES target genes and finds genes that correspond to specific biological functions. To perform this analysis, BGI first maps all genes to GO terms in the database[1], which calculates the number of genes for each term. The hypergeometric test is then performed to identify significantly enriched GO terms in the input gene list. The analysis was based on GO::TermFinder[2] and was performed using the following algorithm:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (6)$$

Here, in the equation, N is the number of all genes with GO annotations; n is the number of DES target genes in N; M is the number of all genes annotated with a specific GO term; and m is the number of DES target genes in M. The *P*-value was corrected by the Bonferroni method (Ludbrook, 1998); a corrected *P*-value ≤ 0.05 was taken as the threshold. GO terms fulfilling this condition were defined as significantly enriched.

Pathway Enrichment Analysis

KEGG (Kanehisa et al., 2007) was used to perform pathway enrichment analysis to identify significantly enriched metabolic or signal transduction pathways in DES target genes when compared with the whole genome.

The formula was as for GO analysis. N is the number of all genes with KEGG annotations; n is the number of DES target genes in N; M is the number of all genes annotated with a specific pathway; m is the number of DES target genes in M. The *P*-value was corrected by the Bonferroni method (Weisstein, 2004); a corrected *P*-value <0.05 was taken as the threshold. KEGG terms fulfilling this condition were defined as significantly enriched.

RESULTS

General Data Information

The first results of our experiments turn out to be the evaluation of the guinea pig models used for further analyses. Based on the scoring criteria, we screened out qualified animals with a total score of ≥ 5 as candidate guinea pigs for further grouping and sequencing. Then, we sequenced the microRNAs from four group: model establishment group (negative control); models adding antagonists against HR4 (positive control); group P-ONE for models adding peptide P-ONE and group P-TWO for models adding peptide P-TWO. After sequencing and data preprocessing, we firstly summarized the numbers detected small non-coding RNAs from each group shown in **Table 3**. After such data, we can identify that:

[1] http://www.geneontology.org/

[2] http://www.yeastgenome.org/help/analyze/go-term-finder

TABLE 3 | Summary of detected small non-coding RNAs for each sample.

Sample name	Known miRNA count	Novel miRNA count	Known piRNA count	Novel piRNA count	Known siRNA count	Novel siRNA count
Negative Control	266	1976	0	3467	0	0
Positive Control	264	1044	0	3618	0	0
P-ONE	317	768	0	1308	0	0
P-TWO	289	3294	0	26333	0	0

(1) Small non-coding RNAs (sncRNAs) have quite different distribution patterns in different samples, indicating their different biological status;
(2) Most of samples have similar number of known microRNAs, indicating effective microRNA may be stable and may not participate in related regulations;
(3) No siRNAs have been identified in all the samples.

To verify the distribution pattern, the first step is to verify the quality of small RNA sequencing. Therefore, we firstly showed the sequencing qualities length distribution of small RNAs among different samples (**Figures 1**, **2**). According to such two figures, it's easy for us to confirm that:

(1) Our sequencing is of high quality among all the samples: generally, sequencing with unstable quality along the genomic position or with averaged quality lower than 20 are regarded as low quality sequencing data. Our sequencing data has a stable quality greater than 35, ensuring the reliability of our further analysis;
(2) The identification of small RNAs is quite effective using our experimental and computational methods;
(3) Such small RNA sequencing results can be processed for further analysis.

Annotation of Small RNAs

After filtering, the next result obtained from analyses turned out to be the annotation name and genome locations of such identified small RNAs. Clean tags were mapped to sRNA database such as miRBase and Rfam. **Table 4** lists separate mapping rate for each sample and **Figure 3** shows the distribution of tags. The proportion of all kinds of sRNA is shown in **Figure 3**. According to **Figure 3**, different sample groups have quite different distribution of small RNA subtypes but they do share some specific prosperities:

(1) Most of the identified small RNAs can be mapped to the genome.
(2) There still remain various unknown small RNAs for further identification and function exploration with different proportions in different samples.
(3) Among those genomic derived small RNAs, most of such RNAs derived from genetic repeats and intergenic regions.

Based on the annotation of small RNAs, we summarized the number and distribution patterns of small RNAs that have already been confirmed and validated before, trying to reveal potential functional small RNA contribution on allergic rhinitis.

Prediction of Unknown Small RNAs

After the annotation of small RNAs, there still remain a lot of unknown tags and small RNAs. Therefore, it's quite necessary for us to identify new participators for the pathogenesis of allergic rhinitis at small RNA level. The identification/prediction of new small RNAs may not only help us enrich feature candidates for distribution comparison, but also predicted potential functional new small RNAs. Here, we used effective software : miRDeep2 (Friedländer et al., 2008) (for animals) and RIPmiR (Breakfield et al., 2012) (for plants) to predict some unknown small RNAs (microRNAs and piRNAs) based their architectural features.

Expression Identification of Small RNAs

The small RNA expression level is calculated by using TPM, which is standardized for comparison.

Target Prediction of MicroRNAs Using Two Typical Computational Software

The target gene/transcripts of microRNAs may actually reflect the biological functions and significance of microRNAs. We can use two effective software (RNAhybrid and miRanda) to get the target gene of miRNA, extract intersection or union of target gene as final prediction result. The combined target result as shows in **Figure 4**. According to the prediction results, RNAhybrid and miRanda shared various predicted targets (2646560), while RNAhybrid can identify more unique targets comparing to miRanda (4492273 vs 894290). The detailed distribution and comparison of such prediction results can be seen in **Figure 4**.

Screening Differentially Expressed piRNAs

Differentially Expressed small RNAs (DESs) screening is aimed to find differentially expressed small RNA between samples and do the further analysis. We use DEGseq and ExpDiff methods to do this analysis on piRNAs. The DESs in each pairwise as shown in **Figure 5**.

Screening Differentially Expressed miRNAs

Similar with the identification of differentially expressed miRNAs, using software like DEGseq and ExpDiff, we also identified differentially expressed microRNAs in different groups. The distribution of differentially expressed microRNAs in each pairwise as shown in **Figure 6**. According to the figure, we still focused on the differentially expressed miRNA pattern of With DESs, we perform hierarchical clustering of three

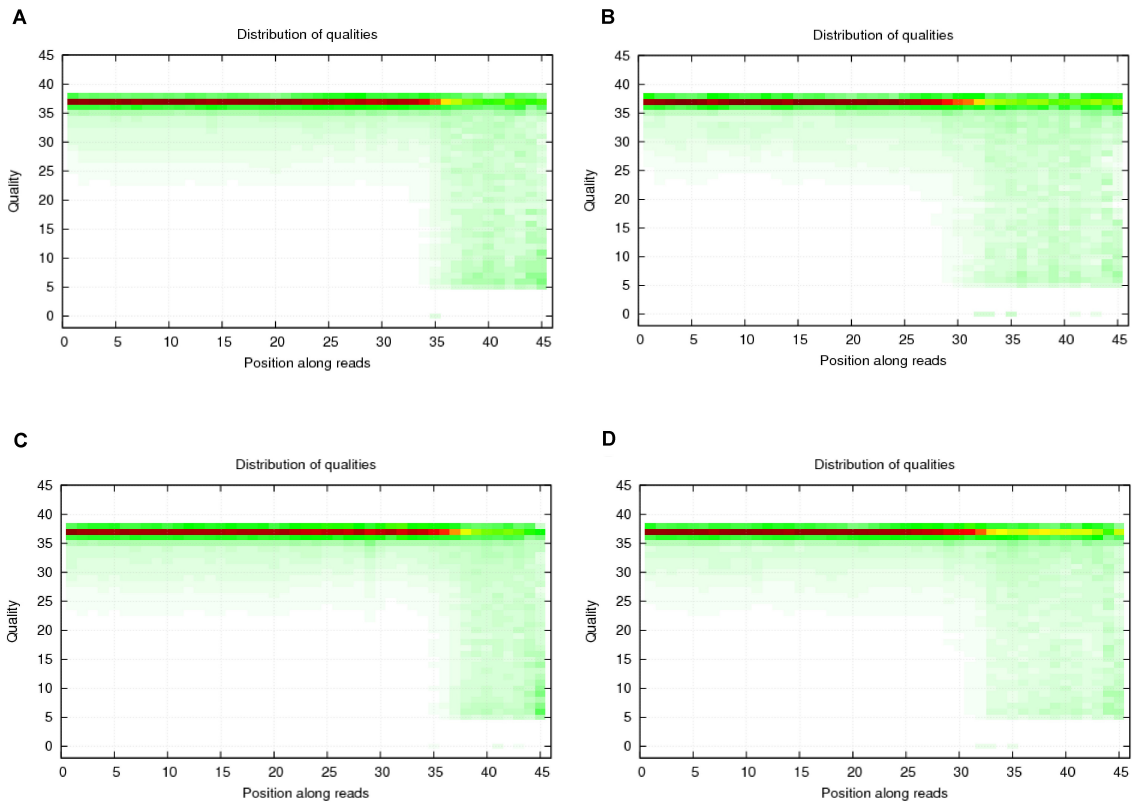

FIGURE 1 | Quality distribution of four sequencing results. **(A)** Negative control group; **(B)** Positive control group; **(C)** P-ONE group; **(D)** P-TWO group. From such four bar plots, we can confirm that all sequencing results are of high quality (greater than 20), satisfying the requirements for further processing and analyses.

comparisons: negative controls and P-ONE; negative controls and P-TWO and negative controls and positive controls. Based on such comparation, we identified various expression statistics at miRNA level:

(1) MicroRNAs have similar alteration pattern in positive controls and P-ONE group, implying that via microRNAs, the therapeutic mechanisms of P-ONE may share some specific regulatory processes with the traditional HR4-based therapeutics.
(2) However, P-TWO may have totally different regulatory mechanism considering its specific different alteration pattern comparing to P-ONE and positive control.

DESs Target Prediction

As we have described in the Methods, we also identified some target of the DESs. The DESs target were performed by using several software.

Gene Ontology Enrichment Analysis of DESs Targets

According to previous analyses, we identified thousands of genes targeted by differential expressed miRNAs. However, it's impossible and unreasonable to analyze the biological effects of such genes one by one. To show the detailed correlations between genes targeted by differentially expressed microRNAs and AR therapeutic effects, here, we introduced gene ontology (The Gene Ontology Consortium, 2016) and KEGG terms (Kanehisa et al., 2016, 2018) to describe the functional distribution of such targeted genes.

Based on the methods we described in Methods, we further performed Gene Ontology (GO) enrichment analysis (The Gene Ontology Consortium, 2016) with screened DESs target genes. GO functional classification is listed to help understanding the distribution of gene functions of the specie from the macro level. To reveal the detailed pharmacological effects of P-ONE and P-TWO, we chose three comparison to show with GO functional classification box plot. The comparison can be seen in **Figure 7**. Comparing the GO classification box plot, it's easy to find out that DES target genes may enrich in similar pattern under three therapeutic conditions, implying that such three therapeutic methods (HR4 antagonist, P-ONE and P-TWO) may have similar pharmacological mechanisms and microRNA may play an irreplaceable role during such processes.

Pathway Enrichment Analysis of DESs Targets

Genes usually interact with each other to play roles in certain biological functions. We perform pathway enrichment analysis

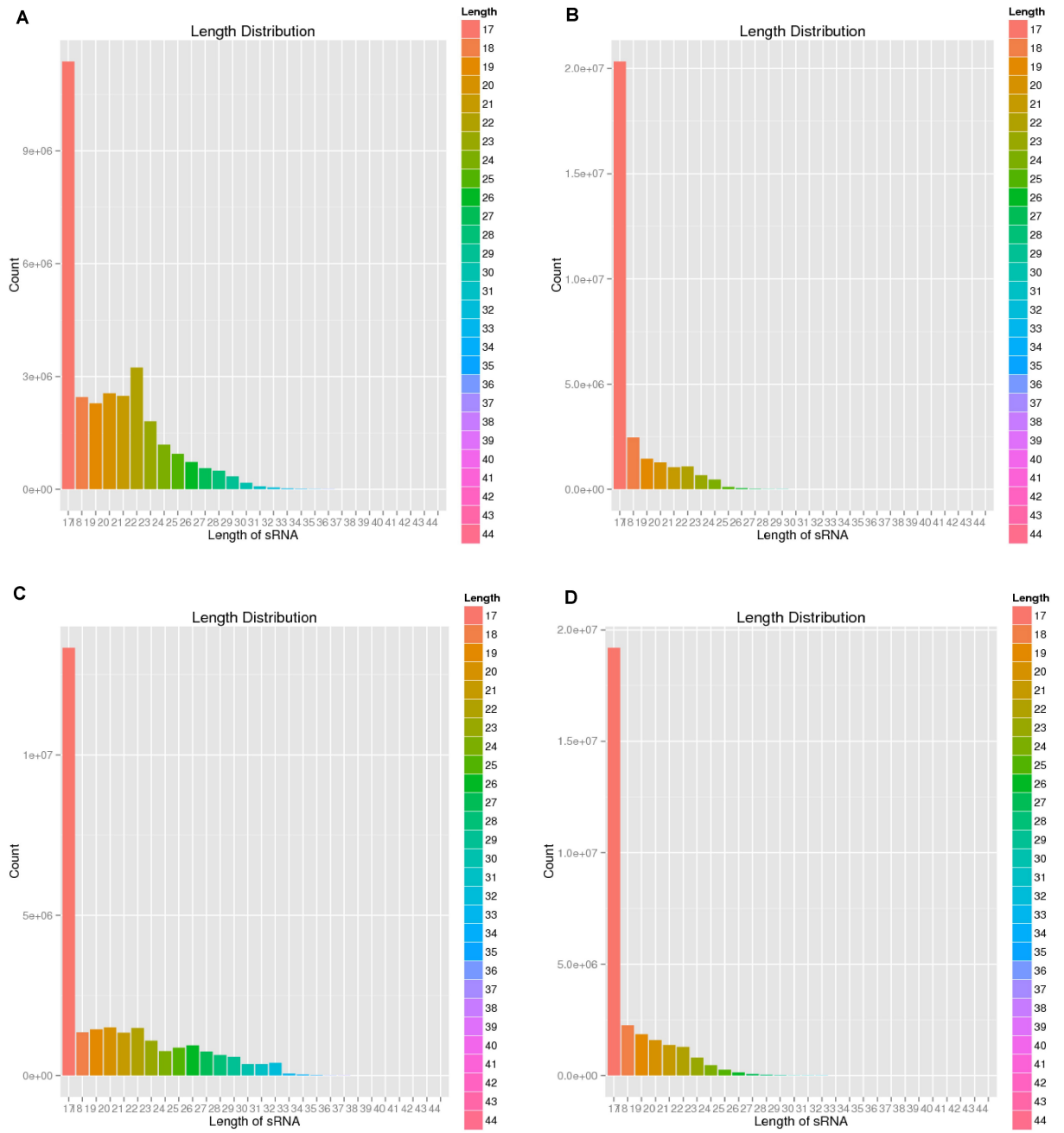

FIGURE 2 | Length distribution of four sequencing results. **(A)** Negative control group; **(B)** Positive control group; **(C)** P-ONE group); **(D)** P-TWO group. From such four bar plots, most of the small RNAs have reasonable length less than 18 nt, corresponding with the general distribution of small RNAs' length. Therefore, such results validated the high-quality of our sequencing and the accurate identification of small RNAs.

of DESs target genes based on KEGG database (Kanehisa et al., 2016, 2018) and generate a report for DESs target genes in each pairwise, respectively. In addition, we generate a scatter plot for the top 20 of KEGG enrichment results as **Figure 8** and a bar plot for the statistics of KEGG terms types as **Figure 9**.

According to the KEGG enrichment figures, we can summarize the different functional enrichment pattern under three therapeutic conditions (HR4 antagonist, P-ONE and P-TWO):

(1) The detailed KEGG enrichment pattern under three conditions are different involving different regulatory pathways.
(2) Some specific pathways like pathways in cancer, TGF-beta signaling pathway and focal adhesion are shared in all

TABLE 4 | Summary of detected tags for each sample.

Sample name	Total tag	Mapped tag	Percentage (%)
Negative Control	21100237	15900803	75.36
Positive Control	25112110	23343614	92.96
P-ONE	27733405	14780247	53.29
P-TWO	23575452	19846022	84.18

the three groups, indicating the potential contribution of such pathways for the pharmacological effects of such three treatment methods.

(3) Still, there are various specific pathways that is differentially enriched in three groups. For instance, PI3K signaling pathway is only enriched in positive control group (HR4 antagonist) and P-TWO treatment group, but not P-ONE treatment group, revealing the potential differences among such three therapeutic methods.

DISCUSSION

Here, as we have presented above, we accomplished a systematic analysis on the small RNA (piRNAs and small RNAs) distribution pattern and potential targeting functional distribution pattern under different therapeutic conditions against AR. To further discuss the underlying therapeutic mechanisms of two reported epitopes against H4 receptor for allergic rhinitis treatment at the microRNA level and try to reveal their potential pharmacological mechanisms by targeting H4 receptors, we divided our discussion in two parts : (1) discussion on the differential small RNA distribution patterns; (2) discussion on functional clustering of genes targeted by the differential expressed microRNAs.

Discussion on the Differential Small RNA Distribution Patterns

As we have shown in **Figures 5**, **6**, it is obvious to see that at piRNA level, although there is differential expression patterns in P-ONE and P-TWO, however, the positive control does not show alterations at piRNA level, indicating that such alteration induced by P-ONE and P-TWO may not be directly correlated with targeting HR4 and therapeutic effects on AR. According to recent publications, no direct reports indicate that piRNAs may play effective role in the regulation of HR4 during the pathogenesis of AR, further explaining the specific pattern of piRNAs in the positive control groups. However, there are various publications, in deed confirmed that piRNAs may contribute to the pathogenesis of AR via some specific regulatory mechanisms like interacting with PTEN (Phosphatase and Tensin homolog)

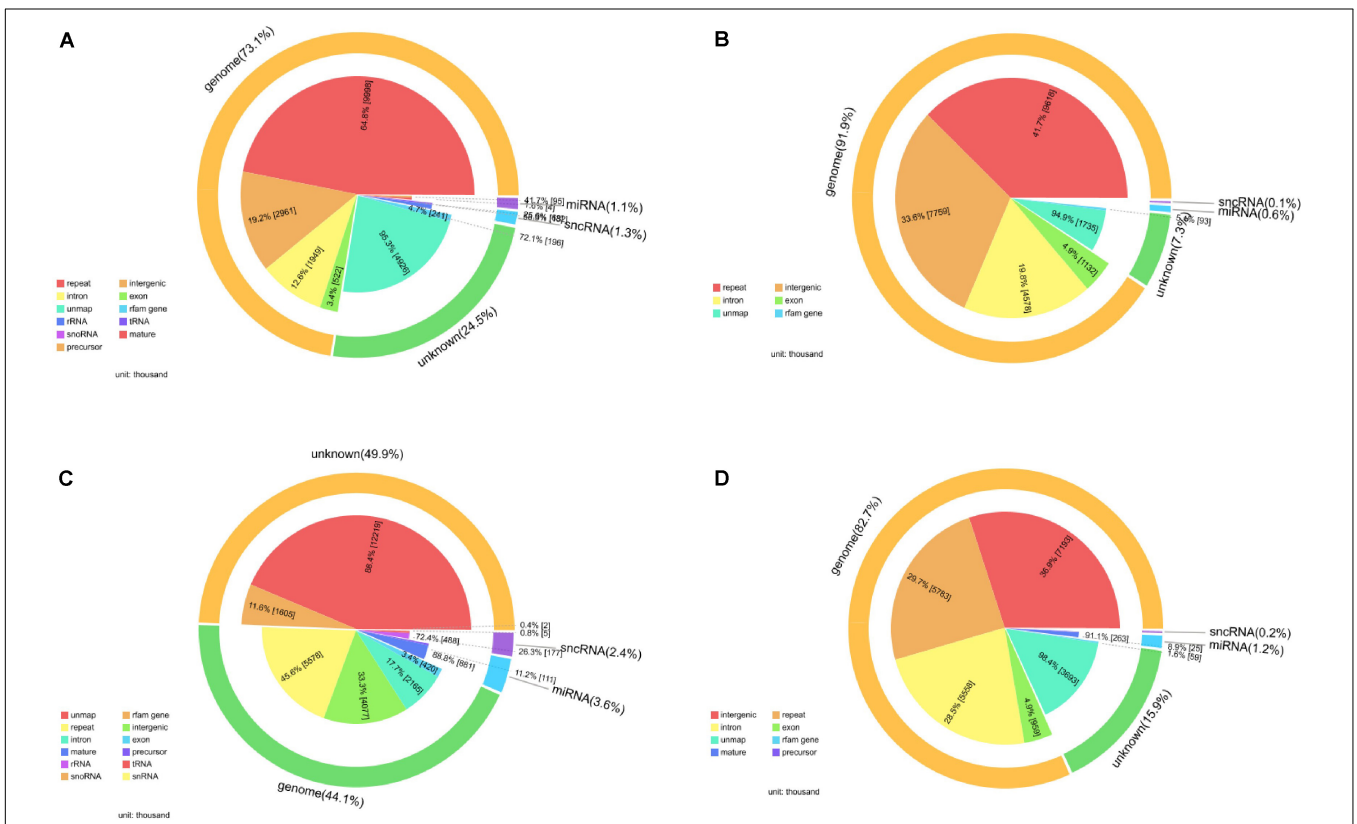

FIGURE 3 | Catalog of small RNAs in four sequencing groups. **(A)** Negative control group; **(B)** positive control group; **(C)** P-ONE group; **(D)** P-TWO group. For all the four samples, most of the so-called small RNAs all come from the genome especially the intergenic regions and only about 5% of all the small RNAs can be defined as microRNA or small non-coding RNAs. And for all the samples, there still remain thousands of unknown small RNAs for further exploration.

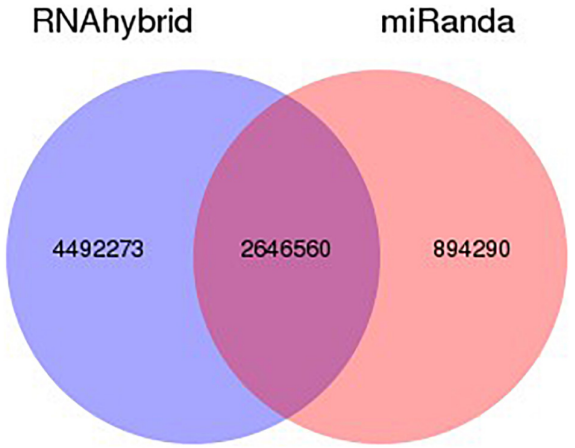

FIGURE 4 | Venn statistics of filtered predicted targets. Using RNAhybrid and miRanda, we identified potential targets for our identified miRNAs. According to the prediction results, among all the predicted targets, most of the targets are shared by both RNAhybrid and miRanda. Comparing such two methods, RNAhybrid may work better with more unique prediction results.

(Alexandrova et al., 2016) and PI3k (Phosphoinositide 3-kinase) (Alexandrova et al., 2016; Narożna et al., 2017). Considering that traditional HR4 antagonists only block HR4 by physical binding, however, our newly identified peptides block HR4 biological functions by triggering specific immune response against HR4. Therefore, although also targeting HR4, such two peptides may also have some unique therapeutic contributions on AR, probably via PTEN or PI3K associated biological processes. The detailed biological mechanisms may still need further molecular and cell biology studies to reveal. What's more, actually, the distribution patterns of P-ONE and P-TWO are also quite different, P-TWO has greater effects on the regulation of piRNAs, indicating that such two peptides may still trigger different immune response and have different pharmacological mechanisms against AR.

Different from the distribution pattern of piRNAs, the distributions of miRNAs are quite similar between positive group and P-ONEp group, indicating that at miRNA level, such two methods may have similar therapeutic effects on AR. However, as for the P-TWO group, the distribution of up-regulated and down regulated microRNAs are reversed. More microRNAs turn out to be up-regulated in such pattern. Such phenotype cannot be explained now. However, at least, such results indicate that P-TWO has quite different pharmacological effects on microRNA level comparing to P-ONE and traditional HR4 antagonists.

All in all, summarized from such figures, we can conclude that:

(1) At piRNA level, P-ONE, P-TWO and traditional HR4 antagonists have totally different expression pattern, indicating their different regulatory effects and pharmacological mechanisms.

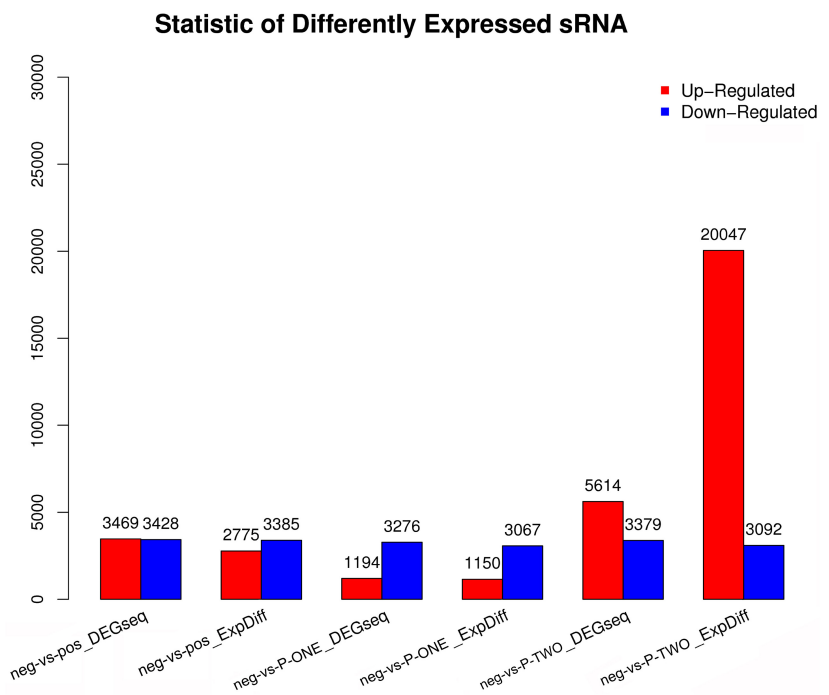

FIGURE 5 | Statistic of differential expressed sRNAs. Using DEGseq nad ExpDiff, we identified the differentially expressed piRNAs from four groups: 1) negative control group (neg); 2) positive control group (pos); 3) P-ONE group; 4) P-TWO group. *X*-axis represents pairwise and *Y*-axis means number of screened DESs. Blue bar denotes down-regulated and orange bar for the up-regulated. According to the figure, comparing to the control group, P-TWO have quite a large group of upregulated pi-RNAs, indicating that P-TWO may have specific pharmacological effects via microRNA regulation.

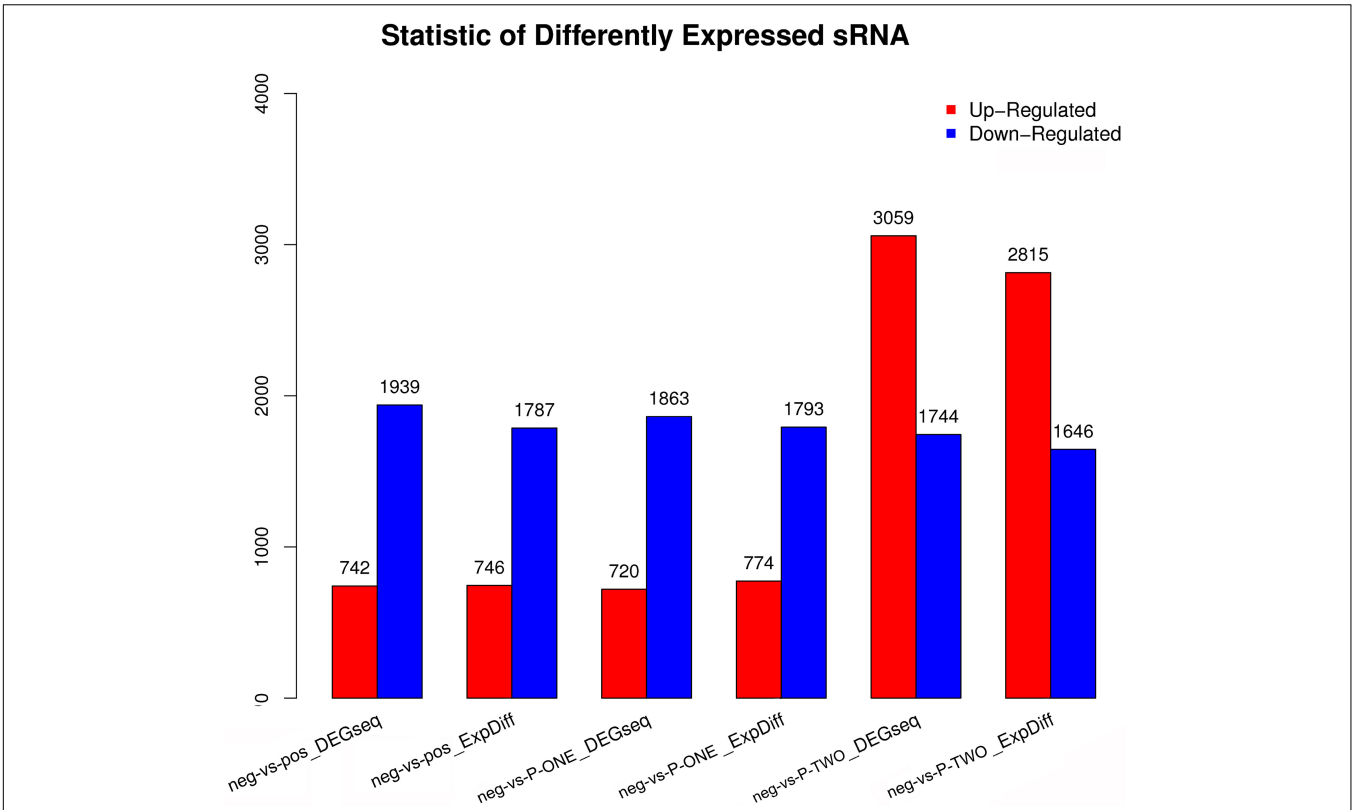

FIGURE 6 | Statistic of differential expressed miRNAs. Using DEGseq nad ExpDiff, we identified the differentially expressed miRNAs from four gourps: 1) negative control group (neg); 2) positive control group (pos); 3) P-ONE group; 4) P-TWO group. *X*-axis represents pairwise and *Y*-axis means number of screened DESs. Blue bar denotes down-regulated and orange bar for the upregulated. According to the figure, P-ONE and traditional HR4-targeted method may have similar regulatory mechanisms via microRNAs but P-TWO may have its specific pharmacological effects and mechanisms via microRNA regulation.

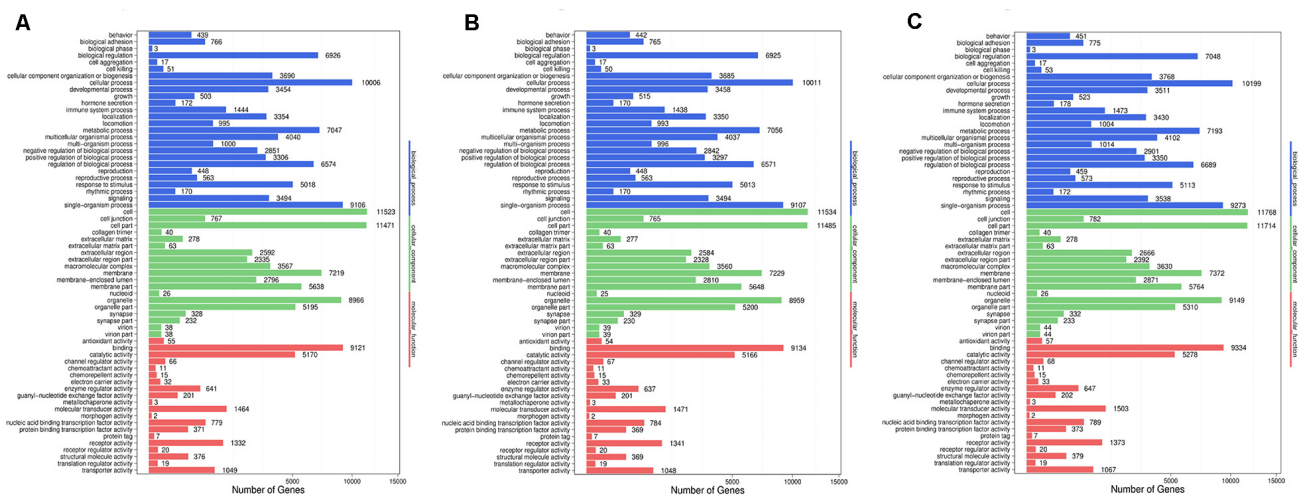

FIGURE 7 | Go functional classification of DES target genes from three comparison. **(A)** Comparison between negative control and positive control. **(B)** Comparison between negative control and P-ONE. **(C)** Comparison between negative control and P-TWO. *X*-axis means number of DEGs (the number is presented by its square root value). *Y*-axis represents GO terms. All GO terms are grouped in to three ontologies: blue is for biological processes; brown is for cellular component and orange is for molecular functions. According to three plots, DES target genes may enrich in similar pattern under three therapeutic conditions, implying that such three therapeutic methods (HR4 antagonist, P-ONE and P-TWO) may have similar pharmacological mechanisms and microRNA may play an irreplaceable role during such processes.

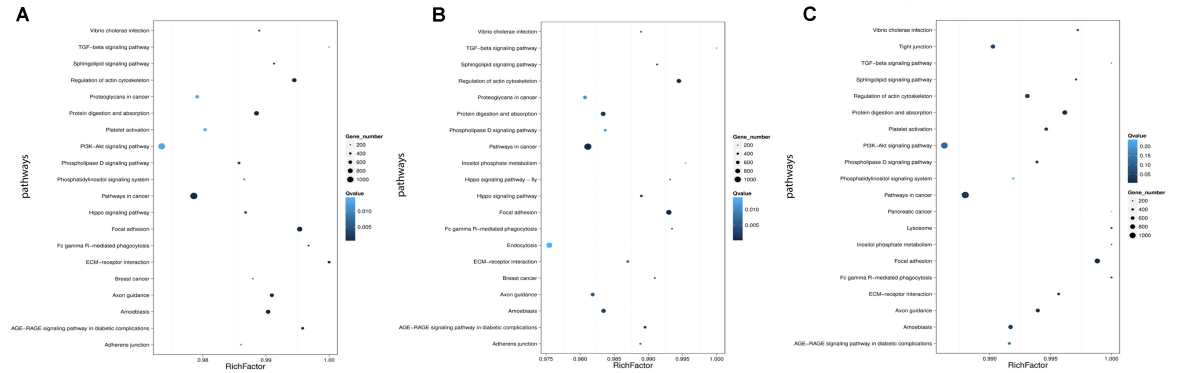

FIGURE 8 | Statistics of pathway enrichment in each pairwise. **(A)** Comparison between negative control group and positive control group. **(B)** Comparison between negative control group and P-ONE. **(C)** Comparison between negative control group and P-TWO. Rich Factor is the ratio of DESs target genes numbers annotated in this pathway term to all gene numbers annotated in this pathway term. Greater Rich Fator means greater intensiveness. *Q*-value is corrected *P*-value ranging from 0 to 1, and less *Q*-value means greater intensiveness. We just display the top 20 of enriched pathway terms.

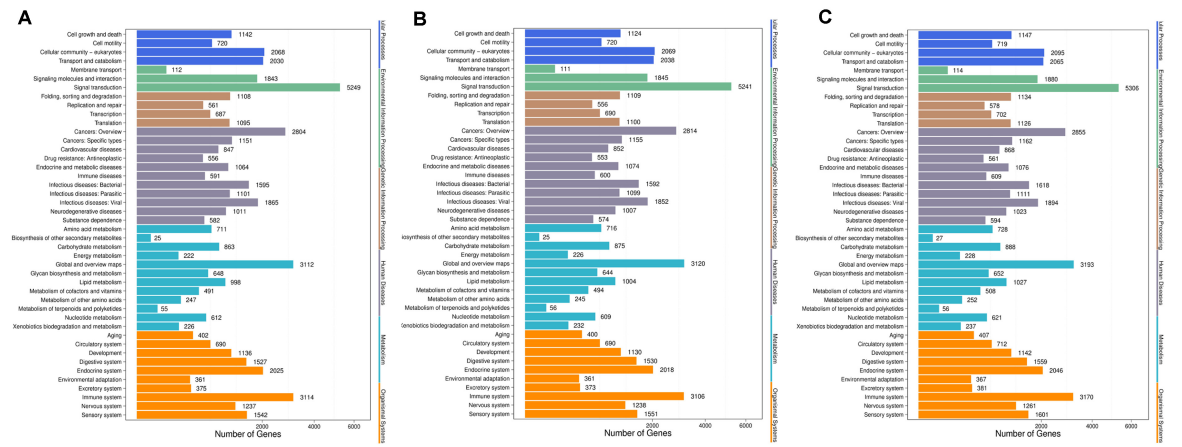

FIGURE 9 | KEGG classification of each pairwise. **(A)** Comparison between negative control group and positive control group. **(B)** Comparison between negative control group and P-ONE. **(C)** Comparison between negative control group and P-TWO. *X*-axis means number of DEGs. *Y*-axis represents second KEGG pathway terms. All second pathway terms are grouped in top pathway terms indicated in different color.

(2) At miRNA level, P-ONE may have similar therapeutic effects with HR4 antagonists but P-TWO has quite unique therapeutic effects on such level.
(3) The typical alteration of small RNA expression level confirmed that small RNAs in deed play an irreplaceable role during the therapy of AR and participate in the pharmacological mechanisms of such medicine.
(4) Also, in terms of methodology, software DEGseq and ExpDiff may have quite comparable results.

Discussion on Functional Clustering of Genes Targeted by the Differential Expressed MicroRNAs

Apart from such phenotypic discussion on the expression comparison of small RNAs in different groups, using gene ontology and KEGG annotation and clustering, we also identified some specific enrichment patterns under different therapeutic conditions, helping reveal the potential pharmacological effects of P-ONE and P-TWO comparing to traditional HR4 antagonists.

Here, firstly, we focused on **Figure 7** describing the results of gene ontology enrichment analyses. Based on the gene ontology classification, we can summarize that the microRNA target expression pattern is quite similar under such three treatment conditions. Therefore, according to such results, although some regulatory details of P-ONE and P-TWO are different from traditional HR4 antagonists, actually, the comprehensive regulatory effects of such two peptides may still be the same at microRNA regulatory level. Further, such results also confirmed that new drugs like P-ONE and P-TWO only affect similar biological processes comparing with previous HR4 antagonists. Therefore, such two peptides may also be safe to be used in further therapies.

Apart from gene ontology, we also focused on the KEGG annotation and clustering results. Based on **Figure 8**, we

presented the pathway enrichment pattern in each pairwise. Here, we identified some specific KEGG pathways that differentially enriched in different experimental groups.

Firstly, there are still some shared KEGG pathways that have been identified under all the same conditions, indicating its specific role for AR pathogenesis and therapies at microRNA regulatory level. For instance, TGF-beta signaling pathway, according to recent publications, such biological process has been widely reported to be a specific pathological pathway for AR. Early in 1992, researchers in the United States have confirmed that TGF beta 1 as a core regulator in such signaling pathway contribute to the pathogenesis of chronically inflammation in human upper airway tissues, related to the onset of allergic rhinitis (Ohno et al., 1992). Further in 2002, another independent study further confirmed the pathogenesis of allergic rhinitis is directly correlated with TGF-beta effects (Benson et al., 2002). Therefore, the identification of such pathway by all the three therapeutic treatment confirmed that such two new medicine also relied on interfering one of the most significant pathways of AR to cure such disease. What's more, more recent publications (Akdis et al., 2005; Jutel et al., 2006; Kucuksezer et al., 2013) on TGF-beta and allergic rhinitis also indicate that TGF-beta is associated with the abnormal immune responses of AR, corresponding with the designed principal of P-ONE and P-TWO which is triggering antigen-specific immune response against HR4.

Apart from such shared biological processes, we also identified some effective biological processes that is only recognized by P-ONE and P-TWO. For P-ONE, endocytosis is a unique pathway with quite low *Q*-value and has not been identified by group positive control and P-TWO. In 2019, a specific publication (Blanco-Pérez et al., 2019)confirmed that a unique pattern of endocytosis mediated allergen fusion contributing to the relief of specific allergies, implying that endocytosis may also contribute to the pathogenesis of AR. The functional enrichment of P-ONE associated microRNA targets may indicate that P-ONE may potential inhibit abnormal allergic effects by interfering allergen fusion, presenting a new theory for the pharmacological effects of P-ONE. Similarly, as for P-TWO, there are still some detailed biological processes and pathways that are uniquely identified in such group. For instance, the lysosome, although with a relatively high *q*-value, recent publications (Ring and Munehen, 1983; Kohno et al., 1987; Liu et al., 2005) also reported that such biological process also regulated the abnormal immune response of AR. In 2005, a specific histopathological study (Liu et al., 2005) on allergic rhinitis confirmed that another drug named as Centipeda minima treats AR by interfering lysosome associated biological processes. Therefore, the enrichment of microRNA targets in such biological process may indicate that P-TWO, our new peptide drug may interact with lysosome associated biological processes and interfere the pathogenesis of AR under certain mechanisms.

Further, we identified the KEGG classification pattern for each pairwise. Although we have identified various unique KEGG pathways for each comparison, the general classification pattern of such three pariwises are quite similar with each other, implying the general therapeutic effects contributed by microRNA regulation and the safety of our new drugs P-ONE and P-TWO.

All in all, as we have mentioned analyzed above, at the functional level, we can summarize that:

(1) Both P-ONE and P-TWO has similar general and comprehensive therapeutic effects comparing to traditional HR4 antagonists at microRNA regulation level according to gene ontology analyses.
(2) The general pharmacological effects of P-ONE and P-TWO are similar with those of traditional HR4 antagonists at microRNA regulatory level. Therefore, P-ONE and P-TWO may be safe to be applied in clinics considering its systematic effects in vivo.
(3) According to KEGG pathway enrichment analyses, there are still some differential regulatory effects of different treatment strategies at microRNA regulatory level. The biological foundations of differential therapeutic effects induced by P-ONE and P-TWO have all been supported by recent publications.
(4) Some specific pathways like endocytosis, lysosomes, hippo signaling pathway and inositol phosphate metabolism may be significant and specific pharmacological mechanisms for our new drugs P-ONE and P-TWO comparing with previously widely reported HR4 antagonists.

CONCLUSION

Relied on stable AR models, we identified the pharmacological effects of our two new candidate peptide drugs P-ONE and P-TWO on the small RNA level comparing to traditional HR4 targeting antagonists. Based on the small RNA profiling results, we firstly confirmed that P-ONE, P-TWO and traditional HR4 targeting antagonists have specific therapeutic on AR at microRNA level. Apart from that, the comprehensive effects of such three treatment strategies are quite similar. For details, based on KEGG pathway enrichment analysis, we also identified some unique pharmacological effects of new drugs P-ONE and P-TWO. All in all, using small RNA sequencing techniques, for the first time, we compared the pharmacological effects of P-ONE, P-TWO and traditional drugs and revealed both the similarities and the differences of such strategies at small RNA regulatory level, laying a solid foundation for the comprehensive understanding of the new drugs' pharmacological mechanisms and the potential pathogenesis of AR.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in NCBI SRA. The study number of SRA database is SRP278422 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study~=~SRP278422), and the BioProject number is PRJNA658395 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA658395).

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) of Jilin University approved all animal procedures [permit number: SYXK(2014-0012)].

AUTHOR CONTRIBUTIONS

LL: conception or design of the work. L-FA: data collection. Z-DL: data analysis and interpretation. L-FA, Z-DL, and LL: manuscript drafting and final approval of the version to be published. All authors contributed to the article and approved the submitted version.

FUNDING

The work presented in this report is the subject of two patents filed by Jilin University (CN201510382851.1 [P].2015-11-04 and CN201510382781.X [P].2015-11-11). This study was supported by the National Natural Science Foundation of China (81100702), the Health and Family Planning Foundation of Jilin Province (20152046), and the Science and Technology Development Plan Foundation of Jilin Province (20160101070JC).

REFERENCES

Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4:e05005.

Akdis, M., Blaser, K., and Akdis, C. A. (2005). T regulatory cells in allergy: novel concepts in the pathogenesis, prevention, and treatment of allergic diseases. *J. Allergy Clin. Immunol.* 116, 961–968. doi: 10.1016/j.jaci.2005.09.004

Al Suleimani, M., Ying, D., and Walker, M. J. (2007). A comprehensive model of allergic rhinitis in guinea pigs. *J. Pharmacol. Toxicol. Methods* 55, 127–134. doi: 10.1016/j.vascn.2006.05.005

Alexandrova, E., Miglino, N., Hashim, A., Nassa, G., Stellato, C., Tamm, M., et al. (2016). Small RNA profiling reveals deregulated phosphatase and tensin homolog (PTEN)/phosphoinositide 3-kinase (PI3K)/Akt pathway in bronchial smooth muscle cells from asthmatic patients. *J. Allergy Clin. Immunol.* 137, 58–67. doi: 10.1016/j.jaci.2015.05.031

Atkinson, G. (2002). Analysis of repeated measurements in physical therapy research: multiple comparisons amongst level means and multi-factorial designs. *Phys. Ther. Sport* 3, 191–203. doi: 10.1054/ptsp.2002.0123

Audic, S., and Claverie, J.-M. (1997). The significance of digital gene expression profiles. *Genome Res.* 7, 986–995. doi: 10.1101/gr.7.10.986

Baran-Gale, J., Kurtz, C. L., Erdos, M. R., Sison, C., Young, A., Fannin, E. E., et al. (2015). Addressing bias in small RNA library preparation for sequencing: a new protocol recovers MicroRNAs that evade capture by current methods. *Front. Genet.* 6:352. doi: 10.3389/fgene.2015.00352

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

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* 29, 1165–1188.

Benson, M., Carlsson, B., Carlsson, L. M., Mostad, P., Svensson, P.-A., and Cardell, L.-O. (2002). DNA microarray analysis of transforming growth factor-β and related transcripts in nasal biopsies from patients with allergic rhinitis. *Cytokine* 18, 20–25. doi: 10.1006/cyto.2002.1012

Blanco-Pérez, F., Papp, G., Goretzki, A., Möller, T., Anzaghe, M., and Schülke, S. (2019). Adjuvant allergen fusion proteins as novel tools for the treatment of Type I allergies. *Arch. Immunol. Ther. Exp.* 67, 273–293. doi: 10.1007/s00005-019-00551-8

Bousquet, J., Arnavielhe, S., Bedbrook, A., Fonseca, J., Morais Almeida, M., Todo Bom, A., et al. (2018). The Allergic Rhinitis and its Impact on Asthma (ARIA) score of allergic rhinitis using mobile technology correlates with quality of life: the MASK study. *Allergy* 73, 505–510.

Breakfield, N. W., Corcoran, D. L., Petricka, J. J., Shen, J., Sae-Seaw, J., Rubio-Somoza, I., et al. (2012). High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. *Genome Res.* 22, 163–176. doi: 10.1101/gr.123547.111

Broide, D. H. (2010). Allergic rhinitis: pathophysiology. *Allergy Asthma Proc.* 31, 370–374. doi: 10.2500/aap.2010.31.3388

Buschmann, D., Haberberger, A., Kirchner, B., Spornraft, M., Riedmaier, I., Schelling, G., et al. (2016). Toward reliable biomarker signatures in the age of liquid biopsies - how to standardize the small RNA-Seq workflow. *Nucleic Acids Res.* 44, 5995–6018. doi: 10.1093/nar/gkw545

Cardell, L.-O., Olsson, P., Andersson, M., Welin, K.-O., Svensson, J., Tennvall, G. R., et al. (2016). TOTALL: high cost of allergic rhinitis—a national Swedish population-based questionnaire study. *NPJ Prim. Care Respirat. Med.* 26:15082.

Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/nar/gkp1137

Fahlgren, N., and Carrington, J. C. (2010). miRNA target prediction in plants. *Methods Mol. Biol.* 592, 51–57. doi: 10.1007/978-1-60327-005-2_4

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., et al. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* 26:407. doi: 10.1038/nbt1394

Fuchs, R. T., Sun, Z., Zhuang, F., and Robb, G. B. (2015). Bias in ligation-based small RNA sequencing library construction is determined by adaptor and RNA structure. *PLoS One* 10:e0126049. doi: 10.1371/journal.pone.0126049

García-Martín, E., García-Menaya, J., Sanchez, B., Martínez, C., Rosendo, R., and Agúndez, J. (2007). Polymorphisms of histamine-metabolizing enzymes and clinical manifestations of asthma and allergic rhinitis. *Clin. Exper. Allergy* 37, 1175–1182. doi: 10.1111/j.1365-2222.2007.02769.x

Han, Y.-Y., Forno, E., Gogna, M., and Celedón, J. C. (2016). Obesity and rhinitis in a nationwide study of children and adults in the United States. *J. Allergy Clin. Immunol.* 137, 1460–1465. doi: 10.1016/j.jaci.2015.12.1307

Hill, D. A., Grundmeier, R. W., Ram, G., and Spergel, J. M. (2016). The epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children: a retrospective cohort study. *BMC Pediatr.* 16:133. doi: 10.1186/s12887-016-0673-z

Jiang, H., and Wong, W. H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 25, 1026–1032. doi: 10.1093/bioinformatics/btp113

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human microRNA targets. *PLoS Biol.* 2:e363. doi: 10.1371/journal.pone.000363

Jung, S., Lee, S.-Y., Yoon, J., Cho, H.-J., Kim, Y.-H., Suh, D. I., et al. (2020). Risk factors and comorbidities associated with the allergic rhinitis phenotype in children according to the ARIA classification. *Allergy Asthma Immunol. Res.* 12, 72–85. doi: 10.4168/aair.2020.12.1.72

Jutel, M., Blaser, K., and Akdis, C. A. (2006). "The role of histamine in regulation of immune responses," in *Allergy and Asthma in Modern Society: A Scientific Approach*, ed. R. Crameri (Berlin: Karger Publishers), 174–187. doi: 10.1159/000090280

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36(Suppl._1), D480–D484.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2018). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* 47, D590–D595.

Kohno, H., Inoue, H., Seyama, Y., Yamashita, S., and Akasu, M. (1987). Mode of the anti-allergic action of cepharanthine on an experimental model of allergic rhinitis. Nihon yakurigaku zasshi. *Folia Pharmacol. Jpn.* 90, 205–211. doi: 10.1254/fpj.90.205

Kolde, R., and Kolde, M. R. (2015). *Package 'Pheatmap'. R Package 1.7*.

Krüger, J., and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.* 34(Suppl._2), W451–W454.

Kucuksezer, U. C., Ozdemir, C., Akdis, M., and Akdis, C. A. (2013). Mechanisms of immune tolerance to allergens in children. *Korea. J. Pediatr.* 56:505. doi: 10.3345/kjp.2013.56.12.505

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25.

Ledford, D. K., and Lockey, R. F. (2016). Aspirin or nonsteroidal anti-inflammatory drug-exacerbated chronic rhinosinusitis. *J. Allergy Clin. Immunol.* 4, 590–598. doi: 10.1016/j.jaip.2016.04.011

Liu, Z., Yu, H., Wen, S., and Liu, Y. (2005). Histopathological study on allergic rhinitis treated with Centipeda minima. *China J. Chin. Mater. Med.* 30, 292–294.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clin. Exper. Pharmacol. Physiol.* 25, 1032–1037. doi: 10.1111/j.1440-1681.1998.tb02179.x

Lundberg, K., Broos, S., Greiff, L., Borrebaeck, C. A., and Lindstedt, M. (2011). Histamine H4 receptor antagonism inhibits allergen-specific T-cell responses mediated by human dendritic cells. *Eur. J. Pharmacol.* 651, 197–204. doi: 10.1016/j.ejphar.2010.10.065

Maoz-Segal, R., Machnes-Maayan, D., Veksler-Offengenden, I., Frizinsky, S., Hajyahia, S., and Agmon-Levin, N. (2019). "Local allergic rhinitis: an old story but a new entity," in *Rhinosinusitis*, eds B. S. Gendeh and M. Turkalj (London: IntechOpen). doi: 10.5772/intechopen.86212

Maziere, P., and Enright, A. J. (2007). Prediction of microRNA targets. *Drug Discov. Today* 12, 452–458.

Meltzer, E. O., Malmstrom, K., Lu, S., Prenner, B. M., Wei, L. X., Weinstein, S. F., et al. (2000). Concomitant montelukast and loratadine as treatment for seasonal allergic rhinitis: a randomized, placebo-controlled clinical trial. *J. Allergy Clin. Immunol.* 105, 917–922. doi: 10.1067/mai.2000.106040

Narita, S.-I., Asakura, K., Shirasaki, H., Isobe, M., Ogasawara, H., Saito, H., et al. (1998). Effects of cyclosporin A and glucocorticosteroids on antigen-induced hypersensitivity to histamine in a guinea pig model of allergic rhinitis. *Inflamm. Res.* 47, 62–66. doi: 10.1007/s000110050274

Narożna, B., Langwiński, W., and Szczepankiewicz, A. (2017). Non-coding RNAs in pediatric airway diseases. *Genes* 8:348. doi: 10.3390/genes8120348

Nawrocki, E. P., and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29, 2933–2935. doi: 10.1093/bioinformatics/btt509

Ohno, I., Lea, R., Flanders, K., Clark, D., Banwatt, D., Dolovich, J., et al. (1992). Eosinophils in chronically inflamed human upper airway tissues express transforming growth factor beta 1 gene (TGF beta 1). *J. Clin. Invest.* 89, 1662–1668. doi: 10.1172/jci115764

Peng, H., Wang, J., Ye, X. Y., Cheng, J., Huang, C. Z., Li, L. Y., et al. (2019). Histamine H4 receptor regulates IL-6 and INF-γ secretion in native monocytes from healthy subjects and patients with allergic rhinitis. *Clin. Transl. Allergy* 9, 1–4.

Price, D. B., Smith, P. K., Harvey, R. J., Carney, A. S., Kritikos, V., Bosnic-Anticevich, S. Z., et al. (2018). Real-life treatment of rhinitis in Australia: a historical cohort study of prescription and overthe-counter therapies for patients with and without additional respiratory disease. *Pragmat. Obs. Res.* 9, 43–54. doi: 10.2147/por.s153266

Ring, J., and Munehen, J. L. (1983). Decreased release of lysosomal enzymes from peripheral leukocytes of patients with atopic dermatitis. *J. Am. Acad. Dermatol.* 8, 378–385. doi: 10.1016/s0190-9622(83)70043-5

Scholkopf, B., and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond*. Cambridge, MA: MIT press.

Settipane, R. A., and Charnock, D. R. (2016). "Epidemiology of rhinitis: allergic and nonallergic," in *Nonallergic Rhinitis*, eds J. N. Baraniuk, and D. J. Shusterman (Boca Raton, FL: CRC Press), 45–56. doi: 10.3109/9781420021172-6

Shiraishi, Y., Jia, Y., Domenico, J., Joetham, A., Karasuyama, H., Takeda, K., et al. (2013). Sequential engagement of FcεRI on mast cells and basophil histamine H4 receptor and FcεRI in allergic rhinitis. *J. Immunol.* 190, 539–548. doi: 10.4049/jimmunol.1202049

Smith, P., Price, D., Harvey, R., Carney, A. S., Kritikos, V., Bosnic-Anticevich, S. Z., et al. (2017). Medication-related costs of rhinitis in Australia: a NostraData cross-sectional study of pharmacy purchases. *J. Asthma Allergy* 10:153. doi: 10.2147/jaa.s128431

Specjalski, K., Maciejewska, A., Pawłowski, R., Chełmińska, M., and Jassem, E. (2016). Changes in the expression of microRNA in the buildup phase of wasp venom immunotherapy: a pilot study. *Int. Arch. Allergy Immunol.* 170, 97–100. doi: 10.1159/000447637

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Takahashi, Y., Kagawa, Y., Izawa, K., Ono, R., Akagi, M., and Kamei, C. (2009). Effect of histamine H4 receptor antagonist on allergic rhinitis in mice. *Int. Immunopharmacol.* 9, 734–738. doi: 10.1016/j.intimp.2009.02.011

The Gene Ontology Consortium (2016). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338.

Turner, P. J., Worm, M., Ansotegui, I. J., El-Gamal, Y., Rivas, M. F., Fineman, S., et al. (2019). Time to revisit the definition and clinical criteria for anaphylaxis? *World Allergy Organ. J.* 12:100066. doi: 10.1016/j.waojou.2019.100066

Walter, M., Kottke, T., and Stark, H. (2011). The histamine H4 receptor: targeting inflammatory disorders. *Eur. J. Pharmacol.* 668, 1–5. doi: 10.1016/j.ejphar.2011.06.029

Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., et al. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinform.* 15:419. doi: 10.1186/s12859-014-0419-6

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2009). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26, 136–138. doi: 10.1093/bioinformatics/btp612

Wang, Y., Sha, J., Wang, H., An, L., Liu, T., and Li, L. (2018). P-FN12, an H4R-based epitope vaccine screened by phage display, regulates the Th1/Th2 balance in rat allergic rhinitis. *Mol. Ther. Methods Clin. Dev.* 11, 83–91. doi: 10.1016/j.omtm.2018.09.004

Weisstein, E. W. (2004). *Bonferroni Correction*. Available online at: https://mathworld.wolfram.com/BonferroniCorrection.html (accessed September 1, 2020).

Wu, H.-J., Ma, Y.-K., Chen, T., Wang, M., and Wang, X.-J. (2012). PsRobot: a web-based plant small RNA meta-analysis toolbox. *Nucleic Acids Res.* 40, W22–W28.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., et al. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 30:e15.

Zeyrek, D., Tanac, R., Altinoz, S., Berdeli, A., Gulen, F., Koksoy, H., et al. (2008). FcγRIIIa-V/F 158 polymorphism in Turkish children with asthma bronchiale and allergic rhinitis. *Pediatr. Allergy Immunol.* 19, 20–24. doi: 10.1111/j.1399-3038.2007.00553.x

Zhao, X. J. (1993). Experimental models of nasal hypersensitive reaction. *Zhonghua er bi yan hou ke za zhi* 28, 17–18, 58–59.

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 Physiology

ORIGINAL RESEARCH
published: 16 September 2020
doi: 10.3389/fphys.2020.573433

Proteomic Analysis of Atrial Appendages Revealed the Pathophysiological Changes of Atrial Fibrillation

Ban Liu[1†], Xiang Li[2†], Cuimei Zhao[3†], Yuliang Wang[4†], Mengwei Lv[5,6], Xin Shi[7], Chunyan Han[1], Pratik Pandey[1], Chunhua Qian[8], Changfa Guo[9]* and Yangyang Zhang[6]**

[1] Department of Cardiology, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [2] Department of Cardiology, The First Affiliated Hospital of Chongqing Medical University, Chongqing Medical University, Chongqing, China, [3] Department of Cardiology, Tongji Hospital, Tongji University School of Medicine, Shanghai, China, [4] Department of Immunology, School of Basic Medical Science, Nanjing Medical University, Nanjing, China, [5] Shanghai East Hospital of Clinical Medical College, Nanjing Medical University, Shanghai, China, [6] Department of Cardiovascular Surgery, Shanghai East Hospital, Tongji University School of Medicine, Shanghai, China, [7] Department of Pediatric Cardiology, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China, [8] Department of Endocrinology and Metabolism, Shanghai Tenth People's Hospital, Tongji University School of Medicine, Shanghai, China, [9] Department of Cardiovascular Surgery, Zhongshan Hospital, Fudan University, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Ye Yang,
Nanjing University of Chinese Medicine, China
Xue Liang,
The Fifth Affiliated Hospital of Guangzhou Medical University, China

***Correspondence:**
Chunhua Qian
chqian2003@126.com
Changfa Guo
guo.changfa@zs-hospital.sh.cn
Yangyang Zhang
zhangyangyang_wy@vip.sina.com

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *17 June 2020*
Accepted: *13 August 2020*
Published: *16 September 2020*

Citation:
Liu B, Li X, Zhao C, Wang Y, Lv M, Shi X, Han C, Pandey P, Qian C, Guo C and Zhang Y (2020) Proteomic Analysis of Atrial Appendages Revealed the Pathophysiological Changes of Atrial Fibrillation. Front. Physiol. 11:573433. doi: 10.3389/fphys.2020.573433

Atrial fibrillation (AF), known as the most common arrhythmia in the developed world, affects 1.5–2.0% of the population. Numerous basic studies have been carried out to identify the roles of electric and structural remodeling in the pathophysiological changes of AF, but more explorations are required to further understand the mechanisms of AF development. Proteomics enables researchers to identify protein alterations responsible for the pathological developing progresses of diseases. Compared to the genome, the proteome is closely related to the disease phenotype and can better manifest the progression of diseases. In this study, AF patients proteomically analyzed to identify possible mechanisms. Totally 20 patients undergoing cardiac surgery (10 with paroxysmal AF and 10 with persistent AF) and 10 healthy subjects were recruited. The differentially expressed proteins identified here included AKR1A1, LYZ, H2AFY, DDAH1, FGA,FGB, LAMB1, LAMC1, MYL2, MYBPC3, MYL5, MYH10, HNRNPU, DKK3, COPS7A, YWHAQ, and PAICS. These proteins were mainly involved in the development of structural remodeling. The differently expressed proteins may provide a new perspective for the pathological process of AF, and may enable useful targets for drug interference. Nevertheless, more research in terms of multi-omics is required to investigate possible implicated molecular pathways of AF development.

Keywords: atrial fibrillation, proteomics, proteins, structural remodeling, mechanism

INTRODUCTION

Atrial fibrillation (AF), known as the most common arrhythmia in the developed world, attacks 1.5–2.0% of the population. In the population aged over 40 years, the lifetime risk for AF is about 25% both in genders (Heeringa et al., 2006). The incidence of AF has risen about threefold with the aging population during the next 50 years, which progressively increases economic burden (Steinberg, 2004; Miyasaka et al., 2006). AF is characterized electrocardiographically by low-amplitude baseline

oscillations as supraventricular arrhythmia. The fibrillatory waves, namely f waves, originate from the fibrillating atria and are accompanied by an irregular ventricular rhythm. AF mainly causes cardiovascular mortality and morbidity (Heeringa et al., 2006). A variety of cardiac diseases and conditions may cause atrial remodeling and consequently lead to AF development, but AF may also contribute to atrial remodeling owing to the progressiveness of the arrhythmia (Wakili et al., 2011).

These remodeling approaches include structural remodeling characterized as atrial fibrosis (Frustaci et al., 1997) and atrial adipose (Hatem and Sanders, 2014), electrical remodeling featured by changes in ion channels and gap junction proteins (Lai et al., 1999), and endocardial and metabolic remodeling (Schild et al., 2006; Jeganathan et al., 2017). Numerous basic studies have been conducted to explore the roles of electric, structural and contractile remodeling in the pathophysiological changes of AF. Nevertheless, further explorations are required to better understand the mechanisms of AF development.

Various techniques, especially "omics" techniques, have been applied to identify the molecular targets and mechanisms that mediate AF-related remodeling. Proteomics is one "omics" technique to study large-scale gene expression at the protein level, and enables researchers to identify protein alterations responsible for the pathological developing progresses of diseases. The proteome determines the cell phenotype and variations that may change cell and tissue functions. Compared to the genome, the proteome is closely related to the disease phenotype and can better manifest the progression of diseases.

In this study, AF patients were categorized into two groups according to the duration of AF. Paroxysmal AF was termed as terminating spontaneously within 7 days, and permanent AF was defined as persisting for more than 1 year. We compared the proteomics between subjects with sinus rhythm (SR) and patients with AF to demonstrate the pathophysiological changes.

MATERIALS AND METHODS

Patients and Tissue Preparation

Thirty subjects were enrolled and divided into three groups, including 10 healthy subjects with SR (Group1, G1), 10 patients with paroxysmal AF (Group2, G2), and 10 patients with permanent AF (Group3, G3). The 10 healthy subjects with SR were all males and aged between 25 and 38 years old. All AF patients were subjected to physical examination and clinical evaluation, including medical history, routine blood test, electrocardiography (ECG), chest CT, and echocardiography. Exclusion criteria were valvular heart disease, coronary artery disease, chronic heart failure, myocarditis, cardiomyopathy, chronic pulmonary heart disease, or hyperthyroidism.

Protocol for sample collection was adhered to the Human Ethics Committee of Shanghai East Hospital (DI:0402017). This study complied with the Helsinki Declaration. Prior to operation of fibrillation ablation, written informed consents were obtained from all enrolled patients. The left atrial appendage (LAA) was resected during isolated surgical ablation, and tissue samples were collected from the abandoned LAA. Normal LAA samples were collected from healthy male donors. Collected tissues were frozen in −80°C liquid nitrogen before further processing.

Protein Extraction

The extraction of proteins from atrial tissues followed previous protocols (Waller et al., 2013). Briefly, about 20 mg of atrial tissues were cut on ice and homogenized in a buffer, containing 100 mM Tris, 4% SDS, and maintaining PH 7.6. Protease and phosphatase inhibitors from Meck were added in the buffer. The mixture was sonicated for 5 s at 15% amplitude on ice and paused for 5 s for 2 min of working time on a JY92-IIDN instrument (Ningbo Scientz Biotechnology Co., Ltd., China). The proteins were denatured and condensed for 5 min at 95°C circumstance afterword. The mixture was centrifuged at 14,000 g for 10 min to remove the insoluble debris and retain the supernatant for proteomic experiments. The bicinchoninic acid (BCA) assay was performed to determine the concentration of protein. All protein samples were stored at −80°C for further experiment.

Label-Free Proteomic Analysis

Protein digestion was performed by Filter-aided sample preparation (FASP) (Wisniewski et al., 2009). Briefly, protein extraction 200 μg was mixed with a reducing buffer (1 M DTT) to 100 mM DTT concentration as total, incubated for 1 h at 56°C afterword. Then the protein samples were washed twice with 200 mL of a UA buffer (pH 8.5, 8 Murea in 0.1 M Tris-HCl), adding 50 mM iodoacetamide in the tube to alkylate in the darkness for 30 min. The mixture was washed firstly with the 100 mL UA buffer and secondly with ammonium bicarbonate 50 mM for three times. All resulting solutions were centrifuged at 25°C for 12,000 g. Protein samples were digested with trypsin (Promega) at 37°C for 18 hr, with a 1:50 (w/w) concentration in 50 mM ammonium bicarbonate. Then, peptide samples were centrifuged to elute. The BCA protein procedure was used to determine peptide concentration. Peptides were desalted and dried for further procedure.

For proteomic analysis, nanoflow HPLC Easy-nLC 1000 system (Thermo Fisher Scientific) was used to separate about 1 μg peptides at 300 nL/min with a 70-min LC gradient. Proteomic analyses were conducted on an Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific). The positive ion mode at 1,900 V was set as spray voltage and the ion transfer tube at 275°C was also set. Xcalibur was used to perform data-dependent acquisition. The orbitrap mass analyzer, with a RF lens 60%, resolution of 60,000 @ m/z 200, maximum IT 50 ms and AGC target 4e5, was used to perform the MS1 full scan. HCD fragmentation, with a resolution of 15,000 @m/z 200, maximum IT 150 ms in a 3 s cycle time and AGC target 2e5, was used to generate top-speed MS2 scans. 1.2 m/z was set as isolation window. The HCD collision energy and the dynamic exclusion time were set at 30% and 60 s separately. MS2 analysis were selected by precursors charged at state 2–6.

Database for Proteomic Analysis

MaxQuant 1.6.1.0, containing 172,418 sequences (downloaded in July, 2019), was used to analyze all mass spectra. Enzyme specification was used in search. The fixed modification was

performed as carbamidomethylation of cysteine, while variable modification was carried out by N-terminal acetylation and oxidation of methionine. In the initial scan and the main search setting at 6 ppm, mass tolerances for fragment ions and precursor were set at 0.02 Da and 20 ppm respectively.

The Andromeda search engine, integrating into Maxquant, was used to search tandem MS. Seven amino acids was set as cutoff of minimum peptide length, while two amino acids was set as maximum permissible missed cleavage. Maximal FDR was set at 0.01for proteins, peptide spectral match and site. Two sequence-unique peptides was set as minimum identification.

The label-free quantitation (LFQ) was analyzed by the Andromeda search engine. The quantification results of Maxquant protein and peptide were imported for further analysis. Comparing with controls, differentially expressed proteins in patients were defined as significant change if the ratios were $\geq$2 or $\leq$0.5 ($P < 0.05$).

Protein-Protein Interaction (PPI) Network Analysis

Proteins and their interactive functions form the backbone of cellular biology. The PPIs were identified and characterized to necessarily understand the physiology and efficacy in the organism. The connective network was demonstrated for full understanding of cellular machinery. STRING 11.0 (Szklarczyk et al., 2019)[1] covering more than 5,090 organisms was used to analyze PPIs. The biological characteristics of high-throughput transcriptome data was identified by Gene ontology (GO) analysis in defining protein products. GO[2] consortium was used to identify the pathways involved. For molecular function in terms of GO analysis, $p < 0.05$ was considered significant.

RESULTS

Patient Characteristics

The baseline characteristics of AF patients, with paroxysmal or permanent AF, were shown in **Table 1**. All AF patients received transthoracic echocardiography to rule out heart failure, defined as left ventricular ejection fraction or LVEF $\geq$50%, and valvular heart disease. Color Doppler echocardiography measured left atrial diameter before fibrillation ablation. All subjects went through coronary CT angiography (CCTA) to rule out coronary heart disease.

Differentially Expression of Proteins

Three groups of specimens were detected by liquid chromatography-tandem mass spectrometry (LC-MS/MS) and analyzed by the LFQ proteomics. This method quantified 3,911 proteins. The proteome of LAAs was examined to compare the different changes in protein expressions between healthy controls and AF patients, using LFQ intensities. The differentially expressed proteins, compared in pairs between three groups, were shown in the heat map (**Figure 1A**). Totally 17 differentially expressed proteins with significant difference were identified with

[1]https://string-db.org/

[2]http://www.geneontology.org/

TABLE 1 | Baseline Characteristics of patients with paroxysmal or permanent AF.

No.	Type of AF	Gender	Age (Year)	Height (M)	Weight (Kg)	Hyper lipidemia	Smoking	Hyper tension	T2DM	LVEF (%)	CCTA	LAD (mm)	Duration of AF (Year)
1	Paroxysmal	Male	69	1.69	76	No	No	Yes	No	70	Negative	40	/
2	Paroxysmal	Male	63	1.7	64	No	No	No	No	59	Negative	46	/
3	Paroxysmal	Male	63	1.7	70	No	No	No	No	66	Negative	39	/
4	Paroxysmal	Male	69	1.73	67	No	No	Yes	No	67	Negative	46	/
5	Paroxysmal	Male	69	1.65	75	No	No	No	No	70	Negative	36	/
6	Paroxysmal	Male	61	1.76	76	No	No	Yes	Yes	60	Negative	42	/
7	Paroxysmal	Male	64	1.68	52	No	Yes	No	No	64	Negative	40	/
8	Paroxysmal	Male	64	1.81	71	No	No	Yes	Yes	63	Negative	39	/
9	Paroxysmal	Male	61	1.67	87	No	Yes	Yes	No	62	Negative	37	/
10	Paroxysmal	Male	66	1.73	82	No	No	Yes	No	63	Negative	42	/
11	Persistent	Male	63	1.76	86	No	No	Yes	No	57	Negative	46	2.5
12	Persistent	Male	63	1.78	80	No	No	No	No	68	Negative	55	3
13	Persistent	Male	64	1.7	70	No	No	No	No	67	Negative	41	4
14	Persistent	Male	64	1.64	84	No	No	Yes	No	55	Negative	48	2
15	Persistent	Male	65	1.69	73	No	No	Yes	No	69	Negative	55	3.5
16	Persistent	Male	66	1.68	66	No	No	Yes	No	64	Negative	45	4
17	Persistent	Male	67	1.75	80	No	No	Yes	Yes	59	Negative	47	2.5
18	Persistent	Male	67	1.65	73	No	Yes	Yes	No	59	Negative	47	3
19	Persistent	Male	63	1.64	61	No	No	No	No	73	Negative	49	2
20	Persistent	Male	67	1.78	90	No	No	Yes	No	70	Negative	58	2.5

AF, Atrial fibrillation; T2DM, type 2 diabetes mellitus; LVFE, left ventricular ejection fraction; LAD, Left atrial diameter.

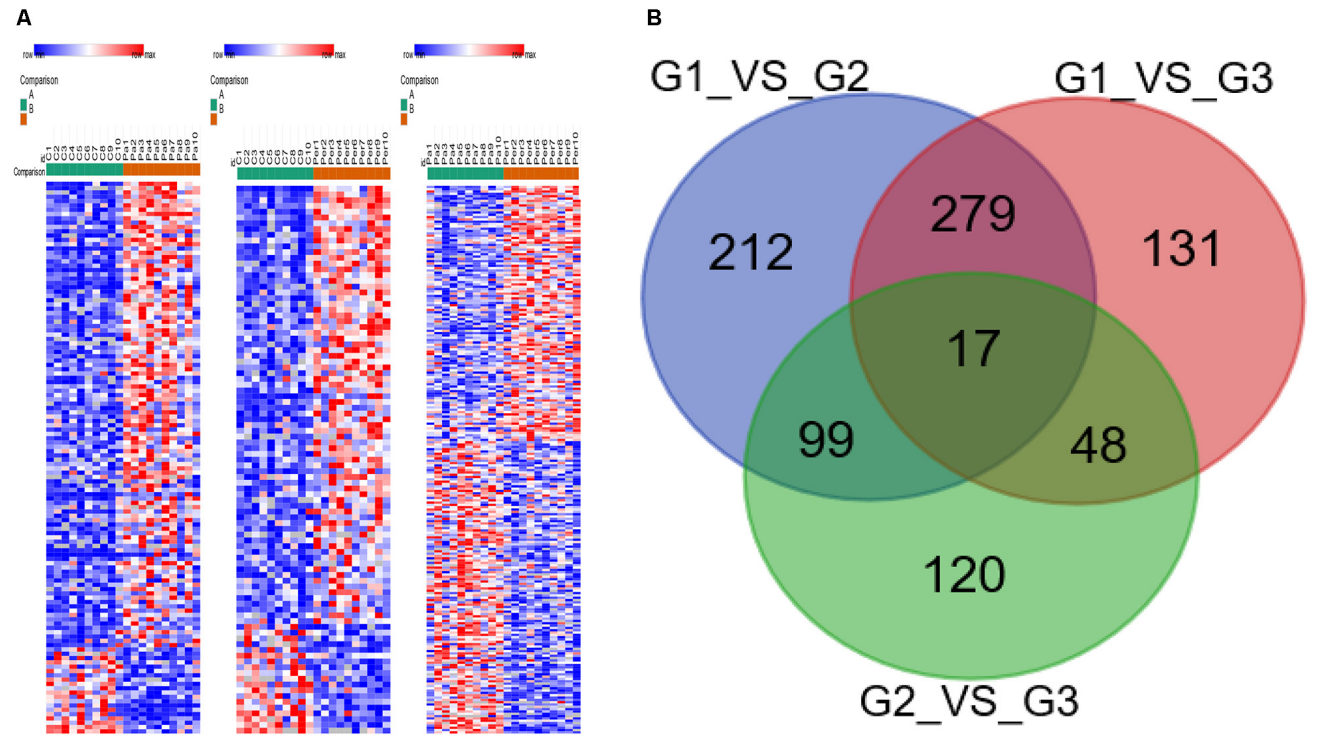

FIGURE 1 | Heatmap visualization of the differently expressed proteins identified in healthy controls, paroxysmal AF patients, and permanent AF patients. **(A)** Heatmap showing the differential protein expression profiles compared in pairs among three groups. **(B)** The differently expressed proteins identified in three groups by Venn diagram via comparing in pairs.

a gradient change among healthy controls, paroxysmal AF group, and permanent AF group via comparing in pairs (**Figure 1B**).

Functions of the Identified Proteins

The 17 proteins were divided into three major groups according to their different functions: association with cytoskeleton and protein binding, with chromatin binding, and with oxidative stress (**Table 2**).

PPI Network

The STRING analysis was used to establish a PPI network involving the 17 differentially expressed proteins (**Figure 2**). This network contained 17 nodes and 12 edges. The average node degree was 1.41. In the network analysis, the clustering coefficient (cc) was 0.5, and PPI enrichment *p*-value was 4.94e-05, which was practically negligible.

Molecular function (MF), cellular component (CC), and biological process (BP) were all analyzed by the GO consortium database. MF analysis suggested that most of the differently expressed proteins participated in structural component, protein binding, and chromatin DNA binding (**Figure 3A**). BP analysis demonstrated these proteins were mostly involved in myocyte activity, development, metabolism, post-translational protein modification, cell-substrate interaction, and apoptotic regulation (**Figure 3B**). CC analysis showed cellular structural components (**Figure 3C**).

DISCUSSION

AF is the major cause of thrombotic stroke (Vergara and Della Bella, 2014). Though AF is a major cause of mortality and morbidity and there are decades of basic and clinical studies, its fundamental mechanisms and effective treatment are still unknown. Patients with paroxysmal AF suffer less than 7 days of self-terminating episodes, but mostly progress to persistent AF, lasting more than 7 days (Kerr et al., 2005). AF lasting over 12 months is termed "long-term persistent AF" or permanent AF. AF leads to structural and electrical remodeling of the atria, while the underlying mechanisms are scarcely acquainted and remain to be revealed. Proteins are essential in cellular function and biological component, and make up to about 50% of the structural component of mammalian cells (Milo, 2013). The proteome represents the entire set of proteins expressed based on cellular genome at a specific time point, while various cellular processes and disease developments are always manifested with different protein levels (Mann et al., 2013). In brief, characterizing proteomes and specific proteins have almost been a new approach to understand the cell function mechanism and disease development.

In this study, 30 LAA samples underwent proteomic analysis, and 17 differently expressed proteins were identified between healthy subjects and patients with AF after compared in pairs between three groups, which means gradient changes with the development of AF. With time progressing, the atrial remodeling

TABLE 2 | Differently expressed proteins identified by proteomic analysis.

No.	Protein name	Gene	Accession no.	Function
1	Epididymis secretory protein Li 6	AKR1A1	V9HWI0	Cardiac necrosis
2	Lysozyme	LYZ	B2R4C5	Regulating apoptosis and K(ATP) ion channel
3	Core histone macro-H2A.1	H2AFY	O75367	Promoter-specific chromatin binding, oxidative stress
4	Dimethylargininedimethy laminohydrolase 1	DDAH1	B1AKK2	Sarcolemma of cardiomyocytes
5	Fibrinogen alpha chain	FGA	P02671	Structural molecule activity, metabolism
6	Epididymis secretory sperm binding protein Li 78p	FGB	V9HVY1	Structural molecule activity
7	Laminin subunit beta-1	LAMB1	P07942	Structural molecule activity
8	Laminin gamma 1	LAMC1	A0A024R972	Structural molecule activity
9	MYL2 protein	MYL2	Q6IB42	Structural molecule activity
10	Mutant cardiac myosin-binding protein C	MYBPC3	B6D425	Structural molecule activity
11	Myosin light chain 5	MYL5	D6RA88	Structural molecule activity
12	Myosin-10	MYH10	P35580	Actin binding
13	Heterogeneous nuclear ribonucleoprotein U	HNRNPU	Q00839	Actin binding
14	Dickkopf-related protein 3	DKK3	Q9UBP4	Cardiac hypertrophy and fibrosis
15	COP9 signalosome complex subunit 7a	COPS7A	Q9UBW8	Cytosol of cardiomyocytes, cardiac proteinopathy
16	14-3-3 protein theta	YWHAQ	P27348	Adaptor protein, regulating electric channel activity
17	Multifunctional protein ADE2	PAICS	E9PBS1	Purine biosynthesis, unclear

continuously occurs, and paroxysmal AF evolves into permanent AF (Jalife and Kaur, 2015). All these differently expressed proteins were grouped according to their functions in AF development and progression, including proteins associated with apoptosis, with cytoskeleton and protein binding, with oxidative stress, and with ion channel regulation.

Cardiomyocytes Necrosis and Apoptosis

AKR1A1 belongs to the aldo/keto reductase superfamily, consisting of more than 40 known proteins and enzymes. This superfamily is also involved in the reduction of xenobiotic and biogenic aldehydes, virtually presenting every tissue, and is known as aldehyde reductase. AKR1A1 protein levels increased in cardiac tissues with more vacuole formation and severe necrosis. These results suggest that AKR1A1 protein participates in DOX-induced cardiotoxicity (Zhou et al., 2016). LYZ levels were elevated in cardiac sarcoidosis patients with intractable heart failure and refractory arrhythmias (Odawara et al., 2019). LYZ may participate in the apoptosis in the isolated hearts of rats (Kim et al., 2010).

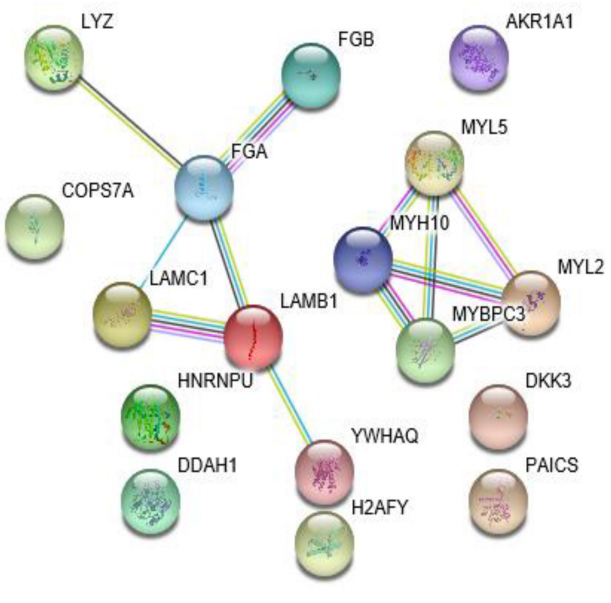

FIGURE 2 | The protein-protein interaction network for the 17 identified proteins.

Oxidative Stress

H2AFY, belonging to histone H2A family, supersedes conventional H2A histones with a subset of nucleosomes. Histones, basic nuclear proteins in eukaryotes, constitute the nucleosome structure of the chromosomal fiber. Human zinc finger RNA-binding protein is regulated in macrophage differentiation by preventing aberrant splicing of H2AFY, and controls interferon signaling. H2AFY may participate in transcriptional response to infection (Haque et al., 2018). H2AFY is related to inflammation in healthy subjects exposed to ultrafine carbon particles, and especially changes in the glucose metabolism and cardiovascular system (Huang et al., 2010).

Cytoskeleton and Protein Binding

Twelve proteins differently expressed between healthy controls and AF patients, correlating to cytoskeletal structure, were identified. DDAH1 attenuates ventricular remodeling and cardiac hypertrophy under stress conditions via regulating subcellular NO signaling (Xu et al., 2017). FGA participates in left ventricular diastolic dysfunction as a core protein in β3-adrenergic receptor knockout mice. FGA may potentially relate to the cardiac muscle contraction and actin cytoskeleton organization (Yang et al., 2019). FGB mutation can elevate the

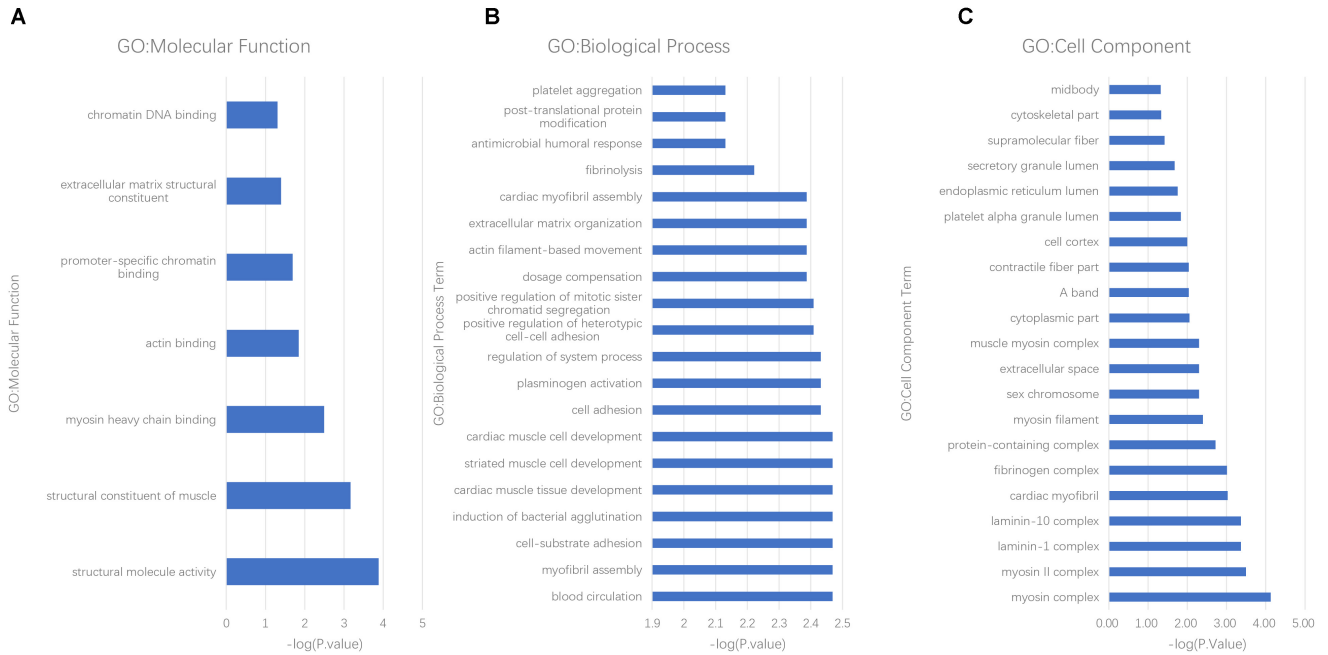

FIGURE 3 | GO analysis of differently expressed proteins in AF patients. **(A)** molecular function, **(B)** biological process, and **(C)** cell components.

level of plasma fibrinogen in AF patients, and thereby played a role in cardioembolic stroke (Hu et al., 2017).

LAMB1 and LAMC1 belonging to an extracellular matrix glycoprotein family constitute non-collagenous basement membranes. LAMB1 is moderately expressed in heart basement membranes (Cotrufo et al., 2005). Coding exons of LAMB1, LAMB4 and PIK3CG were screened in dilated cardiomyopathy (Schonberger et al., 2005). LAMC1-deficient cardiomyocytes lacked basement membranes, leading to hormonal regulation and electrical activity (Malan et al., 2009).

MYL2 triggers contraction by phosphorylation of the regulatory light chain. Mutations in this gene are related to hypertrophic cardiomyopathy. MYBPC3, a myosin- associated protein, consists in the cross-bridge-bearing zone of A bands in striated muscles, and is expressed exclusively in heart muscles. Genetic testing discovered the prevalence of MYBPC3 and MYL2in patients with hypertrophic cardiomyopathy and AF (Bongini et al., 2016). MYL5 is a component of the hexameric ATPase cellular motor protein myosin. MYH10, belonging to the superfamily of myosins, is a conventional non-muscle myosin. MYH10 is an actin-dependent motor protein, regulating cytokinesis, cell polarity, and cell motility. Mutations in MYH10 are associated with cardiac developmental defects (Takeda et al., 2003; Lo et al., 2004).

HNRNPU, belonging to a protein superfamily, binds nucleic acids and functions in the nucleus by the formation of ribonucleoprotein complexes with heterogeneous nuclear RNA. Mice lacking HNRNPU developed lethal dilated cardiomyopathy, which presented disorganized cardiomyocytes, abnormal excitation-contraction coupling activities, and impaired contractility (Ye et al., 2015). DKK3, belonging to the Dickkopf family as a secreted protein, plays an important role in heart development. DKK3 presents cardioprotective effect in pathological cardiac hypertrophy via regulating the ASK1-JNK/p38 signaling pathway (Zhang et al., 2014). COPS7A, a component of the COP9 signalosome, may participate in regulating the degradation of a bona fide misfolded and a surrogate protein in the myocardial cytosol, while COPS8 hypomorphism may impair autophagosome and exacerbate cardiac proteinopathy (Liu et al., 2016).

Besides the structural functions above, YWHAQ was suggested to amplify and prolong the activity of beta-adrenergic stimulated HERG channel by affecting IKr activity in ventricular repolarization (Choe et al., 2006). LYZ may participate in K (ATP) ion channel activity in the isolated hearts of rats, in addition to apoptosis of cardiomyocyte (Kim et al., 2010).

Besides all the 16 differently expressed proteins discussed above, the cardiac function of PAICS, which may participate in purine biosynthesis, is unclear.

AF is commonly associated with structural and electrical atrial remodeling. Structural atrial remodeling mainly includes degenerative processes, such as apoptosis and fibrosis, and alteration of cellular structural expression. Oxidative stress acts as an interdependent signaling pathway leading to cardiac fibrosis (Schotten et al., 2011).

From the above analysis, we speculate that most differently expressed proteins participate in structural remodeling and some may further develop to atrial electrical remodeling by structural remodeling or direct electrophysiologic consequence.

In this study, label-free proteomics analysis identified 17 AF-associated proteins, which were mostly correlated to structural atrial remodeling. It has been well-demonstrated for decades

that AF represents atrial myocardium hypertrophy, atrial cavity dilatation, and apoptosis of atrial cardiomyocytes, and replaces with fibrotic tissue focus or diffusion. Whether the development of arrhythmia precedes or follows the structural remodeling is unclear. Underlying this sophisticated multiple process is a complex network of molecular correlation. In the study, a comprehensive PPI network was generated from the proteomics approach to define the molecular functions participated in AF development. As mentioned above, these proteins were structural components of cardiomyocyte, and structural remodeling was presumed to play a crucial part in AF development.

The differentially expressed proteins in AF patients require further investigation to understand their exact roles in the pathological process of AF. For further functional research, candidate proteins will be selected by stringent bioinformatics analysis, which may provide vital information to investigators for future research. The joint analysis of multi-omics analysis will be carried out to reveal the regulatory mechanism of AF.

Limitations

The sample size of investigated subjects was small, owing to the difficulty in obtaining LAA samples. The healthy controls were younger than AF patients on average, which may result in inconsistency of the samples. In addition, we investigated human samples with idiopathic disease, but experiments were hardly carried out to modulate the protein levels. Although the left atrium is the key player in AF, only left atrial appendage tissues can be resected during cardiac ablation, which cannot fully represent the pathological changes of AF in the left atrium and cannot thoroughly explain the mechanism of AF.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Shanghai East Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ, CG, and CQ conceived and designed the experiments. BL, YW, and ML performed the experiments. XS, CH, and PP analyzed the data. BL, XL, and CZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Key Research and Development Program (Grant No. 2018-YFC-1312505 to YZ), the National Natural Science Foundation of China (Grant No. 81770408 to CG), Program of Outstanding Young Scientists of Tongji Hospital of Tongji University (Grant No. HBRC1803), and Project supported by Clinical Research Project of Tongji Hospital of Tongji University [Grant No. ITJ(QN)1803].

REFERENCES

Bongini, C., Ferrantini, C., Girolami, F., Coppini, R., Arretini, A., Targetti, M., et al. (2016). Impact of genotype on the occurrence of atrial fibrillation in patients with hypertrophic cardiomyopathy. *Am. J. Cardiol.* 117, 1151–1159. doi: 10.1016/j.amjcard.2015.12.058

Choe, C. U., Schulze-Bahr, E., Neu, A., Xu, J., Zhu, Z. I., Sauter, K., et al. (2006). C-terminal HERG (LQT2) mutations disrupt IKr channel regulation through 14-3-3epsilon. *Hum. Mol. Genet.* 15, 2888–2902. doi: 10.1093/hmg/ddl230

Cotrufo, M., De Santo, L., Della Corte, A., Di Meglio, F., Guerra, G., Quarto, C., et al. (2005). Basal lamina structural alterations in human asymmetric aneurismatic aorta. *Eur. J. Histochem. [EJH]* 49, 363–370. doi: 10.4081/964

Frustaci, A., Chimenti, C., Bellocci, F., Morgante, E., Russo, M. A., and Maseri, A. (1997). Histological substrate of atrial biopsies in patients with lone atrial fibrillation. *Circulation* 96, 1180–1184. doi: 10.1161/01.cir.96.4.1180

Haque, N., Ouda, R., Chen, C., Ozato, K., and Hogg, J. R. (2018). ZFR coordinates crosstalk between RNA decay and transcription in innate immunity. *Nat. Commun.* 9:1145.

Hatem, S. N., and Sanders, P. (2014). Epicardial adipose tissue and atrial fibrillation. *Cardiovasc. Res.* 102, 205–213. doi: 10.1093/cvr/cvu045

Heeringa, J., van der Kuip, D. A., Hofman, A., Kors, J. A., van Herpen, G., Stricker, B. H., et al. (2006). Prevalence, incidence and lifetime risk of atrial fibrillation: the rotterdam study. *Eur. Heart J.* 27, 949–953. doi: 10.1093/eurheartj/ehi825

Hu, X., Wang, J., Li, Y., Wu, J., Qiao, S., Xu, S., et al. (2017). The beta-fibrinogen gene 455G/A polymorphism associated with cardioembolic stroke in atrial fibrillation with low CHA2DS2-VaSc score. *Sci. Rep.* 7:17517.

Huang, Y. C., Schmitt, M., Yang, Z., Que, L. G., Stewart, J. C., Frampton, M. W., et al. (2010). Gene expression profile in circulating mononuclear cells after exposure to ultrafine carbon particles. *Inhalat. Toxicol.* 22, 835–846. doi: 10.3109/08958378.2010.486419

Jalife, J., and Kaur, K. (2015). Atrial remodeling, fibrosis, and atrial fibrillation. *Trends Cardiovas. Med.* 25, 475–484. doi: 10.1016/j.tcm.2014.12.015

Jeganathan, J., Saraf, R., Mahmood, F., Pal, A., Bhasin, M. K., Huang, T., et al. (2017). Mitochondrial dysfunction in atrial tissue of patients developing postoperative atrial fibrillation. *Ann. Thorac. Surg.* 104, 1547–1555. doi: 10.1016/j.athoracsur.2017.04.060

Kerr, C. R., Humphries, K. H., Talajic, M., Klein, G. J., Connolly, S. J., Green, M., et al. (2005). Progression to chronic atrial fibrillation after the initial diagnosis of paroxysmal atrial fibrillation: results from the Canadian registry of atrial fibrillation. *Am. Heart J.* 149, 489–496. doi: 10.1016/j.ahj.2004.09.053

Kim, Y. J., Lim, H. J., and Choi, S. U. (2010). Effect of propofol on cardiac function and gene expression after ischemic-reperfusion in isolated rat heart. *Korean J. Anesthesiol.* 58, 153–161. doi: 10.4097/kjae.2010.58.2.153

Lai, L. P., Su, M. J., Lin, J. L., Lin, F. Y., Tsai, C. H., Chen, Y. S., et al. (1999). Down-regulation of L-type calcium channel and sarcoplasmic reticular Ca(2+)-ATPase mRNA in human atrial fibrillation without significant change in the mRNA of ryanodine receptor, calsequestrin and phospholamban: an insight into the mechanism of atrial electrical remodeling. *J. Am. CollCardiol.* 33, 1231–1237. doi: 10.1016/s0735-1097(99)00008-x

Liu, J., Su, H., and Wang, X. (2016). The COP9 signalosome coerces autophagy and the ubiquitin-proteasome system to police the heart. *Autophagy* 12, 601–602. doi: 10.1080/15548627.2015.1136773

Lo, C. M., Buxton, D. B., Chua, G. C., Dembo, M., Adelstein, R. S., and Wang, Y. L. (2004). Nonmuscle myosin IIb is involved in the guidance of fibroblast migration. *Mol. Biol. Cell* 15, 982–989.

Malan, D., Reppel, M., Dobrowolski, R., Roell, W., Smyth, N., Hescheler, J., et al. (2009). Lack of laminin gamma1 in embryonic stem cell-derived cardiomyocytes causes inhomogeneous electrical spreading despite intact differentiation and function. *Stem Cells (Dayton, Ohio)* 27, 88–99. doi: 10.1634/stemcells.2008-0335

Mann, M., Kulak, N. A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. *Mol. Cell* 49, 583–590. doi: 10.1016/j.molcel.2013.01.029

Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* 35, 1050–1055. doi: 10.1002/bies.201300066

Miyasaka, Y., Barnes, M. E., Gersh, B. J., Cha, S. S., Bailey, K. R., Abhayaratna, W. P., et al. (2006). Secular trends in incidence of atrial fibrillation in olmsted county, minnesota, 1980 to 2000, and implications on the projections for future prevalence. *Circulation* 114, 119–125. doi: 10.1161/circulationaha.105.595140

Odawara, K., Inoue, T., and Hirooka, Y. (2019). Effective steroid therapy in an elderly patient with cardiac sarcoidosis and severe left ventricular dysfunction. *J. Cardiol. Cases* 19, 165–168. doi: 10.1016/j.jccase.2018.12.019

Schild, L., Bukowska, A., Gardemann, A., Polczyk, P., Keilhoff, G., Täger, M., et al. (2006). Rapid pacing of embryoid bodies impairs mitochondrial ATP synthesis by a calcium-dependent mechanism–a model of in vitro differentiated cardiomyocytes to study molecular effects of tachycardia. *Biochimica et Biophysicaacta* 1762, 608–615. doi: 10.1016/j.bbadis.2006.03.005

Schonberger, J., Kuhler, L., Martins, E., Lindner, T. H., Silva-Cardoso, J., and Zimmer, M. (2005). A novel locus for autosomal-dominant dilated cardiomyopathy maps to chromosome 7q22.3-31.1. *Hum. Genet.* 118, 451–457. doi: 10.1007/s00439-005-0064-2

Schotten, U., Verheule, S., Kirchhof, P., and Goette, A. (2011). Pathophysiological mechanisms of atrial fibrillation: a translational appraisal. *Physiol. Rev.* 91, 265–325. doi: 10.1152/physrev.00031.2009

Steinberg, J. S. (2004). Atrial fibrillation: an emerging epidemic? *Heart (British Cardiac Society)* 90, 239–240. doi: 10.1136/hrt.2003.014720

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Takeda, K., Kishi, H., Ma, X., Yu, Z. X., and Adelstein, R. S. (2003). Ablation and mutation of nonmuscle myosin heavy chain II-B results in a defect in cardiac myocyte cytokinesis. *Circ. Res.* 93, 330–337. doi: 10.1161/01.res.0000089256.00309.cb

Vergara, P., and Della Bella, P. (2014). Management of atrial fibrillation. *F1000prime Rep.* 6:22.

Wakili, R., Voigt, N., Kaab, S., Dobrev, D., and Nattel, S. (2011). Recent advances in the molecular pathophysiology of atrial fibrillation. *J. Clin. Invest.* 121, 2955–2968. doi: 10.1172/jci46315

Waller, A. P., George, M., Kalyanasundaram, A., Kang, C., Periasamy, M., Hu, K., et al. (2013). GLUT12 functions as a basal and insulin-independent glucose transporter in the heart. *Biochimica et Biophysicaacta* 1832, 121–127. doi: 10.1016/j.bbadis.2012.09.013

Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat. Methods* 6, 359–362. doi: 10.1038/nmeth.1322

Xu, X., Zhang, P., Kwak, D., Fassett, J., Yue, W., Atzler, D., et al. (2017). Cardiomyocyte dimethylarginine dimethylaminohydrolase-1 (DDAH1) plays an important role in attenuating ventricular hypertrophy and dysfunction. *Basic Res. Cardiol.* 112:55.

Yang, W., Wei, X., Su, X., Shen, Y., Jin, W., and Fang, Y. (2019). Depletion of beta3-adrenergic receptor induces left ventricular diastolic dysfunction via potential regulation of energy metabolism and cardiac contraction. *Gene* 697, 1–10. doi: 10.1016/j.gene.2019.02.038

Ye, J., Beetz, N., O'Keeffe, S., Tapia, J. C., Macpherson, L., Chen, W. V., et al. (2015). hnRNP U protein is required for normal pre-mRNA splicing and postnatal heart development and function. *Proc. Natl. Acad. Sci. U S A.* 112, E3020–E3029.

Zhang, Y., Liu, Y., Zhu, X. H., Zhang, X. D., Jiang, D. S., Bian, Z. Y., et al. (2014). Dickkopf-3 attenuates pressure overload-induced cardiac remodelling. *Cardiovasc. Res.* 102, 35–45. doi: 10.1093/cvr/cvu004

Zhou, Z. Y., Wan, L. L., Yang, Q. J., Han, Y. L., Li, D., Lu, J., et al. (2016). Nilotinib reverses ABCB1/P-glycoprotein-mediated multidrug resistance but increases cardiotoxicity of doxorubicin in a MDR xenograft model. *Toxicol. Lett.* 259, 124–132. doi: 10.1016/j.toxlet.2016.07.710

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

frontiers in Genetics

METHODS
published: 25 September 2020
doi: 10.3389/fgene.2020.591461

RWRNET: A Gene Regulatory Network Inference Algorithm Using Random Walk With Restart

Wei Liu[1,2], Xingen Sun[1], Li Peng[3], Lili Zhou[1], Hui Lin[1] and Yi Jiang[1]*

[1] School of Computer Science, Xiangtan University, Xiangtan, China, [2] Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan, China, [3] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China

Inferring gene regulatory networks from expression data is essential in identifying complex regulatory relationships among genes and revealing the mechanism of certain diseases. Various computation methods have been developed for inferring gene regulatory networks. However, these methods focus on the local topology of the network rather than on the global topology. From network optimisation standpoint, emphasising the global topology of the network also reduces redundant regulatory relationships. In this study, we propose a novel network inference algorithm using Random Walk with Restart (RWRNET) that combines local and global topology relationships. The method first captures the local topology through three elements of random walk and then combines the local topology with the global topology by Random Walk with Restart. The Markov Blanket discovery algorithm is then used to deal with isolated genes. The proposed method is compared with several state-of-the-art methods on the basis of six benchmark datasets. Experimental results demonstrated the effectiveness of the proposed method.

Keywords: gene regulatory networks, random walk with restart, local topology, global topology, Markov Blanket discovery algorithm

OPEN ACCESS

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

Reviewed by:
Padhmanand Sudhakar,
Earlham Institute, United Kingdom
Xiangzheng Fu,
Hunan University, China

****Correspondence:***
Wei Liu
lw2001184@163.com

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

Received: *04 August 2020*
Accepted: *02 September 2020*
Published: *25 September 2020*

Citation:
Liu W, Sun X, Peng L, Zhou L, Lin H and Jiang Y (2020) RWRNET: A Gene Regulatory Network Inference Algorithm Using Random Walk With Restart. Front. Genet. 11:591461. doi: 10.3389/fgene.2020.591461

INTRODUCTION

Inferring accurate gene regulatory networks (GRNs) is an exciting but difficult topic in the field of bioinformatics. Inferring accurate GRNs is not only helpful to understanding complex regulatory relationships between genes in cells but also to understanding relationships between genes and diseases (Lv and Bao, 2009; Altay and Emmert-Streib, 2010; Tang et al., 2015). With the development of high-throughput technologies, huge gene expression data have been produced from which researchers can infer GRNs (Maetschke et al., 2014; Liu, 2015).

Numerous network inference methods for inferring accurate GRNs have been developed. These methods can be classified into two categories: model-based and similarity-based methods. Model-based methods, which mainly include Boolean network model, differential equation model and Bayesian network model, usually infer GRNs through a computational model. The Boolean network model is a simple discrete model that contributes to understanding various states of cells, such as proliferation, differentiation and apoptosis (Huang, 1999; Lim et al., 2016; Zhou et al., 2016). However, the Boolean network model cannot be applied in networks with complex regulatory relationships. The differential equation model is a continuous network model that can accurately

describe the dynamic characteristics of GRNs. The expression level of genes in differential equation is determined by related genes and regulatory equations, thus allowing the underlying phenomena of organisms to be accurately described (Alter et al., 2000; Cantone et al., 2009; Honkela et al., 2010; Huppenkothen et al., 2017). The Bayesian network model is a popular graphical model of probability. In this model, the dependencies between genes are described by a directed acyclic graph. The Bayesian network model is superior to other models in terms of dealing with noise and prior knowledge, but it has high computational complexity (Tan et al., 2011; Betliński and Ślęzak, 2012; Shi et al., 2016).

Similarity-based methods, which primarily include correlation-based and information theory-based methods, identify regulatory relationships by measuring the dependencies between genes (Li et al., 2011). In correlation-based methods, the dependencies are determined by the degree of co-expression. Typical measurement methods include Pearson's correlation coefficient, Euclidean distance and partial correlation coefficient (de la Fuente et al., 2004; Saito et al., 2011; Fukushima, 2013; Ruyssinck et al., 2014; Mohamed Salleh et al., 2015; Ghosh and Barman, 2016). However, these measurement methods cannot identify complex dependencies, such as non-linear dependencies (Wang and Huang, 2014). Information theory-based methods can capture complex non-linear regulatory relationships (Brunel et al., 2010; Mousavian et al., 2016). Mutual information (MI) is first used in information theory to measure the similarity between signals and later used in the field of biology to measure regulatory relationships between genes. Classical methods include Relevance Network (RN), Minimum Redundancy Network (MRNET), Path Consistency Algorithm based on Conditional Mutual Information (PCA-CMI) and Redundancy Reduction in the MRNET algorithm (RRMRNET). RN (Butte and Kohane, 2000; Kuzmanovski et al., 2018) is one of the earliest methods that used MI to measure relationships. MRNET (Meyer et al., 2008) is a feature selection method. In MRNET, a feature selection strategy is adopted in selecting regulatory relationships. Although non-linear regulatory relationships can be measured by MI, it cannot distinguish indirect regulatory relationships (Margolin et al., 2006). To overcome this limitation, Zhang et al. (2012) proposed PCA-CMI, in which MI is replaced by conditional mutual information (CMI). However, CMI tends to underestimate the relationship between genes, so Zhang et al. (2015) proposed conditional mutual inclusive information (CMI2) to solve the problem of underestimation of CMI. To improve accuracy, Liu et al. (2017) proposed RRMRNET on the basis of MRNET, in which two strategies are implemented in eliminating redundant regulatory relationships.

In addition, several machine learning-based methods, such as tree-based ensemble regression and neural network-based inference methods, have been applied in this field (Huynh-Thu et al., 2010; Huynh-Thu and Sanguinetti, 2015; Petralia et al., 2015; Raza and Alam, 2016). Researchers have also noticed that several regulatory relationships do not occur in every cell. Thus, the GRN should be defined in specific cells and situations (Moignard et al., 2015; Moris et al., 2016). Therefore, network inference methods based on single-cell expression data have attracted people's interest, which has led to the development of computational and statistical methods that are aimed at discovering new insights into cell state transitions (Bendall et al., 2014; Trapnell et al., 2014; Pina et al., 2015; Rue and Martinez Arias, 2015). The use of single-cell expression data to infer networks has many advantages. With the development of single-cell technology, the amount of data we can use will increase, which can effectively alleviate the defects of high-dimensional and low-sample gene expression data (Macosko et al., 2015). However, obtaining the time-series data of single cells is currently impossible. Notably, these methods infer the regulatory relationship based on the similarity between the transcriptional states of genes and usually provide strong assumptions, which are often unconvincing. However, several methods can still be used for network reasoning using single-cell expression data (Bendall et al., 2014; Trapnell et al., 2014; Haghverdi et al., 2016; Moris et al., 2016; Reid and Wernisch, 2016).

Although these aforementioned methods have extensively promoted GRN research, they still have certain shortcomings. For example, model-based methods usually have high computational complexity. Most similarity-based methods consider relationships between only two and not all genes at a time. Moreover, these methods usually focus on the surrounding information rather than on the global topology of network, thus resulting in numerous redundant regulatory relationships. Therefore, the present study mainly concentrates on inferring GRNs by combining local and global topologies.

Random Walk with Restart (RWR) is an improvement of the Random Walk (RW). RWR is widely used in the field of bioinformatics because it can capture multivariate relationships between nodes and explores the global topology of networks (Rosvall and Bergstrom, 2008; Athanasiadis et al., 2017; Peng et al., 2018; Valdeolivas et al., 2019). Chen et al. (2012) used RWR to determine associations between diseases and miRNAs. Sun et al. (2014) verified the robustness of RWR for parameter selection. Luo et al. (2016) proposed a new computational approach, MBiRW, that uses a combination of similarity measures and a double random Walk (BiRW) algorithm to identify potential new indications for a particular drug. Yu et al. (2017) provided a comprehensive framework for predicting new HCC drugs based on multi-source random walk.

To address the limitations in gene network inference, we propose a novel network inference algorithm using RWR (RWRNET). The restart probability, initial probability vector and roaming network in RWR is first improved to apply it in network inference. Second, the improved RWR is used in inferring network structure. Finally, the Markov–Blanket discovery algorithm IPC-MB is used to optimise the network structure to obtain the final gene network. The main contributions of this study are described as follows:

(1) We improve the three key elements of RWR. First, the proposed method obtains the restart probability and initial probability vector according to node connectivity and functional modularity and then captures the local

topology structure of the network. Second, a roaming network construction method is proposed for reducing the complexity of regulatory relationships among genes.
(2) We use a Markov–Blanket discovery algorithm (IPC-MB) to deal with isolated genes in the network that are generated by the RWR process.
(3) Extensive experiments are conducted to evaluate the performance of RWRNET. Experimental results confirmed that RWRNET is an effective network inference method.

THEORY

In this section, we review the concepts of (conditional) mutual information, RWR and Markov–Blanket that are related to the proposed method.

(Conditional) Mutual Information

Mutual information is an information measurement in information theory. MI can be regarded as the information shared by two random variables or the reduction of uncertainty due to a known random variable. The MI between random variable X and Y is defined as follows:

$$MI(X, Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \quad (1)$$

where $p(x, y)$ is the joint distribution of X and Y; while $p(X)$ and $p(Y)$ represent the marginal probability functions of X and Y, respectively.

Conditional mutual information (CMI) is a variant of MI. CMI represents the information shared between variable X and variable Y under the influence of variable Z. The CMI between variable X and variable Y is defined as follows:

$$CMI(X, Y|Z) = \sum_{x \in X, y \in Y, z \in Z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)\, p(y|z)} \quad (2)$$

where $p(x, y, z)$ is the joint distribution of X, Y and Z, $p(x|z)$ is the marginal distribution of variable X when variable Z occurs; and $p(x, y, z)$ is the joint distribution of X, Y under the influence of variable Z .

Random Walk With Restart

Random Walk with Restart is an improvement of RW. RWR contains a parameter α as the restart probability, and $1 - \alpha$ represents the probability of a walker moves from a node to an adjacent node. The RWR of graph can be defined by assigning a transition probability to each edge. In this way, a walker can jump from one node to another, and the sequence of nodes visited by the walker is called RWR. Let $p_{t+1}(j)$ denote the probability that walker locates at j-th node when it come to a stable state, then the formula is:

$$p_{t+1} = (1 - \alpha)\, W p_t + \alpha p_0 \quad (3)$$

where $W = [a_{ij}]_{N \times N}$ is the transition probability matrix, a_{ij} is the transition probability from the i-th node to the j-th node; and p_0 represents the initial probability vector of $N \times 1$, in which the i-th element is 1 and the others are zero. N is the number of nodes in the graph.

Markov–Blanket

This section introduces Markov-Blanket (MB). In the complete set U of random variables, for a given variable $X \in U$ and variable set $MB \in U$ ($X \notin MB$), the following exists:

$$X \perp \{U - MB - \{X\}\} | MB \quad (4)$$

that is, if the variable X and the set $\{\text{U-MB} - \{X\}\}$ are independent of each other under MB, then the minimum variable set MB that can meet the above conditions is called MB of X.

METHODS

In this study, we propose an effective network inference method (i.e., RWRNET). To apply RWR in GRNs, we improve its three key elements, namely, restart probability, initial probability vector and roaming network. Then the RWR is used to infer network structure. Finally, we use IPC-MB to optimise the network structure. **Figure 1** presents the flowchart of RWRNET. Specific details are discussed in the following sections. At the same time, we have uploaded the source code (MATLAB format) to the Internet, and readers can view it by visiting the link[1].

Improvements of RWR

This section mainly introduces specific improvements to the three elements of RWR (i.e., restart probability, initial probability vector and roaming network) when RWR is applied in GRNs. First, the restart probability and initial probability vector are determined according to node connectivity and functional modularity to capture the network topology. Second, a roaming network is constructed using the asymmetric MI ranking strategy to reduce the complexity of regulatory relationships among genes. Specific details are described as follows.

Calculation of Restart Probability

Different nodes in a network have different connectivity, which reflects network topology structure to some extent. Laplacian Eigenmaps is an effective way to obtain network topology, because it can map high-dimensional data to low-dimensional data and ensure their similarity to the original data as much as possible. Applying discrete Laplacian Eigenmaps to the graph network can obtain the Laplacian matrix L. And the pseudo inverse L^+ of L is a valid kernel that can provides a similarity measure between nodes. On the basis of L^+, the average commute time ACT (g_i,g_j) between gene g_i and gene g_j can be then defined as

$$\text{ACT}(g_i, g_j) = L^+(g_i, g_i) + L^+(g_j, g_j) - 2L^+(g_i, g_j) \quad (5)$$

$$L = D^{-\frac{1}{2}} (D - W) D^{-\frac{1}{2}} \quad (6)$$

[1]https://github.com/Dam-1517/RWRNET

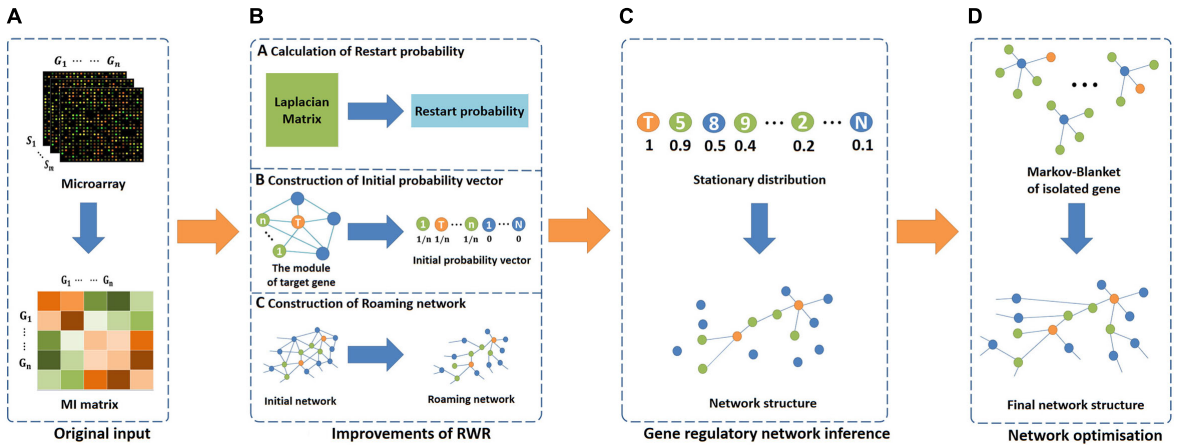

FIGURE 1 | Flowchart of RWRNET. This flowchart consists of four parts, namely, the original input stage, algorithm improvement stage, network inference stage and network optimisation stage. **(A)** In the original input stage, the MI matrix was obtained through gene microarrays data. **(B)** In the algorithm improvement stage, three key elements of RWR were improved. The first was the restart probability, which was calculated by the pseudo inverse of the Laplacian matrix. Then, the initial probability vector was improved next. Different genes have various initial probability vectors, depending on the functional module. In the figure, the centres of modules are represented by orange nodes, modular genes are indicated by green nodes, and others are denoted by blue nodes. Finally, the roaming network. In this paper, the network represented by the mutual information matrix was considered a fully connected network, and the roaming network was obtained by adjusting this network. **(C)** In the network inference stage, RWR was executed to obtain a stationary distribution, and the gene regulatory network was inferred based on the stationary distribution. **(D)** In the network optimisation stage, the MB for each isolated gene was discovered to establish a relationship between the isolated genes and others.

where W is the adjacency matrix of graph, which is MI matrix in this paper; and $D = diag(a_{\mathrm{i}}.)$ with $d_{\mathrm{ii}} = [D]_{\mathrm{ii}} = a_{\mathrm{i}}. = \sum_{\mathrm{j}=1}^{n} a_{\mathrm{ij}}$; $ACT(g_{\mathrm{i}}, g_{\mathrm{j}})$ describes the average number of steps that particles moves from g_{i} to g_{j} and then back to g_{i} .

The average commute time increases when the number of paths connecting the two points increases and when the length of paths decreases. According to this idea, the average commute frequency $ACF(g_{\mathrm{i}}, g_{\mathrm{j}})$ and restart probability α can be defined as follows:

$$\mathrm{ACF}(g_i, g_j) = \begin{cases} 1 & , g_i = g_j \\ \frac{1}{ACT(g_i, g_j)} & , g_i \neq g_j \end{cases} \quad (7)$$

$$\alpha = \frac{1}{N^2} \sum_{g_i \in G} \sum_{g_j \in \mathrm{G}} \mathrm{ACF}(\mathrm{g_i}, \mathrm{g_j}) \quad (8)$$

where $G = \{g_1, g_2, \cdots, g_{\mathrm{N}}\}$ is the set of genes, and N denotes the number of genes.

Construction of Initial Probability Vector

Gene regulatory networks is scale-free network in which only a few genes have regulatory relationships with numerous genes. These genes have substantial expression levels and form their own modules according to different functions. Genes in the same module are closely related not only to each other but also to genes in other modules. In addition, although RWR can obtain the global information of the network, taking only these genes as starting nodes is insufficient. Therefore, the functional module of these genes is used as starting nodes to obtain sufficient information in this paper.

In this study, the sum of MI between one gene and another is used to represent its expression level. The genes whose expression level is higher than the average expression level are selected as the centre of functional module. At the same time, due to the influence of noise on gene expression data, genes with low expression levels less than $MEAN\,(EL) - STD(EL)$ are also selected to fully consider the surrounding information. These genes then put together to form a set C that includes not only the genes with high expression levels but also genes with abnormally low expression levels. The expression level EL and the set C are defined as follows:

$$\mathrm{EL}\left(g_i\right) = \sum_{g_j \in \{G - g_i\}} MI(g_i, g_j) \quad (9)$$

$$C = \left\{ g_i | EL\left(g_i\right) > MEAN\,(EL) \quad or \quad EL\left(g_i\right) \right\} \left\{ < MEAN(EL) - STD\,(EL) \quad , g_i \in G \right\} \quad (10)$$

where $MEAN(EL)$ is the average expression level, and $STD\,(EL)$ represents the standard deviation.

Finally, for each gene g_{i} in the set C, the top $\log n$ genes with the largest $MI(g_{\mathrm{i}}, g_{\mathrm{j}})$ are selected as the functional module $module_{g_{\mathrm{i}}}$. Based on these modules, the initial probability vector p_0 can be constructed according to the following strategy: for each gene g_{i} in G, if g_{i} is an element of C, then the elements of g_{i}-corresponding and module-corresponding have a value of non-zero, with their sum equals to 1. Otherwise, only g_{i} -corresponding is 1, whereas the others are zero.

Construction of Roaming Network

Although GRN is sparse, the regulatory relationships among genes are extremely complicated. Therefore,

several classical methods have introduced redundant regulatory relationships when inferring network structure. To reduce the complexity of regulatory relationships while maintaining the local topology, we propose a novel method for constructing the roaming network. The basic idea is to use the asymmetry of MI ranking to adjust the relationships between genes, thereby weakening those that are not closely related. The roaming network (i.e., transition probability matrix) W can be constructed using the following formulas:

$$W(g_i, g_j) = Rank_{g_i g_j} * MI(g_i, g_j) \tag{11}$$

$$Rank_{g_i g_j} = \begin{cases} 1\ ,\ if\ MI(g_i, g_j) \geq \overline{MI_{g_i}} \\ 1 - \frac{R_{g_i g_j}}{N}\ ,\ if\ MI(g_i, g_j) < \overline{MI_{g_i}}\ and\ MI(g_j, g_i) \geq \overline{MI_{g_j}} \\ 0.1\ ,\ if\ MI(g_i, g_j) < \overline{MI_{g_i}}\ and\ MI(g_j, g_i) < \overline{MI_{g_j}} \end{cases} \tag{12}$$

where $Rank_{g_i g_j}$ is the attenuation factor, which represents the attenuation degree of regulatory relationships; $R_{g_i g_j}$ is the MI ranking of g_j among the genes connected with g_i . $\overline{MI_{g_i}}$ represents the average MI between gene g_i and others. As depicted by the formulas, the regulatory relationship between g_i and g_j is determined by $R_{g_i g_j}$ when $MI(g_i, g_j) < \overline{MI_{g_i}}$ and $MI(g_j, g_i) \geq \overline{MI_{g_j}}$. The lower the ranking, the higher the attenuation degree will be. If $MI(g_i, g_j) \geq \overline{MI_{g_i}}$, the regulatory relationship between g_i and g_j will not be weakened; if $MI(g_i, g_j) < \overline{MI_{g_i}}$ and $MI(g_j, g_i) < \overline{MI_{g_j}}$, the relationship between them will be weakened by 0.1 times.

Gene Regulatory Network Inference Based on RWR

This section covers network inference on RWR. Specific details are discussed below.

The first stage involves initialisation of regulatory relationships. In this stage, we obtain the MI matrix $(MI_{ij})_{N \times N}$ from the gene microarrays expression data that contain N genes and M samples. This matrix is then taken as the input of the method.

The second stage entails implementation of RWR. Given the restart probability, normalised transition probability matrix and appropriate initial probability vector, RWR can be performed on the roaming network for each gene g_i to obtain the stationary distribution p_{t+1} . Considering time efficiency and accuracy, when $|p_{t+1} - p_t| < 10^{-6}$, p_{t+1} is stable, $p_{t+1}^{(g_i)}(g_j)$ represents the probability that g_i finds g_j .

The final stage concerns GRNs inference. In this stage, stationary distribution is multiplied to transition probability to obtain the final score $MIP^{(g_i)}(g_j)$:

$$MIP^{(g_i)}(g_j) = p_{t+1}^{(g_i)}(g_j) * W(g_i, g_j) \tag{13}$$

Based on the final score, GRNs can be inferred according to the following formula:

$$NETWORK(g_i, g_j) = \begin{cases} 1\ ,\ if\ \ MIP^{(g_i)}(g_j) > Threshold\ (g_i) \\ 0\ ,\ otherwise \end{cases} \tag{14}$$

$$Threshold\ (g_i) = \frac{3\alpha}{4} \sum_{g_j \in G} MIP^{(g_i)}(g_j) \tag{15}$$

where $NETWORK(g_i, g_j)$ represents the regulatory relationship between g_i and g_j; $Threshold(g_i)$ is an adaptive threshold for g_i . In this paper, the threshold of each gene is automatically determined by its prediction results based on the following reasons. The prediction results of each gene were obtained by executing the RWR with different initial probability vectors, and different amounts of information were generated by each execution of RWR. Therefore, the prediction results obtained from different genes cannot be compared and cannot be processed with a fixed threshold. To this end, Eq. 15 was designed to screen the regulatory relationships for each gene. $\sum MIP^{(g_i)}(g_j)$ was selected as the major component of formula to simultaneously consider the effect of the predicted relationship between all genes and the target gene on the results. However, the regulatory relationship cannot be screened out if only one major component is used. Therefore, we added a factor of $3\alpha/4$, which represents the information occupancy of the target gene. Equation 15 indicates that only when the predictive relationship between a gene and the target gene exceeds the total information that the target gene holds can the real regulatory relationship between them be considered.

Network Optimisation Based on IPC-MB

Given that each gene in GRNs has a unique role, no gene should be isolated. However, RWR cannot handle isolated nodes. Therefore, the isolated nodes are processed by a Markov–Blanket discovery algorithm (IPC-MB) to optimise the network structure. IPC-MB is a classical feature selection algorithm (Fu and Desmarais, 2008). Its main idea is involves eliminating redundant and irrelevant regulatory relationships according to conditional independence to find genes that have direct regulatory relationships with the target gene, CMI stands for the conditional independence in this article. The basic idea is look for a parent–child set (PC) and a spouse set. These sets are then merged to obtain the Markov–Blanket (MB) of the target gene. However, since the genes in the spouse set are actually redundant, we will not use all of Markov-Blanket, but only use the parent-child set (PC). Finally, on the basis of PC, the regulatory relationships between isolated genes and genes in the PC are established to obtain optimised GRNs.

To describe the proposed method comprehensively, **Table 1** summarises the complete RWRNET. As shown in the table, Lines 2–10 of the pseudo code are the improvements of RWR, including calculating the restart probability, construction of a

TABLE 1 | Gene Regulatory Network Inference Algorithm Using Random Walk with Restart.

Algorithm: RWRNET

Input: Gene microarrays data $G = \{g_1, \cdots, g_N\}$

Output: A gene regulatory network

1: Construct a MI matrix *MI* according to Eq. 1;

2: Calculate restart probability α using Eq. 8;

3: Construct transition probability matrix *W* using Eq. 11;

4: Calculate gene expression level *EL* (g_i) for each gene using Eq. 9;

5: Select centres of functional module and put them into set *C* according to Eq. 10;

6: Construct functional modules:
$module_{g_1} = \{g_1\}$, $module_{g_2} = \{g_2\}$, $\cdots$, $module_{g_N} = \{g_N\}$;

7: For each gene $g_i \in C$ do

8: Rank the genes g_j in $\{G - g_i\}$ according to $MI(g_i, g_j)$ in descending order to form ranking list *MIL*;

9: $module_{g_i} \leftarrow$ the top $\log N$ genes in *MIL*;

10: End For

11: For each gene $g_i \in G$ do

12: Construct initial probability vector $p_0^{(g_i)}$ according to $module_{g_i}$;

13: $p_{t+1}^{(g_i)} = RWR(\alpha, W, p_0^{(g_i)})$;

14: Calculate final score $MIP^{(g_i)}$ according to Eq. 13;

15: End For

16: Infer network using Eq. 14;

17: Process isolated genes based on IPC-MB;

18: Return the optimised gene regulatory network.

roaming network and search for functional modules to construct the initial probability vector. In Lines 11–16, RWR was used to infer the initial network structure. The 17th line was used in IPC-MB to optimise the network structure.

EXPERIMENT

In this section, we introduce the datasets and evaluation metrics used to evaluate RWRNET performance. In the experiment, the performance of RWRNET was compared with that of different methods, namely, CLR, ARACNE, MRNET, MIDER, MI3, MRMSn, PCA-CMI, and RRMRNET, based on information theory. Among these methods, MI3 and MIDER can infer regulatory directions. However, RWRNET does not infer regulatory directions. Hence, we ignored the regulatory direction during the comparisons.

Datasets

During the experiment, the proposed and other methods were tested and compared in terms of six datasets. The test datasets were divided into simulated and real data, which included the reaction chain data, DREAM3 yeast gene expression data and SOS data. The reaction chain data were downloaded from the KEGG database[2]. The reaction chain data were time-series data. The DREAM3 yeast gene expression data were downloaded from the DREAM3 challenge project[3]. The DREAM3 challenge project provided three types of data; the null-mutant gene knockout data were selected in this article. The SOS data were downloaded from E. coli database[4]. The SOS data were interference data, that is, the measurement data obtained through a series of transcription interference. **Table 2** provides a summary of the details of the above six datasets.

The reaction chain with four species datasets comes from a small linear chain of chemical reactions (Samoilov, 1997). The dataset contained four variables, each of which contained 100 samples. The real network of the reaction chain included of four nodes and three edges.

The reaction chain with eight species datasets comes from a small linear chain of chemical reactions (Samoilov et al., 2001). The dataset contained eight variables, each of which contained 250 samples. The real network of the reaction chain included of eight nodes and seven edges.

The Dream3-10 gene dataset is from a yeast network in DREAM3 (Marbach et al., 2010). The dataset contained 10 genes, each of which contained 10 samples. The corresponding real network structure included of 10 nodes and 10 edges.

The Dream3-50 gene dataset is from a yeast network in DREAM3 (Marbach et al., 2010). The dataset contained 50 genes, each of which contained 50 samples. The corresponding real network structure included 50 nodes and 50 edges.

The Dream3-100 gene dataset is also from a yeast network in DREAM3 (Margolin et al., 2006). The dataset contained 100 genes, each of which contained 100 samples. The corresponding real network structure included 100 nodes and 166 edges.

The SOS dataset is from an SOS network (Ronen et al., 2002). The dataset contained nine genes, each of which contained nine samples. The corresponding real network structure included nine nodes and 24 edges.

[2]https://www.genome.jp/kegg/

[3]http://dreamchallenges.org/project-list/

[4]http://regulondb.ccg.unam.mx/index.jsp/

TABLE 2 | Descriptions of the datasets in our experiments.

Datasets	Variables	Samples	Type	Network nodes	Network edges
Reaction chain with four species	4	100	Simulated	4	3
Reaction chain with eight species	8	250	Simulated	8	7
DREAM3-10 genes	10	10	Simulated	10	10
DREAM3-50 genes	50	50	Simulated	50	77
DREAM3-100 genes	100	100	Simulated	100	166
SOS	9	9	Real	9	24

Evaluation Metrics

To verify the effectiveness of the proposed method, we used four evaluation metrics: true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV) and accuracy rate (ACC). TP, FP, TN and FN denote the number of true positives, false positives, true negatives and false negatives, respectively. These four evaluation metrics are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

$$PPV = \frac{TP}{TP + FP} \tag{18}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{19}$$

RESULTS

Results of the Chain Structure Network

To verify whether the proposed method has an effect on special networks, such as chain structure network, we selected the expression data of chain structure network with sizes of four and eight as the test datasets.

First, we tested the proposed method on the chain structure network with a size of four. **Table 3** shows the performance of RWRNET and other methods in this dataset. Like most methods, RWRNET achieved perfect performance (PPV = 1, ACC = 1) in this dataset.

To verify further the effectiveness of the proposed method, we selected a chain structure network with a size of eight for testing. **Table 4** shows the performance of all methods. RWRNET, CLR and ARACNE predicted six correct regulatory relationships (TP = 6), only one missing regulatory relationship and one redundant regulatory relationship (FP = 1). Compared with the performance of the other methods, RWRNET predicted the most regulatory relationships, and its FPR performance was only worse than that of MIDER. However, MIDER achieved FPR = 0 at the cost of TPR. Hence, our proposed method still offered great advantages. To intuitively explain the advantages of RWRNET, we show the network structure inferred by all methods (**Figure 2**). The first network in the figure is the true network structure, the second network is the network structure inferred by RWRNET, and the other networks are the network structures inferred by comparison method. The figure shows that the network structure inferred by CLR, RRMRNET, ARACNE, and MIDER was the closest to the true network, whereas the results obtained by MRNET, PCA-CMI, and MI3 contained considerable redundant control relationships. RWRNET missed X1–X8 and incorrectly linked X8 to other genes, similar to the other methods. Only MI3, MIDER and PCA-CMI were able to predict X1–X8. However, MIDER missed X3–X4 and X5–X6, MI3 and PCA-CMI introduced excessive redundant regulatory relationships. In summary, the proposed method showed excellent performance. Finally, by combining the performance of RWRNET in these two datasets, we learned that RWRNET is suitable for special networks.

TABLE 3 | Comparison of the different methods' performances in the reaction chain with four species dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	3	0	1	0	1	1
ARACNE	3	0	1	0	1	1
MRNET	3	1	1	0.33	0.75	0.833
MI3	2	3	0.667	1	0.4	0.333
MIDER	3	0	1	0	1	1
MRMSn	3	0	1	0	1	1
RRMRNET	3	0	1	0	1	1
PCA-CMI	3	1	1	0.333	0.75	0.833
RWRNET	3	0	1	0	1	1

TABLE 4 | Comparison of the different methods' performances in the reaction chain with eight species dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	6	1	0.857	0.048	0.857	0.929
ARACNE	6	1	0.857	0.048	0.857	0.929
MRNET	6	9	0.857	0.429	0.4	0.643
MI3	2	11	0.286	0.524	0.154	0.429
MIDER	5	0	0.714	0	1	0.929
MRMSn	–	–	–	–	–	–
RRMRNET	6	2	0.857	0.095	0.75	0.893
PCA-CMI	6	16	0.857	0.762	0.273	0.393
RWRNET	6	1	0.857	0.048	0.857	0.929

Results of the DREAM3 Challenge Network

To demonstrate that the proposed method can be used to infer GRNs from simulated dataset, we tested it in DREAM3. The DREAM3 Challenge Network is a version of the DREAM project that provides various gene expression datasets and corresponding golden networks to evaluate the performance of the inferred model. The gene expression dataset provided by DREAM3 is a simulation dataset. We used yeast gene expression data with a size of 10, 50, and 100 as the test datasets.

First, we tested the proposed method in the yeast gene expression dataset with a size of 10. A comparative analysis of different methods is summarised in **Table 5**. RRMRNET had the best performance (PPV = 1, ACC = 1). MRMSn and PCA-CMI identified nine correct regulatory relationships (TP = 9), whereas RWRNET identified eight regulatory relationships only (TP = 8) and introduced a redundant regulatory relationship (FP = 1). To analyse visually the gap between RWRNET and other methods, we showed the network structure they inferred (**Figure 3**). The figure contains nine networks. The first network is a standard network, and the one on the right of the standard network is the network inferred by RWRNET. Like most other methods, RWRNET missed G4–G9 and predicted G2–G9 incorrectly

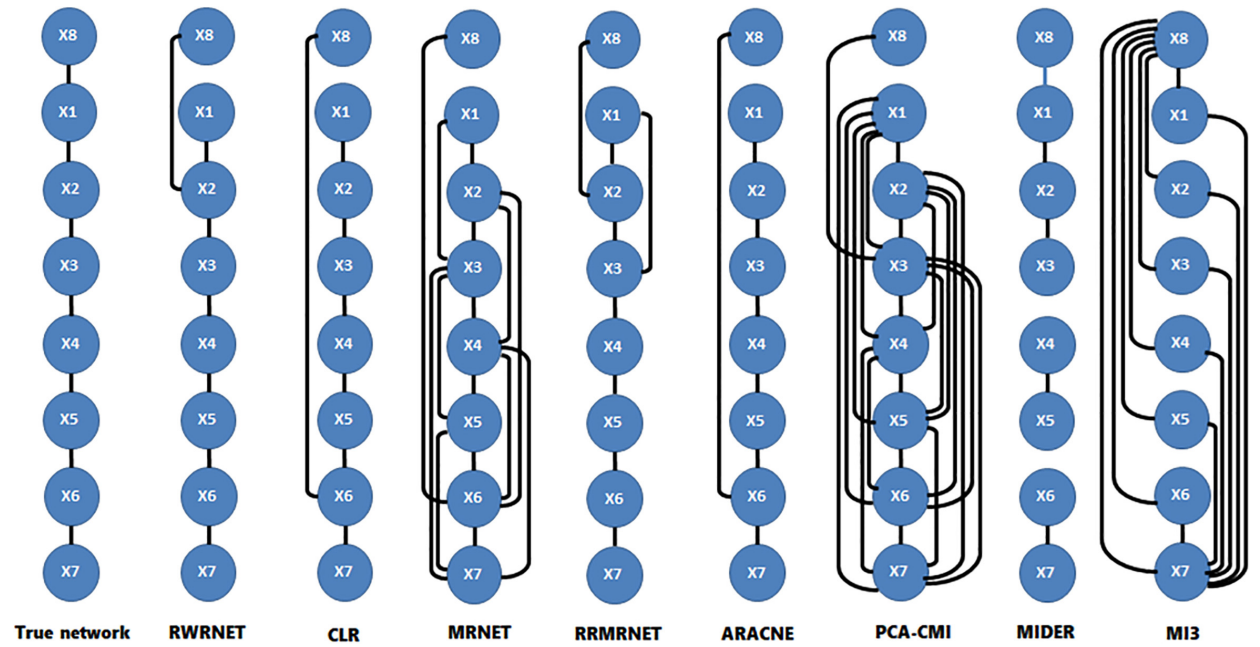

FIGURE 2 | Comparison of the different methods in the reaction chain with eight species dataset.

TABLE 5 | Comparison of the different methods' performances in the Dream3-10 gene dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	6	10	0.6	0.286	0.375	0.689
ARACNE	6	6	0.6	0.171	0.5	0.778
MRNET	6	12	0.6	0.343	0.333	0.644
MI3	8	6	0.8	0.171	0.571	0.822
MIDER	–	–	–	–	–	–
MRMSn	9	1	0.9	0.029	0.9	0.956
RRMRNET	10	0	1	0	1	1
PCA-CMI	9	1	0.9	0.029	0.9	0.956
RWRNET	8	1	0.8	0.029	0.889	0.933

probably because of noise in the data. Unfortunately, RWRNET also missed G3–G5. Similar to RWRNET, the network structure inferred by MI3 lost G3–G5 because MI3 cannot recognise the triangle relationship between G1, G3, and G5. Similarly, the loss of G3–G5 in our proposed method may have been caused by the complex network structures between G1, G3 and G5. Although RWRNET did not perform as well as the RRMRNET, MRMSn and PCA-CMI, it still performed well in terms of these four metrics compared with CLR, ARACNE, MRNET, and MI3.

We then tested the performance of the proposed method in the yeast gene expression dataset with a size of 50 (**Table 6**). The TPR of the proposed method was 0.377, whereas that of the others was between 0.052 and 0.494. RRMRNET was the only method that performed better than RWRNET in terms of TPR. The FPR of the proposed method was only 0.014, whereas the minimum FPR of the other methods was 0.015. The proposed method clearly identified correct regulatory relationships and avoided redundant regulatory relationships (TP = 29, FP = 16). In addition, the proposed method outperformed the other methods in all metrics, especially with an ACC of 0.948. In summary, the proposed method evidently performed better than the other methods.

Finally, we tested the performance of proposed method in the yeast gene expression dataset with a size of 100 (**Table 7**). The performance of RWRNET was superior to that of CLR, ARACNE, MRNET, MI3 and MIDER in all metrics. Compared with RRMRNET and PCA-CMI, RWRNET selected about 65 correct regulatory relationships (TP = 65) and introduced 50 redundant regulatory relationships (FP = 50). Although the TPR of RWRNET was not the highest (TPR = 0.392), its FPR was only 0.01. To sum up, the proposed method was considerably reduced the number of redundant regulatory relationships. Therefore, our method achieved the best performance in terms of PPV (PPV = 0.565) and ACC (ACC = 0.969).

In conclusion, RWRNET achieved a good performance in the DREAM3 challenge network dataset. The proposed method predicted as many correct regulatory relationships as possible while introducing the least redundant regulatory relationships. These features indicate that our method may be more advantageous than the other methods in inferring large-scale networks.

Results of SOS Network in *E. coli*

Finally, we tested the performance of our method in the SOS network in *E. coli*. The SOS network is a signal pathway in the SOS DNA repair system, which has been experimentally confirmed and is often used to test the effectiveness of various methods in real networks. For gene expression data, we chose

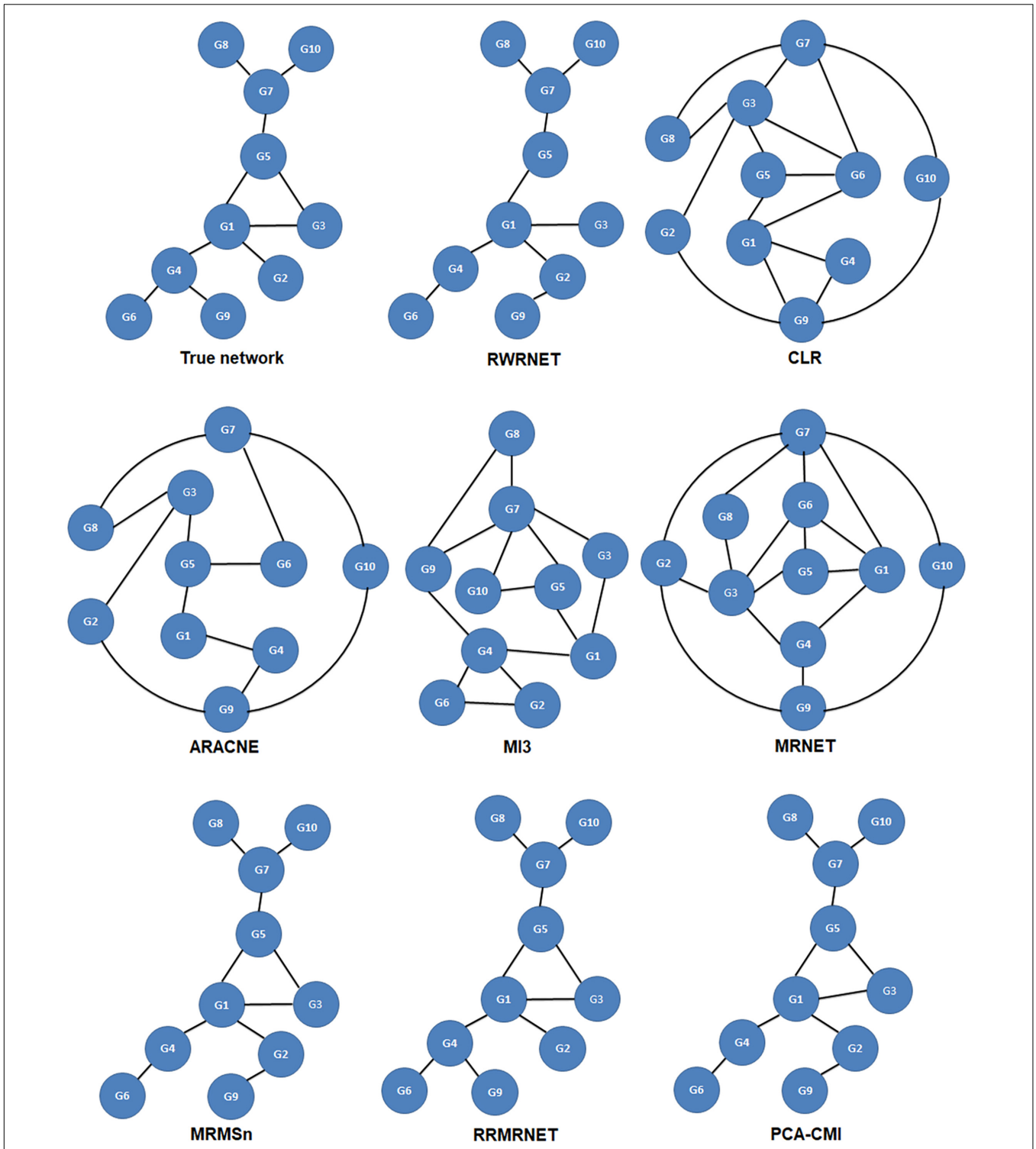

FIGURE 3 | Comparison of the different methods in the Dream3-10 gene dataset.

interference data, which were obtained through a series of transcription interference measurements.

The performance of all methods are analysed visually in **Table 8**. The performance of the proposed method was superior to that of the other methods, except for PCA-CMI in terms of ACC. In addition, the performance of the proposed method was the best in terms of PPV. At the same time, RWRNET had the best performance in terms of FPR, indicating that our method

TABLE 6 | Comparison of the different methods' performances in the Dream3-50 gene dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	19	165	0.247	0.144	0.103	0.818
ARACNE	13	125	0.169	0.109	0.094	0.846
MRNET	21	215	0.273	0.187	0.089	0.779
MI3	21	68	0.273	0.059	0.236	0.899
MIDER	4	79	0.052	0.069	0.048	0.876
MRMSn	21	17	0.273	0.015	0.553	0.94
RRMRNET	38	56	0.494	0.049	0.404	0.922
PCA-CMI	25	19	0.325	0.017	0.568	0.942
RWRNET	29	16	0.377	0.014	0.644	0.948

TABLE 7 | Comparison of the different methods' performances in the Dream3-100 gene dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	39	713	0.235	0.149	0.052	0.830
ARACNE	20	417	0.121	0.087	0.046	0.886
MRNET	49	984	0.295	0.206	0.047	0.778
MI3	27	165	0.163	0.035	0.141	0.939
MIDER	13	80	0.078	0.017	0.140	0.953
MRMSn	–	–	–	–	–	–
RRMRNET	92	238	0.554	0.05	0.28	0.937
PCA-CMI	70	64	0.422	0.013	0.522	0.968
RWRNET	65	50	0.392	0.01	0.565	0.969

TABLE 8 | Comparison of the different methods' performances in the SOS dataset.

	TP	FP	TPR	FPR	PPV	ACC
CLR	12	5	0.5	0.417	0.706	0.528
ARACNE	7	3	0.292	0.25	0.7	0.444
MRNET	17	6	0.708	0.5	0.739	0.639
MI3	9	5	0.375	0.417	0.643	0.444
MIDER	–	–	–	–	–	–
MRMSn	10	2	0.417	0.167	0.833	0.556
RRMRNET	10	2	0.417	0.167	0.833	0.556
PCA-CMI	19	3	0.92	0.25	0.84	0.778
RWRNET	15	1	0.625	0.083	0.938	0.722

introduced fewer redundant regulatory relationships than the others. A real network usually has a complex network structure and close regulatory relationships. Thus, inferring a real network is difficult. However, compared with the other methods, the proposed method performed well in the SOS network, especially in identifying redundant regulatory relationships. This result demonstrated that our method can effectively reduce network complexity and thus it is suitable for inferring real networks.

DISCUSSION

In this article, we emphasised that combining local topology with global topology can be used to improve the accuracy of network inference. However, existing methods usually focus on local topology rather than on global topology. Given that RWR is a global search algorithm, we used it to obtain the global topology of the network. To confirm that RWR can be better applied to GRNs, we improved its three key elements. First, we constructed restart probability and initial probability vector on the basis of network characteristics and regulatory mechanisms to obtain the local topology structure. Second, we adopted the asymmetric ranking strategy in constructing the roaming network to reduce the complexity of regulatory relationships. Finally, we used IPC-MB to optimise the network structure. Thus, the proposed method (RWRNET) could theoretically infer accurate GRNs.

RWRNET was tested on simulated and real datasets. In simulated datasets, the proposed method achieved excellent performance. In the reaction chain with four species, the network structure inferred by RWRNET was exactly the same as the true network. In the reaction chain with eight species, the Dream3-50 gene dataset and the Dream3-100 gene dataset, RWRNET accomplished superior performance. In the Dream3-50 gene dataset, its PPV was 0.644 and ACC was 0.948, indicating that the proposed method had a relatively good effect. These results showed that combining local topology with global topology can effectively improve the accuracy of network inference. In real datasets, RWRNET also achieved satisfactory results. Under the premise that RWRNET obtained enough regulatory relationships (TP = 15), the redundant regulatory relationships it introduced were the least (FP = 1) possibly because the processing of roaming networks reduced the effects of complex regulatory relationships on RWR. Interestingly, RWRNET performed unsatisfactorily compared with the other network inference methods in the Dream3-10 gene dataset and SOS dataset. Two possible reasons can be offered: the complexity of network structure and the amount of noise in the data. In the Dream3-10 gene network, RWRNET missed G3–G5 because of the triangular relationship between G1, G3, and G5 that increased the complexity of the network structure. Moreover, the SOS network had a lot of noise that negatively affected the performance of the proposed method.

RWRNET was tested on networks of different sizes (i.e., different numbers of variables), containing 4, 8, 9, 10, 50, and 100 genes. The experimental results show that RWRNET achieved good performance on the six different scale networks. As shown in **Tables 3–8**, except for networks of sizes 9 and 10, the performance of RWRNET showed an upward trend with the increase in the number of genes (the number of variables) in the network. Especially in networks with sizes of 50 and 100, RWRNET achieved good results in terms of the PPV and ACC metrics. Thus, combining global topology with local topology can effectively improve the accuracy of network inference.

The performance of RWRNET was also compared with that of other gene network inference methods in terms of different evaluation metrics. Results showed that RWRNET performed better than the other methods for most datasets. In the Dream3-10 Gene Network and SOS Network datasets, RWRNET did not perform as well as PCA-CMI. Although the performance of RWRNET in these datasets was not satisfactory,

it nevertheless considerably reduced the number of redundant regulatory relationships, indicating that the global topology relationships of the network can also improve the performance of network inference.

CONCLUSION

In this study, we proposed a novel network inference method based on information theory and RWR. We improved the three key elements of RWR to infer GRNs by using the proposed method. Restart probability was calculated, initial probability vector was constructed to adapt to network characteristics and regulatory mechanisms as much as possible to capture the network topology accurately. Moreover, a roaming network construction algorithm based on asymmetric ranking was proposed. This algorithm effectively reduced the effects of complex regulatory relationships on RWR. Finally, the local topology was combined with the global topology through RWR to infer the network structure. IPC-MB was used to deal with isolated nodes and optimise the network structure. The proposed method was tested in six standard network datasets, and its performance was compared with that of eight state-of-the-art methods based on information theory. Experimental results confirmed that the proposed method can efficiently and accurately infer GRNs.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

WL and XS implemented the experiments, analysed the result, and wrote the manuscript. LP and LZ analysed the result. HL and YJ provided the constructive discussions and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant No. 61902125), Natural Science Foundation of Hunan Province (2019JJ50187), and Scientific Research Project of Hunan Education Department (Grant Nos. 18B209 and 19C1788).

REFERENCES

Altay, G., and Emmert-Streib, F. (2010). Inferring the conservative causal core of gene regulatory networks. *BMC Syst. Biol.* 4:132. doi: 10.1186/1752-0509-4-132

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10101–10106. doi: 10.1073/pnas.97.18.10101

Athanasiadis, E., Bourdakou, M., and Spyrou, G. (2017). D-Map: random walking on gene network inference maps towards differential avenue discovery. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 484–490. doi: 10.1109/TCBB.2016.2535267

Bendall, S. C., Davis, K. L., Amir, E. A., Tadmor, M. D., Simonds, E. F., Chen, T. J., et al. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* 157, 714–725. doi: 10.1016/j.cell.2014.04.005

Betliński, P., and Ślęzak, D. (2012). "The problem of finding the sparsest bayesian network for an input data set is NP-Hard," in *Proceedings of the Foundations of Intelligent Systems, ISMIS 2012. Lecture Notes in Computer Science*, eds L. Chen, A. Felfernig, J. Liu, and Z. W. Raś (Heidelberg: Springer), 21–30.

Brunel, H., Gallardo-Chacon, J. J., Buil, A., Vallverdu, M., Soria, J. M., Caminal, P., et al. (2010). MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics* 26, 1811–1818. doi: 10.1093/bioinformatics/btq273

Butte, A. J., and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 2000, 418–429. doi: 10.1142/9789814447331_0040

Cantone, I., Marucci, L., Iorio, F., Ricci, M. A., Belcastro, V., Bansal, M., et al. (2009). A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* 137, 172–181. doi: 10.1016/j.cell.2009.01.055

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

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574. doi: 10.1093/bioinformatics/bth445

Fu, S., and Desmarais, M. C. (2008). "Fast markov blanket discovery algorithm via local learning within single pass," in *Advances in Artificial Intelligence. Canadian AI 2008. Lecture Notes in Computer Science*, Vol. 5032, ed. S. Bergler (Berlin: Springer), 96–107.

Fukushima, A. (2013). DiffCorr: an R package to analyze and visualize differential correlations in biological networks. *Gene* 518, 209–214. doi: 10.1016/j.gene.2012.11.028

Ghosh, A., and Barman, S. (2016). Application of Euclidean distance measurement and principal component analysis for gene identification. *Gene* 583, 112–120. doi: 10.1016/j.gene.2016.02.015

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., and Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods* 13, 845–848. doi: 10.1038/nmeth.3971

Honkela, A., Girardot, C., Gustafson, E. H., Liu, Y. H., Furlong, E. E., Lawrence, N. D., et al. (2010). Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. U.S.A.* 107, 7793–7798. doi: 10.1073/pnas.0914285107

Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.* 77, 469–480. doi: 10.1007/s001099900023

Huppenkothen, D., Heil, L. M., Hogg, D. W., and Mueller, A. (2017). Using machine learning to explore the long-term evolution of GRS 1915+105. *Month. Not. R. Astronom. Soc.* 466, 2364–2377. doi: 10.1093/mnras/stw3190

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 5:e012776. doi: 10.1371/journal.pone.0012776

Huynh-Thu, V. A., and Sanguinetti, G. (2015). Combining tree-based and dynamical systems for the inference of gene regulatory networks. *Bioinformatics* 31, 1614–1622. doi: 10.1093/bioinformatics/btu863

Kuzmanovski, V., Todorovski, L., and Dzeroski, S. (2018). Extensive evaluation of the generalized relevance network approach to inferring gene regulatory networks. *Gigascience* 7:giy118. doi: 10.1093/gigascience/giy118

Li, Z., Li, P., Krishnan, A., and Liu, J. (2011). Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* 27, 2686–2691. doi: 10.1093/bioinformatics/btr454

Lim, C. Y., Wang, H., Woodhouse, S., Piterman, N., Wernisch, L., Fisher, J., et al. (2016). BTR: training asynchronous Boolean models using single-cell expression data. *BMC Bioinform.* 17:355. doi: 10.1186/s12859-016-1235-y

Liu, W., Zhu, W., Liao, B., Chen, H., Ren, S., and Cai, L. (2017). Improving gene regulatory network structure using redundancy reduction in the MRNET algorithm. *RSC Adv.* 7, 23222–23233. doi: 10.1039/C7RA01557G

Liu, Z. P. (2015). Reverse engineering of genome-wide gene regulatory networks from gene expression data. *Curr. Genom.* 16, 3–22. doi: 10.2174/1389202915666141110210634

Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F. X., et al. (2016). Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 32, 2664–2671. doi: 10.1093/bioinformatics/btw228

Lv, Y., and Bao, E. (2009). Apoptosis induced in chicken embryo fibroblasts in vitro by a polyinosinic:polycytidylic acid copolymer. *Toxicol. Vitro* 23, 1360–1364. doi: 10.1016/j.tiv.2009.06.026

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Maetschke, S. R., Madhamshettiwar, P. B., Davis, M. J., and Ragan, M. A. (2014). Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Brief Bioinform.* 15, 195–211. doi: 10.1093/bib/bbt034

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6286–6291. doi: 10.1073/pnas.0913357107

Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I., and Califano, A. (2006). Reverse engineering cellular networks. *Nat. Protoc.* 1, 662–671. doi: 10.1038/nprot.2006.106

Meyer, P. E., Lafitte, F., and Bontempi, G. (2008). minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinform.* 9:461. doi: 10.1186/1471-2105-9-461

Mohamed Salleh, F. H., Arif, S. M., Zainudin, S., and Firdaus-Raih, M. (2015). Reconstructing gene regulatory networks from knock-out data using Gaussian noise model and pearson correlation coefficient. *Comput. Biol. Chem.* 59(Pt B), 3–14. doi: 10.1016/j.compbiolchem.2015.04.012

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., et al. (2015). Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol.* 33, 269–276. doi: 10.1038/nbt.3154

Moris, N., Pina, C., and Arias, A. M. (2016). Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* 17, 693–703. doi: 10.1038/nrg.2016.98

Mousavian, Z., Kavousi, K., and Masoudi-Nejad, A. (2016). Information theory in systems biology. Part I: gene regulatory and metabolic networks. *Semin. Cell Dev. Biol.* 51, 3–13. doi: 10.1016/j.semcdb.2015.12.007

Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., et al. (2018). Improving the measurement of semantic similarity by combining geneontology and co-functional network: a random walk based approach. *BMC Syst. Biol.* 12:18. doi: 10.1186/s12918-018-0539-0

Petralia, F., Wang, P., Yang, J., and Tu, Z. (2015). Integrative random forest for gene regulatory network inference. *Bioinformatics* 31, i197–i205. doi: 10.1093/bioinformatics/btv268

Pina, C., Teles, J., Fugazza, C., May, G., Wang, D., Guo, Y., et al. (2015). Single-cell network analysis identifies DDIT3 as a nodal lineage regulator in hematopoiesis. *Cell Rep.* 11, 1503–1510. doi: 10.1016/j.celrep.2015.05.016

Raza, K., and Alam, M. (2016). Recurrent neural network based hybrid model for reconstructing gene regulatory network. *Comput. Biol. Chem.* 64, 322–334. doi: 10.1016/j.compbiolchem.2016.08.002

Reid, J. E., and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. *Bioinformatics* 32, 2973–2980. doi: 10.1093/bioinformatics/btw372

Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. (2002). Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. U.S.A.* 99, 10555–10560. doi: 10.1073/pnas.152046799

Rosvall, M., and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci. U.S.A.* 105, 1118–1123. doi: 10.1073/pnas.0706851105

Rue, P., and Martinez Arias, A. (2015). Cell dynamics and gene expression control in tissue homeostasis and development. *Mol. Syst. Biol.* 11:792. doi: 10.15252/msb.20145549

Ruyssinck, J., Huynh-Thu, V. A., Geurts, P., Dhaene, T., Demeester, P., and Saeys, Y. (2014). NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One* 9:e92709. doi: 10.1371/journal.pone.0092709

Saito, S., Hirokawa, T., and Horimoto, K. (2011). Discovery of chemical compound groups with common structures by a network analysis approach (affinity prediction method). *J. Chem. Inf. Model.* 51, 61–68. doi: 10.1021/ci100262s

Samoilov, M. (1997). *Reconstruction and Functional Analysis of General Chemical Reactions and Reaction Networks*, Ph. D. thesis, Stanford University, Stanford, CA.

Samoilov, M., Arkin, A., and Ross, J. (2001). On the deduction of chemical reaction pathways from measurements of time series of concentrations. *Chaos* 11, 108–114. doi: 10.1063/1.1336499

Shi, M., Shen, W., Wang, H. Q., and Chong, Y. (2016). Adaptive modelling of gene regulatory network using Bayesian information criterion-guided sparse regression approach. *IET Syst. Biol.* 10, 252–259. doi: 10.1049/iet-syb.2016.0005

Sun, J., Shi, H., Wang, Z., Zhang, C., Liu, L., Wang, L., et al. (2014). Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* 10, 2074–2081. doi: 10.1039/c3mb70608g

Tan, M., Alshalalfa, M., Alhajj, R., and Polat, F. (2011). Influence of prior knowledge in constraint-based learning of gene regulatory networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 130–142. doi: 10.1109/TCBB.2009.58

Tang, W. W., Dietmann, S., Irie, N., Leitch, H. G., Floros, V. I., Bradshaw, C. R., et al. (2015). A unique gene regulatory network resets the human germline epigenome for development. *Cell* 161, 1453–1467. doi: 10.1016/j.cell.2015.04.053

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 35, 497–505. doi: 10.1093/bioinformatics/bty637

Wang, Y. X., and Huang, H. (2014). Review on statistical methods for gene network reconstruction using expression data. *J. Theor. Biol.* 362, 53–61. doi: 10.1016/j.jtbi.2014.03.040

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

Zhang, X., Zhao, J., Hao, J. K., Zhao, X. M., and Chen, L. (2015). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.* 43:e31. doi: 10.1093/nar/gku1315

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., et al. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28, 98–104. doi: 10.1093/bioinformatics/btr626

Zhou, J. X., Samal, A., d'Herouel, A. F., Price, N. D., and Huang, S. (2016). Relative stability of network states in Boolean network models of gene regulation in development. *Biosystems* 142–143, 15–24. doi: 10.1016/j.biosystems.2016.03.002

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: 02 October 2020
doi: 10.3389/fgene.2020.00820

Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning

Qijuan Gao[1†], Xiu Jin[1†], Enhua Xia[2], Xiangwei Wu[3], Lichuan Gu[4], Hanwei Yan[5], Yingchun Xia[4] and Shaowen Li[1]*

[1] Anhui Province Key Laboratory of Smart Agricultural Technology and Equipment, Anhui Agriculture University, Hefei, China, [2] State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China, [3] School of Resources and Environment, Anhui Agricultural University, Hefei, China, [4] School of Information and Computer Science, Anhui Agricultural University, Hefei, China, [5] Key Laboratory of Crop Biology of Anhui Province, Anhui Agricultural University, Hefei, China

OPEN ACCESS

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

Reviewed by:
Jun Jiang,
Fudan University, China
Jing Ding,
Nanjing Agricultural University, China
Xiaohui Zhang,
Nanjing University, China

****Correspondence:***
Shaowen Li
shaowenli@ahau.edu.cn
orcid.org/0000-0002-1118-1922

[†] These authors have contributed equally to this work

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

Received: *09 June 2020*
Accepted: *08 July 2020*
Published: *02 October 2020*

Citation:
Gao Q, Jin X, Xia E, Wu X, Gu L, Yan H, Xia Y and Li S (2020) Identification of Orphan Genes in Unbalanced Datasets Based on Ensemble Learning. Front. Genet. 11:820. doi: 10.3389/fgene.2020.00820

Orphan genes are associated with regulatory patterns, but experimental methods for identifying orphan genes are both time-consuming and expensive. Designing an accurate and robust classification model to detect orphan and non-orphan genes in unbalanced distribution datasets poses a particularly huge challenge. Synthetic minority over-sampling algorithms (SMOTE) are selected in a preliminary step to deal with unbalanced gene datasets. To identify orphan genes in balanced and unbalanced *Arabidopsis thaliana* gene datasets, SMOTE algorithms were then combined with traditional and advanced ensemble classified algorithms respectively, using Support Vector Machine, Random Forest (RF), AdaBoost (adaptive boosting), GBDT (gradient boosting decision tree), and XGBoost (extreme gradient boosting). After comparing the performance of these ensemble models, SMOTE algorithms with XGBoost achieved an F1 score of 0.94 with the balanced *A. thaliana* gene datasets, but a lower score with the unbalanced datasets. The proposed ensemble method combines different balanced data algorithms including Borderline SMOTE (BSMOTE), Adaptive Synthetic Sampling (ADSYN), SMOTE-Tomek, and SMOTE-ENN with the XGBoost model separately. The performances of the SMOTE-ENN-XGBoost model, which combined over-sampling and under-sampling algorithms with XGBoost, achieved higher predictive accuracy than the other balanced algorithms with XGBoost models. Thus, SMOTE-ENN-XGBoost provides a theoretical basis for developing evaluation criteria for identifying orphan genes in unbalanced and biological datasets.

Keywords: unbalanced dataset, ensemble learning, orphan genes, XGBoost model, two-class

INTRODUCTION

The process of identifying orphan genes is an emerging field. Orphan genes play critical roles in the evolution of species and the adaptability of the environment (Davies and Davies, 2010; Donoghue et al., 2011; Huang, 2013; Cooper, 2014; Gao et al., 2014). In most plant species, orphan genes make up about 10–20% of the number of genes (Khalturin et al., 2009; Tautz and Domazet-Loso, 2011), and each species has a specific proportion of orphan genes (Khalturin et al., 2009;

Arendsee et al., 2014), Many attempts have been made to identify orphan genes in multiple species or taxa and to analyze their functions. The whole genome and transcriptome sequences of many species have been published, including those of *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2002), O*ryza sativa* (Goff et al., 2002), Populus (Tuskan et al., 2006), and the discovery of orphan genes among these sequences has helped to clarify the special biological characteristics and environmental adaptability of angiosperm. For example, the *A. thaliana* orphan genes qua-quine starch (*QQS*) alter the carbon and nitrogen content of the plant, increasing the protein content and decreasing the starch content (Li et al., 2009; Arendsee et al., 2014); the wheat, *TaFROG* (*Triticum aestivum* fusarium resistance orphan gene) contributes to disease resistance genes for crop-breeding programs (Perochon et al., 2015); and the rice orphan gene *GN2* (GRAINS NO. 2) can affect plant height and rice yield (Chen et al., 2017).

Currently, orphan genes are detected mainly by comparison of genome and transcriptome sequences of related species using BLAST (Blast-Basic Local Alignment Search Tool; Altschul et al., 1990; Tollriera et al., 2009). However, this approach requires large server resources and time, and common problems with complexity and timeliness occur (Ye et al., 2012).

Computational technology and machine learning (ML) algorithms are widely used in the detection of orphan genes in big datasets. The method of ML can be used to make two kinds of field classification from an enormous genome dataset (Libbrecht and Noble, 2015; Syahrani, 2019). Orphan genes are widely distributed in plant species and generally exhibit significant differences in gene length, the number of exons, GC content, and expression level compared to protein-coding genes (Donoghue et al., 2011; Neme and Tautz, 2013; Yang et al., 2013; Arendsee et al., 2014; Xu et al., 2015; Ma et al., 2020). In systems biology, traditional classification methods, such as Support Vector Machines (SVMs; Zhu et al., 2009) or Random Forest (RF; Pang et al., 2006; Dimitrakopoulos et al., 2016) have been applied in the classification scheme. More recently, ensemble classification algorithms have achieved remarkable results in the fields of biology and medicine (Chen and Guestrin, 2016).

Additionally, the number of orphan genes is much less than the numbers of non-orphan gene datasets, therefore unbalanced datasets pose significant problems for developers of classifiers. The original method of over-sampling and under-sampling (Drummond and Holte, 2003; Chen and Guestrin, 2016) can help address the problems of an unbalanced dataset (Weiss, 2004; Zhou and Liu, 2006). In over-sampling methods, the synthetic minority over-sampling technique (SMOTE) (Demidova and Klyueva, 2017) can add new minority class examples, but the deleted information of majority samples may contain representative information of the majority class. Then, the improved SMOTE which combines with edited nearest neighbors (SMOTE-ENN) algorithm (Zhang et al., 2019), is used in the K-nearest neighbor (KNN) method to classify the sampled dataset, by the theory of over-sampling and under-sampling.

The bagging and boosting methods are two important approaches to ensemble learning (Breiman, 1996) that can improve the accuracy of a model significantly. The boosting family algorithm adaptively fits a series of weak models and combines them. Because the number of minority samples in an unbalanced dataset is small, they are easily misclassified, so the results of the previous classifier determine the parameters of the later model and let the next classifier focus on training the last misclassified sample. Therefore, the Boosting family algorithm pays more attention to samples that are difficult to classify, which can effectively improve the prediction accuracy.

In the study described in this manuscript, over-sampling and under-sampling algorithms were introduced to clean up unbalanced data (Chawla et al., 2002). Representative serial classified algorithms of the Boosting family are AdaBoost (adaptive boosting), GBDT (gradient boosting decision tree), XGBoost (extreme gradient boosting), and the representative parallel classified algorithm are SVM and RF. The performance of these five classification models with over-sampling SMOTE is better than those with single classifiers. The relevant features of the whole gene sequencing of *A. thaliana* were designed as a model for the identification and prediction of orphan genes. The result could show that balancing algorithms play a more effective guiding role in identifying the orphan genes in a species.

MATERIALS AND METHODS

Data Processing Method for Unbalanced Data

Data preprocessing is the first step for data mining and affects the result. Preprocessing includes data discretization, missing values, attribute coding, and data standard regularization. In practice, each industry has unique data characteristics, so different methods are used to analyze the data and perform preprocessing.

The processing of unbalanced data describes classes with obviously uneven distribution. The traditional method used random over-sampling to increase the number of small-class samples to achieve a consistent number. Because this method achieves balance by a single random over-sampling strategy of copying data, the added repeated data will increase the complexity of data training and induce over-fitting.

To deal with the problem of unbalanced data classification, some algorithms have been used effectively to improve the performance of classification. Common methods for processing datasets included mainly: over-sampling and under-sampling, or a combination of under-sampling and over-sampling.

Over-Sampling SMOTE and Borderline SMOTE

To solve the problem of over-fitting associated with unbalanced data when the learning information is not generalized, Chawla et al. (2002) proposed the SMOTE algorithm for preprocessing over-sampling data of synthetic minority categories. SMOTE was designed based on a random over-sampling method in the feature space. By analyzing data with few categories, many new data are generated by linear interpolation and added to the original data set. SMOTE first selects each sample from the minority samples successively as the root sample for the synthesis of the

new sample. Then according to the up-sampling rate n, SMOTE randomly selects one of K (K is generally odd, such as $K = 5$) neighboring samples of the same category, which is used as an auxiliary sample to synthesize a new sample and repeated n times. Finally, linear interpolation is performed between the sample and each auxiliary sample to generate n synthesized samples. The basic flow of the algorithm is:

(i) Find K samples of the nearest neighbor for each sample x_i, whose label is "1";
(ii) A sample x_j belonging with few categories is selected randomly from K;
(iii) Linearly interpolate randomly between x_i and x_j to construct a new minority sample.

The SMOTE algorithm effectively solves the problem of over-fitting caused by the blind replication of random over-sampling techniques. However, the selection of the nearest neighbor sample in step 1 exits is purposeless. Users need to determine the number of K values of the neighbor samples themselves, so it is difficult to determine the optimal value. Additionally, the newly synthesized samples may fall into the sample area labeled "0," which confuses the boundaries between them and interferes with the correct classification of the data.

Therefore, to address these two problems, Wang et al. (2015) proposed Borderline SMOTE (an over-sampling method in unbalanced datasets learning), which is an improved over-sampling algorithm based on SMOTE. By finding suitable areas that can better reflect the characteristics of the data to be interpolated, the problem of sample overlap can be solved. The Borderline SMOTE algorithm uses only a few samples on the boundary to synthesize new samples, thereby improving the internal distribution of samples.

Adaptive Synthetic Sampling

Adaptive Synthetic Sampling adaptively generates different numbers of sampling samples according to data distribution (He et al., 2008). The basic flow of the algorithm is below:

(i) Calculate the number of samples to be synthesized, as follows: $G = (m_l - m_s) \times \beta$, where m_l is the number of majority samples, andm_s is the number of minority samples. If $\beta = 1$, the number of positive and negative samples is the same after sampling, indicating that the data is balanced at this time.
(ii) Calculate the number of K nearest neighbor value of each minority sample, Δ is the number of majority samples in the K neighbors, the formula is as follows: $r_i = \Delta_i/K$, where Δ_i is the number of majority samples in K nearest neighbors, $i = 1,2,3.....$, m_s
(iii) To normalize r_i, the formula is $\hat{r} = r_i / \sum_{i=1}^{m_s} r_i$
(iv) According to the sample weights, calculate the number of new samples that need to be generated for every minority sample. The formula is $g = \hat{r} \times G$.

Select one sample from the K neighbors around each data with the label "1" to be synthesized, calculate the number to be generated according to g the formulas$s_i = x_i + (x_{zi} - x_i) \times \lambda$, where s_i is the synthetic sample, x_i is the ith minority samples, and x_{zi} is a random number of the minority sample $\lambda \in [0,1]$ selected from the K nearest neighbors of x_i .

Combining Algorithms

Apart from using a single under-sampling or over-sampling method, two resampling methods can be combined. For example, SMOTE-ENN (Zhang et al., 2019), ENN is an under-sampling method focusing on eliminating noise samples, which is added to the pipeline after SMOTE to obtain cleaner combined samples. For each combined sample, its nearest-neighbors are computed according to the Euclidean distance. These samples will be removed whose most KNN samples are different from other classes (shown in **Figure 1**).

SMOTE-Tomek (Batista et al., 2004) also combine SMOTE with Tome-links (Tomek), a data cleaning method to handle the overlapping parts, which are difficult to classify for a few classes and most surrounding samples. A Tome link can be defined as follows: given that sample x and y belong to two classes, and be the distance between x to y as $d(x,y)$. If there is not a sample z, such as $d(x,z) < d(x,y)$ or $d(y,z) < d(x,y)$, $A(x,y)$ pair is called a Tome link.

Ensemble Learning Methods

The main idea of the ensemble learning algorithm is to construct multiple classifiers with weak performance and use a certain strategy to combine them into a classifier with strong generalization performance. Consequently, the performance of the ensemble is better than that of a single classifier.

This study created two classification models for unbalanced datasets and used Python to build five integrated learning models of SVM, RF, AdaBoost, GBDT, and XGBoost and conducted comparative experiments to find the optimal model. XGBoost performed best in the classification, Five kinds of balanced data learning methods of resampling: SMOTE, BSMOTE, ADASYN, SMOTE-ENN, and SMOTE-Tomek, were then combined with XGBoost to build an ensemble model that produced excellent classification results (Lemaitre et al., 2017; Wu et al., 2018).

XGBoost was modified by adding regular items to the GBDT algorithm that can predict the orphan gene binary classification problem and increase the calculation speed. XGBoost uses the gradient boosting algorithm of the based learner classification and regression tree (CART) to calculate the complexity of the leaf nodes of each tree and uses the gradient descent algorithm to minimize the loss for finding the optimal prediction score, thus avoiding over-fitting the learned model and effectively controlling the complexity of the model (Chen and Guestrin, 2016).

The derivation process is as follows:

(i) Objective function: $\text{obj}(\theta) = \sum_{i}^{n} l\left(y_i, \widehat{y_i}\right) + \sum_{k=1}^{K} \Omega(f_k)$
(ii) Using the first and second derivatives, the Taylor formula expands:

$$\text{obj}^{(t)} = [\sum_{i}^{n} l\left(y_i, \widehat{y_i}^{t-1}\right) + g_i f_t(x_i)] + \Omega\left(f_t\right) + constant$$

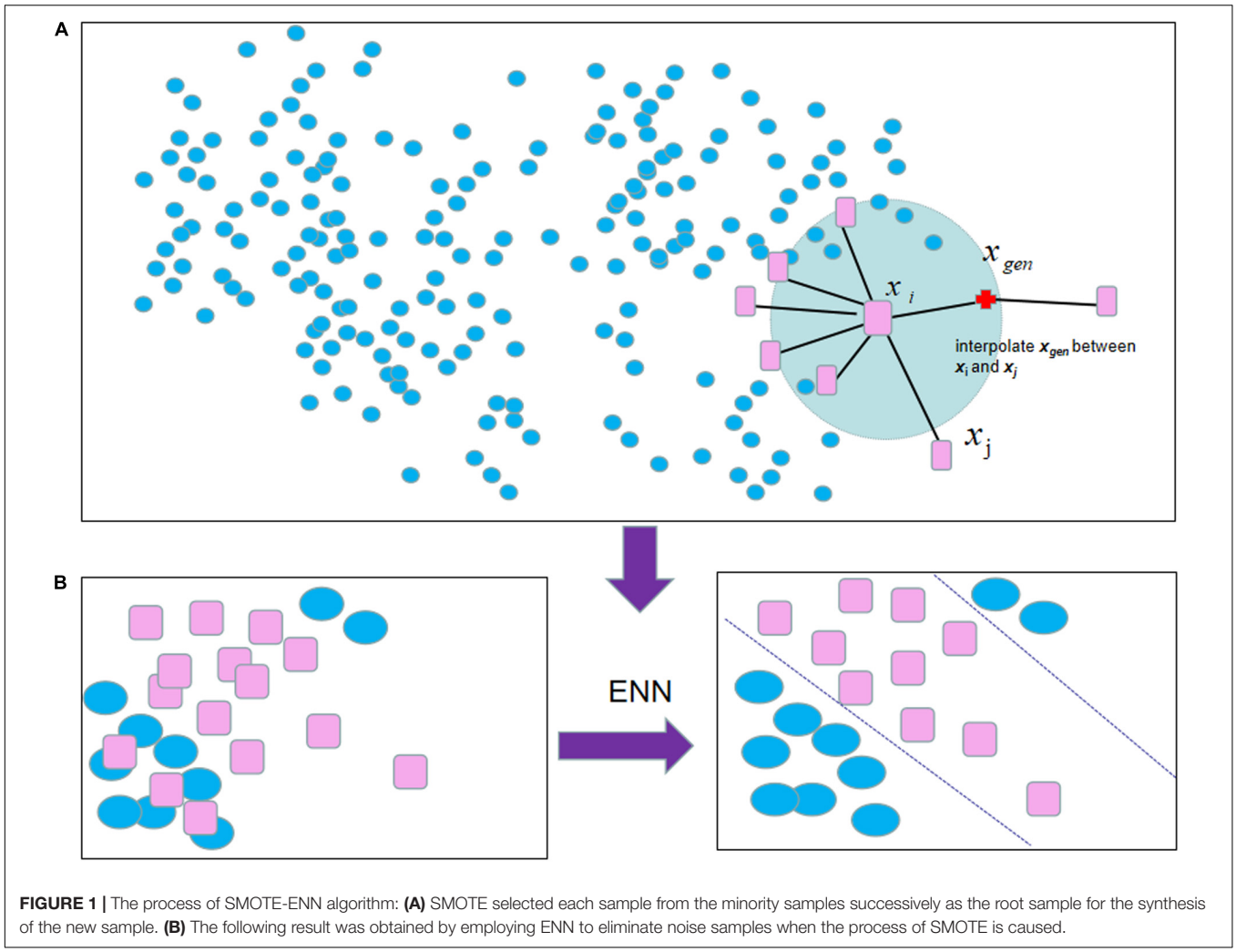

FIGURE 1 | The process of SMOTE-ENN algorithm: **(A)** SMOTE selected each sample from the minority samples successively as the root sample for the synthesis of the new sample. **(B)** The following result was obtained by employing ENN to eliminate noise samples when the process of SMOTE is caused.

(iii) Measuring the complexity of the decision tree as:$\Omega\left(\mathrm{f}\right) = \gamma\mathrm{T} + \frac{1}{2}\lambda\sum_{\mathrm{j}=1}^{T} w_{\mathrm{j}}^2$, where T is the number of leaf nodes in the decision tree, and w is the prediction result corresponding to the leaf node.

(iv) Substituting the above two steps into the objective function (1), it is organized as:

$$\mathrm{obj}^{(t)} \approx \sum_{\mathrm{i}=1}^{n}[g_{\mathrm{i}}w_{q(x_{\mathrm{i}})} + \frac{1}{2}\left(h_{\mathrm{i}}w_{q(x_{\mathrm{i}})}^2\right)] + \gamma\mathrm{T} + \frac{1}{2}\sum_{\mathrm{j}=1}^{T} w_{\mathrm{j}}^2$$

$$= \sum_{\mathrm{j}=1}^{T}\left[G_{\mathrm{j}}w_{\mathrm{j}} + \frac{1}{2}\left(H_{\mathrm{j}} + \lambda\right)w_{\mathrm{j}}^2\right] + \gamma\mathrm{T}$$

(v) Then, $I_{\mathrm{j}} = \left\{i|q\left(x_{\mathrm{i}}\right) = j\right\}$, represents the sample set belonging to the j-th leaf node.

$$G_{\mathrm{j}} = \sum_{\mathrm{i}\in\mathrm{I}_{\mathrm{j}}} g_{\mathrm{i}}, \quad H_{\mathrm{j}} = \sum_{\mathrm{i}\in\mathrm{I}_{\mathrm{j}}} h_{\mathrm{i}},$$

(vi) To minimize the objective function, let the derivative be 0 and find the optimal prediction score for each leaf node:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

(vii) Substitute the objective function again to get its minimum value:

$$\mathrm{obj}^{(t)} = -\frac{1}{2}\sum_{J=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T$$

(viii) Find the optimization goal of each layer of the build tree through obj to find the optimal tree structure, and split the left and right subtrees as:

$$\mathrm{Gain}\left(\phi\right) = \frac{1}{2}\left[\frac{\left(\sum_{\mathrm{i}\subseteq\mathrm{I}_{\mathrm{L}}} g_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}\subseteq\mathrm{I}_{\mathrm{L}}} h_{\mathrm{i}} + \lambda} + \frac{\left(\sum_{\mathrm{i}\subseteq\mathrm{I}_{\mathrm{R}}} g_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}\subseteq\mathrm{I}_{\mathrm{R}}} h_{\mathrm{i}} + \lambda} - \frac{\left(\sum_{\mathrm{i}\subseteq\mathrm{I}} g_{\mathrm{i}}\right)^2}{\sum_{\mathrm{i}\subseteq\mathrm{I}} h_{\mathrm{i}} + \lambda}\right] - \gamma$$

TABLE 1 | Binary confusion matrix.

	Real positive	Real negative
Predict positive	TP	FP
Predict negative	FN	TN

Confusion Matrix

The confusion matrix (error matrix) is a matrix table (shown in **Table 1**) that is used to judge whether a sample is 0 or 1 and reflects the accuracy of classification. The results of the classification model are analyzed using four basic indicators: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The prediction classification model that gives the best results will have a large number of TPs and TNs and a small number of TPs and TNs.

(i) True positive (TP): the actual value of the model is the orphan genes, so the model predicts the number of orphan genes.
(ii) False positive (FP): the actual value of the model is the orphan gene, but the model predicts the number of non-orphan genes.
(iii) False negative (FN): the true value of the model is non-orphan genes, so the model predicts the number of orphan genes.
(iv) True negative (TN): the true value of the model is non-orphan genes, but the model predicts the number of non-orphan genes.

TABLE 2 | Training and testing datasets used to design and evaluate the model classifiers.

Class	Train dataset	Test dataset	Original dataset
None-orphan genes	24833	6208	31041
Orphan genes	1427	357	1784

Recall, Precision, and F1 Value as Performance Indicators

A large number of confusion matrix statistics make it difficult to measure the pros and cons of a model. Therefore, we added using Recall, Precision, and F1-score, as performance indicators to better evaluate the performance of the model:

(i) Recall rate (accuracy rate of positive samples):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

(ii) Precision (precision rate of positive samples):

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

(iii) F1-score value:

$$\text{F1}_{\text{SCORE}} = \frac{2\text{PR}}{\text{P} + \text{R}}$$

ROC Curve and AUC Value

The receiver operating characteristic (ROC) curve reflects the probability of identifying correct and wrong results according to different thresholds. The curve passes (0, 0) and (1, 1), and the validity of the model is generally determined by the diagonal of the curve in the upper left section of the graph.

The AUC value is the value of the area under the ROC curve, which is generally between 0.5 and 1. The quantized index value can better compare the performance of the classifiers: a high performance classifier AUC value is close to 1.

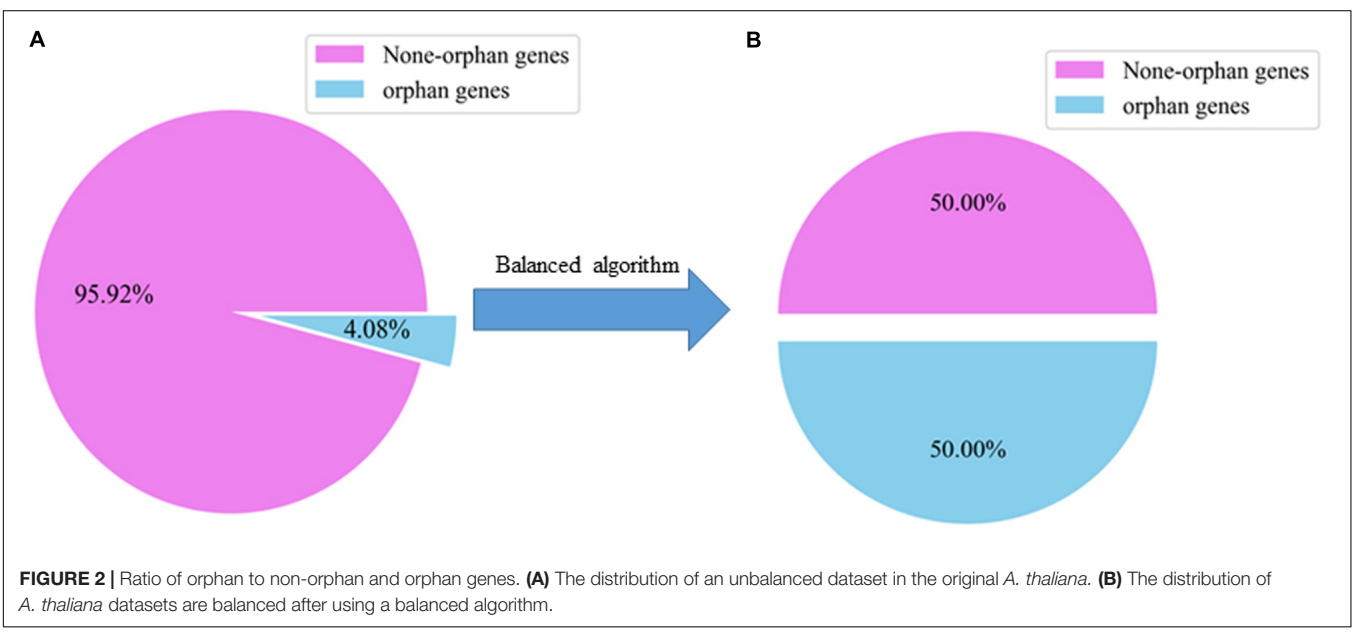

FIGURE 2 | Ratio of orphan to non-orphan and orphan genes. **(A)** The distribution of an unbalanced dataset in the original *A. thaliana*. **(B)** The distribution of *A. thaliana* datasets are balanced after using a balanced algorithm.

TABLE 3 | Compute time compared among Adaboost, GBDT, XGBoost models with SMOTE algorithm.

Traing model	Time (s)
AdaBoost	11.7
GBDT	10.3
XGBoost	0.3

TABLE 4 | F1 scores of GBDT, Adaboost, XGBoost models with the SMOTE algorithm on test datasets.

n_estimator	Learning_rate	Testing Algorithm (%)		
		GBDT	AdaBoost	XGBoost
200	0.2	90	87.6	93
200	0.1	89	88	92
200	0.01	87	87.4	88
150	0.2	90	87.9	93
150	0.1	89	87.4	91
150	0.01	87	87.4	88
100	0.2	89	87.5	92
100	0.1	88	87.5	90
100	0.01	87	87.5	88

RESULTS

Collating Feature Data of Orphan and Non-orphan Genes

The whole genome data of the angiosperm *A. thaliana* were obtained from The Arabidopsis Information Resource (TAIR8) dataset ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR8_genome_release, which contained a total of 32825 gene sequences. The known orphan genes of *A. thaliana* downloaded from the public website https://www.biomedcentral.com/content/supplementary/1471-2148-10-%2041-S2.TXT (Lin et al., 2010). The protein sequences and coding sequences were downloaded from TAIR. GC percent, protein length, molecular mass, protein isoelectric point (pI), average exon number were selected.

The six features of the protein and coding sequences were recorded as V1-V6 (Perochon et al., 2015; Shah, 2018; Ji et al., 2019). The class of orphan genes is recorded as a *Class* problem, where the label of orphan genes is recorded as 1 and the non-orphan genes are recorded as 0, combined with V1–V6 features (Ji et al., 2019; Li et al., 2019).

Analyzing Orphan and Non-orphan Gene Dataset

There were 32825 samples in the gene datasets, but only about 4.08% of them were orphan genes, so the distribution of orphan and non-orphan samples was uneven. We evaluated whether the models can identify the orphan genes. For traditional ML classification algorithms, the premise is that the amount of data between categories is balanced, or that the cost of misclassification for each category is the same. Therefore, the direct application of many algorithms leads to more predictions being made for the category with a larger number.

To solve the problem, of unbalanced data sets, we first used over-sampling to copy small sample data, which increased the number of categories with fewer samples. This method balanced the numbers of orphan and non-orphan samples to improve the learning ability of the classifier. The random sampling method was used to divide the samples into training and testing sets with a ratio of 8:2 which is the same ratio as the original dataset (**Table 2**).

The training set was used to design the model, and the test set was used to test the performance of the model. The Precision, Recall, F1, and AUC evaluation indicators were used to compare the model classifiers to determine the effectiveness of the models and select the best model.

We used SMOTE to balance the numbers of orphan and non-orphan genes in the original *A. thaliana* gene dataset shown in **Figure 2**.

Training Model Using Ensemble Learning Methods

Among the ensemble learning methods, some members of the Boosting family, such as AdaBoost., GBDT, XGBoost, can be used to train classifying models, which can save the compute time remarkably (**Table 3**).

Two parameters, train_node and learning_rate were considered to reduce the complexity in modeling. However, selecting the best parameters for the ensemble learning algorithms is important to avoid an over-fitting problem. For this study, we set the learning_rate as 0.01, 0.1, and 0.2 and train_node as 100, 150, 200 to compute the F1 score.

AdaBoost, GBDT, XGBoost with the two parameters are used to classify the samples in the training and testing datasets (**Table 2**). The results are shown in **Table 4**.

Overall, the XGBoost with SMOTE performed better than AdaBoost and GBDT models with SMOTE.

Performance of Different Models With Balanced and Unbalanced Datasets

Five models, SVM, RF, GBDT, AdaBoost, and XGBoost were used as baseline classifiers to distinguish orphan and non-orphan genes in the unbalanced and balanced *A. thaliana* gene datasets. The results are shown in **Table 5**.

Overall, the five models produced better results with the balanced datasets. However, the accuracy of the models with the balanced datasets was lower than with the unbalanced dataset, which indicates the classification of orphan genes was towards the majority samples of non-orphan genes. These results clearly show that designing models using unbalanced datasets will lead to significant inaccuracies, which cannot identify orphan genes VS non-orphan genes precisely. This indicates the importance of using a balancing algorithm to balance datasets in the first step of the classification process.

TABLE 5 | Performance of models in distinguishing orphan vs. non-orphan genes in *A. thaliana* gene balanced and unbalanced datasets with 8:2 training-testing ratios.

Best Model	Unbalanced datasets (%)					Balanced datasets (SMOTE) (%)				
	Accuracy	Precision	Recall	F1	AUC	Accuracy	Precision	Recall	F1	AUC
SVM	97	78	47	58	74	83	83	83	83	88
RF	96	47	58	52	93	84	77	98	86	95
GBDT	96	60	59	60	94	87	87	87	87	94
Adaboost	97	56	73	45	93	87	87	86	89	95
XGBoost	97	81	50	62	94	92	91	95	93	97

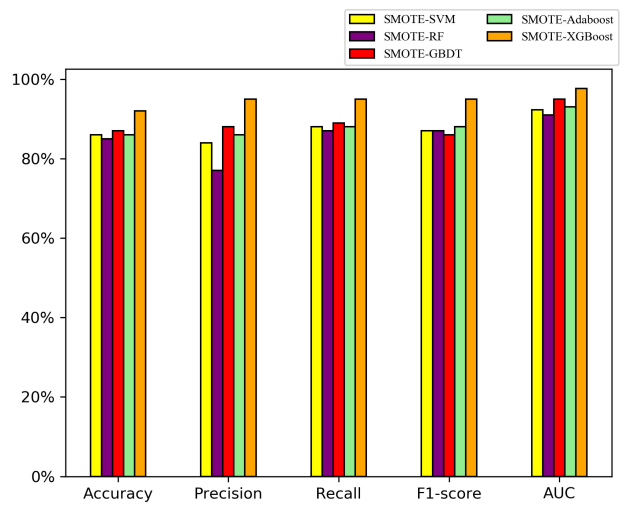

FIGURE 3 | Performance indices of five classifiers model with the SMOTE algorithm on the testing dataset to distinguishing orphan and non-orphan genes after balancing the distribution of *A. thaliana* gene dataset.

On the balanced *A. thaliana* gene dataset, the performance indices of five classifier models on the testing datasets are shown in **Figure 3**. Overall, the ensemble models were better than the single classifiers, as determined by the performance indicators, among them, the AUC and precision values of XGBoost, GBDT, AdaBoost with SMOTE were higher than SVM, RF with SMOTE algorithm. Particularly, XGBoost with SMOTE produced the highest results among all classifier models (*t*-test, $P < 0.05$). In particular, the F1 value indicates that the XGBoost model can distinguish orphan genes and non-orphan genes precisely.

We found that the ROC curve of SMOTE-XGBoost completely wrapped the ROC curves of the other methods, and the Precision-Recall (PR) curve confirmed that XGBoost produced the best performance among the five balancing algorithm methods (**Figure 4**).

The PR curve (**Figure 4**) indicated that when the classification threshold was near 1, all the samples were classified as non-orphan genes, and the Precision and Recall values were 0 at this time. When the classification threshold was 0.9, there were no FPs, so the Precision was 1, which means all the genes were classified as orphans. Because the number of TPs was small, the Recall was small and the Precision value declined continually. When the threshold declined to 0, all the samples were classified as non-orphan genes, meaning that the Precision will not be 0, because there were no FNs, and the Recall value was 1. This indicates that the prediction result is reasonable.

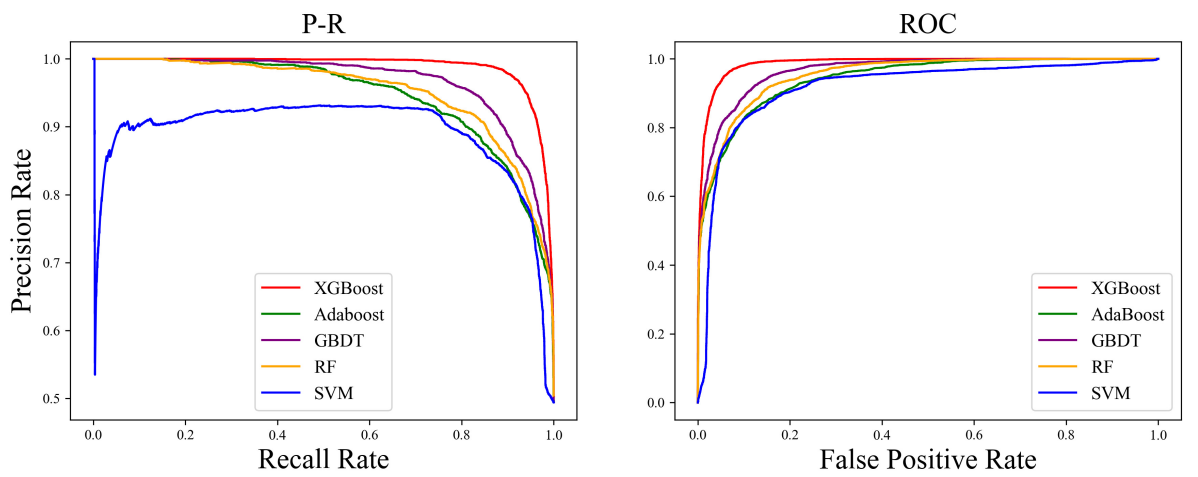

FIGURE 4 | Precision-Recall (PR) curve and ROC curve "True" area and for the five classifiers with an unbalanced dataset.

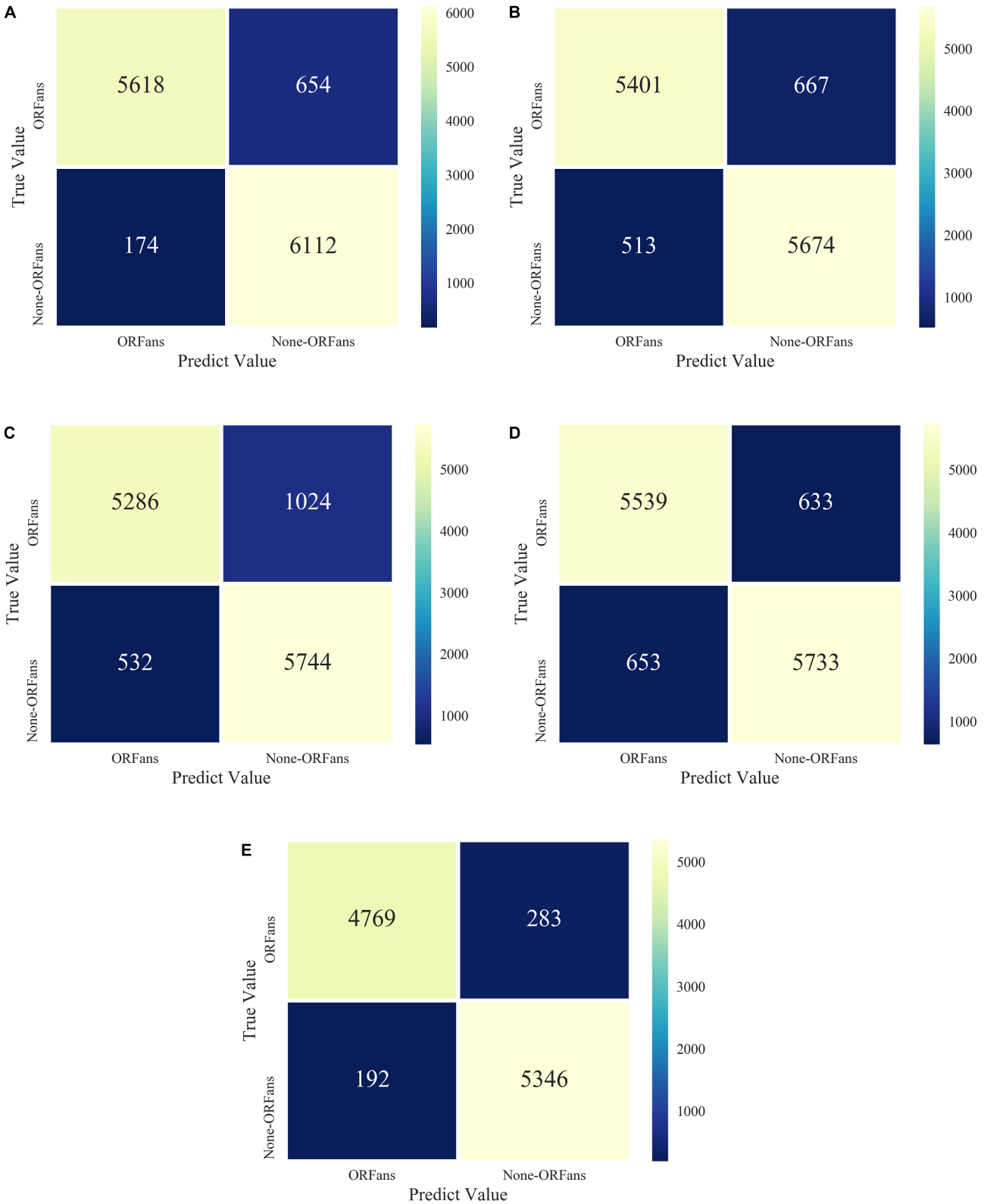

FIGURE 5 | Model confusion matrices for XGBoost combined with five over-sampling models: **(A)** ADASYN-XGB; **(B)** SMOTE-Tomek-XGB; **(C)** BSMOTE-XGB; **(D)** SMOTE-XGB; **(E)** SMOTE-ENN-XGB.

TABLE 6 | Performance indices of the ensemble of composite XGBoost classifiers.

Evaluation value	ADASYN-XGB (%)	BSMOTE-XGB (%)	SMOTE-XGB (%)	SMOTE-ENN-XGB (%)	SMOTE-Tomek-XGB (%)
Accuracy	85	92	88	95	89
Precision	83	89	87	94	88
Recall	89	97	89	95	90
F1	86	93	88	95	89
AUC	92	97	95	98	96

Performance of XGboost With Different Balanced Algorithm Methods

We also tested five different models, XGBoost combined with a balanced algorithm including SMOTE, BSMOTE, ADASYN, SMOTE-Tomek, SMOTE-ENN, to further explore the result of the unbalanced datasets. The results of the confusion matrices of five models are shown in **Figure 5**. The performance of the SMOTE-ENN-XGBoost model is better and the predicted value is higher,which indicates fewer incorrect classifiers.

The performance indices of the five balanced algorithms with ensemble XGBoost classifiers models are shown in **Table 6**. The ensemble SMOTE-ENN-XGB model had the highest among the other ensemble models to predict orphan genes (ORFans).

Therefore, the SMOTE-ENN-XGBoost model is used to classify and analyze the orphan genes in unbalanced datasets and applied to the actual predictions.

DISCUSSION

Our research indicates that in the classification of orphan vs Non-orphan genes the ML method is preferred because the traditional biological method is time-consuming and labor-intense. Since the orphan genes of plant species have similar characteristics, we selected 6 features of the *A. thaliana* dataset to build training and testing models (Donoghue et al., 2011).

The datasets of orphan genes and non-orphan genes are often unbalanced, which tends to produce a bias towards majority samples. To overcome this problem, we combined over-sampling and under-sampling algorithms, making the trained model with balanced datasets, which improves the generalization ability of the model, and eventually, the precision, recall, F1, and AUC for the test set are significantly increased. To further compare the result of the evaluation, the balanced algorithm combines classifying learning algorithms, RF, SVM, Adaboost, GBDT, XGBoost, which have similar improved results. Furthermore, the boosting methods containing Adaboost, GBDT, XGBoost have a better performance than those that use RF and SVM. Thus, ensemble boosting learning models are an important method in advancing the identification of orphan genes and non-orphan genes in unbalanced datasets. At the same time, the same training node and learning_rate parameters were automatically used for parallel computing among the boosting methods, which revealed that the XGBoost model was more practical than other models for classifying orphan genes. In particular, since it saves time and labor, classifying orphan versus non-orphan genes experimentally in this way could benefit this field and future studies.

To increase the precision of these ensemble models, we compared five different balanced algorithms including SMOTE, BSMOTE, ADASYN, SMOTE-Tomek, SOMTE-ENN combing with XGBoost models. SMOTE-ENN with XGBoost has a better evaluation result, especially the value of Recall. In this paper, we propose the SMOTE-ENN-XGBoost model for efficiently identifying unbalanced datasets of orphan genes. We built the SMOTE-ENN-XGBoost model to classify genes by predicting 0 or 1 values. The results showed that the ensemble classifiers method classified the orphan and non-orphan genes more precisely than the single classifiers, and among the five ensemble models with XGBoost, the SMOTE-ENN-XGBoost model performed best.

This study provides a new method for the identification of unbalanced datasets of orphan genes, which can be applied in the classification of unbalanced biological datasets. Meanwhile, the method can support the evolution of species.

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**.

AUTHOR CONTRIBUTIONS

QG and XJ: development of methodology. HY and YX: sample collection. QG, XJ, EX, and XW: analysis and interpretation of data. QG, XJ, LG, and SL: writing, review, and revision of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the State Key Laboratory of Tea Plant Biology and Utilization (Grant Number

SKLTOF20190101), the National Science and Technology Support Program (Grant Number 2015BAD04B0302), and the International S&Y Cooperation Project of the China Ministry of Agriculture (Grant Numbers 2015-Z44 and 2016-X34).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00820/full#supplementary-material

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Arabidopsis Genome Initiative (2002). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–813. doi: 10.1038/35048692

Arendsee, Z. W., Li, L., and Wurtele, E. S. (2014). Coming of age: orphan genes in plants. *Trends Plant Sci.* 19, 698–708. doi: 10.1016/j.tplants.2014.07.003

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *Sigkdd Expl.* 6, 20–29. doi: 10.1145/1007730.1007735

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 26, 123–140. doi: 10.1007/bf00058655

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

Chen, H., Tang, Y., Liu, J., Tan, L., Jiang, J., Wang, M., et al. (2017). Emergence of a Novel Chimeric Gene Underlying Grain Number in Rice. *Genetics* 205, 993–1002. doi: 10.1534/genetics.116.188201

Chen, T., and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System," in *knowledge discovery and data mining ACM SIGKDD International Conference on knowledge discovery and data mining*, Washingtom, DC: University of Washington Vol. 2016, 785–794.

Cooper, E. D. (2014). Horizontal gene transfer: accidental inheritance drives adaptation. *Curr. Biol.* 24, R562–R564. doi: 10.1016/j.cub.2014.04.042

Davies, J., and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* 74, 417–433. doi: 10.1128/MMBR.0001610

Demidova, L., and Klyueva, I. (2017). "SVM classification: Optimization with the SMOTE algorithm for the class imbalance problem," in *Paper presented at the mediterranean conference on embedded computing*, (New Jersey: IEEE).

Dimitrakopoulos, G. N., Balomenos, P., Vrahatis, A. G., Sgarbas, K. N., and Bezerianos, A. (2016). Identifying disease network perturbations through regression on gene expression and pathway topology analysis. *Int. Conferen. IEEE Engin. Med. Biol. Soc.* 2016, 5969–5972.

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., and Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evol. Biol.* 11:47. doi: 10.1186/1471-2148-11-47

Drummond, C., and Holte, R. C. (2003). "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Workshop Notes ICML Workshop Learn*. Washington, DC.

Gao, C., Ren, X., Mason, A. S., Liu, H., Xiao, M., Li, J., et al. (2014). Horizontal gene transfer in plants. *Funct. Integr. Genom.* 14, 23–29. doi: 10.1007/s10142-013-0345340

Goff, S. A., Ricke, D. O., Lan, T., Presting, G. G., Wang, R., Dunn, M., et al. (2002). A draft séquence of the rice genome (Oryza sativa L. *ssp. japonica) : The rice genome*. *Science* 296, 79–92. doi: 10.1126/science.1068037

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. New Jersey: IEEE, 1322–1328.

Huang, J. (2013). Horizontal gene transfer in eukaryotes: the weak-link model. *Bioessays* 35, 868–875. doi: 10.1002/bies.201300007

Ji, X., Tong, W., Liu, Z., and Shi, T. (2019). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. *Front. Genet.* 10:600. doi: 10.3389/fgene.2019.00600

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., and Bosch, T. C. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Gen.* 25, 404–413. doi: 10.1016/j.tig.2009.07.006

Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* 18, 559–563.

Li, L., Foster, C. M., Gan, Q., Nettleton, D., James, M. G., Myers, A. M., et al. (2009). Identification of the novel protein QQS as a component of the starch metabolic network in Arabidopsis leaves. *Plant J.* 58, 485–498. doi: 10.1111/j.1365-313X.2009.03793.x

Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Front. Genet.* 10:1077. doi: 10.3389/fgene.2019.01077

Libbrecht, M., and Noble, W. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Lin, H. N., Moghe, G., Ouyang, S., Iezzoni, A., Shiu, S. H., Gu, X., et al. (2010). Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. *BMC Evol. Biol.* 10:41. doi: 10.1186/1471-2148-10-41

Ma, S., Yuan, Y., Tao, Y., Jia, H., and Ma, Z. (2020). Identification, characterization and expression analysis of lineage-specific genes within Triticeae. *Genomics* 112, 1343–1350. doi: 10.1016/j.ygeno.2019.08.003

Neme, R., and Tautz, D. (2013). Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* 14:117. doi: 10.1186/1471-2164-14-117

Pang, H., Lin, A. P., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., et al. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics* 22, 2028–2036. doi: 10.1093/bioinformatics/btl344

Perochon, A., Jia, J. G., Kahla, A., Arunachalam, C., Scofield, S. R., Bowden, S., et al. (2015). TaFROG Encodes a Pooideae Orphan Protein That Interacts with SnRK1 and Enhances Resistance to the Mycotoxigenic Fungus Fusarium graminearum. *Plant Physiol.* 169, 2895–2906. doi: 10.1104/pp.15.01056

Shah, R. (2018). *Identification and characterization of orphan genes in rice (Oryza sativa japonica) to understand novel traits driving evolutionary adaptation and crop improvement. Creative Components*. America: IOWA State University.

Syahrani, I. M. (2019). Comparation Analysis of Ensemble Technique With Boosting(Xgboost) and Bagging (Randomforest) For Classify Splice Junction DNA Sequence Category. *J. Penel. Pos dan Inform.* 9, 27–36. doi: 10.17933/jppi.2019.090103

Tautz, D., and Domazet-Loso, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* 12, 692–702. doi: 10.1038/nrg3053

Tollriera, M., Castelo, R., Bellora, N., and Alba, M. M. (2009). Evolution of primate orphan proteins. *Biochem. Syst. Ecol.* 37, 778–782. doi: 10.1042/bst0370778

Tuskan, G. A., Difazio, S. P., Jansson, S., Bohlmann, J., Grigoriev, I. V., Hellsten, U., et al. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. *& Gray)*. *Science* 313, 1596–1604.

Wang, K. J., Adrian, A. M., Chen, K. H., and Wang, K. M. (2015). A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: a case study in Taiwan. *Comput. Meth. Progr. Biomed.* 119, 63–76. doi: 10.1016/j.cmpb.2015.03.003

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *Sigkdd Explor.* 6, 7–19. doi: 10.1145/1007730.1007734

Wu, Z., Lin, W., and Ji, Y. (2018). *An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics*. New Jersey: *IEEE*, 8394–8402.

Xu, Y., Wu, G., Hao, B., Chen, L., Deng, X., and Xu, Q. (2015). Identification, characterization and expression analysis of lineage-specific genes within sweet orange (Citrus sinensis). *BMC Genomics* 16:995. doi: 10.1186/s12864-015-2211-z

Yang, L., Zou, M., Fu, B., and He, S. (2013). Genome-wide identification, characterization, and expression analysis of lineage-specific genes within zebrafish. *BMC Genomics* 14:65. doi: 10.1186/1471-2164-14-65

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13134

Zhang, X., Ran, J., and Mi, J. (2019). "An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic," in *Paper presented at the international conference on computer science and network technology*, (New Jersey: IEEE).

Zhou, Z. H., and Liu, X. Y. (2006).). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Know. Data Engin.* 18, 63–77. doi: 10.1109/Tkde.2006.17

Zhu, Y., Shen, X., and Pan, W. (2009). Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics.* 10:S21. doi: 10.1186/1471-2105-10-S1-S21

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 October 2020
doi: 10.3389/fgene.2020.594796

QIMCMDA: MiRNA-Disease Association Prediction by q-Kernel Information and Matrix Completion

Lin Wang[1], Yaguang Chen[1], Naiqian Zhang[1], Wei Chen[1], Yusen Zhang[1] and Rui Gao[2]**

[1] School of Mathematics and Statistics, Shandong University, Jinan, China, [2] School of Control Science and Engineering, Shandong University, Jinan, China

Studies have shown that microRNAs (miRNAs) are closely associated with many human diseases, but we have not yet fully understand the role and potential molecular mechanisms of miRNAs in the process of disease development. However, ordinary biological experiments often require higher costs, and computational methods can be used to quickly and effectively predict the potential miRNA-disease association effect at a lower cost, and can be used as a useful reference for experimental methods. For miRNA-disease association prediction, we have proposed a new method called Matrix completion algorithm based on q-kernel information (QIMCMDA). We use fivefold cross-validation and leave-one-out cross-validation to prove the effectiveness of QIMCMDA. LOOCV shows that AUC can reach 0.9235, and its performance is significantly better than other commonly used technologies. In addition, we applied QIMCMDA to case studies of three human diseases, and the results show that our method performs well in inferring potential interaction between miRNAs and diseases. It is expected that QIMCMDA will become an excellent supplement in the field of biomedical research in the future.

Keywords: microRNA-disease interaction, association prediction, heterogeneous omics data, q-kernel neighborhood similarity, matrix factorization

OPEN ACCESS

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

Reviewed by:
Qi Zhao,
University of Science and Technology Liaoning, China
Pingjian Ding,
University of South China, China
Cheng Liang,
Shandong Normal University, China

****Correspondence:***
Yusen Zhang
zhangys@sdu.edu.cn
Rui Gao
gaorui@sdu.edu.cn

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

Received: *14 August 2020*
Accepted: *21 September 2020*
Published: *22 October 2020*

Citation:
Wang L, Chen Y, Zhang N, Chen W, Zhang Y and Gao R (2020) QIMCMDA: MiRNA-Disease Association Prediction by q-Kernel Information and Matrix Completion. Front. Genet. 11:594796. doi: 10.3389/fgene.2020.594796

INTRODUCTION

MicroRNAs (miRNAs) are a type of single-stranded small non-coding RNA (∼22 nt) that play an important role in gene regression by interfering with post-transcriptional regulation (Filipowicz et al., 2008; Bartel, 2009). Lee et al. (1993) discovered the first miRNA lin-4 in *Caenorhabditis elegans*, and since then, 1000s of currently annotated miRNAs have been found in various species from plants, animals to viruses (Jopling et al., 2005; Kozomara and Griffiths-Jones, 2011). More and more evidence have shown that miRNA is an important component in cells and may play an important role in a variety of biological processes including cell growth (Ambros, 2003), immune response (Taganov et al., 2006), cell proliferation and differentiation (Chen et al., 2004, 2006), cell development, cell cycle regulation (Carleton et al., 2007), inflammation (Urbich et al., 2008), apoptosis (Petrocca et al., 2008), and stress response (Leung and Sharp, 2010). Many studies have shown that miRNA abnormalities are associated with various human diseases, such as cancer, Alzheimer's disease, and diabetes (Iorio et al., 2005; Nunez-Iglesias et al., 2010; Catto et al., 2011; Guay et al., 2011; Farazi et al., 2013). For example, there is evidence that MicroRNA-155 regulates colon

cancer cell proliferation, cell cycle, apoptosis, migration, and targets CBL (Yu et al., 2017). miR-21 negatively regulates Pdcd4 and inhibits TPA-induced tumor transformation (Asangani et al., 2008). MicroRNA-494 has become a major epigenetic regulator in aggressive human hepatocellular carcinoma neoplasms (Chuang et al., 2005). miR-146a is a tumor suppressor that inhibits NF-κB activity related to the promotion and inhibition of tumor growth (Li et al., 2014b). This makes miRNAs increasingly recognized as key regulators in gene expression (Niu et al., 2019). Finding the association of miRNA-disease is an important field of biomedicine. It not only helps humans understand the mechanism of diseases, but also helps the discovery, prognosis, diagnosis, treatment, and prevention of human complex diseases (Calin and Croce, 2006; Tricoli and Jacobson, 2007; Cho, 2010; Jiang et al., 2010).

However, the identification of miRNA-disease associations using traditional biological methods is often costly (Chen et al., 2018). Therefore, the use of mathematical and computational tools to predict potential miRNA-disease associations based on various experimentally validated association datasets is a hot issue. Through the integration and collection of data from a large number of biological experiments, there are now multiple databases related to miRNA-disease relationships such as HMDD and dbDEMC (Lu et al., 2008; Yang et al., 2010; Li et al., 2014a). In recent years, a large number of miRNA-disease association prediction methods have been proposed. For instance, Chen and Yan (2014) proposed a regularized least squares model (RLSMDA) to predict miRNA-disease associations. This model is a semi-supervised model that learns in the miRNA space and disease space respectively, and then combines to get the final prediction score. However, it should be pointed out that the parameter selection of this model is more difficult, and the combined form of the two spatial scores can be improved in the end. Xu et al. (2011) proposed a method based on support vector machine (SVM) to predict the interaction between miRNA and the disease. However, the current database rarely provides data for non-cancer miRNAs. Therefore, the main problem of the model is the lack of negative samples, which will make the supervised learning model unsuitable for the prediction of large-scale disease-miRNA interactions. Obtaining large numbers of negatively associated samples is still difficult (Guan et al., 2020). Chen et al. (2012) adopted restart random walk (RWRMDA) to predict the potential miRNA-disease interaction, which restarted the known miRNA-disease interaction network, using random walks on miRNA functional similarity network to predict potential miRNA-disease interaction. However, this method is not applicable to the prediction of new diseases that are not related to any miRNA. Chen (2018) introduced the induction matrix completion model (IMCMDA) for the prediction of miRNA disease association based on the known miRNA-disease association matrix, miRNA functional similarity and disease semantic similarity matrix. However, this method is too sensitive to the noise in the data, which affects its performance. Chen et al. (2016b) introduced the model of Within and Between Score for MiRNA-Disease Association prediction (WBSMDA) by a combination of integrated similarity and known miRNA-disease associations. Chen et al. (2018) introduced the MiRNA-disease association prediction (TLHNMDA) model based on three-layer heterogeneous network inference, which integrates multi-level data about miRNA, disease, lncRNA and their associated information into three layers heterogeneous network to determine the relationship between miRNA and disease Potential biological connection. Zhao et al. (2018) proposed a novel computational model of Symmetric Non-negative Matrix Factorization for MiRNA-Disease Association prediction (SNMFMDA) to reveal the relation of miRNA-disease pairs. Compared to the direct use of the integrated similarity in previous computational models, the integrated similarity needs to be interpolated by symmetric non-negative matrix factorization (SymNMF) before application in SNMFMDA. Jihwan Ha et al. (2020) present IMIPMF, a novel method for predicting miRNA–disease associations using probabilistic matrix factorization (PMF), which is a machine learning technique that is widely used in recommender systems. Zhu et al. (2020) proposed a new computational model based on biased heat conduction for MiRNA-Disease Association prediction (BHCMDA),which can achieve the AUC of 0.8890 in LOOCV.

We hope to use a simple and effective method for prediction. Here, we proposed a new matrix completion algorithm based on the q-kernel function to predict new miRNA disease associations (QIMCMDA). This model used miRNA q-kernel similarity, disease q-kernel similarity, known miRNA disease associations, and miRNA functional similarity. A matrix decomposition algorithm based on KL divergence was used to complement missing miRNA-disease associations. Here we used the receiver operating characteristic (ROC) curve as an evaluation index to evaluate the effectiveness of QIMCMDA. For known miRNA-disease associations downloaded from HMDD V2.0, the relevant data was cross-validated using the method of leave-one-out cross-validation (LOOCV) and fivefold cross-validation, and compared with the four previous classic methods (TLHNMDA, WBSMDA, RLSMDA, and IMCMDA). In addition, case studies were conducted on three common human diseases (Breast Neoplasms, Carcinoma Hepatocellular, Colon Neoplasms). All candidate miRNAs for these three diseases were ranked according to the predicted scores of QIMCMDA. Then the top 50 predicted miRNAs of these three diseases were verified in dbDEMC and HMDD 3.2 respectively. As a result, 46, 45, and 48 of the top 50 potentially relevant miRNAs for the three diseases were confirmed. These results indicated the effectiveness of QIMCMDA in predicting potential miRNA-disease associations.

MATERIALS AND METHODS

Human MiRNA-Disease Associations

In this study, we used human disease-miRNA associations in the HMDD v2.0 database, the dataset contains 383 diseases, 495 miRNAs, and 5430 high-quality experimentally verified human miRNA-diseases associations (Chen et al., 2018). We defined the

adjacency matrix $A \in R^{nd^* nm}$ as follows:

$$A\left(d\left(i\right), m\left(j\right)\right) = \begin{cases} 1 & diseased\ (i)\ has\ association\ with\ miRNA\ m(j) \\ 0 & diseased\ (i)\ has\ no\ association\ with\ miRNA\ m(j) \end{cases} \quad (1)$$

MiRNA Functional Similarity

MiRNA functional similarity score was calculated by Wang et al. (2010) based on the hypothesis that similarly functional miRNAs tend to be associated with diseases with similar phenotypes. Thanks to their work, we obtained from http://www.cuilab.cn/files/images/cuilab/misim.zip downloaded the data. We constructed a matrix *FS*, where the matrix $FS(m(i), m(j))$ represents the functional similarity between miRNAs$m(i)$and $m(j)$.

Disease Semantic Similarity

Disease Semantic Similarity 1

A Directed Acyclic Graph (*DAG*) was constructed to describe a disease based on the MeSH descriptors downloaded from the National Library of Medicine (Lipscomb, 2000). The *DAG* of disease *D* included not only the ancestor nodes of *D* and *D* itself but also the direct edges from parent nodes to child nodes. The semantic score of disease *D* could be defined by the following equation:

$$DV1(D) = \sum_{d \in T(D)} D1_D(d) \quad (2)$$

we defined the contribution score of disease *d* in *DAG(D)* to the disease *D* by:

$$\begin{cases} D1_D(d) = 1 & if\ d = D \\ D1_D(d) = max\left\{\triangle^* D1_D(d') | d' \in\ children\ of\ d\right\} & if\ d \neq D \end{cases} \quad (3)$$

Δ is the semantic contribution factor. The contribution score of disease is decreased as the distance between *D* and other diseases increases. Based on the assumption that two diseases with larger shared area of their *DAGs* may have greater similarity score, the semantic similarity score between disease $d(i)$ and disease $d(j)$ could be defined by the following equation:

$$SS1(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D1_{(d(i))}(t) + D1_{d(j)}(t))}{DV1(d(i)) + DV1(d(j))} \quad (4)$$

Disease Semantic Similarity 2

From above formula (3), it is easy to see that the diseases in the same layer of *DAG(D)* will make the same contribution to the semantic value of *D*. Moreover, for diseases in the same layer of *DAG(D)*, it is reasonable to assume that the diseases appeared in fewer *DAGs* will be more specific than those diseases appeared in more *DAGs*. Hence, to protrude the contribution of these more specific diseases, the contribution of the node d in *T(D)* to the semantic value of the disease *D* could be obtained according to the following formula as well (Chen, 2018):

$$D2_D(d) = -\log\left[\frac{\text{the number of DAGs containing d}}{\text{the number of diseases}}\right] \quad (5)$$

Based on the above formula, the semantic value of the disease *D* could be obtained according to the following formula as well:

$$DV2(D) = \sum_{d \in T(D)} D2_D(d) \quad (6)$$

Hence, the semantic similarity between two diseases d_i and d_j could be obtained according to the following formula as well:

$$SS1(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D2_{(d(i))}(t) + D2_{d(j)}(t))}{DV2(d(i)) + DV2(d(j))} \quad (7)$$

q-Kernel Similarity

Many contributions indicate that the performance of kernel-based learning algorithms largely depends on the choice of kernel (Chapelle et al., 2002; Lanckriet et al., 2002; Nogayama et al., 2003). Boughorbel also proved through experiments that in some applications, kernels with only positive conditions may be better than most classical kernels (Boujemaa et al., 2005). Based on this theory, Zhang et al. (2019) designed a variety of q-Kernel Functions, such as Non-Linear q-Kernel, Gaussian q-Kernel, Laplacian q-Kernel, Rational Quadratic q-Kernel, Multiquadric q-Kernel, Inverse Multiquadric q-Kernel, Wave q-Kernel, and so on. A q-analog is a mathematical expression parameterized by a quantity q that generalizes a known expression and reduces to the known expression. Therefore, after a long period of trial, we have chosen the inverse quadratic square q kernel function as the main method for calculating similarity.

Here we introduce a q-Kernel function (inverse multiquadric q-Kernel) and construct a q-Kernel similarity. Based on the assumption that similar miRNAs are more likely to exhibit interactions with similar diseases and vice versa. The q-Kernel similarity is used to calculate the kernel similarity of miRNA and disease, respectively, based on known miRNA- diseases. The value range of the two parameters c and q of the function is between 0 and 1.

$$H_q(x, y) = \frac{1}{1-q}\left(q^{-\frac{1}{c}} - q^{-\frac{1}{\sqrt{||x-y||^2 + c^2}}}\right) \quad (8)$$

Similarity Calculation of miRNA Based on q-Kernel

In previous work, we obtained a similarity network between two miRNAs. But the integrity of this network is only 0.2058, and too many missing values make it impossible for us to use this network directly. Here, the q-kernel function is used to complete the matrix. First, the obtained q-kernel distance needs to be normalized and scaled to [0,1], because the similarity network value of the previous miRNA is between [0,1]. Then we used the 1-H_q to convert the kernel distance into the similarity and a q-kernel similarity network of miRNA is obtained, which is called QM. The similarity of MiRNA is constructed as follows:

$$S_m\left(m\left(i\right), m\left(j\right)\right) \begin{cases} \omega FS\left(m\left(i\right), m\left(j\right)\right) + (1-\omega) QM(m(i), & \\ \qquad m(j)) & m\left(i\right)\ and\ m\left(j\right)\ has\ similarity \\ QM\left(m\left(i\right), m\left(j\right)\right) & otherwise \end{cases} \quad (9)$$

The ω is a weighting parameter defined as limiting the effect of FS and QM on miRNA similarity. Set ω to 0.01 through training. The greater similarity between miRNAs, the more similar the miRNAs are.

Network Similarity Calculation for Diseases Based on q-Kernel

We used the same method as the miRNA similarity network to build the disease similarity network QD. Then integrated QD with disease semantic similarities *SS1* and *SS2*:

$$S_d\left(d\left(i\right), d\left(j\right)\right) = \begin{cases} \omega SS(d(i), d(j)) + (1-\omega)QD(d(i), d(j)) \\ \qquad\qquad d(i) and\ d(j) \\ \qquad\qquad has\ similarity \\ QD(d\left(i\right), d(j)) \qquad\quad otherwise \end{cases} \quad (10)$$

$$SS(d(i), d(j)) = \frac{SS1(d(i), d(j)) + SS2(d(i), d(j))}{2} \quad (11)$$

We set the parameter values of c and q through training, that is, c = 0.1 and q = 0.6. Finally, we obtained two kernel similarity matrices, S_m and S_d .

Matrix Completion

After integrated various known data and similarity calculations of q-kernel, we can obtain human miRNA-disease correlation matrix A (Matrix density is 0.028), disease similarity matrixS_d, miRNA similarity matrix S_m. Our goal is to deduce undiscovered miRNA-disease associations based on this known information. Here we use $S_d \in R^{nd*nd}$ as the feature matrix of nd diseases, and $S_m \in R^{nm*nm}$ as the feature matrix for miRNAs. $S_d(i)$denote the feature vector of disease $d(i)$, and $S_m(j)$ denote the feature vector of miRNA $m(j)$. The main idea of QIMCMDA is to complement the two feature matrices S_d and S_m by the similarity of the q-kernel, and then supplement the missing elements under the restriction of the association matrix A to obtain the potential associations. Finally, the recovery matrix Z is obtained, and the form of Z is $Z = S_d W H^T S_m$. where $W \in R^{nd*r}$ and$H \in R^{r*nm}$, r is the desired rank which is equal tomin$(rank(W), rank(H))$. The parameter r mainly affects the convergence speed of the algorithm, and has little effect on the results. The matrices W and H can be obtained as a solution to the following optimization problems.

$$\min_{W,H} \emptyset = \sum_{i=1}^{nd}\sum_{j=1}^{nm}(A_{ij}\ln\frac{A_{ij}}{S_d*W*H*S_m} - A_{ij} + (S_d * W*H*S_m)_{ij})$$

$$s.t. W \geq 0, H \geq 0 \quad (13)$$

W and H were set to random dense matrices, and then the alternating gradient descent method is used to update iterations W and H.

$$W \leftarrow \frac{W^*\left[(\frac{S_d*A}{S_d*W*H*S_m})*S_m*H'\right]}{S_d*ONES*S_m*H'} \quad (14)$$

$$H \leftarrow \frac{H^*\left[W'*S_d*(\frac{A^*S_m}{S_d*W*H*S_m})\right]}{W'*S_d*ONES*S_m} \quad (15)$$

Through the alternating gradient descent algorithm, W and H will stabilize and stop the iteration after reaching the maximum number of iterations. Here, the maximum number of iterations is set to 100. *ONES* is a matrix, all its elements are 1. It is used to multiply two matrixes of different ranks. We can use W and H to calculate the predicted score between disease $d(i)$ and miRNA $m(j)$ by the following formula (Symbol meaning can refer to **Table 1**).

$$Score(d(i), m(j)) = S_d(i)WHS_m(j) \quad (16)$$

The specific implementation process of QIMCMDA is shown in **Figure 1**.

TABLE 1 | Notations.

Symbol	Description
nm	number of miRNAs
nd	number of diseases
$A \in R^{nd*nm}$	miRNA-diseases associations matrix
$S_m \in R^{nm*nm}$	miRNA similarity matrix
$S_d \in R^{nd*nd}$	disease similarity matrix
$W \in R^{nd*r}$	alternating iteration matrix in matrix factorization
$H \in R^{r*nm}$	alternating iteration matrix in matrix factorization

RESULTS

We used 5,430 miRNA-disease associations from HMDD v2.0 as the gold standard dataset, and we used LOOCV and fivefold CV to test the effectiveness of QIMCMDA. In addition, QIMCMDA will be compared with four other methods IMCMDA (Chen, 2018), RLSMDA (Chen and Yan, 2014), TLHNMDA (Chen et al., 2018), WBSMDA (Chen et al., 2016b) to evaluate the predictive ability of QIMCMDA (see **Table 2**). In the framework of the LOOCV evaluation, 5430 miRNA-disease associations in the data set are considered as test samples one by one, the other remaining samples are considered as training samples, and samples with unknown associations are considered as candidate samples. Through the calculation of the model, we can obtain the prediction score, and then rank and record according to the prediction score. The process of fivefold CV is similar to LOOCV. The miRNA-disease association of the golden data set was randomly divided into five groups, one of which was selected as the test set in turn, and the rest as the training set. Candidate sample settings are the same as LOOCV. Then rank and record the predicted scores for each test sample. **Figure 2** shows a comparison of the prediction performance based on the overall AUC value of LOOCV. As a result of LOOCV, the AUC of QIMCMDA is 0.9235, and the AUC values obtained by IMCMDA, RLSMDA, TLHNMDA and WBSMDA are 0.8378, 0.8193, 0.8795, 0.8010, respectively. For fivefold QIMCMDA, IMCMDA, RLSMDA, TLHNMDA and WBSMDA 10 times were performed, and the average AUC and standard deviation were recorded as 0.9170 ± 0.0006, 0.8311 ± 0.0006, 0.7814 ± 0.0020, 0.8735 ± 0.0010,0.7980 ± 0.0009, respectively (see **Figure 3**).

Parameter Analysis

There are several hyper-parameters in QIMCMDA that need to be tuned, i.e., c, q, w, k. We use a random search strategy

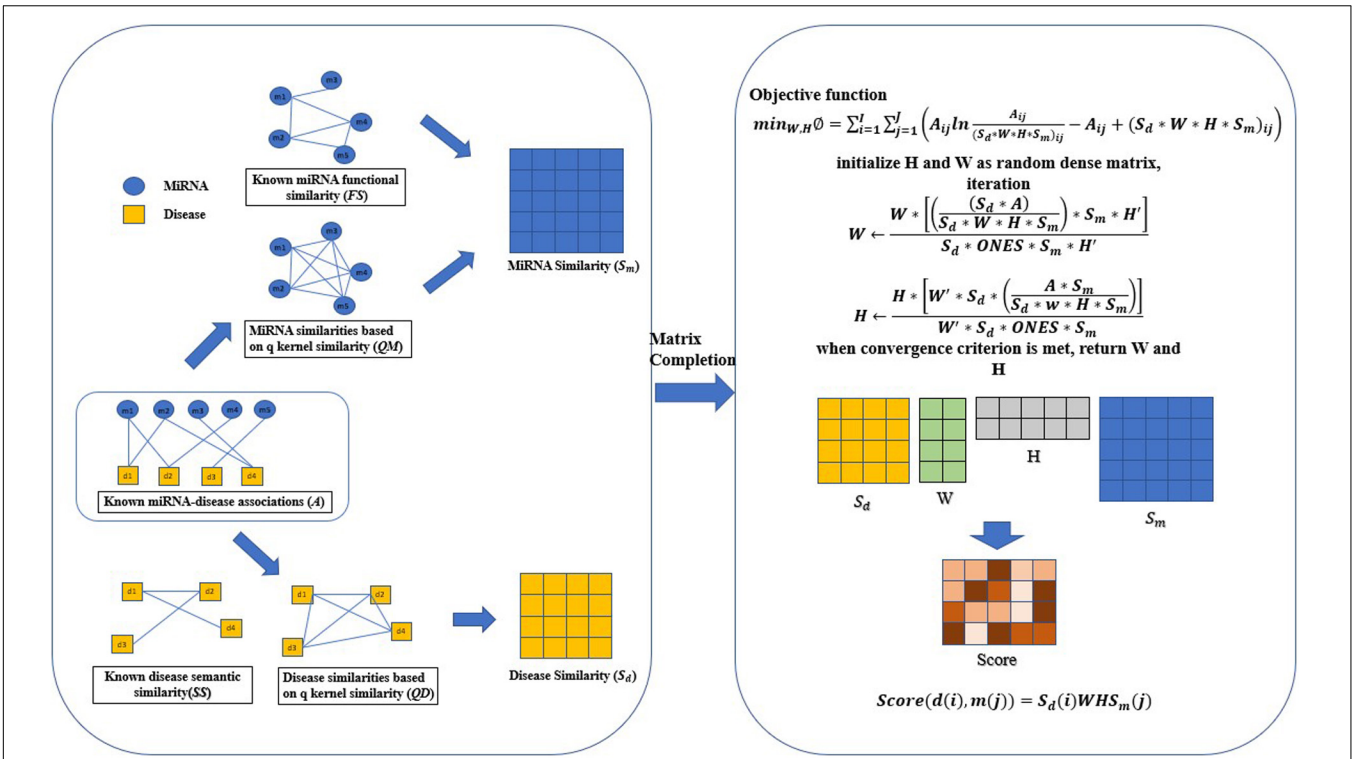

FIGURE 1 | Flowchart of QIMCMDA model to infer the potential miRNA-disease associations. ONES is an all-ones matrix of rank nd*nm.

TABLE 2 | Under the fivefold CV and LOOCV verification framework, the performance of QIMCMDA and other benchmark methods.

Methods	LOOCV	Fivefold CV
QIMCMDA	0.9235	0.9170 ± 0.0006
IMCMDA	0.8378	0.8311 ± 0.0006
RLSMDA	0.8193	0.7814 ± 0.0020
TLHNMDA	0.8795	0.8735 ± 0.0010
WBSMDA	0.8010	0.7980 ± 0.0009

to select hyper-parameters from fixed ranges (Zhang et al., 2020). c and q are parameters for adjusting the q-Kernel function. In this study, the value of c is selected from {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1}, and the value of q is selected from {0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9}. q can't be equal to 1. ω is the weight parameter used to integrate similarity. Here, ω is selected from {0.01,0.05,0.1,0.15,0.2,0.3,0.4,0.5,0.8,1}. Next, we show the influence of the these parameters under the fivefold CV.

The k is a potential feature size. In our test, the impact of this variable is actually very small, but we still decided to use PCA to calculate the cumulative contribution rate to obtain the most appropriate k value. This method is in the paper by Wang et al. (2017). It has been well-verified. In this article, the cumulative contribution rate of 95% is used to select the PC, and the final k is set 114.

ω is a weight parameter used to integrate the similarity matrix. **Figure 4** shows the effect of changes in ω on AUC when other parameters are fixed. When ω = 0.01, AUC takes the maximum value. When c = 0.1, q = 0.6, the model can achieve the best effect (see **Figure 5**).

Case Study

In this article, we used case studies to further demonstrate the effectiveness of QIMCMDA. We performed case studies on three diseases: Breast Neoplasms, Carcinoma Hepatocellular, and Colon Neoplasms. These diseases were selected in our case study because they all have high incidence and insignificant early symptoms. In addition, they have been considered as case studies in many previous publications (Guan et al., 2020). Our case study used HMDD v2.0 as the training database for QIMCMDA. HMDD 3.2 and dbDEMC (Lu et al., 2008; Yang et al., 2010; Li et al., 2014a) serve as validation databases to confirm the predicted potential associations. Compared with the previous 2.0 version, the 3.2 version contains more than double the association between human diseases and miRNAs, the classification of evidence is more clear, and there is a clear third-party annotation for each association. The differentially expressed miRNA database (dbDEMC) in human cancer is a comprehensive database microRNA (miRNA) designed to store and display differentially expressed human cancers detected by high-throughput methods. The database collected a total of 209 newly released data sets from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). The current version contains data from 436 biological experiments, including 2224 differentially expressed miRNAs in

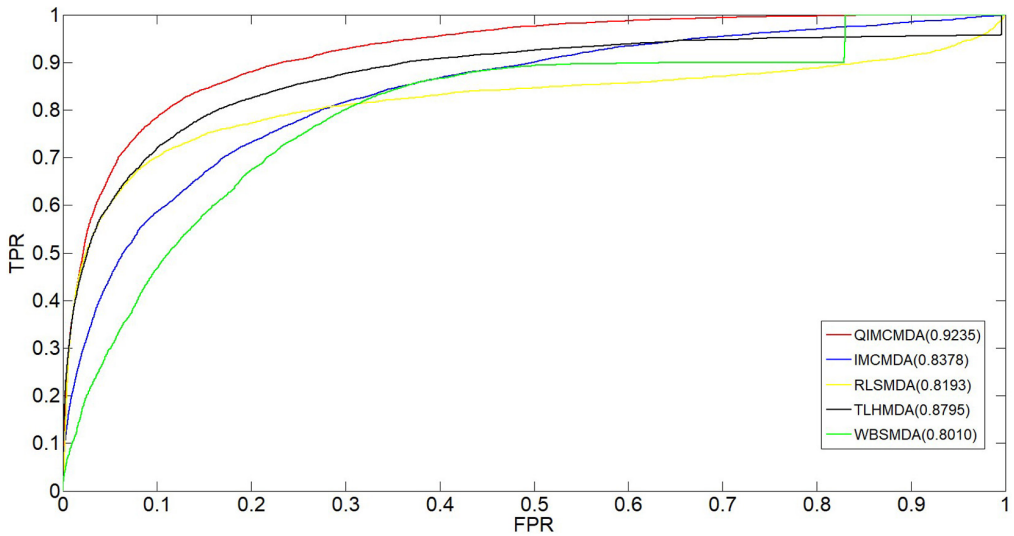

FIGURE 2 | Performance comparison between QIMCMDA and other benchmark methods (RLSMDA, IMCMDA, TLHNMDA, WBSMDA) on AUC of LOOCV.

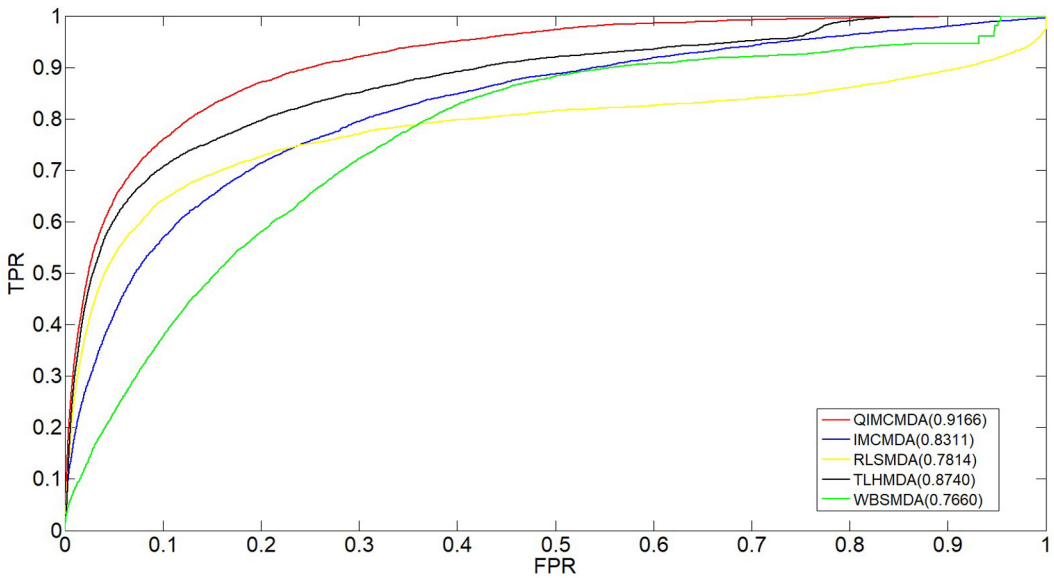

FIGURE 3 | Performance comparison between QIMCMDA and other benchmark methods (RLSMDA, IMCMDA, TLHNMDA, WBSMDA) on AUC of fiveflod CV.

36 cancer types. We only perform ranking verification on candidate miRNAs of interest, so training samples are not in the final result. In other words, the miRNA disease associations obtained from the predicted list do not overlap with the known 5430 associations.

Breast Neoplasms is one of the most common malignancies in women. With more than 2 million new cases worldwide each year, it ranks second among the world's major cancer types (Jemal et al., 2017). More than half of these cases occurred in industrialized countries (Parkin et al., 2005). It was one of the leading causes of death among women aged 20–59 (Siegel et al., 2015). With the development of biological technology, researchers have found more miRNAs related to Breast Neoplasms. Our results are supported by third-party annotations in two databases, HMDD3.2 and dbDEMC. For example, miR-150 and miR-372 can promote the proliferation and growth of Breast Neoplasms cells by targeting the pro-apoptotic purinergic P2X7 receptor and LATS2 respectively (Huang et al., 2017; Cheng et al., 2018). MicroRNA-130a targets RAB5A to inhibit the proliferation, invasion and migration of Breast Neoplasms cells (Pan et al., 2015). miR-494 targets CXCR4 through the Wnt/β-catenin signaling pathway, thereby inhibiting Breast Neoplasms progression *in vitro* (Song et al., 2015). The increased miR-451 expression may negatively regulate Bcl-2 mRNA and protein expression, which in turn affects caspase 3 protein expression and accelerates Breast Neoplasms cell apoptosis (Gu et al., 2015). MiR-449a inhibits cell migration and invasion in Breast Neoplasms by targeting PLAGL2

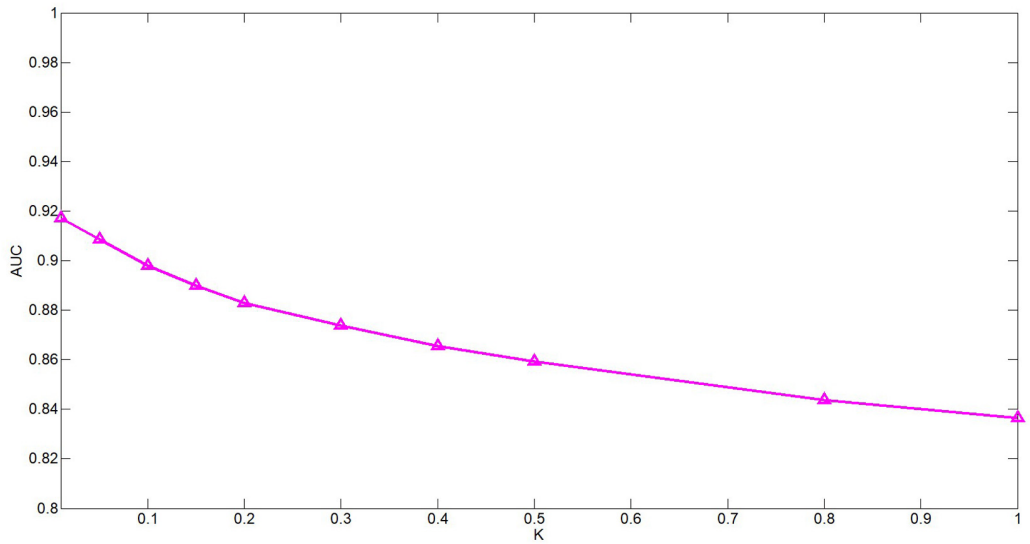

FIGURE 4 | Performance of QIMCMDA with different values of ω under fivefold CV.

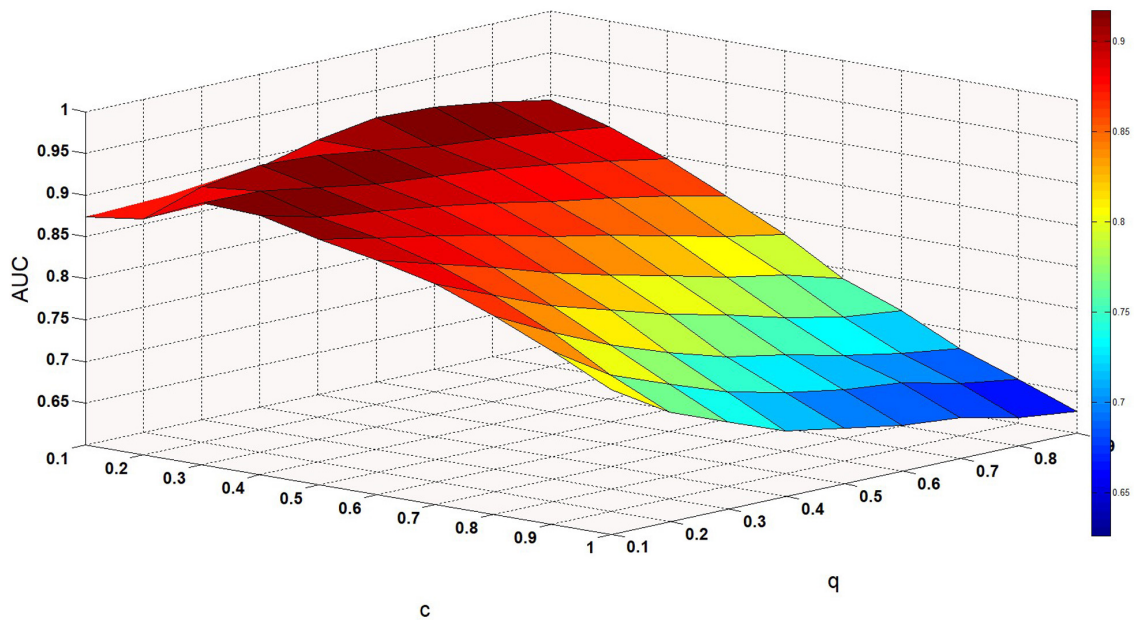

FIGURE 5 | Performance of QIMCMDA with different values of c and q under fivefold CV.

(Wang et al., 2018). We selected the top 50 in the results and verified them with two databases, HMDD 3.2 and dbDEMC. It was found that 10 of the first 10 predictions and 46 miRNAs of the first 50 predictions were verified (see **Table 3**).

Hepatocellular carcinoma (HCC), one of the most common malignancies worldwide (Yegin et al., 2016), was also the main cause of cancer in men under 60 in China (Chen et al., 2016a). MiRNAs have important roles in the treatment of HCC and have been corroborated. For example, related *in vitro* experiments have further confirmed the anti-tumor effect of miR-132 in HCC (Liu et al., 2015; Zhang et al., 2016). The newly identified miR-429-CRKL axis represents a new potential therapeutic target for HCC therapy (Guo et al., 2018). MicroRNA-23b inhibits epithelial–mesenchymal transition (EMT) and metastasis of Hepatocellular Carcinoma by targeting Pyk2 (Cao et al., 2017). MicroRNA-494 is a major epigenetic regulator of microRNAs for multiple invasion inhibitors by targeting 10 11 translocation 1 in aggressive human Hepatocellular Carcinoma (Chuang et al., 2005). MicroRNA-340 inhibits the proliferation and invasion of Hepatocellular Carcinoma cells by targeting JAK1 (Yuan et al., 2017). Therefore, 10 of the top 10 predicted miRNAs

TABLE 3 | Prediction results of the top 50 predicted Breast Neoplasms-related miRNAs based on known associations in HMDD V2.0.

miRNA	Evidence	miRNA	Evidence
hsa-mir-151	HMDD3.2	hsa-mir-663	dbDEMC
hsa-mir-30e	HMDD3.2	hsa-mir-382	dbDEMC
hsa-mir-92b	HMDD3.2	hsa-mir-494	HMDD3.2
hsa-mir-451	HMDD3.2	hsa-mir-575	HMDD3.2
hsa-mir-130a	HMDD3.2	hsa-mir-658	dbDEMC
hsa-mir-192	HMDD3.2	hsa-mir-181d	dbDEMC
hsa-mir-98	HMDD3.2	hsa-mir-376a	HMDD3.2
hsa-mir-372	HMDD3.2	hsa-mir-211	dbDEMC
hsa-mir-32	HMDD3.2	hsa-mir-484	HMDD3.2
hsa-mir-106a	HMDD3.2	hsa-mir-455	Unconfirmed
hsa-mir-130b	HMDD3.2	hsa-mir-432	dbDEMC
hsa-mir-99b	dbDEMC	hsa-mir-381	HMDD3.2
hsa-mir-95	dbDEMC	hsa-mir-99a	HMDD3.2
hsa-mir-28	dbDEMC	hsa-mir-154	dbDEMC
hsa-mir-150	HMDD3.2	hsa-mir-523	dbDEMC
hsa-mir-186	dbDEMC	hsa-mir-526b	HMDD3.2
hsa-mir-15b	HMDD3.2	hsa-mir-507	Unconfirmed
hsa-mir-142	HMDD3.2	hsa-mir-525	Unconfirmed
hsa-mir-449b	dbDEMC	hsa-mir-660	HMDD3.2
hsa-mir-198	dbDEMC	hsa-mir-181c	HMDD3.2
hsa-mir-196b	HMDD3.2	hsa-mir-300	dbDEMC
hsa-mir-491	HMDD3.2	hsa-mir-297	dbDEMC
hsa-mir-449a	HMDD3.2	hsa-mir-136	dbDEMC
hsa-mir-424	HMDD3.2	hsa-mir-331	HMDD3.2
hsa-mir-212	HMDD3.2	hsa-mir-512	Unconfirmed

TABLE 4 | Prediction results of the top 50 predicted Carcinoma Hepatocellular-related miRNAs based on known associations in HMDD V2.0.

miRNA	Evidence	miRNA	Evidence
hsa-mir-132	HMDD3.2	hsa-mir-516a	unconfirmed
hsa-mir-429	HMDD3.2	hsa-mir-663	dbDEMC
hsa-mir-34b	HMDD3.2	hsa-mir-340	HMDD3.2
hsa-mir-151	HMDD3.2	hsa-mir-28	dbDEMC
hsa-mir-30e	HMDD3.2	hsa-mir-186	HMDD3.2
hsa-mir-367	HMDD3.2	hsa-mir-575	HMDD3.2
hsa-mir-339	dbDEMC	hsa-mir-658	dbDEMC
hsa-mir-9	HMDD3.2	hsa-mir-452	HMDD3.2
hsa-mir-215	HMDD3.2	hsa-mir-193b	HMDD3.2
hsa-mir-451	HMDD3.2	hsa-mir-196b	dbDEMC
hsa-mir-194	HMDD3.2	hsa-mir-494	HMDD3.2
hsa-mir-302a	dbDEMC	hsa-mir-449a	HMDD3.2
hsa-mir-32	HMDD3.2	hsa-mir-424	HMDD3.2
hsa-mir-204	HMDD3.2	hsa-mir-520c	HMDD3.2
hsa-mir-135b	HMDD3.2	hsa-mir-382	unconfirmed
hsa-mir-95	HMDD3.2	hsa-mir-301b	dbDEMC
hsa-mir-488	dbDEMC	hsa-mir-510	unconfirmed
hsa-mir-302d	HMDD3.2	hsa-mir-376c	unconfirmed
hsa-mir-23b	HMDD3.2	hsa-mir-455	HMDD3.2
hsa-mir-133a	HMDD3.2	hsa-mir-206	HMDD3.2
hsa-mir-299	HMDD3.2	hsa-mir-137	HMDD3.2
hsa-mir-143	HMDD3.2	hsa-mir-211	HMDD3.2
hsa-mir-153	HMDD3.2	hsa-mir-154	HMDD3.2
hsa-mir-516b	Unconfirmed	hsa-mir-27b	HMDD3.2
hsa-mir-383	dbDEMC	hsa-mir-523	dbDEMC

TABLE 5 | Prediction results of the top 50 predicted Colon Neoplasms-related miRNAs based on known associations in HMDD V2.0.

miRNA	Evidence	miRNA	Evidence
hsa-mir-143	HMDD3.2	hsa-mir-200b	HMDD3.2
hsa-mir-106b	HMDD3.2	hsa-mir-24	HMDD3.2
hsa-mir-21	HMDD3.2	hsa-mir-1	HMDD3.2
hsa-mir-128	HMDD3.2	hsa-mir-205	HMDD3.2
hsa-mir-18a	HMDD3.2	hsa-mir-29b	HMDD3.2
hsa-mir-9	dbDEMC	hsa-let-7b	HMDD3.2
hsa-mir-155	HMDD3.2	hsa-mir-31	HMDD3.2
hsa-mir-181a	HMDD3.2	hsa-mir-223	HMDD3.2
hsa-mir-494	unconfirmed	hsa-let-7c	HMDD3.2
hsa-mir-483	HMDD3.2	hsa-mir-15a	HMDD3.2
hsa-let-7a	HMDD3.2	hsa-mir-200c	HMDD3.2
hsa-mir-125b	HMDD3.2	hsa-mir-222	HMDD3.2
hsa-mir-146a	HMDD3.2	hsa-mir-199a	HMDD3.2
hsa-mir-34a	HMDD3.2	hsa-mir-30b	HMDD3.2
hsa-mir-210	HMDD3.2	hsa-mir-141	HMDD3.2
hsa-mir-16	HMDD3.2	hsa-mir-200a	HMDD3.2
hsa-mir-146b	dbDEMC	hsa-let-7e	HMDD3.2
hsa-mir-221	HMDD3.2	hsa-mir-196a	HMDD3.2
hsa-mir-93	HMDD3.2	hsa-mir-142	HMDD3.2
hsa-mir-92a	HMDD3.2	hsa-let-7f	HMDD3.2
hsa-mir-20b	dbDEMC	hsa-mir-34c	Unconfirmed
hsa-mir-19a	HMDD3.2	hsa-let-7i	HMDD3.2
hsa-mir-29a	HMDD3.2	hsa-let-7d	HMDD3.2
hsa-mir-18b	HMDD3.2	hsa-let-7g	HMDD3.2

and 45 of the top 50 predicted miRNAs were confirmed by experimental literature from the dbDEMC and HMDD3.2 (see **Table 4**).

Colon Neoplasms are the most common type of gastrointestinal cancer (Jemal et al., 2011; Ogata-Kawata et al., 2014). Siegel et al. (2018), there were 97,220 new cases in the United States alone, and approximately 50,630 patients died. A variety of miRNAs have been experimentally confirmed to be associated with colon neoplasms. For example, MicroRNA-155 regulates Colon Neoplasms cell proliferation, cell cycle, apoptosis, migration and targets CBL (Yu et al., 2017). MicroRNA-21 induces stem cells by down-regulating transforming growth factor beta receptor 2 (TGFbetaR2) in Colon Neoplasms cells (Yu et al., 2012). Let-7 is also involved in the development of Colon Neoplasms (Williams, 2008). MicroRNA-221 promotes Colon Neoplasms cell proliferation *in vitro* (Sun et al., 2011). MicroRNA-34a inhibits the migration and invasion of Colon Neoplasms cells by targeting Fra-1 (Wu et al., 2012). Verification of dbDEMC and HMDD3.2 confirmed 10 of the first 10 predictions and 48 miRNAs of the first 50 predictions (see **Table 5**).

DISCUSSION

Research on the potential prediction of miRNA-disease associations will help us to understand the pathogenesis and treatment of the disease more deeply. Especially for cancer, targeted therapy by regulating miRNA may be a breakthrough point for future treatment.

In this paper, we developed an algorithm for miRNA-disease association prediction (QIMCMDA), which mainly introduced the q-kernel function to complete the similarity information required. The QIMCMDA model is based on the known miRNA disease association and miRNA functional similarity network. First, calculated and completed the miRNA similarity network and the disease similarity network using the q-kernel function. Then used the matrix decomposition method to calculate the prediction score for each sample, and finally sort the scores. The AUC of QIMCMDA based on LOOCV is 0.9235, showing better performance than previous methods. In addition, experimental literature has confirmed the validity of potential miRNA-disease association predictions for three major human diseases: Breast Neoplasms, Carcinoma Hepatocellular, Colon Neoplasms).

The reasons for the reliable performance of QIMCMDA are as follows: the key advantage of QIMCMDA is that it utilizes the functional similarity of known miRNAs in combination with q-kernel similarity as features of diseases and miRNAs to complete the association of missing miRNAs and diseases. And the use of alternating gradient descent algorithm to search for the optimal solution can ensure the reliability of disease feature vectors and miRNA feature vectors. In addition, the overall complexity of our method from the construction of the network to the final prediction score calculation is low, and the operation is simple and easy to reproduce. QIMCMDA has a short running time and is suitable for large-scale data research. It is a simple and effective method. Finally, QIMCMDA is a semi-supervised model that does not require negative samples, reducing the difficulty of model construction. Compared with methods that require a large number of negative samples, our method has some advantages. However, QIMCMDA currently has some limitations. First of all, there are inevitable noises and outliers in the known materials we use. Second, QIMCMDA used the KL divergence as an error function, which is unstable due to noise and outliers. With the development of the times, database construction will become more and more perfect. As the number of associated data increases, our predictions will become more accurate. In addition, for miRNA or disease without any known associations, our method may be less effective, because the calculation of q-kernel is mainly based on known associations. In the future, we can use a large amount of biological data to further increase the reliability and practicability of the model prediction. And our method can be practiced in other fields such as the interaction between microorganisms and diseases or the interaction between drugs and targets.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

LW and YZ conceived the study. LW, YZ, and YC developed the prediction method and designed the experiments. LW analyzed the result and wrote the manuscript. NZ and WC optimized the flow chart and manuscript structure. All authors reviewed and improved the manuscript.

FUNDING

This work has been supported by the National Natural Science Foundation of China (under Grant Nos. 61877064, U1806202, and 61533011).

REFERENCES

Ambros, V. (2003). MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* 113, 673–676. doi: 10.1016/S0092-8674(03)00428-8

Asangani, I. A., Rasheed, S. A. K., Nikolova, D. A., Leupold, J. H., Colburn, N. H., Post, S., et al. (2008). MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pdcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene* 27, 2128–2136. doi: 10.1038/sj.onc.1210856

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002

Boujemaa, N., Tarel, J., and Boughorbel, S. (2005). "Conditionally positive definite kernels for svm based image recognition," in *IEEE International Conference on Multimedia and Expo(ICME)*, Amsterdam, 113–116. doi: 10.1109/ICME.2005.1521373

Calin, G. A., and Croce, C. M. (2006). MicroRNA signatures in human cancers. *Nat. Rev. Cancer.* 6, 857–866. doi: 10.1038/nrc1997

Cao, J., Liu, J. K., Long, J. Y., Fu, J., Huang, L., Li, J., et al. (2017). MicroRNA-23b suppresses epithelial-mesenchymal transition (EMT) and metastasis in hepatocellular carcinoma via targeting Pyk2. *Biomed. Pharmacother.* 17:30 doi: 10.1016/j.biopha.2017.02.030

Carleton, M., Cleary, M. A., and Linsley, P. S. (2007). MicroRNAs and cell cycle regulation. *Cell Cycle.* 6, 2127–2132. doi: 10.4161/cc.6.17.4641

Catto, J. W. F., Alcaraz, A., Bjartell, A. S., White, R. D. V., Evans, C. P., Fussel, S., et al. (2011). MicroRNA in prostate, bladder, and kidney Cancer: a systematic review. *Eur. Urol.* 59, 671–681. doi: 10.1016/j.eururo.2011.01.044

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Mach. Learn.* 46, 131–159. doi: 10.1023/A:1012450327387

Chen, C. Z., Li, L., Lodish, H. F., and Bartel, D. P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. *Science* 303:903. doi: 10.1126/science.1091903

Chen, J. F., Mandel, E. M., Thomson, J. M., Wu, Q., and Wang, D. Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nature* 38:1725. doi: 10.1038/ng1725

Chen, W., Zheng, R., Baade, P. D., Zhang, S., Zeng, H., Bray, F., et al. (2016a). Cancer statistics in China, 2015. *CA Cancer J. Clin.* 66, 115–132. doi: 10.3322/caac.21338

Chen, X. (2018). IMCMDA: predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi: 10.1093/bioinformatics/bty503

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

Chen, X., Qu, J., and Yin, J. (2018). TLHNMDA: triple layer heterogeneous network based inference for MiRNA-Disease association prediction. *Front. Genet.* 18:234. doi: 10.3389/fgene.2018.00234

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L. X., Liu, Y., et al. (2016b). WBSMDA: within and between score for MiRNA-disease association prediction. *Sci. Rep.* 6:21106. doi: 10.1038/srep21106

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

Cheng, X. Y., Chen, J. Q., and Huang, Z. (2018). MiR-372 promotes breast cancer cell proliferation by directly targeting LATS2. *Exp. Therap. Med.* 15:5761. doi: 10.3892/etm.2018.5761

Cho, W. C. (2010). MicroRNAs: potential biomarkers for cancer diagnosis, prognosis and targets for therapy. *Int. J. Biochem. Cell Biol.* 42, 1273–1281. doi: 10.1016/j.biocel.2009.12.014

Chuang, K. H., Whitney-Miller, C. L., Chu, C.-Y., Zhou, Z., Dokus, K. M., Schmit, S., et al. (2005). MicroRNA-494 is a master epigenetic regulator of multiple invasion-suppressor microRNAs by targeting ten eleven translocation 1 in invasive human hepatocellular carcinoma neoplasms. *Hepatology* 62:27816. doi: 10.1002/hep.27816

Farazi, T. A., Hoell, J. I., Morozov, P., and Tuschl, T. (2013). MicroRNAs in human cancer. *Adv. Exp. Med. Biol.* 774, 1–20. doi: 10.1007/978-94-007-5590-1_1

Filipowicz, W., Bhattacharyya, S. N., and Sonenberg, N. (2008). Mechanisms of posttranscriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* 9, 102–114. doi: 10.1038/nrg2290

Gu, X., Li, J. Y., Guo, J., Li, P. S., and Zhang, W. H. (2015). Influence of MiR-451 on drug resistances of paclitaxel-resistant breast cancer cell line. *Med. Sci. Monit.* 21:894475. doi: 10.12659/MSM.894475

Guan, N. N., Wang, C. C., Zhang, L., Huang, L., Li, J. Q., and Piao, X. (2020). In silico prediction of potential miRNA−disease association using an integrative bioinformatics approach based on kernel fusion. *J. Cell Mol. Med.* 24, 573–587. doi: 10.1111/jcmm.14765

Guay, C., Roggli, E., Nesca, V., Jacovetti, C., and Regazzi, R. (2011). Diabetes mellitus, a microRNA-related disease? *Transl. Res.* 157, 253–264. doi: 10.1016/j.trsl.2011.01.009

Guo, C. M., Zhao, D. T., Zhang, Q. L., Liu, S. Q., and Sun, M. S. (2018). MiR-429 suppresses tumor migration and invasion by targeting CRKL in hepatocellular carcinoma via inhibiting Raf/MEK/ERK pathway and epithelial-mesenchymal transition. *Sci. Rep.* 18:8 doi: 10.1038/s41598-018-20258-8

Ha, J., Park, C. H., Park, C. Y., and Park, S. (2020). IMIPMF: inferring miRNA-disease interactions using probabilistic matrix factorization. *J. Biomed. Inform.* 102:103358. doi: 10.1016/j.jbi.2019.103358

Huang, S. Y., Chen, Y. S., Wu, W., Ouyang, N. Y., Chen, J. N., Li, H. Y., et al. (2017). MiR-150 promotes human breast cancer growth and malignant behavior by targeting the pro-apoptotic purinergic P2X7 receptor. *PLoS One* 8:707. doi: 10.1371/journal.pone.0080707

Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., et al. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res.* 65:70657070. doi: 10.1158/0008-5472.CAN-05-1783

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Wardm, E., Forman, D., et al. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107

Jemal, A., Ward, E. M., Johnson, C. J., Cronin, K. A., Ma, J., Ryerson, B., et al. (2017). Annual report to the nation on the status of cancer, 1975-2014, featuring survival. *J. Natl. Cancer Inst.* 17:30. doi: 10.1093/jnci/djx030

Jiang, Q. H., Hao, Y. Y., Wang, G., Juan, L. R., Zhang, T. J., Teng, M. X., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BioMed. Central* 4:S2. doi: 10.1186/1752-0509-4-S1-S2

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 309, 1577–1581. doi: 10.1126/science.1113329

Kozomara, A., and Griffiths-Jones, S. (2011). miRbase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157. doi: 10.1093/nar/gkq1027

Lanckriet, G. R. G., Christianini, N., Bartlett, P. L., Ghaoui, L. E., and Jordan, M. I. (2002). "Learning the kernel matrix with semi-definite programming," in *Nineteenth International Conference on Machine Learning*, Sydney, 323–330. doi: 10.1023/B:JODS.0000012018.62090.a7

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75:90529. doi: 10.1016/0092-8674(93)90529-Y

Leung, A. K., and Sharp, P. A. (2010). MicroRNA functions in stress responses. *Mol. Cell.* 40, 205–215. doi: 10.1016/j.molcel.2010.09.027

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014a). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Li, Y., Zhang, Z., Mao, Y., Jin, M., Jing, F., Ye, Z., et al. (2014b). A genetic variant in MiR146a modifies digestive system Cancer risk: a meta-analysis. *Asian Pac. J. Cancer Prev.* 15, 145–150. doi: 10.7314/APJCP.2014.15.1.145

Lipscomb, C. E. (2000). Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* 88, 265–266. doi: 10.0000/PMID10928714

Liu, K., Li, X. L., Cao, Y. C., Ge, Y. Y., Wang, J. M., and Shi, B. (2015). MiR-132 inhibits cell proliferation, invasion and migration of hepatocellular carcinoma by targeting PIK3R3. *Int. J. Oncol.* 15:3112 doi: 10.3892/ijo.2015.3112

Lu, M., Zhang, Q. P., Deng, M., Miao, J., Guo, Y. H., Gao, W., et al. (2008). An analysis of human MicroRNA and disease associations. *PLoS One* 3:3420. doi: 10.1371/journal.pone.0003420

Niu, Y. W., Wang, G. H., Yan, G. Y., and Chen, X. (2019). Integrating random walk and binary regression to identify novel miRNA-disease association. *BMC Bioinformatics* 20:59. doi: 10.1186/s12859-019-2640-9

Nogayama, T., Takahashi, H., and Muramatsu, M. (2003). Generalization of kernel pca and automatic parameter tuning. *Techn. Report Ieice Prmu* 103, 43–48.

Nunez-Iglesias, J., Liu, C. C., Morgan, T. E., Finch, C. E., and Zhou, X. J. (2010). Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. *PLoS One* 5:8898. doi: 10.1371/journal.pone.0008898

Ogata-Kawata, H., Izumiya, M., Kurioka, D., Honma, Y., Yamada, Y., Furuta, K., et al. (2014). Circulating exosomal microRNAs as biomarkers of colon cancer. *PLoS One* 14:921. doi: 10.1371/journal.pone.0092921

Pan, Y. Q., Wang, R. J., Zhang, F. W., Chen, Y. L., Lv, Q. F., Long, G., et al. (2015). MicroRNA-130a inhibits cell proliferation, invasion and migration in human breast cancer by targeting the RAB5A. *Int. J. Clin. Exp. Pathol.* 8, 384–393.

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J. Clin.* 55, 74–108. doi: 10.3322/canjclin.55.2.74

Petrocca, F., Visone, R., Onelli, M. R., Shah, M. H., Nicoloso, M. S., Martino, I. D., et al. (2008). E2F1-regulated microRNAs impair TGFb-dependent cell-cycle arrest and apoptosis in gastric cancer. *Cancer Cell* 13, 272–286. doi: 10.1016/j.ccr.2008.02.013

Siegel, R. L., Kimberly D M., and Jemal, A.. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Siegel, R. L., Miller, K. D., and Jemal, A. (2015). Cancer statistics, 2015. *CA Cancer J. Clin.* 65, 5–29. doi: 10.3322/caac.21208

Song, L. Q., Liu, D., Wang, B. F., He, J. J., Zhang, Q. Q., Dai, Z. J., et al. (2015). MiR-494 suppresses the progression of breast cancer in vitro by targeting CXCR4 through the Wnt/β-catenin signaling pathway. *Oncol. Rep.* 34:3965. doi: 10.3892/or.2015.3965

Sun, K., Wang, W., Lei, S. T., Wu, C. T., and Li, G. X. (2011). MicroRNA-221 promotes colon carcinoma cell proliferation in vitro by inhibiting CDKN1C/p57 expression. *J. South. Med. Univ.* 11:2011. doi: 10.1038/cmi.2011.4

Taganov, K. D., Boldin, M. P., Chang, K. J., and Baltimore, D. (2006). NF-jB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc. Natl. Acad. Sci. U.S.A.* 103, 12481–12486. doi: 10.1073/pnas.0605298103

Tricoli, J. V., and Jacobson, J. W. (2007). MicroRNA: potential for Cancer detection, diagnosis, and prognosis. *Cancer Res.* 67, 4553–4555. doi: 10.1158/0008-5472.CAN-07-0563

Urbich, C., Kuehbacher, A., and Dimmeler, S. (2008). Role of microRNAs in vascular diseases, inflammation, and angiogenesis. *Cardiovasc Res.* 79, 581–588. doi: 10.1093/cvr/cvn156

Wang, D., Wang, J., Lu, M., and Song, F. (2010). Qinghua cui. inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26:241. doi: 10.1093/bioinformatics/btq241

Wang, H. L., Xiao, Y., Wu, L., and Ma, D. C. (2018). Comprehensive circular RNA profiling reveals the regulatory role of the circRNA-000911/miR-449a pathway in breast carcinogenesis. *Int. J. Oncol.* 18:4265. doi: 10.3892/ijo.2018.4265

Wang, W. J., Chen, X., Jiao, P. F., and Jin, D. (2017). Similarity-based regularized latent feature model for link prediction in bipartite networks. *Sci. Rep.* 7:9. doi: 10.1038/s41598-017-17157-9

Williams, A. E. (2008). Functional aspects of animal microRNAs. *Cell. Mol. Life Sci.* 8:9. doi: 10.1007/s00018-007-7355-9

Wu, J. M., Wu, G., Lv, L., Ren, Y. F., Zhang, X. J., Xue, Y. F., et al. (2012). MicroRNA-34a inhibits migration and invasion of colon cancer cells via targeting to Fra-1. *Carcinogenesis* 12:304. doi: 10.1093/carcin/bgr304

Xu, J., Li, C. X., Lv, J. Y., Li, Y. S., Huan, R., Xiao, Y., et al. (2011). Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol. Cancer Therap.* 10:55. doi: 10.1158/1535-7163.MCT-11-0055

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomic* 10:S5. doi: 10.1186/1471-2164-11-S4-S5

Yegin, E. G., Oymaci, E., Karatay, E., and Coker, A. (2016). Progress in surgical and nonsurgical approaches for hepatocellular carcinoma treatment. *Hepatobiliary Pancreat Dis. Int.* 15, 234–256. doi: 10.1016/S1499-3872(16)60097-8

Yu, H., Xu, W. L., Gong, F. C., Chi, B. R., Chen, J. Y., and Zhou, L. (2017). MicroRNA-155 regulates the proliferation, cell cycle, apoptosis and migration of colon cancer cells and targets CBL. *Exp. Therap. Med.* 14:5085. doi: 10.3892/etm.2017.5085

Yu, Y. J., Kanwar, S. S., Patel, B., Ohta, P. S., Nautiyal, J., Sarkar, F. H., et al. (2012). MicroRNA-21 induces stemness by downregulating transforming growth factor beta receptor 2 (TGFβR2) in colon cancer cells. *Carcinogenesis* 12:246. doi: 10.1093/carcin/bgr246

Yuan, J. Y., Ji, H. X., Xiao, F., Lin, Z. P., Zhao, X.J., Wang, Z. C., et al. (2017). MicroRNA-340 inhibits the proliferation and invasion of hepatocellular carcinoma cells by targeting JAK1. *Biochem. Biophys. Res. Commun.* 17:102. doi: 10.1016/j.bbrc.2016.12.102

Zhang, X., Tang, W., Chen, G., Ren, F. H., Liang, H. W., Dang, Y. W., et al. (2016). An encapsulation of gene signatures for hepatocellular carcinoma, MicroRNA-132 predicted target genes and the corresponding overlaps. *PLoS One* 16:e0159498 doi: 10.1371/journal.pone.0159498

Zhang, Y. S., Pang, D. L., Wang, J. H., and Zhang, J. L. (2019). *qkerntool: Q-Kernel-Based and Conditionally Negative Definite Kernel-Based Machine Learning Tools*. Available online at: https://cran.r-project.org/package=qkerntool (accessed April 13, 2019).

Zhang, Z. C., Zhang, X. F., Wu, M., Ou-Yang, L., Zhao, X. M., and Li, X. L. (2020). A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics (Oxf Engl)* 36:157. doi: 10.1093/bioinformatics/btaa157

Zhao, Y., Chen, X., and Yin, J. (2018). A Novel computational method for the identification of potential miRNA-disease association based on symmetric non-negative matrix factorization and kronecker regularized least square. *Front. Genet.* 9:324. doi: 10.3389/fgene.2018.00324

Zhu, X., Wang, X., Zhao, H., Pei, T., Kuang, L., and Wang, L. (2020). BHCMDA: a new biased heat conduction based method for potential MiRNA-disease association prediction. *Front. Genet.* 11:384. doi: 10.3389/fgene.2020.00384

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 Neuroscience

ORIGINAL RESEARCH
published: 28 October 2020
doi: 10.3389/fnins.2020.586197

Multimodal Glioma Image Segmentation Using Dual Encoder Structure and Channel Spatial Attention Block

Run Su [1,2], Jinhuai Liu [1,2], Deyun Zhang [3], Chuandong Cheng [4,5,6] and Mingquan Ye [7]*

[1] Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, [2] Science Island Branch of Graduate School, University of Science and Technology of China, Hefei, China, [3] School of Engineering, Anhui Agricultural University, Hefei, China, [4] Department of Neurosurgery, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China, [5] Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China, [6] Anhui Province Key Laboratory of Brain Function and Brain Disease, Hefei, China, [7] School of Medical Information, Wannan Medical College, Wuhu, China

Multimodal medical images provide significant amounts of complementary semantic information. Therefore, multimodal medical imaging has been widely used in the segmentation of gliomas through computational neural networks. However, inputting images from different sources directly to the network does not achieve the best segmentation effect. This paper describes a convolutional neural network called F-S-Net that **f**uses the information from multimodal medical images and uses the semantic information contained within these images for glioma **s**egmentation. The architecture of F-S-Net is formed by cascading two sub-networks. The first sub-network projects the multimodal medical images into the same semantic space, which ensures they have the same semantic metric. The second sub-network uses a dual encoder structure (DES) and a channel spatial attention block (CSAB) to extract more detailed information and focus on the lesion area. DES and CSAB are integrated into U-Net architectures. A multimodal glioma dataset collected by Yijishan Hospital of Wannan Medical College is used to train and evaluate the network. F-S-Net is found to achieve a dice coefficient of 0.9052 and Jaccard similarity of 0.8280, outperforming several previous segmentation methods.

Keywords: medical image fusion, glioma segmentation, fully convolutional neural networks, DES, CSAB, F-S-Net

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:
Mengran Zhou,
Anhui University of Science and Technology, China
Huaqing Zhu,
Anhui Medical University, China

***Correspondence:**
Jinhuai Liu
jhliu@iim.ac.cn

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Neuroscience

Received: *22 July 2020*
Accepted: *15 September 2020*
Published: *28 October 2020*

Citation:
Su R, Liu J, Zhang D, Cheng C and Ye M (2020) Multimodal Glioma Image Segmentation Using Dual Encoder Structure and Channel Spatial Attention Block. Front. Neurosci. 14:586197. doi: 10.3389/fnins.2020.586197

1. INTRODUCTION

Gliomas, which arise from the canceration of gliocyte in the brain and myelon, are the most common form of cancer in the skull, accounting for 80% of malignant brain tumors (Ostrom et al., 2014). The incidence ranges from 3 to 8 per 100,000 people and the fatality rate is high. Hence, the early diagnosis and treatment of gliomas are very important. The presence of gliomas can also cause complications such as increased intracranial pressure, brain edema, brain hernia, and psychosis. The size, location, and type of a glioma are determined by segmenting the affected region from other normal brain tissue. Accurate segmentation plays an important role in the diagnosis and treatment of gliomas. However, manual delineation practices not only require significant anatomical knowledge, but are also expensive, time consuming, and inaccurate. The automatic segmentation of gliomas would allow doctors to detect the growth of brain tumors earlier and provide additional information for the generation of treatment plans. Bi et al. (2019) believed that

artificial intelligence could improve the role of current standard diagnostic imaging technology by refining the preoperative classification of brain tumors above the level achievable by experts. Automatic segmentation based on computer-assisted intervention provides a steady solution for the treatment of gliomas, and is an effective tool in reducing the time required for the accurate detection, location, and delineation of tumor regions. Hence, it is necessary to automatically segment gliomas from medical images.

In recent years, methods based on deep learning (LeCun et al., 2015) have made significant breakthroughs in image classification (Krizhevsky et al., 2012; Rawat and Wang, 2017), image segmentation (Badrinarayanan et al., 2017; Garcia-Garcia et al., 2017), object detection (Ren et al., 2015; Zhao et al., 2019), object tracking (Li et al., 2018; Ristani and Tomasi, 2018), image captioning (Anderson et al., 2018; Hossain et al., 2019), and other fields (Hu et al., 2020). These breakthroughs have promoted the development of deep learning methods in the field of medical image analysis (Litjens et al., 2017; Altaf et al., 2019; Esteva et al., 2019). One of the best-known architectures for medical image segmentation is U-Net, initially proposed by Ronneberger et al. (2015), in which the backbone is a fully convolutional network (FCN) (Long et al., 2015). U-Net has received widespread attention from researchers in the field of medical image segmentation, and many improvements to U-Net have since been proposed (Alom et al., 2018; Oktay et al., 2018; Zhou et al., 2018). For example, Milletari et al. (2016) proposed V-Net for processing 3D medical images, whereby residual learning is employed to improve the convergence speed of the network and random nonlinear transformation and histogram matching are used for data augmentation. Milletari et al. also proposed the dice loss technique based on dice coefficients. Cheng et al. (2019) obtained a multilevel glioma segmentation network by combining an attention mechanism and atrous convolution with 3D U-Net. Chen et al. (2018b) used 3D U-Net and separable 3D convolution to build a separable 3D U-Net architecture. A multiscale masked 3D U-Net was proposed by Xu et al. (2018). The input to their network is a superimposed multiscale map, and multiscale information is obtained from the 3D ASPP layer.

Although methods based on deep learning have been widely used in this field, the current approaches have some disadvantages. Usually, researchers combine multimodal or multisequence medical images to obtain better segmentation accuracy (Kamnitsas et al., 2017b; Chen et al., 2018b; Xu et al., 2018; Zhao et al., 2018; Cheng et al., 2019). The multimodal medical images are input directly into the network for learning. However, the semantic conflicts between multimodal medical images cannot be completely avoided, and these may have a certain impact on the segmentation results. The method of image fusion can integrate valuable information from multimodal medical images, and the fusion results are typically more comprehensive than the original images (Liu et al., 2017). To date, there have been few reports on the segmentation of gliomas based on multimodal medical image fusion.

Another disadvantage of existing methods is that U-Net variants do not improve the basic architecture of U-Net. In particular, the features of the medical images are extracted by a single encoder. This means there may be a loss of feature information. Therefore, it is necessary for networks to obtain and retain more useful features.

In this paper, we propose F-S-Net, which combines image fusion technology to obtain images with richer semantic information. F-S-Net consists of two sub-networks: a **f**usion sub-network and a **s**egmentation sub-network. The fusion sub-network projects images obtained from computed tomography (CT) and magnetic resonance imaging (MRI) into the same semantic space for fusion. Compared with the original images, the fused image contains more semantic information for segmentation. To improve the segmentation performance, the segmentation sub-network uses a dual encoder structure (DES) and a channel spatial attention block (CSAB) to perform image segmentation. Based on the U-Net architecture, DES and CSAB use different sizes of convolution kernel to extract more effective features and focus on the lesion area. In the process of skip-connection, a 1×1 convolution and a concatenation operation are used to achieve better feature fusion. This method is conducive to feature extraction and utilization, and can achieve good performance. DES and CSAB are integrated into the networks based on the U-Net framework, and are found to improve the segmentation result. Experiments show that the cascaded networks proposed in this paper achieve better performance than existing approaches.

The contributions of this study are as follows:

1. A DES is constructed by increasing the width of the encoder. The proposed structure uses convolution kernels of different sizes to extract more effective features from images.
2. Our CSAB is constructed by combining channel attention and spatial attention mechanisms in the U-Net architecture. The proposed attention mechanism can be easily integrated into other networks that use the U-Net framework.
3. The proposed F-S-Net is formed by combining two sub-networks. One sub-network fuses CT and MRI images to enhance the semantic information of the images, while the other is used to segment gliomas accurately from the fused image.
4. Clinical glioma imaging data were collected from Yijishan Hospital of Wannan Medical College. The labels of each image were annotated by professional medical staff. The collected dataset provides a valuable tool for further research.
5. Extensive comparison experiments were conducted based on the collected dataset to demonstrate that the proposed method obtains the best segmentation performance among several deep segmentation methods.

2. RELATED WORK

Convolutional neural networks (CNNs) are a common architecture for glioma segmentation, especially the encoder–decoder model.

Wang et al. (2017) trained each tumor sub-region by using networks with similar architectures and cascading these

networks. The input to each network was the output from the previous network. However, some loss of global information might be caused by the way the gliomas are progressively segmented. Kamnitsas et al. (2017a) reported better results using ensembles of multiple models and architectures (EMMA). In particular, EMMA combined the DeepMedic (Kamnitsas et al., 2017b), FCN, and U-Net models and synthesized their segmentation results. The strong performance of EMMA helped Kamnitsa et al. to win the BraTS Challenge in 2017. However, EMMA does not offer end-to-end training, and the final result is affected by the accumulation of errors. Unlike most researchers, Isensee et al. (2018) demonstrated that competitive performance could be achieved with a few minor modifications to a generic U-Net. They reduced the number of feature maps before sampling from the decoder, and used additional training data to produce some improvements in terms of tumor enhancement. Myronenko (2018) won the BraTS 2018 challenge with a segmentation network based on the encoder–decoder architecture. An asymmetric encoder is used to extract features, and then two decoders segment the brain tumor and reconstruct the input image, respectively. The first decoder outputs the segmentation results from three tumor sub-regions, while the second uses a variational auto-encoder (VAE) to reconstruct the input image. The VAE branch only reconstructs the input images during the training stage. Jiang et al. (2019) achieved the best results in the 2019 BraTS challenge. They proposed a U-Net-based cascade network that is divided into two stages. In the first stage, a variant of U-Net produces an unshaped result. In the second stage, improved performance is obtained by increasing the width of the decoder. In fact, their network uses two decoders that are structurally similar, but have some differences in their up-sampling procedures: one decoder uses deconvolution while the other uses trilinear interpolation. Although multimodal medical imaging has been widely used in glioma segmentation, few researchers have considered the processing of multimodal medical images. This is a clear gap in the research, as the results might be affected by the different semantic information contained in multimodal medical images.

3. METHODS

This section describes the proposed F-S-Net architecture in detail. F-S-Net consists of two sub-networks, a fusion sub-network and a segmentation sub-network. The fusion sub-network uses multimodal images to obtain more detailed medical images with a wealth of semantic information. After processing the corresponding CT and MRI images, the fusion results are input to the segmentation sub-network. The segmentation sub-network uses a dual encoder architecture to extract detailed features from the lesion area. Different sizes of convolutional kernel are used to process images on parallel paths. At the same time, an attention mechanism is integrated into the CSAB module among the skip-connection processing. The final result is obtained by segmenting the fused results.

3.1. F-S-Net

Multimodal medical images have been widely used in medical image analysis tasks. As multimodal images contain different semantic information, image fusion technology is used to map the semantic information from the multimodal images to the same semantic space, including image structure information and edge information. Therefore, F-S-Net incorporates medical image fusion technology. The proposed network architecture is shown in **Figure 1**.

F-S-Net is divided into two stages. In the first stage, the fusion sub-network is used to fuse CT and MRI images. As the semantic information from various multimodal images is combined, this process provides more detailed medical images for segmentation networks. In the second stage, the fused image is input into the segmentation sub-network. The CSAB and DES modules are used in the segmentation sub-network based on the U-Net architecture. **Figure 2** shows the structure of the fusion sub-network (Fan et al., 2019). The E_θ and D_ϕ of fusion sub-network are follows the structure of U-Net. E_θ is used to generate the fusion results. D_ϕ is used to reconstruct the input. The loss value is determined by the input, fusion results, and reconstruction results. The loss function of the fusion sub-network has been modified by us. The details of the loss function are described in section 3.4. D_ϕ is used during the training stage. The segmentation sub-network architecture is a typical encoder–decoder structure, as shown in **Figure 3**. The segmentation sub-network consists of two encoders (left side) and a decoder (right side). The two encoders use convolution kernels of different sizes. In the skip-connection process, the attention mechanism is used to enable the network to extract the features of a specific area and perform feature fusion. The decoder is the same as in U-Net. The network takes input images of 256×256 pixels, and outputs images of the same size. The network can obtain more comprehensive and consistent medical images, and perform better segmentation tasks, after multimodal image fusion. The results are generated by minimizing the loss value.

3.2. Channel Spatial Attention Block

The attention mechanism is derived from the study of human vision. In computer vision, the attention mechanism allows the system to ignore irrelevant information and focus on important information. Combining channel attention, spatial attention, and the structural features of U-Net gives the CSAB module. This module enhances the salient features of the up-sampling process by applying an attention weight to the high- and low-dimensional features. The proposed structure is shown in **Figure 4**. The input feature maps x and g are scaled using the attention coefficient (α_3) computed in CSAB. Areas of concern are selected by analyzing the different types of attention weights provided by x and g.

Given an intermediate feature map x, $g \in R^{C\times H\times W}$ as input, CSAB obtains two intermediate 1D channel attention weights α_1, $\alpha_2 \in R^{C\times 1\times 1}$ and an intermediate 2D spatial attention weight $\alpha_3 \in R^{1\times H\times W}$. **Figure 4** describes the calculation for each attention module. The overall attention process can be summarized as:

$$g^l = \alpha_1(g) \otimes g \tag{1}$$

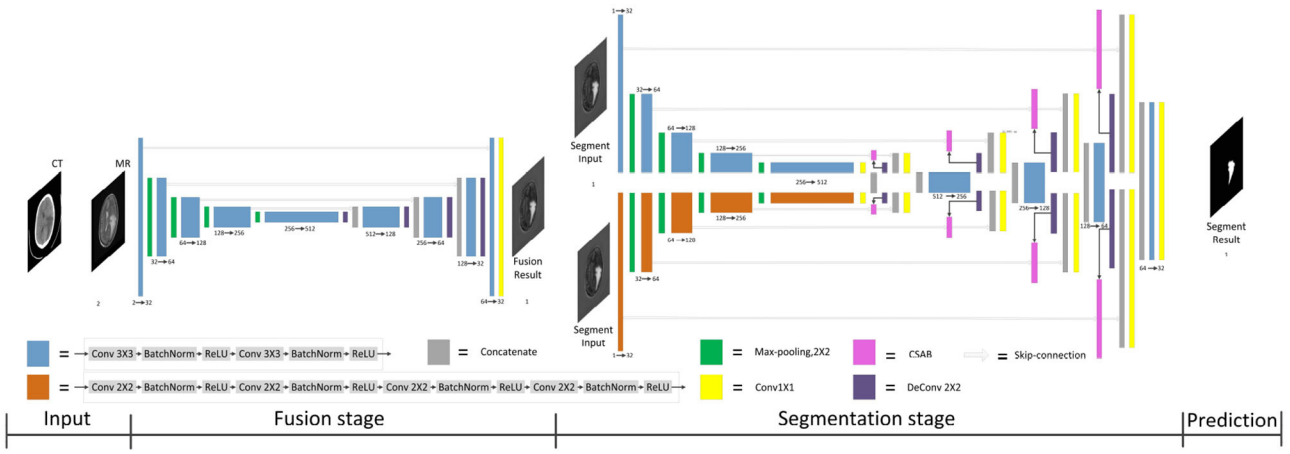

FIGURE 1 | Architecture of the proposed F-S-Net for glioma segmentation. The corresponding CT and MRI images are processed by the fusion sub-network. The fusion images are copied into the segmentation sub-network, which uses convolution kernels of different sizes to process images on parallel paths. The final results of segmentation are output by the decoder.

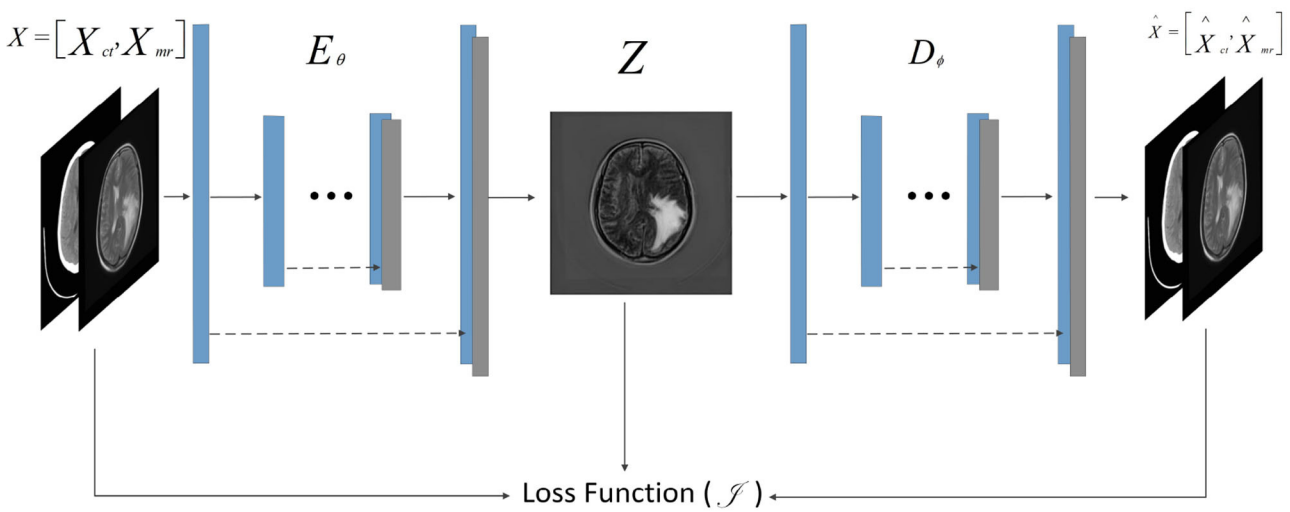

FIGURE 2 | Architecture of fusion sub-network for glioma segmentation. The input size of the network is 256×256. The E_θ and D_ϕ are follows the structure of U-Net. The D_ϕ is used only during training to reconstruct the input.

$$x^l = \alpha_2(x) \otimes x \tag{2}$$

$$f = \alpha_3(g^l, x^l) \otimes x^l \otimes g^l \tag{3}$$

$$F = w(Cat[f, x]) + b \tag{4}$$

where $\otimes$ denotes element-wise multiplication. F is the final output obtained by 1×1 convolution after fusing f and feature x.

3.2.1. Channel Attention Block

The channel attention weight is produced from high- and low-dimensional features using the relationship among the features. Four different spatial context descriptions, g_{max}, g_{avg}, x_{max}, and x_{avg}, are obtained using average pooling and maximum pooling operations on the feature map. These four characteristics are entered into a small network for further processing. The output feature vectors of the small network are merged using a concatenation operation. Finally, the channel attention weights $\alpha_1(g)$ and $\alpha_2(x)$ are obtained after the dimension has been reduced by 1×1 convolution. The channel attention is calculated as follows:

$$g_{max} = MaxPool(g) \tag{5}$$

$$g_{avg} = AvgPool(g) \tag{6}$$

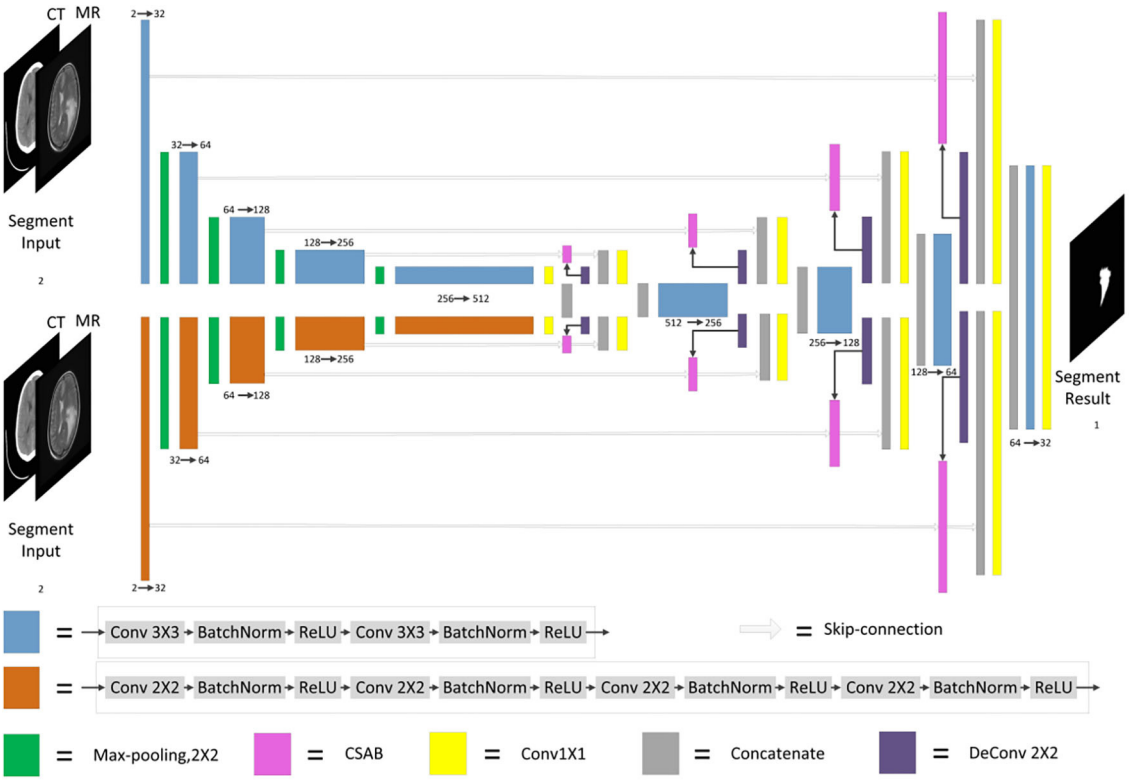

FIGURE 3 | Architecture of segmentation sub-network for glioma segmentation. The input size of the network is 256×256. Each different box denotes the different operations. The number of channels is denoted on the boxes. The parallel pathways process each feature using different sizes of convolution kernel, which are combined at the end of the encoder.

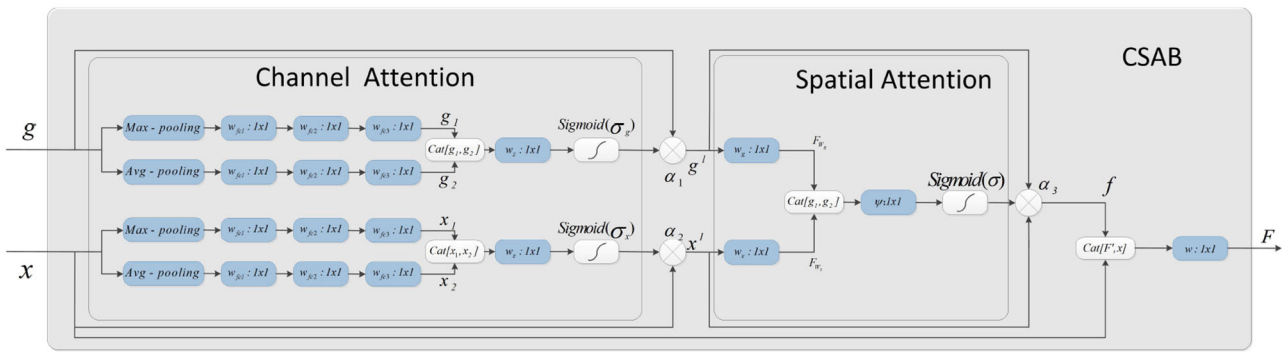

FIGURE 4 | Schematic of the proposed CSAB. CSAB is composed of Channel Attention Block (CAB) and Spatial Attention Block (SAB). CAB applies channel attention weight to *x* from encoder part and *g* from upsampled, respectively. x^l and g^l are the feature maps of CAB output. x^l and g^l are fused in the spatial attention block. Spatial attention weights are applied to the result of feature fusion. *f* is the feature maps of SAB output. *F* is the feature map by CSAB output.

$$x_{max} = MaxPool(x) \tag{7}$$

$$x_{avg} = AvgPool(x) \tag{8}$$

$$g_1 = w_{fc3}(w_{fc2}(w_{fc1}(g_{max}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{9}$$

$$g_2 = w_{fc3}(w_{fc2}(w_{fc1}(g_{avg}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{10}$$

$$x_1 = w_{fc3}(w_{fc2}(w_{fc1}(x_{max}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{11}$$

$$x_2 = w_{fc3}(w_{fc2}(w_{fc1}(x_{avg}) + b_{fc1}) + b_{fc2}) + b_{fc3} \tag{12}$$

$$\alpha_1(g) = \sigma_g(w_z(Cat[g_1, g_2]) + b_z) \tag{13}$$

$$\alpha_2(x) = \sigma_x(w_z(Cat[x_1, x_2]) + b_z) \tag{14}$$

where σ_g and σ_x denote the sigmoid function, $W_{fc1} \in R^{C/8\times1\times1}$, $W_{fc2} \in R^{C/8\times1\times1}$, $W_{fc3} \in R^{C\times1\times1}$, and $W_z \in R^{C\times1\times1}$. W_{fc1}, W_{fc2}, W_{fc3}, and W_z denote the weight of each convolution. The rectified linear units (ReLU) activation function is followed by W_{fc1}, W_{fc2}, and W_{fc3}.

3.2.2. Spatial Attention Block

The spatial attention map is generated from $\alpha_1(g)$ and $\alpha_2(x)$ using the relationship among the features. The attention coefficient, $\alpha_3 \in [0, 1]$, suppresses the expression of irrelevant regions in the input. In addition, the attention coefficient can highlight features that are useful for the task.

In the spatial attention block, the high- and low-dimensional features are subjected to 1×1 convolution to obtain two features: $F_{W_g} \in R^{C\times H\times W}$ and $F_{W_x} \in R^{C\times H\times W}$. The concatenation operation then performs feature fusion. Finally, the spatial attention map of $\alpha_3 \in R^{1\times H\times W}$ is generated by 1×1 convolution. The output of the spatial attention block (SAB) is the element-wise multiplication of the input feature graph and the attention coefficient. The spatial attention is calculated as follows:

$$F_{W_g} = w_g(g^l) + b_g \tag{15}$$

$$F_{W_x} = w_x(x^l) + b_x \tag{16}$$

$$f = \alpha_3(g^l, x^l)\otimes x^l \otimes g^l = \sigma(\psi(Cat[F_{W_g}, F_{W_x}]) + b)\otimes x^l \otimes g^l \tag{17}$$

where σ denotes the sigmoid function. W_g, W_x, and ψ represent the convolution kernel weights, and b_g, b_x, and b are the bias terms.

3.3. Dual Encoder Structure

The DES is developed by extending the encoder of U-Net. Two different encoders are used to extract features from images, and the convolution kernel size of the two encoders is different. One encoder has a convolution kernel size of 3×3, while the other has a convolution kernel size of 2×2. The encoder with a convolution kernel size of 3×3 is consistent with U-Net. Each layer consists of two 3×3 convolutions, followed by batch normalization (BN) and ReLU activation. The encoder with a convolution kernel of 2×2 is different from that of U-Net. Each layer consists of four 2×2 convolutions, each followed by BN and ReLU activation. The padding of the four 2×2 convolutions is 0101. The number of initial filters is 32. More feature information is obtained from images that use convolution kernels of different sizes. In addition, more significant information will be input to the decoder through the parallel paths design.

As the encoder has been expanded, it is necessary to fuse the features of each path when the features are input into the decoder. The output of CSAB is fused with the features obtained by up-sampling. Then, 1×1 convolution is used to reduce the dimension of the fused features. Finally, the processed features are input into the decoder. The two features from the encoder are processed separately. This approach is conducive to the integration of low- and high-dimensional information. The experimental results of the optimization procedure demonstrate the effectiveness of our structure. The structure designed in this study is shown in **Figure 5**.

Let X_1 and X_2 be features extracted by the encoder. F_1 and F_2 are the features output by CSAB, respectively, and g is the feature obtained after up-sampling. F_1 and F_2 are fused with g, and the features connected by skip-connection are subjected to 1×1 convolution for dimension reduction, resulting in x_{13} and x_{23}. These two features are fused after dimensionality reduction to obtain X, which is input to the decoder. X is computed as follows:

$$X = Cat[x_{13}, x_{23}] \tag{18}$$

where F_1, F_2, x_{13}, and x_{23} are given by:

$$F_1 = Att(g, x_1) \tag{19}$$

$$F_2 = Att(g, x_2) \tag{20}$$

$$x_{13} = W_{x_1}(Cat[F_1, g]) + b_1 \tag{21}$$

$$x_{23} = W_{x_2}(Cat[F_2, g]) + b_2 \tag{22}$$

DES has two advantages. First, the convolution kernels of the two encoders are 3×3 and 2×2, respectively. This strategy can extract more different features, which is beneficial to the segmentation task. Secondly, the features processed during the skip-connection ensure more complete information fusion. We have not made any major changes to the U-Net architecture. Therefore, our DES can be extended to most networks that are based on the U-Net architecture.

3.4. Loss Function

The loss function consists of three terms:

$$L_{total} = 0.02 * (L_{MSE} + L_{SSIM}) + L_{BCE} \tag{23}$$

L_{MSE} is the mean squared error (MSE) loss between the reconstructed output I_i and the input image O_i:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(I_i - O_i)^2 \tag{24}$$

where N is the number of epochs.

L_{SSIM} is calculated as:

$$L_{SSIM} = \frac{1}{N}\sum_{i=1}^{N}(1 - SSIM(O_i, F_i)) \tag{25}$$

where SSIM(·) represents the structural similarity between two images (Wang et al., 2004). F_i represents the fused image.

L_{BCE} is the binary cross-entropy (BCE) loss applied to the segmentation output P_i and the segmentation mask T_i:

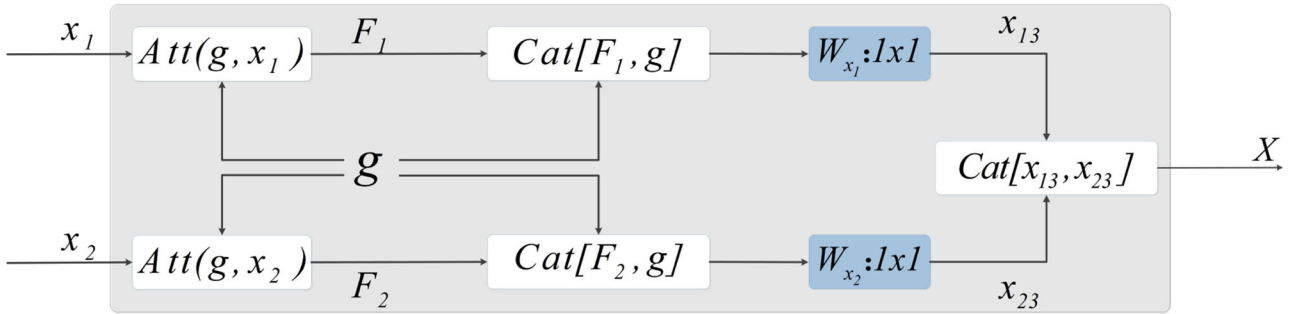

FIGURE 5 | Method of skip-connection. First, the attention weight is applied by CSAB to the output characteristics of the encoder. The features obtained from the up-sampling are then fused with the features after applying the attention weight. Subsequently, 1×1 convolution is used to reduce the dimension of the fused features. The final output is obtained by concatenating the two features.

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}(T_i \log(P_i) + (1 - T_i)\log(1 - P_i)) \quad (26)$$

L_{MSE} and L_{SSIM} are the loss functions of the fusion sub-network, and L_{BCE} is the loss function for the segmentation sub-network. Since calculations of loss function is different, the loss functions must be balanced. The proposed model is trained with $\eta = 1$ and $\gamma = 1$. η represents the loss weight of the fusion sub-network. γ represents the loss weight of the segmentation sub-network. The loss curves are shown in **Figure 6**, from which we can learn that fusion loss is bigger than segmentation. To balance the loss weights between fusion and segmentation sub-networks, the loss weight in Equation (23) are set to $\eta = 0.02$ and $\gamma = 1$.

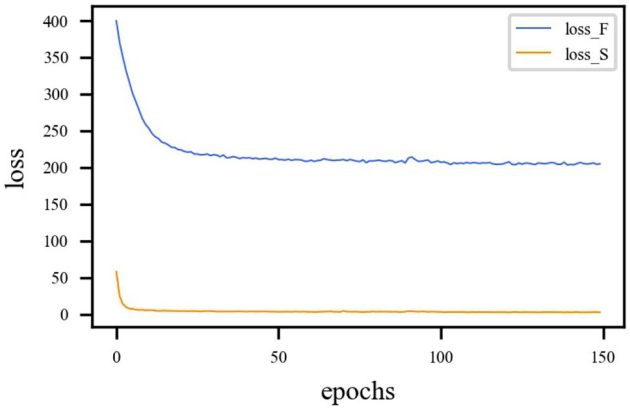

FIGURE 6 | Loss curve of different sub-network. The blue line represents the loss function curve of the fusion sub-network. The orange line represents the loss function curve of the segmentation sub-network. Fusion loss is bigger than segmentation loss. Therefore, weights must be applied to balance the two sub-networks.

4. RESULTS

4.1. Experimental Environment

A 12 GB NVIDIA Titan X (Pascal) was used for training and evaluation. The system was running Windows 10 with an Intel Xeon CPU with 64 GB RAM. The program was written on Pycharm and is based on the Pytorch (Paszke et al., 2019) framework.

4.2. Dataset

The dataset contains clinical imaging data from 26 patients with brain gliomas examined at Yijishan Hospital of Wannan Medical College. The clinical image data consist of CT and T2-weighted MRI scans from glioma patients, of which nine images were acquired from low-grade glioma patients and 17 images were obtained from high-grade glioma patients. These are brain scans before treatment. After slicing the data, 860 pieces of CT and MRI images were obtained. Registration was completed after slicing. In addition, an expert was invited from the First Affiliated Hospital of the University of Science and Technology of China to manual delineate the whole tumor area. The data are shown in **Figure 7**.

Data augmentation was used to improve the generalization ability and robustness of the models. As the image size may change after data augmentation, the images were resampled to 256×256 pixels. Finally, the dataset was randomly divided into a training dataset (60%), validation dataset (20%), and test dataset (20%).

4.3. Evaluation Measures

The accuracy rate (ACC), positive predictive value (PPV), Jaccard similarity (JS), and dice coefficient (DC) were used as evaluation indexes. These metrics were calculated as follows:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (27)$$

$$PPV = \frac{TP}{TP + FP} \quad (28)$$

$$JS = \frac{TP}{FP + TP + FN} \quad (29)$$

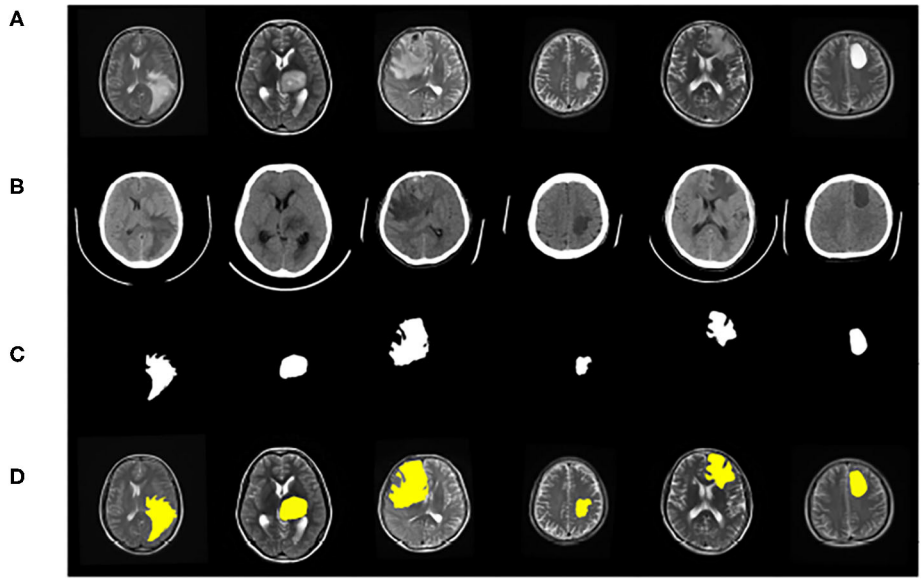

FIGURE 7 | Example of image modalities and ground truth in the multimodal glioma dataset. **(A)** Shows a head scan CT. **(B)** Shows a T2-weighted MRI. **(C)** Shows the ground truth. **(D)** Shows the mergence result of **(B,C)**.

TABLE 1 | JS and DC for F-S-Net with different numbers of kernels and optimizers.

Number of convolution kernels	JS	DC
Adam + (32)	0.8070	0.8922
Adabound + (32)	0.8172	0.8975
SGD + (32)	0.7936	0.8839
Adabound + (16)	0.8040	0.8902

$$DC = \frac{2 * TP}{2 * TP + FP + FN} \tag{30}$$

where TP (true positive) represents the number of foreground pixels that are correctly classified as foreground (tumor region), TN (true negative) represents the number of background pixels that are correctly classified as background (non-tumor region), FP (false positive) represents the number of background pixels that are correctly identified as foreground, and FN (false negative) represents the number of foreground pixels that are incorrectly classified as background.

ACC is used to represent the classification accuracy of the classifier. PPV represents the proportion of true positives in all positive cases. JS reflects the ratio of the common area of the matched element to the split result. Any imprecise segmentation, whether under- or over-segmentation, will cause the JS to decrease. DC calculates the similarity between the prediction results and the ground truth to evaluate the performance of the model.

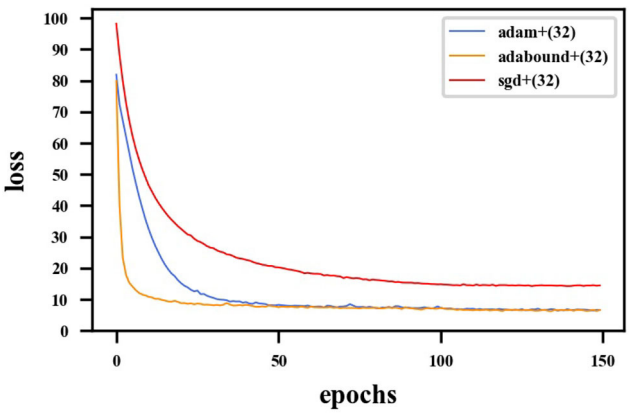

FIGURE 8 | Loss curve for F-S-Net with different optimizers. The red line represents loss function curve of the SGD optimizer. The blue line represents the loss function curve of the Adam optimizer. The orange line represents the loss function curve of the Adabound optimizer. The optimizer of Adabound has the fastest rate of convergence.

4.4. Training Optimization

First, the appropriate numbers of optimizers and convolution kernels were determined. Stochastic gradient descent (SGD) (Robbins and Monro, 1951) has been widely applied in the field of deep learning, while adaptive moment estimation (Kingma and Ba, 2014) offers better optimization performance. Adabound (Luo et al., 2019) dynamically crops the learning rate so that the algorithm is closer to Adam in the early stages of training and closer to SGD at the end. For CNNs, the receptive field and number of channels on the receptive field determine the

performance of the network. The convolution kernels considered in the experiments had the following structures: (16) 1-16-32-64-128-256-128-64-32-16-1; (32) 1-32-64-128-256-512-256-128-64-32-1. Four experimental groups were examined in the experiments: (1) Adam + (32), (2) Adabound + (32), (3) SGD + (32), and (4) Adabound + (16). The number of training epochs was set to 150, the batch size was set to 4, the weight decay was set to 5×10^{-8}, and the learning rate decreased by 0.1 after the 100th epoch. The experimental results are presented in **Table 1**. The loss curve is shown in **Figure 8**. In **Figure 8**, Adabound converges faster than the other optimizers. On the independent test dataset, the DCs of SGD, Adabound, and Adam are 0.8839, 0.8975, and 0.8922, respectively. Based on these results, Adabound and structure (32) were used in subsequent experiments.

Convolution kernels of different sizes have different receptive fields. The convolution kernel size of one encoder was kept the same as that in U-Net, while the convolution kernel size of the other encoder was modified as follows: (1) The 3×3 convolution of the amplified path was replaced by 5×5 convolution. (2) The two 3×3 convolutions were kept unchanged. (3) The 3×3 convolution of the amplified path was replaced by two 2×2 convolutions. Note that the padding is different when using 2×2 convolution. The experimental results presented in **Table 2** show that replacing a set of 3×3 convolutions with a set of 2×2 convolutions produces a better effect.

TABLE 2 | DC and JS for F-S-Net with different sizes of kernels and optimizers in the encoder–decoder for test dataset.

Sizes of convolution kernel	JS	DC
3×3-3×3	0.8172	0.8975
3×3-5×5	0.8226	0.9019
3×3-2×2(0101)	0.8234	0.9023
3×3-2×2(1010)	0.8226	0.9014

0101 and 1010 are the settings for the padding in each 2×2 convolution block.

It is necessary to modify the skip-connection to adapt to the inputs of the two encoders. An increase in skip-connection input would inevitably require feature fusion and dimensionality reduction. The order of 1×1 convolution and feature fusion may affect the performance of the network. Therefore, the four different structures shown in **Figure 9** were constructed.

Experiments were performed using the above four structures. The final experimental results are presented in **Table 3**, showing that better results are obtained by the skip-connection and dimension reduction of the two paths, respectively.

4.5. Ablation Analysis of Proposed Methods

The experimental results of the proposed structures with non-fusion and fusion were compared. It is clear that the improved structure and combination of modules are effective in enhancing the glioma segmentation results. The hyperparameters were set according to the previous optimization experiment. The training and testing samples for the experiment were taken from the glioma dataset. The fusion results in **Figure 10** clearly represent the overall area of the tumor, which makes the image features more obvious. The glioma can be accurately segmented and the network captures the specific outline and edge details of the lesion area in the image. **Table 4** presents the experimental results from using the proposed architecture.

DECSAU-Net is the segmentation sub-network in F-S-Net. When the proposed modules are removed, the network

TABLE 3 | DC and JS for F-S-Net with different types of skip-connection.

Skip-connection type	JS	DC
a	0.8234	0.9023
b	0.8019	0.8883
c	0.8109	0.8947
d	0.8280	0.9052

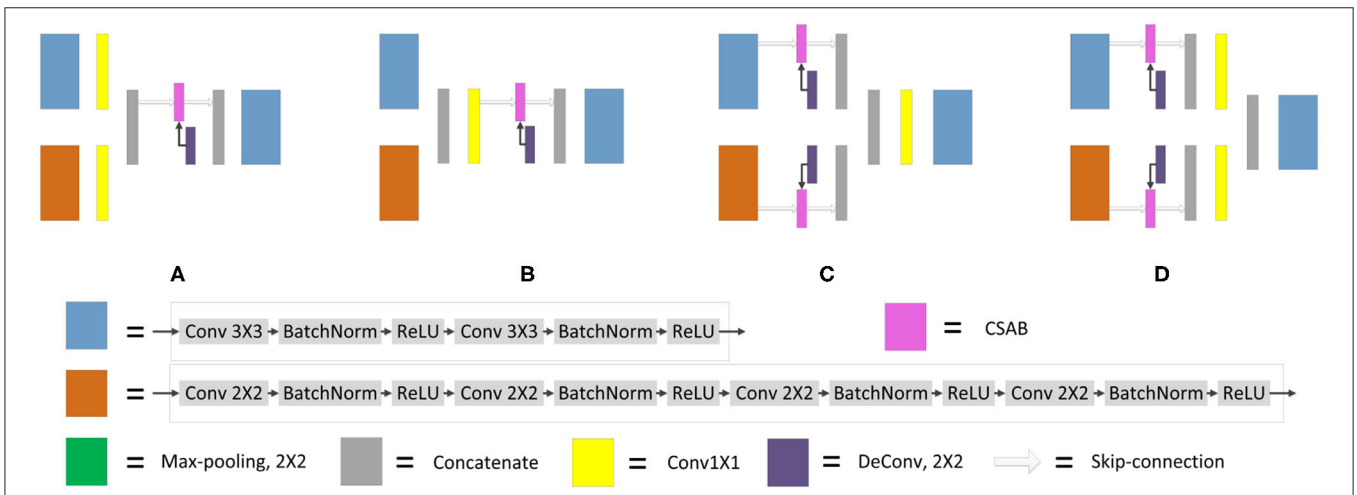

FIGURE 9 | Different types of skip-connection. The skip-connection of F-S-Net is divided into three parts: (**Cu**) CSAB and up-sampling. (**1×1**) 1×1 convolution. (**Cat**) concatenation. **(A)** (**1×1**) + (**Cat**) + (**Cu**). **(B)** (**Cat**) + (**1×1**) + (**Cu**). **(C)** (**Cu**) + (**Cat**) + (**1×1**). **(D)** (**Cu**) + (**1×1**) + (**Cat**).

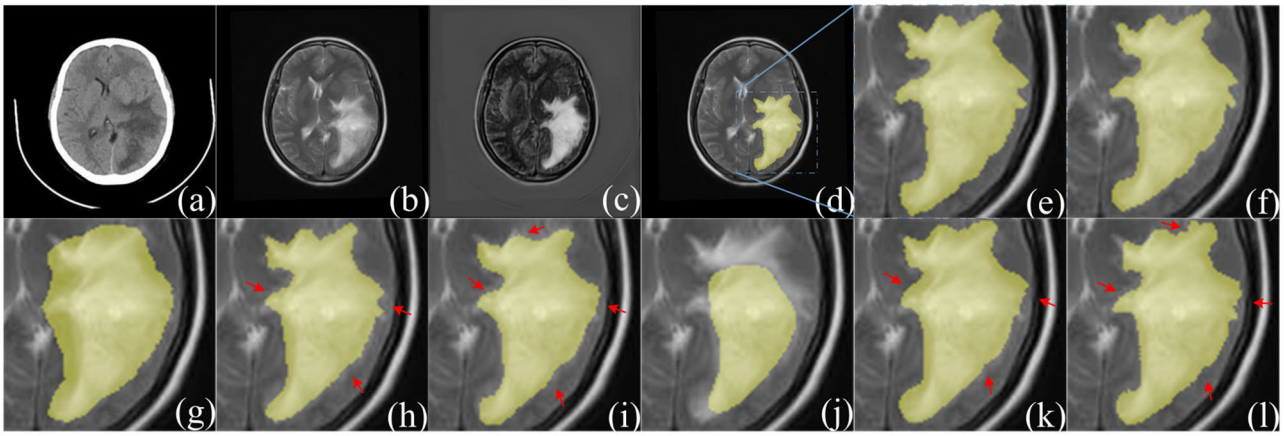

FIGURE 10 | Comparison of segmentation results between F-S-Net and other networks. **(a,b)** source images before fusion, **(c)** fusion result. Compared with **(a,b)**, the features in **(c)** are more obvious. The ground-truth glioma segmentation **(e)** is highlighted in **(d)**. Similarly, other model predictions are compared with those of F-S-Net **(f)**. **(g–l)** Are the results given by FCN8s, SegNet, DeeplabV3+, U-Net, R2U-Net, and AttU-Net, respectively. The missed dense predictions by other models are highlighted with red arrows.

TABLE 4 | Evaluation metric for ablation analysis of our methods with test dataset.

Model	ACC	PPV	JS	DC
U-Net (Ronneberger et al., 2015)	0.9916	0.8001	0.7656	0.8656
DECSAU-Net (Ours)	0.9938	0.8624	0.8193	0.8994
F-S-Net (Ours)	0.9943	0.9054	0.8280	0.9052

TABLE 5 | Evaluation metrics for different network architectures.

Model	ACC	PPV	JS	DC
FCN8s (Long et al., 2015)	0.9885	0.7714	0.6980	0.8197
SegNet (Badrinarayanan et al., 2017)	0.9931	0.8428	0.8039	0.8890
DeeplabV3+ (Chen et al., 2018a)	0.9931	0.8328	0.8066	0.8914
U-Net (Ronneberger et al., 2015)	0.9916	0.8001	0.7656	0.8656
R2U-Net (Alom et al., 2018)	0.9932	0.8472	0.8040	0.8905
AttU-Net (Oktay et al., 2018)	0.9934	0.8586	0.8087	0.8932
DECSAU-Net (Ours)	0.9938	0.8624	0.8193	0.8994
F-S-Net (Ours)	0.9943	0.9054	0.8280	0.9052

architecture is the same as the standard U-Net. Comparing the network models with and without DES and CSAB, it can be seen that the inclusion of DES and CSAB results in better performance. The PPV of DECSAU-Net is about 0.0623 higher than that of U-Net. The JS and DC values are about 0.0537 and 0.0338 higher, respectively. A comparison with U-Net shows that DES and CSAB improve the results of U-Net.

The results achieved with non-fusion and fusion approaches are now compared. The DC of the fused image is about 0.0058 higher than that of the image before fusion. The PPV of glioma segmentation after fusion is also higher at 0.9054. The difference in JS values shows that the result obtained after fusion is more similar to the ground truth. In general, the higher DC and JS values demonstrate that the segmentation is more accurate after fusion.

4.6. Comparison With Other Methods

Table 5 compares the performance of different network architectures with that of the proposed F-S-Net after normalizing and enhancing the glioma data on the same test dataset. **Figure 10** shows the glioma segmentation results, which can be used to compare F-S-Net with other networks.

Several medical image segmentation architectures (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Alom et al., 2018; Chen et al., 2018a; Oktay et al., 2018) are outperformed by F-S-Net in both evaluations. The results in **Table 5** indicate that F-S-Net is more effective for performing accurate glioma segmentation. Compared with other network architectures, our method is more conducive to the segmentation of lesions as it maps multimodal medical images into the same semantic space. The advantage of F-S-Net is that the fusion of multimodal images makes the semantic information more conspicuous, and DES and CSAB allow the network to achieve a better segmentation effect.

5. DISCUSSION

Segmenting gliomas directly from CT or MRI images is a challenging task. In addition, the blurred edges of adjacent bones, blood vessels, or surgical packaging materials greatly increase the difficulty of segmentation.

Currently, most researchers directly input multimodal images into a network for learning. To the best of our knowledge, there are few reports on the segmentation of gliomas based on multimodal medical image fusion. To bridge this gap, F-S-Net has been proposed based on medical image fusion technology. Fusion and segmentation sub-networks are cascaded for end-to-end training, and two

new structures, DES and CSAB, are proposed based on the structural characteristics of U-Net. The basic idea of F-S-Net is to use fusion technology to produce images with more semantic information for the segmentation network, so as to obtain better segmentation results. DES and CSAB extract more detailed features and force the network to focus on the lesion area. Our work builds on existing techniques, such as CT and MRI image fusion. Medical image fusion techniques are not specifically designed for the segmentation task, but can provide images with richer semantic information for segmentation.

The most important innovation described in this paper is the ability to perform the task of glioma segmentation using image fusion. In the field of medical image analysis, better performance is often achieved by combining different technologies. The results in **Table 4** demonstrate the effectiveness of DES and CSAB, while those in **Tables 4**, **5** demonstrate the improvement offered by using fusion technology for segmentation. Our network has the following advantages. First, image fusion can enrich the information available by integrating information between multimodal medical images. This method improves the quality of the image and facilitates the segmentation task. Second, the convolution kernels of different sizes in DES allow the network to obtain richer features. This helps to focus attention on the area of interest, and then obtains a better segmentation effect. Third, CSAB makes the network focus on the lesion area by applying different attention weights to the features. Our method not only integrates the complementary information from different modalities, but also extracts more detailed features. The experimental results show that F-S-Net outperforms several existing methods.

In summary, our proposed method will be helpful in allowing clinicians to diagnose and treat gliomas. More detailed segmentation results provide doctors with more complete boundary information of the tumor, and can better guide the resulting operations. In addition, better segmentation contributes to the reconstruction of the image data, which can provide more information for future monitoring and treatment planning. Our method overcomes the problem of incomplete semantic information and achieves good performance. The combination of segmentation and other medical imaging technologies will be explored in the future. This may improve clinical guidance in the diagnosis and treatment of glioma patients.

6. CONCLUSION

Glioma segmentation is a challenging and significant task in medical image segmentation. Based on medical image fusion technology, a cascade network was proposed to automatically segment gliomas from CT and MRI images. Our network obtained a DC of 0.9052 on the test dataset. Experimental results show that the combination of image fusion and image segmentation is effective. Our model provides a new method and a new idea for glioma segmentation based on deep learning, and is beneficial to the clinical diagnosis and treatment of patients. The proposed network is not only applicable to the segmentation of gliomas, but could also be easily applied to other medical image segmentation tasks.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because, datasets are open in the future. Requests to access the datasets should be directed to Deyun Zhang, zhangdeyun2016@163.com.

AUTHOR CONTRIBUTIONS

RS and JL: conceptualization. RS, CC, and MY: data curation. RS: methodology, project administration, visualization, and writing (original draft). RS, JL, and DZ: validation and writing (review and editing). All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by two research grants: (1) Science and Technology Project grant from Anhui Province (Grant Nos. 1508085QHl84 and 201904a07020098). (2) Fundamental Research Fund for the Central Universities (Grant No. WK 9110000032).

ACKNOWLEDGMENTS

The authors are grateful for the glioma dataset provided by Yijishan Hospital of Wannan Medical College. Thanks are also due for the ground truth glioma dataset provided by the First Affiliated Hospital of University of Science and Technology of China.

REFERENCES

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-net (R2u-net) for medical image segmentation. *arXiv [Preprint].* arXiv:1802.06955. doi: 10.1109/NAECON.2018.8556686

Altaf, F., Islam, S. M., Akhtar, N., and Janjua, N. K. (2019). Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access* 7, 99540–99572. doi: 10.1109/ACCESS.2019.2929365

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 6077–6086. doi: 10.1109/CVPR.2018.00636

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., et al. (2019). Artificial intelligence in cancer imaging: clinical challenges and applications. *Cancer J. Clin.* 69, 127–157. doi: 10.3322/caac.21552

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018a). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 801–818.

Chen, W., Liu, B., Peng, S., Sun, J., and Qiao, X. (2018b). "S3D-Unet: separable 3D U-net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 358–368. doi: 10.1007/978-3-030-11726-9_32

Cheng, J., Liu, J., Liu, L., Pan, Y., and Wang, J. (2019). "Multi-level glioma segmentation using 3D U-net combined attention mechanism with atrous convolution," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA), 1031–1036. doi: 10.1109/BIBM47256.2019.8983092

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Fan, F., Huang, Y., Wang, L., Xiong, X., Jiang, Z., Zhang, Z., et al. (2019). A semantic-based medical image fusion approach. *arXiv [Preprint]*. arXiv:1906.00225. Available online at: https://arxiv.org/abs/1906.00225

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., and Garcia-Rodriguez, J. (2017). A review on deep learning techniques applied to semantic segmentation. *arXiv [Preprint]*. arXiv:1704.06857. doi: 10.1016/j.asoc.2018.05.018

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., and Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Comput. Surveys* 51, 1–36. doi: 10.1145/3295748

Hu, S., Chen, P., Gu, P., and Wang, B. (2020). A deep learning-based chemical system for QSAR prediction. *IEEE J. Biomed. Health Inform.* 24, 3020–3028. doi: 10.1109/JBHI.2020.2977009

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). "No new-net," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 234–244.

Jiang, Z., Ding, C., Liu, M., and Tao, D. (2019). "Two-stage cascaded U-net: 1st place solution to brats challenge 2019 segmentation task," in *BrainLes@MICCAI* (Shenzhen). doi: 10.1007/978-3-030-46640-4_22

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., et al. (2017a). "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop* (Quebec City, QC: Springer), 450–462. doi: 10.1007/978-3-319-75238-9_38

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017b). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]*. arXiv:1412.6980. Available online at: https://arxiv.org/abs/1412.6980

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Li, P., Wang, D., Wang, L., and Lu, H. (2018). Deep visual tracking: review and experimental comparison. *Pattern Recogn.* 76, 323–338. doi: 10.1016/j.patcog.2017.11.007

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, Y., Chen, X., Cheng, J., and Peng, H. (2017). "A medical image fusion method based on convolutional neural networks," in *2017 20th International Conference on Information Fusion (Fusion)* (Xi'an), 1–7. doi: 10.23919/ICIF.2017.8009769

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Luo, L., Xiong, Y., Liu, Y., and Sun, X. (2019). Adaptive gradient methods with dynamic bound of learning rate. *arXiv [Preprint]*. arXiv:1902.09843. Available online at: https://arxiv.org/abs/1902.09843

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford University), 565–571. doi: 10.1109/3DV.2016.79

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Granada: Springer), 311–320. doi: 10.1007/978-3-030-11726-9_28

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-net: Learning where to look for the pancreas. *arXiv [Preprint]*. arXiv:1804.03999. Available online at: https://arxiv.org/abs/1804.03999

Ostrom, Q. T., Bauchet, L., Davis, F. G., Deltour, I., Fisher, J. L., Langer, C. E., et al. (2014). The epidemiology of glioma in adults: a "state of the science" review. *Neuro-oncology* 16, 896–913. doi: 10.1093/neuonc/nou087

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (Vancouver, BC), 8024–8035.

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/neco_a_00990

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (Montreal, QC), 91–99.

Ristani, E., and Tomasi, C. (2018). "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City), 6036–6046. doi: 10.1109/CVPR.2018.00632

Robbins, H., and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* 22, 400–407. doi: 10.1214/aoms/1177729586

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Wang, G., Li, W., Ourselin, S., and Vercauteren, T. (2017). "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks," in *International MICCAI Brainlesion Workshop* (Quebec City, QC: Springer), 178–190.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/TIP.2003.819861

Xu, Y., Gong, M., Fu, H., Tao, D., Zhang, K., and Batmanghelich, K. (2018). "Multi-scale masked 3-D U-net for brain tumor segmentation," in *International MICCAI Brainlesion Workshop* (Springer), 222–233. doi: 10.1007/978-3-030-11726-9_20

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., and Fan, Y. (2018). A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. *Med. Image Anal.* 43, 98–111. doi: 10.1016/j.media.2017.10.002

Zhao, Z.-Q., Zheng, P., Xu, S.-T., and Wu, X. (2019). Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3212–3232. doi: 10.1109/TNNLS.2018.2876865

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). "Unet++: A nested U-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Granada: Springer), 3–11. doi: 10.1007/978-3-030-00889-5_1

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 Physiology

ORIGINAL RESEARCH
published: 11 November 2020
doi: 10.3389/fphys.2020.604241

Multiple Feature Selection Strategies Identified Novel Cardiac Gene Expression Signature for Heart Failure

Dan Li[1], Hong Lin[2] and Luyifei Li[1]*

[1]Department of Cardiovascular Medicine, First Hospital Affiliated to Harbin Medical University, Harbin, China, [2]Internal Medicine-Cardiovascular Department, Harbin Chest Hospital, Harbin, China

Heart failure (HF) is a serious condition in which the support of blood pumped by the heart is insufficient to meet the demands of body at a normal cardiac filling pressure. Approximately 26 million patients worldwide are suffering from heart failure and about 17–45% of patients with heart failure die within 1-year, and the majority die within 5-years admitted to a hospital. The molecular mechanisms underlying the progression of heart failure have been poorly studied. We compared the gene expression profiles between patients with heart failure (n = 177) and without heart failure (n = 136) using multiple feature selection strategies and identified 38 HF signature genes. The support vector machine (SVM) classifier based on these 38 genes evaluated with leave-one-out cross validation (LOOCV) achieved great performance with sensitivity of 0.983 and specificity of 0.963. The network analysis suggested that the hub gene *SMOC2* may play important roles in HF. Other genes, such as *FCN3*, *HMGN2*, and *SERPINA3*, also showed great promises. Our results can facilitate the early detection of heart failure and can reveal its molecular mechanisms.

Keywords: heart failure, microarray, biomarker, network, molecular mechanism

OPEN ACCESS

Edited by:
Tao Huang,
Chinese Academy of Sciences (CAS), China

Reviewed by:
Hui Huang,
Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, China
Yuan-Lin Zheng,
Jiangsu Normal University, China

****Correspondence:***
Dan Li
lihao325657@163.com

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *09 September 2020*
Accepted: *15 October 2020*
Published: *11 November 2020*

Citation:
Li D, Lin H and Li L (2020) Multiple Feature Selection Strategies Identified Novel Cardiac Gene Expression Signature for Heart Failure. Front. Physiol. 11:604241. doi: 10.3389/fphys.2020.604241

INTRODUCTION

Heart failure (HF) is a serious condition in which the support of blood pumped by the heart is insufficient to meet the demands of body at a normal cardiac filling pressure (Ramachandra et al., 2020). Defined as a syndrome with high morbidity and mortality, HF is the major cause of death and a serious threat to human health for a long period (Jarcho, 2020). Approximately 26 million patients worldwide are suffering from heart failure, and the society faces the long-term great stresses on patients, medical stuff, and medical systems (Bowen et al., 2020). About 17–45% of patients with heart failure die within 1 year, and the majority die within 5 years admitted to a hospital in worldwide (Davison and Cotter, 2015; Zhou et al., 2020). However, the survival rates for patients with HF have improved in many parts of the world in recent years along with the advanced therapies and patient management systems. Heart failure is a complex disease, and so many factors are responsible that it is hard to blame it on one specific issue (McMurray and Pfeffer, 2005).

Over the past decades, the genetic causes and molecular mechanism underlying the progression of heart failure have been partially illustrated. Most previous studies in heart failure are limited by inadequate biological samples from patients with heart failure (Prohászka et al., 2013). Since then, studies have focused on the molecular mechanism of heart failure by virtue of animal models in combination with molecular biological techniques. Previous studies suggested that classification of disease status for HF is much important for the decision of treatment and improvement of prognosis (van Oort et al., 2011). They have discovered that novel gene biomarkers play a vital role in various diseases depending on the leapfrog development of RNA-Seq technology (Asakura and Kitakaze, 2009). According to previous reports, the specific gene expression is related to the pathological conditions of HF.

Liu et al. (2015) collected six samples from three controls, one ischemic heart disease (ISCH), and two dilated cardiomyopathies (DCMs) and used RNA-Seq to filter novel gene signatures for HF, and precisely categorize HF status in larger samples of 313 patients. Vigil-Garcia et al. (2020) selected novel genes induced during pathological cardiac hypertrophy that are relevant for human HF through cardiomyocyte-specific gene expression analysis. These results recognized PFKP as a novel potential therapeutic target to prohibit the succession of HF. Tan et al. (2002) used microarrays to describe gene expression fingerprints of HF etiologies based on seven non-failing human hearts and eight failing human hearts with a diagnosis of end-stage dilated cardiomyopathy. Zhou et al. (2020) proposed that valosin-containing protein could protect the heart against pressure overload-induced heart failure using RNA-Seq and a comprehensive bioinformatics analysis. Kittleson et al. (2004) used microarrays of 48 myocardial samples and gene expression profiling to predict biomarkers in determining prognosis and response to therapy in HF precisely. All these studies were based on microarrays, which have been the remarkable method for gene expression studies because of their ability to filter thousands of transcripts.

In our study, we tried to detect the novel HF signature genes and their networks from previous transcriptomic data which included the gene expression profiles in patients with heart failure (n = 177) and without heart failure (n = 136) using advanced bioinformatics methods. Compared with previous studies, which are intended to find the biomarker for HF put the focus on separated gene, our study focused on the linkage among them. We built the support vector machine (SVM) model with the application of multiple feature selection methods: Monte Carlo Feature Selection (MCFS; Draminski et al., 2008; Chen et al., 2018a, 2020; Pan et al., 2019b; Li et al., 2020a) and incremental feature selection (IFS; Zhang et al., 2016; Chen et al., 2018b, 2020; Wang et al., 2018; Pan et al., 2019a). What is more, we used the Search Tool for the Retrieval of Interacting Genes (STRING) database (Szklarczyk et al., 2018) to explore the protein interaction networks. A remarkable result of our study is that 38 selected genes can serve as novel biomarkers for HF and can conduce to revealing the pathological mechanism of HF.

MATERIALS AND METHODS

The Microarray Data of Heart Failure Patients

We downloaded the microarray gene expression data of 177 patients with heart failure and 136 patients without heart failure from Gene Expression Omnibus (GEO) at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57338 (Liu et al., 2015). The expression levels of 33,297 probes corresponding to 20,254 genes in the cardiac tissue were measured with Affymetrix Human Gene 1.1 ST Array. The probes corresponding to the same gene were averaged to obtain the gene expression levels, and the gene expression levels were quantile normalized using function normalize.quantiles from R/Bioconductor package preprocessCore[1] to minimize the systematic variance. The normalized data were used for further feature selections.

Select the Genes Based on Their Importance to Classify the Heart Failure Patients

There have been many methods for identifying differentially expressed genes (DEGs), such as t-test. But such methods only consider the distribution of one gene each time, and do not consider the relationship among genes (Tao et al., 2020). That leads to two limitations: (1) The distribution difference of a gene is not equivalent to its classification ability; and (2) The combinations of the most significant DEGs may not have good performance since they may be redundant and do not help each other to achieve a better performance. Therefore, we adopted machine learning based multiple feature selection strategies to objectively select the optimal heart failure signature. The machine learning-based methods have been widely used and achieved great success in biomarker discovery (Wang and Huang, 2019; Li et al., 2020a,b; Yuan et al., 2020; Zhang et al., 2020a,b; Zhu et al., 2020).

The proposed multiple feature selection strategies can be summarized as **Figure 1**. First, the expression profiles of 20,254 genes in 177 patients with heart failure and 136 patients without heart failure were normalized. Second, we randomly selected many subset data to construct the classification trees using Monte Carlo strategy (Draminski et al., 2008; Chen et al., 2018a, 2020; Pan et al., 2019b; Li et al., 2020a). To perform MCFS, we used the dmLab software version 2.3.0 from https://home.ipipan.waw.pl/m.draminski/mcfs.html. Third, all these trees were ensembled to calculate the classification importance of the genes. The important genes should appear in a large number of trees and be able to correctly classify the samples into right groups.

[1]https://bioconductor.org/packages/preprocessCore/

Fourth, the top ranked genes (1,000 in this study) were further analyzed using IFS strategy (Zhang et al., 2016; Chen et al., 2018b, 2019; Wang et al., 2018; Pan et al., 2019a). Each time, a gene set including the top K most important genes (K = 1, 2, 3, …, 1,000) was used to train a SVM model, and its performance was evaluated with leave-one-out cross validation (LOOCV; Li and Huang, 2018). To build the SVM, we used the function svm from R package e1071.[2] Fifth, the optimal heart failure signature was the gene set with the best performance. If the IFS curve did not reach its peak or the plateau area and kept increasing as the number of genes increased, more top genes should be analyzed. Sixth, to better understand the underlying regulatory mechanisms of the signature and increase the interpretability of the signature, we constructed the signature network based on STRING database version 11.0 (http://string-db.org; Szklarczyk et al., 2018; Shi et al., 2020).

RESULTS

The Optimal Heart Failure Signature Identification

We adopted multiple feature selection strategies (**Figure 1**) to identify the optimal heart failure signature. It integrated the strategies of MCFS and IFS. Step A was data preprocessing. MCFS included Steps B and C. IFS included Steps D and E. Step F was to interpret the biological mechanisms of the signature. As demonstrated in **Figure 1D**, the actual IFS curve was shown in **Figure 2**. The highest LOOCV accuracy was 0.974 when the top 38 MCFS genes were used to train the SVM model. Therefore, these 38 genes were chosen as the optimal heart failure signature, which was shown in **Table 1**. The confusion matrix of the 38 optimal heart failure signature genes which compared the actual class labels and precited class labels of all samples were given in **Table 2**. Their LOOCV sensitivity, specificity, and accuracy were 0.983, 0.963, and 0.974, respectively. The performance was great.

[2]https://CRAN.R-project.org/package=e1071

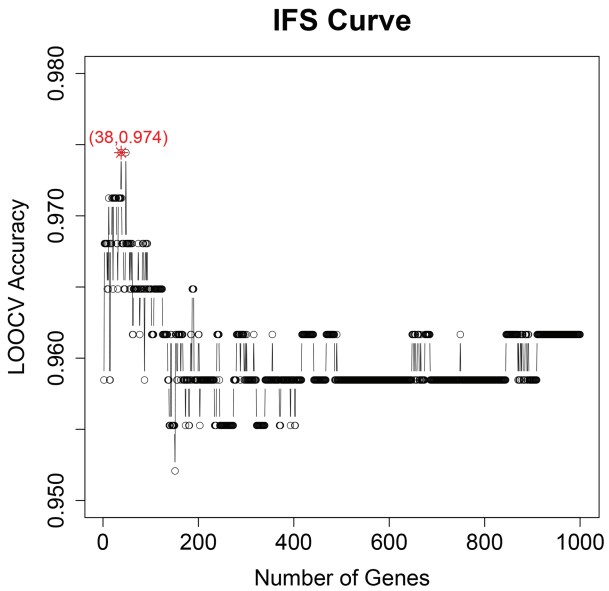

FIGURE 2 | The IFS curve of optimal heart failure signature identification. It showed the relationship between the number of genes (x) and their LOOCV accuracy (y). The peak accuracy was 0.974 when 38 genes were used. Therefore, the 38 genes were chosen as optimal heart failure signature.

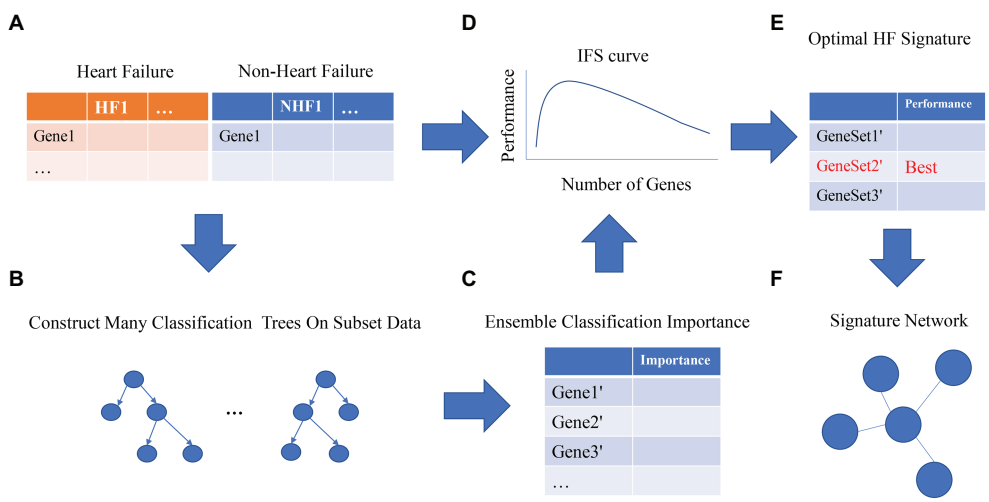

FIGURE 1 | The workflow for optimal heart failure signature identification. The workflow integrated the strategies of Monte Carlo Feature Selection (MCFS) and incremental feature selection (IFS). Step **(A)**: data preprocessing. Steps **(B,C)**: MCFS. Steps **(D,E)**: IFS. Step **(F)**: signature network.

TABLE 1 | The 38 optimal heart failure (HF) signature genes.

Rank	Gene symbol	Full name	Importance
1	HMGN2	High mobility group nucleosomal binding domain 2	0.571
2	HMOX2	Heme oxygenase 2	0.527
3	SERPINA3	Serpin family A member 3	0.499
4	TUBA3D	Tubulin alpha 3d	0.489
5	ECM2	Extracellular matrix protein 2	0.481
6	FREM1	FRAS1 related extracellular matrix 1	0.461
7	FCN3	Ficolin 3	0.458
8	ZMAT1	Zinc finger matrin-type 1	0.405
9	SMOC2	SPARC related modular calcium binding 2	0.386
10	CSDC2	Cold shock domain containing C2	0.383
11	LCN6	Lipocalin 6	0.359
12	LUM	Lumican	0.356
13	FURIN	Furin, paired basic amino acid cleaving enzyme	0.349
14	LAD1	Ladinin 1	0.338
15	MNS1	Meiosis specific nuclear structural 1	0.338
16	ASPN	Asporin	0.317
17	FRZB	Frizzled related protein	0.310
18	GGT5	Gamma-glutamyltransferase 5	0.296
19	TUBA3E	Tubulin alpha 3e	0.293
20	PDE5A	Phosphodiesterase 5A	0.292
21	ISLR	Immunoglobulin superfamily containing leucine rich repeat	0.289
22	S1PR3	Sphingosine-1-phosphate receptor 3	0.279
23	SFRP4	Secreted frizzled related protein 4	0.271
24	APBB3	Amyloid beta precursor protein binding family B member 3	0.270
25	USP31	Ubiquitin specific peptidase 31	0.268
26	SLCO4A1	Solute carrier organic anion transporter family member 4A1	0.251
27	VSIG4	V-set and immunoglobulin domain containing 4	0.251
28	KCNN3	Potassium calcium-activated channel Subfamily N member 3	0.250
29	FAM58A	CCNQ cyclin Q cyclin Q	0.248
30	AP3M2	Adaptor related protein complex 3 subunit mu 2	0.247
31	C15orf59	INSYN1 inhibitory synaptic factor 1	0.243
32	BTN3A1	Butyrophilin subfamily 3 member A1	0.243
33	ZDHHC16	Zinc finger DHHC-type containing 16	0.241
34	CD163	CD163 molecule	0.238
35	SEMA4B	Semaphorin 4B	0.237
36	ST6GALNAC3	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 3	0.228
37	TTC3	Tetratricopeptide repeat domain 3	0.228
38	MATN2	Matrilin 2	0.219

The Expression Pattern of the 38 Genes in Patients With HF and Without HF

We plotted the heatmap of the 38 genes in 177 patients with heart failure (HF) and 136 patients without heart failure [non-heart failure (NHF)] in **Figure 3**. It can be seen that most samples were clustered into the correct groups. Only very few samples were misclustered. Within the 38 genes, 17 genes (*ZMAT1*, *APBB3*, *MNS1*, *AP3M2*, *BTN3A1*, *KCNN3*, *TTC3*, *SMOC2*, *LUM*, *ASPN*, *FRZB*, *SFRP4*, *MATN2*, *ISLR*, *PDE5A*, *ECM2*, and *FREM1*) were highly expressed in HF and 20 genes (*FAM58A*, *CSDC2*, *C15orf59*, *S1PR3*, *VSIG4*, *CD163*, *SEMA4B*, *SLCO4A1*, *SERPINA3*, *GGT5*, *FURIN*, *ZDHHC16*, *LAD1*, *USP31*, *TUBA3D*, *TUBA3E*, *ST6GALNAC3*, *LCN6*, *HMOX2*, and *FCN3*) were lowly expressed in HF.

TABLE 2 | The confusion matrix of the 38 optimal heart failure signature genes.

	Predicted HF	Predicted NHF
Actual HF	174	3
Actual NHF	5	131

The Network of the 38 Genes

Signature genes were not necessarily key regulators. They could be only markers. But if the signature genes have clear biological functions, they certainly can be better interpreted. Therefore, as we stated in **Figure 1F**, we searched the interaction among the STRING database (https://string-db.org/; Szklarczyk et al., 2018) and plotted the networks of the 38 genes in **Figure 4**. It can be seen that *SMOC2* is located in the hub position of the network.

SMOC2, a member of the SPARC family, which is highly expressed during embryogenesis and wound healing. Previous studies recognized that inflammatory pathways were generally dysregulated in right ventricular failure (RVF) tissue. Williams et al. (2018) analyzed mRNA datasets of human non-failing and failing heart samples from patients, and concluded that *SMOC2* was differentially expressed. *SMOC2* could be a potential significance factor that altered remodeling and inflammation for further study in the mechanism of HF. Laugier et al. (2017) found that *SMOC2*, involved in matrix remodeling, is potentially associated with the increased T-helper 1 cytokine-mediated inflammatory damage in heart, using genome-wide cardiac DNA methylation on global gene expression in myocardial samples in chronic Chagas disease cardiomyopathy, which is an inflammatory cardiomyopathy presenting with heart failure and arrhythmia.

DISCUSSION

In the present study, 38 genes were selected from our prediction model of SVM, implying strong relevance with the pathological mechanisms of HF. After literature retrieval and utilization, several evidences and analysis results have been retrieved to validate the dependability and reality of our analysis.

FCN3, a member of ficolin/opsonin p35 lectin family which consists of a collagen-like domain and a fibrinogen-like domain, which were found in all human serum. Prohászka et al. (2013) reported that the main initiator molecules of the lectin complement pathway *MBL*, *FCN2*, and *FCN3* were related to chronic heart failure (CHF). Low *FCN3* levels were related to decreased concentrations of complement factor C3 and increased complement activation product C3a (Prohászka et al., 2013). They also provided evidence for a significant association of low *FCN3* levels with advanced HF and outcome (Prohászka et al., 2013). *FCN3* is reported to be increased in microvesicles obtained

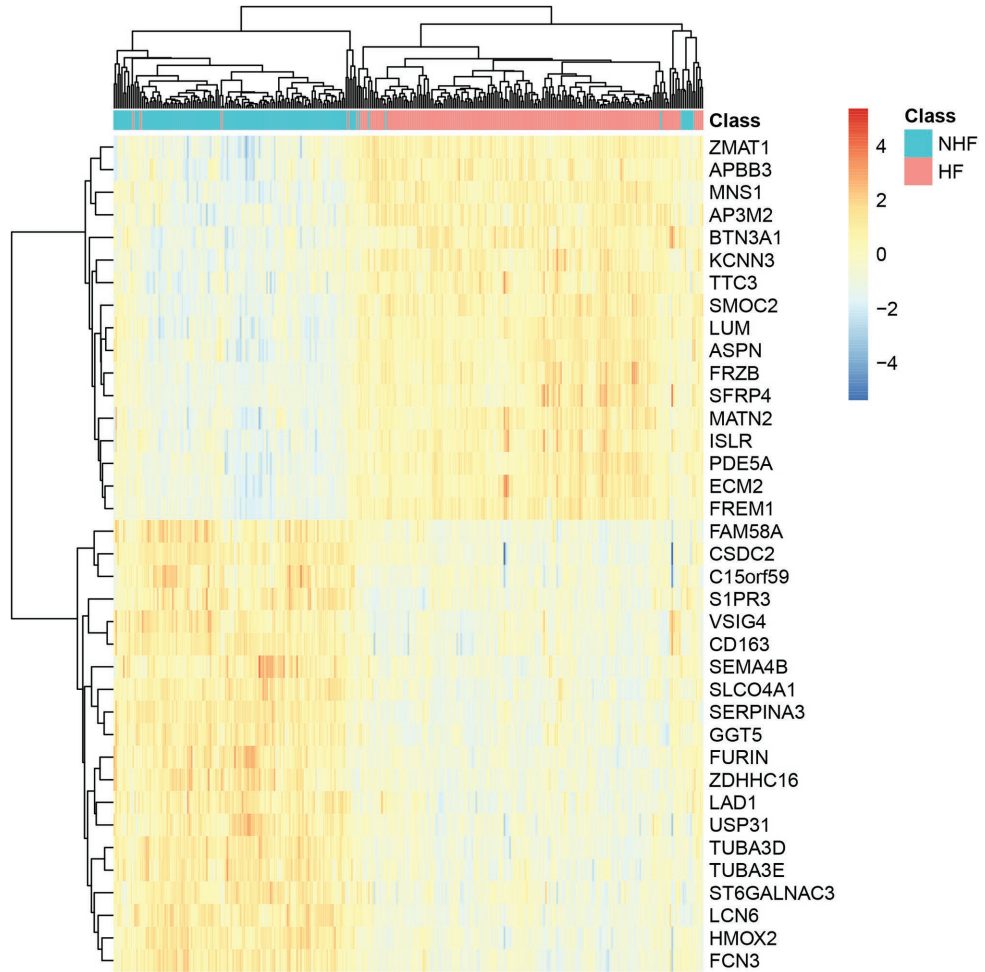

FIGURE 3 | The heatmap of the 38 genes in 177 HF and 136 non-heart failure (NHF) patients. Most samples were clustered into the correct groups. Only very few samples were misclustered. Within the 38 genes, 17 genes were highly expressed in HF, and 20 genes were lowly expressed in HF.

from activated platelets and abdominal aortic aneurysm (AAA) tissue (Fernandez-García et al., 2017). There is an obvious relationship between increased *FCN3* plasma levels and AAA presence and progression.

HMGN2 binds nucleosomal DNA and is associated with transcriptionally active chromatin, which is the top-ranked feature recognized by our bioinformatics analysis. HMGN protein family could regulate chromatin structure and could influence epigenetic modifications. *HMGN2* regulates active and bivalent genes by promoting an epigenetic landscape of active histone modifications at promoters and enhancers (Garza-Manero et al., 2019). *HMGN2* protected corticogenesis *via* maintaining global chromatin accessibility at promoter regions, thus ensuring proper transcriptome regulation (Apelt et al., 2020; Gao et al., 2020). There are few studies to certificate the role of HMGN2 in the progress of HF.

SERPINA3 also called Alpha-1-Antichymotrypsin or ACT, is first discovered as a plasma protease inhibitor and a member of the serine protease inhibitor (Jiang et al., 2020). Previous study showed that *SERPINA3* emerged as a responsible cardiac secreted factor that is increased in HF patients could be the most robust and promising culprit and were related to long-term mortality. Additionally, several researches thought that mineralocorticoid receptor antagonists (MRAs) were associated to *SERPINA3* (Meijers et al., 2018). Gene expression of *SERPINA3* was significantly increased in the HF group. In circulating plasma, the level of *SERPINA3* in the HF group was confirmed significant increase by ELISA analysis. These results suggested that *SERPINA3* might play an important role in the progression of HF (Zhao et al., 2020). Asakura and Kitakaze (2009) proved that *SERPINA3* might become novel diagnostic and therapeutic targets linked to the pathophysiology of HF using seven microarray datasets previously reported.

Due to the length limitation of the article, we cannot describe all 38 selected genes in detail. After detailed literature review, we found that all the above-mentioned genes play a vital role in the progression of HF, which also verifies the reliability of

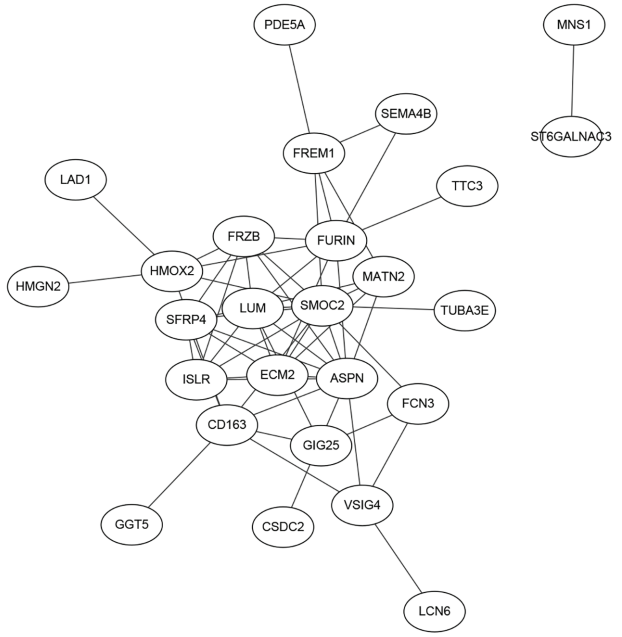

FIGURE 4 | The network of the 38 genes searched on the Search Tool for the Retrieval of Interacting Genes (STRING) database. We searched the interactions of the 38 genes on the STRING database and plotted the network. It can be seen that *SMOC2* located in the hub position of the network and its functions in HF was intensively studied.

our prediction model. We believe that these 38 selected genes are meaningful in the development of HF. They will contribute to the study of molecular mechanism, diagnosis, and treatment of HF, and will play an enlightening role in the future molecular biology research.

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.

AUTHOR CONTRIBUTIONS

DL: conception and design, administrative support, and provision of study materials or patients. HL: collection and assembly of data. HL and LL: data analysis and interpretation. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Scientific Research Project of Heilongjiang Health and Family Planning Commission (2017-017).

REFERENCES

Apelt, K., Zoutendijk, I., Gout, D. Y., Wondergem, A. P., van den Heuvel, D., and Luijsterburg, M. S. (2020). Human HMGN1 and HMGN2 are not required for transcription-coupled DNA repair. *Sci. Rep.* 10:4332. doi: 10.1038/s41598-020-61243-4

Asakura, M., and Kitakaze, M. (2009). Global gene expression profiling in the failing myocardium. *Circ. J.* 73, 1568–1576. doi: 10.1253/circj.cj-09-0465

Bowen, R. E. S., Graetz, T. J., Emmert, D. A., and Avidan, M. S. (2020). Statistics of heart failure and mechanical circulatory support in 2020. *Ann. Transl. Med.* 8:827. doi: 10.21037/atm-20-1127

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

Chen, L., Pan, X., Guo, W., Gan, Z., Zhang, Y. -H., Niu, Z., et al. (2020). Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* 112, 2524–2534. doi: 10.1016/j.ygeno.2020.02.004

Chen, L., Pan, X., Hu, X., Zhang, Y. -H., Wang, S., Huang, T., et al. (2018b). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

Chen, L., Pan, X., Zeng, T., Zhang, Y., Huang, T., and Cai, Y. (2019). Identifying essential signature genes and expression rules associated with distinctive development stages of early embryonic cells. *IEEE Access* 7, 128570–128578. doi: 10.1109/ACCESS.2019.2939556

Davison, B., and Cotter, G. (2015). Why is heart failure so important in the 21st century? *Eur. J. Heart Fail.* 17, 122–124. doi: 10.1002/ejhf.219

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

Fernandez-García, C. E., Burillo, E., Lindholt, J. S., Martinez-Lopez, D., Pilely, K., Mazzeo, C., et al. (2017). Association of ficolin-3 with abdominal aortic aneurysm presence and progression. *J. Thromb. Haemost.* 15, 575–585. doi: 10.1111/jth.13608

Gao, X. L., Tian, W. J., Liu, B., Wu, J., Xie, W., and Shen, Q. (2020). High-mobility group nucleosomal binding domain 2 protects against microcephaly by maintaining global chromatin accessibility during corticogenesis. *J. Biol. Chem.* 295, 468–480. doi: 10.1074/jbc.RA119.010616

Garza-Manero, S., Sindi, A. A. A., Mohan, G., Rehbini, O., Jeantet, V. H. M., Bailo, M., et al. (2019). Maintenance of active chromatin states by HMGN2 is required for stem cell identity in a pluripotent stem cell model. *Epigenetics Chromatin* 12:73. doi: 10.1186/s13072-019-0320-7

Jarcho, J. A. (2020). More evidence for SGLT2 inhibitors in heart failure. *N. Engl. J. Med.* 383, 1481–1482. doi: 10.1056/NEJMe2027915

Jiang, Z., Guo, N., and Hong, K. (2020). A three-tiered integrative analysis of transcriptional data reveals the shared pathways related to heart failure from different aetiologies. *J. Cell. Mol. Med.* 24, 9085–9096. doi: 10.1111/jcmm.15544

Kittleson, M. M., Ye, S. Q., Irizarry, R. A., Minhas, K. M., Edness, G., Conte, J. V., et al. (2004). Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. *Circulation* 110, 3444–3451. doi: 10.1161/01.Cir.0000148178.19465.11

Laugier, L., Frade, A. F., Ferreira, F. M., Baron, M. A., Teixeira, P. C., Cabantous, S., et al. (2017). Whole-genome cardiac DNA methylation fingerprint and gene expression analysis provide new insights in the pathogenesis of chronic Chagas disease cardiomyopathy. *Clin. Infect. Dis.* 65, 1103–1111. doi: 10.1093/cid/cix506

Li, J., and Huang, T. (2018). Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies. *Biochim. Biophys. Acta Mol. basis Dis.* 1864, 2241–2246. doi: 10.1016/j.bbadis.2017.10.036

Li, J., Lu, L., Zhang, Y. -H., Xu, Y., Liu, M., Feng, K., et al. (2020a). Identification of leukemia stem cell expression signatures through Monte Carlo Feature Selection strategy and support vector machine. *Cancer Gene Ther.* 27, 56–69. doi: 10.1038/s41417-019-0105-y

Li, J., Xu, Q., Wu, M., Huang, T., and Wang, Y. (2020b). Pan-cancer classification based on self-normalizing neural networks and feature selection. *Front. Bioeng. Biotechnol.* 8:766. doi: 10.3389/fbioe.2020.00766

Liu, Y., Morley, M., Brandimarto, J., Hannenhalli, S., Hu, Y., Ashley, E. A., et al. (2015). RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics* 105, 83–89. doi: 10.1016/j.ygeno.2014.12.002

McMurray, J. J., and Pfeffer, M. A. (2005). Heart failure. *Lancet* 365, 1877–1889. doi: 10.1016/S0140-6736(05)66621-4

Meijers, W. C., Maglione, M., Bakker, S. J. L., Oberhuber, R., Kieneker, L. M., de Jong, S., et al. (2018). Heart failure stimulates tumor growth by circulating factors. *Circulation* 138, 678–691. doi: 10.1161/circulationaha.117.030816

Pan, X., Chen, L., Feng, K. Y., Hu, X. H., Zhang, Y. H., Kong, X. Y., et al. (2019a). Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms. *Int. J. Mol. Sci.* 20:2185. doi: 10.3390/ijms20092185

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

Prohászka, Z., Munthe-Fog, L., Ueland, T., Gombos, T., Yndestad, A., Förhécz, Z., et al. (2013). Association of ficolin-3 with severity and outcome of chronic heart failure. *PLoS One* 8:e60976. doi: 10.1371/journal.pone.0060976

Ramachandra, C. J. A., Hernandez-Resendiz, S., Crespo-Avilan, G. E., Lin, Y. H., and Hausenloy, D. J. (2020). Mitochondria in acute myocardial infarction and cardioprotection. *EBioMedicine* 57:102884. doi: 10.1016/j.ebiom.2020.102884

Shi, X., Shao, X., Liu, B., Lv, M., Pandey, P., Guo, C., et al. (2020). Genome-wide screening of functional long noncoding RNAs in the epicardial adipose tissues of atrial fibrillation. *Biochim. Biophys. Acta Mol. Basis Dis.* 1866:165757. doi: 10.1016/j.bbadis.2020.165757

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2018). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Tan, F. L., Moravec, C. S., Li, J., Apperson-Hansen, C., McCarthy, P. M., Young, J. B., et al. (2002). The gene expression fingerprint of human heart failure. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11387–11392. doi: 10.1073/pnas.162370099

Tao, X., Wu, X., Huang, T., and Mu, D. (2020). Identification and analysis of dysfunctional genes and pathways in $CD8^+$ T cells of non-small cell lung cancer based on RNA sequencing. *Front. Genet.* 11:352. doi: 10.3389/fgene.2020.00352

van Oort, R. J., Garbino, A., Wang, W., Dixit, S. S., Landstrom, A. P., Gaur, N., et al. (2011). Disrupted junctional membrane complexes and hyperactive ryanodine receptors after acute junctophilin knockdown in mice. *Circulation* 123, 979–988. doi: 10.1161/Circulationaha.110.006437

Vigil-Garcia, M., Demkes, C. J., Eding, J. E. C., Versteeg, D., de Ruiter, H., Perini, I., et al. (2020). Gene expression profiling of hypertrophic cardiomyocytes identifies new players in pathological remodeling. *Cardiovasc. Res.* cvaa233. doi: 10.1093/cvr/cvaa233 [Epub ahead of print]

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

Wang, D., Li, J. R., Zhang, Y. H., Chen, L., Huang, T., and Cai, Y. D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 9:155. doi: 10.3390/genes9030155

Williams, J. L., Cavus, O., Loccoh, E. C., Adelman, S., Daugherty, J. C., Smith, S. A., et al. (2018). Defining the molecular signatures of human right heart failure. *Life Sci.* 196, 118–126. doi: 10.1016/j.lfs.2018.01.021

Yuan, F., Pan, X., Zeng, T., Zhang, Y. -H., Chen, L., Gan, Z., et al. (2020). Identifying cell-type specific genes and expression rules based on single-cell transcriptomic atlas data. *Front. Bioeng. Biotechnol.* 8:350. doi: 10.3389/fbioe.2020.00350

Zhang, Y. -H., Pan, X., Zeng, T., Chen, L., Huang, T., and Cai, Y. -D. (2020b). Identifying the RNA signatures of coronary artery disease from combined lncRNA and mRNA expression profiles. *Genomics* 112, 4945–4958. doi: 10.1016/j.ygeno.2020.09.016

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochim. Biophys. Acta* 1860, 2750–2755. doi: 10.1016/j.bbagen.2016.06.003

Zhang, S., Zeng, T., Hu, B., Zhang, Y. -H., Feng, K., Chen, L., et al. (2020a). Discriminating origin tissues of tumor cell lines by methylation signatures and dys-methylated rules. *Front. Bioeng. Biotechnol.* 8:507. doi: 10.3389/fbioe.2020.00507

Zhao, L., Guo, Z., Wang, P., Zheng, M., Yang, X., Liu, Y., et al. (2020). Proteomics of epicardial adipose tissue in patients with heart failure. *J. Cell. Mol. Med.* 24, 511–520. doi: 10.1111/jcmm.14758

Zhou, N., Chen, X., Xi, J., Ma, B., Leimena, C., Stoll, S., et al. (2020). Genomic characterization reveals novel mechanisms underlying the valosin-containing protein-mediated cardiac protection against heart failure. *Redox Biol.* 36:101662. doi: 10.1016/j.redox.2020.101662

Zhu, J., Yan, Q., Wang, J., Chen, Y., Ye, Q., Wang, Z., et al. (2020). The key genes for perineural invasion in pancreatic ductal adenocarcinoma identified with Monte-Carlo Feature Selection method. *Front. Genet.* 11:554502. doi: 10.3389/fgene.2020.554502

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: 16 November 2020
doi: 10.3389/fgene.2020.598185

Prediction of Potential Associations Between miRNAs and Diseases Based on Matrix Decomposition

Pengcheng Sun, Shuyan Yang, Ye Cao, Rongjie Cheng and Shiyu Han**

Department of Obstetrics and Gynecology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

It is known that miRNA plays an increasingly important role in many physiological processes. Disease-related miRNAs could be potential biomarkers for clinical diagnosis, prognosis, and treatment. Therefore, accurately inferring potential miRNAs related to diseases has become a hot topic in the bioinformatics community recently. In this study, we proposed a mathematical model based on matrix decomposition, named MFMDA, to identify potential miRNA–disease associations by integrating known miRNA and disease-related data, similarities between miRNAs and between diseases. We also compared MFMDA with some of the latest algorithms in several established miRNA disease databases. MFMDA reached an AUC of 0.9061 in the fivefold cross-validation. The experimental results show that MFMDA effectively infers novel miRNA–disease associations. In addition, we conducted case studies by applying MFMDA to three types of high-risk human cancers. While most predicted miRNAs are confirmed by external databases of experimental literature, we also identified a few novel disease-related miRNAs for further experimental validation.

Keywords: miRNA, matrix decomposition (MFMDA), endometrial cancer, miRNA–disease association, computational prediction model

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd., China

Reviewed by:
Ali Salehzadeh-Yazdi,
University of Rostock, Germany
JunLin Xu,
Hunan University, China
Lan Yu,
Inner Mongolia People's Hospital, China

****Correspondence:***
Shiyu Han
shiyuhan62@163.com
Rongjie Cheng
rongjie_jie@126.com

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

Received: *24 August 2020*
Accepted: *22 October 2020*
Published: *16 November 2020*

Citation:
Sun P, Yang S, Cao Y, Cheng R and Han S (2020) Prediction of Potential Associations Between miRNAs and Diseases Based on Matrix Decomposition. Front. Genet. 11:598185. doi: 10.3389/fgene.2020.598185

INTRODUCTION

Non-coding RNA (ncRNA) is a type of RNA that cannot be translated into protein. Although ncRNA cannot be translated into protein, its target gene can be regulated at the post-transcriptional level, thereby affecting disease (Hammond, 2015). A large amount of research evidence indicates that mutations and disorders of ncRNA are important causes of disease. Therefore, the identification of disease-related ncRNA has become an important topic in the field of biological research in recent years. ncRNA is a huge family and can be divided into housekeeper ncRNA and regulatory ncRNA (Kapranov et al., 2007; Lindsay et al., 2017). Housekeeping ncRNA is closely related to cell function, mainly involved in gene translation, gene splicing, gene modification, etc. The main function of regulating ncRNA is to regulate the expression level of genes. As regulatory ncRNA, miRNA is a class of non-coding single-stranded RNA molecules with a length of 22 nucleotides encoded by endogenous genes. They participate in the regulation of post-transcriptional gene expression in animals and plants (Taft et al., 2007; Chen et al., 2015). So far, 28645 miRNA molecules have been found in animals, plants, and viruses. Most miRNA genes exist in the genome in the form of single copies, multiple copies, or gene clusters (Wang and Chang, 2011).

In recent years, more and more studies have shown that miRNA plays a huge role in the process of cell differentiation, biological development, and disease development, which has also attracted more researchers' attention (Xu et al., 2004; Jiang et al., 2012; Li et al., 2014; Kang et al., 2020). With further in-depth research on the mechanism of action of miRNA, and the use of the latest high-throughput technologies such as miRNA chips to study the relationship between miRNA and disease, people will make higher eukaryote gene expression regulation Network understanding has improved to a new level (Cui et al., 2006). This will also make miRNA a new biological marker for disease diagnosis; it may also make this molecule a drug target, or simulate this molecule for new drug development, which will likely provide a new treatment for human diseases (Goh et al., 2016).

However, using biological experiments to identify disease-associated miRNAs is expensive and time-consuming, and it is blind. Therefore, there is an urgent need for simple and effective computational prediction models for predicting disease-related miRNAs. With the rapid development of high-throughput sequencing technology, more and more omics data are published, which also provides data support for the study of computational prediction models (Yi et al., 2017).In recent years, many scholars have proposed some effective computational models for predicting miRNA related to complex diseases. According to their respective implementation strategies, we can roughly divide these methods into machine-based computational prediction methods and network-based computational prediction methods (Zou et al., 2016).

Machine learning-based computational prediction methods predict the association of potential miRNAs with the disease can be divided into supervised-based machine learning methods and semi-supervised-based machine learning methods. The method based on supervision is mainly based on labeling sample set and label-less sample set to construct a machine learning model. Jiang et al. extracted feature sets based on known and unknown associations for training support vector machine (SVM) classifiers to predict potential miRNAs and disease associations, and achieved comparative prediction performance through cross-validation (Maly et al., 2019). Qu et al. (Zou et al., 2015) developed a new calculation method based on the KATZ model to predict MiRNA disease association (KATZMDA) by integrating multiple data sources. Based on the known miRNA–disease association in the HMDD database, Li et al. (2017) developed a MiRNA–disease association prediction model (MCMDA) called the matrix completion algorithm. The MCMDA model uses a matrix completion algorithm to update the adjacency matrix of known miRNA–disease associations and further predict potential associations. Xu et al. (Chen et al., 2018) proposed a method based on low-rank matrix completion to predict miRNA–disease association (LRMCMDA). LRMCMDA first constructs negative samples based on known associations, and then uses a low-rank matrix to complete the model to infer all miRNA and disease associations. Cross-validation shows that the model has obtained reliable prediction performance. However, although this supervised machine learning method uses different ways to define negative sample data, it is difficult to deal with the actual situation in any way, which will affect the prediction performance. In order to overcome this limitation, Chen and Yan (2014) proposed a least-squares-based semi-supervised machine learning method for predicting the association of potential miRNAs with disease, referred to as RLSMDA for short. The RLSMDA method constructs a continuous classifier function, and the predicted value reflects the probability score between specific miRNAs and specific diseases. This method can obtain the predicted values of all miRNAs and diseases at the same time, and does not require negative sample data. In addition, the RLSMDA method can also predict miRNAs associated with isolated diseases. Xu et al. (2019) designed a set of probabilistic matrix decomposition algorithms by integrating the similarity of miRNAs with diseases, using known correlation matrices and integrated similarity matrices to identify miRNAs that are potentially related to diseases. Luo et al. (2017) proposed a semi-supervised method called KRLSM to reveal the association between miRNA and disease. Machine learning has been a hot topic in recent years, and some machine learning methods can be used to solve this problem. Despite the outstanding contributions made by existing methods, there is still room for improvement in prediction accuracy.

In addition to machine learning-based methods, network-based methods to predict disease-related miRNAs have also attracted the attention of many researchers. Such methods are mainly based on a common biological hypothesis, "miRNAs with similar functions are more likely to be associated with disease phenotypes with similar functions, and vice versa" (Jiang et al., 2010). Based on this basic assumption, Jiang et al. proposed a new method that uses Bayesian models to integrate genomic data to rank disease-related miRNAs. Chen et al. (2012) adopted the global network similarity measure and proposed an improved restart-based random walk model (RWRMDA) to predict the association between miRNAs and disease. Yet, this method is not suitable for predicting new disease-related miRNAs. Xuan et al. (2013) integrated the information entropy of disease entries and the similarity of disease phenotypes to measure the functional similarity of diseases and miRNAs, and gave greater weight to miRNAs belonging to the same family or the same cluster class, and proposed a k-nearest neighbor prediction model (HDMP) is used to predict disease-related miRNAs. This method has obtained reliable prediction performance, but also cannot predict miRNAs associated with isolated diseases. Later, Xuan et al. (Banys-Paluchowski et al., 2015) further proposed the MIDP method based on random walk. In this model, by assigning different weights to known and unknown nodes, the prior information of the topology is effectively integrated. In addition, the extended conversion on the double-layer network of miRNA diseases makes it possible to predict miRNAs associated with isolated diseases. You et al. (2017) proposed a path-based miRNA–disease association (PBMDA) prediction model by integrating known human miRNA–disease associations, miRNA functional similarities, disease semantic similarities, and Gaussian interaction profiles for miRNA and disease similarities. The model constructs a heterogeneous graph composed of three interrelated subgraphs, and further uses a depth-first search algorithm to infer potential miRNA–disease associations.

The results show that reliable performance is obtained. Gu et al. (2016) created a network consistency projection algorithm to identify potential associations (NCPMDA) by integrating similarity networks and association networks. The biggest advantage of these methods is that they can predict isolated miRNAs associated with disease, but the performance obtained is not very satisfactory.

Although research on miRNA disease association prediction models has made some progress, there is still room to further improve the prediction performance of the model. In this study, we propose a predictive model called matrix decomposition, which fully considers the similarity between miRNAs and the similarity between diseases. In order to evaluate the effectiveness of MFMDA, we tested it using a global fivefold and local LOOCV framework. MFMDA is superior to the benchmark algorithm used for comparison, and achieves reliable performance in the framework of fivefold CV and local LOOCV (AUC 0.9061 and 0.7933) in the HMDD (V2.0) data set. To further prove the superiority of MFMDA, we analyzed three common diseases. Based on the analysis of the test results, we can find that 18 of the top 30 potential miRNAs related to the three diseases predicted by MFMDA have been confirmed by other databases.

MATERIALS AND METHODS

Human Disease–miRNA Interactome Network

In the past few decades, as the technology has matured, a large number of omics data have been published, including a large number of pairs related to miRNA diseases. Here, we use the known miRNAs and disease-associated data set HMDD V2.0 as the benchmark dataset (Huang et al., 2019a). The data set contains 495 miRNAs and 383 diseases and 5430 experimentally verified human-disease-related pairs. We use the adjacency matrix A to represent this confirmed association. Specifically, if the disease $d\,(i)$ was previously associated with miRNA $m\,(j)$, the value of A_{ij} is 1; otherwise, the corresponding position is set to 0.

miRNA Functions Similarly

Based on previous research, it is not difficult to find that miRNAs with similar functions are more likely to be related to similar diseases (Wang et al., 2010). Under this assumption, the miRNA functional similarity score was calculated[1]. Therefore, we constructed a functional similarity matrix FS between miRNAs based on these data, where $FS(m\,(i), m\,(j))$ represents the similarity between miRNA $m\,(i)$ and another miRNA $m\,(j)$.

Disease Semantic Similarity

Semantic similarity is a common way to express the similarity of diseases in this field. MFMDA uses a layered directed acyclic graph (DAG) to calculate the similarity between two diseases (Wang et al., 2010). Specifically, for disease d, let $DAG_d = (d, T_d, E_d)$ be a DAG, where T_d represents the ancestor node set of d (including itself) and E_d represents the hierarchical connection between diseases defined by the MeSH disease tree structure of the National Library of Medicine. For any $t \in T_d$, MFMDA defines the semantic contribution of disease t to d as:

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

Where Δ is the semantic decay factor, which is set to 0.5 in the iterative equation according to previous researches (Dong et al., 2019; Marcuello et al., 2019). Therefore, the semantic similarity between the diseases d_1 and d_2 can be defined as:

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

Gaussian Similarity of miRNA and Disease

Among various similarity measurement algorithms, Gaussian similarity is a very good measurement method, which has been widely used in various fields. Let $VP(m_i)$ be the vector related to miRNA $m_i$in Y, i.e., the i^{th} column of Y. Then, the Gaussian similarity between the diseases m_i and $m_j$is calculated as follows:

$$KM\left(r_i, r_j\right) = exp(-\gamma_m||VP\,(r_i) - VP(r_j)||^2) \quad (3)$$

Where γ_m is the adjustment parameter of the bandwidth (van Laarhoven et al., 2011). The update rule of parameter γ_m is as follows:

$$\gamma_m = \gamma_m^{'} / \left(\frac{1}{nm}\sum_{i=1}^{nm} ||VP\,(r_i)\,||^2\right) \quad (4)$$

Similarly, the Gaussian similarity between miRNAs can be defined as follows:

$$KD\left(d_i, d_j\right) = exp(-\gamma_d||VP\left(d_i\right) - VP(d_j)||^2) \quad (5)$$

$$\gamma_d = \gamma_d^{'} / \left(\frac{1}{nd}\sum_{i=1}^{nd} ||VP\left(d_i\right)\,||^2\right) \quad (6)$$

Integrated Similarity for Diseases and miRNAs

In order to obtain a more comprehensive disease similarity, the semantic similarity of the disease is combined with the Gaussian interactive contour kernel similarity through the following piecewise function to obtain the final similarity between the diseases:

$$S_d\left(d_i, d_j\right) = \begin{cases} D\left(d_i, d_j\right) & d_i\ and\ d_j\ has\ semantic\ similarity \\ KD\left(d_i, d_j\right) & otherwise \end{cases} \quad (7)$$

Similarly, the similarity between miRNAs can also be redefined as:

$$S_m\left(\mathrm{m}_i, m_j\right) = \begin{cases} FS\left(\mathrm{m}_i, m_j\right) & r_i\ and\ r_j\ has\ functional\ similarity \\ KM\left(\mathrm{m}_i, m_j\right) & otherwise \end{cases} \quad (8)$$

[1] http://www.cuilab.cn/files/images/cuilab/misim.zip

MFMDA

Matrix factorization (MF) is an effective technique that has been widely used in data representation (Huang and Zheng, 2006; Hosoda et al., 2009; Zheng et al., 2009; Xu et al., 2020). It aims to find two matrices whose product provides the best approximation to the original matrix. Given a miRNAs–diseases association matrix, MF can be decomposed into two matrices $Y = R^{n\times m}$, that is, $W \in R^{n\times k}$ and $H \in R^{m\times k}$,and $Y \approx UV^T$. Here, we use mathematical formulas to express the potential association prediction problem between diseases and miRNAs as the following objective function:

$$\min_{U,V} ||I\cdot(Y - WH^T)||_F^2 \tag{9}$$

where $||\cdots||_F^2$ represents the Frobenius norm and $\cdot$ denotes the Hadamard product of two matrices, that is, the multiplication of the corresponding elements of the matrix, and $I_{ij} = 0$ if the entry (i, j) in Y is missing, and 1 otherwise.

The standard MF in Eq. 2 is just to find two matrices, and their product tries to approximate the original matrix. However, the effects caused by the similarity between miRNAs and diseases are ignored. Suppose the functions of the two miRNAs are very similar, and at the same time, the diseases implicitly learned that they should have a similar distance in the vector space. The diseases dimension is the same. For the same reason, the miRNAs size can also use this idea to constrain the drug's implicit representation. That is, if the two diseases are similar, the distance of the miRNAs in the low-dimensional vector space should also be small.

$$\begin{aligned}\min_{U,V} ||I\cdot\left(Y - WH^T\right)||_F^2 + \lambda_l\left(||W||_F^2 + ||H||_F^2\right)\\ +\lambda_v\sum_{i,p=1}^{n}||w_i - w_p||^2 S_{i,p}^{m*}\\ +\lambda_d\sum_{j,k=1}^{m}||h_j - h_k||^2 S_{j,k}^{d*}\end{aligned} \tag{10}$$

where λ_l, λ_d, and λ_v are the regularization coefficients; w_i and h_j are the ith and jth rows of W and H, respectively. S^{v*} is the hidden social similarity between miRNAs and S^{d*} is the hidden social similarity between diseases.

Optimization

In order to solve the local optimal solution problem of Eq. 3, we use the gradient descent algorithm to solve. According to the nature of the Frobenius norm, the corresponding Lagrange function L_E of Eq. 2 can be redefined as:

$$\begin{aligned}L_E = Tr\left(I\cdot\left(YY^T - 2*YHW^T + WH^THW^T\right)\right) +\\ \lambda_l Tr\left(WW^T\right) + \lambda_l\, Tr\left(HH^T\right) + \lambda_m Tr\left(W^T L_m W\right) +\\ \lambda_d Tr\left(H^T L_d H\right) + Tr\left(\emptyset W^T\right) + Tr\left(\psi H^T\right)\end{aligned} \tag{11}$$

where $Tr(\cdots)$ represents the trace of a matrix; $L_m = D_m - S^{m*}$ and $L_d = D_d - S^{d*}$ are the graph Laplacian matrices for S^{m*} and S^{d*}, respectively; and D_m and D_m are the diagonal matrices whose entries are row (or column) sums of S^{m*} and S^{d*}, respectively.

The partial derivatives of the above functions with respect to W and H are:

$$\begin{aligned}\frac{\partial L_E}{\partial W} &= -2YH + 2WH^TH + 2\lambda_l W + 2\lambda_m L_m W + \emptyset\\ \frac{\partial L_E}{\partial H} &= -2Y^TW + 2HW^TW + 2\lambda_l H + 2\lambda_d L_d H + \psi\end{aligned} \tag{12}$$

According to the solution conditions of Karush–Kuhn–Tucker (KKT) (Facchinei et al., 2013), we can make $\emptyset_{ik}w_{ik} = 0$ and $\psi_{jk}h_{jk} = 0$, thus obtain the following equations for w and h:

$$\begin{aligned}-(YH)_{ik}\,w_{ik} + \left(WH^TH\right)_{ik}w_{ik} + (\lambda_l W)_{ik}\,w_{ik} +\\ \left(\lambda_m(D_m - S^{m*})W\right)_{ik}w_{ik} = 0\\ -\left(Y^TW\right)_{jk}h_{jk} + \left(HW^TW\right)_{jk}h_{jk} + (\lambda_l H)_{jk}\,h_{jk} +\\ \left(\lambda_d(D_d - S^{d*})H\right)_{jk}h_{jk} = 0.\end{aligned} \tag{13}$$

Therefore, we get the w_{ik} and h_{jk} update rules as follows:

$$\begin{aligned}w_{ik} &= w_{ik}\frac{(YH + \lambda_m S^{m*}W)_{ik}}{(WH^TH + \lambda_l W + \lambda_m D_m W)_{ik}}\\ h_{jk} &= h_{jk}\frac{(Y^TW + \lambda_d S^{d*}H)_{jk}}{(HW^TW + \lambda_l H + \lambda_d D_d H)_{jk}}\end{aligned} \tag{14}$$

The matrices W and H are updated based on Eq. 3 until convergence. Finally, we can obtain the predicted miRNAs–diseases association matrix as $Y^* = WH^T$, and determine the priority of potential miRNAs and disease according to the value in the matrix Y^*. In principle, the miRNAs with the highest grade in Y^* are more likely to be associated with the disease. The flow chart of MFMDA is shown in **Figure 1**.

RESULTS

Evaluation of Prediction Performance

There are many performance indicators for evaluating prediction models. In this field, ROC curve and AUC value, PR curve, and AUPR value are usually used to evaluate the performance of the algorithm (Chen and Huang, 2017; Chen et al., 2020).

The ROC curve, also called receiver operating characteristic curve or susceptibility curve, is a comprehensive indicator reflecting sensitivity and specificity. The ROC curve graphically reveals the correlation between sensitivity and specificity. By setting different thresholds, a series of corresponding sensitivities and specificities are calculated, and then plotted with the true positive rate on the ordinate and false positive rate on the abscissa curve. The simple assumption is that for binary classification problems (only two types, positive and negative samples), the

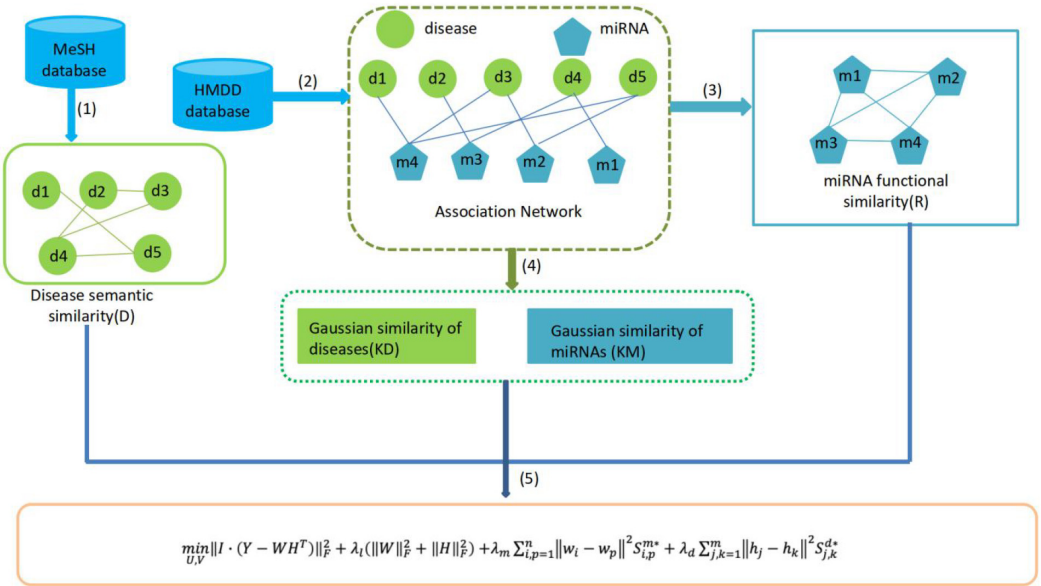

FIGURE 1 | Diagram of MFMDA for predicting potential miRNA–disease associations.

calculation methods of TPR and FPR are shown in Eq. 15.

$$\text{TPR} = \frac{TP}{TP + FN}, \ \text{FPR} = \frac{FP}{TN + FP} \tag{15}$$

TP refers to the number of positive samples that are correctly predicted, that is, the number of positive samples that are predicted as positive samples; FP refers to the number of positive samples that are incorrectly predicted, that is, the number of negative samples that are predicted to be positive samples; the number of negative samples correctly predicted, that is, the number of negative samples predicted as negative samples; FN refers to the number of negative samples that are incorrectly predicted, that is, the number of positive samples predicted as negative samples. The area under the line of the ROC curve is AUC. The more convex the ROC curve, the closer to the upper left corner. The larger the AUC value, the better the prediction performance. The AUC value is generally between 0.5 and 1. The AUC value of 0.5 is the effect of random prediction. The AUC value of 1 has the best performance and the perfect classifier, that is, it can correct all positive and negative classes.

The PR curve calculates a series of accuracy and recall by setting different thresholds, and then draws the curve as the precision ordinate and recall as the abscissa. The precision and recall are calculated into the formulas 16:

$$\text{precision} = \frac{TP}{TP + FP}, \ \text{FPR} = \frac{TP}{TP + FN}. \tag{16}$$

The PR curve reflects the correlation between accuracy and recall. The area under the PR curve is AUPR. The larger the AUPR value, the better the performance.

Comparison With Other Methods

We further compared the prediction performance of the MFMDA model with four benchmark prediction models (i.e., LRMCMDA, IMCMDA, NCPMDA, and RLSMDA). LRMCMDA and IMCMDA belong to the matrix completion algorithm, and have achieved good predictive performance in this field. NCPMDA is a network projection algorithm, which is one of the representatives of algorithms based on network prediction. RLSMDA is a semi-supervised learning method based on the Regularized Least Squares (RLS) framework, which represents a good opportunity to learn learning algorithms. Since the data used in this study are all from the public data set HMDD2.0, all the parameters of the comparison algorithm will also use the parameters given by the original author.

Performance on Predicting miRNA–Disease Association

We applied MFMDA, LRMCMDA, IMCMDA, NCPMDA, and RLSMDA to HMDD V2.0 miRNA–disease association data, which contains 5430 unique associations between 495 miRNAs and 383 diseases, and draws their ROC curves of the global fivefold CV in **Figure 2A**. As can be seen, the AUCs of MFMDA, LRMCMDA, IMCMDA, NCPMDA, and RLSMDA are 0.9061, 0.8883, 0.8364, 0.8637, and 0.8326, respectively, indicating that MFMDA performed best in predicting miRNA–disease associations.

However, considering the limited number of known and experimentally verified miRNA–disease associations, it is too arbitrary to use AUC to evaluate the performance of prediction methods. Therefore, we also include the exact recall (PR) curve and the AUPR in **Figure 2B** to supplement performance evaluation. As shown in **Figure 2B**, the AUPR of MMFDA,

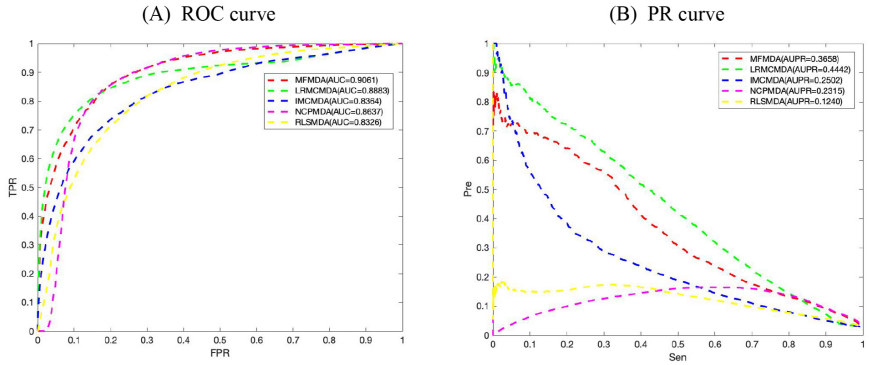

FIGURE 2 | Comparison of MFMDA with four best performers for miRNA–disease associations. **(A)** ROC curves for fivefold cross validation. **(B)** Precision–recall (PR) curve for fivefold cross validation.

LRMCMDA, IMCMDA, NCPMDA, and RLSMDA are 0.3658, 0.4442, 0.2502, 0.2315, and 0.1240, which again shows that MFMDA performs better than most algorithms in predicting miRNA–disease associations and can be a supplement to the existing computational prediction model.

Predicting Novel Disease-Related miRNAs

For a new disease, if it can find its related miRNAs, it will provide a great help for people to understand the pathogenesis of the disease. Therefore, we performed CV_d experiment to test the performance of MFMDA in predicting miRNAs associated to a novel disease d. In CV_d: CV on disease d_i, we remove all the known miRNA–disease association of the disease d_i (column vectors in matrix $Y \in R^{m \times n}$) and build prediction model (for inferring the deleted associations) using the remaining data. As shown in **Figure 3**, the AUC value obtained by MFMDA is second only to LRMCMDA, which also indicates that MFMDA is also relatively good at predicting miRNAs related to new diseases. Of course, although LRMCMDA is more effective at predicting new disease-related miRNAs, LRMCMDA uses network projection to construct negative samples. This method of constructing negative samples will be affected by the size of the data set, which will affect its prediction performance. Presumably, MFMDA is a semi-supervised algorithm, it does not need to construct negative samples and the prediction performance is relatively stable.

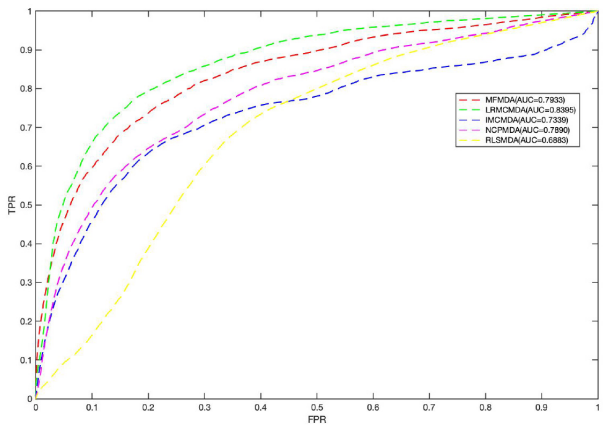

FIGURE 3 | Comparison between MFMDA and benchmark algorithms based on local LOOCV.

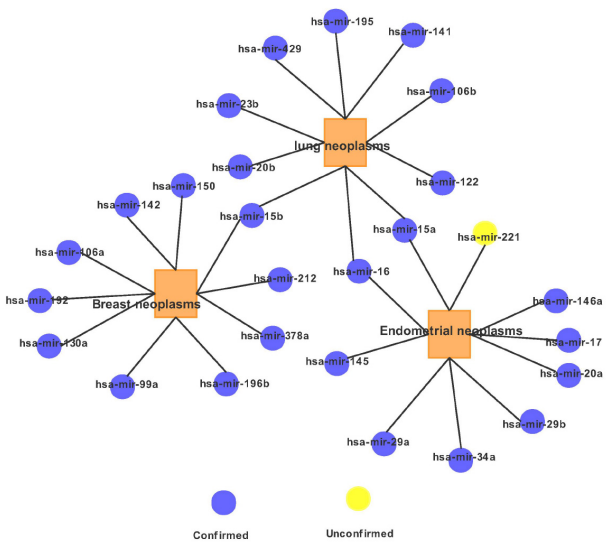

FIGURE 4 | Network of the top 10 predicted associations for the three diseases via MFMDA.

Finally, we explored the effect of the disease similarity and miRNA similarity on prediction performance. Specifically, we performed global fivefold CV with parameters $\lambda_m$or λ_d from 0.2 to 1 and a step size of 0.2 (**Table 1**). We can see that the two similarities really help predict performance. However, as the parameters continue to increase, the performance of the prediction is constantly decreasing.

TABLE 1 | Prediction AUCs of MFMDA at different choices of parameters.

MFMDA	$\lambda_m = \lambda_d = 0.2$	$\lambda_m = \lambda_d = 0.4$	$\lambda_m = \lambda_d = 0.6$	$\lambda_m = \lambda_d = 0.8$	$\lambda_m = \lambda_d = 1$
AUC	0.9061	0.9058	0.9013	0.8924	0.8912

TABLE 2 | The top 10 potential miRNA candidates detected by MFMDA for endometrial neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Endometrial neoplasms	9	1	hsa-mir-146a	HMDD V3.0	6	hsa-mir-34a	HMDD V3.0
		2	hsa-mir-221	Unconfirmed	7	hsa-mir-29a	HMDD V3.0
		3	hsa-mir-20a	HMDD V3.0	8	hsa-mir-145	HMDD V3.0
		4	hsa-mir-17	HMDD V3.0	9	hsa-mir-15a	HMDD V3.0
		5	hsa-mir-16	HMDD V3.0	10	hsa-mir-29b	HMDD V3.0

Case Study

Next, three disease case studies were conducted to further validate the predictive power of the new miRNA disease pairs discovered by MFMDA. We first use the verified HMDD V2.0 pair as a training sample. For each predicted disease, the corresponding unverified miRNA is ranked according to the predicted score. Then, according to the other three well-known databases dbDEMC2.0 (Yang et al., 2017), miR2Disease (Jiang et al., 2009), and HMDD V3.0 (Huang et al., 2019b), the top 10 candidate miRNAs in the prediction list were examined.

Endometrial cancer is a group of epithelial malignant tumors that occur in the endometrium, and it occurs in perimenopausal and postmenopausal women. Endometrial cancer is one of the most common tumors of the female reproductive system. There are nearly 200,000 new cases each year, and it is the third most common gynecological malignant tumor that causes death. Earlier studies have shown that the differential expression of miRNA in endometrial adenocarcinoma can play a key auxiliary role in understanding the diagnosis and treatment of endometrial adenocarcinoma (Jurcevic et al., 2014). Therefore, in this study, we used MFMDA to identify potential miRNAs associated with endometrial adenocarcinoma. Nine of the top 10 miRNAs found were confirmed by at least one external database (see **Table 2**).

In the second case study, we still choose the tumor that belongs to women with high incidence, namely, breast tumor. Breast tumors are malignant tumors that occur in the epithelial tissue of the breast glands. Currently, the treatment is mainly based on clinical and pathological features. Targeted therapy and personalized therapy are the ultimate goals. Related studies have shown that the occurrence of breast tumors is also related to abnormalities of related miRNAs. For example, an abnormal increase in miR-22 may promote the occurrence and metastasis of breast cancer and lead to a higher degree of tumor malignancy.

TABLE 3 | The top 10 potential miRNA candidates detected by MFMDA for breast neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Breast neoplasms	10	1	hsa-mir-150	dbDEMC 2.0	6	hsa-mir-130a	dbDEMC 2.0
		2	hsa-mir-142	dbDEMC 2.0	7	hsa-mir-99a	dbDEMC 2.0
		3	hsa-mir-15b	dbDEMC 2.0	8	hsa-mir-196b	dbDEMC 2.0
		4	hsa-mir-106a	dbDEMC 2.0	9	hsa-mir-378a	dbDEMC 2.0
		5	hsa-mir-192	dbDEMC 2.0	10	hsa-mir-212	dbDEMC 2.0

TABLE 4 | The top 10 potential miRNA candidates detected by MFMDA for lung neoplasms.

Cancer	No. of confirmed miRNAs	Top 10 ranked predictions					
		Rank	miRNAs	Evidences	Rank	miRNAs	Evidences
Lung neoplasms	9	1	hsa-mir-16	miR2Disease	6	hsa-mir-141	miR2Disease
		2	hsa-mir-122	dbDEMC 2.0	7	hsa-mir-195	miR2Disease
		3	hsa-mir-15a	dbDEMC 2.0	8	hsa-mir-429	miR2Disease
		4	hsa-mir-15b	Unconfirmed	9	hsa-mir-23b	dbDEMC 2.0
		5	hsa-mir-106b	dbDEMC 2.0	10	hsa-mir-20b	dbDEMC 2.0

Therefore, predicting miRNAs related to breast tumors through related algorithms will also provide corresponding help for human breast cancer treatment. As shown in **Table 3**, we found that the top 10 miRNAs predicted by MFMDA related to breast cancer have all been confirmed by relevant databases.

Finally, we conduct prediction studies on miRNAs associated with lung tumors. Lung cancer is one of the fastest growing morbidity and mortality rates, and the most threatening to the health and life of the population. In the past 50 years, many countries have reported that the incidence and mortality of lung cancer have increased significantly. The incidence and mortality of lung cancer in men accounted for the first place in all malignant tumors, the incidence in women accounted for the second place, and the mortality rate took the second place. Despite the important therapeutic value of chemotherapy, surgery is still the only way to treat lung cancer. There is an urgent need to find potential biomarkers that respond strongly to clinical observations. The researchers found that the expression level of miR-99a is related to the clinicopathological factors of lung cancer and lymph node metastasis. Identifying more miRNAs related to lung cancer helps to accurately assess clinical outcomes. Therefore, we conducted a lung cancer case study based on MFMDA. In the prediction list, nine of the top 10 predicted miRNAs confirmed their association with lung tumors (see **Table 4**).

For a clear view, we illustrate in **Figure 4** the association network of the top 10 predicted miRNA candidates for the three diseases. It is worth noting that some top candidates were found to be related to several diseases. For example: hsa-mir-15a has not only been shown to be related to the occurrence of endometrial neoplasms, but also has a certain relationship with lung neoplasms.

DISCUSSION

A large number of studies have shown that miRNA plays an increasingly important role in many physiological processes. Researchers are trying to identify disease-related miRNAs as valuable biomarkers that can be used for clinical measurement, diagnosis, prognosis, and treatment. Therefore, accurately inferring potential miRNAs related to diseases can help us study the pathogenesis of diseases and find more effective treatments. In this study, we proposed a mathematical model based on MF (MFMDA) to identify potential miRNA–disease associations. First, MFMDA not only uses known miRNA and disease-related data, but also integrates the similarities between miRNA and disease. Second, the model is a semi-supervised model, which does not rely on negative samples. Finally, in the process of solving the model, we use the alternating gradient descent algorithm to find the optimal solution to ensure a stable decomposition matrix. Experimental results show that, compared with other methods, MFDMA can effectively improve performance and is a powerful tool for discovering the association of potential diseases with miRNA. However, this method still has some limitations; we need to further optimize. For example, the similarity measure between diseases and miRNAs used by MFMDA is too single and may not be the best choice. How to integrate multiple omics information more effectively to improve prediction performance is also worthy of further research.

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

SH and RC designed the study. PS collected and wrote the manuscript. SY and YC reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Heilongjiang Postdoctoral Fund (No. LBH-Z18190), Wutong Tree Foundation of The Fourth Affiliated Hospital of Harbin Medical University (No. HYDSYWTS201904), and Heilongjiang Youth Science Foundation (No. QC2018100).

REFERENCES

Banys-Paluchowski, M., Schneck, H., Blassl, C., Schultz, S., Meier-Stiegen, F., Niederacher, D., et al. (2015). Prognostic relevance of circulating tumor cells in molecular subtypes of breast cancer. *Geburtshilfe Frauenheilkd.* 75, 232–237. doi: 10.1055/s-0035-1545788

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction[J]. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

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

Chen, X., Sun, L. G., and Zhao, Y. (2020). Ncmcmda: mirna–disease association prediction through neighborhood constraint matrix completion[J]. *Brief. Bioinform.* [Epub ahead of print].

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

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., Dai, Q., et al. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5:11338.

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

Cui, Q., Yu, Z., Purisima, E. O., and Wang, E. (2006). Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.* 2:46. doi: 10.1038/msb4100089

Dong, J., Zhu, D., Tang, X., Qiu, X., Lu, D., Li, B., et al. (2019). Detection of circulating tumor cell molecular subtype in pulmonary vein predicting prognosis of stage i-iii non-small cell lung cancer patients. *Front. Oncol.* 9:1139. doi: 10.3389/fonc.2019.01139

Facchinei, F., Kanzow, C., and Sagratella, S. (2013). Solving quasi-variational inequalities via their KKT conditions. *Math. Programm.* 144, 369–412. doi: 10.1007/s10107-013-0637-0

Goh, J. N., Loo, S. Y., Datta, A., Siveen, K. S., Yap, W. N., Cai, W., et al. (2016). microRNAs in breast cancer: regulatory roles governing the hallmarks of cancer. *Biol. Rev. Camb. Philos. Soc.* 91, 409–428. doi: 10.1111/brv.12176

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6:36054.

Hammond, S. M. (2015). An overview of microRNAs. *Adv. Drug Deliv. Rev.* 87, 3–14. doi: 10.1007/978-3-319-03725-7_1

Hosoda, K., Watanabe, M., Wersing, H., Körner, E., Tsujino, H., Tamura, H., et al. (2009). A model for learning topographically organized parts-based representations of objects in visual cortex: topographic nonnegative matrix factorization. *Neural Comput.* 21, 2605–2633. doi: 10.1162/neco.2009.03-08-722

Huang, D., and Zheng, C. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019a). HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res.* 47, D1013–D1017.

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019b). HMDD v3. 0: a database for experimentally supported human microRNA–disease associations[J]. *Nucleic Acids Res.* 47, D1013–D1017.

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4:S2. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37, D98–D104.

Jiang, W., Chen, X., Liao, M., Li, W., Lian, B., Wang, L., et al. (2012). Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Sci. Rep.* 2:282.

Jurcevic, S., Olsson, B., and Klinga-Levan, K. (2014). MicroRNA expression in human endometrial adenocarcinoma. *Cancer Cell Int.* 14:88.

Kang, B. J., Ra, S. W., Lee, K., Lim, S., Son, S. H., Ahn, J. J., et al. (2020). Circulating tumor cell number is associated with primary tumor volume in patients with lung adenocarcinoma. *Tuberc. Respir. Dis.* 83, 61–70. doi: 10.4046/trd.2019.0048

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

Li, J., Zhang, Y., Wang, Y., Zhang, C., Wang, Q., Shi, X., et al. (2014). Functional combination strategy for prioritization of human miRNA target. *Gene* 533, 132–141. doi: 10.1016/j.gene.2013.09.106

Li, J. Q., Rong, Z. H., Chen, X., Yan, G. Y., and You, Z. H. (2017). MCMDA Matrix completion for MiRNA-disease association. *Oncotarget* 8, 21187–21199. doi: 10.18632/oncotarget.15061

Lindsay, C. R., Faugeroux, V., Michiels, S., Pailler, E., Facchinetti, F., Ou, D., et al. (2017). A prospective examination of circulating tumor cell profiles in non-small-cell lung cancer molecular subgroups. *Ann. Oncol.* 28, 1523–1531. doi: 10.1093/annonc/mdx156

Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017). Predicting MicroRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data. *IEEE Access.* 5, 2503–2513. doi: 10.1109/access.2017.2672600

Maly, V., Maly, O., Kolostova, K., and Bobek, V. (2019). Circulating tumor cells in diagnosis and treatment of lung cancer. *In Vivo* 33, 1027–1037. doi: 10.21873/invivo.11571

Marcuello, M., Vymetalkova, V., Neves, R. P. L., Duran-Sanchon, S., Vedeld, H. M., Tham, E., et al. (2019). Circulating biomarkers for early detection and clinical management of colorectal cancer. *Mol. Aspects Med.* 69, 107–122.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29, 288–299. doi: 10.1002/bies.20544

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

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

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell.* 43, 904–914. doi: 10.1016/j.molcel.2011.08.018

Xu, J., Cai, L., Liao, B., Zhu, W., Wang, P., Meng, Y., et al. (2019). Identifying potential miRNAs-disease associations with probability matrix factorization. *Front. Genet.* 10:1234. doi: 10.3389/fgene.2019.01234

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi: 10.1093/bioinformatics/btaa109

Xu, P., Guo, M., and Hay, B. A. (2004). MicroRNAs and the regulation of cell death. *Trends Genet.* 20, 617–624. doi: 10.1016/j.tig.2004.09.010

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One* 8:e70204. doi: 10.1371/journal.pone.0070204

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* 45, D812–D818.

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

You, Z. H., Huang, Z. A., Zhu, Z., Yan, G. Y., Li, Z. W., Wen, Z., et al. (2017). PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005455. doi: 10.1371/journal.pcbi.1005455

Zheng, C., Huang, D. S., Zhang, L., and Kong, X. Z. (2009). Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Trans. Inform. Technol. Biomed.* 13, 599–607. doi: 10.1109/titb.2009.2018115

Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., et al. (2015). Prediction of MicroRNA-disease associations based on social network analysis methods. *Biomed. Res. Int.* 2015:810514.

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics* 15, 55–64.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 November 2020
doi: 10.3389/fgene.2020.600097

Integrative Analysis of Genomics and Transcriptome Data to Identify Regulation Networks in Female Osteoporosis

Xianzuo Zhang[1†], *Kun Chen*[1†], *Xiaoxuan Chen*[2], *Nikolaos Kourkoumelis*[3], *Guoyuan Li*[1], *Bing Wang*[4*] *and Chen Zhu*[1*]

[1] Department of Orthopedics, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China, [2] College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China, [3] Department of Medical Physics, School of Health Sciences, University of Ioannina, Ioannina, Greece, [4] School of Electrical and Information Engineering, Anhui University of Technology, Ma'anshan, China

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co. Ltd, China

Reviewed by:
Prashanth N. Suravajhala,
Birla Institute of Scientific Research, India
Lorena Aguilar Arnal,
National Autonomous University of Mexico, Mexico

***Correspondence:**
Bing Wang
wangb@ahut.edu.cn
Chen Zhu
zhuchena@ustc.edu.cn

[†] These authors have contributed equally to this work

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

Received: 28 August 2020
Accepted: 28 October 2020
Published: 30 November 2020

Citation:
Zhang X, Chen K, Chen X, Kourkoumelis N, Li G, Wang B and Zhu C (2020) Integrative Analysis of Genomics and Transcriptome Data to Identify Regulation Networks in Female Osteoporosis. Front. Genet. 11:600097. doi: 10.3389/fgene.2020.600097

Background: Osteoporosis is a highly heritable skeletal muscle disease. However, the genetic mechanisms mediating the pathogenesis of osteoporosis remain unclear. Accordingly, in this study, we aimed to clarify the transcriptional regulation and heritability underlying the onset of osteoporosis.

Methods: Transcriptome gene expression data were obtained from the Gene Expression Omnibus database. Microarray data from peripheral blood monocytes of 73 Caucasian women with high and low bone mineral density (BMD) were analyzed. Differentially expressed messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs) were identified. Differences in BMD were then attributed to several gene modules using weighted gene co-expression network analysis (WGCNA). LncRNA/mRNA regulatory networks were constructed based on the WGCNA and subjected to functional enrichment analysis.

Results: In total, 3,355 mRNAs and 999 lncRNAs were identified as differentially expressed genes between patients with high and low BMD. The WGCNA yielded three gene modules, including 26 lncRNAs and 55 mRNAs as hub genes in the blue module, 36 lncRNAs and 31 mRNAs as hub genes in the turquoise module, and 56 mRNAs and 30 lncRNAs as hub genes in the brown module. *JUN* and *ACSL5* were subsequently identified in the modular gene network. After functional pathway enrichment, 40 lncRNAs and 16 mRNAs were found to be related to differences in BMD. All three modules were enriched in metabolic pathways. Finally, mRNA/lncRNA/pathway networks were constructed using the identified regulatory networks of lncRNAs/mRNAs and pathway enrichment relationships.

Conclusion: The mRNAs and lncRNAs identified in this WGCNA could be novel clinical targets in the diagnosis and management of osteoporosis. Our findings may help elucidate the complex interactions between transcripts and non-coding RNAs and provide novel perspectives on the regulatory mechanisms of osteoporosis.

Keywords: osteoporosis, WGCNA (Weighted Gene Co-expression Network Analyses), pathway, biomarker, systems biology, LncRNA-long noncoding RNA

INTRODUCTION

Osteoporosis is a systemic disease of the musculoskeletal system. Its main pathophysiological characteristics are decreased bone mass, destruction of bone tissue microstructure, increased bone fragility, and increased fracture risk (Ensrud and Crandall, 2017). According to the National Health and Nutrition Examination Survey III, there are more than 9.9 million patients with osteoporosis in the United States of America, and 1.5 million patients suffer from osteoporotic fractures each year (Sahni et al., 2009). The social costs associated with osteoporosis are expected to rise as the population ages (Ruza et al., 2013). Affected by many factors, such as menopause, women are especially susceptible to osteoporosis (Baccaro et al., 2015). A large-scale epidemiological survey in 2006 showed that among people over 50 years old, the prevalence of osteoporosis in men was 14.4%, whereas that in women was as high as 20.7%(Chen et al., 2016). The lifetime risk of osteoporotic fractures in women is as high as 40%, which is significantly higher than the combined risks of breast cancer, endometrial cancer, and ovarian cancer (Ganji et al., 2019).

Osteoporosis is a disabling disease with insidious onset. In most patients, no symptoms are detected during the early to middle stages of illness. However, sudden osteoporotic fracture can lead to lifelong disability. Early detection and treatment can significantly improve survival rates and quality of life in patients with osteoporosis. However, our understanding of the pathogenesis of osteoporosis is not sufficient. Although many factors, such as oxidative stress (Zhou et al., 2016; Geng et al., 2019) and altered estrogen signaling (Sapir-Koren and Livshits, 2017), have been shown to contribute to osteoporosis, specific biomarkers for the early diagnosis and treatment of this disease have not yet been identified.

Despite the success of proteomics analyses for screening of molecular targets in osteoporosis (Xu et al., 2018; Saad, 2020), transcriptomic studies are now attracting much attention. Previous studies have shown that long non-coding RNAs (lncRNAs) are involved in the regulation of a series of biological processes, such as the occurrence and development of osteoporosis (Zhao et al., 2017; Zhou et al., 2019; Zhang et al., 2020). lncRNAs can directly interfere with messenger RNA (mRNA) transcription or form an endogenous competitive network with microRNAs (miRNAs) to regulate transcription (Zhang et al., 2020). The regulation mechanisms of lncRNA have been less studied compared with the more mature studies on miRNAs (Hupkes et al., 2014; You et al., 2016; Shao, 2017; Wang et al., 2018; Cui et al., 2019). Therefore, further research on the lncRNA/mRNA regulatory network in osteoporosis is needed for better dissemination.

Like most chronic diseases, osteoporosis is determined by a combination of genetic and environmental factors (Ongphiphadhanakul, 2007). The heritability of bone density is thought to be 50–85% (Ralston, 2010). However, all single genetic pathogenic factors discovered to date can explain <6% of heritability, including loci discovered by genome-wide association studies (GWASs) (Liu et al., 2014). In addition, the two-dimensional role of genes is limited. Therefore, building networks may improve our ability to discover the remaining heritability factors in patients with osteoporosis.

Most studies of osteoporosis have focused on screening for differentially expressed genes (DEGs) to identify biomarkers (Liu et al., 2014; Xia et al., 2017; Zhou et al., 2018a, 2019; Zhang et al., 2020). However, few studies have explored the relevance of genes that share a high degree of functional interconnection and are regulated in a similar fashion. Weighted gene co-expression network analysis (WGCNA), a systems biology method, is particularly useful in this context and may help establish free-scale gene co-expression networks to identify the associations between different gene sets or between gene sets and clinical features (Qian et al., 2019). Notably, WGCNA has been broadly used to identify hub genes linked with clinical features in different diseases, such as breast cancer (Li et al., 2019), heart failure (Niu et al., 2019), and osteoporosis (Farber, 2010; Chen et al., 2016; Zhang et al., 2016; Qian et al., 2019).

In the current study, WGCNA and other approaches were used to analyze microarray data from blood monocytes collected from pre-and postmenopausal women with low or high bone mineral density (BMD) to characterize the key genes associated with osteoporosis. We then constructed a regulatory network containing key mRNAs and lncRNAs based on the co-expression relationships. Our findings improve our understanding of the biological relationships between osteoporosis and genetics and identify novel potential gene targets for the diagnosis and treatment of osteoporosis.

METHODS

Datasets and Samples

Data of this experiment are obtained and processed in the following ways (**Figure 1**). The microarray dataset GSE56814 was downloaded using the GEOquery package with R version (The R Foundation for Statistical Computing, Vienna, Austria) from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The gene expression microarray was based on the GPL5175 platform (Affymetrix Human Exon 1.0 ST Array). Subjects for the study were enrolled in a previous microarray-based transcriptome-wide profiling study of peripheral blood monocytes in 73 Caucasian females (47–56 years old) (Liu et al., 2015). Briefly, the patients included 42 women with high BMD (aged 52.9 $\pm$ 2.3 years, Z-score = 1.38 $\pm$ 0.49) and 31 women with low BMD (aged 51.4 $\pm$ 2.6 years, Z-score = -1.05 ± 0.51; **Table 1**). The raw files of gene profiles were downloaded and processed with the Robust Multi-array Average (RMA) algorithm. The nsFilter algorithm was used to filter the data for the subsequent WGCNA.

Annotation of lncRNAs From the Gene Expression Microarray Profile

LncRNAs were annotated from the gene expression microarray profile in two steps. First, we used the BLAST software to align the probes in GPL5175 to the mRNA database, which was selected from the overlap of coding RNAs in NCBI and Ensembl.

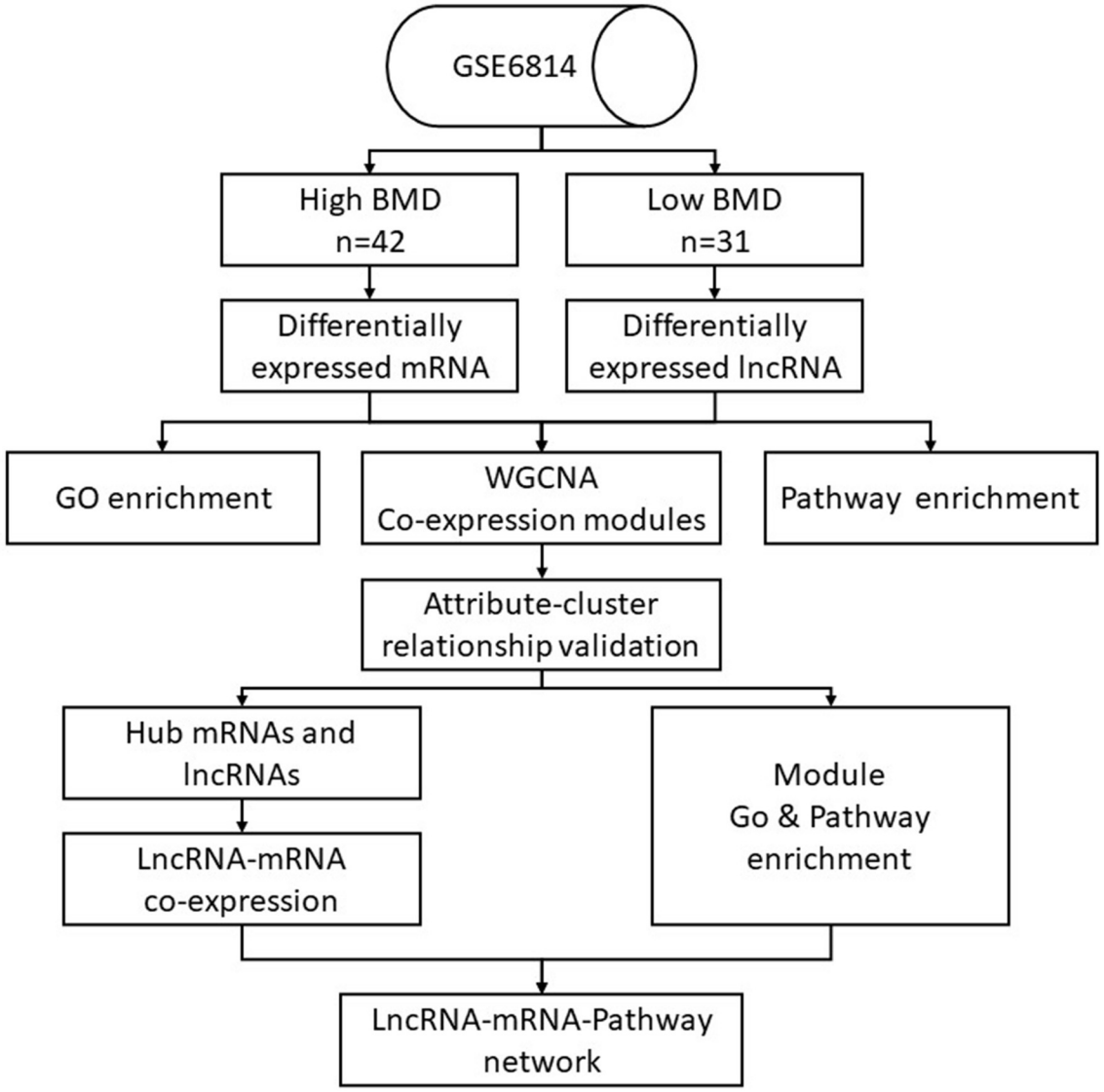

FIGURE 1 | Data management flowchart of the study.

TABLE 1 | Demographic characteristics of the patient samples.

	N	Age	BMD*
High BMD	42	52.9 ± 2.3	1.38 ± 0.49
Low BMD	31	51.4 ± 2.6	−1.05 ± 0.51
Total	73	52.3 ± 2.4	0.34 ± 0.50

**Hip BMD Z-score.*

Second, probes that could not be aligned to the mRNA database in the first step were further aligned to the lncRNA database, which included non-coding RNAs longer than 200 nucleotides collected from the NCBI, Ensembl, Refseq, and NONCODEv5 databases. Sequences were considered matching if they showed at least 90% identity. In both steps, the cutoff value was set to e-value <10e−5.

Identification and Visualization of Differentially Expressed mRNAs and lncRNAs

A random variance model *t*-test, which could effectively increase the degrees of freedom for small samples, was used to filter differentially expressed mRNAs and lncRNAs between patients with high and low BMD (Wright and Simon, 2003). After significance and false discovery rate (FDR) analyses, we selected DEGs according to the *p* value threshold and absolute value of fold change (FC). Results with a *p* value of <0.05 with |FC| >1.2 were considered significantly different (Yang et al., 2005).

For visualization, the differentially expressed mRNAs and lncRNAs were clustered using a hierarchical cluster algorithm with average linkage and Spearman's rank correlation distance, as provided by the EPCLUST software (http://ep.ebi.ac.uk/EP/EPCLUST/). Clustering was performed using the methods outlined in a previous publication (Misha et al., 2004). The results were visualized using heatmaps and dendrograms.

Functional Enrichment Analysis

Gene ontology (GO) analysis, which organizes genes into hierarchical categories and uncovers gene regulatory networks based on biological processes and molecular functions, was used to analyze the main functions of DEGs (Gene Ontology, 2006). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was then used to identify the significant pathways for these genes (Kanehisa et al., 2004). The Database for Annotation, Visualization and Integrated Discovery (DAVID; https://david.ncifcrf.gov/) provides a comprehensive set of functional annotation tools to analyze high-throughput gene functions. GO and KEGG pathway enrichment analyses were performed using DAVID. We were only interested in biological processes and KEGG pathways showing significance according to the following parameters: $p < 0.05$, FDR < 0.05, and enrichment score >1.5.

WGCNA

WGCNA is an analysis method for complex samples and is used to mine module information from chip data (Wan et al., 2018). In the current study, WGCNA was performed using a freely accessible R package. To minimize the loss of statistical information, the top 25% of mRNAs from the absolute median deviation and the top 10% lncRNAs were selected for WGCNA. The Pearson coefficient between any two genes was calculated. Subsequently, the correlation coefficients took multiple powers of N so that the connections between genes in the network align with the scale-free network distribution. A one-step function was performed to construct the network and detect consensus modules. Additionally, we constructed a hierarchical clustering tree using the correlation coefficient between genes. Gene modules are indicated as different branches on the clustering tree, and different colors were used to distinguish the modules.

Interaction Analysis of the Co-expression Modules

Interaction analysis of co-expression modules was performed as previously described Qian et al. (2019). Briefly, we calculated the eigengene adjacency based on similar co-expression in modules, and specific interactions among modules were evaluated using the flashClust function (Langfelder et al., 2012). A heatmap was established to elucidate the correlations among different modules.

Construction of the lncRNA–mRNA Weighted Network

Using the modules obtained with WGCNA, hub genes were extracted as the top 100 genes in the module. Hub genes with high connectivity are usually regulatory factors located upstream of regulatory networks, whereas genes with low connectivity are usually located downstream of regulatory networks (e.g., transporters and catalytic enzymes). Thus, the co-expression relationships among hub genes were calculated, and the co-expression of lncRNAs/mRNAs among the top 50 hub genes, as well as the co-expression of mRNAs/mRNAs among the top 150 hub genes, was selected to construct a co-expression network. Interactions between lncRNAs and mRNAs were identified by calculating the Pearson correlation coefficient of differentially expressed mRNAs and lncRNAs with a cutoff $|cor| > 0.5$. All interactions were identified using a p.adjust value <0.01. Next, lncRNA/mRNA regulatory networks were constructed using the Cytoscape software.

Construction of the lncRNA/mRNA Pathway Weighted Co-expression Network

The lncRNA/mRNA pathway network was constructed based on the regulatory relationship of lncRNAs/mRNAs and the significant pathways involved in the regulation of mRNAs. The primary objective of this analysis was to identify the signaling pathways regulated by lncRNAs to predict possible mechanisms of lncRNAs in disease.

Statistical Analysis

Data were analyzed using the SPSS 23.0 software (SPSS, Chicago, IL, USA). The random variance model *t*-test was performed using BRB-ArrayTools (v4.6, http://linus.nci.nih.gov/BRB-ArrayTools.html) (Wright and Simon, 2003). Because the sample size was limited, the adjusted p values were too large after multiple testing controls. We used a raw p <0.05 as the threshold for nominally significant differential expression. Notably, multiple testing adjustment with an FDR <0.05 was used to filter enriched GO and KEGG pathways.

RESULTS

Differentially Expressed mRNAs and lncRNAs

With an FC cutoff value >1.2 and $p < 0.05$, 3,355 mRNAs (**Figures 2A,C**) and 999 lncRNAs (**Figures 2B,D**) were identified as differentially expressed between patients with high and low hip BMD; these were selected as candidate genes for subsequent WGCNA. The pathway analysis reveals that the up-/down-regulated DEGs were primarily enriched in metabolic pathways (**Figures 3A,B**). The GO analysis found that up-regulated DEGs were enriched in terms of apoptotic process, G-protein coupled receptor signaling pathway, negative regulation of transcription from RNA polymerase II promoter, etc. Furthermore, the down-regulated ones were enriched in transcription, DNA-templated, G-protein coupled receptor signaling pathway, DNA-templated regulation of transcription, etc. (**Figures 3C,D**).

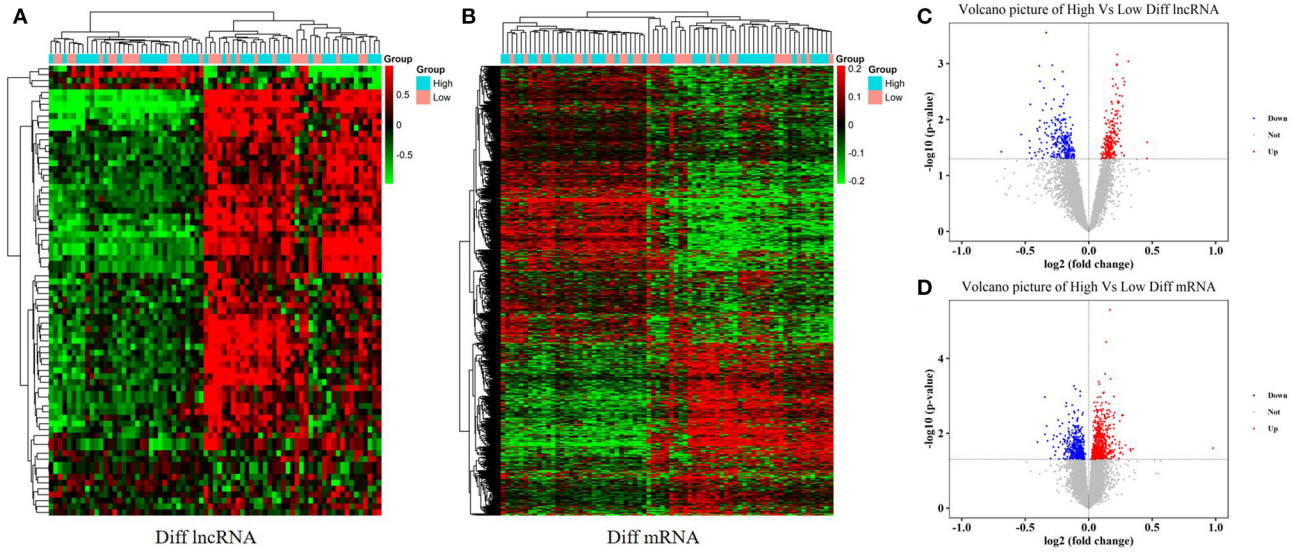

FIGURE 2 | Differentially expressed mRNAs and lncRNAs between high and low hip BMD subjects. **(A,B)** The heatmaps represent hierarchical clustering for differentially expressed lncRNAs and mRNAs. **(C,D)** Volcano plots of significantly differently expressed genes (DEGs).

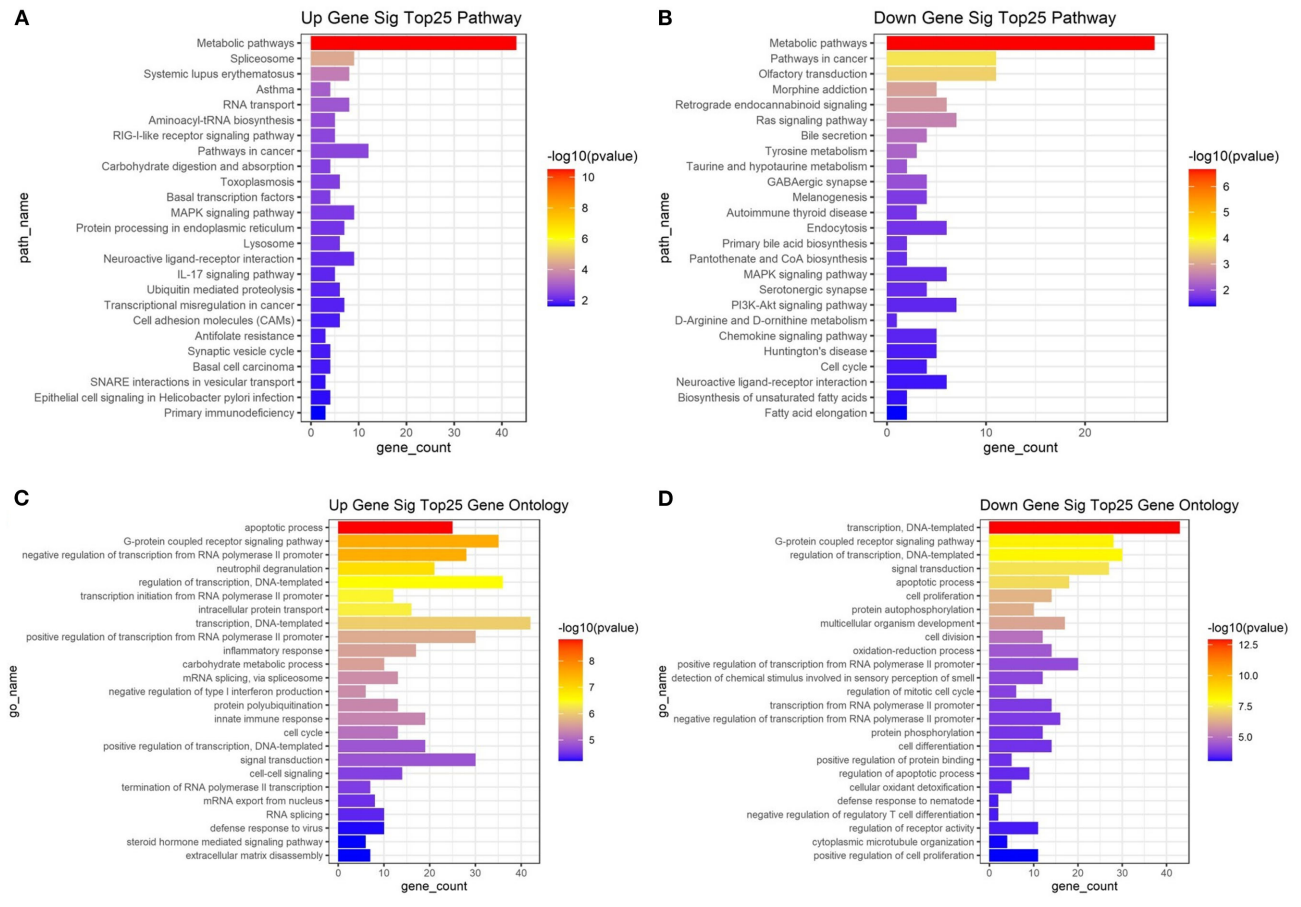

FIGURE 3 | Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis of the top 25 up **(A)**/down **(B)**-regulated pathways enriched in differentially expressed genes between high/low BMD subjects. Top 25 up **(C)**/down **(D)**-regulated biological processes enriched in differentially expressed genes between high/low BMD subjects.

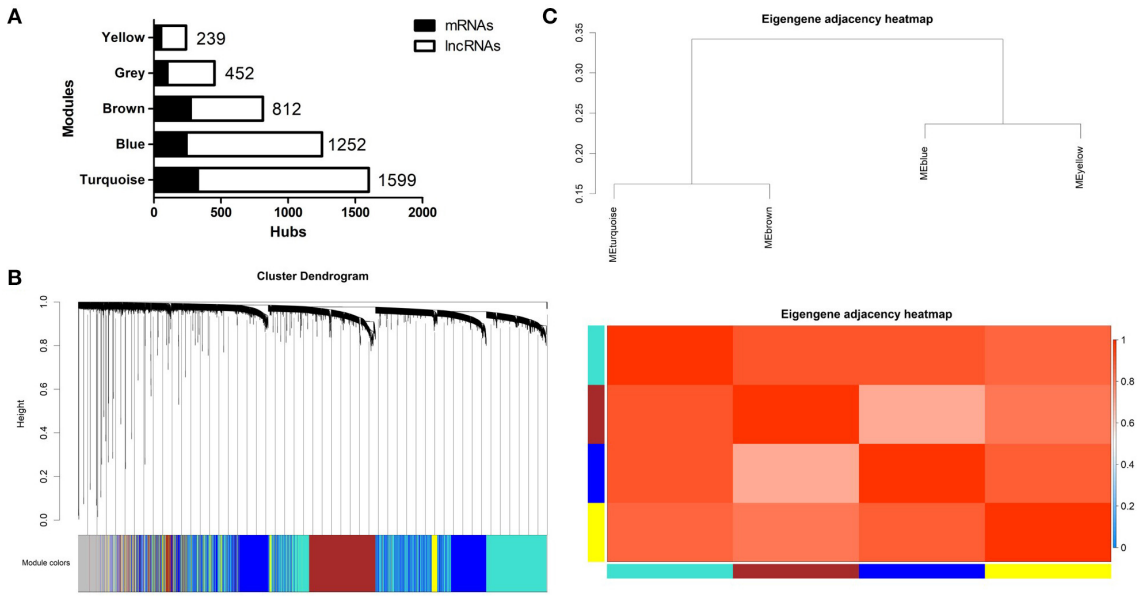

FIGURE 4 | (A) Size and composition of the modules in the WGCNA network. **(B)** Dendrogram obtained by clustering the dissimilarity based on consensus topological overlap with the corresponding module colors indicated by the color row. **(C)** Eigengene adjacencies of heatmap. Red shows high adjacency.

Establishing Weighted lncRNA/mRNA Co-expression Networks and Identification of Soft Threshold Power

A lncRNA/mRNA co-expression network was established from the newly generated set of mRNAs and lncRNAs. First, we performed cluster analysis on the selected mRNAs and lncRNAs. The results showed that no outlier existed in the sample; thus, there was no need to remove any outliers. Second, we used the R package to check the integrity of the data and constructed a network topology to determine the soft thresholding power. A soft threshold power of 6.5 was used to define the adjacency matrix, which was processed using the criteria of approximate scale-free topology. Third, the adjacent and topological matrices were obtained through the soft thresholding power. According to the topological matrix, genes were clustered through dissimilarity. Next, a dynamic shearing method was used to separate the cluster dendrogram into four modules, each indicated by a different color (turquoise, blue, brown, or yellow; gray was used for genes that did not fit into a distinct group). The largest module was the turquoise module, followed by the blue module. The size and composition of the modules are shown in **Figure 4A**. Of all selected genes, 351 mRNAs and 101 lncRNAs failed to fit within a distinct group and were assigned to the gray module (**Figure 4B**). After generating an eigengene adjacency heatmap (**Figure 4C**) to explore the correlations between modules, we found that the regulation directions of these modules were consistent. The modules showed a significantly positive correlation in patients with high BMD and a negative correlation in premenopausal women with low BMD. However, the correlation was not significant in postmenopausal women except that the gray module in patients with high BMD showed a correlation coefficient of 0.25 ($p = 0.03$).

Functional Analyses and Pathway Enrichment of Different Modules

To determine whether the modules were composed of functionally similar genes and to understand the functional significance of the network modules, GO term and KEGG pathway enrichment analyses were performed. The enrichment results from the yellow module were not significant because there were few genes in this module. The GO results of all three modules were enriched in the positive regulation of transcription from RNA polymerase II promoter, DNA-templated transcription, and their regulatory mechanisms. Specifically, genes in the blue module were highly enriched in cell surface receptor signaling pathway, chemical synaptic transmission, ion transmembrane transport, multicellular organism development, neutrophil degranulation, and regulation of receptor activity. The turquoise module was associated with calcium ion transmembrane transport, cell adhesion, cell proliferation, cellular protein metabolic process, membrane depolarization, and microtubule-based movement. The brown module was associated with bicarbonate transport, cell cycle arrest, cell differentiation, cell division, cell migration, oxidation–reduction process, and rRNA processing. The top 20 GO terms for the three modules are shown in **Figure 5**. mRNA pathway enrichment was also analyzed. Notably, all three modules were significantly enriched in metabolic pathways and neuroactive ligand–receptor interactions. The turquoise module was specifically associated with purine metabolism, necroptosis, inflammatory mediator regulation of TRP channels, alcoholism,

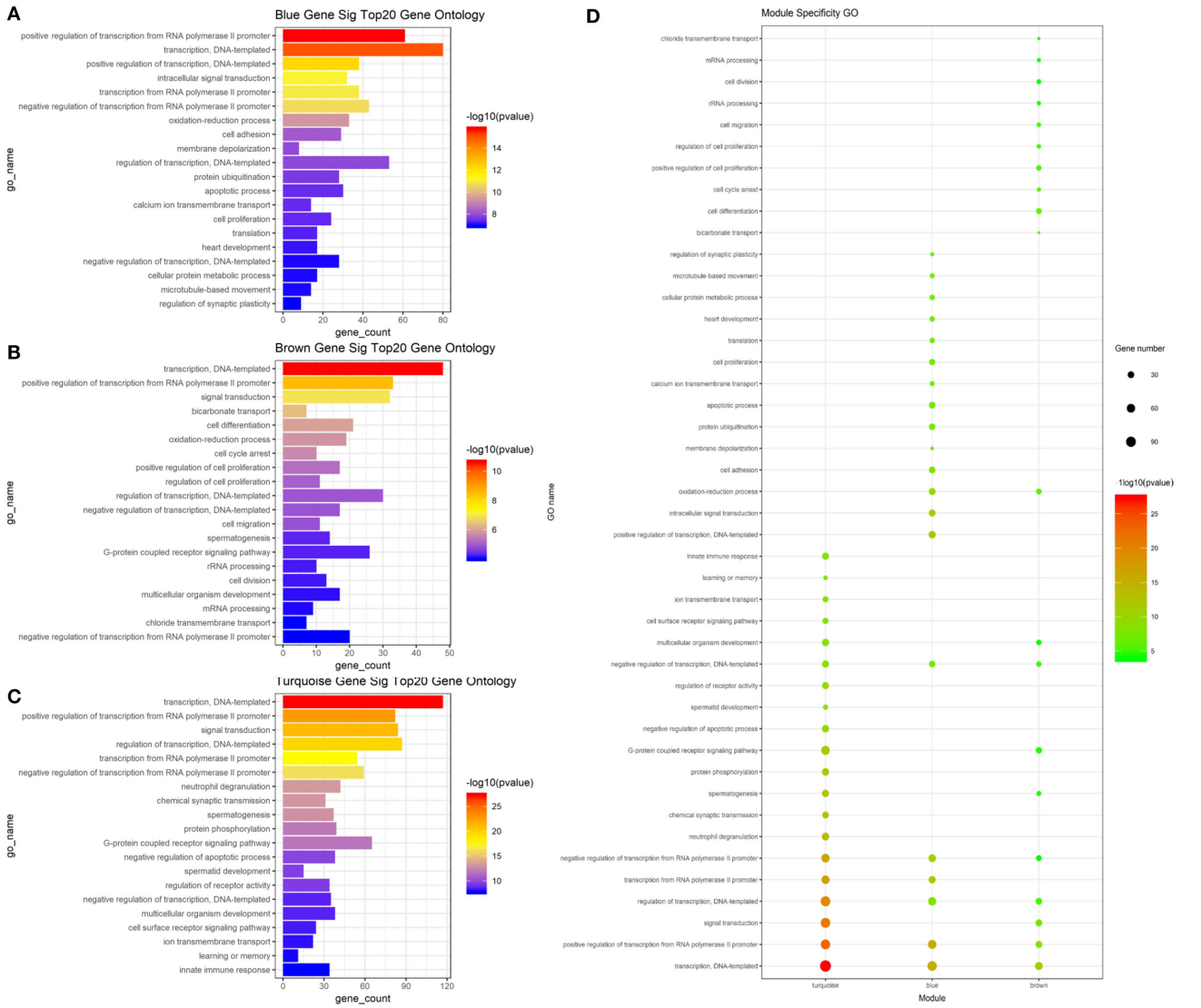

FIGURE 5 | Result of GO analysis about mRNA based on WGCNA. **(A)** Top 20 enriched gene ontologies in blue module **(B)**. Top 20 enriched gene ontologies in brown module. **(C)** Top 20 enriched gene ontologies in turquoise module. **(D)** Module specificity gene ontologies. The vertical axis shows the results of GO analysis, and the horizontal axis shows the different modules. The color of the round shape shows transitional values (log10 of q value), and the size shows the number of genes that were enriched. mRNA, messenger RNA; WGCNA, weighted gene co-expression network analysis.

and the hypoxia-inducible factor-1 signaling pathway. The blue module was associated with the Rap1 signaling pathway, calcium signaling pathway, NOD-like receptor signaling pathway, phosphatidylinositol 3-kinase/Akt signaling pathway, and sphingolipid signaling pathway. The brown module was associated with protein processing in the endoplasmic reticulum, tyrosine metabolism, glycerophospholipid metabolism, cell cycle, and metabolism of xenobiotics by cytochrome P450. The top 20 pathways for each module are shown in **Figure 6**.

WGCNA Hub Gene Identification

Hub genes are usually key regulators, such as transcription factors, and are worthy of in-depth analysis and mining. In the blue module, we found 26 lncRNAs and 55 mRNAs as hub genes (**Figure 7A**). We analyzed the functions of these hub genes and found that these genes were mainly involved in response to muscle stretch (e.g., *JUN* and *MAPK14*), biotic stimulus (e.g., *IFITM3*), and ventricular system development (e.g., *HYDIN* and *ARMC4*). The cell components were enriched in the cytoplasm (e.g., *BCAS3*, *CD248*, *DNAJC17*, *GCN1*, and *GLE1*), endoplasmic reticulum (e.g., *ALG12*, *NECAB3*, *UVRAG*, *CERS2*, and *KCNMA1*), and endoplasmic reticulum membrane (e.g., *ALG12*, *NECAB3*, *CERS2*, and *PCYT1A*). The molecular functions were mainly enriched in ubiquitin protein ligase binding (e.g., *FAF2*, *ABTB1*, and *UBE2N*). We also observed 36 lncRNAs and 31 mRNAs as hub nodes in the turquoise module

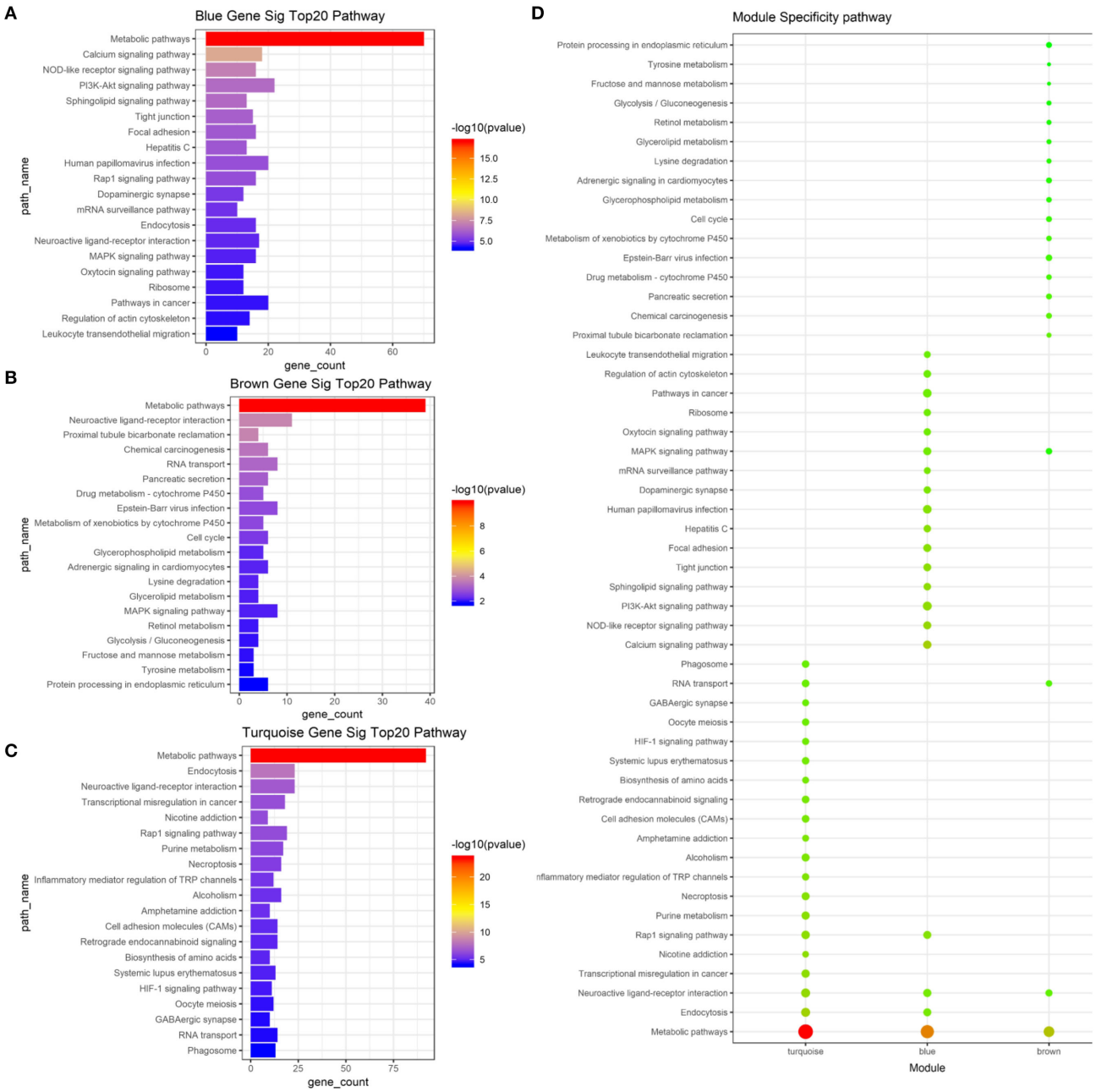

FIGURE 6 | Result of pathway analysis about mRNA based on WGCNA. **(A)** Top 20 enriched pathways in blue module **(B)**. Top 20 enriched pathways in brown module. **(C)** Top 20 enriched pathways in turquoise module. **(D)** Module specificity pathways. The vertical axis shows the results of pathway analysis, and the horizontal axis shows the different modules. The color of the round shape shows transitional values (log10 of q value), and the size shows the number of genes that were enriched. mRNA, messenger RNA; WGCNA, weighted gene co-expression network analysis.

(**Figure 7B**) and 56 mRNAs and 30 lncRNAs as hub nodes in the brown module (**Figure 7C**).

Construction of lncRNA/mRNA Pathway Co-expression Networks

To uncover the possible mechanisms of lncRNA-mediated regulation of signaling pathways, we selected a number of pathways with significant differences in the turquoise, blue, and brown modules and associated them with the lncRNA/mRNA co-expression network. In the pathway co-expression network, the blue module had 3 mRNAs and 24 lncRNAs (**Figure 7D**), the brown module had 4 mRNAs and 11 lncRNAs (**Figure 7E**), and the turquoise module had 9 mRNAs and 5 lncRNAs (**Figure 7F**). In the blue module, XR_001739541.1 was linked to *MRPS10*,

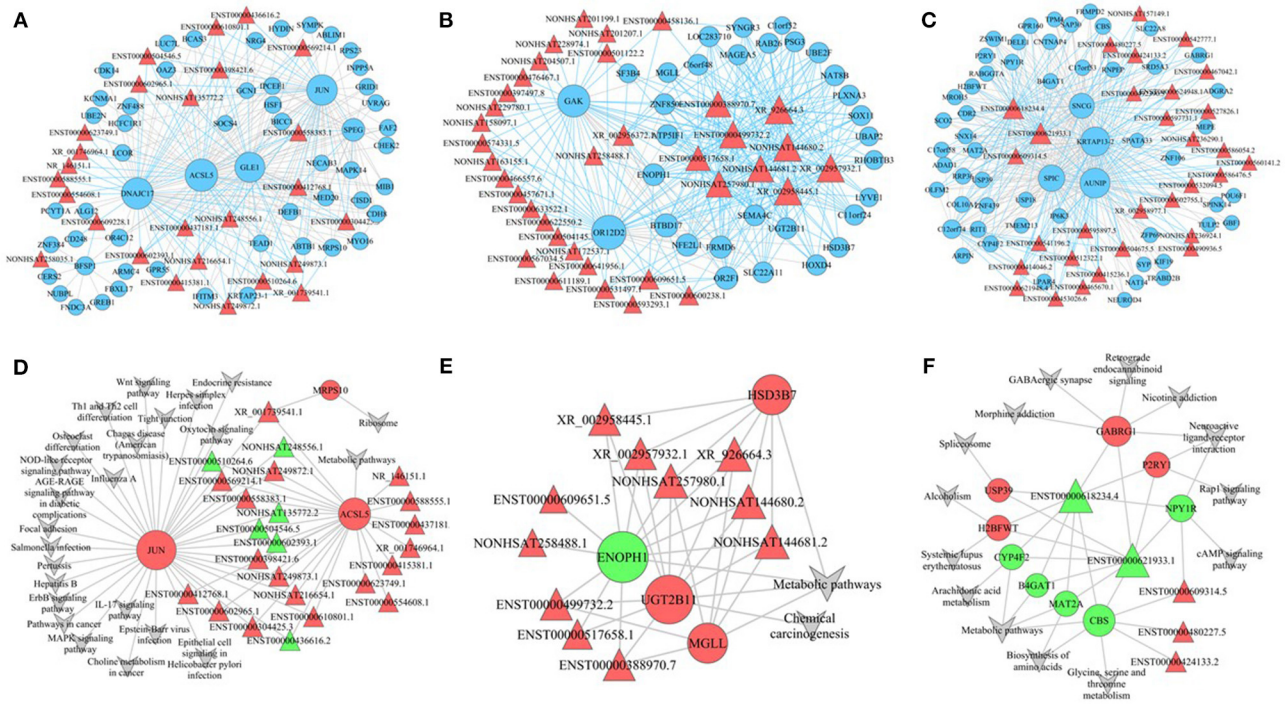

FIGURE 7 | Established lncRNA–mRNA network and lncRNA–mRNA pathway. **(A–C)** lncRNA–mRNA network of genes in blue **(A)**, brown **(B)**, and turquoise **(C)** modules. **(D–F)** lncRNA–mRNA pathway of genes in blue **(D)**, brown **(E)**, and turquoise **(F)** modules.

ACSL5, and *JUN* and was therefore enriched in the ribosome pathway and metabolic pathway. Sixteen lncRNAs, including NONHSAT249872.1 and ENST00000510264.6, were linked to two mRNAs (*JUN* and *ACSL5*) and were enriched in pathways, such as the NOD-like receptor signaling pathway, mitogen-activated protein kinase signaling pathway, Wnt signaling pathway, ErbB signaling pathway, osteoclast differentiation, and metabolic pathways. Seven other lncRNAs were linked to *ACSL5* and were enriched in metabolic pathways. In the brown module, XR_002958445.1, XR_002957932.1, NONHSAT257980.1, XR_926664.3, NONHSAT144580.2, and NONHSAT144681.2 were connected to *HSD3B7*, *ENH1*, *UGT2B11*, and *MGLL* mRNAs and were therefore enriched in metabolic pathways and chemical carcinogenesis. In the turquoise module, ENST00000618234.4 and ENST00000621933.1 were linked to nine mRNAs (*USP39*, *H2BFWT*, *CYP4F2*, *B4GAT1*, *MAT2A*, *CBS*, *GABRG1*, *P2RY1*, and *NPY1R*) and enriched in pathways, such as GABAergic synapse, retrograde endocannabinoid signaling, Rap1 signaling pathways, and cAMP signaling pathway. The lncRNAs ENST00000609314.5, ENST00000480227.5, and ENST00000424133.2 were also involved in the turquoise modular lncRNA/mRNA pathway co-expression network.

DISCUSSION

Osteoporosis is a common and complex systemic bone disease, and women are especially susceptible to this disease. The onset of osteoporosis is insidious, and the disease often remains undetected in the early stages. However, once a secondary osteoporotic fracture occurs, many complications can occur, and the prognosis is poor. Therefore, many researchers have investigated the molecular diagnosis, treatment targets, and genetic regulation of osteoporosis. In a previous study, Liu showed that *DAXX* and *PLK3*, which are related to induction of apoptosis, were down-regulated in patients with a low BMD among a cohort of 73 Caucasian females (Liu et al., 2015). Based on the same microarray dataset available online, Zhou performed GWAS and found 29 potential transcription factors for up-regulated genes and 9 transcription factors for down-regulated genes (Zhou et al., 2018a). They further investigated the relationships between mRNAs and lncRNAs using two approaches and claimed that 26 candidate lncRNAs may regulate mRNA expression (Zhou et al., 2019). After correcting for crosstalk effects, they identified several significant enriched pathways involved in BMD regulation (Zhou et al., 2018b). Moreover, Xia established a meta-analysis using the microarray datasets GSE56815 and GSE56814 and found 10 potential pathogenic genes of osteoporosis (Xia et al., 2017).

In this study, we found 4,354 DEGs in the peripheral blood chips of patients with high or low BMD in the hip; these included 3,355 mRNAs and 999 differentially expressed lncRNAs. In contrast to previous studies based on protein–protein interaction (PPI) networks, we employed WGCNA to aggregate genes with common expression characteristics into modules. This systemic

biology method helped free-scale gene co-expression networks to identify associations without previous PPI knowledge (Zheng et al., 2020). The WGCNA co-expression networks revealed three gene modules consisting of 40 lncRNAs and 16 mRNAs, which were significantly related to the level of BMD. In a previous study, Qian (Qian et al., 2019) found 12 genes as hub genes in 80 Caucasian females. Another WGCNA study identified six genes from 26 healthy young Chinese females (Farber, 2010). Zhang et al. found seven genes that were significantly down- or up-regulated using traditional comparative analysis, WGCNA, and gene set enrichment analysis (Zhang et al., 2016). Chen constructed a WGCNA co-expression network composed of BMD GWAS genes and found two functional gene modules and nine interesting genes. Of note, the genes identified in the current study did not overlap in these previous studies. We attribute the observed discrepancy to differences in patients and ethnicity; these potential differences should be investigated further. The differentially expressed mRNAs and lncRNAs were primarily involved in metabolic pathways, including glycerophospholipid metabolism, lysine degradation, and glycerolipid metabolism.

Our study and previous studies have established possible targets for the treatment of osteoporosis, such as JUN (Ralston, 2010; Zhou et al., 2019). JUN belongs to the AP-1 family of transcription factors, which includes c-Fos, Fra1, Fra2, JunB, and JunD. JUN expression was significantly up-regulated in dental pulp stem cells induced to undergo osteogenic differentiation (Guo et al., 2018). Higher concentrations of glucocorticoids impair osteogenesis by inhibiting JUN expression and human bone marrow mesenchymal stem cell (BMSC) proliferation, which can be driven by glucocorticoid receptor and AP-1 crosstalk (Carcamo-Orive et al., 2010). Moreover, our recent study showed that JUN can drive bone formation by expanding osteoprogenitor populations and forcing them into the bone fate, providing a rationale for future clinical applications (Lerbs et al., 2020).

Long-chain fatty acyl-CoA synthetases 5 (ACSL5) is an isozyme of the long-chain fatty-acid-coenzyme A ligase family. It is a regulatory enzyme that converts free long-chain fatty acids into fatty acyl-CoA esters and thereby plays key roles in lipid biosynthesis and fatty acid degradation. Currently, there is no evidence that ACSL5 expression is involved in osteoporosis; however, the presence of ACSL5 is obviously related to disorders of glucose metabolism. High glucocorticoid concentrations impair osteogenesis (Carcamo-Orive et al., 2010) and induce the activation of osteoclast proliferation and differentiation (Wongdee and Charoenphandhu, 2011). In addition, ACSL5 may also be an important mediator in apoptosis (Xia et al., 2016). Further studies are needed to assess the potential roles of this protein in the pathophysiological process of osteoporosis.

The lncRNA/mRNA regulatory networks were further constructed using high connectivity hub genes in the WGCNA co-expression network. Compared with nodes with low connectivity, nodes with high connectivity play more important roles in the entire transcription network and are more likely to be upstream regulators. According to the above-mentioned regulatory relationships of lncRNAs/mRNAs and the significantly involved pathways, we further constructed a network of pathways in which lncRNAs could regulate mRNAs through co-expression and thereby play roles in these pathways. Notably, metabolic pathways were significantly enriched in all three functional gene modules. Bone formation is known to be dependent on the supply of metabolites to monocytes in the bone marrow (Bidwell et al., 2013). Additionally, the balance of bone metabolism depends on the coordination of bone formation and resorption, and this process requires information exchange between different types of cells. For example, the lncRNA *Bmncr* is a key regulator of age-related osteogenic niche alteration and cell fate switch of BMSCs (Li et al., 2018). Moreover, the lncRNA *ODSM* functions as a competing endogenous RNA in the lncRNA *ODSM*/*miR-139-3p*/*ELK1* pathway and has important functions in osteoblast differentiation and apoptosis (Wang et al., 2018). Further studies are needed to explore the molecular mechanisms through which lncRNAs act as transcription factors to regulate osteoporosis (Zhang et al., 2020). It is worth noting that the above-mentioned molecular targets may be indirectly related to the BMD phenotype. In the process of establishing the aforementioned weighted lncRNA/mRNA co-expression networks, there was no observed direct quantitative relationship with the level of BMD. These genes aggregate to form modules through co-expression relationships. They have significant correlations and may participate in certain biological processes together. Not all genes in these modules are directly related to the level of BMD, which makes it difficult for us to interpret the experimental results within the context of BMD levels. Therefore, it is necessary to construct a network relationship, find the hub genes, and conduct further *in vitro* validations.

There were some limitations to this study. First, this study was based purely on microarray datasets, and we did not obtain any data directly from *in vivo* experiments. Thus, further studies are needed to confirm the observed molecular mechanisms. Second, when selecting the phenotype of osteoporosis, we used BMD as the only indicator. Because phenotype identification can directly influence patient grouping and is crucial to the construction of gene networks, additional indicators (e.g., bone geometric parameters, bone size, and compressive strength index of the femoral neck) should be evaluated in further studies in order to obtain a complete picture of osteoporosis. Third, this study did not compare the obtained results in female osteoporosis with male cases, because there are few samples of male osteoporosis in the public database, and the platforms are not the same. It is worth noting that the above-mentioned biomarkers were all found in female database samples; therefore, we may not be able to extrapolate these conclusions to samples of male patients. Previous studies have shown that miRNAs are gender-dependent as molecular targets of BMD (Kelch et al., 2017). Finally, this study is based on gene expression from blood monocytes. This is far removed from therapeutic application in musculoskeletal diseases. Further validation on bone samples should be done in future research.

In conclusion, in this study, we identified differentially expressed mRNAs and lncRNAs in existing microarray profile

data. A WGCNA was constructed and yielded three significant modules associated with differences in BMD. Enrichment analysis indicated that the modules were primarily enriched in metabolic pathways, such as glycerophospholipid metabolism, lysine degradation, and glycerolipid metabolism. Several hub genes, including *JUN* and *ACSL5*, were found and may represent potential biomarkers or clinical targets for osteoporosis. In addition, a comprehensive lncRNA/mRNA-pathway regulatory network was built to elucidate the complex interactions between the transcripts and non-coding RNAs. Our findings provided a novel perspective on the regulatory mechanisms of osteoporosis.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <GEO data base (http://www.ncbi.nlm.nih.gov/geo/) GSE56814>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of the First Affiliated Hospital of USTC. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XZ conceived the idea and designed the project. XC performed the data analysis. XZ and KC wrote the paper. NK, GL, and BW revised the manuscript. GL and CZ gained institutional funding. CZ provided administrative support. All authors have read and approved the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 81871788 and 81902201), the project for Science and Technology leader of Anhui Province (Grant No. 2018H177), the Scientific Research Fund of Anhui Education (Grant No. 2017jyxm1097), the Anhui Provincial Postdoctoral Science Foundation (Grant No. 2019B302), the Key Research and Development Plan of Anhui Province (Grant No. 912278014064), and the Fundamental Research Funds for the Central Universities (Grant No. WK9110000093).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.600097/full#supplementary-material

REFERENCES

Baccaro, L. F., Conde, D. M., Costa-Paiva, L., and Pinto-Neto, A. M. (2015). The epidemiology and management of postmenopausal osteoporosis: a viewpoint from Brazil. *Clin. Interv. Aging*. 10, 583–591. doi: 10.2147/CIA.S54614

Bidwell, J. P., Alvarez, M. B., Hood, M., and Childress, P. (2013). Functional impairment of bone formation in the pathogenesis of osteoporosis: the bone marrow regenerative competence. *Curr. Osteoporos. Rep.* 11, 117–125. doi: 10.1007/s11914-013-0139-2

Carcamo-Orive, I., Gaztelumendi, A., Delgado, J., Tejados, N., Dorronsoro, A., Fernandez-Rueda, J., et al. (2010). Regulation of human bone marrow stromal cell proliferation and differentiation capacity by glucocorticoid receptor and AP-1 crosstalk. *J. Bone Miner Res.* 25, 2115–2125. doi: 10.1002/jbmr.120

Chen, P., Li, Z., and Hu, Y. (2016). Prevalence of osteoporosis in China: a meta-analysis and systematic review. *BMC Public Health*. 16:1039. doi: 10.1186/s12889-016-3712-7

Cui, Q., Xing, J., Yu, M., Wang, Y., Xu, J., Gu, Y., et al. (2019). Mmu-miR-185 depletion promotes osteogenic differentiation and suppresses bone loss in osteoporosis through the Bgn-mediated BMP/Smad pathway. *Cell Death Dis.* 10:172. doi: 10.1038/s41419-019-1428-1

Ensrud, K. E., and Crandall, C. J. (2017). Osteoporosis. *Ann. Intern. Med.* 167, ITC17–ITC32. doi: 10.7326/AITC201708010

Farber, C. R. (2010). Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data. *J. Bone Miner Res.* 25, 2359–2367. doi: 10.1002/jbmr.138

Ganji, R., Moghbeli, M., Sadeghi, R., Bayat, G., and Ganji, A. (2019). Prevalence of osteoporosis and osteopenia in men and premenopausal women with celiac disease: a systematic review. *Nutr J*. 18:9. doi: 10.1186/s12937-019-0434-6

Gene Ontology, C. (2006). The gene ontology (GO) project in 2006. *Nucleic Acids Res.* 34(Database issue):D322–D326. doi: 10.1093/nar/gkj021

Geng, Q., Gao, H., Yang, R., Guo, K., and Miao, D. (2019). Pyrroloquinoline quinone prevents estrogen deficiency-induced osteoporosis by inhibiting oxidative stress and osteocyte senescence. *Int. J. Biol. Sci.* 15, 58–68. doi: 10.7150/ijbs.25783

Guo, T., Cao, G., Li, Y., Zhang, Z., Nor, J. E., Clarkson, B. H., et al. (2018). Signals in stem cell differentiation on fluorapatite-modified scaffolds. *J. Dent. Res.* 97, 1331–1338. doi: 10.1177/0022034518788037

Hupkes, M., Sotoca, A. M., Hendriks, J. M., Van Zoelen, E. J., and Dechering, K. J. (2014). MicroRNA miR-378 promotes BMP2-induced osteogenic differentiation of mesenchymal progenitor cells. *BMC Mol. Biol.* 15:1. doi: 10.1186/1471-2199-15-1

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32(Database issue): D277–D280. doi: 10.1093/nar/gkh063

Kelch, S., Balmayor, E. R., Seeliger, C., Vester, H., Kirschke, J. S., and Van Griensven, M. (2017). miRNAs in bone tissue correlate to bone mineral density and circulating miRNAs are gender independent in osteoporotic patients. *Sci Rep*. 7:15861. doi: 10.1038/s41598-017-16113-x

Langfelder, P., Horvath, S., and Fast, R. (2012). Functions for robust correlations and hierarchical clustering. *J. Stat. Softw*. 46: e00027–16. doi: 10.18637/jss.v046.i11

Lerbs, T., Cui, L., Muscat, C., Saleem, A., Van Neste, C., Domizi, P., et al. (2020). Expansion of bone precursors through jun as a novel treatment for osteoporosis-associated fractures. *Stem Cell Rep.* 14, 603–613. doi: 10.1016/j.stemcr.2020.02.009

Li, C. J., Xiao, Y., Yang, M., Su, T., Sun, X., Guo, Q., et al. (2018). Long noncoding RNA Bmncr regulates mesenchymal stem cell fate during skeletal aging. *J. Clin. Invest*. 128, 5251–5266. doi: 10.1172/JCI99044

Li, J., Liu, C., Chen, Y., Gao, C., Wang, M., Ma, X., et al. (2019). Tumor characterization in breast cancer identifies immune-relevant gene signatures associated with prognosis. *Front. Genet*. 10:1119. doi: 10.3389/fgene.2019.01119

Liu, Y. J., Zhang, L., Papasian, C. J., and Deng, H. W. (2014). Genome-wide association studies for osteoporosis: a 2013 update. *J. Bone Metab*. 21, 99–116. doi: 10.11005/jbm.2014.21.2.99

Liu, Y. Z., Zhou, Y., Zhang, L., Li, J., Tian, Q., Zhang, J. G., et al. (2015). Attenuated monocyte apoptosis, a new mechanism for osteoporosis suggested by a transcriptome-wide expression study of monocytes. *PLoS ONE* 10:e0116792. doi: 10.1371/journal.pone.0116792

Misha, K., Patrick, K., Culhane, A. C., Steffen, D., Jan, I., Christine, K. R., et al. (2004). Expression profiler: next generation–an online platform for analysis of microarray data. *Nucleic Acids Research* 32(Web Server issue):465–470. doi: 10.1093/nar/gkh470

Niu, X., Zhang, J., Zhang, L., Hou, Y., Pu, S., Chu, A., et al. (2019). Weighted gene co-expression network analysis identifies critical genes in the development of heart failure after acute myocardial infarction. *Front. Genet*. 10:1214. doi: 10.3389/fgene.2019.01214

Ongphiphadhanakul, B. (2007). Osteoporosis: the role of genetics and the environment. *Forum Nutr*. 60, 158–167. doi: 10.1159/000107166

Qian, G. F., Yuan, L. S., Chen, M., Ye, D., Chen, G. P., Zhang, Z., et al. (2019). PPWD1 is associated with the occurrence of postmenopausal osteoporosis as determined by weighted gene coexpression network analysis. *Mol Med Rep*. 20, 3202–3214. doi: 10.3892/mmr.2019.10570

Ralston, S. H. (2010). Genetics of osteoporosis. *Ann. N. Y. Acad. Sci*. 1192, 181–199. doi: 10.1111/j.1749-6632.2009.05317.x

Ruza, I., Mirfakhraee, S., Orwoll, E., and Gruntmanis, U. (2013). Clinical experience with intravenous zoledronic acid in the treatment of male osteoporosis: evidence and opinions. *Ther. Adv. Musculoskelet Dis*. 5, 182–198. doi: 10.1177/1759720X13485829

Saad, F. A. (2020). Novel insights into the complex architecture of osteoporosis molecular genetics. *Ann. N. Y. Acad. Sci*. 1462, 37–52. doi: 10.1111/nyas.14231

Sahni, S., Hannan, M. T., Blumberg, J., Cupples, L. A., Kiel, D. P., and Tucker, K. L. (2009). Protective effect of total carotenoid and lycopene intake on the risk of hip fracture: a 17-year follow-up from the framingham osteoporosis study. *J. Bone Miner Res*. 24, 1086–1094. doi: 10.1359/jbmr.090102

Sapir-Koren, R., and Livshits, G. (2017). Postmenopausal osteoporosis in rheumatoid arthritis: the estrogen deficiency-immune mechanisms link. *Bone* 103, 102–115. doi: 10.1016/j.bone.2017.06.020

Shao, M. (2017). Construction of an miRNA-regulated pathway network reveals candidate biomarkers for postmenopausal osteoporosis. *Comput. Math. Methods Med*. 2017:9426280. doi: 10.1155/2017/9426280

Wan, Q., Tang, J., Han, Y., and Wang, D. (2018). Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma. *Exp. Eye Res*. 166, 13–20. doi: 10.1016/j.exer.2017.10.007

Wang, Y., Wang, K., Hu, Z., Zhou, H., Zhang, L., Wang, H., et al. (2018). MicroRNA-139-3p regulates osteoblast differentiation and apoptosis by targeting ELK1 and interacting with long noncoding RNA ODSM. *Cell Death Dis*. 9:1107. doi: 10.1038/s41419-018-1153-1

Wongdee, K., and Charoenphandhu, N. (2011). Osteoporosis in diabetes mellitus: possible cellular and molecular mechanisms. *World J. Diabetes*. 2, 41–48. doi: 10.4239/wjd.v2.i3.41

Wright, G. W., and Simon, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19, 2448–2455. doi: 10.1093/bioinformatics/btg345

Xia, B., Li, Y., Zhou, J., Tian, B., and Feng, L. (2017). Identification of potential pathogenic genes associated with osteoporosis. *Bone Joint Res*. 6, 640–648. doi: 10.1302/2046-3758.612.BJR-2017-0102.R1

Xia, Q., Chesi, A., Manduchi, E., Johnston, B. T., Lu, S., Leonard, M. E., et al. (2016). The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. *Diabetologia*. 59, 2360–2368. doi: 10.1007/s00125-016-4077-2

Xu, R., Yallowitz, A., Qin, A., Wu, Z., Shin, D. Y., et al. (2018). Targeting skeletal endothelium to ameliorate bone loss. *Nat. Med*. 24, 823–833. doi: 10.1038/s41591-018-0020-z

Yang, H., Crawford, N., Lukes, L., Finney, R., Lancaster, M., and Hunter, K. W. (2005). Metastasis predictive signature profiles pre-exist in normal tissues. *Clin. Exp. Metastasis* 22, 593–603. doi: 10.1007/s10585-005-6244-6

You, L., Pan, L., Chen, L., Gu, W., and Chen, J. (2016). MiR-27a is essential for the shift from osteogenic differentiation to adipogenic differentiation of mesenchymal stem cells in postmenopausal osteoporosis. *Cell Physiol. Biochem.* 39, 253–265. doi: 10.1159/000445621

Zhang, L., Liu, Y. Z., Zeng, Y., Zhu, W., Zhao, Y. C., Zhang, J. G., et al. (2016). Network-based proteomic analysis for postmenopausal osteoporosis in Caucasian females. *Proteomics* 16, 12–28. doi: 10.1002/pmic.201500005

Zhang, X., Liang, H., Kourkoumelis, N., Wu, Z., Li, G., and Shang, X. (2020). Comprehensive analysis of lncRNA and miRNA expression profiles and ceRNA network construction in osteoporosis. *Calcif Tissue Int*. 106, 343–354. doi: 10.1007/s00223-019-00643-9

Zhao, X., Liu, Y., and Yu, S. (2017). Long noncoding RNA AWPPH promotes hepatocellular carcinoma progression through YBX1 and serves as a prognostic biomarker. *Biochim. Biophys. Acta Mol. Basis Dis*. 1863, 1805–1816. doi: 10.1016/j.bbadis.2017.04.014

Zheng, J. N., Li, Y., Yan, Y. M., Shi, H., Zou, T. T., Shao, W. Q., Wang, Q. (2020). Identification and validation of key genes associated with systemic sclerosis-related pulmonary hypertension. *Front. Genet.* 11:816. doi: 10.3389/fgene.2020.00816

Zhou, Q., Zhu, L., Zhang, D., Li, N., Li, Q., Dai, P. (2016). Oxidative stress-related biomarkers in postmenopausal osteoporosis: a systematic review and meta-analyses. *Dis Markers*. 2016:7067984. doi: 10.1155/2016/7067984

Zhou, Y., Gao, Y., Xu, C., Shen, H., Tian, Q., Deng, H. W. A. (2018b). Novel approach for correction of crosstalk effects in pathway analysis and its application in osteoporosis research. *Sci. Rep.* 8:668. doi: 10.1038/s41598-018-19196-2

Zhou, Y., Xu, C., Zhu, W., He, H., Zhang, L., Tang, B., et al. (2019). Long noncoding RNA analyses for osteoporosis risk in caucasian women. *Calcif Tissue Int*. 105, 183–192. doi: 10.1007/s00223-019-00555-8

Zhou, Y., Zhu, W., Zhang, L., Zeng, Y., Xu, C., Tian, Q., et al. (2018a). Transcriptomic data identified key transcription factors for osteoporosis in caucasian women. *Calcif Tissue Int*. 103, 581–588. doi: 10.1007/s00223-018-0457-6

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 December 2020
doi: 10.3389/fgene.2020.617136

Revealing the Interactions Between Diabetes, Diabetes-Related Diseases, and Cancers Based on the Network Connectivity of Their Related Genes

Lijuan Zhu[1], Ju Xiang[2,3], Qiuling Wang[4], Ailan Wang[5], Chao Li[5], Geng Tian[5,6], Huajun Zhang[1] and Size Chen[7]**

[1] College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, China, [2] Neuroscience Research Center, Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [3] School of Computer Science and Engineering, Central South University, Changsha, China, [4] Department of Endocrinology, The Affiliated Yantai Yuhuangding Hospital of Qingdao University, Yantai, China, [5] Geneis Beijing Co., Ltd., Beijing, China, [6] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [7] Department of Oncology, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangdong Provincial Engineering Research Center for Esophageal Cancer Precision Treatment, Guangzhou, China

OPEN ACCESS

Edited by:
Tao Huang, Chinese Academy of Sciences (CAS), China

Reviewed by:
Yulin Zhang, Shandong University of Science and Technology, China
Jujuan Zhuang, Dalian Maritime University, China

***Correspondence:**
Huajun Zhang
huajunzhang@zjnu.cn
Size Chen
chensize@gdpu.edu.cn

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

Received: *14 October 2020*
Accepted: *18 November 2020*
Published: *14 December 2020*

Citation:
Zhu L, Xiang J, Wang Q, Wang A, Li C, Tian G, Zhang H and Chen S (2020) Revealing the Interactions Between Diabetes, Diabetes-Related Diseases, and Cancers Based on the Network Connectivity of Their Related Genes. Front. Genet. 11:617136. doi: 10.3389/fgene.2020.617136

Diabetes-related diseases (DRDs), especially cancers pose a big threat to public health. Although people have explored pathological pathways of a few common DRDs, there is a lack of systematic studies on important biological processes (BPs) connecting diabetes and its related diseases/cancers. We have proposed and compared 10 protein–protein interaction (PPI)-based computational methods to study the connections between diabetes and 254 diseases, among which a method called DIconnectivity_eDMN performs the best in the sense that it infers a disease rank (according to its relation with diabetes) most consistent with that by literature mining. DIconnectivity_eDMN takes diabetes-related genes, other disease-related genes, a PPI network, and genes in BPs as input. It first maps genes in a BP into the PPI network to construct a BP-related subnetwork, which is expanded (in the whole PPI network) by a random walk with restart (RWR) process to generate a so-called expanded modularized network (eMN). Since the numbers of known disease genes are not high, an RWR process is also performed to generate an expanded disease-related gene list. For each eMN and disease, the expanded diabetes-related genes and disease-related genes are mapped onto the eMN. The association between diabetes and the disease is measured by the reachability of their genes on all eMNs, in which the reachability is estimated by a method similar to the Kolmogorov–Smirnov (KS) test. DIconnectivity_eDMN achieves an area under receiver operating characteristic curve (AUC) of 0.71 for predicting both Type 1 DRDs and Type 2 DRDs. In addition, DIconnectivity_eDMN reveals important BPs connecting diabetes and DRDs. For example, "respiratory system development" and "regulation of mRNA metabolic process" are critical in associating Type 1 diabetes (T1D) and many Type 1 DRDs. It is also found that the average proportion of diabetes-related genes interacting with DRDs is higher than that of non-DRDs.

Keywords: diabetes-related disease, PPI network, biological process, network connectivity, network modules

INTRODUCTION

With the increasing of human life-span, the incidence of diabetes is rapidly increasing, which presents a big threat to public health all over the world (Naslafkih and Sestier, 2003). According to a statistics from the International Diabetes Federation, approximately 415 million people worldwide suffered from diabetes in 2015, and the incidence is still increasing at a terrifying rate. By 2040, this number is estimated to exceed 640 million (International Diabetes Federation, 2015). Diabetes is a metabolic disease characterized by chronic hyperglycemia, which includes two forms, namely, Type 1 diabetes (T1D) and Type 2 diabetes (T2D). T2D accounts for about 85% of the diabetes incidences. Besides genetic factors, insulin resistance is a major risk factor for both T1D and T2D (Fourlanos et al., 2004). T1D and T2D also have a few common complications including damage to the kidneys, nerves, and cardiovascular systems, which may result in diabetes-related diseases (DRDs) like renal diseases (Papatheodorou et al., 2016, 2018). In general, DRDs can be divided into three categories: (1) microvascular disease, (2) macrovascular disease, and (3) miscellaneous complications. Microvascular disease mainly includes eye disease, kidney disease, and neuropathy; macrovascular disease mainly contains cardiovascular diseases; while miscellaneous complications include depression (Nouwen et al., 2011), dementia (Cukierman et al., 2005), and so on.

At present, people have explored the pathogenesis and pathological pathways of many DRDs. For example, inflammation, extracellular matrix expansion, oxidative stress, DNA damage, and vascular and nerve dysfunction are common pathways for the development of diabetic nephropathy (Wada and Makino, 2013; Jenkins et al., 2015; Zhang et al., 2018a); endothelial dysfunction and inflammation are involved in the development of diabetic vascular disease (Paneni et al., 2013); inflammation, endothelial dysfunction, and hypercoagulability are related to each other and play an important role in the occurrence of diabetic vascular disease (Domingueti et al., 2016). Though it is clear that certain biomarkers and biological pathways are involved in many DRDs, there is no systematic study summarizing DRD-associated common pathways, and pathways specific to the interaction between diabetes and specific DRDs.

With the development of high-throughput sequencing techniques, there are a lot of studies on genes and networks associated with diabetes and other diseases. For example, Ding et al. (2019) identified the core genes of T2D based on biological information, such as protein–protein interaction (PPI) network and microarray data. Zhang et al. (2018b) identified genes related to proliferative diabetic retinopathy based on PPI network and the random walk with restart (RWR) algorithm. Jiang et al. identified key genes and biological pathways related to diabetic nephropathy based on PPI network and microarray data (Jiang et al., 2015; Liu and Li, 2019; Song et al., 2019). The more and more accessible disease-related genes together with other important biological information, such as PPI data, gene expression data, and gene ontology (GO) data, provide us a unique opportunity for studying the interaction between diabetes and DRDs at the network level.

In this paper, we have proposed and compared 10 network-based computational methods to study the connections between diabetes and 254 diseases&vitamin D, which can generally be grouped into four categories, namely (1) DIcd based on the closest distance; (2) DIoverlap based on gene set overlap; (3) DINet based on random walk and gene set enrichment; (4) DIconnectivity based on cut edges between gene sets. Using these methods, we aim to predict DRDs, and perform a comprehensive analysis on important biological pathways associated with DRDs.

RESULTS

We have proposed four categories of algorithms to study the connections between diabetes and other diseases&vitamin D, namely, DIoverlap, DIcd, DINet, and DIconnectivity, all of which are based on PPI network/subnet. Since the diabetes-disease related genes might be enriched in a few biological processes (BPs) (Nigro et al., 2014), we also studied the connections based on BP modularized networks (MNs). The MN is constructed by mapping genes in each GO BP to the reference PPI network. In addition, we further expand each MN by an RWR procedure to construct the expanded MN (eMN). In our study, we set the expansion fold N to 3.

An Overview of DIoverlap, DIcd, DINet, and DIconnectivity

DIoverlap is the Jaccard coefficient between diabetes and disease gene set. We applied this algorithm to three types of networks including the whole network, MN, and eMN, corresponding to DIoverlap-Whole network, DIoverlap-MN, and DIoverlap-eMN, respectively. We define the mean of Jaccard coefficients across the MNs/eMNs as the evaluation standard for DIoverlap-MN/DIoverlap-eMN. An overview of other three algorithms DIcd, DINet, and DIconnectivity is presented in **Figure 1**. For each algorithm, the disease-related genes were mapped to the network first. DIcd is the closest distance from diabetes genes to disease genes on PPI network (see **Figure 1A**). The major steps of DINet are shown in **Figure 1B**, which is similar to GeroNet algorithm (Yang et al., 2016). For DINet algorithm, the diabetes and disease genes were mapped to each eMN and the connection between the two mapped gene sets was estimated using RWR and gene set enrichment analysis (GSEA); the significance of the connection was evaluated by a permutation analysis, in which the diabetes genes are randomly permuted, and the significance p-value is adjusted for multiple testing; the connection between diabetes and disease/vitamin D is evaluated by the minimum adjusted p-value. The details of each step are presented in Section "Materials and Methods."

DIconnectivity (**Figure 1C**) calculates the number of interactions between diabetes and disease gene set. We applied this algorithm to three types of networks including the whole network, MN, and eMN, corresponding to DIconnectivity-Whole network, DIconnectivity-MN, and DIconnectivity-eMN, respectively. We define the mean of interaction numbers across

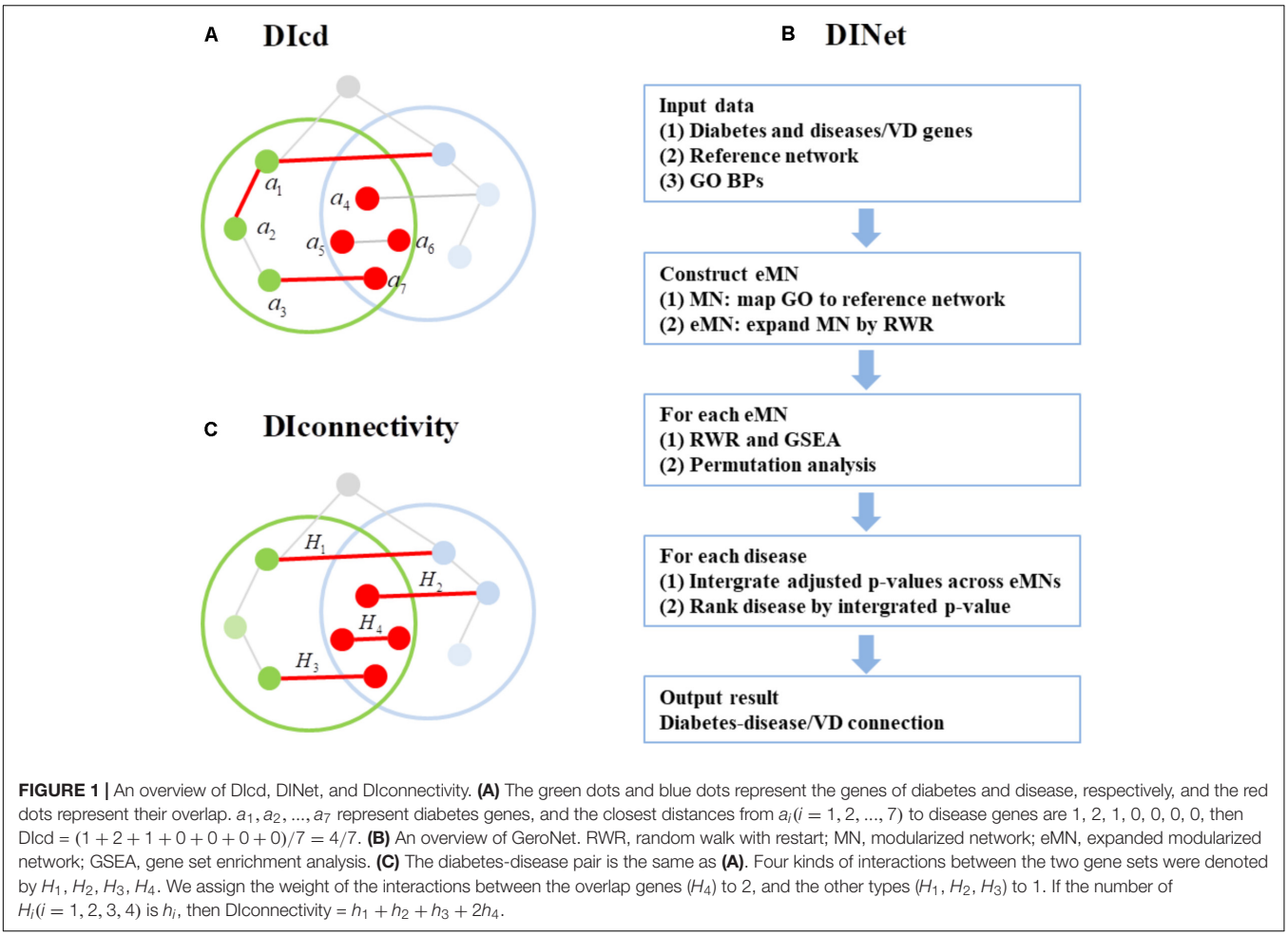

FIGURE 1 | An overview of DIcd, DINet, and DIconnectivity. **(A)** The green dots and blue dots represent the genes of diabetes and disease, respectively, and the red dots represent their overlap. $a_1, a_2, ..., a_7$ represent diabetes genes, and the closest distances from $a_i (i = 1, 2, ..., 7)$ to disease genes are 1, 2, 1, 0, 0, 0, 0, then DIcd = $(1 + 2 + 1 + 0 + 0 + 0 + 0)/7 = 4/7$. **(B)** An overview of GeroNet. RWR, random walk with restart; MN, modularized network; eMN, expanded modularized network; GSEA, gene set enrichment analysis. **(C)** The diabetes-disease pair is the same as **(A)**. Four kinds of interactions between the two gene sets were denoted by H_1, H_2, H_3, H_4. We assign the weight of the interactions between the overlap genes (H_4) to 2, and the other types (H_1, H_2, H_3) to 1. If the number of $H_i (i = 1, 2, 3, 4)$ is h_i, then DIconnectivity = $h_1 + h_2 + h_3 + 2h_4$.

the MNs/eMNs as the evaluation standard for DIconnectivity-MN/DIconnectivity-eMN. In addition, DIconnectivity-eDMN calculates the interaction number between the expand diabetes and disease gene set on eMNs, and the gene sets are expanded by RWR and GSEA.

Collection of Diabetes and Disease&vitamin D Genes, Reference PPI Network, GO BPs, and DRD Classification

We used diabetes/diseases genes collected from Enrichr as our input genes, and the genes of T1D/T2D/254 diseases were obtained by merging genes with the same human terms. Owing to some of the T1D/T2D/254 diseases also contain mouse or rat genes, we constructed two datasets: one of which only considers the human genes, called the H_Dataset, and the other one considers the genes of these three species, called HMR_Dataset. The vitamin D genes are obtained from GO terms which are related to vitamin D (i.e., the GO terms contain the word "vitamin D") and the number of this gene set is 57. The number of disease genes in H_Dataset ranges from 298 to 3875 and a full list of disease&vitamin D genes is provided in **Supplementary Dataset S1**, while the number in HMR_Dataset ranges from 298 to 4134 and the gene list is provided in **Supplementary Dataset S2**. Besides, the number of T1D/T2D genes in H_Dataset is 355/2109, and the number is 2288/3521 in HMR_Dataset.

We used the PPI network compiled by Menche et al. as the reference network, and considered 3367 GO BPs to define MNs (see section "Materials and Methods"). We annotated the diseases&vitamin D as being either diabetes-related or non-diabetes related based on literature mining. 41 diseases&vitamin D were annotated as DRD1s (**Supplementary Table S1**) and 29 diseases&vitamin D were annotated as DRD2s (**Supplementary Table S2**).

Comparison of DIoverlap, DIcd, DINet, and DIconnectivity

We used 10 methods to study the diabetes-disease&vitamin D connections based on PPI network/subnet, which are DIoverlap-Whole network, DIoverlap-MN, DIoverlap-eMN, DIoverlap-eDMN, DIcd, DINet, DIconnectivity-Whole network, DIconnectivity-MN, DIconnectivity-eMN, and DIconnectivity-eDMN. For DIoverlap-MN and DIconnectivity-MN, we only considered the MNs with the numbers of diabetes and

disease/vitamin D mapping genes greater than 5, while for DIoverlap-eMN, DIconnectivity-eMN, DIoverlap-eDMN, and DIconnectivity-eDMN, we only considered the eMNs, which are expanded by these MNs. In addition, for DIoverlap-eDMN and DIconnectivity-eDMN, we also performed permutation training of eMNs (see **Supplementary Material**). For DINet, we considered the eMNs with the numbers of diabetes and disease/vitamin D mapping genes greater than 5. We compared the methods according to the accuracy of predicting the DRD1s/DRD2s. To quantify the performance, we calculated the area under the receiver operating characteristic curve (AUROC or simply AUC) for each method, a commonly used statistics to characterize the overall performance of a predictive model. For DINet, we tested nine values for parameter (i.e., 0.1, 0.2, . . ., 0.9) to get the best prediction result; for DIoverlap-eDMN/DIconnectivity-eDMN, we tested 10 values for expansion fold N (i.e., 1, 2, . . ., 10) on diabetes and diseases&vitamin D genes, and denoted the corresponding methods as DIoverlap-eDMN_EN/DIconnectivity-eDMN_EN. For T1D/T2D, DIconnectivity-eDMN_E3/DIconnectivity-eDMN_E4 performed the best with AUC of 0.71/0.71 on HMR_Dataset (**Figure 2**). In **Figure 2**, we only plotted the AUC result of each method under the optimal parameter (if parameter is included), and the parameter training results of different methods are listed in **Supplementary Table S3**.

Diabetes Related Diseases Predicted by DIconnectivity-eDMN

We used the best performing method DIconnectivity_eDMN to predict the connections between diabetes and diseases&vitamin D, and the predicted ranking list of all 254 diseases&vitamin D related to T1D/T2D is provided in **Supplementary Datasets S3, S4**. It should be noted that we only considered the eMNs whose numbers of interactions between gene sets were greater than 0 for each diabetes-disease pair. In order to find significant related diseases, we converted the DIconnectivity into z-score statistics and calculated the *p*-values and then the diseases with *p*-values less than 0.05 were significant DRDs (**Table 1**). Finally, we found 22 significant related diseases of T1D/T2D. Among these DRDs, bacterial infection, acute myocardial infarction, atherosclerosis, osteoarthritis, and obesity are well-known DRDs. For bacterial infections, the mechanism of the susceptibility is the influence of glycemia on polymorphonuclear cell functions, such as urinary tract infection, "diabetic foot," or "infectious cellulitis" (Schubert and Heesemann, 1995). Besides, certain infections (i.e., respiratory and foot infections) are overrepresented in the diabetic population and are associated with a higher risk of infection-related mortality (Pearson-Stuttard et al., 2016). On the one hand, diabetes increases the risk of acute myocardial infarction; on the other hand, acute myocardial infarction is the major cause of morbidity and mortality in diabetic patients (Echouffo-Tcheugui et al., 2018). The statistics from US centers for disease control and prevention (CDC[1]) also note that heart disease is the leading cause of death among people with diabetes. Diabetes is also associated with elevated odds of having osteoarthritis, which is the most frequent disease in individuals with diabetes (Rehling et al., 2019). The relationship between diabetes and obesity is more obvious (Weyer et al., 2001; Okada-Iwabu et al., 2013). According to the latest statistics from CDC, 89% of diabetes patients in the United States are overweight or obese (body mass index > 25 kg/m^2). In Brazilian, 75% of the T2D patients are overweight, and 30% of them are obese (Gomes et al., 2006).

It should be noted that for both T1D and T2D, systemic lupus erythematosus ranked first, which is associated with an increased risk of development of diabetes (Chung et al., 2007; Jiang et al., 2018). A cohort study in Toronto documented that women with SLE had a significantly higher prevalence of diabetes than the age-matched healthy controls (5 versus 1%) (Bruce et al., 2003). Therefore, we can conclude that SLE patients may develop diabetes. Followed in the list are breast cancer and asthma. According to Cancer Research UK[2], women with diabetes have an increased risk of breast cancer. In addition, some studies have shown that diabetes not only increases the risk of breast cancer (Liao et al., 2011), but also increases the risk of breast cancer death (Luo et al., 2014; Bronsveld et al., 2015). The published data on disease occurrence showed that there was a strong positive association between T1D and asthma in Europe and elsewhere (Stene and Nafstad, 2001). Similarly, T2D has attracted attention as a risk factor for asthma (Murakami et al., 2019). Followed in the list are various types of psychiatric disorders, neurodegenerative diseases, and cancers. According to CDC, the complications of diabetes include heart disease, nerve damage, and mental health. On the other hand, some studies have shown that bipolar disorder (McIntyre et al., 2005), schizophrenia (Hoffman, 2017), and autism spectrum disorder (Alhowikan et al., 2019) also increase the prevalence of diabetes. In addition, high blood sugar can cause neuropathy (nerve damage) throughout your body, and some studies also suggested that there was an association between diabetes and the neurodegenerative diseases multiple sclerosis and amyotrophic lateral sclerosis (Mariosa et al., 2015; Tettey et al., 2015). Additionally, Cancer Research UK notes that people with diabetes have an increased risk of pancreatic cancer[3]. What is more, several studies show a higher risk of womb cancer in women with diabetes[4]. We should also note that diabetes is one of the common comorbidities of ulcerative colitis (Maconi et al., 2014) and cystic fibrosis (Prentice et al., 2016; Hart et al., 2018).

DIconnectivity-eDMN can effectively rank some recognized DRDs at the top of the list, but there are still some obvious related diseases that are relatively backward, such as diabetic nephropathy of T1D (48th), insulin resistance of T2D (68th), and even put some diabetes related diseases at the bottom of the list, such as vitamin D (255th) and morbid obesity (240th) of T1D/T2D. Such a ranking error may be due to incomplete genes in our network or diseases&vitamin D. In addition, some DRDs were not defined as DRDs, but we did find evidence to support

[1]https://www.cdc.gov/diabetes/managing/problems.html

[2]https://www.cancerresearchuk.org/about-cancer/breast-cancer/risks-causes/risk-factors

[3]https://www.cancerresearchuk.org/about-cancer/pancreatic-cancer/risks-causes

[4]https://www.cancerresearchuk.org/about-cancer/womb-cancer/risks-causes

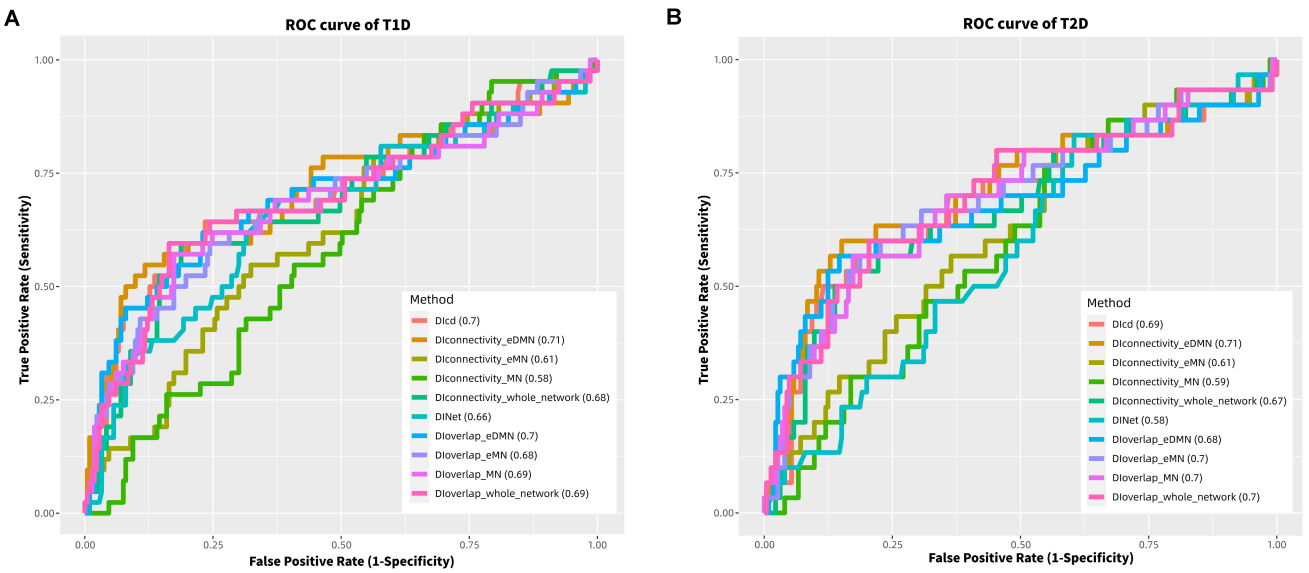

FIGURE 2 | Comparison of different methods based on AUC of ROC in T1D **(A)** and T2D **(B)**. DIcd is based on whole PPI network; DINet is based on expanded modularized network (eMN); DIconnectivity_whole network represents DIconnectivity based on whole PPI network; DIconnectivity_MN represents DIconnectivity based on modularized network (MN); DIconnectivity_eMN represents DIconnectivity based on expanded modularized network (eMN); DIoverlap is defined similarly.

TABLE 1 | The significant diabetes-related diseases inferred by DIconnectivity-eDMN.

Disease	*p*-value	DRD1	Disease	*p*-value	DRD1
Type 1 diabetes					
Systemic lupus erythematosus	1.24 E-04	1	Endometrial cancer	1.06 E-02	0
Breast cancer	4.45 E-04	0	Acute myocardial infarction	1.13 E-02	1
Bacterial infection	1.07 E-03	1	Endometriosis	1.58 E-02	0
Asthma	1.08 E-03	1	Cystic fibrosis	1.62 E-02	1
Ulcerative colitis	1.64 E-03	1	Huntington's disease	2.29 E-02	0
Bipolar disorder	2.16 E-03	0	Multiple sclerosis	3.26E-02	1
Crown's disease	2.49 E-03	1	Pancreatic cancer	3.58 E-02	1
Polycystic ovary syndrome	2.67 E-03	1	Osteoarthritis	4.14 E-02	0
Hypoxia	2.96 E-03	1	Obesity	4.31 E-02	1
Schizophrenia	3.77 E-03	0	Amyotrophic lateral sclerosis	4.36 E-02	0
Autism spectrum disorder	6.54E-03	0	Prostate cancer	4.84 E-02	0
Disease	***p*-value**	**DRD2**	**Disease**	***p*-value**	**DRD2**
Type 2 diabetes					
Systemic lupus erythematosus	6.52 E-04	0	Schizophrenia	1.10 E-02	0
Bacterial infection	7.74 E-04	1	Endometriosis	1.16 E-02	0
Asthma	1.72 E-03	0	Autism spectrum disorder	1.70 E-02	0
Breast cancer	1.73 E-03	1	Cystic fibrosis	1.76 E-02	0
Crown's disease	2.13 E-03	1	Pancreatic cancer	2.66 E-02	1
Hypoxia	4.08E-03	1	Osteoarthritis	2.90 E-02	0
Bipolar disorder	5.09 E-03	0	Multiple sclerosis	3.07 E-02	0
Ulcerative colitis	5.64 E-03	0	Alzheimer's disease	3.28 E-02	0
Polycystic ovary syndrome	6.99 E-03	1	Obesity	3.35E-02	1
Endometrial cancer	7.33 E-03	0	Huntington's disease	4.76 E-02	0
Acute myocardial infarction	1.01 E-02	1	Atherosclerosis	4.95 E-02	1

their connections, such as bipolar disorder, endometrial cancer, and osteoarthritis.

Functional Subnets Connecting Diabetes and Diseases&vitamin D

For each diabetes-disease/vitamin D connection, we consider eMNs satisfying the following two conditions: (1) the eMNs are under the optimal permutation result; (2) the interaction numbers between diabetes and disease/vitamin D mapping genes in the eMNs are greater than 0. In the 255 T1D/T2D-disease&vitamin D connections, the number of eMNs ranges from 295/298 to 3123/3291, total 427,349/431,778 eMNs. Generally speaking, not all subnets play an important role in the diabetes-disease&vitamin D connections, so we identify the significant eMNs for each diabetes-disease/vitamin D connection with permutation analysis method, and the specific steps are as follows: (1) permute diabetes genes in each eMN for 100 times to calculate the null distribution of DIconnectivity with DIconnectivity-eDMN_E3 for T1D and DIconnectivity-eDMN_E4 for T2D and (2) convert the DIconnectivity to a z-score statistic based on this null distribution, then a p-value is estimated and adjusted for multiple testing. We consider the eMNs with FDR $\leq$ 0.05 are significant, and the number of significant eMNs for T1D/T2D-disease&vitamin D connections ranges from 46/0 to 1908/1284, a total of 214,545 ($\sim$50.2%)/84,165 ($\sim$19.5%) significant eMNs.

Functional Subnets Connecting T1D and Diseases&vitamin D

It is worth noting that different eMNs have different frequencies to connect diabetes and diseases&vitamin D, that is, some eMNs are involved in multiple diabetes-disease&vitamin D connections, and some only affect a few or specific ones. In order to study eMN frequency in the T1D-disease&vitamin D connections, we calculated the frequencies of all significant eMNs for each connection, and the average frequency (AF) was used as its eMN frequency. Among 255 T1D-disease&vitamin D (42 DRD1s and 213 non-DRD1s) connections (AF$\in$[102, 201]), there are 92 connections with AF less than 150, of which 23 are DRD1s involved and 69 are non-DRD1s involved. This shows that 55% of T1D-DRD1 connections have an eMN frequency of less than 150, while for non-DRD1s, this proportion is only 32%. Obviously, the smaller the eMN frequency, the higher the specificity, and then we can conclude that DRD1s have higher eMN specificity to connect T1D compared to non-DRD1s. The AFs of 42 connections (DRD1s involved) are plotted in **Figure 3A**, and from the figure, we can see that some well-known DRD1s have low frequencies (e.g., morbid obesity and diabetic nephropathy). The higher frequent diseases include heart diseases (e.g., cardiomyopathy and atherosclerosis) and inflammatory diseases (e.g., colitis and eczema), which suggests that the connections between T1D and DRD1s may be mediated by eMNs with very different frequencies.

In order to further search for specific eMNs and non-specific eMNs of T1D-DRD1 connections, we defined the specific index SP (SP = KF/AF, 0 < KF$\leq$42, 0 < AF$\leq$255, 0 < SP$\leq$1),

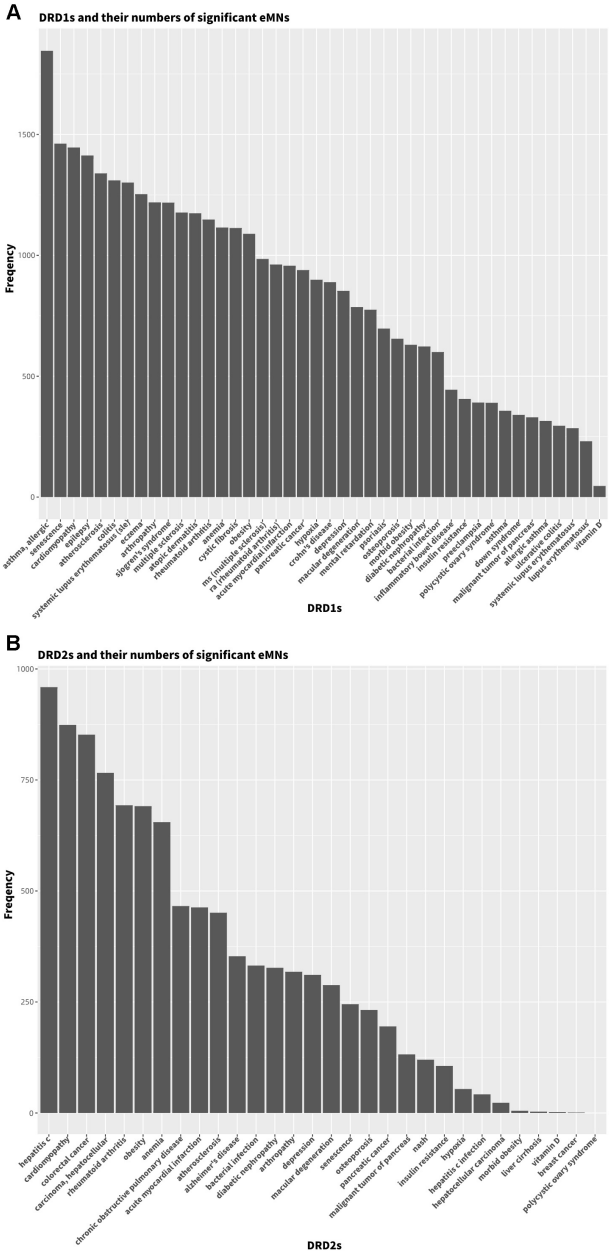

FIGURE 3 | The numbers of significant eMNs in T1D-DRD1 connections **(A)** and T2D-DRD2 connections **(B)**.

where KF is the frequency of significant eMN in the range of T1D-DRD1 connections. It is easy to know that when AF is closer to KF and KF is closer to 42, the eMN specificity is higher. Therefore, we set the SP threshold to 0.3 (**Supplementary Table S4**), i.e., the eMN with SP greater than 0.3 is defined as specific eMN, otherwise non-specific eMN. We sorted the specific eMNs according to KF from large to small, and list the top 20 non-specific eMNs and specific eMNs in **Table 2**. Non-specific eMNs include BPs such as "GO:0060070_canonical Wnt signaling pathway" and "GO:0060828_regulation of canonical

TABLE 2 | Top 20 non-specific and specific eMNs (KF from large to small) of T1D-DRD1connections.

Non-specific eMNs	KF	AF	SP
GO:0060070_canonical Wnt signaling pathway	41	237	0.172995781
GO:0060828_regulation of canonical Wnt signaling pathway	41	226	0.181415929
GO:2000027_regulation of animal organ morphogenesis	41	224	0.183035714
GO:0000226_microtubule cytoskeleton organization	40	244	0.163934426
GO:0022604_regulation of cell morphogenesis	40	240	0.166666667
GO:0051090_regulation of DNA-binding transcription factor activity	40	243	0.164609053
GO:0050769_positive regulation of neurogenesis	40	243	0.164609053
GO:0051047_positive regulation of secretion	40	243	0.164609053
GO:0042391_regulation of membrane potential	40	249	0.16064257
GO:0002793_positive regulation of peptide secretion	40	233	0.17167382
GO:0016055_Wnt signaling pathway	39	241	0.161825726
GO:0198738_cell-cell signaling by wnt	39	240	0.1625
GO:0030111_regulation of Wnt signaling pathway	39	226	0.172566372
GO:0016050_vesicle organization	39	224	0.174107143
GO:0022412_cellular process involved in reproduction in multicellular organism	39	229	0.170305677
GO:0050804_modulation of chemical synaptic transmission	39	240	0.1625
GO:0051091_positive regulation of DNA-binding transcription factor activity	39	235	0.165957447
GO:0072001_renal system development	39	243	0.160493827
GO:0001822_kidney development	39	234	0.166666667
GO:0001655_urogenital system development	39	222	0.175675676
Specific eMNs			
GO:0060541_respiratory system development	28	91	0.307692308
GO:1903311_regulation of mRNA metabolic process	23	75	0.306666667
GO:1902105_regulation of leukocyte differentiation	20	59	0.338983051
GO:0052548_regulation of endopeptidase activity	20	66	0.303030303
GO:0007517_muscle organ development	18	53	0.339622642
GO:0071383_cellular response to steroid hormone stimulus	18	47	0.382978723
GO:0031100_animal organ regeneration	18	56	0.321428571
GO:0060537_muscle tissue development	17	49	0.346938776
GO:0009267_cellular response to starvation	17	48	0.354166667
GO:0003007_heart morphogenesis	17	51	0.333333333
GO:0021782_glial cell development	16	52	0.307692308
GO:0048545_response to steroid hormone	16	53	0.301886792
GO:0002521_leukocyte differentiation	16	42	0.380952381
GO:0071901_negative regulation of protein serine/threonine kinase activity	16	42	0.380952381
GO:0048771_tissue remodeling	16	46	0.347826087
GO:0042110_T cell activation	16	52	0.307692308
GO:0048732_gland development	16	51	0.31372549
GO:0043434_response to peptide hormone	15	46	0.326086957
GO:0051169_nuclear transport	15	42	0.357142857
GO:0036473_cell death in response to oxidative stress	15	44	0.340909091

Wnt signaling pathway." According to **Table 2**, there are 41 DRD1s (42 in total) and 196 non-DRD1s (213 in total) that are significantly related to the eMN "GO:0060070_canonical Wnt signaling pathway," and there are 41 (98%) DRD1s and 185 (87%) non-DRD1s that are significantly related to the eMN "GO:0060828_regulation of canonical Wnt signaling pathway." The Wnt signaling pathway has been reported to be associated with glucose and lipid metabolism (Qin et al., 2018). Besides, many studies have shown that the Wnt signaling pathway is related to the pathogenesis of diabetic nephropathy (Kavanagh et al., 2011) and diabetic retinopathy (Chen and Ma, 2017). In the non-specific eMNs, except for multiple pathways related to Wnt signaling (GO:0016055_Wnt signaling pathway, GO:0198738_cell-cell signaling by wnt, GO:0030111_regulation of Wnt signaling pathway), there are also eMNs related to kidney development, such as GO:0072001_renal system development, GO:0001822_kidney development, and GO:0001655_urogenital system development.

Specific eMNs include BPs such as "respiratory system development" and "regulation of mRNA metabolic process." There are 28 (68%) DRD1s and 63 (30%) non-DRD1s that are significantly related to "respiratory system development," and there are 23 (55%) DRD1s and 52 (24%) non-DRD1s that are significantly related to "regulation of mRNA metabolic process." Related studies have shown that respiratory control imbalance is common in T1D patients (Bianchi et al., 2017). The available evidence shows that diabetes usually changes metabolites such as glucose, fructose, amino acids, and lipids through metabolic pathways (Arneth et al., 2019). In addition, the well-known specific eMNs of diabetes, insulin related BPs (GO:0032868_response to insulin, GO:0032869_cellular response to insulin stimulus) are also in the list (Brezar et al., 2011).

Functional Subnets Connecting T2D and Diseases&vitamin D

We conducted a similar analysis for T2D. Among 255 T2D-diseases&vitamin D (30 DRD2s and 225 non-DRD2s) connections (AF∈[10, 209]), there are 94 connections with AF less than 100, of which 16 are DRD2s involved and 78 are non-DRD2s involved. This shows that 53% of T2D-DRD2 connections have an eMN frequency of less than 100, while for non-DRD2s, this proportion is only 35%. The AFs of 30 connections (DRD2s involved) are plotted in **Figure 3B**, and from the figure, we can see that high frequent diseases include obesity and some heart diseases (cardiomyopathy, acute myocardial infarction, and atherosclerosis).

We set the SP threshold to 0.2 (**Supplementary Table S5**) to define specific eMNs and non-specific eMNs, and list the top 20 of them in **Table 3**. The non-specific eMN with the largest KF is "GO:0000226_microtubule cytoskeleton organization," and there are 22 DRD2s (30 in total) and 178 non-DRD2s (225 in total) that are significantly related to it. Studies have found that microtubule polymerization may play an important role in glucose transport (Taneja and Priyadarshini, 2018). It is worth noting that pathways related to Wnt signaling are also significantly related to T2D, such

TABLE 3 | Top 20 non-specific and specific eMNs (KF from large to small) of T2D-DRD2 connections.

Non-specific eMNs	KF	AF	SP
GO:0000226_microtubule cytoskeleton organization	22	200	0.11
GO:0051052_regulation of DNA metabolic process	22	206	0.106796117
GO:0016570_histone modification	22	184	0.119565217
GO:0198738_cell-cell signaling by wnt	22	182	0.120879121
GO:0016055_Wnt signaling pathway	22	178	0.123595506
GO:0090068_positive regulation of cell cycle process	22	191	0.115183246
GO:0048285_organelle fission	22	209	0.105263158
GO:0045787_positive regulation of cell cycle	22	195	0.112820513
GO:0045930_negative regulation of mitotic cell cycle	22	187	0.117647059
GO:0034660_ncRNA metabolic process	21	176	0.119318182
GO:0051260_protein homooligomerization	21	165	0.127272727
GO:0000082_G1/S transition of mitotic cell cycle	21	180	0.116666667
GO:0072331_signal transduction by p53 class mediator	21	176	0.119318182
GO:1901987_regulation of cell cycle phase transition	21	196	0.107142857
GO:0031396_regulation of protein ubiquitination	21	167	0.125748503
GO:1901990_regulation of mitotic cell cycle phase transition	21	189	0.111111111
GO:0060249_anatomical structure homeostasis	20	202	0.099009901
GO:0016569_covalent chromatin modification	20	144	0.138888889
GO:0060070_canonical Wnt signaling pathway	20	172	0.11627907
GO:0044843_cell cycle G1/S phase transition	20	166	0.120481928
Specific eMNs			
GO:0061138_morphogenesis of a branching epithelium	14	67	0.208955224
GO:0007626_locomotory behavior	13	60	0.216666667
GO:0001890_placenta development	13	65	0.2
GO:0007162_negative regulation of cell adhesion	12	55	0.218181818
GO:0001894_tissue homeostasis	12	50	0.24
GO:0060562_epithelial tube morphogenesis	12	43	0.279069767
GO:0034101_erythrocyte homeostasis	12	55	0.218181818
GO:0048469_cell maturation	12	52	0.230769231
GO:0009267_cellular response to starvation	11	45	0.244444444
GO:0007179_transforming growth factor beta receptor signaling pathway	11	44	0.25
GO:0042594_response to starvation	11	43	0.255813953
GO:0048762_mesenchymal cell differentiation	11	42	0.261904762
GO:0051100_negative regulation of binding	11	53	0.20754717
GO:0051047_positive regulation of secretion	11	39	0.282051282
GO:0030098_lymphocyte differentiation	11	50	0.22
GO:0001558_regulation of cell growth	11	46	0.239130435
GO:0006732_coenzyme metabolic process	11	42	0.261904762
GO:0032259_methylation	10	41	0.243902439
GO:0090287_regulation of cellular response to growth factor stimulus	10	46	0.217391304
GO:0019359_nicotinamide nucleotide biosynthetic process	10	41	0.243902439

as GO:0198738_cell-cell signaling by wnt, GO:0016055_Wnt signaling pathway, GO:0030111_regulation of Wnt signaling pathway, and GO:0060828_regulation of canonical Wnt signaling pathway, and there are evidences that the Wnt signaling pathway is a key pathway for the occurrence of T2D (Lee et al., 2008; Liu et al., 2018). Therefore, we can conclude that both T1D and T2D are significantly related to the Wnt signaling pathway. On the other hand, the Wnt signaling pathway is also related to the development of some DRD2s, for example, miR-128-3p aggravates cardiovascular calcification and insulin resistance in T2D rats by downregulating ISL1 through the activation of the Wnt pathway (Wang et al., 2019). The specific eMN with the largest KF is "GO:0061138_morphogenesis of a branching epithelium," and studies have found that branching morphogenesis is a critical step in the development of many epithelial organs, for example, lung (Carter et al., 2014; Goodwin et al., 2019), kidney (Basson et al., 2006), and breast, besides, breast epithelial branch morphogenesis may be related to breast cancer (Kessenbrock et al., 2017). In addition, the similar BPs of morphogenesis of a branching epithelium (GO:0060562_epithelial tube morphogenesis ranked 6, GO:0001763_morphogenesis of a branching structure ranked 22 and GO:0048754_branching morphogenesis of an epithelial tube ranked 25) are also in the specific eMN list, which further indicates that the BP of morphogenesis of a branching epithelium structure is important for T2D-DRD2 connections.

Key Connectors Mediating Diabetes-Disease Connections in Significant Subnets

We performed key connector analysis (KCA) to infer key genes that connect diabetes and DRDs in selected eMNs. The detailed information of KCA is provided in Section "Materials and Methods." We selected two common diabetes-disease connections including T1D-bacterial infection and T2D-obesity as case studies to illustrate the key connectors (**Figure 4**). In **Figure 4**, we only show the subnet consisting of key connectors and their neighboring genes for a better view.

The T1D-bacterial infection connection is most significant in the eMN corresponding to "GO:0016055_Wnt signaling pathway." In this eMN, there are 111 T1D genes and 58 bacterial infection genes, and the number of overlap between them is 30. We analyzed these 30 common genes with key driver analysis (KDA), and the key connector gene HSPA8 was obtained (**Figure 4A**). Studies have shown that HSPA8 binds bacterial lipopolysaccharide (LPS) and mediates LPS-induced inflammatory response (Yahata et al., 2000; Triantafilou et al., 2001). Similarly, T2D-obesity connection is the most significant in the eMN corresponding to "GO:0035107_appendage morphogenesis." In this eMN, there are 84 T2D genes and 30 obesity genes, and the number of overlap between them is 23. We analyzed these 23 common genes, and got the key connector genes TCF4, CTNNB1 and CEBPB (**Figure 4B**). TCF4 (TCF7L2) is the strongest T2D candidate gene discovered to date, and it also plays a key role in the development and function of adipose tissue (Chen et al., 2018). CTNNB1 (β-catenin) is a key regulator

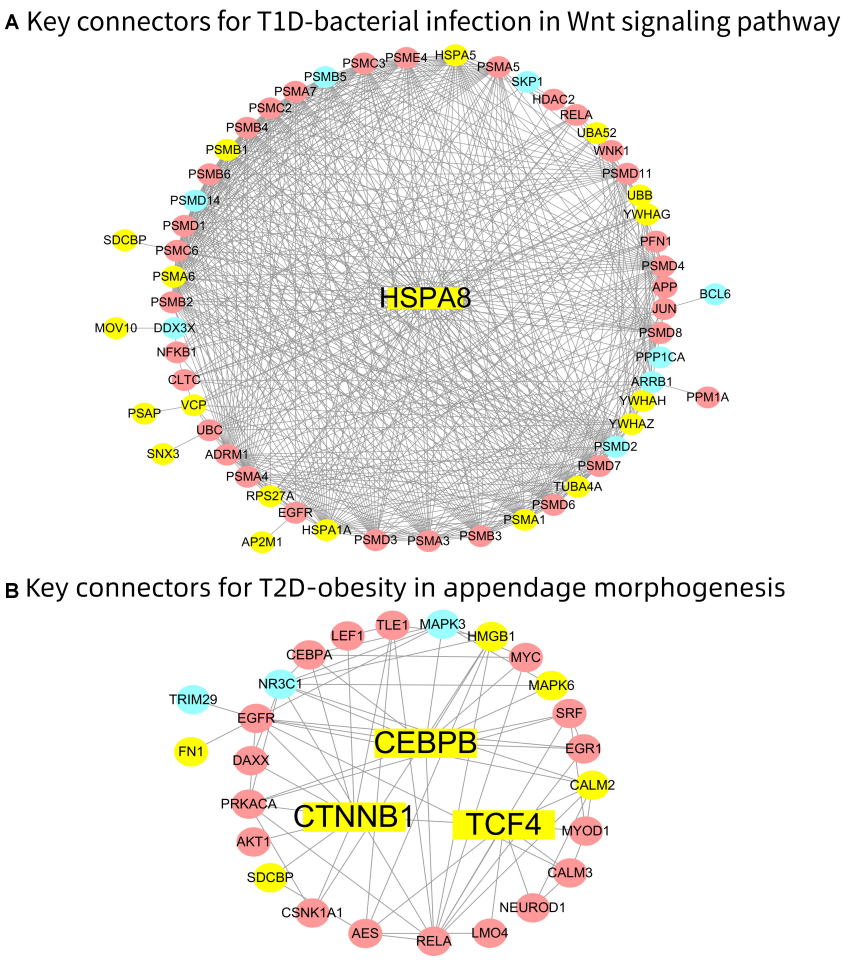

FIGURE 4 | Key connectors of **(A)** T1D-bacterial infection in Wnt signaling pathway and **(B)** T2D-obesity in appendage morphogenesis. We use node shape to denote key connectors: (1) square represents the key connectors; (2) circle represents diabetes and disease genes. We use fill color to denote diabetes and disease information: (1) red represents diabetes gene; (2) blue represents disease gene; and (3) yellow represents the overlapping diabetes and disease gene.

of fat expansion and human obesity (Chen et al., 2020). Besides, studies have shown that co-administration of insulin and leptin to pancreatic islet-derived mesenchymal stem cell (PID-MSC) leads to the co-development of insulin and leptin resistance, and the differentiation signaling is mainly mediated by CTNNB1 and Tub (Ercin et al., 2018). In addition, the repressed expression of CEBPB has been found in obesity and T2D in adipose tissue (Li et al., 2016).

DISCUSSION

The Effect of the Number of Disease Genes on the Result

In the predicted diabetes-related diseases&vitamin D list, some well-known diseases are not ranked at the top (e.g., diabetic nephropathy, insulin resistance) and we speculate that it may be caused by the difference in the number of known disease genes. We sorted the numbers of diseases&vitamin D genes from large to small, and found that diabetic nephropathy ranked 118 and vitamin D ranked 255. In order to prove our conjecture, we expanded (limited) the genes of each disease to 500 and predicted DRDs with DIconnectivity_eDMN method on HMR_Dataset (**Supplementary Table S6**). We found that the predicted top five related diseases of T1D and T2D are the same, all of which are "essential thrombocytemia," "bacterial infection," "osteoporosis," "diabetic nephropathy," "insulin resistance." This result shows that the number of known disease genes indeed affects the ranking of related diseases to a certain extent. Additionally, vitamin D has the least number of genes and is predicted at the bottom of the list, which further proves our conjecture.

Diabetes Genes Play an Important Role in the Diabetes-Disease Connections

We calculated the proportion of T1D genes involved in the DIconnectivity (DIconnectivity-Whole network) for each T1D-DRD1 connection, and found that the average proportion was 0.64, while in T1D-non-DRD1 connections, the proportion was

only 0.54. Similarly, in T2D-DRD2 connections and T2D-non-DRD2 connections, the average proportions were 0.59 and 0.51, respectively. In addition, we also calculated the proportion of disease genes involved in the corresponding DIconnectivity, and found that the average proportions of DRD1s and non-DRD1s were 0.80 and the average proportions of DRD2s and non-DRD2s were 0.88 and 0.87, respectively. The average proportion of disease genes is higher than that of diabetes, but the proportions is the same for DRDs and non-DRDs, which shows that diabetes plays a key role in diabetes-disease connections.

For the shortest path method, we considered three distance measures (Guney et al., 2016): (1) the shortest distance $d_s(A, S)$, $d_s(A, S) = \frac{1}{||S||}\sum_{a\in A}\frac{1}{||A||}\sum_{s\in S} d(a, s)$, where A is diabetes gene set, S is disease gene set, and $d(a, s)$ is the shortest path length between nodes a and s in PPI network; (2) the closest distance $d_c(S, A)$, $d_c(S, A)=\frac{1}{||S||}\sum_{s\in S}\min_{a\in A} d(s, a)$, $d(s, a)=d(a, s)$; (3) the closest distance $d_c(A, S)$, $d_c(A, S)=\frac{1}{||A||}\sum_{a\in A}\min_{s\in S} d(a, s)$. We found that $d_c(A, S)$ has the best results (**Supplementary Table S7**). Among these three methods, $d_s(A, S)$ considers all genes of diabetes and disease, $d_c(S, A)$ only considers all genes of disease, and $d_c(A, S)$ only considers all genes of diabetes. Therefore, we can conclude that diabetes plays a more important role in diabetes-disease connections.

The Important Genes and Distances in the Diabetes-Disease Connections

DIoverlap method takes the intersection between diabetes and disease gene sets as a criterion for measuring their connection. In essence, it only considers the genes with distances of 0; DIconnectivity method considers the genes with distances of 0 and 1; DIcd method considers all diabetes genes regardless of distance. Among the three methods, DIconnectivity_eDMN performs best, which shows that the genes with distances of 0 and 1 play an important role in the diabetes-disease connections.

The Impact of BP Redundancy

In order to evaluate the impact of BP redundancy on prediction results, we calculated the semantic similarity among 3367 BP terms using R software package GOSemSim, of which 1141/359/59 terms have semantic similarity less than 0.8/0.7/0.6. Too high similarity and few terms are not our selection criteria, so we adopted the optimal method DIconnectivity-eDMN to predict DRDs again based on eMNs with similarity less than 0.7. Through the training of DIconnectivity-eDMN_EN ($N = 1, 2,\ldots, 10$), we found that DIconnectivity-eDMN_E3/DIconnectivity-eDMN_E4 has the best prediction for DRD1s/DRD2s with AUC of 0.70/0.71. Therefore, we can conclude that removing a few highly similar terms has very little impact on the prediction effect.

MATERIALS AND METHODS

Database

We downloaded the upregulated and downregulated gene files of diabetes/diseases Disease_Perturbations_ from_GEO_up.txt (**Supplementary Dataset S5**) and Disease_ Perturbations_ from_GEO_down.txt (**Supplementary Dataset S6**) from Enrichr[5]. Enrichr is a comprehensive resource for curated gene sets, currently containing 180,184 annotated gene sets from 102 gene set libraries (Kuleshov et al., 2016). Terms of these two files are the same, but the corresponding genes are different, so we first merge the upregulated and downregulated genes of each term, and get a total of 839 terms of human, mouse, and rat. In addition, since some diabetes/diseases terms are the same but only the case of the first letter is different, so we merged the same human terms of diabetes/diseases. Finally, we obtained a list of genes for 254 diseases and T1D/T2D (**Supplementary Dataset S1**). Besides, we also extracted vitamin D genes from GO terms which were related to vitamin D. In addition, we found that diabetes and some of 254 diseases not only contain human term genes, but also mouse or rat term genes, so we constructed another dataset by adding them to the corresponding disease gene set (**Supplementary Dataset S2**).

We used the human PPI network compiled by Menche et al. as the reference PPI network (Guney et al., 2016), and conducted research based on its largest connected subnet, which consists of 13,329 proteins and 141,150 protein interactions.

Gene ontology terms were obtained based on R software package GO.db. We consider GO BPs containing 30–500 genes, and ignore either very small or overly large functional gene sets. Finally, we obtained 3367 GO BPs to generate various network modules.

Diabetes–Disease/Vitamin D Connection Annotation

We adopted literature mining approach to annotate whether a disease/vitamin D is diabetes-related. Specifically, we ranked diseases&vitamin D based on their Jaccard indices between their names and the term "type 1 diabetes" ("type 2 diabetes") in PubMed abstracts published from 2008 to 2019. The PubMed abstracts containing the term "type 1 diabetes" ("type 2 diabetes") from 2008 to 2019 were retrieved using Entrez Programming Utilities[6]. The term "type 1 diabetes" corresponds to "type 1 diabetes" [MeSH Terms] OR "type 1 diabetes" [All Fields] in PubMed, which is a superset of the term "type 1 diabetes." The co-occurrence of disease and diabetes was evaluated by the following equation:

$$Jaccard(disease, diabetes) = \frac{|PubMedID_{disease} \cap PubMedID_{diabetes}|}{|PubMedID_{disease} \cup PubMedID_{diabetes}|}$$

Where $PubMedID_{disease}$ and $PubMedID_{diabetes}$ were the PubMed IDs containing the disease name and the term "diabetes," respectively.

According to previous study, some diseases are indeed associated with diabetes, such as diabetic nephropathy, obesity, and bacterial infection (Forbes and Cooper, 2013). We used the minimum Jaccard coefficient of these diseases as the threshold, and selected the diseases&vitamin D with Jaccard coefficient larger than threshold as DRDs. Finally, we obtained

[5]https://maayanlab.cloud/Enrichr/

[6]http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db~=~pubmed&term~=~[type1diabetes]+AND+2008:2019[pdat]&retmax~=~999999

41 diseases&vitamin D that are defined as DRD1s and 29 diseases&vitamin D that are defined as DRD2s.

Four Categories of Algorithms

We used four algorithms to identify the connections between diabetes and diseases&vitamin D, namely DIoverlap rank the diseases&vitamin D by calculating the Jaccard coefficient, DIcd performed by using the closest distance. DINet based on a procedure similar to gene set enrichment analysis and an RWR procedure, and DIconnectivity based on the number of interactions between diabetes and diseases&vitamin D genes.

DIcd

DIcd is a closest distance method: let A and S denote diabetes and disease gene set, respectively, and $d_c(A, S)$ is the closest distance from A to S. Given two nodes $a \in A$ and $s \in S$, the shortest path length between a and s in the network is represented by $d(a, s)$, then we define $d_c(A, S)$ as follows:

$$d_c(A, S) = \frac{1}{||A||} \sum_{a \in A} \min_{s \in S} d(a, s)$$

It should be noted that the smaller the value of DIcd, the higher the connection between diabetes and disease.

DIoverlap

DIoverlap is the Jaccard coefficient between diabetes and disease gene set, and the larger the value, the higher the connection between them.

DINet

DINet is similar to the GeroNet (Yang et al., 2016, 2017) and it consists of three steps: (1) Generate expanded network modules (eMN), (2) Calculate the enrichment scores on eMNs follow a method similar to GSEA, and (3) Calculate the significance of enrichment score based on permutation test.

Step 1: To generate expanded network modules (eMN), we map each GO BP to the reference PPI network to generate the corresponding MN, which is further expanded by an RWR (see **Supplementary Material**) until it reaches N times the original gene size and the maximum does not exceed 500 genes.

Step 2: To calculate the diabetes-disease enrichment score on an eMN, we first map the two gene sets to the eMN and perform two RWR expansions by setting the two mapped gene sets as seeds, which will rank all genes in the eMN, respectively. We go through the sorted gene list of eMN based on disease (diabetes) gene seed, if we encounter a gene that is not a diabetes (disease) gene, $-\sqrt{\frac{G}{N-G}}$ is added to the score, where N is the number of genes for the network, and G is the number of diabetes (disease) genes; otherwise, $\sqrt{\frac{N-G}{G}}$ is added. This generates a curve and the peak value is defined as $ES_1(ES_2)$. The enrichment score is defined as the weighted sum of scores

$$ES_\beta = \beta ES_1 + (1-\beta)ES_2, \quad 0 < \beta < 1.$$

Step 3: To calculate the significance of enrichment score, we permute diabetes genes in the eMN for 100 times to calculate the null distribution of enrichment scores and convert the ES_β to a z-score statistic based on this null distribution, then a p-value is estimated and adjusted for multiple testing. For each diabetes-disease connection, the significance is defined as the minimum adjusted p-value of eMN. The diseases are then ranked based on their significances, and the more significant the disease, the more diabetes-related.

DIconnectivity

DIconnectivity is the weighted sum of interaction numbers between diabetes and disease gene set, which is based on the idea of cut edge. We can divide the interactions between the two gene sets into four categories: (1)H_1: one gene involved in the interaction is disease/VD gene and the other gene is diabetes gene; (2)H_2: one gene is disease/VD gene, and the other is an overlap gene (both a disease gene and a diabetes gene); (3)H_3: One gene is a diabetes gene, and the other is an overlap gene; (4) H_4: the two genes are both overlap genes. We give the weight of the number of $H_i (i = 1, 2, 3)$ as 1, and the weight of H_4 as 2 (see **Figure 1C**). In addition, we also proposed DIconnectivity-eDMN method, which calculates the weighted sum of interaction numbers between the expand diabetes and disease gene set. The gene sets are expanded based on RWR and GSEA: (1) In Step 2 of DINet, we can obtain the score of each diabetes/disease gene; (2) Sort the diabetes/disease genes in descending order according to their scores; (3) The top n genes are defined as expanded diabetes/disease genes (Hu et al., 2018), and n is N times the original gene size and the maximum does not exceed the number of eMN genes. For each diabetes-disease pair, its DIconnectivity is defined as the mean of interaction numbers across eMNs. The larger the value is, the higher the connection between them.

Key Connector Analysis

We adopted the KDA software package (Zhang and Zhu, 2013) to identify key connectors in PPI network. KDA was originally designed to identify "key regulators" in a directed regulatory network. When applied to undirected networks like PPI networks, we consider the key nodes as "key connectors" since they do not necessarily contain the directional information (Zhang and Zhu, 2013). Such key connectors function more like a "hub" gene, instead of being considered as "master regulators." Specifically, KDA takes a set of genes G and an undirected gene network N as inputs. It has two searching strategies, namely, dynamic neighborhood search (DNS) and static neighborhood search (SNS) for identifying key connectors. We adopted DNS in this study: (1) It first generates a subnet N_G consisting of all nodes in N with no more than $L (L = 2$ in this study) steps away from the nodes in G. (2) For each gene g in N_G, DNS then searches for genes with distances no more than $h = 1, 2, ..., H (H = 2$ in this study) in N_G. The set of genes (not including g) is denoted by $N_G(HLN_{g,h})$. The hypergeometric test is then used to calculate the enrichment between $N_G(HLN_{g,h})$ and G with the genes in N_G as background for each h. The final enrichment p-value of each gene g is calculated as the minimum p-value across h layers. (3) The Bonferroni correction is performed to adjust for multiple

testing and the genes with significant Bonferroni p-values ($\leq$0.05) are outputted as key connectors.

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

SC, HZ, and LZ conceived the concept of the work. LZ, JX, QW, AW, CL, GT, and HZ performed the experiments. LZ wrote the manuscript. All authors approved the final version of this manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (Grant Nos. 11971439 and 61702054), the Training Program for Excellent Young Innovators of Changsha (Grant No. kq1905045), and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 2019zzts279).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.617136/full#supplementary-material
DIconnectivity-eDMN is available at https://github.com/zhulijuan123/connectivity.

REFERENCES

Alhowikan, A. M., Al-Ayadhi, L. Y., and Halepoto, D. M. (2019). Impact of environmental pollution, dietary factors and diabetes mellitus on Autism Spectrum Disorder (ASD). *Pak. J. Med. Sci.* 35, 1179–1184. doi: 10.12669/pjms.35.4.269

Arneth, B., Arneth, R., and Shams, M. (2019). Metabolomics of Type 1 and Type 2 Diabetes. *Int. J. Mol. Sci.* 20:2467. doi: 10.3390/ijms20102467

Basson, M. A., Watson-Johnson, J., Shakya, R., Akbulut, S., Hyink, D., Costantini, F. D., et al. (2006). Branching morphogenesis of the ureteric epithelium during kidney development is coordinated by the opposing functions of GDNF and Sprouty1. *Dev. Biol.* 299, 466–477. doi: 10.1016/j.ydbio.2006.08.051

Bianchi, L., Porta, C., Rinaldi, A., Gazzaruso, C., Fratino, P., DeCata, P., et al. (2017). Integrated cardiovascular/respiratory control in type 1 diabetes evidences functional imbalance: possible role of hypoxia. *Int. J. Cardiol.* 244, 254–259. doi: 10.1016/j.ijcard.2017.06.047

Brezar, V., Carel, J. C., Boitard, C., and Mallone, R. (2011). Beyond the hormone: insulin as an autoimmune target in type 1 diabetes. *Endocr. Rev.* 32, 623–669. doi: 10.1210/er.2011-0010

Bronsveld, H. K., ter Braak, B., Karlstad, Ø, Vestergaard, P., Starup-Linde, J., Bazelier, M. T., et al. (2015). Treatment with insulin (analogues) and breast cancer risk in diabetics; a systematic review and meta-analysis of in vitro, animal and human evidence. *Breast Cancer Res.* 17:100. doi: 10.1186/s13058-015-0611-2

Bruce, I. N., Urowitz, M. B., Gladman, D. D., Ibañez, D., and Steiner, G. (2003). Risk factors for coronary heart disease in women with systemic lupus erythematosus: the Toronto Risk Factor Study. *Arthritis Rheum.* 48, 3159–3167. doi: 10.1002/art.11296

Carter, E., Miron-Buchacra, G., Goldoni, S., Danahay, H., Westwick, J., Watson, M. L., et al. (2014). Phosphoinositide 3-kinase alpha-dependent regulation of branching morphogenesis in murine embryonic lung: evidence for a role in determining morphogenic properties of FGF7. *PLoS One* 9:e113555. doi: 10.1371/journal.pone.0113555

Chen, M., Lu, P., Ma, Q., Cao, Y., Chen, N., Li, W., et al. (2020). CTNNB1/β-catenin dysfunction contributes to adiposity by regulating the cross-talk of mature adipocytes and preadipocytes. *Sci. Adv.* 6:eaax9605. doi: 10.1126/sciadv.aax9605

Chen, Q., and Ma, J. X. (2017). Canonical Wnt signaling in diabetic retinopathy. *Vis. Res.* 139, 47–58. doi: 10.1016/j.visres.2017.02.007

Chen, X., Ayala, I., Shannon, C., Fourcaudot, M., Acharya, N. K., Jenkinson, C. P., et al. (2018). The diabetes gene and Wnt pathway effector TCF7L2 regulates adipocyte development and function. *Diabetes Metab. Res. Rev.* 67, 554–568. doi: 10.2337/db17-0318

Chung, C. P., Avalos, I., Oeser, A., Gebretsadik, T., Shintani, A., Raggi, P., et al. (2007). High prevalence of the metabolic syndrome in patients with systemic lupus erythematosus: association with disease characteristics and cardiovascular risk factors. *Ann. Rheum. Dis.* 66, 208–214. doi: 10.1136/ard.2006.054973

Cukierman, T., Gerstein, H. C., and Williamson, J. D. (2005). Cognitive decline and dementia in diabetes–systematic overview of prospective observational studies. *Diabetologia* 48, 2460–2469. doi: 10.1007/s00125-005-0023-4

Ding, L., Fan, L., Xu, X., Fu, J., and Xue, Y. (2019). Identification of core genes and pathways in type 2 diabetes mellitus by bioinformatics analysis. *Mol. Med. Rep.* 20, 2597–2608. doi: 10.3892/mmr.2019.10522

Domingueti, C. P., Dusse, L. M., Carvalho, Md, de Sousa, L. P., Gomes, K. B., et al. (2016). Diabetes mellitus: the linkage between oxidative stress, inflammation, hypercoagulability and vascular complications. *J. Diabetes Compl.* 30, 738–745. doi: 10.1016/j.jdiacomp.2015.12.018

Echouffo-Tcheugui, J. B., Kolte, D., Khera, S., Aronow, H. D., Abbott, J. D., Bhatt, D. L., et al. (2018). Diabetes Mellitus and cardiogenic shock complicating acute myocardial infarction. *Am. J. Med.* 131, 778–786. doi: 10.1016/j.amjmed.2018.03.004

Ercin, M., Sancar-Bas, S., Bolkent, S., and Gezginci-Oktayoglu, S. (2018). Tub and β-catenin play a key role in insulin and leptin resistance-induced pancreatic beta-cell differentiation. *Biochim. Biophys. Acta Mol. Cell Res.* 1865, 1934–1944. doi: 10.1016/j.bbamcr.2018.09.010

Forbes, J. M., and Cooper, M. E. (2013). Mechanisms of diabetic complications. *Physiol. Rev.* 93, 137–188. doi: 10.1152/physrev.00045.2011

Fourlanos, S., Narendran, P., Byrnes, G. B., Colman, P. G., and Harrison, L. C. (2004). Insulin resistance is a risk factor for progression to type 1 diabetes. *Diabetologia* 47, 1661–1667. doi: 10.1007/s00125-004-1507-3

Gomes, M. B., Giannella Neto, D., Mendonça, Tambascia, M. A., Fonseca, R. M., Réa, R. R., et al. (eds) (2006). Nationwide multicenter study on the prevalence of overweight and obesity in type 2 diabetes mellitus in the Brazilian population. *Arq. Bras. Endocrinol. Metabol.* 50, 136–144. doi: 10.1590/s0004-27302006000100019

Goodwin, K., Mao, S., Guyomar, T., Miller, E., Radisky, D. C., Košmrlj, A., et al. (2019). Smooth muscle differentiation shapes domain branches during mouse lung development. *Development* 146:dev181172. doi: 10.1242/dev.181172

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

Hart, N. J., Aramandla, R., Poffenberger, G., Fayolle, C., Thames, A. H., Bautista, A., et al. (2018). Cystic fibrosis-related diabetes is caused by islet loss and inflammation. *JCI Insight* 3:e98240. doi: 10.1172/jci.insight.98240

Hoffman, R. P. (2017). The complex inter-relationship between diabetes and schizophrenia. *Curr. Diabetes Rev.* 13, 528–532. doi: 10.2174/1573399812666161201205322

Hu, K., Hu, J. B., Tang, L., Xiang, J., Ma, J. L., Gao, Y. Y., et al. (2018). Predicting disease-related genes by path structure and community structure in protein–protein networks. *J. Stat. Mech. Theory and Experiment* 2018:100001. doi: 10.1088/1742-5468/aae02b

International Diabetes Federation (2015). *IDF Diabetes Atlas*, 7th Edn. Brussels: International Diabetes Federation.

Jenkins, A. J., Joglekar, M. V., Hardikar, A. A., Keech, A. C., O'Neal, D. N., and Januszewski, A. S. (2015). Biomarkers in diabetic retinopathy. *Rev. Diabet. Stud.* 12, 159–195. doi: 10.1900/RDS.2015.12.159

Jiang, M. Y., Hwang, J. C., and Feng, I. J. (2018). Impact of diabetes mellitus on the risk of end-stage renal disease in patients with systemic lupus erythematosus. *Sci. Rep.* 8:6008. doi: 10.1038/s41598-018-24529-2

Jiang, Z. S., Jia, H. X., Xing, W. J., Han, C. D., Wang, J., Zhang, Z. J., et al. (2015). Investigation of several biomarkers associated with diabetic nephropathy. *Exp. Clin. Endocrinol. Diabetes* 123, 1–6. doi: 10.1055/s-0034-1385875

Kavanagh, D. H., Savage, D. A., Patterson, C. C., McKnight, A. J., Crean, J. K., Maxwell, A. P., et al. (2011). Association analysis of canonical Wnt signalling genes in diabetic nephropathy. *PLoS One* 6:e23904. doi: 10.1371/journal.pone.0023904

Kessenbrock, K., Smith, P., Steenbeek, S. C., Pervolarakis, N., Kumar, R., Minami, Y., et al. (2017). Diverse regulation of mammary epithelial growth and branching morphogenesis through noncanonical Wnt signaling. *Proc. Natl. Acad. Sci. U.S.A.* 114, 3121–3126. doi: 10.1073/pnas.1701464114

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Lee, S. H., Demeterco, C., Geron, I., Abrahamsson, A., Levine, F., and Itkin-Ansari, P. (2008). Islet specific Wnt activation in human type II diabetes. *Exp. Diabetes Res.* 2008:728763. doi: 10.1155/2008/728763

Li, J. W., Lee, H. M., Wang, Y., Tong, A. H., Yip, K. Y., Tsui, S. K., et al. (2016). Interactome-transcriptome analysis discovers signatures complementary to GWAS Loci of Type 2 Diabetes. *Sci. Rep.* 6:35228. doi: 10.1038/srep35228

Liao, S., Li, J., Wei, W., Wang, L., Zhang, Y., Li, J., et al. (2011). Association between diabetes mellitus and breast cancer risk: a meta-analysis of the literature. *Asian Pac. J. Cancer Prev.* 12, 1061–1065.

Liu, L. B., Chen, X. D., Zhou, X. Y., and Zhu, Q. (2018). The Wnt antagonist and secreted frizzled-related protein 5: implications on lipid metabolism, inflammation, and type 2 diabetes mellitus. *Biosci. Rep.* 38:BSR20180011. doi: 10.1042/BSR20180011

Liu, X., and Li, X. (2019). Key genes involved in diabetic nephropathy investigated by microarray analysis. *J. Comput. Biol.* 26, 1438–1447. doi: 10.1089/cmb.2019.0182

Luo, J., Virnig, B., Hendryx, M., Wen, S., Chelebowski, R., Chen, C., et al. (2014). Diabetes, diabetes treatment and breast cancer prognosis. *Breast Cancer Res. Treat.* 148, 153–162. doi: 10.1007/s10549-014-3146-9

Maconi, G., Furfaro, F., Sciurti, R., Bezzio, C., Ardizzone, S., and de Franchis, R. (2014). Glucose intolerance and diabetes mellitus in ulcerative colitis: pathogenetic and therapeutic implications. *World J. Gastroenterol.* 20, 3507–3515. doi: 10.3748/wjg.v20.i13.3507

Mariosa, D., Kamel, F., Bellocco, R., Ye, W., and Fang, F. (2015). Association between diabetes and amyotrophic lateral sclerosis in Sweden. *Eur. J. Neurol.* 22, 1436–1442. doi: 10.1111/ene.12632

McIntyre, R. S., Konarski, J. Z., Misener, V. L., and Kennedy, S. H. (2005). Bipolar disorder and diabetes mellitus: epidemiology, etiology, and treatment implications. *Ann. Clin. Psychiatry* 17, 83–93. doi: 10.1080/10401230590932380

Murakami, D., Anan, F., Masaki, T., Umeno, Y., Shigenaga, T., Eshima, N., et al. (2019). Visceral fat accumulation is associated with asthma in patients with Type 2 Diabetes. *J. Diabetes Res.* 2019:3129286. doi: 10.1155/2019/3129286

Naslafkih, A., and Sestier, F. (2003). Diabetes mellitus related morbidity, risk of hospitalization and disability. *J. Insur. Med.* 35, 102–113.

Nigro, E., Scudiero, O., Monaco, M. L., Palmieri, A., Mazzarella, G., Costagliola, C., et al. (2014). New insight into adiponectin role in obesity and obesity-related diseases. *Biomed Res. Int.* 2014:658913. doi: 10.1155/2014/658913

Nouwen, A., Nefs, G., Caramlau, I., Connock, M., Winkley, K., Lloyd, C. E., et al. (2011). Prevalence of depression in individuals with impaired glucose metabolism or undiagnosed diabetes: a systematic review and meta-analysis of the European Depression in Diabetes (EDID) Research Consortium. *Diabetes Care* 34, 752–762. doi: 10.2337/dc10-1414

Okada-Iwabu, M., Yamauchi, T., Iwabu, M., Honma, T., Hamagami, K., Matsuda, K., et al. (2013). A small-molecule AdipoR agonist for type 2 diabetes and short life in obesity. *Nature* 503, 493–499. doi: 10.1038/nature12656

Paneni, F., Beckman, J. A., Creager, M. A., and Cosentino, F. (2013). Diabetes and vascular disease: pathophysiology, clinical consequences, and medical therapy: part I. *Eur. Heart J.* 34, 2436–2443. doi: 10.1093/eurheartj/eht149

Papatheodorou, K., Banach, M., Bekiari, E., Rizzo, M., and Edmonds, M. (2018). Complications of diabetes 2017. *J Diabetes Res.* 2018:3086167. doi: 10.1155/2018/3086167

Papatheodorou, K., Papanas, N., Banach, M., Papazoglou, D., and Edmonds, M. (2016). Complications of diabetes 2016. *J. Diabetes Res.* 2016:6989453. doi: 10.1155/2016/6989453

Pearson-Stuttard, J., Blundell, S., Harris, T., Cook, D. G., and Critchley, J. (2016). Diabetes and infection: assessing the association with glycaemic control in population-based studies. *Lancet Diabetes Endocrinol.* 4, 148–158. doi: 10.1016/S2213-8587(15)00379-4

Prentice, B., Hameed, S., Verge, C. F., Ooi, C. Y., Jaffe, A., and Widger, J. (2016). Diagnosing cystic fibrosis-related diabetes: current methods and challenges. *Expert Rev. Respir. Med.* 10, 799–811. doi: 10.1080/17476348.2016.1190646

Qin, Y., Chen, M., Yang, Y., Zhou, X. R., Shao, S. Y., Wang, D. W., et al. (2018). Liraglutide improves hepatic insulin resistance via the canonical Wnt signaling pathway. *Mol. Med. Rep.* 17, 7372–7380. doi: 10.3892/mmr.2018.8737

Rehling, T., Bjørkman, A. D., Andersen, M. B., Ekholm, O., and Molsted, S. (2019). Diabetes is associated with musculoskeletal pain, osteoarthritis, osteoporosis, and rheumatoid arthritis. *J. Diabetes Res.* 2019:6324348. doi: 10.1155/2019/6324348

Schubert, S., and Heesemann, J. (1995). Infections in diabetes mellitus. *Immun. Infekt.* 23, 200–204. doi: 10.1016/0928-8244(95)00069-1

Song, X., Gong, M., Chen, Y., Liu, H., and Zhang, J. (2019). Nine hub genes as the potential indicator for the clinical outcome of diabetic nephropathy. *J. Cell. Physiol.* 234, 1461–1468. doi: 10.1002/jcp.26958

Stene, L. C., and Nafstad, P. (2001). Relation between occurrence of type 1 diabetes and asthma. *Lancet* 357, 607–608. doi: 10.1016/S0140-6736(00)04067-8

Taneja, N., and Priyadarshini. (2018). Mass spectrometric analysis of proteins of L6 skeletal muscle cells under different glucose conditions, and Vitamin D supplementation. *Protein Pept. Lett.* 25, 356–361. doi: 10.2174/0929866525666180406142128

Tettey, P., Simpson, S. Jr., Taylor, B. V., and van der Mei, I. A. (2015). The co-occurrence of multiple sclerosis and type 1 diabetes: shared aetiologic features and clinical implication for MS aetiology. *J. Neurol. Sci.* 348, 126–131. doi: 10.1016/j.jns.2014.11.019

Triantafilou, K., Triantafilou, M., and Dedrick, R. L. (2001). A CD14-independent LPS receptor cluster. *Nat. Immunol.* 2, 338–345. doi: 10.1038/86342

Wada, J., and Makino, H. (2013). Inflammation and the pathogenesis of diabetic nephropathy. *Clin. Sci.* 124, 139–152. doi: 10.1042/CS20120198

Wang, X. Y., Zhang, X. Z., Li, F., and Ji, Q. R. (2019). MiR-128-3p accelerates cardiovascular calcification and insulin resistance through ISL1-dependent Wnt pathway in type 2 diabetes mellitus rats. *J. Cell. Physiol.* 234, 4997–5010. doi: 10.1002/jcp.27300

Weyer, C., Funahashi, T., Tanaka, S., Hotta, K., Matsuzawa, Y., Pratley, R. E., et al. (2001). Hypoadiponectinemia in obesity and type 2 diabetes: close association with insulin resistance and hyperinsulinemia. *J. Clin. Endocr. Metab.* 86, 1930–1935. doi: 10.1210/jcem.86.5.7463

Yahata, T., de Caestecker, M. P., Lechleider, R. J., Andriole, S., Roberts, A. B., Isselbacher, K. J., et al. (2000). The MSG1 non-DNA-binding transactivator binds to the p300/CBP coactivators, enhancing their functional link to the Smad transcription factors. *J. Biol. Chem.* 275, 8825–8834. doi: 10.1074/jbc.275.12.8825

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

Yang, J., Qiu, J., Wang, K., Zhu, L., Fan, J., Zheng, D., et al. (2017). Using molecular functional networks to manifest connections between obesity and obesity-related diseases. *Oncotarget.* 8, 85136–85149. doi: 10.18632/oncotarget.19490

Zhang, B., and Zhu, J. (2013). "Identification of key causal regulators in gene networks," in *Proceedings of the World Congress on Engineering 2013 WCE 2013*, Vol. 2, London.

Zhang, J., Liu, J., and Qin, X. (2018a). Advances in early biomarkers of diabetic nephropathy. *Rev. Assoc. Med. Bras.* 64, 85–92. doi: 10.1590/1806-9282.64.01.85

Zhang, J., Suo, Y., Liu, M., and Xu, X. (2018b). Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein-protein interaction network. *Biochim. Biophys. Acta Mol. Basis Dis.* 1864, 2369–2375. doi: 10.1016/j.bbadis.2017.11.017

Conflict of Interest: AW, CL, and GT were employed by Geneis Beijing Co., Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 January 2021
doi: 10.3389/fgene.2020.628539

Cell Type-Specific Predictive Models Perform Prioritization of Genes and Gene Sets Associated With Autism

Jinting Guan[1,2], Yang Wang[1], Yiping Lin[1], Qingyang Yin[1], Yibo Zhuang[3] and Guoli Ji[1,2,4]*

[1] Department of Automation, Xiamen University, Xiamen, China, [2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China, [3] Xiamen YLZ Yihui Technology Co., Ltd., Xiamen, China, [4] Innovation Center for Cell Signaling Network, Xiamen University, Xiamen, China

Bulk transcriptomic analyses of autism spectrum disorder (ASD) have revealed dysregulated pathways, while the brain cell type-specific molecular pathology of ASD still needs to be studied. Machine learning-based studies can be conducted for ASD, prioritizing high-confidence gene candidates and promoting the design of effective interventions. Using human brain nucleus gene expression of ASD and controls, we construct cell type-specific predictive models for ASD based on individual genes and gene sets, respectively, to screen cell type-specific ASD-associated genes and gene sets. These two kinds of predictive models can predict the diagnosis of a nucleus with known cell type. Then, we construct a multi-label predictive model for predicting the cell type and diagnosis of a nucleus at the same time. Our findings suggest that layer 2/3 and layer 4 excitatory neurons, layer 5/6 cortico-cortical projection neurons, parvalbumin interneurons, and protoplasmic astrocytes are preferentially affected in ASD. The functions of genes with predictive power for ASD are different and the top important genes are distinct across different cells, highlighting the cell-type heterogeneity of ASD. The constructed predictive models can promote the diagnosis of ASD, and the prioritized cell type-specific ASD-associated genes and gene sets may be used as potential biomarkers of ASD.

Keywords: autism spectrum disorder, cell type-specific, predictive model, gene set, biomarker

OPEN ACCESS

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

Reviewed by:
Jiebiao Wang,
University of Pittsburgh, United States
Qian Du,
University of Nebraska-Lincoln, United States

****Correspondence:***
Jinting Guan
jtguan@xmu.edu.cn

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

Received: *12 November 2020*
Accepted: *16 December 2020*
Published: *15 January 2021*

Citation:
Guan J, Wang Y, Lin Y, Yin Q, Zhuang Y and Ji G (2021) Cell Type-Specific Predictive Models Perform Prioritization of Genes and Gene Sets Associated With Autism. Front. Genet. 11:628539. doi: 10.3389/fgene.2020.628539

INTRODUCTION

Autism spectrum disorder (ASD) represents a group of neurodevelopmental disorders, characterized by substantial phenotypic and genetic heterogeneity. Genetic studies have identified variants that contribute to the risk of developing ASD (Iossifov et al., 2012; Neale et al., 2012; O'Roak et al., 2012; Sanders et al., 2012; De Rubeis et al., 2014; Gaugler et al., 2014; Turner et al., 2016; Satterstrom et al., 2020). However, it remains perplexing how these reported variants lead to the pathogenesis of ASD. A major mode of action is that these genetic variants cause gene expression alternations; direct analysis of gene expression in disease-relevant tissue is thus valuable for understanding the molecular mechanism of ASD. As ASD is believed to result from functional aberrations within brains, bulk transcriptomic analyses between autistic and normal brains have been applied for identifying aberrant gene expression patterns in ASD (Voineagu et al., 2011; Gupta et al., 2014; Guan et al., 2016; Parikshak et al., 2016). However, the brain is a highly heterogeneous

organ including different cell types that are highly interconnected. Genes may demonstrate diverse functions across different brain cell types. In ASD, different functions may be dysregulated and causal genes may be distinct across different cells. Although bulk transcriptomic studies revealed convergence of disease pathology on common pathways, the brain cell type-specific molecular pathology of ASD is still needed to study.

Recently, the newly available single-nucleus RNA-sequencing data of ASD (Velmeshev et al., 2019) makes it possible to study the cell-type heterogeneity of ASD directly. The authors identified differentially expressed (DE) genes between ASD and control groups in a cell type-specific way and analyzed the functions of the cell type-specific DE genes to characterize the heterogeneity of dysregulated gene expression patterns among brain cell types in ASD. As genes interact with others, the integrity of disease gene modules instead of individual genes may determine the manifestation of a disease in cells (Kitsak et al., 2016; Mohammadi et al., 2019). Therefore, in addition to identifying the individual cell type-specific risk genes, it is essential to identify cell type-specific gene sets/modules associated with ASD.

There have been more and more studies evaluating the effectiveness of machine learning for diagnosing ASD, exploring its genetic underpinnings, and designing effective interventions (Hyde et al., 2019). These studies were based on different kinds of datasets, such as behavior evaluation based on Autism Diagnostic Observation Schedule (ADOS) (Duda et al., 2014; Levy et al., 2017) and Autism Diagnostic Interview-Revised (ADI-R) (Wall et al., 2012; Duda et al., 2016), brain images for magnetic resonance image (MRI) (Chen et al., 2011; Heinsfeld et al., 2018) and electroencephalogram (EEG) (Bosl et al., 2018), and genetic profiles (Kong et al., 2012; Cogill and Wang, 2016; Guan et al., 2016; Oh et al., 2017). To detect ASD candidate genes, several predictive models were constructed based on gene expression profiling, including the one built using DE genes between ASD and controls based on gene expression microarrays of blood (Kong et al., 2012) and the one built using aberrant gene expression in ASD based on bulk transcriptomic data of brains (Guan et al., 2016). Actually, for identifying ASD risk genes, genetic and genomic studies were usually performed, such as genome-wide association studies, copy number variation studies, and whole exome sequencing; these methods are expensive and time-consuming, and the generated potential candidate genes are numerous and not easy to be validated (Cogill and Wang, 2016). Gene screening methods based on machine learning can prioritize genes and identify high-confidence candidates, which may provide new insights for the experimental studies.

In this study, to characterize the cell-type heterogeneity of ASD and to take advantage of the potential of gene expression signature being diagnostic biomarkers for ASD, we analyze the human brain nucleus gene expression data of ASD and controls published in Velmeshev et al. (2019) and construct multiple kinds of classification models for ASD using the algorithm of partial least squares (PLS), identifying cell type-specific genes and gene sets associated with ASD. Firstly, we construct cell type-specific predictive models based on individual genes to screen cell type-specific genes associated with ASD. Then, we construct cell type-specific gene set-based predictive models to screen cell type-specific gene sets associated with ASD. These two kinds of predictive models can be applied to predict the diagnosis of a given nucleus with known cell type. Lastly, we further construct a multi-label predictive model for predicting the cell type and diagnosis of a given nucleus at the same time. Our results suggest that it may be feasible to use brain cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our analytical pipeline prioritizes ASD-associated cell type-specific genes and gene sets, highlighting the cell-type heterogeneity of ASD.

MATERIALS AND METHODS

Human Brain Nucleus Gene Expression Data

We used the single-nucleus RNA-seq data published in Velmeshev et al. (2019), which includes 104,559 nuclei from 41 post-mortem tissue samples from the prefrontal cortex and anterior cingulate cortex of 15 ASD patients and 16 control subjects. The nuclei were divided into 17 cell types, including fibrous astrocytes (AST-FB), protoplasmic astrocytes (AST-PP), endothelial, parvalbumin interneurons (IN-PV), somatostatin interneurons (IN-SST), SV2C interneurons (IN-SV2C), VIP interneurons (IN-VIP), layer 2/3 excitatory neurons (L2/3), layer 4 excitatory neurons (L4), layer 5/6 corticofugal projection neurons (L5/6), layer 5/6 cortico-cortical projection neurons (L5/6-CC), microglia, maturing neurons (Neu-mat), NRGN-expressing neurons I (Neu-NRGN-I), NRGN-expressing neurons II (Neu-NRGN-II), oligodendrocytes, and OPC. We downloaded the matrices of raw counts from the website of autism.cells.ucsc.edu. Then, we preprocessed the data with R package of scran (Lun et al., 2016), including the quality control of nuclei and genes, removing a minority of nuclei from different cell cycle phases, and normalizing the gene expression data. Next, nuclear and mitochondrial genes downloaded from Human MitoCarta2.0 (Calvo et al., 2016) were excluded. We used the function of plotExplanatoryVariables in scran to check if any factors, including region, age, sex, PMI (post-mortem interval), RIN (RNA integrity number), Capbatch (10X capture batch), and Seqbatch (sequencing batch), may contribute to the heterogeneity of gene expression. It can calculate the percentage of the variance of the expression values that is explained by the factors for each gene. By checking the distribution of percentages across all genes, we found that the expression profiles of most genes are not strongly associated with the factors and the factors thus do not need to be explicitly modeled in the downstream analyses (Lun et al., 2016). We applied scran to obtain highly variable genes, which include a total of 12,036 genes. We used the expression level of 12,036 genes for downstream analyses, which contains 85,125 nuclei, including 3655, 7085, 1991, 3719, 4190, 1836, 5621, 12,795, 6518, 3402, 4385, 2495, 3532, 589, 1459, 12206, and 9647 nuclei from cell types of AST-FB, AST-PP, endothelial, IN-PV, IN-SST, IN-SV2C, IN-VIP, L2/3, L4, L5/6, L5/6-CC, microglia, Neu-mat, Neu-NRGN-I, Neu-NRGN-II, oligodendrocytes, and OPC, respectively.

Annotated Gene Sets

A total of 913 ASD candidate genes were downloaded from Simons Foundation Autism Research Initiative (SFARI) (release of March 4, 2020), which include 119, 144, 219, and 472 genes from categories of S (syndromic), 1 (high confidence), 2 (strong candidate), and 3 (suggestive evidence). For gene set analysis, three kinds of annotated gene sets from Molecular Signatures Database (MSigDB) (Liberzon et al., 2011) were used, including H: hallmark gene sets, C2: curated gene sets (containing gene sets from chemical and genetic perturbations, and canonical pathways of Biocarta, KEGG, PID, and Reactome), and C5: GO gene sets. By intersecting the genes in gene sets and our analyzed gene expression matrix, we kept 3741 gene sets containing more than 30 overlapping genes.

The Algorithm of Partial Least Squares

Partial least squares (Wold, 1966) regression combines features from principal component analysis and multiple regression. It has the ability to address the problem of modeling multicollinearity, noisy, and even incomplete highly dimensional data (Boulesteix and Strimmer, 2006). PLS can solve both single- and multi-label classification problems. Partial least squares discriminant analysis (PLS-DA) is a PLS regression, with the dependent variable being categorical. Suppose X is an $n \times m$ matrix containing n observations of m genes and Y is an $n \times p$ matrix containing n observations of p response variables, then X and Y can be decomposed by:

$$X = TP^T + E, \qquad Y = UQ^T + F$$

where T and U are $n \times k$ score matrices (called component scores or latent variables) of X and Y, respectively, P and Q are $m \times k$ and $p \times k$ orthogonal loading matrices, and E and F are the residual matrices. The decompositions of X and Y are made so as to maximize the covariance between T and U. Then, based on T, P, U, and Q, we can first fit U and T, and then the linear relationship between X and Y can be obtained.

Recursive Feature Elimination With Cross-Validation

Recursive feature elimination (RFE) (Guyon et al., 2002) is a backward feature selection method, which is a recursive process. It first builds a model using all features based on an algorithm specified, such as PLS in our study, and computes a measure of importance for each feature. The least important features are removed. Then, the model is re-built using the left features, importance scores are computed, and the least important features are removed until the specified number of features is reached. RFE attempts to eliminate dependencies and collinearity that may exist in the model. It requires a specified number of features to keep. To find the optimal number of features, RFE with cross-validation (RFECV) is usually used to score feature subsets of different sizes and select the best scoring one. Then, the optimal feature subset is used to build the final model.

The Construction of Predictive Models

The R package of caret (Kuhn, 2008) was adopted to construct predictive models based on the algorithm of PLS. Firstly, for each cell type, we extracted the gene expression data of nuclei from the cell type and constructed a cell type-specific predictive model. Secondly, for each cell type and each annotated gene set, we extracted the expression data of nuclei from the cell type in the genes included in the gene set and constructed a cell type-specific gene set-based predictive model. These two kinds of predictive models can predict the diagnosis of a nucleus with known cell type. Specifically, we split the extracted gene expression data into a training set and a test set at a ratio of 7:3 using stratified sampling. For the training set, we selected the optimal model by applying 10-fold cross-validation for 10 times and tuning over the model hyperparameter (the number of PLS components) with grid search from 1 to 15 with a step of 1. To evaluate the model performance, the area under the receiver operating characteristic (ROC) curve (denoted as AUC) was used, because this metric can deal well with the problem of label imbalance and not be influenced by the selection of threshold. Then, from the optimal model, we obtained the predictive probability of each nucleus being a nucleus from ASD patients. Next, we used R package of pROC (Robin et al., 2011) to obtain the best threshold on training set and the threshold was used to determine the predictive performances on training set and test set. For each predictive model, we calculated the importance of each gene using the function of *varImp* in caret.

In order to predict the cell type and diagnosis of a given nucleus at the same time, we constructed a multi-label predictive model based on PLS using R package of mlr (Bischl et al., 2016). For each nucleus, we used 18 labels to describe it, with 1 label being the diagnosis and the other 17 cell-type labels obtained using one-hot encoding. We split the whole gene expression data including all cell types and all genes into a training set and a test set at a ratio of 7:3 using stratified sampling. Based on the training set, we selected the optimal model by applying five-fold cross-validation for five times and tuning over the model hyperparameter with grid search from 1 to 15 with a step of 1. Hamming loss, which is the fraction of labels that are predicted incorrectly to the total number of labels, was used as a performance indicator. Then, from the optimal model, we obtained the predictive probability of each nucleus belonging to each label. For the labels of cell types, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cut-off on training set for determining the predictive diagnosis of each nucleus in training and test sets.

RESULTS

Methodological Overview

After normalization, we used the function of plotExplanatoryVariables in scran (Lun et al., 2016) to calculate the percentage of the variance of the expression values that is explained by factors, including region, age, sex, PMI, RIN,

Capbatch (10X capture batch), and Seqbatch (sequencing batch), for each gene (**Supplementary Figure 1A**). We found that the expression profiles of most genes are not strongly associated with the factors and the factors thus do not need to be explicitly modeled in the downstream analyses. Then, we obtained highly variable genes, a total of 12,036 genes, and used their expression level for downstream analyses. The density plot of the percentage of variance explained by each factor across highly variable genes can be seen in **Supplementary Figure 1B**.

Then we constructed multiple kinds of predictive models for ASD. The overview of our analytical method can be seen in **Figure 1**. Firstly, to screen genes associated with ASD in each cell type, we constructed cell type-specific predictive models, which can predict the diagnosis of a nucleus whose cell type is known, using the algorithm of PLS (see section "Materials and Methods"). Specifically, for each cell type, we extracted the gene expression data of the nuclei from the cell type and split the data into training and test sets. We selected the optimal model based on the training set, and then obtained the predictive probability of each nucleus being a nucleus from ASD patients. Next, ROC analysis was performed to obtain the best threshold on training set, and the threshold was used to determine the predictive performance on training and test sets. To prioritize genes, we calculated the importance of each gene in the cell type-specific predictive model. In addition, in order to use less genes to achieve similar performances, we performed RFECV (see section "Materials and Methods") to reduce the number of genes used to re-construct cell type-specific predictive models. The optimal genes obtained using RFECV were denoted as RFE genes, which were used for the downstream analyses to depict the cell-type heterogeneity of ASD.

Secondly, to screen gene sets associated with ASD in each cell type, we constructed cell type-specific gene set-based predictive models using PLS. Specifically, for each cell type and each gene set, we extracted the expression level of the nuclei from the cell type in the genes included in the considered gene set and constructed a predictive model. To prioritize gene sets, we ranked gene sets using their predictive performance on the test set and kept the gene sets whose predictive accuracy (ACC), sensitivity (SN), and specificity (SP) are larger than 70% as cell type-specific gene sets associated with ASD. Besides, for the total genes included in these identified gene sets, we calculated their frequency and averaged importance, and used the genes with top averaged importance to re-construct cell type-specific predictive models.

Lastly, we further constructed a multi-label predictive model using PLS, which can predict the cell type and the diagnosis of a given nuclei at the same time. For the labels of cell types, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cut-off for determining the predictive diagnosis of each nucleus in training and test sets.

Cell Type-Specific Genes Associated With ASD

For each of the 17 cell types, we first constructed a cell type-specific predictive model using all genes (**Table 1** and **Supplementary File 1**). To score genes in each cell type, we calculated the importance of genes and ranked the genes (**Supplementary File 2**). Next, in order to use less genes to achieve similar performances, we used the genes with top 500, 1000, and 1500 importance respectively to construct cell type-specific predictive models. We found out that using top 1000 genes made the model performance better than the one using top 500, while approaching the one using top 1500 genes (**Supplementary File 1**). Therefore, for each cell type, we applied RFECV to reduce the number of genes to up to 1000 and obtain the optimal gene subset, which was then used to re-construct

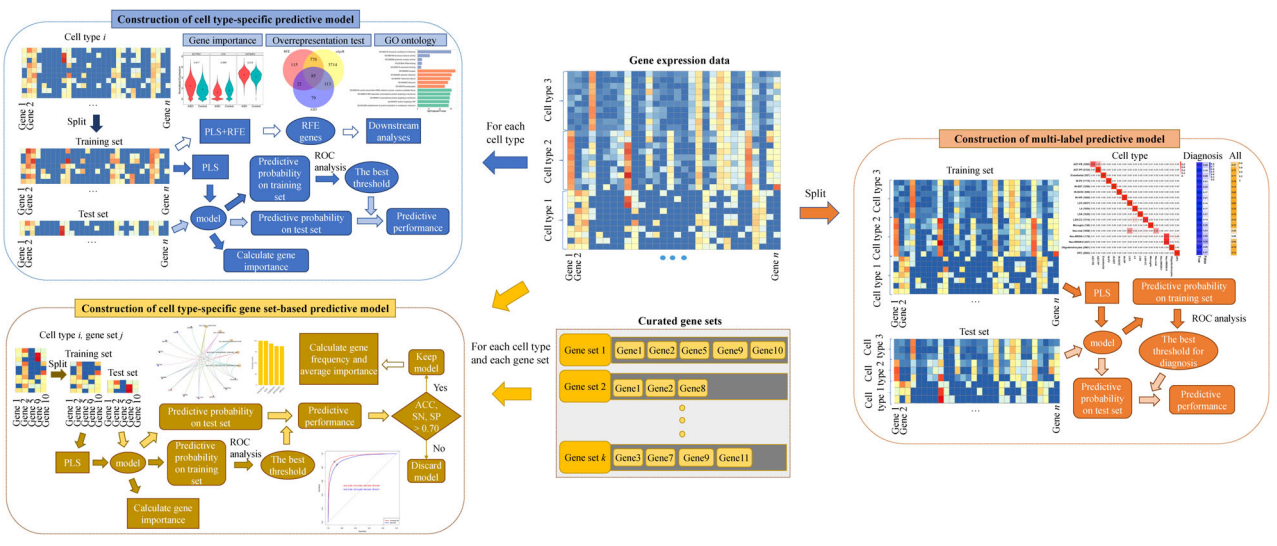

FIGURE 1 | The *methodological overview*.

TABLE 1 | The classification performances of cell type-specific predictive models built using all genes.

Cell type (ASD/control)	Training set				Test set			
	ACC	SN	SP	AUC	ACC	SN	SP	AUC
AST-FB (2033/1622)	0.91	0.92	0.9	0.97	0.72	0.78	0.63	0.79
AST-PP (4749/2336)	0.93	0.93	0.93	0.98	0.84	0.87	0.79	0.90
Endothelial (850/1141)	0.92	0.91	0.92	0.97	0.76	0.70	0.80	0.83
IN-PV (1811/1908)	0.95	0.94	0.96	0.99	0.80	0.77	0.82	0.88
IN-SST (1945/2245)	0.94	0.92	0.95	0.98	0.76	0.70	0.81	0.83
IN-SV2C (990/846)	0.98	0.98	0.97	1.00	0.80	0.83	0.76	0.88
IN-VIP (3098/2523)	0.89	0.88	0.91	0.96	0.79	0.79	0.78	0.86
L2/3 (6962/5833)	0.95	0.95	0.95	0.99	0.89	0.90	0.88	0.96
L4 (3415/3103)	0.93	0.91	0.94	0.98	0.83	0.80	0.87	0.91
L5/6 (1710/1692)	0.93	0.93	0.93	0.98	0.78	0.77	0.80	0.86
L5/6-CC (2279/2106)	0.97	0.98	0.97	1.00	0.85	0.88	0.82	0.93
Microglia (1174/1321)	0.91	0.90	0.93	0.97	0.76	0.73	0.78	0.84
Neu-mat (1853/1679)	0.85	0.82	0.88	0.93	0.75	0.70	0.80	0.83
Neu-NRGN-I (321/268)	0.97	0.99	0.94	0.99	0.69	0.75	0.63	0.74
Neu-NRGN-II (828/631)	0.82	0.86	0.78	0.89	0.63	0.70	0.53	0.68
Oligodendrocytes (4587/7619)	0.83	0.86	0.81	0.91	0.77	0.79	0.75	0.85
OPC (5085/4562)	0.83	0.82	0.84	0.91	0.75	0.74	0.76	0.82

The number of nuclei from ASD and controls are listed. ROC analysis was applied to obtain the AUC and the optimal cut-off point on the training set, and then the optimal cut-off was used to determine the predictive accuracy (ACC), sensitivity (SN), and specificity (SP) on the training and test sets.

a cell type-specific predictive model (see section "Materials and Methods"). The R package of caret (Kuhn, 2008) was adopted to perform PLS-RFE with 10-fold cross-validation for 10 times. The sizes of evaluated gene subsets are from 100 to 1000 with a step of 100. The optimal genes obtained using RFECV were denoted as RFE genes. It is noted that the performances on test sets of the cell type-specific predictive models based on RFE genes approach the ones based on all genes (**Figure 2A** and **Supplementary File 1**); hence, we used the RFE genes for the subsequent analyses in this section.

By examining the number of RFE genes in every cell type (**Table 2**), we found that in several cell types, such as AST-PP, IN-PV, L2/3, L4, and L5/6-CC, there are more RFE genes and the corresponding cell type-specific predictive models have better performances than other cell types (**Figure 2A**). This implies that these cell types may be more vulnerable in ASD and more genes may be dysregulated in these cell types. Then, for each cell type, we also applied edgeR (Robinson et al., 2010) to identify DE genes in ASD compared to controls. It can be seen that in the mentioned cell types above, there are indeed more DE genes, which also indicates that these cell types may be mainly affected by ASD. By performing hypergeometric tests, we found that the RFE genes are significantly overlapped with the DE genes identified by edgeR (**Table 2**). Then, we checked if building cell type-specific predictive models using edgeR genes would be better than the ones using RFE genes, while the model performances using RFE genes are better than the ones using edgeR genes (**Supplementary File 1**). This shows that genes that are not identified by edgeR may have predictive power for ASD. In addition, we also compared the RFE genes with the DE genes identified in the single-nucleus RNA-seq study of ASD (Velmeshev et al., 2019). We found that RFE genes are significantly overlapped with Velmeshev's genes, especially for the cell types of microglia, L2/3, L4, and IN-VIP (**Table 2**). The model performances using RFE genes are significantly better than the ones using Velmeshev's genes (**Supplementary File 1**), which may be because the number of Velmeshev's genes is small. Next, we found that there are more SFARI ASD genes overlapped with RFE genes in neuron-related cell types. We also performed overrepresentation tests between RFE genes and SFARI ASD genes, and found that RFE genes are significantly overlapped with ASD genes (**Table 2**).

For each cell type-specific predictive model built based on RFE genes, we calculated the importance of each RFE gene (**Supplementary File 3**). **Table 2** lists the top RFE genes in each cell type. **Figure 2B** also demonstrates the expression of the top three RFE genes in ASD and control groups for the representative cell types, including AST-PP, endothelial, IN-PV, L2/3, microglia, oligodendrocytes, and OPC. The top genes among different cell types are distinct, implying the cell-type heterogeneity of ASD. However, some top genes appearing in several cell types are of note. For instance, gene *BCYRN1* (brain cytoplasmic RNA 1, a long non-coding RNA) has the largest importance in all excitatory neurons, including L2/3, L4, L5/6, and L5/6-CC. Gene *BCYRN1* is involved in the regulation of synaptogenesis, and there have been several literatures linking *BCYRN1* and Alzheimer's disease, a neurological disease (Wan et al., 2017; Hu et al., 2018), which implies the possible association between *BCYRN1* and ASD. Besides, *BCYRN1* has been prioritized in a blood-based gene expression study of ASD (Ivanov et al., 2015).

To further characterize the cell-type heterogeneity of ASD, we compared the RFE genes across different cells. We performed gene ontology analyses using clusterProfiler (Yu et al., 2012), with background genes set as the genes in the analyzed gene expression matrix. The functions of cell type-specific RFE

TABLE 2 | The overrepresentation tests between RFE genes and differentially expressed genes identified by edgeR, differentially expressed genes identified in the study of Velmeshev et al. (2019), and SFARI ASD genes.

Cell type	Number of RFE genes	Overlapping genes/edgeR genes (FDR-adjusted *P*-value)	Overlapping genes/ASD genes (FDR-adjusted *P*-value)	Overlapping genes/Velmeshev's genes (FDR-adjusted *P*-value)	Top five important genes
AST-FB	200	120/257 (1.5e−158)	22/299 (5.8e−09)	8/11 (1.4e−12)	***DPP10*, *TMSB4X*, *SPARCL1*, *ZFP36L1*, *PCDH9***
AST-PP	1000	667/1464 (0.0e+00)	98/299 (2.2e−34)	33/36 (2.1e−32)	****PTGDS*, *HSPA1A*, *TRPM3*, *RP11-179A16.1*, **CIRBP***
Endothelial	500	115/146 (3.2e−134)	40/299 (6.3e−11)	29/38 (1.5e−32)	***HERC2P3*, **AKAP12*, *TMSB4X*,** *RP11-649A16.1*, ***RPS28***
IN-PV	1000	384/695 (4.2e−251)	103/299 (2.7e−38)	14/14 (1.3e−15)	***AC105402.4*, *MTATP6P1*, *CNTNAP3*, **CIRBP*,** *ARL17B*
IN-SST	1000	549/1346 (5.9e−291)	104/299 (5.3e−39)	16/17 (1.5e−16)	***SST*, *AC105402.4*, *VGF*, *HSPA1A*, *BCYRN1***
IN-SV2C	900	345/616 (7.4e−244)	100/299 (9.7e−40)	9/9 (1.1e−10)	***CCK*, *BCYRN1*, *AC105402.4*, *MEG3*, *HSPB1***
IN-VIP	1000	676/1820 (0.0e+00)	104/299 (5.3e−39)	32/32 (7.1e−35)	***HSPA1A*, *CCK*, **RPS15*, *MEG3*, **RGS12***
L2/3	1000	863/4690 (7.8e−230)	107/299 (2.9e−41)	41/41 (2.0e−44)	***BCYRN1*, *CCK*, **CNTNAP2*, *MEG3*, **CAMK2N1***
L4	1000	715/2477 (1.3e−294)	113/299 (3.7e−46)	40/42 (1.2e−40)	***BCYRN1*, *CCK*, **NCAM2*, *SLC17A7*, *MTATP6P1***
L5/6	900	467/1069 (4.0e−281)	98/299 (2.7e−38)	5/5 (2.8e−06)	***BCYRN1*, *AC105402.4*, *MTATP6P1*, *ATP1B1*, *SLC17A7***
L5/6-CC	1000	701/3183 (7.1e−202)	114/299 (7.5e−47)	7/7 (3.8e−08)	***BCYRN1*, *CCK*, *AC105402.4*,** *RP11-750B16.1*, ***MT-RNR2***
Microglia	200	74/106 (4.9e−112)	20/299 (1.4e−07)	38/49 (2.5e−58)	***FKBP5*, *TMSB4X*, *NEAT1*, *SLC1A3*, *CHN2***
Neu-mat	900	351/476 (1.7e−312)	116/299 (2.9e−53)	1/1 (7.5e−02)	***AC105402.4*,** *XIST,* ***CAMK2N1*, *MEG3*, *ROBO2***
Neu-NRGN-I	100	2/2 (6.8e−05)	12/299 (7.0e−06)	4/6 (8.7e−08)	*RP11-750B16.1,* ******PTMA*, *NRGN*, *GNAO1*,** *TSPAN7*
Neu-NRGN-II	100	7/8 (1.9e−14)	6/299 (3.8e−02)	2/4 (4.4e−04)	***PRNP*, *NRGN*, *STMN1*, *RP11-750B16.1*** *PLP1*
Oligodendrocytes	600	410/1420 (9.4e−253)	57/299 (9.5e−19)	14/14 (1.2e−18)	****PTGDS*, *NRXN1*, *CNDP1*, **ABCA2*, *CREB5***
OPC	900	528/1413 (1.9e−285)	102/299 (2.9e−41)	3/3 (4.4e−04)	***GPC5*, *TMSB4X*, *HSPH1*, **CNTNAP2*, **OLIG1***

*The number of overlapping genes; the number of edgeR genes, Velmeshev's genes, and ASD genes; and the FDR-adjusted hypergeometric test P-values are shown. The genes with top five importance are listed, of which edgeR genes are in boldface, SFARI ASD genes are underlined, and Velmeshev's genes are marked with *.*

genes are different among different cell types (**Supplementary File 4**). For instance, in IN-PV, the enriched GO terms include neuron projection, axon, somatodendritic compartment, and cell part morphogenesis, while in L2/3, the top GO terms are associated with ribosome, cotranslational protein targeting to membrane, and protein localization to endoplasmic reticulum (**Figure 2B**).

Cell Type-Specific Gene Sets Associated With ASD

In addition to screening individual genes associated with ASD, we also constructed cell type-specific gene set-based predictive models to screen ASD-related gene sets. For each cell type and each gene set, we extracted the expression level of the nuclei from the cell type in the genes included in the considered gene set, and constructed a predictive model (see section "Materials and Methods"). We retained the gene sets whose ACC, SN, and SP on test set are larger than 70%, and there are 5, 1, 88, 15, and 137 gene sets identified in cell types of AST-PP, IN-PV, L2/3, L4, and L5/6-CC, respectively (**Supplementary File 5**). **Figure 3A** shows the top five gene sets in each of these five cell types and the performances of corresponding cell type-specific gene set-based predictive models. For AST-PP, the top ASD-associated gene sets include *REACTOME_DISEASE, GO_REGULATION_OF_CELL_POPULATION_PROLIFERATION, GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY, GO_SIGNALING_RECEPTOR_BINDING*, and *GO_ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY*. For other neuron cell types, the ASD-associated gene sets are mostly related to cell junction, synapse, neuron projection, neurogenesis, neuron differentiation, and cell projection organization. By checking the top important genes in each cell type-specific gene set, we found that several genes appear in the majority of the gene sets; for example, gene *HSPA1A* [heat shock protein family A (*HSP70*) member 1A] shows up in all AST-PP specific ASD-associated gene sets (**Figure 3B**). Therefore, for each cell type, we analyzed the frequency of each gene included in the identified gene sets and calculated the averaged importance of genes (**Supplementary File 5**). **Figure 3C** shows the genes with top five averaged importance in each cell type. Gene *HSPA1A* is noted in AST-PP. Actually, heat shock proteins play a central role in the development of neurological disorders, of which *HSP70* family has been shown its functions (Turturici et al., 2011), and *HSPA1A*, a member of *HSP70* family, has already been associated with ASD (Lin et al., 2014). As to gene *CCK* (cholecystokinin), it is prioritized in excitatory neurons, which is a kind of gut peptide hormone. Gut peptide hormones have been found across different brain regions, and many of them are involved with ASD-related deficits (Qi and Zhang, 2020).

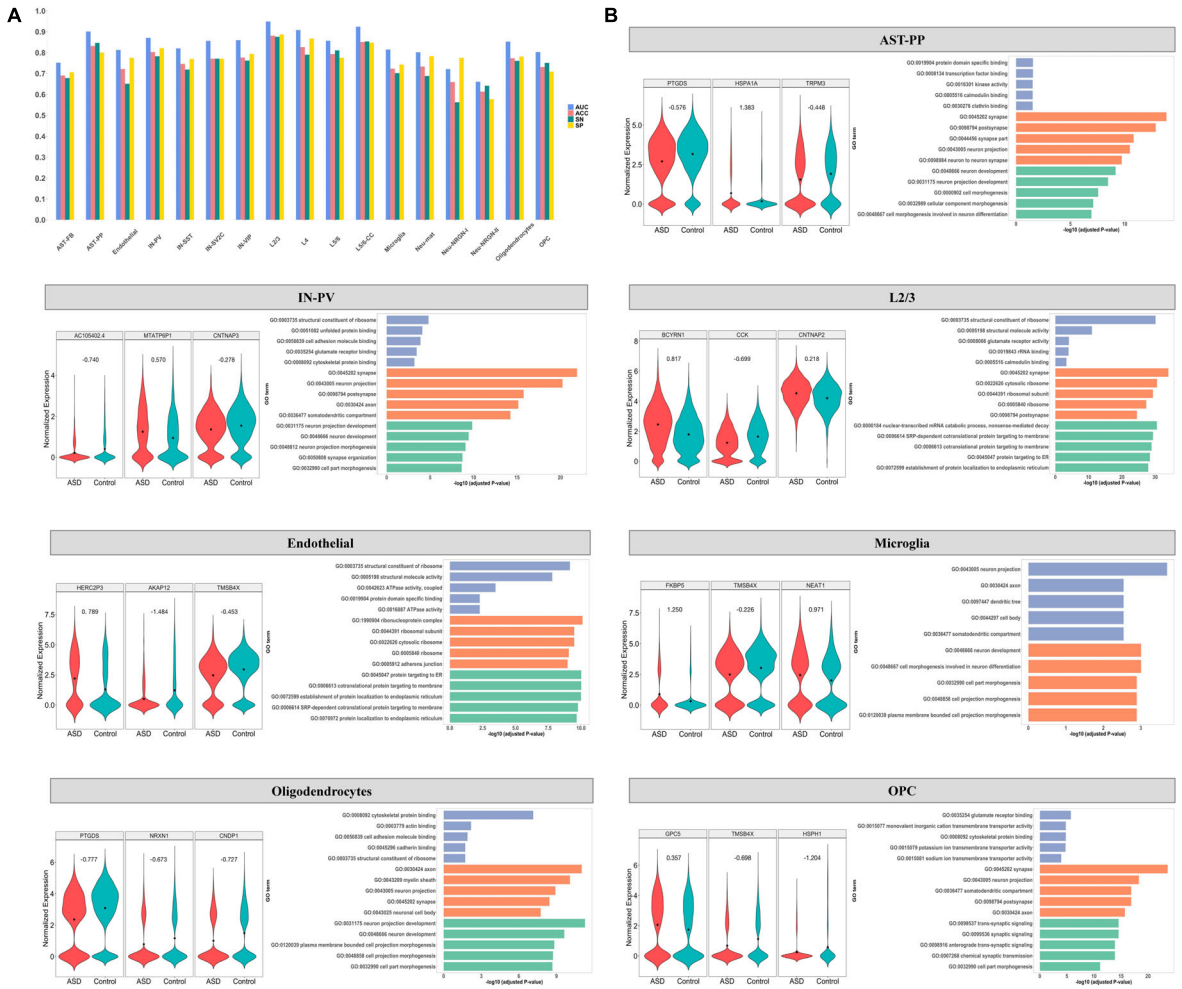

FIGURE 2 | (A) The classification performance on test set of cell type-specific predictive models built using RFE genes. ROC analysis was applied to obtain the AUC and the optimal cut-off point on the training set, and then the optimal cut-off was used to determine the predictive accuracy (ACC), sensitivity (SN), and specificity (SP) on the test set. For the cell types of AST-PP, endothelial, IN-PV, L2/3, microglia, oligodendrocytes, and OPC, **(B)** the expression of the top three important genes in ASD and control groups is shown along with the top enriched GO terms with the RFE genes. The GO terms belonging to molecular functions, cellular component, and biological process are shown in blue, orange, and green, respectively.

Next, based on the genes with averaged importance >10% in corresponding cell types, we re-constructed a cell type-specific predictive model for each of these five cell types. It is noted that their predictive performances are even better than the ones of the cell type-specific gene set-based predictive models (**Figure 3D**). We checked the functions of these genes (**Supplementary File 6**) and found that their functions are distinct, especially among AST-PP, IN-PV, and excitatory neurons (**Figure 3E**). In AST-PP, the top genes are associated with the functions of enzyme-linked receptor protein signaling pathway, transmembrane receptor protein tyrosine kinase signaling pathway, positive regulation of phosphorus and phosphate metabolic process, and cellular component morphogenesis. In IN-PV, the top genes are related to synaptic and postsynaptic membrane, cation channel complex, and neuron projection. As to the cell types of excitatory neurons, the top genes are associated with ribosome, SRP-dependent cotranslational protein targeting to membrane, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, and protein targeting to ER.

A Multi-Label Classification Model Predicting Cell Type and Diagnosis

To predict the cell type and diagnosis of a given nucleus at the same time, we applied PLS to construct a multi-label predictive model (see section "Materials and Methods"). We split the whole gene expression data to a training set and a test set. For the diagnosis label, we extracted the predictive probability of training set and applied ROC analysis to obtain the optimal cut-off for determining the predictive diagnosis of each nucleus in training and test sets. For the cell type labels, the predictive cell type of each nucleus was set as the cell type whose predictive probability is the largest. The Hamming loss of the multi-label predictive model is 0.02, and the accuracy achieves 72.8 with 92.7% accuracy

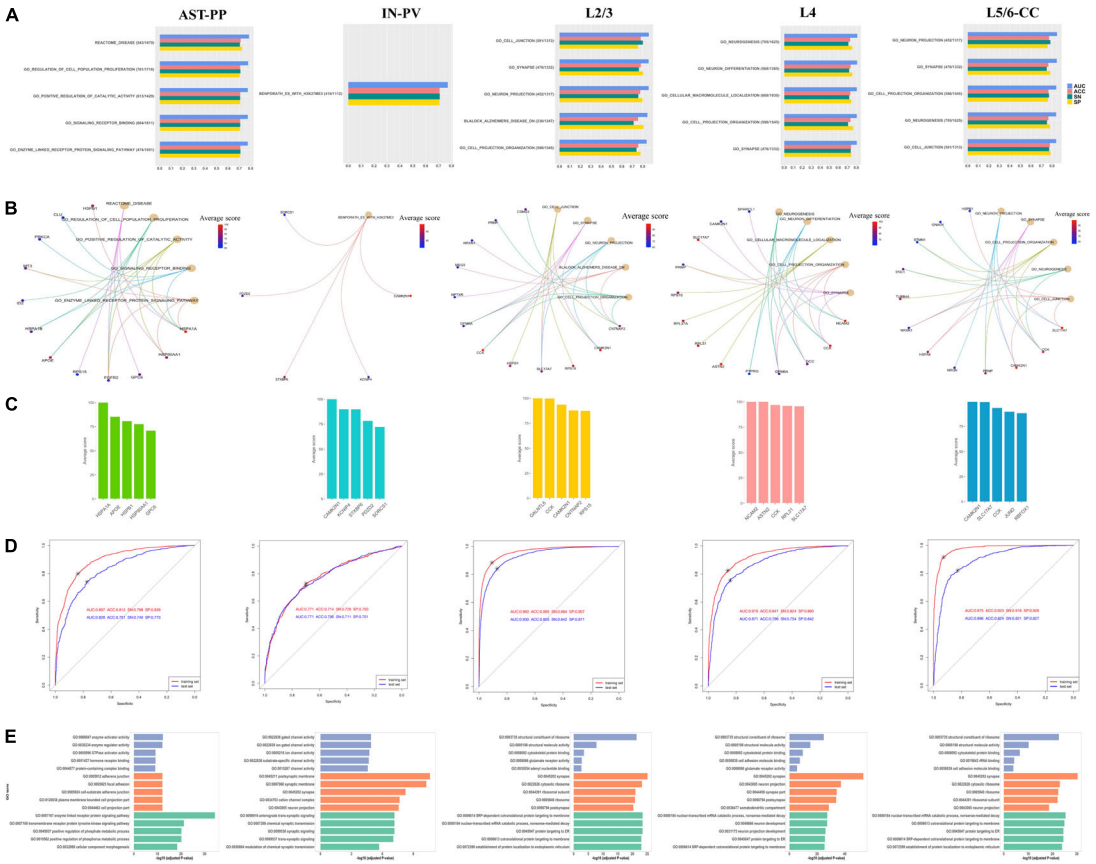

FIGURE 3 | (A) The identified top five gene sets associated with ASD by constructing cell type-specific gene set-based predictive models. The number of overlapping genes between the gene expression data and the gene set, the total number of genes in the gene set, and the performances of corresponding cell type-specific gene set-based predictive models are shown. For each cell type, **(B)** illustrates the top five gene sets and the genes with top five importance in each gene set, and **(C)** plots the genes with top averaged importance. **(D)** The performances of predictive models built using genes with averaged importance >10% and **(E)** the enriched GO terms with these genes. The GO terms belonging to molecular functions, cellular component, and biological process are shown in blue, orange, and green, respectively.

for cell-type labels and 78.5% accuracy for diagnosis label. Then, we examined the predictive performance of the model by cell type and by label. For each cell type, **Figure 4** illustrates the proportion of the number of nuclei predicted as each cell type to the total number of nuclei, the proportion of correct and incorrect predictions for the label of diagnosis, and the proportion of correct predictions for all labels in the test set. It can be seen that for most cell types, the predictive cell types are correct, except for AST-FB, Neu-mat, and Neu-NRGN-I. Because AST-FB and AST-PP are cell clusters of astrocytes and they may have similar gene expression patterns, a part of nuclei from AST-FB is predicted as AST-PP. As both Neu-NRGN-I and Neu-NRGN-II are NRGN-expressing neurons, nuclei from Neu-NRGN-I were mostly predicted as Neu-NRGN-II. As to Neu-mat, more than 40% nuclei were predicted as L2/3, which may indicate that the gene expression patterns between Neu-mat and L2/3 are similar. For most cell types, the predictive accuracy of diagnosis label is larger than 70%, and the top highest accuracy values appear in L2/3, L5/6-CC, IN-SV2C, L4, and AST-PP, showing that these cell types may be more vulnerable in ASD.

DISCUSSION

Genetic studies have identified variants associated with ASD, while the causal variants and the specific cell types in which the disease-risk variants may be active are unclear. Genes may demonstrate diverse functions across different brain cell types. Different functions may be dysregulated and causal genes may be distinct across different brain cells in ASD. Recently, the newly available single-nucleus RNA-sequencing data of ASD (Velmeshev et al., 2019) makes it possible to study the cell-type heterogeneity of ASD directly. The authors identified DE genes between ASD and controls in a cell type-specific way and found that the top DE neuronal genes were identified in L2/3 and IN-VIP, and the top DE genes in non-neuronal cell types were identified in AST-PP and microglia. The relative changes of DE genes in L2/3 and microglia were the most predictive of clinical severity of ASD patients and the cell types that are recurrently affected across multiple patients included L2/3 and L5/6-CC. They concluded that

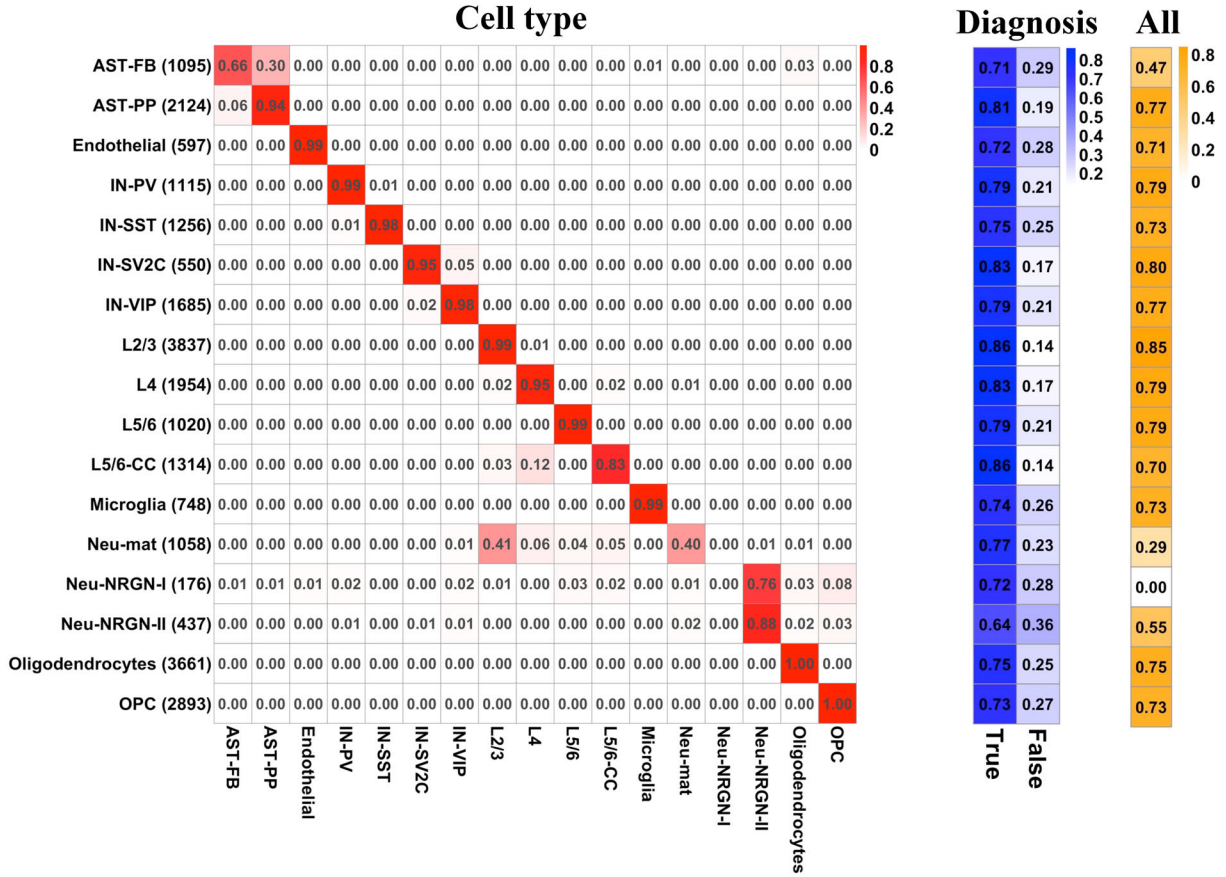

FIGURE 4 | For each cell type, **(left)** the proportion of the number of nuclei predicted as each cell type to the total number of nuclei, **(middle)** the proportion of correct and incorrect predictions for the label of diagnosis, and **(right)** the proportion of correct predictions for all labels in the test set.

synaptic signaling of upper-layer excitatory neurons and the molecular state of microglia are preferentially affected in ASD, and the dysregulation of specific groups of genes in cortico-cortical projection neurons correlates with clinical severity of ASD.

Actually, except for genetic and genomic studies, gene prioritization studies (Kong et al., 2012; Cogill and Wang, 2016; Guan et al., 2016; Oh et al., 2017) can be applied to detect ASD risk genes, which can help to identify high-confidence gene candidates. In this study, to characterize the cell-type heterogeneity of ASD and to identify cell type-specific genes and gene sets associated with ASD, we constructed multiple kinds of predictive models based on the human brain nucleus gene expression data of ASD and controls (Velmeshev et al., 2019). By constructing cell type-specific predictive models based on individual genes, we found that AST-PP, IN-PV, L2/3, L4, and L5/6-CC may be more vulnerable in ASD. They have more RFE genes and the corresponding cell type-specific predictive models have better performances. Actually, they have more DE genes identified by edgeR and more SFARI ASD genes. These indicate that more genes may be dysregulated in these cell types, and these cell types may be mainly affected by ASD. In addition, we also compared the RFE genes with the DE genes identified in the single-nucleus RNA-seq study of ASD (Velmeshev et al., 2019). We found that RFE genes are significantly overlapped with Velmeshev's genes for all cell types, especially for microglia, L2/3, L4, and IN-VIP. The functions of genes with predictive power for ASD are different, and the top important genes are distinct across different cell types, highlighting the cell-type heterogeneity of ASD. However, some genes appearing as top important genes in several cell types are of note. For instance, gene *BCYRN1* has the largest importance in all excitatory neurons, including L2/3, L4, L5/6, and L5/6-CC. Gene *BCYRN1* is involved in the regulation of synaptogenesis, and there have been several literatures linking *BCYRN1* and Alzheimer's disease, a neurological disease (Wan et al., 2017; Hu et al., 2018), which implies the possible association between *BCYRN1* and ASD. Besides, *BCYRN1* has been prioritized in a blood-based gene expression study of ASD (Ivanov et al., 2015).

As genes interact with others, the integrity of disease gene modules instead of individual genes may determine the manifestation of a disease in cells (Kitsak et al., 2016; Mohammadi et al., 2019). Therefore, in addition to identifying the individual cell type-specific risk genes, it is valuable to identify cell type-specific gene sets/modules associated

with ASD. By constructing cell type-specific gene set-based predictive models, we also noted cell types of AST-PP, IN-PV, L2/3, L4, and L5/6-CC. The identified gene sets specific to these cell types are different. For AST-PP, the ASD-associated gene sets include *REACTOME_DISEASE, GO_REGULATION_OF_CELL_POPULATION_PROLIFERATION, GO_POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY, GO_SIGNALING_RECEPTOR_BINDING*, and *GO_ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY*. For the other four neuronal cell types, the ASD-associated gene sets are mostly related to cell junction, synapse, neuron projection, neurogenesis, neuron differentiation, and cell projection organization. We found that gene *HSPA1A* appears as the most important gene in all AST-PP specific ASD-associated gene sets. Actually, heat shock proteins play a central role in the development of neurological disorders, of which the *HSP70* family has been shown its functions (Turturici et al., 2011), and *HSPA1A*, a member of *HSP70* family, has already been associated with ASD (Lin et al., 2014). Gene *CCK* is prioritized in L2/3, L4, and L5/6-CC, which is a kind of gut peptide hormone. Gut peptide hormones have been found across different brain regions, and many of them are involved with ASD-related deficits (Qi and Zhang, 2020).

Overall, we found that the functions of genes with predictive power for ASD are different and the top important genes are distinct across different cell types, depicting the cell-type heterogeneity of ASD. The findings suggest that L2/3, L4, L5/6-CC, AST-PP, and IN-PV are mainly affected in ASD. The results show that it may be feasible to use single cell/nucleus gene expression for ASD detection and the constructed predictive models can promote the diagnosis of ASD. Our method prioritizes ASD-associated cell type-specific genes and gene sets, which may be used as potential biomarkers of ASD, promoting the design of effective interventions.

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. The codes were deposited on Github: https://github.com/JGuan-lab/Cell-type-specific-predictive-model.

AUTHOR CONTRIBUTIONS

JG conceived the study. YW, YL, and QY wrote the codes and analyzed the data. JG, YZ, and GJ interpreted the results. JG and YW wrote the manuscript. All authors approved the final manuscript.

FUNDING

This work has been supported by the National Natural Science Foundation of China (Nos. 61803320 and 61573296), the Fundamental Research Funds for the Central Universities in China (Xiamen University: 202010384099), and the fund with Xiamen YLZ Yihui Technology Co., Ltd. (XDHT2020131A).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.628539/full#supplementary-material

Supplementary Figure 1 | The density plot of the percentage of variance explained by each factor across all genes **(A)** and highly variable genes **(B)**. Each curve denotes one factor.

Supplementary File 1 | The performances of cell type-specific predictive models built based on all genes, top 500, 1000, and 1500 important genes, RFE genes, edgeR genes, and Velmeshev's genes.

Supplementary File 2 | The calculated gene importance in each cell type-specific predictive models built based on all genes.

Supplementary File 3 | The calculated gene importance in each cell type-specific predictive models built based on RFE genes.

Supplementary File 4 | The enriched GO terms with cell type-specific RFE genes.

Supplementary File 5 | The identified ASD-associated cell type-specific gene sets along with their top five important genes and predictive performances. For these gene sets, the frequency and averaged importance of each gene included in the gene sets are listed.

Supplementary File 6 | The enriched GO terms for genes with averaged importance >10% included in the identified ASD-associated cell type-specific gene sets.

REFERENCES

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., et al. (2016). mlr: machine learning in R. *J. Machine Learn. Res.* 17, 5938–5942.

Bosl, W. J., Tager-Flusberg, H., and Nelson, C. A. (2018). EEG analytics for early detection of autism spectrum disorder: a data-driven approach. *Sci. Rep.* 8:6828. doi: 10.1038/s41598-018-24318-x

Boulesteix, A.-L., and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings Bioinform.* 8, 32–44. doi: 10.1093/bib/bbl016

Calvo, S. E., Clauser, K. R., and Mootha, V. K. (2016). MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.* 44, D1251–D1257. doi: 10.1093/nar/gkv1003

Chen, R., Jiao, Y., and Herskovits, E. H. (2011). Structural MRI in autism spectrum disorder. *Pediatric Res.* 69, 63–68. doi: 10.1203/PDR.0b013e318212c2b3

Cogill, S., and Wang, L. (2016). Support vector machine model of developmental brain gene expression data for prioritization of Autism risk gene candidates. *Bioinformatics* 32, 3611–3618. doi: 10.1093/bioinformatics/btw498

De Rubeis, S., He, X., Goldberg, A. P., Poultney, C. S., Samocha, K., Cicek, A. E., et al. (2014). Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* 515, 209–215. doi: 10.1038/nature13772

Duda, M., Daniels, J., and Wall, D. P. (2016). Clinical evaluation of a novel and mobile autism risk assessment. *J. Autism Dev. Dis.* 46, 1953–1961. doi: 10.1007/s10803-016-2718-2714

Duda, M., Kosmicki, J. A., and Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Trans. Psychiatry* 4:e424. doi: 10.1038/tp.2014.65

Gaugler, T., Klei, L., Sanders, S. J., Bodea, C. A., Goldberg, A. P., Lee, A. B., et al. (2014). Most genetic risk for autism resides with common variation. *Nat. Genet.* 46, 881–885. doi: 10.1038/ng.3039

Guan, J., Yang, E., Yang, J., Zeng, Y., Ji, G., and Cai, J. J. (2016). Exploiting aberrant mRNA expression in autism for gene discovery and diagnosis. *Human Genet.* 135, 797–811. doi: 10.1007/s00439-016-1673-1677

Gupta, S., Ellis, S. E., Ashar, F. N., Moes, A., Bader, J. S., Zhan, J., et al. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* 5:5748. doi: 10.1038/ncomms6748

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learn.* 46, 389–422.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hu, G., Niu, F., Humburg, B. A., Liao, K., Bendi, V. S., Callen, S., et al. (2018).). Molecular mechanisms of long noncoding RNAs and their role in disease pathogenesis. *Oncotarget* 9, 18648–18663. doi: 10.18632/oncotarget.24307

Hyde, K. K., Novack, M. N., LaHaye, N., Parlett-Pelleriti, C., Anden, R., Dixon, D. R., et al. (2019). Applications of supervised machine learning in autism spectrum disorder research: a review. *Rev. J. Autism Dev. Dis.* 6, 128–146. doi: 10.1007/s40489-019-00158-x

Iossifov, I., Ronemus, M., Levy, D., Wang, Z., Hakker, I., Rosenbaum, J., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron* 74, 285–299. doi: 10.1016/j.neuron.2012.04.009

Ivanov, H. Y., Stoyanova, V. K., Popov, N. T., Bosheva, M., and Vachev, T. I. J. B. (2015). Blood-based gene expression in children with autism spectrum disorder. *Biodiscovery* 17:e8966.

Kitsak, M., Sharma, A., Menche, J., Guney, E., Ghiassian, S. D., Loscalzo, J., et al. (2016). Tissue specificity of human disease module. *Sci. Rep.* 6:35241. doi: 10.1038/srep35241

Kong, S. W., Collins, C. D., Shimizu-Motohashi, Y., Holm, I. A., Campbell, M. G., Lee, I. H., et al. (2012). Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* 7:e49475. doi: 10.1371/journal.pone.0049475

Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Statist. Software* 28, 1–26.

Levy, S., Duda, M., Haber, N., and Wall, D. P. (2017). Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Mol. Autism* 8:65. doi: 10.1186/s13229-017-0180-186

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lin, M., Zhao, D., Hrabovsky, A., Pedrosa, E., Zheng, D., and Lachman, H. M. (2014). Heat Shock alters the expression of schizophrenia and autism candidate genes in an induced pluripotent stem cell model of the human telencephalon. *PLoS One* 9:e94968. doi: 10.1371/journal.pone.0094968

Lun, A. T. L., McCarthy, D. J., and Marioni, J. C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res.* 5:2122. doi: 10.12688/f1000research.9501.2

Mohammadi, S., Davila-Velderrain, J., and Kellis, M. (2019). Reconstruction of cell-type-specific interactomes at single-cell resolution. *Cell Systems* 9, 559–568.e4. doi: 10.1016/j.cels.2019.10.007

Neale, B. M., Kou, Y., Liu, L., Ma'ayan, A., Samocha, K. E., Sabo, A., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485, 242–245. doi: 10.1038/nature11011

Oh, D. H., Kim, I. B., Kim, S. H., and Ahn, D. H. (2017). Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clin. Psychopharmacol. Neurosci.* 15, 47–52. doi: 10.9758/cpn.2017.15.1.47

O'Roak, B. J., Vives, L., Girirajan, S., Karakoc, E., Krumm, N., Coe, B. P., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485, 246–250. doi: 10.1038/nature10989

Parikshak, N. N., Swarup, V., Belgard, T. G., Irimia, M., Ramaswami, G., Gandal, M. J., et al. (2016). Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. *Nature* 540, 423–427. doi: 10.1038/nature20612

Qi, X.-R., and Zhang, L. (2020). The potential role of gut peptide hormones in autism spectrum disorder. *Front. Cell. Neurosci.* 14:73. doi: 10.3389/fncel.2020.00073

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. doi: 10.1186/1471-2105-12-77

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* 485, 237–241. doi: 10.1038/nature10945

Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., et al. (2020). Large-Scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell* 180, 568–584.e23. doi: 10.1016/j.cell.2019.12.036

Turner, T. N., Hormozdiari, F., Duyzend, M. H., McClymont, S. A., Hook, P. W., Iossifov, I., et al. (2016). Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory DNA. *Am. J. Hum. Genet.* 98, 58–74. doi: 10.1016/j.ajhg.2015.11.023

Turturici, G., Sconzo, G., and Geraci, F. (2011). Hsp70 and its molecular role in nervous system diseases. *Biochem. Res. Int.* 2011:618127. doi: 10.1155/2011/618127

Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., et al. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science* 364, 685–689. doi: 10.1126/science.aav8130

Voineagu, I., Wang, X., Johnston, P., Lowe, J. K., Tian, Y., Horvath, S., et al. (2011). Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474, 380–384. doi: 10.1038/nature10110

Wall, D. P., Dally, R., Luyster, R., Jung, J.-Y., and DeLuca, T. F. (2012). Use of artificial intelligence to shorten the behavioral diagnosis of autism. *PLoS One* 7:e43855. doi: 10.1371/journal.pone.0043855

Wan, P., Su, W., and Zhuo, Y. (2017). The role of long noncoding RNAs in neurodegenerative diseases. *Mol. Neurobiol.* 54, 2012–2021. doi: 10.1007/s12035-016-9793-9796

Wold, H. (1966). "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, ed. P. R. Krishnajah (NewYork, NY: Academic Press), 391–420.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Conflict of Interest: YZ was employed by Xiamen YLZ Yihui Technology Co., Ltd.

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

frontiers in Physiology

ORIGINAL RESEARCH
published: 18 January 2021
doi: 10.3389/fphys.2020.597494

Repositioning Drugs on Human Influenza A Viruses Based on a Novel Nuclear Norm Minimization Method

Hang Liang[1†], Li Zhang[1†], Lina Wang[1], Man Gao[1], Xiangfeng Meng[2], Mengyao Li[2], Junhui Liu[1], Wei Li[1] and Fanzheng Meng[1]*

[1] Pediatric Department of Respiration II, The First Hospital of Jilin University, Changchun, China, [2] Norman Bethune Health Science Center, Jilin University, Changchun, China

Influenza A viruses, especially H3N2 and H1N1 subtypes, are viruses that often spread among humans and cause influenza pandemic. There have been several big influenza pandemics that have caused millions of human deaths in history, and the threat of influenza viruses to public health is still serious nowadays due to the frequent antigenic drift and antigenic shift events. However, only few effective anti-flu drugs have been developed to date. The high development cost, long research and development time, and drug side effects are the major bottlenecks, which could be relieved by drug repositioning. In this study, we proposed a novel antiviral Drug Repositioning method based on minimizing Matrix Nuclear Norm (DRMNN). Specifically, a virus-drug correlation database consisting of 34 viruses and 205 antiviral drugs was first curated from public databases and published literature. Together with drug similarity on chemical structure and virus sequence similarity, we formulated the drug repositioning problem as a low-rank matrix completion problem, which was solved by minimizing the nuclear norm of a matrix with a few regularization terms. DRMNN was compared with three recent association prediction algorithms. The AUC of DRMNN in the global fivefold cross-validation (fivefold CV) is 0.8661, and the AUC in the local leave-one-out cross-validation (LOOCV) is 0.6929. Experiments have shown that DRMNN is better than other algorithms in predicting which drugs are effective against influenza A virus. With H3N2 as an example, 10 drugs most likely to be effective against H3N2 viruses were listed, among which six drugs were reported, in other literature, to have some effect on the viruses. The protein docking experiments between the chemical structure of the prioritized drugs and viral hemagglutinin protein also provided evidence for the potential of the predicted drugs for the treatment of influenza.

Keywords: influenza A viruses, anti-viral drugs, treatment, drug repositioning, hemagglutinin

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis Co., Ltd., China

Reviewed by:
Guohua Huang,
Shaoyang University, China
Ling Tong,
Chifeng Municipal Hospital, China

***Correspondence:**
Fanzheng Meng
mengfanzheng1972@163.com

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Physiology

Received: *21 August 2020*
Accepted: *24 November 2020*
Published: *18 January 2021*

Citation:
Liang H, Zhang L, Wang L, Gao M, Meng X, Li M, Liu J, Li W and Meng F (2021) Repositioning Drugs on Human Influenza A Viruses Based on a Novel Nuclear Norm Minimization Method. Front. Physiol. 11:597494. doi: 10.3389/fphys.2020.597494

INTRODUCTION

Influenza viruses spread quickly and are among the main causes of human death. Influenza is an acute respiratory tract infection caused by influenza viruses that seriously endangers human health. Symptoms include a stuffy nose, cough, sore throat, headache, fever, chills, anorexia, and myalgia. These symptoms are the result of inflammation caused by a viral infection

(Eccles, 2005). Type A influenza viruses are major pathogens for humans. Infection with influenza A viruses usually results in mild and self-limiting illness. For some people, they can cause complications such as pneumonia, bronchitis, sinusitis, and ear infections, leading to serious illness and even death (CDC, 2009). Influenza complications are often associated with secondary bacterial infections, which may be due to the virus inducing a series of changes in the host lung epithelial cells, making them easy to adhere and invade, leading to changes in the immune response (Mccullers, 2006, 2014; Jamieson et al., 2013). Influenza A viruses are evolving very fast, which allows them to regularly produce new strains of human immunodeficiency, leading to periodic pandemics (Taubenberger and Kash, 2010). Among the known 16 hemagglutinin (HA) subtypes and nine neuraminidase (NA) subtypes of influenza A viruses, only H3N2 subtypes and H1N1 subtypes are currently spreading among the population (Webster et al., 1992).

Prevention and treatment of influenza A viruses usually use vaccines or anti-flu chemical drugs. However, the effectiveness of the vaccine is based on the similarity of the vaccine strain to the influenza virus strain that is circulating (Tosh and Poland, 2008). Influenza viruses continue to mutate, and conventional vaccines may not easily prevent or treat influenza outbreaks caused by new viruses. Therefore, the research of anti-influenza chemical drugs is of great significance (Glezen, 2006). Two types of drugs commonly used to prevent or treat influenza A viruses are amantadine and neuraminidase inhibitors (NAIs). Studies have shown that the effectiveness of amantadine is limited by the high prevalence of influenza A virus (H3N2) with the S31N mutation in M2 (Barr et al., 2007; Saito et al., 2007). In 2008, the H1N1 subtype with the H274Y mutation in NA appeared, which raised concerns about the use of oseltamivir (Hauge et al., 2009; Hurt et al., 2009a). On the other hand, the incidence of zanamivir-resistant viruses is low. Chemiluminescence NAI analysis confirmed that the H3N2 subtype with the D151A/V mutation in NA reduces the sensitivity of zanamivir (Sheu et al., 2008). It has been reported that an H1N1 subtype isolate with a new Q136K mutation in NA that is resistant to zanamivir has been isolated in Oceania and Southeast Asia (Hurt et al., 2009b). Burch et al. (2009) commissioned by the National Institute of health and clinical optimization, searched the database of studies on the use of neuraminidase inhibitors in the treatment of seasonal influenza. They presented the results to healthy adults (i.e., adults without known comorbidities) and people at risk for influenza-related complications (Burch et al., 2009). Rohloff et al. (1998) prepared GS-4104, an anti-influenza drug of 3,3-Diaryloxidoles, with a high yield (62–99%), through an isobutyl or substituent reaction.

Nevertheless, the development of new drugs for the prevention and treatment of influenza A viruses is a long process with a high cost. Therefore, repeated use of drugs is a strategy to find specific drugs for the treatment of influenza A viruses among existing drugs. Compared with developing new drugs, it can greatly shorten the time and reduce the cost. However, blindly repeating the use of drugs and randomized clinical trials is risky, and there is still the problem that they are time-consuming and costly. At present, some calculation methods provide new testable hypotheses for the repositioning of systemic drugs (Cheng et al., 2016; Santos et al., 2017). Therefore, more computational methods for drug screening are urgently needed to find drugs that may have therapeutic effects against Influenza A viruses and thereby solve these time-consuming and costly problems.

In this study, we developed a matrix decomposition-based antiviral drug reuse method to predict the efficacity of drugs for the treatment of influenza A virus (H3N2), and the method mainly includes the following four steps: (1) collect and download data about viruses and drugs from the literature; (2) calculate a similar chemical structure of the drugs and similar genetic sequence of the virus; (3) establish a heterogeneous drug-virus network based on the virus and drug-related data, the drug similarity network, and the virus similarity network; (4) use the nuclear norm minimization method to obtain the drug most likely to have a therapeutic effect on the virus. Finally, the experiment evaluated the performance of this method through fivefold CV, and the results showed that DRMNN achieved an average AUC value of 0.8661.

MATERIALS AND METHODS

Human Virus and Drug Interaction Associations

In order to construct a human virus–drug interaction network, we used text mining technology to study a large number of previous documents and screened a drug database, and we finally found 408 confirmed human virus-drug interaction associations, including 34 viruses and 205 drugs. The adjacency matrix variable of the virus-drug interaction network was defined as A. If the drug $d(i)$ has an effect on the virus $v(j)$, then $A(ij)$ is equal to 1, otherwise it is 0. That is:

$$A(ij) = \begin{cases} 1, & \textit{if drug } d(i) \textit{ has an effect on the virus } v(j) \\ 0, & \textit{otherwise} \end{cases}$$

Chemical Structure Similarity of Drugs

The drug discovery process is characterized by a long cycle, high investment, and high risk. In order to shorten the drug development cycle and control the risk and cost of the drug development process, computer-aided drug design (CADD) has become an important tool for new drug development and drug screening. Molecular similarity calculation is widely used in the CADD field. Molecular shape similarity is usually based on the Tanimoto Coefficient (TC). The MACCS fingerprint in Openbabel V2.3.1 software was used to calculate the molecular fingerprint similarity between two drugs, represented by TC. Drugs' chemical structure information was downloaded from the DrugBank database. If the MACCS fragment bit strings of two drug molecules $d(i)$ and $d(j)$ were $m(i)$ and $m(j)$, respectively, a was set as the fingerprints of the two drugs. The similarity between drugs $d(i)$ and $d(j)$ was defined as:

$$DS\left(d\left(i\right), d\left(j\right)\right) = TC = \frac{a}{m(i) + m\left(j\right) - a} \quad (1)$$

The *TC* value ranges from zero (no common bits) to one (all bits are the same), and it can be widely used in various drug development and repositioning processes. The chemical structure similarity matrix of drugs is represented by *DS*. Finally, the calculated drug similarity constitutes a medical chemical structure similarity network.

Viral Similarity

Our understanding of any virus often starts from its sequence. With the development of gene sequencing technology, a lot of multiple sequence comparison software has also emerged. MAFFT is a multi-sequence alignment program (Katoh et al., 2005) that provides a series of alignment methods with the advantages of fast alignment and high accuracy. Therefore, we used MAFFT to calculate the sequence similarity between viruses to express the similarity between viruses. Then, we constructed a viral similarity network and used *VS* to represent the viral similarity matrix.

Human Virus-Drug Interactome Network

We constructed a human virus-drug interactome network by using human virus and drug interaction associations, a network of chemical structure similarity of drugs, and a virus similarity network. Then, the heterogeneous human virus-drug interactome network was treated as a bipartite graph $G(V, D, E)$, where V represented human viruses, D represented drugs, and E was the edges connecting human viruses and drugs. Therefore, the adjacency matrix of the heterogeneous drug-virus network matrix can be defined as:

$$B = \begin{bmatrix} DS & A^T \\ A & VS \end{bmatrix} \tag{2}$$

where A^T is the transposition of A.

DRMNN

An overview of DRMNN was shown in **Figure 1**. The nuclear norm is the sum of the singular values of the matrix, which is used to constrain the low rank of the matrix. For sparse data, the matrix has a low rank and contains a lot of redundant information, which can be used to recover data and extract features. The nuclear norm has been widely used in various fields and has achieved good results (Yang et al., 2019). Generally, when

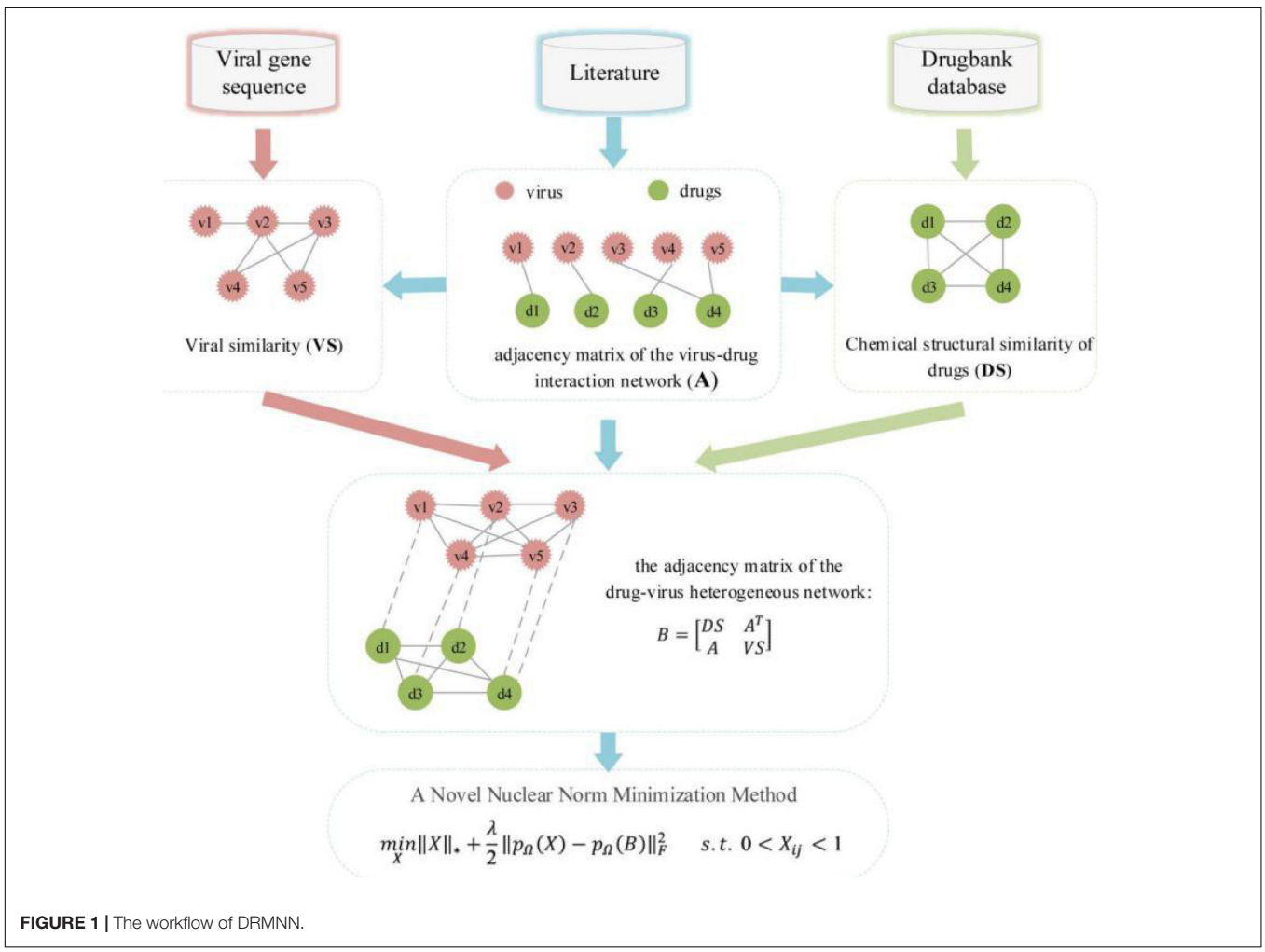

FIGURE 1 | The workflow of DRMNN.

a matrix has a low rank, the kernel norm minimization problem can be expressed as:

$$min\, X||X||_* \tag{3}$$

where $||X||_*$ represents the kernel norm of X, which is defined as the sum of all singular values of X. The kernel norm minimization model is a convex optimization problem.

In order to predict the drug-virus association, the elements in the drug similarity matrix DS and the virus similarity matrix VS are in the interval [0, 1]. The elements in the correlation matrix A are 0 or 1. The predicted value of the unknown entry is expected to be in the range of [0, 1], where a predicted value close to 1 suggests that it may be indicative of an association and vice versa. However, in the above matrix completion model (2), the entries in the completed matrix can be any real values in $(-, +)$. However, it has no practical significance for values greater than 1 and less than 0. Therefore, it is important to add a constraint to the matrix completion model to ensure that the missing elements that are not found are in the interval [0, 1]. In addition, because there may be a lot of "noise" data in the drug and virus data, the drug relocation model should tolerate the potential noise as much as possible. The noise-tolerant matrix completion model is:

$$min\, X||X||_*\ s.t.\ ||p_\Omega(X) - p_\Omega(B)||_F \le \varepsilon \tag{4}$$

where ε is the measurement noise level, $\mathbf{\Omega}$ is a set of index pairs (i, j) containing all known entries in B, and p_Ω is the projection operator on $\mathbf{\Omega}$.

$$(p_\Omega(X))_{ij} = \begin{cases} X_{ij}, & (i,j) \in \mathbf{\Omega} \\ 0, & otherwise \end{cases} \tag{5}$$

However, there are a number of difficulties involved in solving this model with its inequality constraints, for example, how to choose the appropriate model parameters and how to choose an effective solution algorithm. Therefore, we usually replace the inequality constraint model with a regularized model. The introduction of soft regularization can tolerate unknown noise and also make the solution much more convenient to arrive at. Then the model can be rewritten as the following:

$$\min_X ||X||_* + \frac{\lambda}{2}||p_\Omega(X) - p_\Omega(B)||^{2_F} s.t.\ 0 < X_{ij} < 1 \tag{6}$$

where $||\cdot||_F$ denotes the Frobenius norm and λ is the parameter that balances the nuclear specification and the error term. To solve the optimization problem in Eq. (5), we chose the more classic alternating direction multiplier method (ADMM) (Gu et al., 2016). It is worth noting that our objective function is convex. With the introduction of the auxiliary matrix H, the ADMM framework can be optimized in the following equivalent form.

$$\min_X ||X||_* + \frac{\lambda}{2}||p_\Omega(X) - p_\Omega(B)||^{2_F} s.t.\ X = H, 0 < H_{ij} < 1 \tag{7}$$

Therefore, the enhanced Lagrange function becomes the following:

$$\mathcal{L}(H, X, Y, \lambda, \mu) = ||X||_* + \frac{\lambda}{2}||p_\Omega(X) - p_\Omega(B)||^{2_F} + T_\gamma\left(Y^T(X - H)\right) + \frac{\mu}{2}||X - H||^{2_F} \tag{8}$$

where Y is the Lagrange multiplier and $\mu > 0$ is the penalty parameter. The solution process of DRMNN belongs to an iterative solution. Therefore, when we iterate k times, we need to calculate the value of iterations H_{k+1}, Y_{k+1}, and X_{k+1} according to the result of the k^{th} iteration.

Update: Repeat the following steps until there is convergence or a predetermined number of iterations.

Fix X_k and Y_k and calculate a matrix H_{k+1} to minimize Eq. (7).

$$\begin{aligned} H_{k+1} &= arg\, min_{0\le H\le 1}\mathcal{L}(H, X_k, Y_k, \lambda, \mu) \\ &= arg\, min_{0\le H\le 1}\frac{\lambda}{2}||p_\Omega(X) - p_\Omega(B)||^{2_F} + T_\gamma(Y^{T(X_{k-H})}) + \frac{\mu}{2}||X_{k-H}||^{2_F} \end{aligned} \tag{9}$$

Here, H^* is the optimal solution of $argmin_{0\le H\le 1}\mathcal{L}(H, X_k, Y_k, \lambda, \mu)$, if and only if

$$\lambda p_\Omega^*\left(p_\Omega(H^*) - p_\Omega(B)\right) - Y_k - \mu\left(X_{k-H^*}\right) = 0 \tag{10}$$

holds, where p_Ω^* represents the adjoint operator of p_Ω . Then, the closed solution becomes:

$$\begin{aligned} H^* &= \left(\alpha + \frac{\lambda}{\mu}p_\Omega^* p_\Omega\right)^{-1}\left(\frac{1}{\mu}Y_k + \frac{\lambda}{\mu}p^{*\Omega}p_\Omega(B) + X_k\right) \\ &= \left(\alpha - \frac{\lambda}{\mu}p_\Omega^* p_\Omega\right)\left(\frac{1}{\mu}Y_k + \frac{\lambda}{\mu}p^{*\Omega}p_\Omega(B) + X_k\right) \\ &= \left(\frac{1}{\mu}Y_k + \frac{\lambda}{\mu}p_\Omega(B) + X_k\right) - \frac{\lambda}{\mu+\lambda}\left(\frac{1}{\mu}Y_k + \frac{\lambda}{\mu}p_\Omega(B) + X_k\right) \end{aligned} \tag{11}$$

where α is the identity operator. $\left(\alpha + \frac{\lambda}{\mu}p_\Omega^* p_\Omega\right)^{-1}$ denotes the inverse operator of $\left(\alpha + \frac{\lambda}{\mu}p_\Omega^* p_\Omega\right)$, and it is equal to $\left(\alpha - \frac{\lambda}{\mu}p_\Omega^* p_\Omega\right)$. It's worth noting that $p_\Omega^* p_\Omega \quad p_\Omega$. Considering the interval [0,1] constraint, we limit the range of the elements of H_{k+1} to [0, 1] such that

$$\left(H_{k+1}\right)_{ij} = \begin{cases} 1, & H_{ij}^* > 1 \\ H_{ij}^*, & 0 < H_{ij}^* < 1 \\ 0, & H_{ij}^* < 0 \end{cases} \tag{12}$$

Fix H_{k+1} and Y_k and calculate a matrix X_{k+1} to minimize Eq. (7).

$$X_{k+1} = arg\min_{0\le H\le 1} \mathcal{L}\left(H_{k+1}, X, Y_k, \lambda, \mu\right)$$

$$= arg\ min_{0\le H\le 1}||X||_* + \frac{\mu}{2}||X - \left(H_{k+1} - \frac{1}{\mu}Y_k\right)||^{2_F}$$

$$= \frac{\vartheta_1}{\theta}\left(H_{k+1} - \frac{1}{\mu}Y_k\right) \quad (13)$$

where $\vartheta_\tau(X)$ is the singular value shrinkage operator which is defined as:

$$\vartheta_\tau(X) = \int_{i=1}^{\theta_i \ge \tau} (\theta_i - \tau)\,\beta_i \gamma^{T_i} \quad (14)$$

where β_i and γ_i are the left and right singular vectors corresponding to θ_i, respectively. The θ_i are the singular values of X, which are greater than τ .

Fix H_{k+1} and X_{k+1} and calculate a matrix Y_{k+1}.

$$Y_{k+1} = Y_k + \kappa\mu\left(X_{k+1} - H_{k+1}\right) \quad (15)$$

where κ is the learning rate which is set to 1 in this study for simplicity. Iterate according to the above iteration rules until convergence, and finally, we obtain the matrix H_k after convergence. Therefore, the final prediction matrix A^* for potential association between drugs and viruses is

$$A^* \leftarrow \begin{bmatrix} DS^* & A^{*T} \\ A^* & VS^* \end{bmatrix} \leftarrow H_k \quad (16)$$

RESULTS

Indicators of Performance Evaluation

For a binary classification problem, the samples are generally divided into two types: positive samples and negative samples. In dichotomies, therefore, there are usually the following four situations:

TP: True Positives, which means the number from the sample itself that are positive and are predicted to be positive;
FP: False Positives, which means the number of samples that are negative and ultimately predicted to be positive;
TN: True Negatives, which indicates the number of negatives from the sample itself that are also predicted to be negative;
FN: False Negatives, which indicates the number of positives that the sample itself ultimately predicted to be negative.

The commonly used evaluation indicators of classification models are: precision, specificity, and sensitivity. Their calculation formula is as follows:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Specificity(1 - FPR) = \frac{TN}{TN + FP}$$

$$Sensitivity\ (TPR = Recall) = \frac{TP}{TP + FN}$$

The performance evaluation indicators we usually adopt are the ROC curve and area under the ROC curve (AUC value), as well as the PR curve and area under the PR curve (AUPR value). The full name of the ROC is Receiver Operating Characteristic, its abscissa is the false positive rate (FPR), and its ordinate is the true positive rate (TPR). Among them, the false positive rate is the proportion of all negative samples that the classifier incorrectly predicts as positive, also known as 1-specificity. Similarly, the true positive rate refers to the proportion of positive samples correctly identified by the classifier out of all positive samples, which is also called sensitivity. Then we draw the ROC curve based on the TPR and FPR. Calculate the area under the ROC curve and perform a numerical evaluation of the model's performance. The area under the ROC curve is defined as AUC (Area Under Curve). AUC = 0.5 means completely random prediction, and AUC = 1 means a completely accurate prediction. Obviously, the area is less than 1, but the larger the AUC, the better the performance of the classifier. On the other hand, the PR curve is actually made by using precision and recall as variables, where recall is the *x*-coordinate and precision is the *y*-coordinate. AUPR represents the area under the PR curve. The closer AUPR is to 1, the better the prediction's performance will be.

Performance in Predicting Virus-Drug Association

In DRMNN, there are two parameters λ and μ that need to be determined. For λ and μ, they are determined from {0.1, 1, 10, 100}, respectively. We performed fivefold CV on the training data set to determine the parameters and found that when λ = 1 and μ = 10, DRMNN performs best. The AUC results are shown in **Table 1**.

In order to evaluate the prediction performance of DRMNN, we applied DRMNN to the known human virus and drug interaction associations A and used the fivefold CV to evaluate its performance. The specific process was as follows: all known human virus and drug interaction associations were randomly divided into five uncrossed sites with equal size. We used one

TABLE 1 | The AUC values using different λ and μ values in fivefold CV on the dataset.

λ \ μ	0.1	1	10	100
0.1	0.7044	0.6992	0.8087	0.8517
1	0.8116	0.8123	**0.8661**	0.8378
10	0.7919	0.7983	0.8589	0.8318
100	0.7823	0.8006	0.8536	0.8276

The bold value indicates that AUC.

of the parts as a test sample for prediction and the other four parts of the sample as training data to build a predictive model. This process was repeated five times and ended when all samples were predicted once. The results showed that the AUC value was 0.8661. The AUPR value was 0.4442.

At present, there are few algorithms for predicting which drugs will effectively treat influenza A viruses by constructing a network of viruses and drugs. Therefore, in this study, we compared network association prediction algorithms in other fields to explore the performance of DRMNN in predicting drugs that can treat influenza A viruses. NCP was first proposed by Gu et al. (2016) to predict miRNA-disease association. Zou et al. (2018) used this method to predict the association between microorganisms and diseases and achieved good results. NCP is a method based on a general nonparametric network, which belongs to the category of unsupervised learning. Its characteristic is that no negative samples are required. The Random Walk with Restart (RWR) algorithm has advantages. It is not only used for the correlation prediction of binary networks, but also for link prediction of various heterogeneous networks, and in various network correlation predictions, the RWR algorithm shows good predictive performance (Chen et al., 2012). The inductive matrix completion (IMC) algorithm and collaborative matrix factorization algorithm (CMF) (Xu et al., 2020) were more commonly used in prediction problems. The IMC algorithm was originally used to predict the association between drugs and targets, and was finally applied by Chen et al. in miRNA-disease association networks, which also showed good

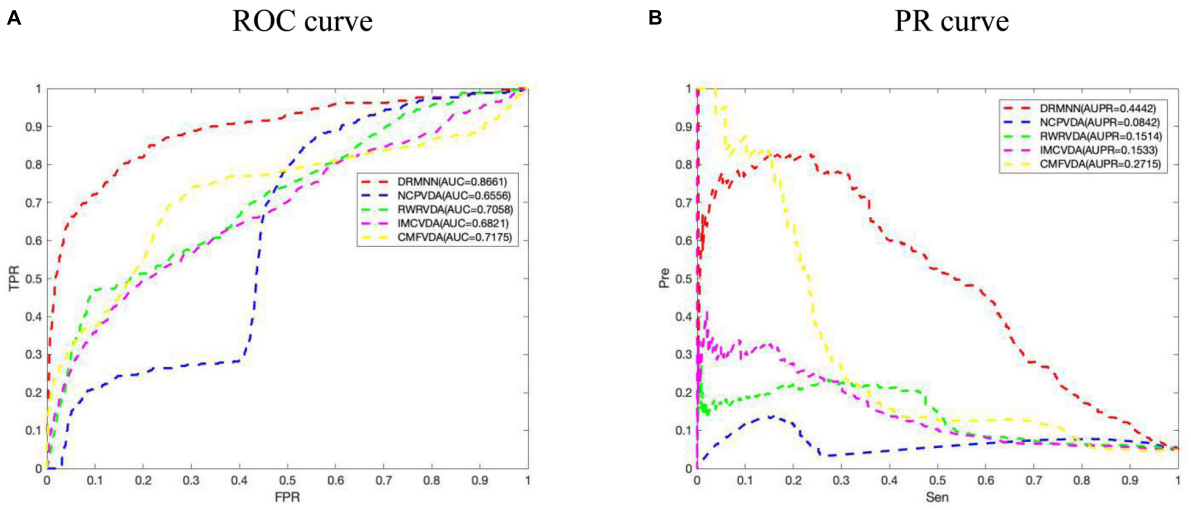

FIGURE 2 | Comparison of predictive performance of DRMNN, NCP, RWR, CMF, and IMC. **(A)** ROC curve and AUC value based on the global fivefold CV. **(B)** PR curve and AUPR values based on global fivefold CV.

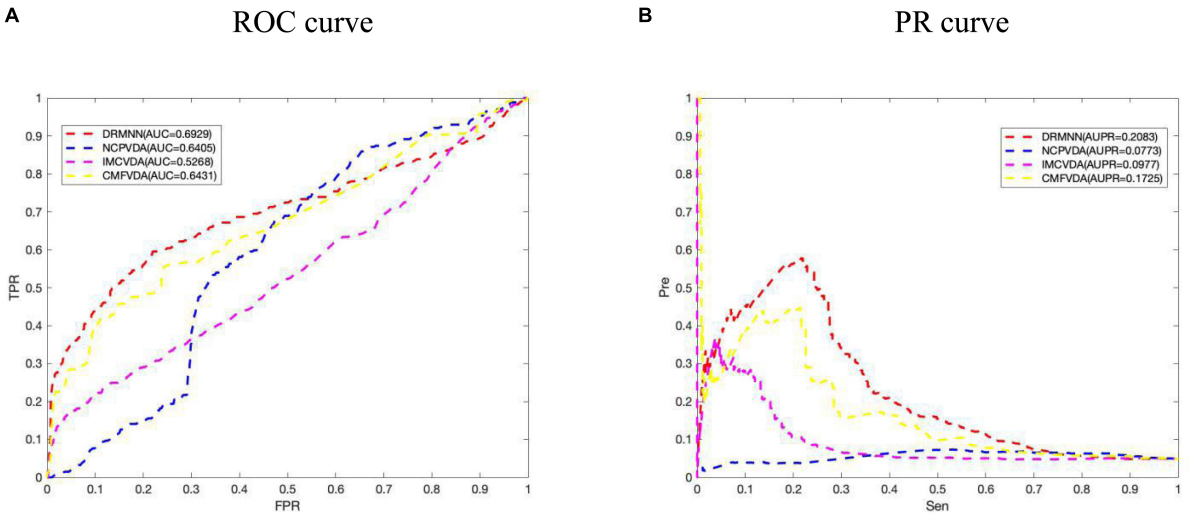

FIGURE 3 | Comparison of predictive performance of DRMNN, NCP, CMF, and IMC. **(A)** ROC curve and AUC value based on the local LOOCV. **(B)** PR curve and AUPRs value based on the local LOOCV.

TABLE 2 | Top 10 possible drugs against influenza A virus (H3N2) predicted by DRMMN.

Virus	Rank	Drug name	Evidence
Influenza A virus (H3N2)	1	Ribavirin	Unconfirmed
	2	Nitazoxanide	Confirmed
	3	Chloroquine	Confirmed
	4	Favipiravir	Unconfirmed
	5	Camostat	Unconfirmed
	6	Mizoribine	Confirmed
	7	Niclosamide	Confirmed
	8	Umifenovir	Confirmed
	9	Mycophenolic acid	Unconfirmed
	10	Amantadine	Confirmed

performance (Chen et al., 2018). In the analysis of all the above methods, RWR and IMC contain parameters that need to be fine-tuned. For all parameters, we select the best parameters by using a global fivefold CV.

We applied DRMNN, NCP, RWR, IMC, and CMF to the 341 associated data between 34 viruses and 205 drugs that were considered. Under the global fivefold CV, the final AUC values of them were 0.8661, 0.6556, 0.7058, 0.6821, and 0.7175, respectively. The ROC curve is shown in **Figure 2A**, indicating that DRMNN showed the best performance in predicting the association between viruses and drugs. We also draw the PR curve in **Figure 2B**. The AUPR of DRMNN, NCP, RWR, CMF, and IMC were 0.4442, 0.0842, 0.1514, 0.1533, and 0.2715, respectively. This once again proves that DRMNN performs best in predicting the treatment of influenza A virus.

In addition, we also carried out a local LOOCV. In particular, for each virus $v(i)$, we removed all known drugs associated with virus $v(i)$, and used the remaining data to build prediction models. But RWR cannot predict new virus-related drugs, so RWR was removed, and only a few other algorithms were compared. The ROC curve and PR curve are shown in **Figure 3**.

TABLE 3 | The binding affinity of the unconfirmed drugs predicted by DRMNN to the target PDB ID: 2VIU.

Drug name	Ribavirin	Favipiravir	Camostat
Free energy of binding (kcal/mol)	−7.4	−5.2	−7.5

The results show that the AUC value and the AUPR value of DRMNN are 0.6929 and 0.2083, respectively, which are much higher than with the other three algorithms. The local LOOCV results also show that DRMNN performed well in predicting potential therapeutic agents for new viruses.

CASE STUDY

Identification of Potential Drugs Against the Influenza A Virus (H3N2)

Accurately determining the drugs to use to treat influenza A virus (H3N2) is the primary task of this study. Through the construction of the viral drug network, we used the DRMNN algorithm to select drugs that may be used to treat influenza A virus (H3N2) and help medical staff choose drugs from a computational perspective. During the construction of the network, the influenza A virus (H3N2) had no association with any drugs. DRMNN was used to predict the probability scores of candidate drugs, and the top 10 drugs for the treatment of influenza A virus (H3N2) are shown in **Table 2**. Nitrazoxanide (ranked 2) can be used to treat *Cryptosporidium parvum* and Giardia infections in children and adults, and it has been licensed in the United States; it is a safe, oral, bioavailable anti-infective drug (White, 2004). In addition to being used to treat protozoan and bacterial infections, thiazoles are also used as a class of broad-spectrum antiviral drugs (Rossignol et al., 2006, 2009a,b; Korba et al., 2008; Elazar et al., 2009; La Frazia et al., 2013). These molecules selectively block the maturation of the viral hemagglutinin through a stage before the resistance to endoglycosidase H digestion and disrupt the intracellular transport and insertion of the HA into the host cell plasma membrane for the correct assembly of the virus and its removal from the host cell to fight off the virus. Studies have found that Nitrazoxanide is effective against influenza A virus (H3N2), which contains the M2 blocker resistance marker S31N (Sleeman et al., 2014). Chloroquine (ranked 3) is a 9-aminoquinolone that can be used to fight malaria and that has biochemical properties that can be used to inhibit virus replication. The report pointed out that chloroquine can inhibit the replication of influenza A virus *in vitro*, and the IC50s of chloroquine to influenza A virus H3N2 are lower than the plasma

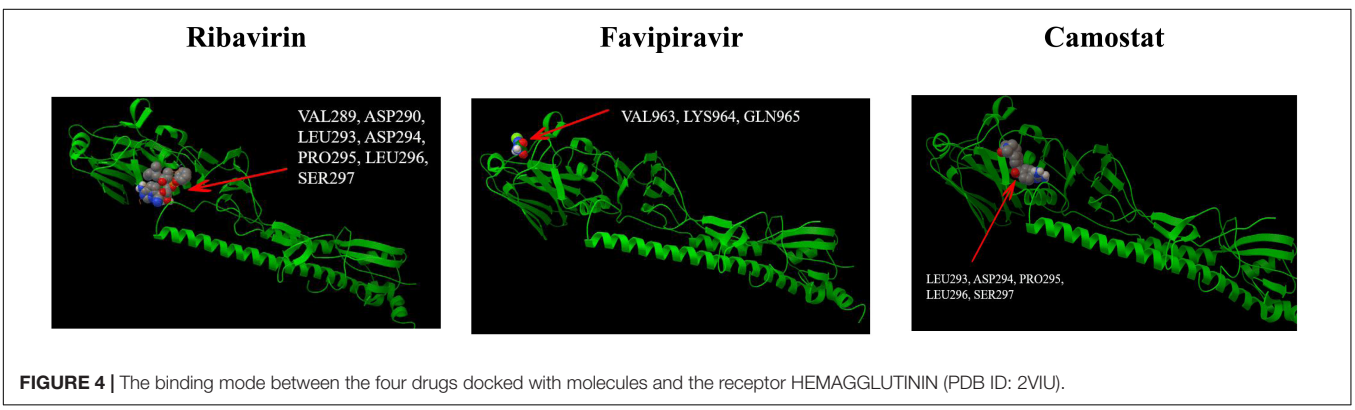

FIGURE 4 | The binding mode between the four drugs docked with molecules and the receptor HEMAGGLUTININ (PDB ID: 2VIU).

concentration reached during acute malaria treatment (Ooi et al., 2006). Umifenovir (ranked 8) is licensed in Russia and is widely used for the prevention or treatment of influenza. Leneva et al. (2019) found that Umifenovir effectively inhibited the replication of antigen-dominant human type A influenza virus using MDCK cell-based enzyme linked immunoadsorption assay (ELISA), and none of the viruses isolated before and during umifenovir treatment showed reduced sensitivity to neuraminidase (NA) inhibitors, suggesting that umifenovir is effective in treating influenza A virus.

Molecular Docking

Molecular docking research has become an economic and modern trend in drug development. It can be used to design known ligands for specific active sites of macromolecules, and it is a method that provides valuable information. The technology-based ligand-protein interaction reveals the possibility of pre-synthesis. In our study, the computer chemistry research of the top five drugs predicted by DRMNN was being blindly connected in online and offline modes. The Autodock 4.2 package[1] was used for offline docking. The X-ray crystal structure of the protein was searched from the RCSB protein database[2]. The PDB ID is a 2VIU macromolecule, which is the receptor binding the domain of influenza A virus (H3N2) complexed with its receptor Hemagglutinin. We used MGL Tools 1.5.6 and Autodock Tool (ADT) to prepare all proteins and ligands. ADT was used to calculate the binding free energy and inhibition constant of the optimal docking complex of the above proteins. **Figure 4** showed the interaction of three unproven drugs predicted by DRMNN with important residues on their receptor Hemagglutinin. The negative combination free energy further indicates the stability of the complex (**Table 3**). This evidence all showed that the drugs predicted by DRMNN are effective in suppressing influenza A viruses.

DISCUSSION

Influenza A viruses have always been among the most important viruses harmful to human health. They can cause acute respiratory infection that is harmful to human health and is one of the main causes of death. To prevent and treat influenza viruses, vaccines or anti-influenza chemicals are usually used. However, traditional vaccines may not easily prevent and treat influenza outbreaks caused by new viruses, while the development of new drugs will require longer time and higher economic costs. Therefore, strategies to find effective drugs among existing drugs can greatly reduce time and cost. In this paper, we propose a method of reuse of antiviral drugs based on the minimum nuclear specification. The method mainly uses data collected from the literature on viruses and drugs, combines the similarity of drug chemical molecules with the similarity of virus gene sequence, and uses DRMNN to obtain the drug most likely to treat influenza A virus (H3N2). After global fivefold CV, DRMNN showed better performance than other methods in determining the treatment of influenza A virus (H3N2). Finally, we obtained the top 10 potential drugs, of which six have been shown to be effective against influenza A virus (H3N2). This method saves the experimental cost and time and provides a powerful reference for preventing and treating influenza A virus.

Although DRMNN has been shown to offer many potential drugs for influenza A virus (H3N2) that may have therapeutic effects, some limitations remain in this study. The DRMNN database contains 13,563 drug entries, and there are thousands of antivirals for broad-spectrum drugs and thousands of viruses for NCBI. Due to the limited amount of data, there are still some biases in the potential drugs we obtain. Therefore, determining how to select effective data to establish greater data integration is an important goal for future research.

Finally, we focused our analyses on influenza in this study. However, it is clear that our method could also be applied to other viruses, for example SARS-CoV-2. The outbreak of SARS-CoV-2 has become a serious pandemic and has caused the deaths of hundreds of thousands of people. Currently, there is no confirmed drug effective against this virus. In the future, we will check the drugs predicted by our method for use against this virus and validate their efficacy through both protein docking and wet-lab experiments.

[1] http://autodock.scripps.edu

[2] www.rscb.org

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

FM conceived, designed, and managed the study. HL and LZ performed the experiments and drafted the manuscript. XM, ML, JL, and WL provided computational support and technical assistance. XM, MG, and LW reviewed the manuscript. All authors approved the final manuscript.

REFERENCES

Barr, I. G., Hurt, A. C., Iannello, P., Tomasov, C., Deed, N., and Komadina, N. (2007). Increased adamantane resistance in influenza A(H3) viruses in Australia and neighbouring countries in 2005. *Antiviral Res.* 73, 112–117. doi: 10.1016/j.antiviral.2006.08.002

Burch, J., Corbett, M., Stock, C., Nicholson, K., Elliot, A. J., Duffy, S., et al. (2009). Prescription of anti-influenza drugs for healthy adults: a systematic review and meta-analysis. *Lancet Infect. Dis.* 9, 537–545. doi: 10.1016/s1473-3099(09)70199-9

CDC. (2009). People at High Risk of Developing Flu-Related Complications at High Risk of Developing Flu-Related Complications. Atlanta, GA: CDC.

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

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018). Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Cheng, F., Murray, J. L., and Rubin, D. H. (2016). Drug repurposing: new treatments for Zika virus infection? *Trends Mol. Med.* 22, 919–921. doi: 10.1016/j.molmed.2016.09.006

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. *Lancet Infect. Dis.* 5, 718–725. doi: 10.1016/s1473-3099(05)70270-x

Elazar, M., Liu, M., McKenna, S. A., Liu, P., Gehrig, E. A., Puglisi, J. D., et al. (2009). The anti-hepatitis C agent nitazoxanide induces phosphorylation of eukaryotic initiation factor 2α via protein kinase activated by double-stranded RNA activation. *Gastroenterology* 137, 1827–1835. doi: 10.1053/j.gastro.2009.07.056

Glezen, W. P. (2006). Influenza control. *N. Engl. J. Med.* 355, 79–81.

Gu, C., Liao, B., Li, X., and Li, K. (2016). Network consistency projection for human miRNA-disease associations inference. *Sci. Rep.* 6:36054.

Hauge, S. H., Dudman, S., Borgen, K., Lackenby, A., and Hungnes, O. (2009). Oseltamivir-resistant influenza viruses A (H1N1), Norway, 2007-08. *Emerg. Infect. Dis.* 15, 155–162. doi: 10.3201/eid1502.081031

Hurt, A. C., Ernest, J., Deng, Y. M., Iannello, P., Besselaar, T. G., Birch, C., et al. (2009a). Emergence and spread of oseltamivir-resistant A(H1N1) influenza viruses in Oceania, South East Asia and South Africa. *Antivir. Res.* 83, 90–93. doi: 10.1016/j.antiviral.2009.03.003

Hurt, A. C., Holien, J. K., Parker, N., Kelso, A., and Barr, I. G. (2009b). Zanamivir-resistant influenza viruses with a novel neuraminidase mutation. *J. Virol.* 83, 10366–10373. doi: 10.1128/jvi.01200-09

Jamieson, A. M., Pasman, L., Yu, S., Gamradt, P., Homer, R. J., Decker, T., et al. (2013). Role of tissue protection in lethal respiratory viral-bacterial coinfection. *Science* 340, 1230–1234. doi: 10.1126/science.1233632

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511–518. doi: 10.1093/nar/gki198

Korba, B. E., Montero, A. B., Farrar, K., Gaye, K., Mukerjee, S., Ayers, M. S., et al. (2008). Nitazoxanide, tizoxanide and other thiazolides are potent inhibitors of hepatitis B virus and hepatitis C virus replication. *Antivir. Res.* 77, 56–63. doi: 10.1016/j.antiviral.2007.08.005

La Frazia, S., Ciucci, A., Arnoldi, F., Coira, M., Gianferretti, P., Angelini, M., et al. (2013). Thiazolides, a new class of antiviral agents effective against rotavirus infection, target viral morphogenesis, inhibiting viroplasm formation. *J. Virol.* 87, 11096–11106. doi: 10.1128/jvi.01213-13

Leneva, I. A., Falynskova, I. N., Makhmudova, N. R., Poromov, A. A., Yatsyshina, S. B., and Maleev, V. V. (2019). Umifenovir susceptibility monitoring and characterization of influenza viruses isolated during ARBITR clinical study. *J. Med. Virol.* 91, 588–597. doi: 10.1002/jmv.25358

Mccullers, J. A. (2006). Insights into the interaction between influenza virus and Pneumococcus. *Clin. Microbiol. Rev.* 19, 571–582. doi: 10.1128/cmr.00058-05

Mccullers, J. A. (2014). The co-pathogenesis of influenza viruses with bacteria in the lung. *Nat. Rev. Microbiol.* 12, 252–262. doi: 10.1038/nrmicro3231

Ooi, E. E., Chew, J. S., Loh, J. P., and Chua, R. C. (2006). In vitro inhibition of human influenza A virus replication by chloroquine. *Virol. J.* 3, 39–39.

Rohloff, J. C., Kent, K. M., Postich, M. J., Becker, M. W., Chapman, H. H., Kelly, D. E., et al. (1998). Practical total synthesis of the anti-influenza drug GS4104. *J. Org. Chem.* 63, 4545–4550.

Rossignol, J., Abu-Zekry, M., Hussein, A., and Santoro, M. G. (2006). Effect of nitazoxanide for treatment of severe rotavirus diarrhoea: randomised double-blind placebo-controlled trial. *Lancet* 368, 124–129. doi: 10.1016/s0140-6736(06)68852-1

Rossignol, J. F., Elfert, A., El-Gohary, Y., and Keeffe, E. B. (2009a). Improved virologic response in chronic hepatitis C genotype 4 treated with nitazoxanide, peginterferon, and ribavirin. *Gastroenterology* 136, 856–862. doi: 10.1053/j.gastro.2008.11.037

Rossignol, J. F., La Frazia, S., Chiappa, L., Ciucci, A., and Santoro, M. G. (2009b). Thiazolides, a new class of anti-influenza molecules targeting viral hemagglutinin at the post-translational level. *J. Biol. Chem.* 284, 29798–29808. doi: 10.1074/jbc.m109.029470

Saito, R., Li, D., and Suzuki, H. (2007). Amantadine-resistant influenza A (H3N2) virus in Japan, 2005-2006. *N. Engl. J. Med.* 356, 312–313. doi: 10.1056/nejmc062989

Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., et al. (2017). A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* 16, 19–34. doi: 10.1038/nrd.2016.230

Sheu, T. G., Deyde, V. M., Okomo-Adhiambo, M., Garten, R. J., Xu, X., Bright, R. A., et al. (2008). Surveillance for neuraminidase inhibitor resistance among human influenza A and B viruses circulating worldwide from 2004 to 2008. *Antimicrob. Agents Chemother.* 52, 3284–3292. doi: 10.1128/aac.00555-08

Sleeman, K., Mishin, V. P., Guo, Z., Garten, R. J., Balish, A., Fry, A. M., et al. (2014). Antiviral susceptibility of variant influenza A(H3N2)v viruses isolated in the United States from 2011 to 2013. *Antimicrob. Agents Chemother.* 58, 2045–2051. doi: 10.1128/aac.02556-13

Taubenberger, J. K., and Kash, J. C. (2010). Influenza virus evolution, host adaptation and pandemic formation. *Cell Host Microbe* 7, 440–451. doi: 10.1016/j.chom.2010.05.009

Tosh, P. K., and Poland, G. A. (2008). Emerging vaccines for influenza. *Expert Opin. Emerg. Drugs* 13, 21–40. doi: 10.1517/14728214.13.1.21

Webster, R. G., Bean, W. J., Gorman, O. T., Chambers, T. M., and Kawaoka, Y. (1992). Evolution and ecology of influenza A viruses. *Microbiol. Res.* 56, 152–179.

White, A. C. (2004). Nitazoxanide: a new broad spectrum antiparasitic agent. *Expert Rev. Anti Infect. Ther.* 2, 43–49. doi: 10.1586/14787210.2.1.43

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi: 10.1093/bioinformatics/btaa109

Yang, M., Luo, H., Li, Y., and Wang, J. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 35, i455–i463.

Zou, S., Zhang, J., and Zhang, Z. (2018). Novel human microbe-disease associations inference based on network consistency projection. *Sci. Rep.* 8:8034.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 January 2021
doi: 10.3389/fgene.2020.596174

Identification of Key Modules and Hub Genes of Annulus Fibrosus in Intervertebral Disc Degeneration

Hantao Wang[1,2], *Wenhui Liu*[3], *Bo Yu*[4], *Xiaosheng Yu*[1] *and Bin Chen*[1]*

[1] Department of Spine Surgery, School of Medicine, Renji Hospital, Shanghai Jiao Tong University, Shanghai, China, [2] Department of Orthopedics, School of Medicine, Renji Hospital, Shanghai Jiao Tong University, Shanghai, China, [3] Plastic & Reconstructive Surgery of the First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [4] Department of Medicine, Lincoln Medical Center, Bronx, NY, United States

Background: Intervertebral disc degeneration impairs the quality of patients lives. Even though there has been development of many therapeutic strategies, most of them remain unsatisfactory due to the limited understanding of the mechanisms that underlie the intervertebral disc degeneration.

Questions/purposes: This study is meant to identify the key modules and hub genes related to the annulus fibrosus in intervertebral disc degeneration (IDD) through: (1) constructing a weighted gene co-expression network; (2) identifying key modules and hub genes; (3) verifying the relationships of key modules and hub genes with IDD; and (4) confirming the expression pattern of hub genes in clinical samples.

Methods: The Gene Expression Omnibus provided 24 sets of annulus fibrosus microarray data. Differentially expressed genes between the annulus fibrosus of degenerative and non-degenerative intervertebral disc samples have gone through the Gene Ontology (GO) and pathway analysis. The construction of a gene network and classification of genes into different modules were conducted through performing Weighted Gene Co-expression Network Analysis. The identification of modules and hub genes that were most related to intervertebral disc degeneration was proceeded. In order to verify the relationships of the module and hub genes with intervertebral disc degeneration, Ingenuity Pathway Analysis was operated. Clinical samples were adopted to help verify the hub gene expression profile.

Results: One thousand one hundred ninety differentially expressed genes were identified. Terms and pathways associated with intervertebral disc degeneration were presented by GO and pathway analysis. The construction of a Weighted Gene Coexpression Network was completed and clustering differentially expressed genes into four modules was also achieved. The module with the lowest *P*-value and the highest absolute correlation coefficient was selected and its relationship with intervertebral disc degeneration was confirmed by Ingenuity Pathway Analysis. The identification of hub genes and the confirmation of their expression profile were also realized.

Conclusions: This study generated a comprehensive overview of the gene networks underlying annulus fibrosus in intervertebral disc degeneration.

OPEN ACCESS

Edited by:
Tao Huang,
Chinese Academy of Sciences (CAS), China

Reviewed by:
Ruidong Zhang,
Inner Mongolia Agricultural University, China
Jun Jiang,
Fudan University, China

***Correspondence:**
Bin Chen
chnbn2003@126.com

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

Received: *18 August 2020*
Accepted: *17 December 2020*
Published: *27 January 2021*

Citation:
Wang H, Liu W, Yu B, Yu X and Chen B (2021) Identification of Key Modules and Hub Genes of Annulus Fibrosus in Intervertebral Disc Degeneration. Front. Genet. 11:596174. doi: 10.3389/fgene.2020.596174

Clinical Relevance: Modules and hub genes identified in this study are highly associated with intervertebral disc degeneration, and may serve as potential therapeutic targets for intervertebral disc degeneration.

Keywords: intervertebral disc degeneration (IDD), weighted gene co-expression network, gene ontology, therapeutic target, annulus fibrosus

INTRODUCTION

Low back pain (LBP), one of the most common musculoskeletal diseases, is estimated that up to 84% of the population suffer from LBP at least once in their life (Walker, 2000; Shen et al., 2018). Intervertebral disc degeneration (IDD), resulting from degenerative and inflammatory changes, promote neurovascular ingrowth into the disc and accounts for between 26 and 42% of LBP (Luoma et al., 2000; Kadow et al., 2015). Current approaches of the treatment of IDD include conservative therapies such as physiotherapy, anti-inflammatory medication, and surgical interventions including spinal fusion and disc arthroplasty. However, the clinical results of these treatments are suboptimal and a comprehensive understanding of the biological causes of IDD is required to develop improved therapies (Rao and Cao, 2014; Kadow et al., 2015).

Several studies have been conducted using microarray to investigate biomarkers and key pathways in IDD. This information not only enhances our understanding of IDD, but also highlights potential therapeutic targets. The Wnt pathway was found to be downregulated in early IDD by Smolders et al. (2013). Gruber et al. identified differentially expressed genes associated with pain, nerves and neurotrophin, and mitochondrial dysfunction, while several aberrantly expressed long non-coding RNAs (lncRNAs) were identified by Gruber et al. (2011, 2012), and Wan et al. (2014). Despite important advances in the clarification of the potential pathogenesis of IDD are achieved using high throughput microarray analysis, this established method has failed to generate a comprehensive overview of the gene network of IDD. A common practice in microarray data analysis is to apply a double filter to differentially expressed genes (DEGs) based on fold changes in expression and *t*-test *P*-values in comparisons between different groups (Zhang and Cao, 2009). However, lists of DEGs fail to elucidate the interactions among genes (Wu et al., 2013). Furthermore, downstream genes usually have greater variance resulting in their higher ranking than upstream genes, which is the key driver of disease (Naylor et al., 2010).

A number of co-expression network algorithms have been developed to investigate interactions among genes, including Weighted Gene Co-expression Network Analysis (WGCNA) (Serin et al., 2016). This algorithm is broadly applied in various fields, including lncRNA profiling of IDD (Langfelder and Horvath, 2008; Chen et al., 2015). WGCNA can be applied to high-throughput microarray or RNA-seq data sets to find clusters (modules) of highly correlated genes, using modular intrinsic genes or in-model central genes to pair these clusters Summarize, correlate modules with each other and with external sample traits, and use them to calculate module membership metrics (Pei et al., 2017). This method has also recently been applied to proteomics and metabolomics data analysis (DiLeo et al., 2011). WGCNA can be used to identify candidate biomarkers or therapeutic targets, and has been used in a variety of human cancers, including colon cancer, uveal melanoma, glioblastoma, liver cancer, and osteosarcoma (Langfelder and Horvath, 2008). This study saw a focus on the gene co-expression network of annulus fibrosus and the principal cause of discogenic symptoms (Kazezian et al., 2015). Integrated bioinformatics methods, including WGCNA, were applied to generate a comprehensive overview of the gene network associated with IDD. Expression of some of the identified hub genes was verified using clinical samples. These hub genes might represent novel therapeutic targets for IDD.

MATERIALS AND METHODS

Date Acquisition and Clinical Samples

Under the accession number GSE70362, the download of the data series was accessed from Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/). Processed data of annulus fibrous were selected for this study. The filtration of DEGs of degenerative (grade III–V) vs. non-degenerative (grade I–II) intervertebral disc samples was achieved using a two-tailed *t*-test. Using DEGs for further analysis not only preserved variance in genetic background, but also reduced unrelated genetic noise. We collected the specific clinical characteristics of clinical samples from 10 patients with IDD, including gender, age, level, and pfirrmann grade.

Patients and Tissues

Ethics Review Board of Renji Hospital (number 2017-003) approved the study trials and the study was performed in accordance with the rules of the China Food and Drug Administration/Good Clinical Practice and the Declaration of Helsinki (2008) of the World Medical Association. All participants or their parents/legal guardians for patients aged under 18 years provided the written informed consent.

From patients with IDD in the Spine Group of Renji Hospital, degenerative intervertebral disc tissue samples were obtained. Patients with IDD combined with infections, tumors, or previous lumbar disc surgery were not included in this study. From patients with accidental fractures, non-degenerative specimens were collected. None of the patients in the non-degenerative group reported any previous lumbar pain. Based on the Pfirrmann grading system, the degenerative condition was evaluated by two independent observers using magnetic

resonance imaging (Pfirrmann et al., 2001). All the intervertebral disc specimens were collected within 1 h after disc excision, rinsed with phosphate-buffered saline and then stored in the RNAstore Reagent DP408-02 (Tiangen Biotech, Beijing, China) at 4°C.

Gene Ontology Analysis and Pathway Analysis

The application of gene ontology (GO) analysis to upregulated and downregulated genes were operated separately (Ashburner et al., 2000). According to Gene Ontology Consortium, GO classifies gene functioned in a species-independent way in line with three aspects: cellular component, molecular function and biological process. GO analysis was performed to determine the GO terms that were over- or under-represented in a given set of genes. GO analysis was performed using ClueGO to generate a visual representation of the enriched terms in a functionally grouped annotation network which reflected the relationships between enriched terms. The leading term in a group was the most significant (Bindea et al., 2009). The *p*-value of enrichment analysis should be <0.05.

By interrogating the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) was used to achieve pathway analysis (Huang et al., 2009; Kanehisa et al., 2017). The *p*-value of enrichment analysis should be < 0.05. The Ingenuity Pathway Analysis Database (IPA, www.Ingenuity.com) was also used for pathway analysis. KEGG pathway analysis is a topology-based approach which takes into account gene interactions whereas IPA is based on gene expression (Khatri et al., 2012). We used a combination of KEGG pathway analysis and IPA to generate more complete and accurate information about the identified DEGs. Up- and down-regulated genes were subjected to KEGG pathway analysis separately. The fold changes in gene expression of the up- and down-regulated genes subjected to IPA. The *p*-value of enrichment analysis should be <0.05.

Weighted Gene Co-expression Network Analysis

Through employing static programming language and environment R, Weighted Gene Co-expression Network Analysis was conducted with WGCNA package (Jiang et al., 2014; Core R, 2015). Only DEGs were included in the WGCNA workflow to minimize noise and reduce the computing burden without causing major information loss (Ghazalpour et al., 2006). The adjacency matrix was calculated based on pairwise Pearson correlation coefficients. WGCNA incorporated the concept that genes interactions occurred following a scale-free distribution pattern (Barabasi, 2009). The *pickSoftThreshold* function was applied to fit the scale-free criterion. Topologic overlap measures, which were a robust measure of networks, were calculated pairwise within the adjacency matrix. The dynamic tree cutting algorithm, an unsupervised hierarchical clustering method, was adopted for clustering with input of topologic overlap measures (Langfelder et al., 2008). In this study, the soft threshold (power) was 4. We used the default parameters in WGCNA algorithm, the maximum size of module was 500, and the minimum size was 30.

Modules can be explained as branches of the clustering tree. In network terminology, a module refers to a group of genes that share similar connection patterns with all other genes outside that group and there are, generally speaking, similar functions existing in genes in the same module (Zhang and Horvath, 2005). The calculation of the main component of module, a module eigengene, was then conducted to summarize the gene expression profiles in the module. In order to identify the modules that were most related to IDD for further analysis, the calculation of correlations between module eigengenes and the degenerative status of samples was operated.

In a scale-free network, hub genes of modules are the most interconnected genes and they serve as the backbones of this network (de Jong et al., 2012). Hub genes in disease-related modules, such as hub lncRNA in IDD, are generally biologically and clinically meaningful (Jiang et al., 2014; Lee et al., 2014; Chen et al., 2015; Wang et al., 2015). Hub genes were determined through ranking intra-modular connectivity and correlation with eigengenes in selected module. Gene co-expression networks of all DEGs and hub genes were visualized using Cytoscape (Shannon et al., 2003).

Ingenuity Pathway Analysis of Selected Modules and Hub Genes

Genes in selected modules were subjected IPA to evaluate their relationship with IDD. GO analysis and KEGG pathway analysis were commonly performed. However, these types of analysis considered only the number of genes in a given set and ignored any values related to genes (Khatri et al., 2012). We undertook a close examination of selected modules including both up- and down-regulated genes. Thus, GO and KEGG analyses of heterogeneous data such as ours were inappropriate. Instead, we performed IPA, which also took into account gene expression levels. This workflow has been widely adopted in many other weighted gene co-expression network analyses (Naylor et al., 2010; Malki et al., 2013). We also adopted the Disease and Biofunction module of IPA which was similar to Go Analysis.

Validation of Hub Gene Expression

To validate the expression pattern of some hub genes, quantitative real-time PCR (qRT-PCR) was conducted. TRIzol reagent (Invitrogen, Carlsbad, CA, USA) was employed to extract RNA based on the instructions of the manufacturer, and qRT-PCR assays were conducted through adopting the ViiA7 Real-Time PCR System (Applied Biosystems, CA, USA) with a thermal profile comprising one min at 95°C for polymerase activation, followed by 40 cycles of 95°C for 15 s and 60°C for 30 s. Expression of target genes was normalized to ß-actin as the endogenous control. For statistical analysis, the calculation of gene expression was processed following the $2^{-\Delta\Delta Ct}$ method, and relative expression data were log2 transformed (Livak and Schmittgen, 2002). The list of sequences of primers used for qRT-PCR amplification was presented in **Supplementary Table 1**.

Statistical Analysis

All quantitative data were represented as mean ± SD. In the mRNA expression experiments (SPSS Statistics Version 22.0; IBM Corp, Armonk, NY, USA), in order to compare control groups with the IDD group, student's *t*-test was operated. Unless otherwise stated, when *P*-values were below 0.05, differences were taken as statistically significant.

RESULTS

Clinical Characteristics of Samples

The specific clinical traits of all samples in GSE70362 were provided in **Supplementary Table 2**. From levels T12–L1 to L4–L5, five pairs of non-degenerative and degenerative annulus fibrous samples were collected for qRT-PCR analysis to confirm hub gene expression. Specific clinical traits of these sample were provided in **Supplementary Table 3**. In the degenerative group, the average age was 48.0 years (range, 33–61 years) with Pfirrmann Grade III–V disc. In the non-degenerative group, the average age was 31.8 years (range, 16–52 years) with Pfirrmann Grade I–II disc.

Gene Ontology Analysis and Pathway Analysis

Altogether 2,636 probes were identified as differentially expressed in comparisons of degenerative and non-degenerative annulus fibrosus tissue samples. According to the annotation file, 1,190 probes were mapped to known genes (464 upregulated and 726 downregulated).

An overview of the GO analysis was presented in **Figure 1** and specific results for up- and downregulated genes are provided in **Supplementary Figures 1, 2**, respectively. For both

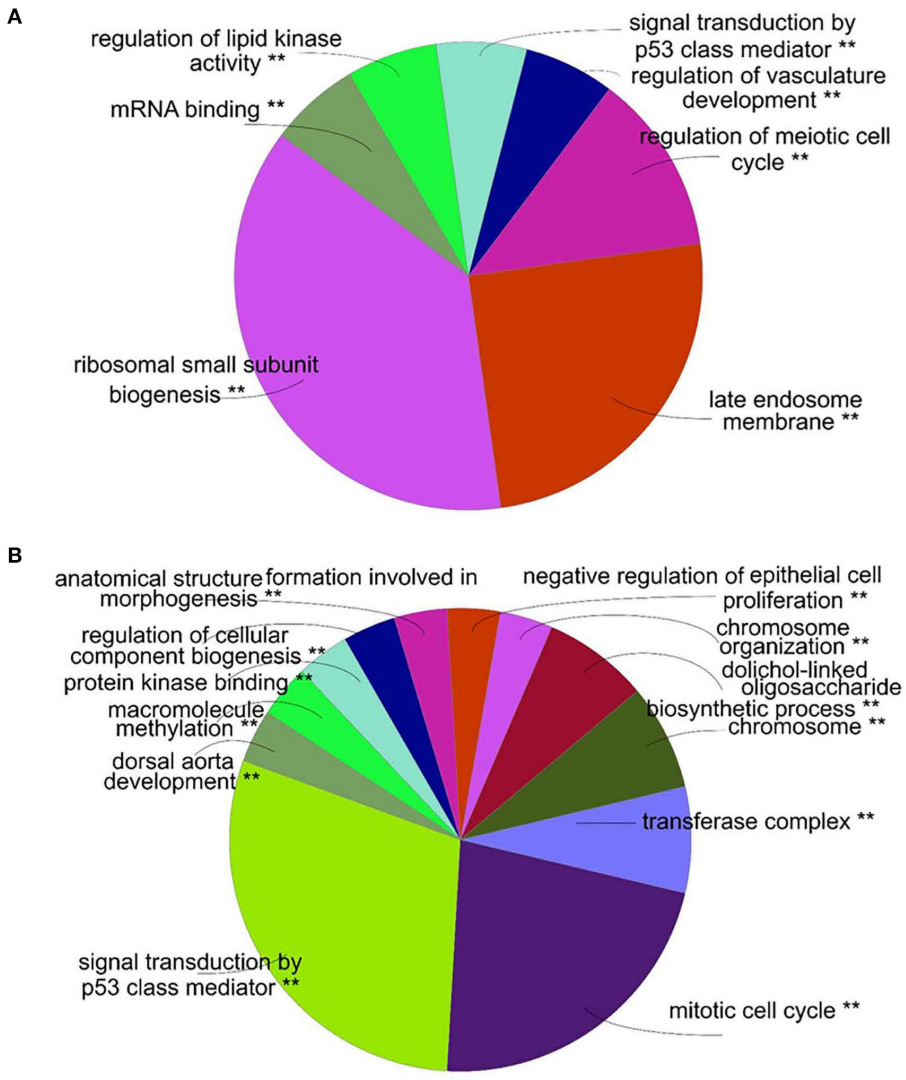

FIGURE 1 | Gene ontology (GO) analysis of differentially expressed genes. **(A)**. GO analysis of upregulated genes by ClueGO. According to the most significant gene in the group, each section of the overview pie-chart represents a group of GO terms with section names allocated. The correlation exists between the size of each section and the number of genes within the group. **(B)**. GO analysis of downregulated genes by ClueGO. $^{**}P < 0.01$.

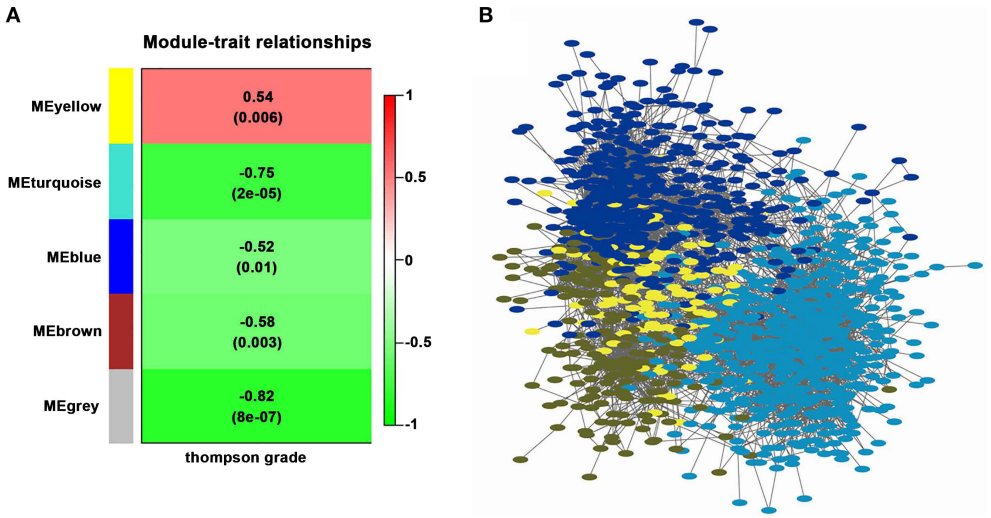

FIGURE 2 | Weighted gene co-expression network of annulus fibrosus in IDD. **(A)**. Correlation between module and Thompson Grade. Each cell represents a module with the corresponding correlation (above) and *P*-value (below). The names of the modules are presented to the left of cells and the bar on the right represents correlation values. **(B)**. Visual representation of Weighted Gene Co-expression Network. The node represents differentially expressed gene, the edge represents the interaction between two genes.

up- and down-regulated genes, signal transduction by a p53 class mediator was enriched, indicating the involvement of apoptosis in IDD. For upregulated genes, regulation of vasculature development was enriched, which was consistent with the vascularization associated with IDD (Freemont et al., 1997).

Results of pathway analysis were presented in **Table 1**. A long list of activated pathways has been generated by IPA and only the IDD-related pathways were presented. The complete IPA results were presented in **Supplementary Table 4**. The TNF signaling pathway, which had a well-established association with IDD, was found to be activated in KEGG pathway analysis of upregulated genes and TNFR1 and TNFR2 signaling activation was identified by IPA (Risbud and Shapiro, 2014). TGF-β signaling was also identified by IPA. Other cytokine signaling pathways including B cell activating factor signaling, IL-1 signaling and IL-6 were also identified (**Supplementary Table 4**) confirming the role of inflammation in IDD. Mismatch repair signaling was identified in both KEGG pathway analysis and IPA. Furthermore, apoptosis signaling activation was identified, thus confirming the results of GO analysis. Axonal guidance signaling was also identified, which highlighted the role of neural ingrowth in IDD (Freemont et al., 1997; Kepler et al., 2013).

Weighted Gene Co-expression Network Analysis

Four modules were generated by WGCNA; these modules were identified by different colors and genes that could not been classified into any modules were shown in gray. In WGCNA's algorithm, some non-clustering genes will be put into a single module, which will be uniformly called "gray." Correlation between modules and Thompson Grade were shown in **Figure 2A**. The module with the lowest *P*-value (shown in turquoise) and the highest absolute correlation coefficient was

TABLE 1 | Pathway analysis of differentially expressed genes.

Term	*P*-value
KEEG UP-REGULATED GENES	
Progesterone-mediated oocyte maturation	0.00561
HTLV-I infection	0.00631
TNF signaling pathway	0.012669
Pathways in cancer	0.013666
Regulation of actin cytoskeleton	0.020803
ErbB signaling pathway	0.027356
Prostate cancer	0.028379
Mismatch repair	0.036276
Acute myeloid leukemia	0.037105
Hepatitis B	0.041927
Choline metabolism in cancer	0.043744
Colorectal cancer	0.047879
KEEG DOWN-REGULATED GENES	
DNA replication	0.034
Cell cycle	0.036
IPA	
TNFR2 signaling	2.75E-05
Mismatch repair in eukaryotes	8.91E-05
Induction of apoptosis by HIV1	9.77E-05
Apoptosis signaling	0.001995
B cell activating factor signaling	0.008913
Axonal guidance signaling	0.009333
HGF signaling	0.01349
Acute phase response signaling	0.016218
TNFR1 signaling	0.017783
TGF-β signaling	0.022387

The analysis of up- and downregulated genes by Kyoto Encyclopedia of Genes and Genomes (KEGG) is shown separately (only pathways with P < 0.05 are shown). Ingenuity pathway analysis (IPA) analysis results are customized to display related pathways only; complete results are shown in ***Supplementary Table 4****.*

considered to be the module which is most related to IDD and selected for further analysis. A visual representation of the whole weighted gene co-expression network was shown in **Figure 2B**. Nodes represent genes and node color indicated module membership. Correlation existed between edges between nodes and topologic overlaps (analogous to correlation), genes and small distances indicate high correlation. There was a tendency for genes within the same module to stay close to each other in the weighted gene co-expression network by visual inspection of **Figure 2B**. The complete results of WGCNA were provided in **Supplementary Table 5**.

The gene is represented by each node while the module membership is indicated by node color. Correlation exists between edges between nodes and topologic overlaps (analogous to correlation) between genes and small distances indicate high correlation. The purpose of this research was to find hub module, and the turquoise module was highly correlated with the disease. Therefore, we focused on the analysis of this module.

By ranking intra-modular connectivity and correlation with the module eigengene, hub genes in the turquoise module were identified. The top hub genes in the turquoise module were represented in **Figure 3**. To be clarified, only the top 30 hub genes were included. Hub genes were represented by nodes and correlation exists between node size and the intra-modular connectivity of the gene. The selection criteria of hub genes was the top 10 genes with the highest connectivity in the co-expression network. DSE, IL17RD, DUSP18, ROBO3, BANK1, MRC2, LGALSL, TFPI, GAP43, and HYAL1, the top ten hub genes, were shown in darker colors.

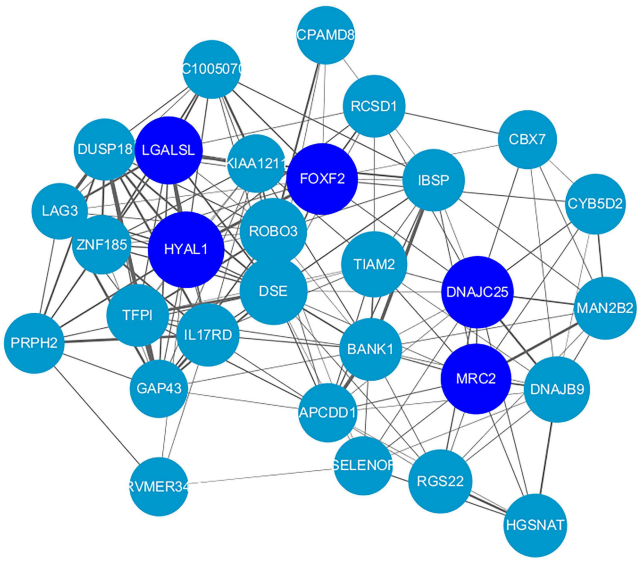

FIGURE 3 | Top 30 hub genes in the turquoise module. Hub genes are represented by nodes and correlation exists between node size and the intra-modular connectivity of the gene. In the center of this network, the top 10 hub genes are located and they are shown in darker colors.

Ingenuity Pathway Analysis of Turquoise Module

To evaluate the relationship between the turquoise module and IDD, IPA was performed. As shown in **Figure 4**, "apoptosis signaling," "factors promoting cardiogenesis invertebrates," "neuregulin signaling," "B cell receptor signaling" "B cell activating factor signaling" and "natural killer cell signaling" were identified by IPA. These pathways highlighted the role of apoptosis, neural ingrowth, vascularization, and inflammatory cytokines in IDD. Highly related pathways were also identified in the Diseases and Bio Functions module (**Supplementary Figure 3**).

Confirmation of Hub Gene Expression

To validate the outcomes of the turquoise module analysis and to identify key genes involved in IDD, RT-PCR was employed to analyze the relative expression levels of 10 hub genes. Comparisons of degenerative disc and matched non-degenerative disc samples revealed significant downregulation of eight genes (IL17RD, DUSP18, GAP43, and HYAL1 [$P < 0.001$]; ROBO3, BANK1, and MRC2 [$P < 0.01$]; LGALSL [$P < 0.05$]) and significant upregulation of two genes (TFPI [$P < 0.001$]; DSE [$P < 0.01$]). This expression profile was consistent with the microarray data (**Figure 5**). To make data from PCR and microarray comparable, expression values of each hub genes were transformed by dividing the average value in whole expression cohort from corresponding data source.

DISCUSSION

This study applied integrated bioinformatics approaches to identify variations in gene expression related to annulus fibrous in degenerative and non-degenerative intervertebral disc tissues. We generated a complete overview of the gene networks to highlight gene modules and hub genes highly related to IDD. The biological and clinical importance of hub genes in weighted gene co-expression networks have been widely reported (Chen et al., 2015). This study identifies the hub genes which may be of importance to the pathogenesis of IDD. By applying this novel method of analysis, the present study not only updates our perspective on the pathogenesis of IDD, but also highlights some hub genes which have the potential to be IDD biomarkers and treatment targets.

GO and pathway analyses revealed differences in annulus fibrous associated with degenerative and non-degenerative intervertebral disc tissue. Some apoptosis, neural ingrowth, vascularization, and inflammation related terms and pathways were identified that were consistent with well-established molecular mechanisms of IDD (Kepler et al., 2013). The pathogenesis of IDD includes cellular oxidative stress, mitochondrial dysfunction and apoptosis (Kang et al., 2019). Endplate chondrocyte apoptosis is an important cause of the pathogenesis of cartilage endplate (CEP) degeneration, leading to the occurrence and development of intervertebral disc degeneration (IDD) (Wu et al., 2010; Li et al., 2014).

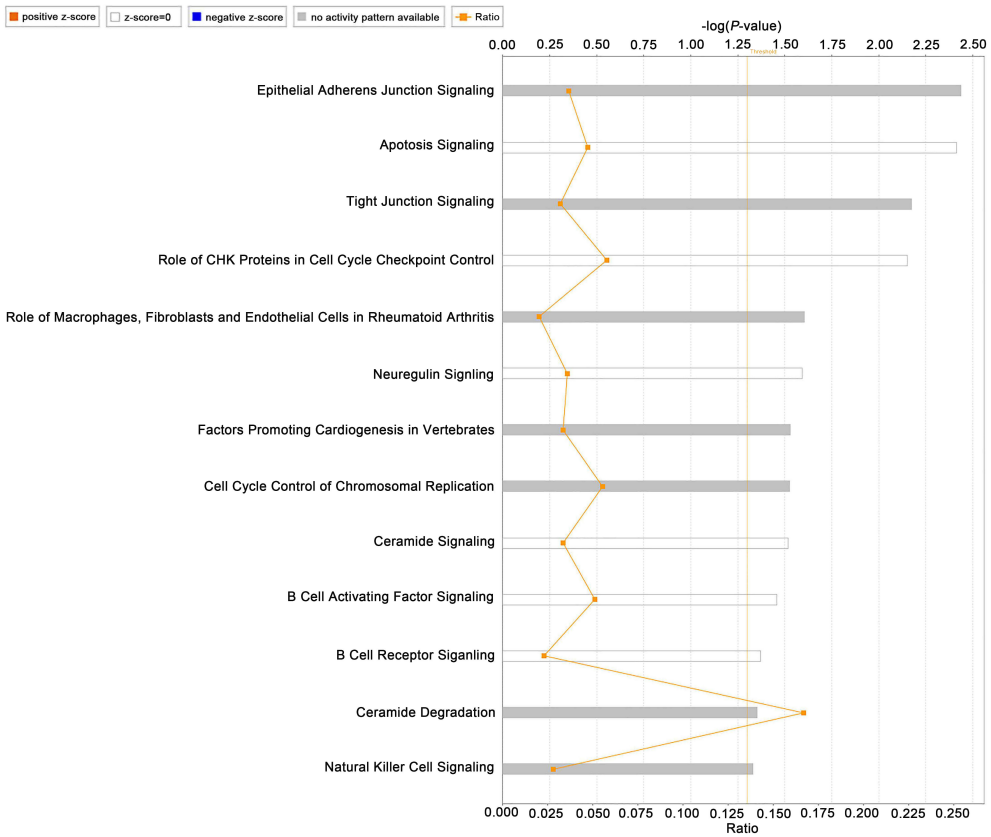

FIGURE 4 | Ingenuity pathway analysis of the turquoise module. Left part shows pathway names and bars represent the corresponding pathway's –log (*P*-value).

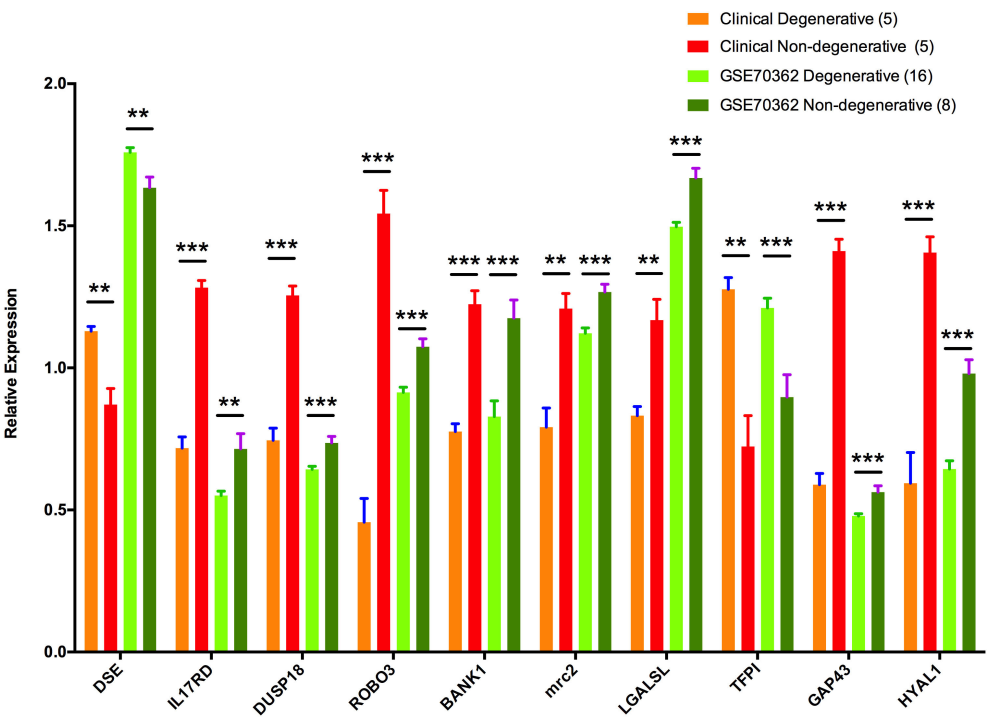

FIGURE 5 | Relative expression levels of 10 hub genes. Expressions of genes in GSE70362 and clinical samples. $^{**}0.001 < P < 0.01$; and $^{***}P < 0.001$.

Nucleus pulposus (NPC) apoptosis is the main factor of IDD. Nucleus pulposus (NP) cell apoptosis is a classic cell characteristic in the IDD process (Xianzhou and Cunxin, 2018). The vascularization of the intervertebral disc is generally considered to be a pathological feature of IDD (Johnson et al., 2007). As IDD progresses, intervertebral disc tend to be increasingly vascularized through angiogenesis. Recent evidence suggests that in addition to abnormal and excessive mechanical loads, inflammation may be a key driver of IDD and low back pain (Sharma, 2018). A study by GSE70362 has identified various dysfunctional cell functions, including cell proliferation and inflammation, and similar findings have been found in this study (Kazezian et al., 2015). Human T lymphovirus type I (HTLV-I) is the inducer of adult T-cell leukemia/lymphoma and HTLV-1-related myelopathy (Sherman et al., 1993). HTLV-I can cause chronic infections that cannot be cured or neutralized by vaccines. Due to HTLV-1 infection, the overall risk of death increases. The research of GSE70362 found that the most important classical pathway induced in degeneration fibrosus was the interferon pathway (Kazezian et al., 2015). Other famous pathways including TNF and TGF-β signaling were also determined in this study (Freemont, 2009). It is well known that the tumor necrosis factor TNF pathway affects the survival of cancer patients (Yi et al., 2018). TNF signal responds to cellular stress and inflammation signals, activates pro-apoptotic pathways and cytokine cascades (Chau et al., 2005). Transforming growth factor-beta (TGF-β) is a cytokine necessary to induce fibrosis and activate cancer stroma (Busch et al., 2015; Chen et al., 2019). The TGF-β signaling pathway plays an important role in many biological processes, including cell growth, differentiation, apoptosis, migration, and the occurrence and development of cancer (Waddell et al., 2004). Four gene modules were generated by WGCNA, and among them, the module that was most highly related to IDD was the turquoise module. Further analysis by IPA validated its tight correlation with IDD. By ranking intra-modular connectivity and correlation with the module eigengene, hub genes in the turquoise module were identified. Using this approach, DSE, IL17RD, DUSP18, ROBO3, BANK1, MRC2, LGALSL, TFPI, GAP43, and HYAL1 were identified as the top 10 hub genes. Hub gene expression profiles were confirmed by RT-PCR analysis.

Hub genes such as DSE, MRC2, and HYAL1 have a considerable effect on extracellular matrix metabolism, alterations in which are a major cause of IDD (Le Maitre et al., 2007). The DSE gene encodes dermatan sulfate epimerase, which regulates the biosynthesis of dermatan sulfate, an important element of the extracellular matrix. Furthermore, DSE-deficient mice have altered collagen structure (Maccarana et al., 2009). MRC2 is a versatile mediator of extracellular matrix metabolism and regulates not only collagen internalization, but also matrix metalloproteinase activity (Bailey et al., 2002; Messaritou et al., 2009; Madsen et al., 2011; Jurgensen et al., 2014). MRC2 also regulates TGF-β function (Caley et al., 2012). The HYAL1 gene encodes lysosomal hyaluronidase, which catalyzes the degradation of hyaluronan, which is one of the major glycosaminoglycans of the extracellular matrix (Lokeshwar et al., 2006). In addition, HYAL1 degenerates chondroitin sulfate, which is also an important component of extracellular matrix (Gushulak et al., 2012).

Neural ingrowth is reported to be involved in the pathogenesis of IDD, and our analysis indicates the involvement of hub genes ROBO3 and GAP43 in this process (Freemont, 2009; Kepler et al., 2013). ROBO3 is proposed to be involved in guiding neuronal axon growth, while GAP43 plays a well-established role in neuronal development and plasticity (Serin et al., 2016).

Inflammation is an essential participant in IDD and both IL17RD and BANK1 are important mediators of inflammatory reactions (Risbud and Shapiro, 2014; Molinos et al., 2015). IL17RD, which interacts with the IL-17 receptor to initiate IL-17 signaling, has been proposed as a therapeutic target in axial spondyloarthritis (Rong et al., 2009; Paine and Ritchlin, 2016). BANK1 mediates B cell signaling is involved in autoimmune disease such as systemic lupus erythematosus (Bernal-Quirós et al., 2013).

The hub gene TFPI may be a versatile participant in IDD based on its ability not only to regulate angiogenesis, but also to induce apoptosis (Hamuro et al., 1998; Amirkhosravi et al., 2007; Fu et al., 2008). Although these hub genes are well-characterized, the remaining two hub genes, DUSP18 and LGALSL, have not been researched extensively. Thus, the potential mechanisms by which these genes participate in the pathogenesis of IDD remain to be clarified. Nevertheless, the close relationships of the other eight hub genes with IDD indicates an important role for DUSP18 and LGALSL.

The major limitation of the present study is the isolated analysis of annulus fibrosus data. Although the formation of vascularized granulation tissue and innervation in annulus fibrosus are the principal causes of discogenic symptoms, the role of nucleus pulposus cannot be ignored (Livak and Schmittgen, 2002). Therefore, the integrated bioinformatics approaches adopted in this study will be used to explore how nucleus pulposus functions in IDD. This combined analysis of annulus fibrosus and nucleus pulposus data will provide a more integrated overview of the gene networks involved in IDD.

In conclusion, the present study was conducted using integrated bioinformatics approaches to generate a comprehensive overview of the gene network associated with annulus fibrosus in IDD. We identified 10 hub genes, DSE, IL17RD, DUSP18, ROBO3, BANK1, MRC2, LGALSL, TFPI, GAP43 and HYAL1, which updated our perspective on the pathogenesis of IDD, and could also serve as novel biomarkers and potential therapeutic targets. In addition, we also explore related signal transduction pathways and interaction networks. IDD is the main contributor to low back pain, which is the main cause of disability worldwide. This study provides clues to the molecular mechanism of IDD for future experimental studies. At the same time, this shows that bioinformatics methods can be used to identify potential targets for other human tumors.

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

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

BC and HW: conception and design. HW and WL: development of methodology. XY: sample collection. HW, BY, and WL: analysis and interpretation of data. HW, WL, and BC: writing, review, and/or revision of the manuscript. All authors contributed to the article and 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.2020.596174/full#supplementary-material

REFERENCES

Amirkhosravi, A., Meyer, T., Amaya, M., Davila, M., Mousa, S. A., Robson, T., et al. (2007). The role of tissue factor pathway inhibitor in tumor growth and metastasis. *Semin. Thromb. Hemost* 33, 643–652. doi: 10.1055/s-2007-991531

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. the gene ontology consortium. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Bailey, L., Wienke, D., Howard, M., Knäuper, V., Isacke, C. M., and Murphy, G. (2002). Investigation of the role of Endo180/urokinase-type plasminogen activator receptor-associated protein as a collagenase 3 (matrix metalloproteinase 13) receptor. *Biochem J.* 363, 67–72. doi: 10.1042/bj3630067

Barabasi, A. L. (2009). Scale-free networks: a decade and beyond. *Science* 325, 412–413. doi: 10.1126/science.1173299

Bernal-Quirós, M., Wu, Y. Y., Alarcón-Riquelme, M. E., and Castillejo-López, C. (2013). BANK1 and BLK act through phospholipase C gamma 2 in B-cell signaling. *PLoS ONE* 8:e59842. doi: 10.1371/journal.pone.0059842

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Busch, S., Acar, A., Magnusson, Y., Gregersson, P., Rydén, L., and Landberg, G. (2015). TGF-beta receptor type-2 expression in cancer-associated fibroblasts regulates breast cancer cell growth and survival and is a prognostic marker in pre-menopausal breast cancer. *Oncogene* 34, 27–38. doi: 10.1038/onc.2013.527

Caley, M. P., Kogianni, G., Adamarek, A., Gronau, J. H., Rodriguez-Teja, M., Fonseca, A. V., et al. (2012). TGFβ1-Endo180-dependent collagen deposition is dysregulated at the tumour-stromal interface in bone metastasis. *J. Pathol.* 226, 775–783. doi: 10.1002/path.3958

Chau, C. H., Clavijo, C. A., Deng, H. T., Zhang, Q., Kim, K. J., Qiu, Y., et al. (2005). Etk/Bmx mediates expression of stress-induced adaptive genes VEGF, PAI-1 and iNOS via multiple signaling cascades in different cell systems. *Am. J. Physiol. Cell Physiol.* 289:C444. doi: 10.1152/ajpcell.00410.2004

Chen, G., Yang, X., Wang, B., Cheng, Z., and Zhao, R. (2019). Human cytomegalovirus promotes the activation of TGF-β1 in human umbilical vein endothelial cells by MMP-2 after endothelial mesenchymal transition. *Adv. Clin. Exp. Med.* 28, 1441–1450. doi: 10.17219/acem/109199

Chen, Y., Ni, H. J., Zhao, Y. C., Chen, K., Li, M., Li, C., et al. (2015). Potential role of lncRNAs in contributing to pathogenesis of intervertebral disc degeneration based on microarray data. *Med. Sci. Monitor* 21, 3449–3458. doi: 10.12659/MSM.894638

Core R, R RDCT, Team R, and Team R. (2015). *A Language and Environment for Statistical, Computing. Computing* 1:12–21.

de Jong, S., Boks, M. P. M., Fuller, T. F., Strengman, E., Janson, E., de Kovel, C. G. F., et al. (2012). A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use, and enriched for brain-expressed, genes. *PLoS ONE* 7:e39498. doi: 10.1371/journal.pone.0039498

DiLeo, M. V., Strahan, G. D., den Bakker, M., and Hoekenga, O. A. (2011). Weighted correlation network analysis 2 (WGCNA) applied to the tomato fruit metabolome. *PLoS ONE* 6:e26683. doi: 10.1371/journal.pone.0026683

Freemont, A., Peacock, T., Goupille, P., Hoyland, J., O'Brien, J., and Jayson, M. (1997). Nerve ingrowth into diseased intervertebral disc in chronic back pain. *Lancet* 350, 178–181. doi: 10.1016/S0140-6736(97)02135-1

Freemont, A. J. (2009). The cellular pathobiology of the degenerate intervertebral disc and discogenic back pain. *Rheumatology* 48, 5–10. doi: 10.1093/rheumatology/ken396

Fu, Y., Zhang, Z., Zhang, G., Liu, Y., Cao, Y., Yu, J., et al. (2008). Adenovirus-mediated gene transfer of tissue factor pathway inhibitor induces apoptosis in vascular smooth muscle cells. *Apoptosis* 13, 634–640. doi: 10.1007/s10495-008-0199-4

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., et al. (2006). Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet.* 2:e130. doi: 10.1371/journal.pgen.0020130

Gruber, H. E., Hoelscher, G. L., Ingram, J. A., and Hanley, E. N. Jr. (2012). Genome-wide analysis of pain-, nerve- and neurotrophin -related gene expression in the degenerating human annulus. *Mol. Pain.* 8:63. doi: 10.1186/1744-8069-8-63

Gruber, H. E., Watts, J. A., Hoelscher, G. L., Bethea, S. F., Ingram, J. A., Zinchenko, N. S., et al. (2011). Mitochondrial gene expression in the human annulus: *in vivo* data from annulus cells and selectively harvested senescent annulus cells. *Spine J.* 11, 782–791. doi: 10.1016/j.spinee.2011.06.012

Gushulak, L., Hemming, R., Martin, D., Seyrantepe, V., Pshezhetsky, A., and Triggs-Raine, B. (2012). Hyaluronidase 1 and β-hexosaminidase have redundant functions in hyaluronan and chondroitin sulfate degradation. *J. Biol. Chem.* 287, 16689–16697. doi: 10.1074/jbc.M112.350447

Hamuro, T., Kamikubo, Y., Nakahara, Y., Miyamoto, S., and Funatsu, A. (1998). Human recombinant tissue factor pathway inhibitor induces apoptosis in cultured human endothelial cells. *FEBS Lett.* 421, 197–202. doi: 10.1016/S0014-5793(97)01559-7

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* 4, 44–57. doi: 10.1038/nprot.2008.211

Jiang, Z. L., Sun, J. W., Dong, H., Luo, O., Zheng, X. B., Obergfell, C., et al. (2014). Transcriptional profiles of bovine *in vivo* pre-implantation development. *BMC Genomics* 15:756. doi: 10.1186/1471-2164-15-756

Johnson, W. E. B., Patterson, A. M., Eisenstein, S. M., and Roberts, S. (2007). The presence of pleiotrophin in the human intervertebral disc is associated with increased vascularization: an immunohistologic study. *Spine* 32, 1295–1302. doi: 10.1097/BRS.0b013e31805b835d

Jurgensen, H. J., Johansson, K., Madsen, D. H., Porse, A., Melander, M. C., Sorensen, K. R., et al. (2014). Complex determinants in specific members of the mannose receptor family govern collagen endocytosis. *J. Biol. Chem.* 289, 7935–7947. doi: 10.1074/jbc.M113.512780

Kadow, T., Sowa, G., Vo, N., and Kang, J. D. (2015). Molecular basis of intervertebral disc degeneration, and herniations: what are the important translational questions? *Clin. Orthopaed. Relat. Res.* 473, 1903–1912. doi: 10.1007/s11999-014-3774-8

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D61. doi: 10.1093/nar/gkw1092

Kang, L., Xiang, Q., Zhan, S., Song, Y., Wang, K., Zhao, K., et al. (2019). Restoration of autophagic flux rescues oxidative damage and mitochondrial dysfunction to protect against intervertebral disc degeneration. *Oxid. Med. Cell. Longev.* 2019:7810320. doi: 10.1155/2019/7810320

Kazezian, Z., Gawri, R., Haglund, L., Ouellet, J., Mwale, F., Tarrant, F., et al. (2015). Gene expression profiling identifies interferon signalling molecules and IGFBP3 in human degenerative annulus fibrosus. *Sci. Rep.* 5:15662. doi: 10.1038/srep15662

Kepler, C. K., Ponnappan, R. K., Tannoury, C. A., Risbud, M. V., and Anderson, D. G. (2013). The molecular basis of intervertebral disc degeneration. *Spine J.* 13, 318–330. doi: 10.1016/j.spinee.2012.12.003

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

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, 719–720. doi: 10.1093/bioinformatics/btm563

Le Maitre, C. L., Pockert, A., Buttle, D. J., Freemont, A. J., and Hoyland, J. A. (2007). Matrix synthesis and degradation in human intervertebral disc degeneration. *Biochem. Soc. Trans*. 35, 652–655. doi: 10.1042/BST0350652

Lee, B., Mazar, J., Aftab, M. N., Qi, F., and Perera, R. J. (2014). Long noncoding RNAs as putative biomarkers for prostate cancer detection. *J. Mol. Diagnostics* 16, 615–626. doi: 10.1016/j.jmoldx.2014.06.009

Li, D., Zhu, B., Ding, L., Lu, W., Xu, G., and Wu, J. (2014). Role of the mitochondrial pathway in serum deprivation-induced apoptosis of rat endplate cells. *Biochem. Biophys. Res. Commun.* 452, 354–360. doi: 10.1016/j.bbrc.2014.08.054

Livak, K. J., and Schmittgen, T. D. (2002). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lokeshwar, V. B., Estrella, V., Lopez, L., Kramer, M., Gomez, P., Soloway, M. S., et al. (2006). HYAL1-v1, an alternatively spliced variant of HYAL1 hyaluronidase: a negative regulator of bladder cancer. *Cancer Res.* 66, 11219–11227. doi: 10.1158/0008-5472.CAN-06-1121

Luoma, K., Riihimäki, H., Luukkonen, R., Raininko, R., Viikari-Juntura, E., and Lamminen, A. (2000). Low back pain in relation to lumbar disc degeneration. *Spine* 25, 487–492. doi: 10.1097/00007632-200002150-00016

Maccarana, M., Kalamajski, S., Kongsgaard, M., Magnusson, S. P., Oldberg, A., and Malmstrom, A. (2009). Dermatan sulfate epimerase 1-Deficient mice have reduced, c.ontent, and changed distribution of iduronic acids in dermatan sulfate and an altered collagen structure in skin. *Mol. Cell. Biol.* 29, 5517–5528. doi: 10.1128/MCB.00430-09

Madsen, D. H., Ingvarsen, S., Jurgensen, H. J., Melander, M. C., Kjoller, L., Moyer, A., et al. (2011). The non-phagocytic route of collagen uptake: a distinct degradation pathway. *J. Biol. Chem.* 286, 26996–27010. doi: 10.1074/jbc.M110.208033

Malki, K., Tosto, M. G., Jumabhoy, I., Lourdusamy, A., Sluyter, F., Craig, I., et al. (2013). Integrative mouse and human mRNA studies using WGCNA nominates novel candidate genes involved in the pathogenesis of major depressive disorder. *Pharmacogenomics* 14, 1979–1990. doi: 10.2217/pgs.13.154

Messaritou, G., East, L., Roghi, C., Isacke, C. M., and Yarwood, H. (2009). Membrane type-1 matrix metalloproteinase activity is regulated by the endocytic collagen receptor Endo180. *J. Cell Sci.* 122, 4042–4048. doi: 10.1242/jcs.044305

Molinos, M., Almeida, C. R., Caldeira, J., Cunha, C., Goncalves, R. M., and Barbosa, M. A. (2015). Inflammation in intervertebral disc degeneration and regeneration. *J. R. Soc. Interface* 12:20150429. doi: 10.1098/rsif.2015.0429

Naylor, P. J., Scott, J., Drummond, J., Bridgewater, L., and Panagiotopoulos, C. (2010). Implementing a whole school physical activity and healthy eating model in rural and remote first nations schools: a process evaluation of action schools! *BC. Rural Remote Health* 10:1296. Available online at: www.rrh.org.au/journal/article/1296

Paine, A., and Ritchlin, C. T. (2016). Targeting the interleukin-23/17 axis in axial spondyloarthritis. *Curr. Opin. Rheumatol.* 28, 359–367. doi: 10.1097/BOR.0000000000000301

Pei, G., Chen, L., and Zhang, W. (2017). WGCNA application to proteomic and metabolomic data analysis. *Methods Enzymol.* 585, 135–158. doi: 10.1016/bs.mie.2016.09.016

Pfirrmann, C. W., Metzdorf, A., Zanetti, M., Hodler, J., and Boos, N. (2001). Magnetic resonance classification of lumbar intervertebral disc degeneration. *Spine* 26, 1873–1878. doi: 10.1097/00007632-200109010-00011

Rao, M. J., and Cao, S. S. (2014). Artificial total disc replacement versus fusion for lumbar degenerative disc disease: a meta-analysis of randomized controlled trials. *Arch. Orthop. Trauma Surg.* 134, 149–158. doi: 10.1007/s00402-013-1905-4

Risbud, M. V., and Shapiro, I. M. (2014). Role of cytokines in intervertebral disc degeneration: pain and disc content. *Nat. Rev. Rheumatol.* 10, 44–56. doi: 10.1038/nrrheum.2013.160

Rong, Z., Wang, A., Li, Z., Ren, Y., Cheng, L., Li, Y., et al. (2009). IL-17RD (Sef or IL-17RLM) interacts with IL-17 receptor and mediates IL-17 signaling. *Cell Res.* 19, 208–215. doi: 10.1038/cr.2008.320

Serin, E. A. R., Harm, N., Hilhorst, H. W. M., and Wilco, L. (2016). Learning from co-expression networks: possibilities and challenges. *Front. Plant. Sci.* 7:444. doi: 10.3389/fpls.2016.00444

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome. Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Sharma, A. (2018). The role of adipokines in intervertebral disc degeneration. *Med. Sci.* 6:34. doi: 10.3390/medsci6020034

Shen, L., Wu, Y., Han, L., and Zhang, H. (2018). Overexpression of growth and differentiation factor-5 inhibits inflammatory factors released by intervertebral disc cells. *Exp. Ther. Med.* 15, 3603–3608. doi: 10.3892/etm.2018.5867

Sherman, M. P., Dube, S., Spicer, T. P., Kane, T. D., Love, J. L., Saksena, N. K., et al. (1993). Sequence analysis of an immunogenic and neutralizing domain of the human T-cell lymphoma/leukemia virus type I gp46 surface membrane protein among various primate T-cell lymphoma/leukemia virus isolates including those from a patient with both HTLV-I-assoc. *Cancer Res.* 53, 6067–6073.

Smolders, L. A., Meij, B. P., Onis, D., Riemers, F. M., Bergknut, N., Wubbolts, R., et al. (2013). Gene expression profiling of early intervertebral disc degeneration reveals a down-regulation of canonical Wnt signaling and caveolin-1 expression: implications for development of regenerative strategies. *Arthritis Res. Ther.* 15:R23. doi: 10.1186/ar4157

Waddell, D. S., Liberati, N. T., Guo, X., Frederick, J. P., and Wang, X. F. (2004). Casein kinase Iepsilon plays a functional role in the transforming growth factor-beta signaling pathway *J. Biol. Chem.* 279, 29236–29246. doi: 10.1074/jbc.M400880200

Walker, B. F. (2000). The prevalence of low back pain: a systematic review of the literature from 1966 to 1998. *J. Spinal Disord.* 13, 205–217. doi: 10.1097/00002517-200006000-00003

Wan, Z. Y., Song, F., Sun, Z., Chen, Y. F., Zhang, W. L., Samartzis, D., et al. (2014). Aberrantly expressed long noncoding RNAs in human intervertebral disc degeneration: a microarray related study. *Arthritis Res. Ther.* 16:465. doi: 10.1186/s13075-014-0465-5

Wang, L., Gong, Y., Chippada-Venkata, U., Heck, M. M., Retz, M., Nawroth, R., et al. (2015). A robust blood gene expression-based prognostic model for castration-resistant prostate cancer. *BMC Med.* 13:201. doi: 10.1186/s12916-015-0442-0

Wu, C., Zhu, J., and Zhang, X. G. (2013). Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1

underlie the molecular difference between smoker and non-smoker lung adenocarcinoma. *BMC Bioinformatics*. 14:365. doi: 10.1186/1471-2105-14-365

Wu, J. P., Zhu, B., Ding, L., Yu, Z. C., and Ye, X. G. (2010). Morphometric analysis of chondrocyte apoptosis and degeneration of vertebral cartilage endplate in rats. *Fudan Univ. J. Med. Sci.* 37, 140–145. doi: 10.3969/j.issn.1672-8467.2010.02.003

Xianzhou, L., and Cunxin, Z. (2018). Endoplasmic reticulum stress participates in the process of high glucose-induced apoptosis in nucleus pulposus cells. *Chin. J. Tissue Eng. Res*. 22, 5778–5784. doi: 10.3969/j.issn.2095-4344.0542

Yi, F., Shi, X., Pei, X., and Wu, X. (2018). Tumor necrosis factor-alpha-308 gene promoter polymorphism associates with survival of cancer patients: a meta-analysis. *Medicine* 97:e13160. doi: 10.1097/MD.0000000000013160

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. 4:Article17. doi: 10.2202/1544-6115.1128

Zhang, S., and Cao, J. (2009). A close examination of double filtering with fold change and T test in microarray analysis. *BMC Bioinformatics*. 10:402. doi: 10.1186/1471-2105-10-402

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: 02 February 2021
doi: 10.3389/fgene.2021.609754

Transcriptomic Signatures and Functional Network Analysis of Chronic Rhinosinusitis With Nasal Polyps

Yun Hao[1,2,3†], *Yan Zhao*[1,2,3†], *Ping Wang*[1,2,3], *Kun Du*[1,2,3], *Ying Li*[1,2,3], *Zhen Yang*[4], *Xiangdong Wang*[1,2,3*] *and Luo Zhang*[1,2,3,5*]

[1] *Department of Otolaryngology Head and Neck Surgery, Beijing TongRen Hospital, Capital Medical University, Beijing, China,* [2] *Department of Allergy, Beijing TongRen Hospital, Capital Medical University, Beijing, China,* [3] *Beijing Key Laboratory of Nasal Diseases, Beijing Institute of Otolaryngology, Beijing, China,* [4] *Shanghai Key Laboratory of Medical Epigenetics, The International Co-laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Pudong Hospital, Institutes of Biomedical Sciences, Fudan University, Shanghai, China,* [5] *Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing, China*

OPEN ACCESS

Edited by:
Minxian Wallace Wang,
Broad Institute, United States

Reviewed by:
Weichen Zhou,
University of Michigan, United States
Sanna Katriina Toppila-Salmi,
University of Helsinki, Finland
Dawei Wu,
Capital Medical University, China

****Correspondence:***
Luo Zhang
dr.luozhang@139.com
Xiangdong Wang
entwxd@vip.sina.com

† These authors have contributed equally to this work

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

Received: *25 September 2020*
Accepted: *07 January 2021*
Published: *02 February 2021*

Citation:
Hao Y, Zhao Y, Wang P, Du K, Li Y, Yang Z, Wang X and Zhang L (2021) Transcriptomic Signatures and Functional Network Analysis of Chronic Rhinosinusitis With Nasal Polyps. Front. Genet. 12:609754. doi: 10.3389/fgene.2021.609754

Chronic rhinosinusitis with nasal polyps (CRSwNP) is a chronic sinonasal inflammatory disease with limited treatment options of corticosteroids, sinus surgery, or both. CRSwNP is frequently associated with allergic rhinitis and asthma, but the molecular mechanisms underlying CRSwNP inflammation are not completely understood. We obtained four gene expression profiles (GSE136825, GSE36830, GSE23552, and GSE72713) from four Gene Expression Omnibus (GEO), which collectively included 65 nasal polyp samples from CRSwNP patients and 54 nasal mucosal samples from healthy controls. Using an integrated analysis approach, we identified 76 co-differentially expressed genes (co-DEGs, including 45 upregulated and 31 downregulated) in CRSwNP patients compared with the healthy controls. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses identified the terms including immune effector process, leukocyte migration, regulation of the inflammatory response, *Staphylococcus aureus* infection, and cytokine-cytokine receptor interaction. protein-protein interaction (PPI) network analysis and real-time quantitative PCR (RT-qPCR) showed that 7 genes might be crucial in CRSwNP pathogenesis. Repurposing drug candidates (Alfadolone, Hydralazine, SC-560, Iopamidol, Iloprost, etc) for CRSwNP treatment were identified from the Connectivity Map (CMap) database. Our results suggest multiple molecular mechanisms, diagnostic biomarkers, potential therapeutic targets, and new repurposing drug candidates for CRSwNP treatment.

Keywords: chronic rhinosinusitis with nasal polyps, differentially expressed genes, hub genes, transcriptomic functional features, drug repurposing bioinformatics analysis of nasal polyps

INTRODUCTION

Chronic rhinosinusitis (CRS) is a common chronic heterogeneous nasal inflammatory disease that is associated with significant morbidity and a decreased quality of life. It affects ∼7 to 27% of adults in European populations, 14% of adults in the United States, and 8% of adults in China (Hastan et al., 2011; Shi et al., 2015; Wang X. et al., 2016). CRS is clinically classified into two

phenotypes according to the presence or absence of nasal polyps: CRS with nasal polyps (CRSwNP) and CRS without nasal polyps (CRSsNP) (Workman et al., 2018). CRSwNP can be classified into 2 distinct immunohistological subtypes based on eosinophil infiltration, eosinophilic CRSwNP (Eos CRSwNP) and non-eosinophilic CRSwNP (non-eos CRSwNP) (Cao et al., 2009). Eos CRSwNP demonstrates Th2 inflammation skewed with a relatively high recurrence and asthma comorbidity rate, while non-eos CRSwNP is characterized by a Th1 or Th17 response and a lower recurrence and asthma comorbidity rate (Zhang et al., 2008; Cao et al., 2009).

Recent studies have demonstrated that defects in the sinonasal epithelial barrier, increased exposure to pathogenic and colonized bacteria, and dysregulation of the host immune system play key roles in CRSwNP pathogenesis (Stevens et al., 2016). However, the inflammatory mechanisms underlying CRSwNP are not completely defined. In this regard, biomarkers that precisely indicate the development and progression of CRSwNP need to be further investigated to develop novel clinical strategies for CRSwNP treatment.

Microarray technology and bioinformatic analysis have emerged as promising, useful tools for screening genetic alterations involved in the development and progression of diseases. Furthermore, over the last decade, next-generation sequencing has produced substantial improvements in quality and yield (Goodwin et al., 2016). However, obtaining reliable results is difficult with both individual microarrays and sequencing due to the lack of samples (Kulasingam and Diamandis, 2008). Therefore, to obtain further insights into the mechanisms underlying the pathogenesis of CRSwNP and to clarify potential therapeutic targets, we analyzed a sufficient number of samples and combined differentially expressed genes (DEGs) derived from multiple microarray datasets with sequence-based data.

We herein aimed to explore the possible molecular mechanisms and biomarkers and propose new drug candidates for CRSwNP by integrating all the public databases for CRSwNP and using bioinformatics analyses of co-differentially expressed genes (co-DEGs) in nasal polyps from CRSwNP patients compared to nasal mucosal tissues from healthy control tissues. We described the transcriptional features, identified the biomarkers, and predicted the drug repurposing candidates, which could provide insights into precise CRSwNP treatment strategies.

MATERIALS AND METHODS

Microarray Studies, Datasets and Characteristics of Clinical Samples From the GEO Data Repository

In the present study, we selected microarray and high-throughput sequencing datasets of nasal tissues from CRSwNP patients in the GEO database using the following keywords: "CRSwNP," "Homo sapiens," and "nasal tissue." Based on these keywords, four CRSwNP datasets (GSE136825, GSE36830, GSE23552, and GSE72713) were downloaded from the repository. Derived from the GPL20301 platform (Illumina HiSeq 4000), GSE136825 includes nasal polyp tissue samples from 42 CRSwNP patients and nasal mucosal samples from 28 healthy controls (Peng et al., 2019). GSE36830 includes nasal polyp tissue samples from 6 CRSwNP patients and nasal mucosal samples from 6 healthy controls evaluated with the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) (Stevens et al., 2015b). GSE23552 is based on the Affymetrix Human Exon 1.0 ST Array and includes nasal polyp tissue samples from 11 CRSwNP patients and nasal mucosal samples from 17 healthy controls (Plager et al., 2010). GSE72713 is based on an Illumina HiSeq 2000 and includes nasal polyp tissue samples from 6 CRSwNP patients and nasal mucosal samples from 3 healthy controls (Wang W. et al., 2016). The details of each dataset are shown in **Table 1** and **Supplementary Table 10**. The flow chart detailing this study protocol is shown in **Figure 1**.

Differential Gene Expression Analysis

First, background correction and standardization were performed for the original GEO datasets using the packages EdgeR and Limma of R software (Ritchie et al., 2015). To determine whether the DEGs distinguished the CRSwNP group from healthy controls, principal coordinate analysis (PCoA) was applied to compare the overall characteristics of DEG communities between the two groups. The PCoA results were extracted and visualized using the Vegan and Ggplot2 packages of R software (version 1.2.5033) (Zhang et al., 2019). Next, differential analysis ($|\log_2 FC| > 1$, adjusted $p < 0.05$) of mRNAs was performed to compare nasal polyp and normal tissue samples with the Limma package of R software. Heatmaps and volcano plots of differentially expressed mRNAs were constructed using the packages Pheatmap and Ggplot2 of R software.

Subsequently, a Venn diagram showing the intersecting DEGs of the four datasets was created with Funrich software (version 3.1.3) (Pathan et al., 2015). The raw data in the four datasets are summarized in the form of a matrix and are shown in **Supplementary Table 1**.

PPI Network Construction

STRING (version 11.0) (http://string-db.org/) was used to identify the PPIs of the intersecting DEGs of the four datasets, with a combined score >0.4 used as the threshold for statistically significant interactions(Szklarczyk et al., 2015). Cytoscape (version 3.7.2) software was used for the PPI network visualization (Shannon et al., 2003). Then, the Molecular Complex Detection (MCODE) plugin, a graph theoretic clustering algorithm finding highly interconnected regions in a given network was used to identify important modules within the PPI network (Bader and Hogue, 2003). For this algorithm, seed vertices are selected and expanded by the density of the cluster. In detail, the degree cutoff was 2, the node score was 0.2, the k-score was 2, and the maximum depth was

TABLE 1 | The details of GEO datasets for CRSwNP.

GSE	PMID	Sample size (*n*)	Technology	Platform	Instrument	Age (y)	Sex, male (*n*%)	Number of DEGs	mRNA	
									Up	Down
GSE136825	PMID: 31439685	CRSwNP: 42 Control: 28	High-Throughput sequencing	GPL20301	Illumina HiSeq 4000	NA	NA	851	507	344
GSE36830	PMID: 26067893	CRSwNP: 6 Control: 6	*In situ* oligonucleotide	GPL570	[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array	CRSwNP: 38 ± 5 Control: 36 ± 6	CRSwNP: 4 (67%) Control: 2 (33%)	286	149	137
GSE23522	PMID: 20625511	CRSwNP: 11 Control: 17	*In situ* oligonucleotide	GPL5175	[HuEx-1_0-st] Affymetrix Human Exon 1.0 ST Array [transcript (gene) version]	CRSwNP: 40 ± 2.788 Control: 31.22 ± 2.913	CRSwNP: 5 (45%) Control: 9 (53%)	459	271	188
GSE72713	PMID: 27216292	CRSwNP: 6 Control: 3	High-Throughput sequencing	GPL11154	Illumina HiSeq 2000 (Homo sapiens)	CRSwNP: 46.8 ± 4.8 Control: 48.7 ± 7.6	CRSwNP: 4 (67%) Control: 1 (67%)	85	21	64

100. A false discovery rate (FDR) < 0.05 was considered statistically significant.

Functional Enrichment and Pathway Analyses

GO functional enrichment analysis is a commonly used method for annotating genes and identifying characteristic biological attributes of high-throughput genome or transcriptome data (Ashburner et al., 2000; Gene Ontology, 2006). KEGG pathway analysis is well-known for its systematic analysis of gene functions in biological pathways, which links genomic information with higher-order functional information (Kanehisa and Goto, 2000). Clusterprofiler package of R software integrates GO functional enrichment and KEGG pathway analyses (Yu et al., 2012). We analyzed the functions and signaling pathways of the intersecting DEGs using GO and KEGG analyses by ClusterProfiler package of R software. GO annotation includes three kinds of functional categories: biological process (BP), cellular component (CC) and molecular function (MF). $P < 0.05$ and $q < 0.2$ were considered statistically significant.

Screening Candidate Small-Molecule Drugs

To screen potential small-molecule drugs related to CRSwNP, the Connectivity Map (CMap) database, an online program for predicting potential drugs that may affect the biological status encoded by specific gene expression markers (https://portals.broadinstitute.org/cmap/), was employed (Lamb et al., 2006). Co-DEGs, which included upregulated and downregulated genes, were uploaded to query the CMap database. The enrichment score indicative of similarity was calculated and ranged from −1 to 1. A positive connectivity score indicated that the drug could induce the expression of the queried gene in CRSwNP, while a negative connectivity score indicated that the drug induced a status similar to that of normal cells, suggesting its potential to treat CRSwNP. The results were ranked by *p*-value.

Patient Recruitment

This study was approved by the Ethics Committee of Beijing TongRen Hospital, Capital Medical University, and written informed consent was obtained from each patient before enrollment. A total of 70 subjects, including 46 patients with CRSwNP and 24 healthy control subjects, were recruited. We collected nasal polyp tissues from patients with CRSwNP and nasal mucosal tissues from control subjects. The diagnosis of CRSwNP was made according to the European Position Paper on Rhinosinusitis and Nasal Polyps 2012 guidelines (Fokkens et al., 2012). Control subjects without other sinonasal diseases were those undergoing septoplasty because of anatomic variations. None of the patients had been treated with corticosteroids, immunomodulatory agents, or antibiotics within 4 weeks before enrollment. The exclusion criteria were as follows: patients with acute infections, acetylsalicylic acid-intolerance, fungal sinusitis, immunodeficiency, coagulation disorder, or cystic fibrosis and pregnant women. Details of the subjects' characteristics are included in **Supplementary Table 7**.

RNA Extraction and Real-Time Quantitative PCR (RT-qPCR)

Total RNA was isolated from nasal polyps of CRSwNP patients and from the nasal mucosa of controls using Tri®-Reagent (Sigma) according to the manufacturer's instructions. The quality of total RNA was assessed with a Nanodrop-2000

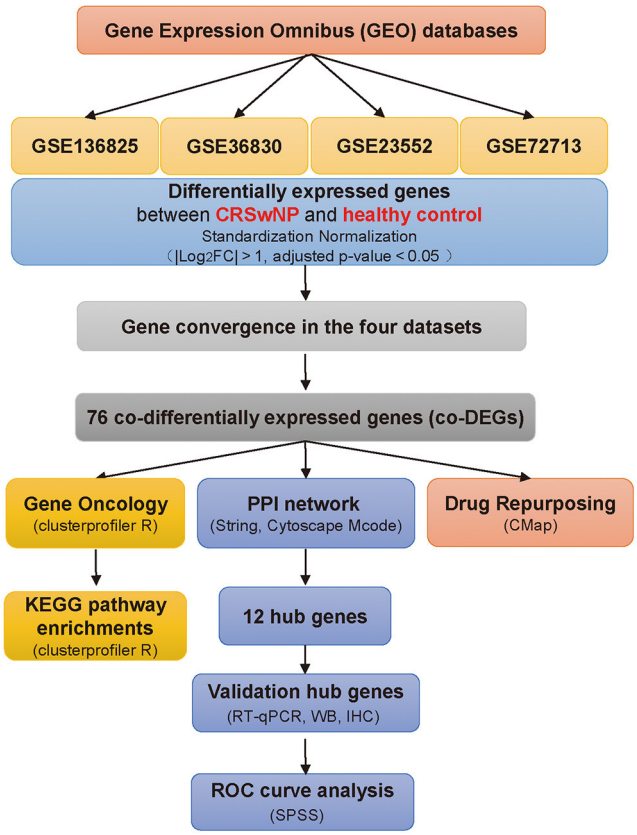

FIGURE 1 | Schematic overview of the study. Flowchart of the study to identify potential factors underlying CRSwNP pathogenesis compared with healthy controls. CRSwNP, chronic rhinosinusitis with nasal polyps; GO, Gene Ontology; FC, fold change; Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; PPI, protein-protein interaction; co-DEGs, co-differentially expressed genes; CMap, Connectivity Map database; RT-qPCR, real-time quantitative PCR; WB, western blot; IHC, immunohistochemistry.

(Thermo Fisher Scientific, Waltham, Mass), and complementary DNA was synthesized from 1 μg of total RNA using PrimeScript RT Master Mix (Abclonal Biotechnology). RT-qPCR was performed by using SYBR Green mix (Abclonal Biotechnology) to assess gene expression levels. Primers are listed in **Supplementary Table 8**.

Western Blot Analysis

Tissues of nasal polyps from CRSwNP patients and the nasal mucosa of controls were homogenized in ice-cold RIPA lysis buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1.0% Triton X-100, 20 mM EDTA, 1 mM Na_3VO_4, 1 mM NaF, and 1 mM PMSF). The protein concentration was measured by bicinchoninic acid (BCA) kit (Beyotime, Shanghai, China). In brief, equal amounts of proteins (16 μg) were loaded on the sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) and transferred to nitro cellulose membranes. The membranes were sequentially incubated with primary antibodies and horseradish peroxidase-conjugated secondary antibodies (described below). The following primary antibodies were used: anti-BTK (1:1000 diluted, ABclonal, A19002, Wuhan, China), anti-HCK (1:1000 diluted, ABclonal, A14537, Wuhan, China), anti-HK3 (1:1000 diluted, ABclonal, A8428, Wuhan, China), anti-NCF2 (1:1000 diluted, ABclonal, A1178, Wuhan, China), anti-NOX2/gp91phox (1:1000 diluted, Abcam, ab80897), anti-FLAP (3:5000 diluted, Abcam, ab53536), and anti-β-actin (1:10000 diluted, Sigma, A5441) at 4°C overnight, they were further immunoblotted with HRP-conjugated IgG antibody (1:5000 diluted, ABclonal, Wuhan, China) at room temperature for 60 min, developed with enhanced chemiluminescence (ECL) substrate (Millipore, Darmstadt, Germany) and chemiluminescence detection by ChemiDocTM MP Imaging System (Bio-Rad, United Kingdom). Band density was quantitated using the Image LabTM software Version 6.0.0 (Bio-Rad, United Kingdom).

Immunohistochemistry Staining

Five-micron thick sections were obtained from blocks of nasal polyps and nasal mucosa from CRSwNP patients and control subjects, dewaxed in xylol and rehydrated in graded ethanol. For antigen retrieval, the slides containing the samples were incubated with citrate buffer (pH 6.0) in a pressure cooker (Zhongshan Jinqiao Biotechnology, Beijing, China). The samples were then treated with freshly prepared 3% hydrogen peroxide in methanol for 20 min and further washed in Tris-buffered saline. The slides were incubated overnight at 4°C with anti-BTK (1:100 diluted, ABclonal, A19002, Wuhan, China), anti-HCK (1:200 diluted, ABclonal, A2083, Wuhan, China), anti-HK3 (1:400 diluted, ABclonal, A8428, Wuhan, China), anti-NCF2 (1:500 diluted, ABclonal, A1178, Wuhan, China), anti-NOX2/gp91phox (1:100 diluted, ABclonal, A19701, Wuhan, China), anti-BFL-1/GRS (1:150 diluted, Abcam, ab45413), anti-FLAP (1:100 diluted, Abcam, ab53536). A polymer system (Zhongshan Jinqiao Biotechnology, Beijing, China) was applied as a secondary antibody conjugated to peroxidase. DAB (3′-diaminobenzidine tetrahydrochloride, Zhongshan Jinqiao Biotechnology, Beijing, China) was used as the chromogen, for 5 min, followed by Harris hematoxylin counterstain. Slides were analyzed under a light microscope (Nikon H600L, Japan) and 5 images were taken for each slide (Nikon NIS software, version 4.60, Nikon, Japan) at high-power (40X objective) field. Representative areas were qualitatively selected for immunostaining analysis. For digital analysis, we used the cell counter function of the ImageJ software (version 1.52), in which we semi-quantitatively determined the average optical density values.

Statistical Analysis

Differences between groups were assessed by ANOVA. In all cases, $P < 0.05$ was considered statistically significant. We drew a receiver operator characteristic (ROC) curve to calculate the area under curve (AUC) to discriminate CRSwNP patients from normal subjects. SPSS 16.0 for Windows (IBM, Chicago, USA) was used for ROC analyses and other statistical analyses were performed using GraphPad Prism 7.0 software (GraphPad Software, La Jolla, CA).

RESULTS

Integrative Analysis of DEGs in CRSwNP Samples From 4 GEO Datasets

To avoid a high proportion of false positives in an individual dataset, multiple-dataset integration was necessary for obtaining reliable results to further investigate the complex molecular mechanisms of CRSwNP. We performed background correction and standardization to reduce variability in four GEO datasets and PCoA, a dimension reduction technique, to present visual coordinates of similarity or differences between the CRSwNP and healthy controls from GEO data (Zhang et al., 2019). PCoA of gene expression in each of the four datasets (GSE136825, GSE36830, GSE23552, and GSE72713) revealed that the samples clustered into two distinct groups (**Figure 2A**).

We divided the genes into different categories according to their biotype (https://www.ncbi.nlm.nih.gov/gene), and Volcano plots were used to display the gene expression data and *p*-value statistics of each of the datasets (**Supplementary Figure 1A**). We identified the genes that were significantly differentially expressed ($|\log_2 FC| > 1$, adjusted $p < 0.05$) in nasal polyps compared to control tissues (**Supplementary Figure 1B**). Then, we integrated the DEGs and identified 76 co-DEGs, including 45 upregulated genes and 31 downregulated genes, derived from the intersections of any three of the four GEO datasets (**Figure 2B** and **Supplementary Table 1**). A cluster heatmap was used to visualize the changes in up- and downregulated genes among 76 co-DEGs (**Figure 2C**), and details of the 76 co-DEGs are shown in **Supplementary Table 1**.

GO Functional Enrichment and KEGG Pathway Analyses of co-DEGs in CRSwNP

To explore the potential functions of co-DEGs, we performed GO functional enrichment and KEGG pathway analyses ($p < 0.05$ and $q < 0.2$). Notably, the BP terms associated with the upregulated genes were regulation of immune effector process, leukocyte migration, regulation of inflammatory response, negative regulation of immune system process, and regulation of leukocyte-mediated immunity. The CC terms associated with the upregulated genes were the external side of plasma membrane, secretory granule membrane, and membrane raft. The MF terms associated with the upregulated genes were phospholipid binding, carboxylic acid binding, organic acid binding, G protein-coupled receptor binding, and cytokine receptor binding (**Figure 3A**). In addition, downregulated genes were also strongly associated with the BP terms organic anion transport, multicellular organismal homeostasis, drug transport, and tissue homeostasis. For downregulated genes, basolateral plasma membrane was found to be the dominant CC term. The significantly enriched MF terms associated with the downregulated genes were metal ion transmembrane transporter activity, secondary active transmembrane transporter activity, and active transmembrane transporter activity (**Figure 3B**).

In the KEGG pathway analysis, the upregulated genes were mainly related to *Staphylococcus aureus* infection, cytokine-cytokine receptor interactions, complement and coagulation cascades, and viral protein interactions with cytokines and cytokine receptors (**Figure 3C**), and the downregulated genes were mainly involved in bile secretion and salivary secretion (**Figure 3D**). The results of GO and KEGG pathway enrichment analyses are also shown in **Supplementary Tables 2–4**.

PPI Networks of co-DEGs and Hub Genes in CRSwNP

To further investigate the biological functions of the co-DEGs, we constructed PPI networks according to the 76 co-DEGs in CRSwNP (**Figure 3E**). The PPI networks contained 57 nodes and 202 edges, and the isolated genes without interactions were removed. The MCODE algorithm was further applied to identify hub genes that were densely associated with each other in the network (**Figure 3F**). We found that 12 hub genes including Arachidonate 5-Lipoxygenase Activating Protein (*ALOX5AP*), Bcl-2-related protein A1 (*BAL2A1*), Tyrosine-protein kinase BTK (*BTK*), Cytochrome b-245 heavy chain (*CYBB*), Neutrophil cytosol factor 2 (*NCF2*), Tyrosine-protein kinase HCK (*HCK*), Hexokinase-3 (*HK3*), Macrophage colony-stimulating factor 1 receptor (*CSF1R*), Pleckstrin (*PLEK*), CMRF35-like molecule 8 (*CD300A*), Integrin beta-2 (*ITGB2*), and fMet-Leu-Phe receptor (*FPR1*) might play prominent roles in interacting with each other in the PPI network, which indicated that these 12 genes might be core molecules in the development of CRSwNP (**Supplementary Tables 5**, **6**). The 12 genes screened from the PPI network were also related to neutrophil-mediated immunity, positive regulation of the innate immune response, positive regulation of the defense response, and *Staphylococcus aureus* infection as determined by GO functional enrichment and KEGG pathway analyses.

Validation of Hub Genes

To further validate the results of bioinformatics analysis, the gene expression levels of the 12 hub genes from PPI network (*ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, HK3, CSF1R, PLEK, CD300A, ITGB2*, and *FPR1*) in nasal polyps from CRSwNPs and nasal mucosa from healthy controls were determined by RT-qPCR. As illustrated in **Figure 4** and **Supplementary Figure 2**, the expression levels of *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK*, and *HK3* were significantly altered in CRSwNP, as identified by the bioinformatics analysis. The other five genes did not show significantly different expression levels in CRSwNP and healthy control samples. Regarding diagnostic prediction quality, the hub genes *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK,* and *HK3* performed well-according to receiver operator characteristic (ROC) analysis (**Figure 4C** and **Supplementary Table 9**). The area under the ROC curves (AUC) of the genes *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK,* and *HK3* are 0.7698, 0.7639, 0.7029, 0.8418, 0.8913, 0.8185, 0.7136, respectively. The AUC of combined detection of the 7 indexes was 0.9354, which was higher than that of each single detection. Both the qPCR and ROC analyses suggest that these seven hub genes could be diagnostic biomarkers for CRSwNP.

Next, we identified the protein level of the seven hub genes (ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3) from nasal polyps from CRSwNPs and nasal mucosa

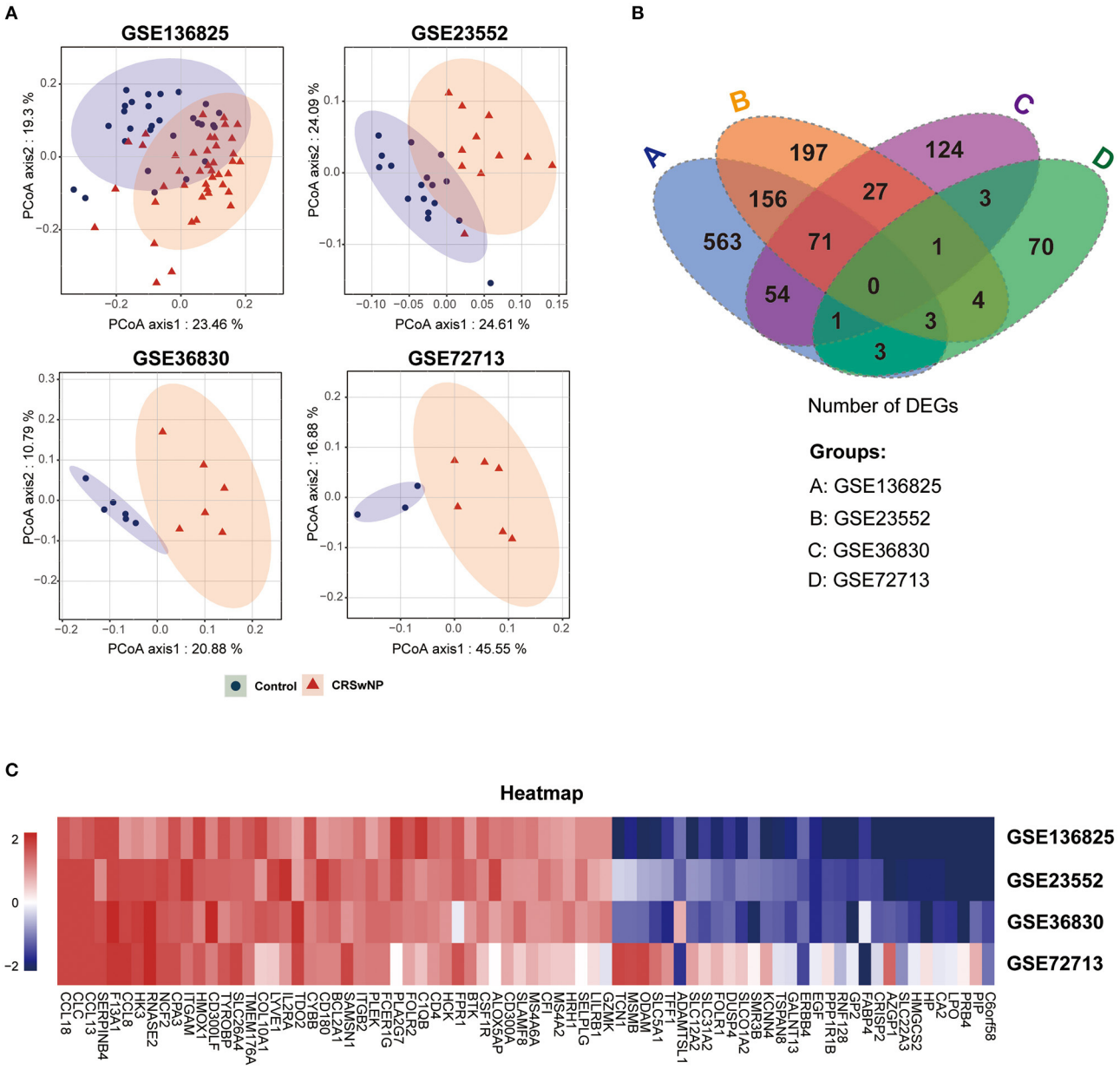

FIGURE 2 | The identification of co-DEGs in four GEO datasets of CRSwNP. **(A)** Principal coordinate analysis (PCoA) of gene expression in CRSwNP and healthy control subjects in four GEO datasets. **(B)** Venn diagram showing the number of integrated genes among the four GEO datasets. **(C)** Gene expression heatmap of the 76 co-DEGs in CRSwNP compared to healthy controls.

from healthy controls. The western blot results showed the expression level of ALOX5AP, BTK, CYBB, NCF2, HCK, and HK3 in CRSwNP was significantly increased in nasal polyps compared to control subjects (**Figure 5**). Moreover, the immunohistochemistry stain results also demonstrated that the protein level of ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 were significantly increased in nasal polyps compared to control subjects (**Figure 6**). We found that ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 were broadly expressed on both epithelial layer and stromal layer in nasal polyp tissues.

Prediction of Potential Novel Drugs for the Treatment of CRSwNP by CMap

To identify potential drugs for CRSwNP treatment, we introduced 45 upregulated co-DEGs and 31 downregulated co-DEGs from the four GEO datasets into the CMap database

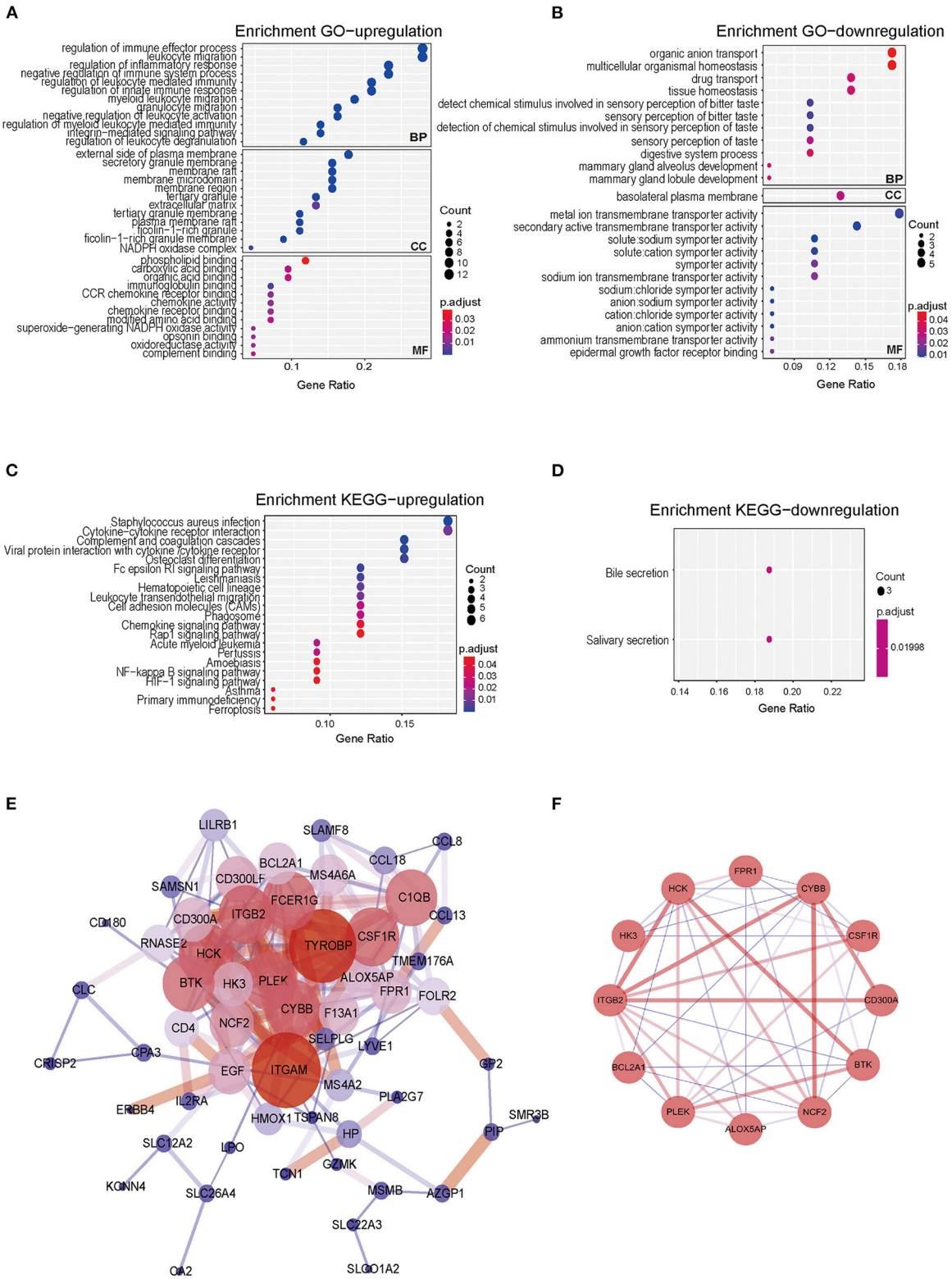

FIGURE 3 | Gene Ontology, KEGG pathway, PPI network of co-DEGs, and hub gene identification analyses in CRSwNP. **(A,B)** Bubble chart showing enriched GO terms for **(A)** upregulated co-DEGs and **(B)** downregulated co-DEGs. **(C,D)** Bubble chart showing enriched KEGG pathways for **(C)** upregulated co-DEGs and **(D)** downregulated co-DEGs. *P*-values < 0.05 and *q*-values< 0.2 were considered statistically significant. **(E)** PPI networks of the 76 co-DEGs from the four GEO datasets of CRSwNP. The node color represents the degree of proteins, and the edge color represents the combined score of proteins. Red represents high, and blue represents low. **(F)** The hub genes of PPI networks.

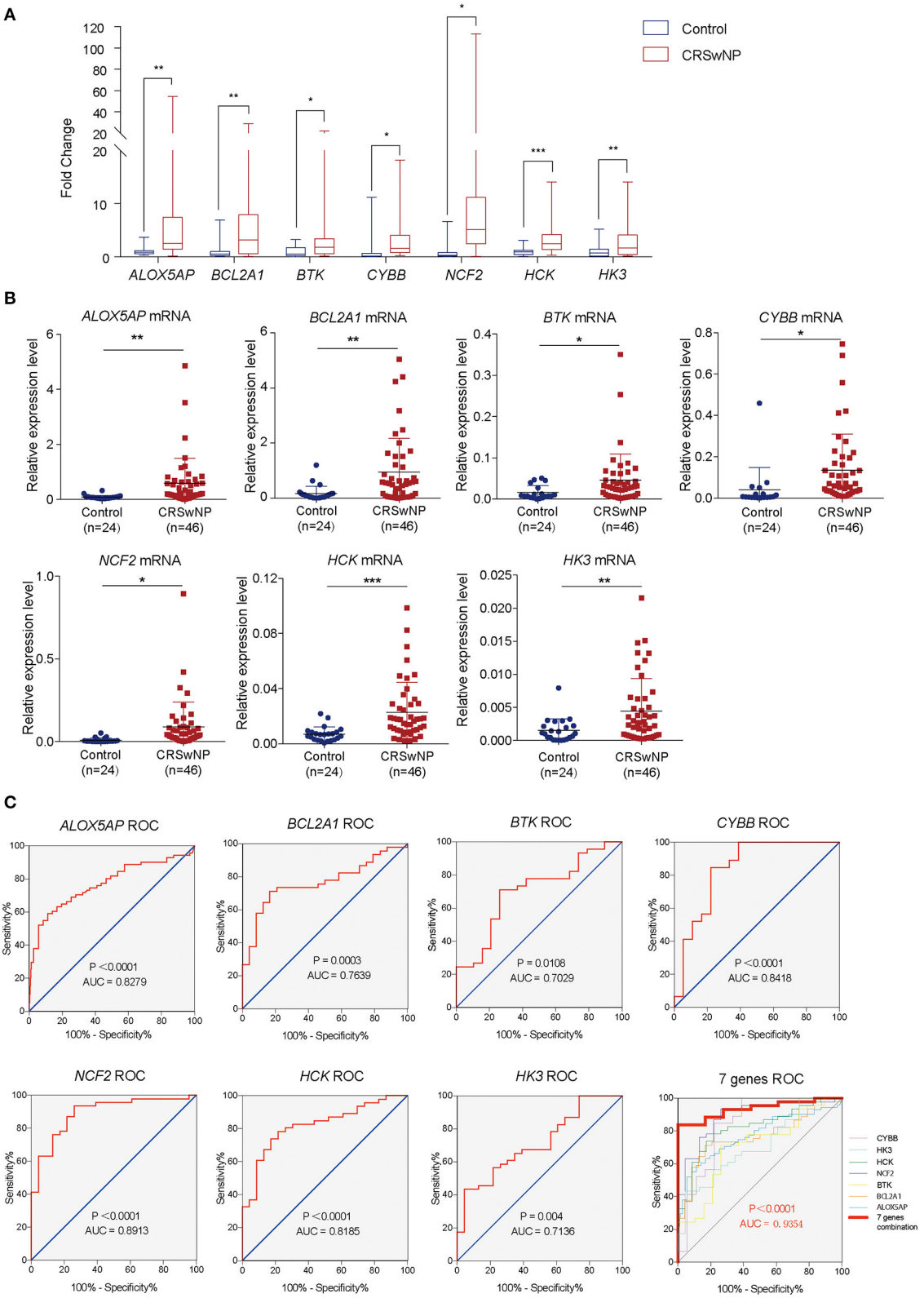

FIGURE 4 | The gene expression levels and diagnostic values of 7 hub genes in CRSwNP. **(A)** Fold changes in the *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK*, and *HK3* genes in CRSwNP as determined by RT-qPCR. **(B)** The relative expression levels of the *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK*, and *HK3* genes in CRSwNP. *GAPDH* was used as a reference. **(C)** ROC curves for testing the hub genes *ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, HK3,* and the combination of 7 hub genes as determined by RT-qPCR. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

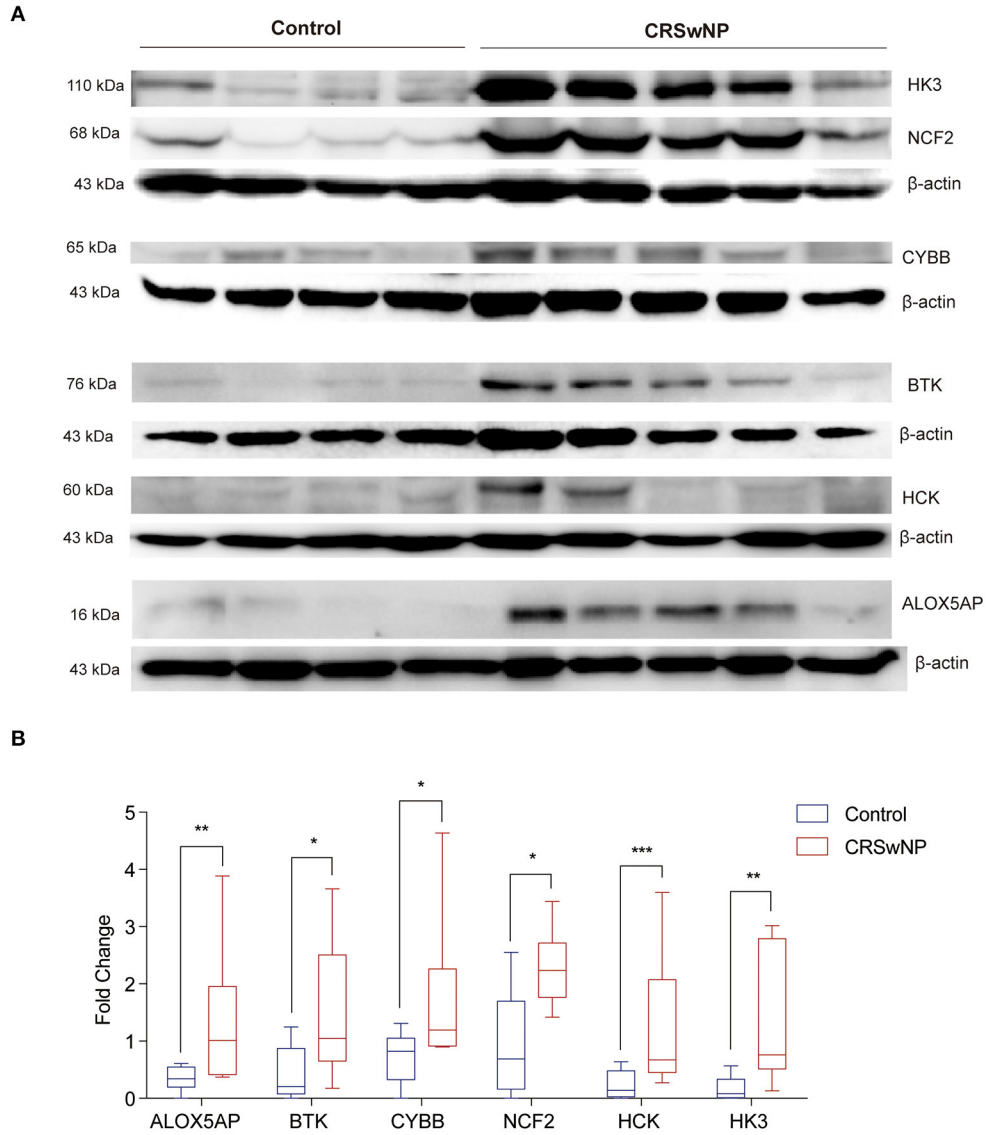

FIGURE 5 | The protein expression levels 6 hub genes were significantly increased in CRSwNP by western blot analysis. **(A)** The expression level of ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 proteins in nasal polyps from CRSwNP patients ($n = 10$) and nasal mucosal tissues from healthy control ($n = 8$). **(B)** Fold changes of relative expression ratio of ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 compare to β-actin in CRSwNP and healthy controls. The expression level of β-actin was used as a reference. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$.

and matched them with targeted drug therapies. The top 15 most significant small molecules (Alfadolone, Hydralazine, SC-560, Iopamidol, Iloprost, Clorgiline, Cefotetan, Etynodiol, Disulfiram, Ketotifen, Florfenicol, Clidinium bromide, Ramifenazone, Nafcillin, Bepridil) and their enrichment value are listed in **Table 2**. These drug repurposing candidates could target co-DEGs in CRSwNP and then affect the expression and function of genes. This provides a novel perspective to explore potential precise targeted drugs for CRSwNP treatment. Further experiments are needed to confirm the efficacy of these drug candidates in CRSwNP.

DISCUSSION

Previous studies were limited to individual datasets or incorrect combinations, while our study integrated all the available public GEO databases of CRSwNP. The CRSwNP groups were independent of the normal control groups in each of the four datasets as determined by PCoA. We identified 76 co-DEGs (45 upregulated and 31 downregulated) among all the GEO data. The PPI network provided an overview illustration of the associations among the 76 co-DEGs, and we identified 7 hub genes not only by mRNA level, also by protein

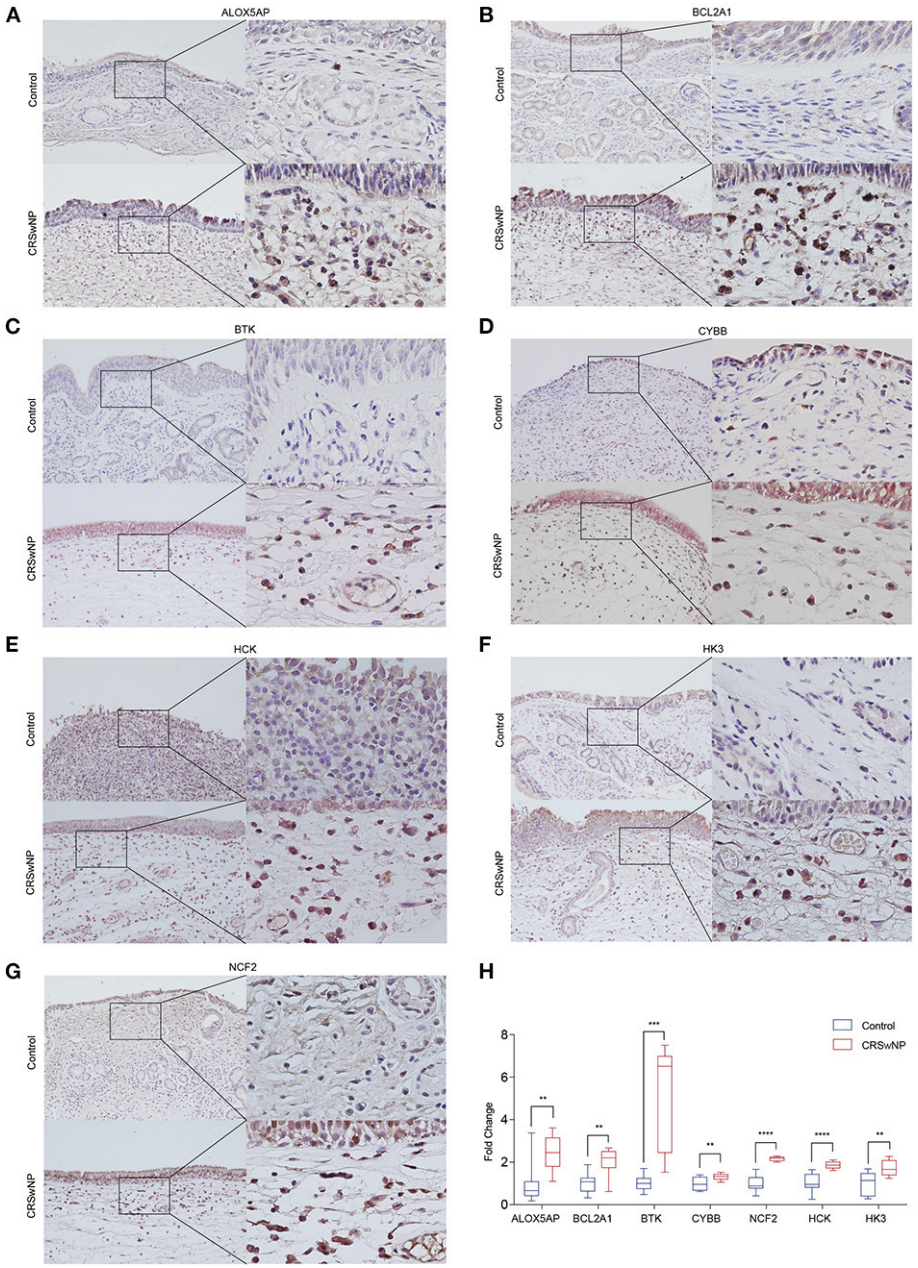

FIGURE 6 | The protein expression levels 7 hub genes were significantly increased in CRSwNP by immunohistochemistry staining. **(A–G)** The expression level and location distribution of ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 proteins in nasal polyps from CRSwNP patients ($n = 11$) and nasal mucosal tissues from healthy control ($n = 7$). **(H),** Fold changes of average optical density value in the ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 protein expression in CRSwNP and healthy controls. $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

expression level that might be biomarkers and key regulators of CRSwNP pathogenesis.

ALOX5AP is an essential regulator of the biosynthesis of leukotriene B4 (Haeggstrom, 2018). Previous studies on the genome-wide gene expression profile of CRSwNP showed increased *ALOX5AP* gene expression levels in the nasal polyps of patients with aspirin-intolerant asthma (Sekigawa et al., 2009) and decreased methylation levels of ALOX5AP in a genome-wide methylation profile of nasal polyps (Cheong et al., 2011). BTK, a member of the Tec family of tyrosine kinases, has been indicated to play crucial roles in B cell development and signal transduction downstream of the high-affinity receptor for IgE (FcεR) on mast cells and basophils in an ovalbumin-induced mouse model of asthma (Phillips et al., 2016). However, there have been no studies

TABLE 2 | Results of CMap analysis of co-DEGs in CRSwNP.

Rank	CMap name	Mean	*N*	Enrichment	*P*-value	CID	Molecular formula	Group
1	Alfadolone	0.715	3	0.942	0.00024	9798416	C21H32O4	Approved
2	Hydralazine	0.27	6	0.752	0.00058	3637	C8H8N4	Approved
3	SC-560	0.631	3	0.905	0.00176	4306515	C17H12ClF3N2O	Experimental
4	Iopamidol	−0.255	4	−0.825	0.00177	65492	C17H22I3N3O8	Approved
5	Iloprost	−0.686	3	−0.899	0.00194	5311181	C22H32O4	Approved
6	Clorgiline	0.287	4	0.8	0.003	4380	C13H15Cl2NO	Experimental
7	Cefotetan	0.484	3	0.877	0.00355	53025	C17H17N7O8S4	Approved
8	Etynodiol	−0.319	4	−0.793	0.00374	14687	C20H28O2	Experimental
9	Disulfiram	0.576	5	0.722	0.00378	3117	C10H20N2S4	Approved
10	Ketotifen	−0.322	4	−0.782	0.00462	3827	C19H19NOS	Approved
11	Florfenicol	0.242	4	0.764	0.00585	114811	C12H14Cl2FNO4S	Approved
12	Clidinium bromide	−0.303	4	−0.753	0.00746	19004	C22H26BrNO3	Approved
13	Ramifenazone	0.62	4	0.731	0.01038	5037	C14H19N3O	Experimental
14	Nafcillin	0.454	4	0.725	0.0115	8982	C21H22N2O5S	Approved
15	Bepridil	0.284	4	0.724	0.01162	2351	C24H34N2O	Approved

CMap, Connectivity Map; CID, Compound ID.

on BTK in chronic nasal diseases. NCF2 (also called p67phox) is a subunit of the multiprotein NADPH oxidase complex, which is an essential component of the innate immune response responsible for effective superoxide production in neutrophils (Thomas, 2017). Another study from our group previously found that p67phox expression was significantly increased in nasal polyp tissue compared with control mucosal tissue (Zheng et al., 2020). A study using a nitric oxide polymerase chain reaction array showed significant upregulation of NCF2 expression in CRS patients who were both *Staphylococcus aureus* biofilm-positive and polyp-positive compared to control subjects (Jardeleza et al., 2013). HCK, a member of the Src family of tyrosine kinases, acts as a key regulator of gene expression in alternatively activated monocytes/macrophages (Bhattacharjee et al., 2011). Similar to NCF2, CYBB (often referred to as p91phox or NOX2) has also been found to be upregulated in CRSwNP (Zheng et al., 2020). HK3 played a functional role in acute promyelocytic leukemia, non-small lung cancer, and colorectal cancer (Federzoni et al., 2014; Wolf et al., 2016; Pudova et al., 2018; Tuo et al., 2020). BCL2A1 is a member of the BCL-2 family of antiapoptotic proteins that is induced by mucin 1 transmembrane C-terminal (MUC1-CT) via the NF-κB p65-dependent signaling pathway (Hiraki et al., 2018). MUC1 has been identified as an anti-inflammatory molecule that could inhibit bacteria- and virus-induced inflammation in airways (Kim and Lillehoj, 2008; Li et al., 2010; Kyo et al., 2012). MUC1-CT also participates in the corticosteroid response in the treatment of CRSwNP (Milara et al., 2015), but relationships between BCL2A1 and CRSwNP remain unknown. Our study has proved the increased expression of ALOX5AP, BCL2A1, BTK, CYBB, NCF2, HCK, and HK3 by mRNA and protein levels in nasal polyps.

Currently, colonization by fungi and bacteria, alterations in mucociliary clearance, abnormalities in the sinonasal epithelial cell barrier and tissue remodeling combined with host innate and adaptive immune responses are known to contribute to the chronic inflammatory and tissue-deforming processes characteristic of CRS (Stevens et al., 2015a). In our study, GO and KEGG results showed that upregulated genes were predominantly enriched for the immune effector process, leukocyte migration, regulation of the inflammatory response, negative regulation of the immune system process, and regulation of leukocyte-mediated immunity. Dysregulation of these processes indicated that increasing exposure to pathogenic and colonized bacteria ultimately caused complicated downstream immune responses and chronic inflammation during the formation and development of nasal polyps. Additionally, downregulated genes were enriched for multicellular organismal homeostasis and tissue homeostasis, which reflected the destruction of the sinonasal epithelial cell barrier and tissue remodeling in CRSwNP. KEGG pathway analysis demonstrated that the upregulated genes were mainly related to *Staphylococcus aureus* infection and cytokine-cytokine receptor interactions. The above pathways are critical biological processes for pathogen invasion, immune effector and inflammatory responses, and tissue homeostasis disorder. It is worth noting that dysregulation of these processes eventually leads to a severe immune response and the formation of nasal polyps.

Although previous studies involved bioinformatic analysis of mRNAs and lncRNAs in CRSwNP (Liu et al., 2019; Zhou et al., 2020), one study included only 12 CRSwNP patients and 9 healthy controls. The other study used nasal tissue data combined with primary human basal cells cultured in an air-liquid interface system, which might be different from nasal polyp tissue. Our study included 65 CRSwNP patients and 54 healthy controls, representing the largest CRSwNP study to date. Additionally, we identified new genes that might be involved in the pathogenesis of CRSwNP. Moreover, we used the CMap database to identify drug repurposing candidates potentially targeting the co-DEGs derived from the four GEO datasets. Repurposing drugs with higher enrichment scores are more

likely to reverse the gene expression changes seen in CRSwNP than those with lower enrichment scores. This work may help to develop new drugs for CRSwNP treatment. Ketotifen is a cycloheptathiophene blocker of histamine H1 receptors and inflammatory mediator release that has been widely used in the treatment of allergic rhinitis, asthma and allergic conjunctivitis, and its common side effects include tiredness, dry mouth, and nausea. Clidinium bromide is a synthetic anticholinergic agent associated with antispasmodic and antisecretory effects on the gastrointestinal tract and has the side effects of dry mouth, dry skin and flushed face. Cefotetan and nafcillin are broad-spectrum cephalosporin antibiotics might rarely cause allergic reactions that include a rash, systemic papules, urticaria, pruritus, and fever. These drugs have not yet been reported as therapies for CRSwNP. Traditionally, CRSwNP treatments include nasal saline irrigation, intranasal corticosteroids, oral antibiotics, oral corticosteroids, and surgery depending on both the site and symptoms of disease (Kariyawasam and Scadding, 2011; Lee, 2015). With the advancement of the concept of CRSwNP endotypes, the management of precision medicine and reductions in recurrence are evolving (Kim and Cho, 2017). Endotyping helps physicians to determine optimal primary therapeutic modality and predict treatment outcomes and risks for comorbidities. Biologics in CRSwNP mainly focus on targeting the type 2 cytokines such as IL-4, IL-5, IL-13, as well as IgE. Combining our study with previous studies, the hub genes increased in CRSwNP might be served as biomarkers. Our study provides new insights which will shift drug discovery toward the personalized and precision medicine treatment approach to enhance CRSwNP therapies. Therefore, further research is needed to explore the potential of new targeted drugs in CRSwNP treatment.

CONCLUSION

In summary, by comprehensively analyzing gene expression profiles, sequencing data from four CRSwNP GEO datasets, identifying key genes and important pathways, and predicting the repurposing drugs for CRSwNP treatment, our study elucidated molecular mechanisms underlying the occurrence and development of CRSwNP to explain its pathogenesis and aid in diagnosis from the perspective of bioinformatics. Our study successfully identified 7 potential genes as key regulators and predicted a series of repurposing drugs to expand CRSwNP treatment. However, more experimental validation is necessary before these data can be translated into the clinic.

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

The studies involving human participants were reviewed and approved by this study was approved by the Ethics Committee of Beijing TongRen Hospital, Capital Medical University (TRECKY2019-050). All subjects who participated in this research provided written informed consent. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LZ and XW conceived and designed the project. YH performed the integrated analysis under the guidance of YZ, ZY, and KD. PW helped collect the GEO database data. YH and YZ did the experiments with the help of PW and YL. YH and YZ wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the national key R&D program of China (2016YFC0905200), the program for the Changjiang scholars and innovative research team (IRT13082), the national natural science foundation of China (81630023, 81970852, 82000962, and 91959106), the Beijing Bai-Qian-Wan talent project (2019A32), and the Public Welfare Development and Reform Pilot Project (2019-10), the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-022).

ACKNOWLEDGMENTS

We thank the researchers who contributed to the GEO datasets and thank the patients and healthy volunteers who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.609754/full#supplementary-material

REFERENCES

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. The gene ontology consortium. *Nat. Genet*. 25, 25–29. doi: 10.1038/75556

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Bhattacharjee, A., Pal, S., Feldman, G. M., and Cathcart, M. K. (2011). Hck is a key regulator of gene expression in alternatively activated human monocytes. *J. Biol. Chem*. 286, 36709–36723. doi: 10.1074/jbc.M111.291492

Cao, P. P., Li, H. B., Wang, B. F., Wang, S. B., You, X. J., Cui, Y. H., et al. (2009). Distinct immunopathologic characteristics of various types of chronic rhinosinusitis in adult Chinese. *J. Allergy. Clin. Immunol.* 124, 484.e471–472. doi: 10.1016/j.jaci.2009.05.017

Cheong, H. S., Park, S. M., Kim, M. O., Park, J. S., Lee, J. Y., Byun, J. Y., et al. (2011). Genome-wide methylation profile of nasal polyps: relation to aspirin hypersensitivity in asthmatics. *Allergy* 66, 637–644. doi: 10.1111/j.1398-9995.2010.02514.x

Federzoni, E. A., Humbert, M., Torbett, B. E., Behre, G., Fey, M. F., and Tschan, M. P. (2014). CEBPA-dependent HK3 and KLF5 expression in primary AML and during AML differentiation. *Sci. Rep*. 4:4261. doi: 10.1038/srep04261

Fokkens, W. J., Lund, V. J., Mullol, J., Bachert, C., Alobid, I., Baroody, F., et al. (2012). European position paper on rhinosinusitis and nasal polyps 2012. *Rhinol. Suppl*. 23, 1–298. doi: 10.4193/Rhino50E2

Gene Ontology, C. (2006). The gene ontology (GO) PROJECT in 2006. *Nucleic Acids Res*. 34, D322–D326. doi: 10.1093/nar/gkj021

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet*. 17, 333–351. doi: 10.1038/nrg.2016.49

Haeggstrom, J. Z. (2018). Leukotriene biosynthetic enzymes as therapeutic targets. *J. Clin. Invest*. 128, 2680–2690. doi: 10.1172/JCI97945

Hastan, D., Fokkens, W. J., Bachert, C., Newson, R. B., Bislimovska, J., Bockelbrink, A., et al. (2011). Chronic rhinosinusitis in Europe–an underestimated disease. A GA(2)LEN study. *Allergy* 66, 1216–1223. doi: 10.1111/j.1398-9995.2011.02646.x

Hiraki, M., Maeda, T., Mehrotra, N., Jin, C., Alam, M., Bouillez, A., et al. (2018). Targeting MUC1-C suppresses BCL2A1 in triple-negative breast cancer. *Signal Transduct. Target Ther*. 3:13. doi: 10.1038/s41392-018-0013-x

Jardeleza, C., Jones, D., Baker, L., Miljkovic, D., Boase, S., Tan, N. C., et al. (2013). Gene expression differences in nitric oxide and reactive oxygen species regulation point to an altered innate immune response in chronic rhinosinusitis. *Int. Forum Allergy Rhinol*. 3, 193–198. doi: 10.1002/alr.21114

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28, 27–30. doi: 10.1093/nar/28.1.27

Kariyawasam, H. H., and Scadding, G. K. (2011). Chronic rhinosinusitis: therapeutic efficacy of anti-inflammatory and antibiotic approaches. *Allergy Asthma Immunol. Res*. 3, 226–235. doi: 10.4168/aair.2011.3.4.226

Kim, D. W., and Cho, S. H. (2017). Emerging endotypes of chronic rhinosinusitis and its application to precision medicine. *Allergy Asthma Immunol. Res*. 9, 299–306. doi: 10.4168/aair.2017.9.4.299

Kim, K. C., and Lillehoj, E. P. (2008). MUC1 mucin: a peacemaker in the lung. *Am. J. Respir. Cell Mol. Biol*. 39, 644–647. doi: 10.1165/rcmb.2008-0169TR

Kulasingam, V., and Diamandis, E. P. (2008). Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol*. 5, 588–599. doi: 10.1038/ncponc1187

Kyo, Y., Kato, K., Park, Y. S., Gajghate, S., Umehara, T., Lillehoj, E. P., et al. (2012). Antiinflammatory role of MUC1 mucin during infection with nontypeable haemophilus influenzae. *Am. J. Respir. Cell Mol. Biol*. 46, 149–156. doi: 10.1165/rcmb.2011-0142OC

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Lee, S. H. (2015). Mechanisms of glucocorticoid action in chronic rhinosinusitis. *Allergy Asthma Immunol. Res*. 7, 534–537. doi: 10.4168/aair.2015.7.6.534

Li, Y., Dinwiddie, D. L., Harrod, K. S., Jiang, Y., and Kim, K. C. (2010). Anti-inflammatory effect of MUC1 during respiratory syncytial virus infection of lung epithelial cells *in vitro*. *Am. J. Physiol. Lung Cell. Mol. Physiol*. 298, L558–563. doi: 10.1152/ajplung.00225.2009

Liu, M., Guo, P., An, J., Guo, C., Lu, F., and Lei, Y. (2019). Genomewide profiling of lncRNA and mRNA expression in CRSwNP. *Mol. Med. Rep*. 19, 3855–3863. doi: 10.3892/mmr.2019.10005

Milara, J., Peiro, T., Armengot, M., Frias, S., Morell, A., Serrano, A., et al. (2015). Mucin 1 downregulation associates with corticosteroid resistance in chronic rhinosinusitis with nasal polyps. *J. Allergy Clin. Immunol*. 135, 470–476. doi: 10.1016/j.jaci.2014.07.011

Pathan, M., Keerthikumar, S., Ang, C. S., Gangoda, L., Quek, C. Y., Williamson, N. A., et al. (2015). FunRich: an open access standalone functional enrichment and interaction network analysis tool. *Proteomics* 15, 2597–2601. doi: 10.1002/pmic.201400515

Peng, Y., Zi, X. X., Tian, T. F., Lee, B., Lum, J., Tang, S. A., et al. (2019). Whole-transcriptome sequencing reveals heightened inflammation and defective host defence responses in chronic rhinosinusitis with nasal polyps. *Eur. Respir. J*. 54:1900732. doi: 10.1183/13993003.00732-2019

Phillips, J. E., Renteria, L., Burns, L., Harris, P., Peng, R., Bauer, C. M., et al. (2016). Btk inhibitor RN983 delivered by dry powder nose-only aerosol inhalation inhibits bronchoconstriction and pulmonary inflammation in the ovalbumin allergic mouse model of asthma. *J. Aerosol Med. Pulm. Drug Deliv*. 29, 233–241. doi: 10.1089/jamp.2015.1210

Plager, D. A., Kahl, J. C., Asmann, Y. W., Nilson, A. E., Pallanch, J. F., Friedman, O., et al. (2010). Gene transcription changes in asthmatic chronic rhinosinusitis with nasal polyps and comparison to those in atopic dermatitis. *PLoS ONE* 5:e11450. doi: 10.1371/journal.pone.0011450

Pudova, E. A., Kudryavtseva, A. V., Fedorova, M. S., Zaretsky, A. R., Shcherbo, D. S., Lukyanova, E. N., et al. (2018). HK3 overexpression associated with epithelial-mesenchymal transition in colorectal cancer. *BMC Genomics* 19 (Suppl. 3):113. doi: 10.1186/s12864-018-4477-4

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 43:e47. doi: 10.1093/nar/gkv007

Sekigawa, T., Tajima, A., Hasegawa, T., Hasegawa, Y., Inoue, H., Sano, Y., et al. (2009). Gene-expression profiles in human nasal polyp tissues and identification of genetic susceptibility in aspirin-intolerant asthma. *Clin. Exp. Allergy* 39, 972–981. doi: 10.1111/j.1365-2222.2009.03229.x

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13, 2498–2504. doi: 10.1101/gr.1239303

Shi, J. B., Fu, Q. L., Zhang, H., Cheng, L., Wang, Y. J., Zhu, D. D., et al. (2015). Epidemiology of chronic rhinosinusitis: results from a cross-sectional survey in seven Chinese cities. *Allergy* 70, 533–539. doi: 10.1111/all.12577

Stevens, W. W., Lee, R. J., Schleimer, R. P., and Cohen, N. A. (2015a). Chronic rhinosinusitis pathogenesis. *J. Allergy Clin. Immunol*. 136, 1442–1453. doi: 10.1016/j.jaci.2015.10.009

Stevens, W. W., Ocampo, C. J., Berdnikovs, S., Sakashita, M., Mahdavinia, M., Suh, L., et al. (2015b). Cytokines in chronic rhinosinusitis. Role in eosinophilia and aspirin-exacerbated respiratory disease. *Am. J. Respir. Crit. Care Med*. 192, 682–694. doi: 10.1164/rccm.201412-2278OC

Stevens, W. W., Schleimer, R. P., and Kern, R. C. (2016). Chronic rhinosinusitis with nasal polyps. *J. Allergy Clin. Immunol. Pract*. 4, 565–572. doi: 10.1016/j.jaip.2016.04.012

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 43 (Database issue), D447–D452. doi: 10.1093/nar/gku1003

Thomas, D. C. (2017). The phagocyte respiratory burst: historical perspectives and recent advances. *Immunol. Lett*. 192, 88–96. doi: 10.1016/j.imlet.2017.08.016

Tuo, Z., Zheng, X., Zong, Y., Li, J., Zou, C., Lv, Y., et al. (2020). HK3 is correlated with immune infiltrates and predicts response to immunotherapy in non-small cell lung cancer. *Clin. Transl. Med*. 10, 319–330. doi: 10.1002/ctm2.6

Wang, W., Gao, Z., Wang, H., Li, T., He, W., Lv, W., et al. (2016). Transcriptome analysis reveals distinct gene expression profiles in eosinophilic and noneosinophilic chronic rhinosinusitis with nasal polyps. *Sci. Rep*. 6:26604. doi: 10.1038/srep26604

Wang, X., Zhang, N., Bo, M., Holtappels, G., Zheng, M., Lou, H., et al. (2016). Diversity of TH cytokine profiles in patients with chronic rhinosinusitis: a multicenter study in Europe, Asia, and Oceania. *J. Allergy Clin. Immunol*. 138, 1344–1353. doi: 10.1016/j.jaci.2016.05.041

Wolf, A. J., Reyes, C. N., Liang, W., Becker, C., Shimada, K., Wheeler, M. L., et al. (2016). Hexokinase is an innate immune receptor for the detection of bacterial peptidoglycan. *Cell* 166, 624–636. doi: 10.1016/j.cell.2016.05.076

Workman, A. D., Kohanski, M. A., and Cohen, N. A. (2018). Biomarkers in chronic rhinosinusitis with nasal polyps. *Immunol. Allergy Clin. North Am*. 38, 679–692. doi: 10.1016/j.iac.2018.06.006

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118

Zhang, N., Van Zele, T., Perez-Novo, C., Van Bruaene, N., Holtappels, G., DeRuyck, N., et al. (2008). Different types of T-effector cells orchestrate mucosal

inflammation in chronic sinus disease. *J. Allergy Clin. Immunol*. 122, 961–968. doi: 10.1016/j.jaci.2008.07.008

Zhang, W., Luo, J., Dong, X., Zhao, S., Hao, Y., Peng, C., et al. (2019). Salivary microbial dysbiosis is associated with systemic inflammatory markers and predicted oral metabolites in non-small cell lung cancer patients. *J. Cancer* 10, 1651–1662. doi: 10.7150/jca.28077

Zheng, K., Hao, J., Xiao, L., Wang, M., Zhao, Y., Fan, D., et al. (2020). Expression of nicotinamide adenine dinucleotide phosphate oxidase in chronic rhinosinusitis with nasal polyps. *Int. Forum Allergy Rhinol*. 10, 646–655. doi: 10.1002/alr.22530

Zhou, X., Zhen, X., Liu, Y., Cui, Z., Yue, Z., Xu, A., et al. (2020). Identification of key modules, hub genes, and noncoding RNAs in chronic rhinosinusitis with nasal polyps by weighted gene coexpression network analysis. *Biomed. Res. Int*. 2020:6140728. doi: 10.1155/2020/6140728

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

The reviewer DW declared a shared affiliation, with no collaboration, with the authors YH, YZ, PW, KD, XW and LZ to the handling Editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 23 March 2021
doi: 10.3389/fgene.2021.594250

Systems Biology Guided Gene Enrichment Approaches Improve Prediction of Chronic Post-surgical Pain After Spine Fusion

Vidya Chidambaran[1], Valentina Pilipenko[2], Anil G. Jegga[3,4], Kristie Geisler[1] and Lisa J. Martin[2,3]*

[1] Department of Anesthesiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States, [2] Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States, [3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States, [4] Department of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Objectives: Incorporation of genetic factors in psychosocial/perioperative models for predicting chronic postsurgical pain (CPSP) is key for personalization of analgesia. However, single variant associations with CPSP have small effect sizes, making polygenic risk assessment important. Unfortunately, pediatric CPSP studies are not sufficiently powered for unbiased genome wide association (GWAS). We previously leveraged systems biology to identify candidate genes associated with CPSP. The goal of this study was to use systems biology prioritized gene enrichment to generate polygenic risk scores (PRS) for improved prediction of CPSP in a prospectively enrolled clinical cohort.

Methods: In a prospectively recruited cohort of 171 adolescents (14.5 ± 1.8 years, 75.4% female) undergoing spine fusion, we collected data about anesthesia/surgical factors, childhood anxiety sensitivity (CASI), acute pain/opioid use, pain outcomes 6–12 months post-surgery and blood (for DNA extraction/genotyping). We previously prioritized candidate genes using computational approaches based on similarity for functional annotations with a literature-derived "training set." In this study, we tested ranked deciles of 1336 prioritized genes for increased representation of variants associated with CPSP, compared to 10,000 randomly selected control sets. Penalized regression (LASSO) was used to select final variants from enriched variant sets for calculation of PRS. PRS incorporated regression models were compared with previously published non-genetic models for predictive accuracy.

Results: Incidence of CPSP in the prospective cohort was 40.4%. 33,104 case and 252,590 control variants were included for association analyses. The smallest gene set enriched for CPSP had 80/1010 variants associated with CPSP ($p < 0.05$), significantly higher than in 10,000 randomly selected control sets ($p = 0.0004$). LASSO selected 20 variants for calculating weighted PRS. Model adjusted for covariates including PRS had AUROC of 0.96 (95% CI: 0.92–0.99) for CPSP prediction, compared to 0.70 (95% CI: 0.59–0.82) for non-genetic model ($p < 0.001$). Odds ratios and positive regression coefficients for the final model were internally validated using bootstrapping:

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis (Beijing) Co., Ltd., China

Reviewed by:
Ali Salehzadeh-Yazdi,
University of Rostock, Germany
Ailan Wang,
Geneis (Beijing) Co., Ltd., China
Samuele Bovo,
University of Bologna, Italy

***Correspondence:**
Vidya Chidambaran
vidya.chidambaran@cchmc.org

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

Received: *12 August 2020*
Accepted: *02 March 2021*
Published: *23 March 2021*

Citation:
Chidambaran V, Pilipenko V, Jegga AG, Geisler K and Martin LJ (2021) Systems Biology Guided Gene Enrichment Approaches Improve Prediction of Chronic Post-surgical Pain After Spine Fusion. Front. Genet. 12:594250. doi: 10.3389/fgene.2021.594250

PRS [OR 1.98 (95% CI: 1.21–3.22); β 0.68 (95% CI: 0.19–0.74)] and CASI [OR 1.33 (95% CI: 1.03–1.72); β 0.29 (0.03–0.38)].

Discussion: Systems biology guided PRS improved predictive accuracy of CPSP risk in a pediatric cohort. They have potential to serve as biomarkers to guide risk stratification and tailored prevention. Findings highlight systems biology approaches for deriving PRS for phenotypes in cohorts less amenable to large scale GWAS.

Keywords: systems biology, genetics, polygenic risk score, chronic post-surgical pain, gene enrichment

INTRODUCTION

Chronic post-surgical pain (CPSP) is an underrecognized and undertreated problem with an incidence of 14.5–38% in children after major surgery, that significantly contributes to prolonged opioid use (Kain et al., 1996; Kehlet et al., 2006; Macrae, 2008; Rabbitts et al., 2017; Harbaugh et al., 2018). CPSP is defined as chronic pain that develops or increases intensity after a surgical procedure and persists beyond healing—at least 3 months after surgery (Werner and Kongsgaard, 2014). It has been recognized as a unique pain state recently in the International Classification of Diseases (ICD-11) (Schug et al., 2019). Chronic pain in adolescents leads to chronic pain in adults, imposes extraordinary annual costs on the health care system (Walker et al., 2010; Parsons et al., 2013), and negatively impacts physical and psychological health, leading to disability and depression (Hunfeld et al., 2001; Kashikar-Zuck et al., 2001; Fletcher et al., 2011). Hence, targeted, individualized preventive and therapeutic measures are needed to decrease CPSP occurrence. Development of such measures is impeded by the inability to accurately predict individual risk for CPSP.

Our previous studies investigating psychological and perioperative factors influencing pediatric CPSP showed that acute postoperative pain, surgical duration and psychological factors, such as those measured by the Childhood anxiety sensitivity index (CASI), are associated with CPSP risk in adolescents undergoing spine surgery (Chidambaran et al., 2017). However, these factors only explain 16% of variability in predicting CPSP, with medium accuracy (C-statistic 0.77). Thus, more accurate and objective biomarkers are needed to guide CPSP prevention and management.

Pain has a heritable component of up to 60%, suggesting incorporation of genetic factors may improve CPSP risk prediction. Our recent systematic literature review of genetic associations with CPSP (Chidambaran et al., 2019) showed that variants of several genes are associated with CPSP. However, any single variant had only a small effect size (Hoofwijk et al., 2016; Chidambaran et al., 2019). Since small effect sizes of single variants explain only a low percentage of the phenotypic variance, any one variant will not be useful at predicting risk. However, as individuals may harbor many variants each contributing modestly to risk, creating a risk score which accounts for the cumulative effect polygenic risk score (PRS) of many variants may better explain risk. PRS profiling has been shown to have translational potential as predictive, prognostic biomarkers (Muranen et al., 2016; Torkamani et al., 2018). Typically, the PRS builds off of the results of genome wide association studies (GWAS), whereby an individual's genetic risk is the sum of all their risk alleles weighted by significance of the corresponding allele (Andersen et al., 2017; Escott-Price et al., 2017). Accurate, generalizable PRS have shown potential to inform clinical practice in several fields (Torkamani et al., 2018; Sugrue and Desikan, 2019). In fact, US Preventive Services Task Force recommended use of PRS for risk prediction and screening prioritization in prostate cancer (Grossman et al., 2018). There is also a push to incorporate PRS in risk assessment for decision-making in cardiovascular disease, breast cancer and Alzheimer's disease (Maas et al., 2016; Knowles and Ashley, 2018; Tan et al., 2018). Richardson et al. (2019) used using UK Biobank data to analyze 162 GWAS-derived PRS for 551 heritable traits, and created an easily accessible web application—"An atlas of polygenic burden associations across the human phenome." Pain was not identified as a phenotype in this atlas.

While CPSP is an important clinical problem the lack of GWAS studies related to pediatric CPSP to inform PRS is a major barrier. The problem is there are no pediatric biobanks to our knowledge with this phenotype. Additionally, pediatric clinical cohorts with well characterized CPSP phenotypes that are adequately powered to achieve GWAS statistical significance are difficult to recruit as they must have surgery and long-term follow-up. Given the lack of GWAS based data and the likelihood of small effect sizes, additional approaches to deriving PRS are required for pediatric CPSP. We recently described a systems-biology approach to identify genes and genetic pathways involved in CPSP (Chidambaran et al., 2020). This approach allows prioritization of functionally associated genes, hence substantially decreases the burden of statistical power for gene association testing and overcomes sample size limitations. We hypothesized that combining systems biology with gene enrichment for associated variants will allow derivation of PRS, which will improve prediction of CPSP risk in conjunction with known psychosocial factors. Our research is unique and novel, and lays the foundation for further research of PRS as predictive biomarkers of chronic pain conditions and less accessible cohorts (Tracey et al., 2019).

MATERIALS AND METHODS

This genomics study has two components: the first being a bioinformatics-driven, systems-biology approach to identify, rank and prioritize new "candidate genes" associated with CPSP,

followed by a gene enrichment and association study in a prospectively recruited surgical cohort with penalized regression for PRS generation and evaluation.

Systems Biology Gene Prioritization

We previously conducted a literature-based systematic review of human clinical studies of genetic associations with CPSP. We conducted a search using electronic databases (including PubMed and MEDLINE) of full-text articles of human clinical studies (limited to English language—clinical trials, multicenter studies, observational studies, and twin studies reported between 01/2002 and 12/2017) evaluating genetic associations with CPSP (Chidambaran et al., 2019). We used the following search terms: ("postoperative pain" OR "postsurgical pain" OR "postoperative pain" OR "postsurgical pain" OR "postoperative analgesia" OR "postoperative opioid" OR "CPSP" OR "chronic postsurgical pain") AND (genetic association OR polymorphism OR variant OR "genotype" OR "Genome wide association" OR "SNP"). We included 21 full-text articles evaluating associations of 69 unique variants/haplotype with CPSP. Of these, variants of 31 genes involved in neurotransmission, pain signaling, immune responses and neuroactive ligand–receptor interaction, were found to be associated with CPSP (**Supplementary Table 1**). The results of the literature review including description of studies, genes, variants and outcomes are detailed elsewhere (Chidambaran et al., 2019). Using the literature derived genes (N = 31) as "training genes," we previously identified novel candidate genes based on their similarity scores ("guilt by association") to the curated training genes using ToppFun application of the Transcriptome Ontology Pathway PubMed based prioritization of genes (ToppGene) Suite, a one-stop portal of computational software tools for gene enrichment (Chen et al., 2009). Pathways based on training and top 10% candidate genes associated with CPSP are described in detail elsewhere (Chidambaran et al., 2020).

Here, as the next step, we used the curated training set (N = 31) and prioritized candidate genes (N = 1305) (henceforth referred to as the "case set" of genes) for association with and gene enrichment for CPSP in a prospective clinical cohort (**Figure 1**).

Prospective Clinical Study

An observational prospective cohort study was conducted in adolescents with idiopathic scoliosis undergoing posterior spine fusion using standard surgical techniques, anesthetic and pain protocols. Studies are registered with ClinicalTrials.gov (Identifier: NCT01839461, NCT01731873), and approved by the Institutional Review Board. Written informed consent was obtained from parents and assent was obtained from children before enrollment.

Inclusion Criteria

Healthy children, age 10–18 years, American Society of Anesthesiologists (ASA) Physical Status $\leq$ 2 (mild systemic disease), diagnosis of idiopathic scoliosis and/or kyphosis, scheduled to undergo elective spinal fusion.

Exclusion Criteria

Pregnant or breastfeeding females, obesity, diagnosis of chronic pain or opioid use in the past 6 months, hepatic/renal disease and/or developmental delays.

Data Collection

Following preoperative data were obtained: demographics (sex, age, race), weight, pain scores (numerical rating scale/0–10 NRS) (von Baeyer, 2009) and home medications. Questionnaires were administered preoperatively to assess functional disability (FDI) (Walker and Greene, 1991) and anxiety sensitivity (CASI) (Silverman et al., 1991). All patients received total intravenous anesthesia (propofol and remifentanil) and midazolam in the intraoperative period, followed by standardized doses of patient controlled analgesia (morphine or hydromorphone) in the postoperative period. Pertinent surgical details (duration and number of vertebral levels fused) and anesthetic data (propofol and remifentanil doses) were collected. Postoperatively, pain scores (every 4 h), doses of morphine equivalents administered [postoperative days (POD) 1 and 2] were recorded. Of note, CASI, surgical duration and acute postoperative pain were associated with CPSP in a sub set of this cohort (Chidambaran et al., 2017). After hospital discharge, at 6–12 months, patients were asked to rate their average pain score (NRS) over the previous week and to answer open-ended questions about nature and site of pain, use of medications/alternative therapies/physician consults for pain, and functional disability (FDI).

CPSP Outcome

CPSP outcome was evaluated as a continuous variable for systems biology prioritization and predictive model development (to maximize power) and dichotomous outcome was used for comparison of predictive models. *CPSP continuous outcome*: Actual NRS pain scores at 6–12 months after surgery. *CPSP dichotomous outcome,* determined by pain score $>$ 3/10 on a 11-point NRS (range 0–10) at 6–12 months after surgery (CPSP = yes) was used for final comparison of non-genetic versus PRS incorporated regression model to evaluate improvement in prediction characteristics. NRS for pain intensity has been validated as a pain measure in children aged 7–17 years (von Baeyer, 2009). NRS pain score $>$ 3 (moderate/severe pain) at 3 months has been described as a predictor for persistence of pain and has been associated with functional disability (Gerbershagen et al., 2011).

DNA Collection and Genotyping

Blood was drawn for genotyping upon intravenous line placement. DNA was isolated on the same day, and frozen at −20°C. Genotyping was done using the Illumina Human Omni5 v41-0 array (85 patients), Human Omni5Exome v41-1 (33 patients) and Infinium Omni5-4-v1 (53 patients). Arrays were changed due to availability of new arrays which had more overall and more functional single nucleotide polymorphisms (SNPs).

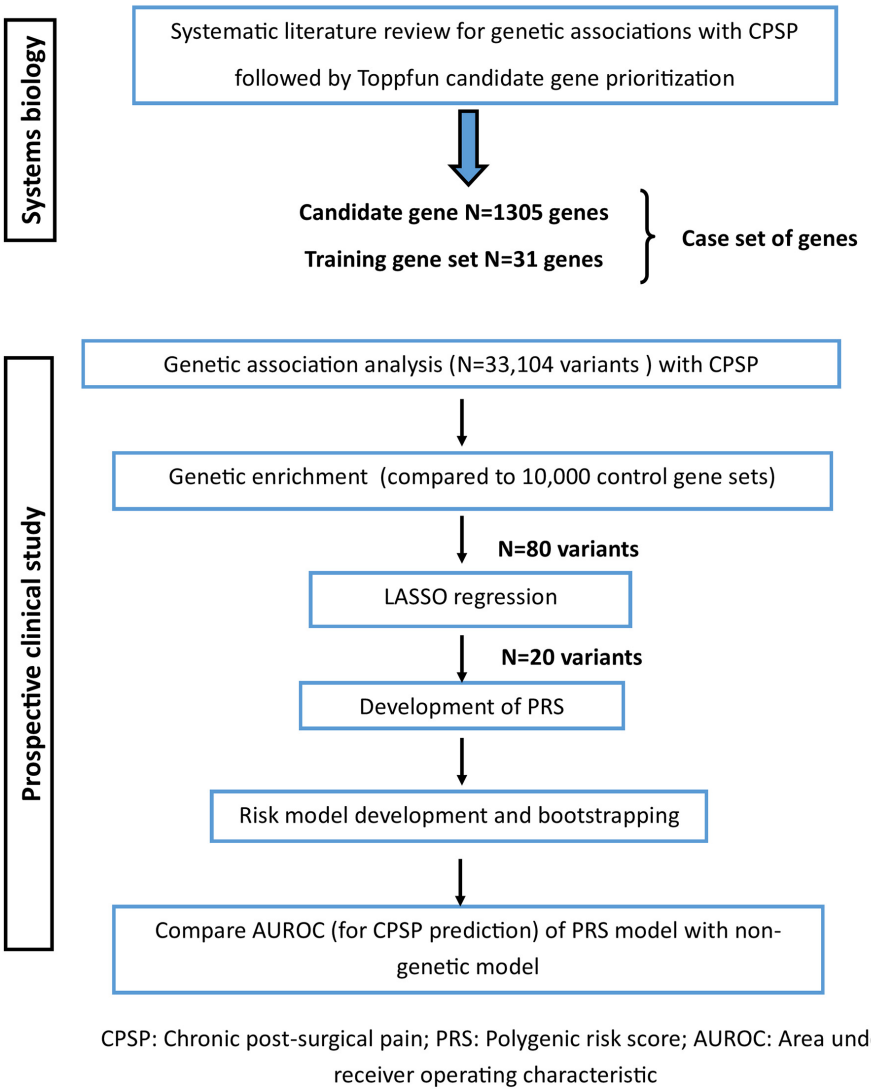

FIGURE 1 | Study flow showing steps involved with gene prioritization using systems biology followed by genetic association analyses in the clinical cohort to derive polygenic risk score based prediction model for chronic post-surgical pain.

Selection of Variants for Comparison of Case/Control Gene Sets

Only SNPs from autosomes were selected for analysis and were annotated using ANNOVAR software (Wang et al., 2010). All samples passed 95% threshold for call rates at genotype and individual levels. Genetic data was assessed for Hardy–Weinberg equilibrium (HWE) by means of goodness of fit χ^2-test with threshold for *p*-values 0.0001 (Wang et al., 2010). SNPs that were not associated with a specific gene according to ANNOVAR annotation were excluded prior to analysis. Low-frequency variants (minor allele frequency less than 10%) were also excluded (**Supplementary Figure 1**). There were 4,186,587 variants on the exome chip initially and 542,313 variants remained after exclusion. SNPs in high linkage disequilibrium (LD) (80%) were pruned out in PLINK (Purcell et al., 2007) using the command –indep-pairwise 50 5 0.8.

Procedure for SNP Selection for PRS

The first step to identify SNPs associated with CPSP was genetic association analyses. The next step was to narrow down the number of significant SNPs by enrichment analysis. The last step for identifying SNPs included in PRS calculation was Least Absolute Shrinkage and Selection Operator (LASSO) regressions. SNPs with non-zero coefficients were selected for PRS.

Genetic Association Analyses

Analyses were conducted using SAS 9.4 (SAS, Cary, NC) and R[1]. Prior to genetic analyses, cryptic relatedness was checked using Graphical Representation of Relationship (GRR) (Abecasis et al., 2001). Principal component analysis was employed to confirm European and African continental ancestry using 482 validated

[1]http://www.R-project.org

ancestry informative markers (Tandon et al., 2011). Concordance with self-reported race was > 95%. Given the concordance, race was used as a covariate in all the models and not principal components. To identify significant SNPs, we used linear models for association of each SNP with CPSP continuous outcome. In all association tests, we used an additive genetic model in which major homozygotes were coded as 0, heterozygotes as 1, and minor homozygotes as 2. Univariate analyses were conducted for CPSP outcomes with initial covariates (demographics, surgical duration, CASI, anesthetic doses, preoperative pain score), as suggested by non-genetic covariates based on our previous findings in a similar cohort (Chidambaran et al., 2017). Covariates significant in univariate analyses ($p < 0.1$) were included for genetic association analyses. PLINK v.07 was used for genetic association tests. Since the association results are only relevant for comparing the significant variants within the ranked case gene sets and those within the control sets for enrichment, they are not reported separately.

Gene Enrichment Analyses

Case gene variants were analyzed as sequence of cumulative sums of ranked variant sets with 10% increment, as has been done in a prior study (Kurowski et al., 2019). The first addend in each sequence was the training gene variant set. For each cumulative sum, we compared the number of associations in our case sets that met the $p < 0.05$ threshold to the number of associations meeting the same criteria in 10,000 matched runs of our control set of genes. SNPs from the control set were selected in the same ratio for minor allele frequency (MAF) as it was observed in the case set. Specifically, we used MAF bands as follows: 10–15%:15–20%:20–30%:30–50%. Empirical *p*-values of resampling tests were computed as follows: we calculated how many samples out of 10,000 had the number of significant SNPs equal to or greater than the number of significant SNPs from the case set and divided this number by 10,000. SNPs in case genes that formed the earliest cumulative group (where the number of significant SNPs were greater than in the matched control group) were considered as a minimal set of variants enriched for associations with corresponding outcomes.

LASSO Regression

To minimize risk of overfitting, we used penalized regression with LASSO in R software (package glmnet) *after* enrichment analyses (Friedman et al., 2010) with CPSP continuous and categorical outcome. SNPs in the genes identified in enrichment analysis were considered for penalized regression. Since penalized regression can be performed only on data without missing values we imputed missing genotypes using Michigan Imputation Server[2]. We imputed chromosomes where SNPs with missing genotypes were located. For each chromosome we submitted two VCF (Variant Call Format) files for subset of white patients and for subset of blacks and with admixture patients. VCF files were obtained from PLINK files using PLINK v1.9. Submitted to the server SNPs had 100% call rate. Both QC and imputation modes were used at the server. Genotypes for subset of white patients were imputed against the 1000G Phase 3 reference panel and the second subset of patients was imputed against the CAAPA African American reference panel. SNPs of interest were extracted from the files with imputed genotypes received from the server. Since SNPs with imputed genotypes overlapped with non-missing genotypes of original data these two types of genotypes (original and imputed) were used for evaluation of imputation accuracy. A controlling penalty parameter lambda for penalized regression was selected via cross-validation approach.

[2]https://imputationserver.sph.umich.edu

PRS Calculation

SNPs with non-zero coefficients in LASSO model were selected for PRS calculation. PRS was calculated as a weighted sum of products between number of risk alleles and their corresponding regression coefficients. The mathematical formula used for PRS calculation was given by the following equation

$$PRS_n = \sum_{i=1}^{m} (|b_i| * R_{i,n})$$

Where i is a number of SNPs, m is an upper range of SNPs participating in PRS calculation, n is a number of patients, PRS_n is a polygenic risk score for n-th patient, b_i is an absolute value of regression coefficient for each out of m SNPs from linear regression models for association of CPSP with a given SNP, $R_{i,n}$ is number of risk alleles for i-th SNP for n-th patient.

Regression Models

We built logistic regression models using stepwise approach including significant non-genetic predictors associated with CPSP ($p < 0.05$ selection criteria), followed by inclusion of PRS. For model performances, we used the area under the receiver operating characteristics (ROC) curve (AUC). AUCs with 95% confidence intervals for clinical and genetic models were used for model comparison in SAS 9.4 (SAS., Cary, NC).

Bootstrapping

While the optimal design for validation is to use an independent sample for validation, given the challenges in collecting such samples, we used the bootstrap method to internally validate the prediction. In this method, new bootstrap samples are generated from the original sample by random drawing with replacement multiple times (Efron, 1979). By bootstrapping across many iterations, the accuracy of parameter estimates can be determined. In this study, we empirically evaluate bias in the regression coefficients from the original model. Bootstrapping bias is a difference between the value obtained by using the original sample and the mean value obtained using bootstrap samples. At each iteration (n = 1,000), a random bootstrap sample (the same size as the original sample) was drawn with replacement from the original sample. Logistic models were generated for each bootstrap sample and bootstrapping results were compared with results from the original model. Regression coefficients and bootstrap confidence intervals are reported as linear terms and equivalent odds ratios. Bootstrapping was performed in R software (R Core Team, 2018) with the package boot (Davison and Hinkley, 1997; Freeman, 1998).

Power Analysis

For the gene set enrichment analyses, our goal was to determine if a set of selected genes/variants were more likely to show association ($p \leq 0.05$) than for a set of variants selected by chance. Out of 33,104 variants, we created deciles of variants, and the rates of associated variants compared each decile to 10,000 randomly selected gene sets of equal size. Based on the one sided proportion test, if we assumed that the background rate for association in the random set was 0.05, in the first decile containing 3310 SNPs, we would have 80% power to detect a difference between the SNPs in the selected genes if they were associated at a rate of 0.064 (OR = 1.3) at alpha = 0.05. Notably, the power calculation for gene enrichment was based on the number of SNPs rather than the number of individuals in the sample because we are comparing the rates of SNPs nominally associated between selected genes and random genes. For individual variants, we would have 80% power to detect an odds ratio as small as 2.1 at alpha = 0.05 and minor allele frequency 0.4. To evaluate the PRS risk score, we evaluated the score in 52 individuals with CPSP and 79 individuals without CPSP. With these numbers we would have 80% power to detect a PRS score difference as small as 2 at alpha = 0.05.

RESULTS

Prospective Cohort Characteristics

Demographics and summary of the variables examined for the prospective cohort are listed in **Table 1**. CPSP outcome was determined for 131 of the 171 patients (∼23% loss to follow-up). The flow diagram for recruitment is presented in **Supplementary Figure 2**. We examined the characteristics of both cohort of subjects lost to follow-up and the cohort of subjects followed for 6–12 months for all pertinent measures included in the models and did not find significant differences in terms of age (p = 0.390), sex (p = 0.361), race (0.906), CASI (p = 0.364), surgical duration (p = 0.322) and preoperative pain (p = 0.879). We found a 40.4% (53/131) incidence of CPSP. CPSP cases had significantly higher preoperative pain scores (p = 0.037) and CASI (p = 0.003) on univariate analyses and these factors were included

TABLE 1 | Baseline and pain follow-up characteristics of the surgical cohort, based on chronic post-surgical outcomes and univariate analyses of perioperative/psychological covariates.

Variable	Entire cohort (*N* = 171)		CPSP (dichotomous outcome)		*p*-value	Pain score at 6–12 months (continuous outcome)	
			CPSP Yes (*N* = 53)	CPSP No (*N* = 78)			
Demographics						Median (IQR)	*p*-value*
Sex F%	75.4%		81.0%	74.4%	0.365	2 (0–4)	0.331
Sex M%	24.6%		19.0%	25.6%		0 (0–4)	
Race (White %)	81.8%		77.4%	84.6%	0.292	1 (0–4)	0.844
Race (Non-white %)	18.2%		22.6%	15.4%		3 (0–4)	
	Mean	***SD***	**Mean (*SD*)**	**Mean (*SD*)**	***p*-value**	**Coefficient (*SE*)**	***p*-value****
Weight (Kg)	57.446	15.256	56.3 (14.2)	57.0 (14.5)	0.781	−0.055 (0.018)	0.323
Age (years)	14.488	1.840	14.7 (1.8)	14.5 (1.8)	0.462	0.184 (0.139)	0.189
Preoperative characteristics							
Preoperative pain score	0.596	1.282	0.3 (0.5)	0.1 (0.3)	0.037	1.210 (0.648)	0.065
CASI	28.552	5.531	30.6 (5.6)	26.8 (4.9)	0.003	0.147 (0.048)	0.003
Surgical/anesthesia characteristics							
Surgical duration	4.816	1.232	5.0 (1.4)	4.8 (1.2)	0.376	0.360 (0.214)	0.095
No. vertebral levels fused	11.506	1.969	11.0 (2.3)	11.6 (1.9)	0.115	0.006 (0.130)	0.963
Propofol dose mg/kg	71.791	27.186	79.5 (27.0)	73.7 (28.7)	0.238	0.014 (0.008)	0.091
Remifentanil dose mcg/kg	113.911	40.891	118.6 (41.5)	115.2 (44.2)	0.563	0.008 (0.006)	0.225
Acute postoperative pain characteristics							
AUC POD1–2	200.327	73.490	222.7 (75.9)	196.7 (66.8)	0.053	0.004 (0.003)	0.697
Morphine meq POD1–2 mg/kg	1.626	0.747	1.6 (0.7)	0.8 (0.1)	0.065	0.646 (0.349)	0.067
Pain follow-up at 6–12 months							
CPSP Y/No %	53/78 (40.5%)						
FDI score	4.485	5.321	6.7 (5.9)	2.3 (4.0)	0.002		
Pain score (NRS)	2.240	2.457	4.6 (2.0)	0.6 (1.0)	<0.001		

** Wilcoxon test.*

*** Simple linear regression.*

CASI, Childhood anxiety sensitivity index; AUC, Area under curve of pain scores over postoperative days (POD) 1 and 2; CPSP, Chronic post-surgical pain; FDI, Functional disability index.

as predictors in the regression model, and covariates for genetic association analyses.

Genetic Enrichment

After quality control and pruning as described under methods, 33,104 case variants and 252,590 control variants were included for covariate adjusted association analyses. Compared to the control set, there was enrichment of SNP associations in the training set for CPSP (**Figure 2**) but not for the other deciles of candidate gene variant sets. Of 1010 variants included in the training set, the number of variants (N = 80) associated with CPSP ($p < 0.05$) was significantly higher than in 10,000 randomly selected control sets (p = 0.0004). These 80 variants were annotated to the following 12 genes:ATXN1 (29); CACNG2 (2); CTSG (2); DRD2 (1); HLA-DQB1 (3); IL10 (1); KCNA1 (1); KCND2 (5); KCNJ3 (3); KCNJ6 (9); KCNK3 (2); PRKCA (22).

LASSO

Before LASSO, we imputed 45 genotypes in all (16 individual SNPs over 26 patients, where missing genotypes ranged from 1 to 10 at an individual level). One SNP rs17843723 form the HLA-DQB1 gene failed imputation and was excluded from consequent analysis. Imputation accuracy was 100% when we compared genotypes detected by chips with imputed genotypes. Number of genotypes for imputation accuracy evaluation was 2,051 (131 patients * 16 SNPs – 45 genotypes with missing values = 2,051 genotypes for accuracy evaluation). After LASSO, when CPSP was a continuous variable, the prediction set was comprised of 53 variants. LASSO regression with CPSP as a categorical variable resulted in 24 variants. We identified 20 variants that had non-zero coefficients in both linear and logistic penalized regression models. Chromosomal location, genetic annotation, function, MAF, odds ratios for CPSP and beta for NRS at 6–12 months with p-values for the LASSO selected variants are provided in **Table 2**. These 20 variants were annotated to nine genes: ATXN1 (7); CACNG2 (1); DRD2 (1); KCNJ3 (2); KCNJ6 (1); KCNK3 (1); PRKCA (7). Of these variants, rs7220480 was imputed for one individual, and rs2891519 and rs200369418 were imputed for three individuals.

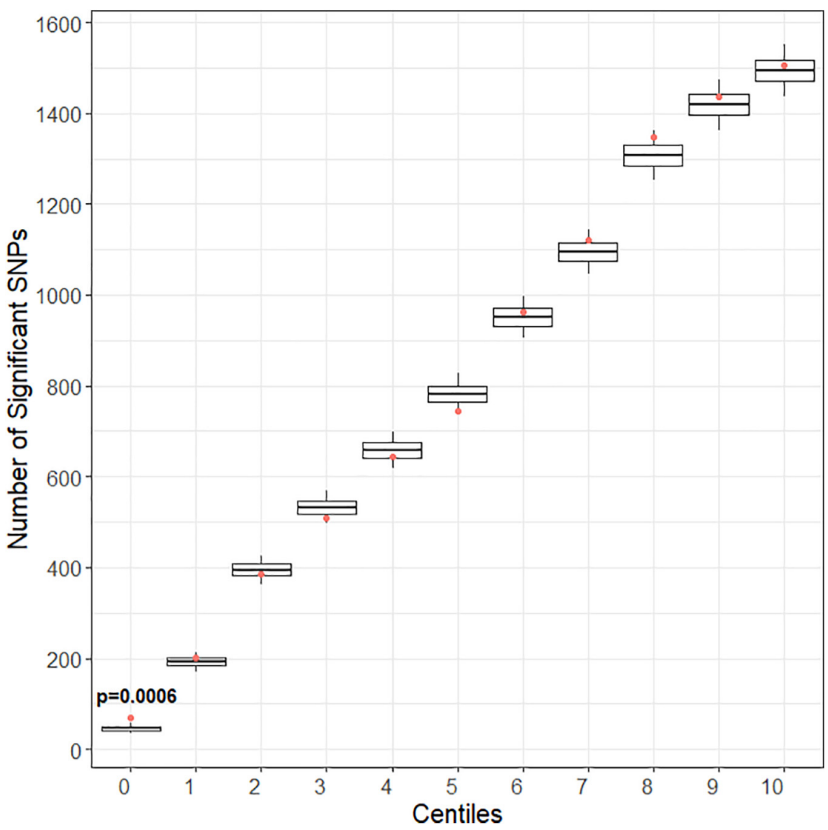

FIGURE 2 | Gene enrichment analyses for pain score at 6–12 months as outcome. Centiles represent the portion of case genes used in the genetic assocaition analysis. 0% includes the training set of gene variants, 10th percentile includes the training list plus the top 10% highest ranked genes, and so forth, vertical axis represents the number of variants. Box plots represent the cumulative number of SNPs with signficant association with pain score at 6–12 months after surgery [chronic post-surgical pain (CPSP) continuous outcome] ($p < 0.05$) in 10,000 runs of control gene variants. The dot indicates the cumulative number of nominal associations ($p < 0.05$) identified for case genes. Enrichment is indicated when a greater number of genetic associations are present in case versus control genes, that is, when the number of associations in case genes (red dot) (80 variants/1010 variants) exceeded the upper 95th percentile threshold in the 10,000 runs of the control set. For CPSP continuous outcome, we see enrichment in the training set of variants ($p < 0.001$). The training set incudes 80 variants showing association with CPSP ($p < 0.05$).

TABLE 2 | Genetic variants and risk alleles with regression coefficients included in the determination of polygenic risk score for prediction of chronic post-surgical pain.

SNP	Observed major allele	Observed minor allele	Gene	#Linear regression weight	*p*-value linear regression	Reference allele	Alternative allele	Function	Chr	Location (GRCh37)	Minor allele frequency
rs62069959	G*	A	PRKCA	2.299	0.001	C	T	Intronic	17	64318923	0.196
rs7125415	G	A*	DRD2	1.657	0.034	C	T	Intronic	11	113000000	0.126
rs61131185	A	G*	ATXN1	1.524	0.011	A	G	Intronic	6	16623387	0.322
rs12665284	G*	A	ATXN1	1.481	0.041	G	A	Intronic	6	16626066	0.146
rs202146909	A*	G	KCNJ3	1.414	0.042	T	C	Intronic	2	156000000	0.193
rs493352	G*	A	ATXN1	1.242	0.031	T	C	Intronic	6	16744169	0.488
rs9754467	A*	G	CACNG2	1.166	0.032	G	A	Intronic	22	37019059	0.222
rs12198202	A*	G	ATXN1	1.064	0.005	T	C	Intronic	6	16679771	0.424
rs11079653	T*	A	PRKCA	0.98	0.011	A	T	Intronic	17	64352329	0.202
rs2850125	G*	A	KCNJ6	0.936	0.046	C	T	Intronic	21	39130114	0.456
rs9914723	G	A*	PRKCA	0.917	0.004	G	A	Intronic	17	64716397	0.196
rs7220480 [1]	A	G*	PRKCA	0.857	0.048	A	G	Intronic	17	64686679	0.406
rs2891519 [2]	G	A*	KCNK3	0.835	0.008	G	A	Downstream	2	26954991	0.220
rs200369418 [2]	A*	C	PRKCA	0.816	0.028	C	A	Intronic	17	64762496	0.500
rs3812204	G	A*	ATXN1	0.789	0.038	G	A	Intronic	6	16698022	0.345
rs4716060	C	A*	ATXN1	0.772	0.038	C	A	Intronic	6	16310456	0.345
rs6459476	A	C*	ATXN1	0.736	0.048	A	C	Intronic	6	16618187	0.348
rs227912	A*	G	PRKCA	0.678	0.049	G	A	Intronic	17	64610729	0.246
rs744214	G*	A	PRKCA	0.634	0.017	G	A	Intronic	17	64334856	0.316
rs1992701	G	A*	KCNJ3	0.584	0.047	C	T	Intronic	2	156000000	0.453

PRKCA (protein kinase C alpha); DRD2 (dopamine receptor D2); ATXN1 (ataxin 1); KCNJ3 (potassium voltage-gated channel subfamily J member 3); CACNG2 (calcium voltage-gated channel auxiliary subunit gamma 2); KCNJ6 (potassium voltage-gated channel subfamily J member 6); KCNK3 (potassium two pore domain channel subfamily K member 3).

#Linear regression coefficients were used to calculate weighted polygenic risk scores; Beta > 0 is the selection criteria per LASSO.

**Risk allele; 1—imputed for one patient; 2—imputed for 3 patients.*

Polygenic Risk Scores

Weighted genetic risk was calculated from the 20 SNPs selected by LASSO regression models. PRS ranged from 10.1 to 30.6 (mean: 21.1; SD 4.0) and were normally distributed. The predicted probability (with 95% CI) of CPSP for a subject having a median (for the cohort) CASI = 28.3 using the regression model is plotted as a function of the PRS in **Figure 3**. The probability of CPSP is higher than 50% at a PRS > 23.06.

Regression Models

The non-genetic full and reduced model are presented in **Table 3**. The genetic model incorporating PRS in the non-genetic reduced model is also presented in **Table 3**. In the final model, both CASI and PRS remained significant predictors with Odds ratio (OR) of 1.37 (95% CI: 1.15–1.65) and 2.16 (95% CI: 1.53–3.05), respectively, for CPSP. In the final model, regression coefficients for CASI and PRS have means and standard errors for linear terms 0.32 ± 0.09 and 0.77 ± 0.18, respectively. Comparison of performance of the predictive model with clinical predictor (CASI) and performance of the predictive model with PRS (PRS and CASI) showed statistically significant higher performance of genetic model. Receiver operating characteristic curve was plotted showing that AUC for genetic model was 0.96 (95% CI: 0.92–0.99) compared to 0.70 (95% CI: 0.59–0.82) for non-genetic model ($p = 0.0001$) (**Figure 4**).

Bootstrapping

The final predictive model was evaluated by bootstrapping. Bootstrapping bias for means of linear terms were positive values for both CASI (0.03) and PRS (0.09) with standard errors 0.13 and 0.25 for means 0.29 and 0.68, respectively. Thus, bootstrap means for linear terms for CASI were 0.29 (0.32 minus 0.03) with 95% confidence interval 0.03–0.38 and for PRS 0.68 (0.77 minus 0.09) with 95% confidence interval 0.19–0.74. Confidence intervals for each regression coefficient obtained using bootstrapping serve as assessments for the model prediction accuracy. OR and 95% CI for CASI and PRS after bootstrapping remained similar to initial model results at 1.33 (95% CI: 1.03–1.72) and 1.98 (1.21–3.22), respectively. Bootstrapping bias means of linear terms, corresponding ORs with 95% CIs for regression coefficients are given in **Table 3**.

DISCUSSION

For phenotypes affected by difficulties in recruiting well powered and well characterized cohorts, novel methodologies are needed to address gaps in objective and accurate predictors. This is especially true for pediatric CPSP as it impedes targeted preventive efforts. By leveraging systems-biology and genetic testing approaches, we conducted enrichment analyses to derive PRS. They were calculated as weighted sum of products between number of risk alleles at 20 variants selected by LASSO

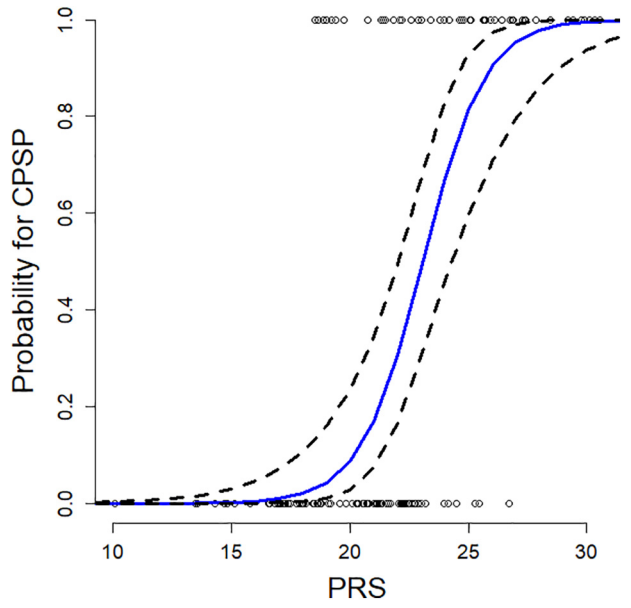

FIGURE 3 | Plot of predicted probability of developing chronic postsurgical pain (CPSP) after spine surgery is presented as a function of polygenic risk score (PRS), at a childhood anxiety sensitivity index (CASI) score of 28.3 (median CASI in the model). The blue line denotes predicted probabilities from the final regression model, and dashed lines the 95% confidence interval, and circles represent observed cases (or outcomes). We see a sigmoid shaped curve with increasing probability of CPSP at PRS > 16, 50% probability at PRS = 23.06 and high probability beyond PRS = 30. Thus, higher the weighted PRS, higher the probability of CPSP.

TABLE 3 | Multiple regression models evaluated for prediction of chronic post-surgical pain (CPSP) and results of bootstrapping.

Independent variable	OR	Lower 95% CI	Upper 95% CI	*P*-values
Full clinical model (AUC = 0.71)				
CASI	1.15	1.04	1.25	0.0038
Preoperative Pain	1.40	0.45	4.33	0.5559
Reduced clinical model (AUC = 0.70)				
CASI	1.15	1.04	1.26	0.0035
Genetic model (AUC = 0.96)				
Independent variable	**OR**	**Lower 95% CI**	**Upper 95% CI**	***P*-values**
CASI	1.37	1.15	1.65	0.0006
Weighted PRS	2.16	1.53	3.05	<0.0001
Bootstrapping results				
	OR (β)	**Lower 95%CI, OR (β)**	**Upper 95%CI OR (β)**	**Bias β**
CASI	1.33 (0.29)	1.03 (0.03)	1.72 (0.38)	0.03
Weighted PRS	1.98 (0.68)	1.21 (0.19)	3.22 (0.74)	0.09

CASI, Childhood anxiety sensitivity index; OR, Odds ratio; β, regression coefficients; AUC, Area under curve; PRS, Polygenic risk score; CI, confidence interval.

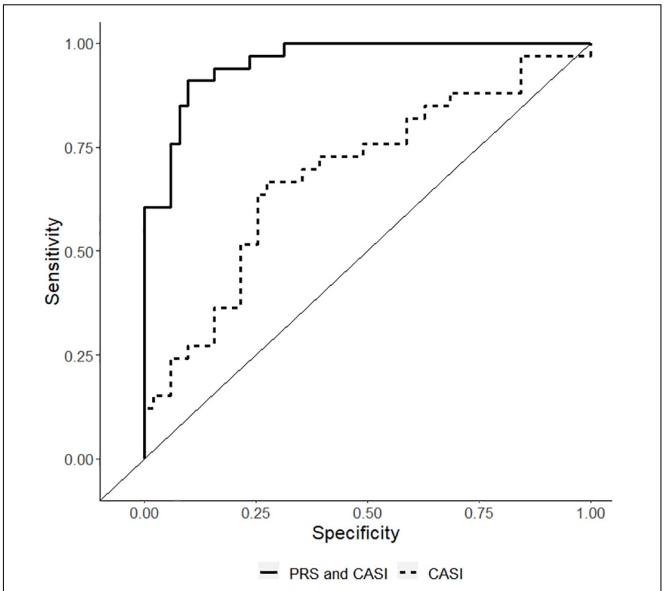

FIGURE 4 | Receiver operating characteristic curve showing the sensitivity/1-specificity for prediction of chronic post-surgical pain using the non-genetic model [including childhood anxiety sensitivity index (CASI) – dashed lines] compared with the prediction using the polygenic risk score final model (PRS and CASI – solid black lines). The area under curve for genetic model is 0.96 (95% CI: 0.92–0.99) compared to 0.70 (95% CI: 0.59–0.82) for non-genetic model (p = 0.0001).

regression, and their corresponding regression coefficients. We used bootstrapping to validate our final model's performance. Two factors—PRS and CASI—remained in the final risk model which predicted CPSP with higher accuracy compared to base non-genetic model (p = 0.0001). Since CPSP is a biopsychosocial phenomenon, it is not surprising that CASI, a psychological construct that measures interpretation of anxiety-related symptoms, remained a major risk predictor along with PRS. Higher anxiety sensitivity is associated with fear of pain, pain interference, which then leads to increased avoidance, disability (Martin et al., 2007) and maladaptive coping styles (Asmundson and Taylor, 1996), thus leading to the persistence of pain. Preoperative assessment of CASI will allow interventions such as education for improved coping, behavioral therapy and possibly use of anti-anxiolytics to temper the pain experience.

Scarcity of available genomic datasets for our phenotype of interest, namely, CPSP, makes GWAS daunting. Systems-biology approaches have been used successfully for identifying gene pathways implicated in other phenotypes (Kurowski et al., 2012; Jegga, 2014; Kurowski et al., 2019) as they allow leveraging known genomic data sources to prioritize functional genes for association, thereby decreasing the statistical burden. In our study, literature derived training sets showed enrichment for CPSP, with genes previously known to play an important role in pain. This either suggests that all relevant genes have been captured by the studies so far or that there are additional genes in very different pathways which need additional larger studies. Importantly, systems biology helped us identify control gene sets which allowed us to refine the optimal variants for PRS

determination by enrichment. Our findings are an important first step in the development of accurate and reliable gene-based biomarkers to predict susceptibility for CPSP. However, these findings will need external validation in unrelated similar and dissimilar surgical cohorts and diverse population structures. In addition, analytic validation of the panel in a CLIA-certified laboratory by re-sequencing and confirmation of the variants is necessary. Nevertheless, there is promising potential for future automated risk decision support based on preemptive genotyping and patient characteristics (CASI). This will allow preemptive preventive strategies to be employed cost-effectively, directed at those with higher risk.

The derived PRS is composed of weighted risk coefficients from 20 variants annotated to 7 genes which (not surprisingly) played a role in CPSP in previous studies: Ataxin-1 (*ATXN1),* Protein Kinase C Alpha (*PRKCA),* calcium channel genes (codes for the G subunit: *CACNG2),* Dopamine receptor gene (*DRD2*) and potassium channel genes (*KCNJ3, KCNJ6, KCNK3).* Potassium and calcium channel genes form the majority of genes involved. This is consistent with knowledge that these channels contribute to activation thresholds and spontaneous or exaggerated neuronal firing in response to noxious stimuli (Cohen and Mao, 2014). CPSP risk 6 months after breast cancer surgery has previously been reported for haplotype A2 rs3111020-rs11895478 G-A of *KCNJ3* and rs2835925 of *KCNJ6* (Langford et al., 2015). Similarly, in another cohort, several variants of the *CACNG2* gene were found to be associated with CPSP at a nominal level after breast cancer surgery (Nissenbaum et al., 2010). *PRKCA* is involved in long-term potentiation, an important process for memory and chronic pain development (Kawasaki et al., 2004; Price and Inyang, 2015). A meta-analysis showed that a recessive model of allele A in rs887797 in *PRKCA* was strongly associated with neuropathic CPSP in adults undergoing joint replacement surgery (Warner et al., 2017). *DRD2* variants were nominally associated with CPSP 4 months after different surgeries (Montes et al., 2015), as well as in chronic pain conditions (migraine) and substance abuse/addiction (Xu et al., 2004; Connor et al., 2007; Todt et al., 2009). Ataxin1 (*ATXN1*) is a gene that may play a role in transcription. Although its role in pain is not known, a study of a multiple surgery cohort found that the A allele at rs179997 of *ATXN1*was associated with CPSP at 4 months (Montes et al., 2015). Although variants selected for PRS in our study are mostly intronic, a functional assessment of the variants informing the PRS is not pertinent for establishing predictive biomarkers. However, intronic sequence alterations could influence gene function via altering binding sites for splicing co-factors or transcriptional enhancer/suppressor elements or may be in linkage with other variants with functional roles.

Since different surgeries are associated with variable pain modalities with different incidences of CPSP, the homogeneity of the surgical cohort in our study is a strength. The well characterized CPSP phenotypes, systematic approaches and bootstrapping add to the robustness of the results. Recent articles discuss clinical implementation of PRS may soon be a reality in cohorts with a higher prior probability of disease, to assist in risk/diagnosis or to inform treatment choices (Lewis and Vassos, 2020). We acknowledge that there are ethical and scientific challenges surrounding clinical implementation of PRS (Martin et al., 2019). Cost-benefit analyses for use of PRS in CPSP will need to consider (a) the prevalence of cohort at risk (In the US alone, 25 million adult and 5 million pediatric major surgeries are conducted per year (specifically, for spine surgery— according to the national scoliosis foundation, about 38,000 spine fusions are conducted in idiopathic scoliosis every year in the United States) (Sieberg et al., 2013) (b) the relative risk of phenotype predicted by PRS (in this study, RR∼2.2), (c) the proportion of surgical population at risk (in this study, ∼40%; the incidence of severe CPSP after major surgery is 2.2%—at a conservative estimate, this translates to 660,000 new cases of CPSP every year in the United States) (Fletcher et al., 2011), (d) the therapeutic response rate (CPSP is potentially preventable), and (e) the cost/impact of the condition being prevented (Gibson, 2019). Recent estimates suggest that CPSP incurs annual direct and indirect costs of US$11,846 and US$29,617, respectively, per patient (Parsons et al., 2013) and negatively impacts quality of life (Hunfeld et al., 2001; Kashikar-Zuck et al., 2001; Fletcher et al., 2011). Furthermore, the decreasing costs of genetic testing indicate that use of PRS will have benefits that outweigh risks/costs. Recent studies investigating preventive strategies like pregabalin have conflicting results (Mishriky et al., 2015; Thapa and Euasobhon, 2018)—this is not necessarily a function of therapeutic inefficacy—but could potentially be due to bias from inclusion of low risk subjects; hence, PRS could potentially improve evaluation of interventional strategies allowing *a priori* assessment of risk.

CONCLUSION

In conclusion, systems biology approaches combined with genetic association testing methodology are useful methods to develop PRS when GWAS approaches are not feasible. PRS holds future potential as a biomarker (simple blood test) that can predict CPSP risk. Given the morbidity associated with CPSP—including the risk for opioid abuse (Brummett et al., 2017), significant rates of chronic opioid dependence after surgery (Lee et al., 2017), the economic burden of CPSP—and decreasing genetic testing costs, we envision PRS to be cost-effective adjunct for risk stratification and clinical decision-making so preventive strategies can be targeted at those with high-risk. Future studies are needed to validate our findings. Our results may also have extended potential in predicting other chronic musculoskeletal pain conditions with similar pathophysiology.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the dbGaP production site at this url: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002105.v1.p1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Cincinnati Childrens Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

VC developed the research concept, oversaw data collection, recruitment, data analyses, and wrote the first draft of the manuscript. VP and LM conducted the statistical genetics analyses and revised the manuscript. AJ conducted the systems biology modeling. KG was the research coordinator who approached and recruited subjects for the prospective data analyses. All authors contributed to the article and approved the submitted version.

FUNDING

This project was supported by the 5K23HD082782 through the Eunice Kennedy Shriver National Institute of Child Health & Human Development, National Institutes of Health (PI: VC).

ACKNOWLEDGMENTS

We would like to thank Maria Ashton for editing the manuscript, and Bobbie Stubbeman, CCRC III for helping with subject recruitment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.594250/full#supplementary-material

REFERENCES

Abecasis, G. R., Cherny, S. S., Cookson, W. O. C., and Cardon, L. R. (2001). GRR: graphical representation of relationship errors. *Bioinform. Appl. Note* 17, 742–743.

Andersen, A. M., Pietrzak, R. H., Kranzler, H. R., Ma, L., Zhou, H., Liu, X., et al. (2017). Polygenic scores for major depressive disorder and risk of alcohol dependencepolygenic risk score analysis for depression and alcohol dependencepolygenic risk score analysis for depression and alcohol dependence. *JAMA Psychiatry* 74, 1153–1160. doi: 10.1001/jamapsychiatry.2017.2269

Asmundson, G. J., and Taylor, S. (1996). Role of anxiety sensitivity in pain-related fear and avoidance. *J. Behav. Med.* 19, 577–586.

Brummett, C. M., Waljee, J. F., Goesling, J., Moser, S., Lin, P., Englesbe, M. J., et al. (2017). New persistent opioid use after minor and major surgical procedures in us adults. *JAMA Surg.* 152, e170504. doi: 10.1001/jamasurg.2017.0504

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Chidambaran, V., Ashton, M., Martin, L. J., and Jegga, A. G. (2020). Systems biology-based approaches to summarize and identify novel genes and pathways associated with acute and chronic postsurgical pain. *J. Clin. Anesth.* 62:109738. doi: 10.1016/j.jclinane.2020.109738

Chidambaran, V., Ding, L., Moore, D. L., Spruance, K., Cudilo, E. M., Pilipenko, V., et al. (2017). Predicting the pain continuum after adolescent idiopathic scoliosis surgery: a prospective cohort study. *Eur. J. Pain* 21, 1252–1265. doi: 10.1002/ejp.1025

Chidambaran, V., Gang, Y., Pilipenko, V., Ashton, M., and Ding, L. (2019). Systematic review and meta-analysis of genetic risk of developing chronic postsurgical pain. *J. Pain* 21, 2–24. doi: 10.1016/j.jpain.2019.05.008

Cohen, S. P., and Mao, J. (2014). Neuropathic pain: mechanisms and their clinical implications. *BMJ* 348:f7656. doi: 10.1136/bmj.f7656

Connor, J. P., Young, R. M., Lawford, B. R., Saunders, J. B., Ritchie, T. L., and Noble, E. P. (2007). Heavy nicotine and alcohol use in alcohol dependence is associated with D2 dopamine receptor (DRD2) polymorphism. *Addict. Behav.* 32, 310–319. doi: 10.1016/j.addbeh.2006.04.006

Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. New York, NY: Cambridge University Press, 582.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7, 1–26. doi: 10.1214/aos/1176344552

Escott-Price, V., Shoai, M., Pither, R., Williams, J., and Hardy, J. (2017). Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol. Aging* 49, 214.e7–214.e11. doi: 10.1016/j.neurobiolaging.2016.07.018

Fletcher, D., Pogatzki-Zahn, E., Zaslansky, R., Meissner, W., and Pain Out, G. (2011). euCPSP: European observational study on chronic post-surgical pain. *Eur. J. Anaesthesiol.* 28, 461–462. doi: 10.1097/EJA.0b013e328344b4cd

Freeman, T. (1998). Bootstrap methods and their applications. *Interfaces* 28, 71–72.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

Gerbershagen, H. J., Rothaug, J., Kalkman, C. J., and Meissner, W. (2011). Determination of moderate-to-severe postoperative pain on the numeric rating scale: a cut-off point analysis applying four different methods. *Br. J. Anaesth.* 107, 619–626. doi: 10.1093/bja/aer195

Gibson, G. (2019). On the utilization of polygenic risk scores for therapeutic targeting. *PLoS Genet.* 15:e1008060. doi: 10.1371/journal.pgen.1008060

Grossman, D. C., Curry, S. J., Owens, D. K., Bibbins-Domingo, K., Caughey, A. B., Davidson, K. W., et al. (2018). Screening for prostate cancer: US preventive services task force recommendation statement. *JAMA* 319, 1901–1913. doi: 10.1001/jama.2018.3710

Harbaugh, C. M., Lee, J. S., Hu, H. M., McCabe, S. E., Voepel-Lewis, T., Englesbe, M. J., et al. (2018). Persistent opioid use among pediatric patients after surgery. *Pediatrics* 141:e20172439. doi: 10.1542/peds.2017-2439

Hoofwijk, D. M., van Reij, R. R., Rutten, B. P., Kenis, G., Buhre, W. F., and Joosten, E. A. (2016). Genetic polymorphisms and their association with the prevalence and severity of chronic postsurgical pain: a systematic review. *Br. J. Anaesth.* 117, 708–719. doi: 10.1093/bja/aew378

Hunfeld, J. A., Perquin, C. W., Duivenvoorden, H. J., Hazebroek-Kampschreur, A. A., Passchier, J., van Suijlekom-Smit, L. W., et al. (2001). Chronic pain and its impact on quality of life in adolescents and their families. *J. Pediatr. Psychol.* 26, 145–153.

Jegga, A. G. (2014). Candidate gene discovery and prioritization in rare diseases. *Methods Mol. Biol.* 1168, 295–312. doi: 10.1007/978-1-4939-0847-9_17

Kain, Z. N., Mayes, L. C., O'Connor, T. Z., and Cicchetti, D. V. (1996). Preoperative anxiety in children. Predictors and outcomes. *Arch. Pediatr. Adolesc. Med.* 150, 1238–1245.

Kashikar-Zuck, S., Goldschneider, K. R., Powers, S. W., Vaught, M. H., and Hershey, A. D. (2001). Depression and functional disability in chronic pediatric pain. *Clin. J. Pain* 17, 341–349.

Kawasaki, Y., Kohno, T., Zhuang, Z. Y., Brenner, G. J., Wang, H., Van Der Meer, C., et al. (2004). Ionotropic and metabotropic receptors, protein kinase A, protein kinase C, and Src contribute to C-fiber-induced ERK activation and cAMP response element-binding protein phosphorylation in dorsal horn

neurons, leading to central sensitization. *J. Neurosci.* 24, 8310–8321. doi: 10.1523/jneurosci.2396-04.2004

Kehlet, H., Jensen, T. S., and Woolf, C. J. (2006). Persistent postsurgical pain: risk factors and prevention. *Lancet* 367, 1618–1625. doi: 10.1016/S0140-6736(06)68700-X

Knowles, J. W., and Ashley, E. A. (2018). Cardiovascular disease: the rise of the genetic risk score. *PLoS Med.* 15:e1002546. doi: 10.1371/journal.pmed.1002546

Kurowski, B., Martin, L. J., and Wade, S. L. (2012). Genetics and outcomes after traumatic brain injury (TBI): what do we know about pediatric TBI? *J. Pediatr. Rehabil. Med.* 5, 217–231. doi: 10.3233/PRM-2012-0214

Kurowski, B. G., Treble-Barna, A., Pilipenko, V., Wade, S. L., Yeates, K. O., Taylor, H. G., et al. (2019). Genetic influences on behavioral outcomes after childhood TBI: a novel systems biology-informed approach. *Front. Genet.* 10:481. doi: 10.3389/fgene.2019.00481

Langford, D. J., Paul, S. M., West, C. M., Dunn, L. B., Levine, J. D., Kober, K. M., et al. (2015). Variations in potassium channel genes are associated with distinct trajectories of persistent breast pain after breast cancer surgery. *Pain* 156, 371–380. doi: 10.1097/01.j.pain.0000460319.87643.11

Lee, J. S., Hu, H. M., Edelman, A. L., Brummett, C. M., Englesbe, M. J., Waljee, J. F., et al. (2017). New persistent opioid use among patients with cancer after curative-intent surgery. *J. Clin. Oncol.* 35, 4042–4049. doi: 10.1200/JCO.2017.74.1363

Lewis, C. M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* 12:44. doi: 10.1186/s13073-020-00742-5

Maas, P., Barrdahl, M., Joshi, A. D., Auer, P. L., Gaudet, M. M., Milne, R. L., et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* 2, 1295–1302. doi: 10.1001/jamaoncol.2016.1025

Macrae, W. A. (2008). Chronic post-surgical pain: 10 years on. *Br. J. Anaesth.* 101, 77–86. doi: 10.1093/bja/aen099

Martin, A. L., McGrath, P. A., Brown, S. C., and Katz, J. (2007). Anxiety sensitivity, fear of pain and pain-related disability in children and adolescents with chronic pain. *Pain Res. Manag.* 12, 267–272.

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Mishriky, B. M., Waldron, N. H., and Habib, A. S. (2015). Impact of pregabalin on acute and persistent postoperative pain: a systematic review and meta-analysis. *Br. J. Anaesth.* 114, 10–31. doi: 10.1093/bja/aeu293

Montes, A., Roca, G., Sabate, S., Lao, J. I., Navarro, A., Cantillo, J., et al. (2015). Genetic and clinical factors associated with chronic postsurgical pain after hernia repair, hysterectomy, and thoracotomy: a two-year multicenter cohort study. *Anesthesiology* 122, 1123–1141. doi: 10.1097/aln.0000000000000611

Muranen, T. A., Mavaddat, N., Khan, S., Fagerholm, R., Pelttari, L., Lee, A., et al. (2016). Polygenic risk score is associated with increased disease risk in 52 Finnish breast cancer families. *Breast Cancer Res. Treat.* 158, 463–469. doi: 10.1007/s10549-016-3897-6

Nissenbaum, J., Devor, M., Seltzer, Z., Gebauer, M., Michaelis, M., Tal, M., et al. (2010). Susceptibility to chronic pain following nerve injury is genetically affected by CACNG2. *Genome Res.* 20, 1180–1190. doi: 10.1101/gr.104976.110

Parsons, B., Schaefer, C., Mann, R., Sadosky, A., Daniel, S., Nalamachu, S., et al. (2013). Economic and humanistic burden of post-trauma and post-surgical neuropathic pain among adults in the United States. *J. Pain Res.* 6, 459–469. doi: 10.2147/JPR.S44939

Price, T. J., and Inyang, K. E. (2015). Commonalities between pain and memory mechanisms and their meaning for understanding chronic pain. *Prog. Mol. Biol. Transl. Sci.* 131, 409–434. doi: 10.1016/bs.pmbts.2014.11.010

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R Core Team (2018). *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabbitts, J. A., Fisher, E., Rosenbloom, B. N., and Palermo, T. M. (2017). Prevalence and predictors of chronic postsurgical pain in children: a systematic review and meta-analysis. *J. Pain* 18, 605–614. doi: 10.1016/j.jpain.2017.03.007

Richardson, T. G., Harrison, S., Hemani, G., and Smith, G. D. (2019). An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* 8:e43657. doi: 10.7554/eLife.43657

Schug, S. A., Lavand'homme, P., Barke, A., Korwisi, B., Rief, W., Treede, R. D., et al. (2019). The IASP classification of chronic pain for ICD-11: chronic postsurgical or posttraumatic pain. *Pain* 160, 45–52. doi: 10.1097/j.pain.0000000000001413

Sieberg, C. B., Simons, L. E., Edelstein, M. R., DeAngelis, M. R., Pielech, M., Sethna, N., et al. (2013). Pain prevalence and trajectories following pediatric spinal fusion surgery. *J. Pain* 14, 1694–1702. doi: 10.1016/j.jpain.2013.09.005

Silverman, W. K., Fleisig, W., Rabian, B., and Peterson, R. A. (1991). Childhood anxiety sensitivity index. *J. Clin. Child Psychol.* 20, 162–168.

Sugrue, L. P., and Desikan, R. S. (2019). What are polygenic scores and why are they important? What are polygenic scores and why are they important? What are polygenic scores and why are they important? *JAMA* 321, 1820–1821. doi: 10.1001/jama.2019.3893

Tan, C. H., Fan, C. C., Mormino, E. C., Sugrue, L. P., Broce, I. J., Hess, C. P., et al. (2018). Polygenic hazard score: an enrichment marker for Alzheimer's associated amyloid and tau deposition. *Acta Neuropathol.* 135, 85–93. doi: 10.1007/s00401-017-1789-4

Tandon, A., Patterson, N., and Reich, D. (2011). Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. *Genet. Epidemiol.* 35, 80–83. doi: 10.1002/gepi.20550

Thapa, P., and Euasobhon, P. (2018). Chronic postsurgical pain: current evidence for prevention and management. *Korean J. Pain* 31, 155–173. doi: 10.3344/kjp.2018.31.3.155

Todt, U., Netzer, C., Toliat, M., Heinze, A., Goebel, I., Nurnberg, P., et al. (2009). New genetic evidence for involvement of the dopamine system in migraine with aura. *Hum. Genet.* 125, 265–279. doi: 10.1007/s00439-009-0623-z

Torkamani, A., Wineinger, N. E., and Topol, E. J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* 19, 581–590. doi: 10.1038/s41576-018-0018-x

Tracey, I., Woolf, C. J., and Andrews, N. A. (2019). Composite pain biomarker signatures for objective assessment and effective treatment. *Neuron* 101, 783–800. doi: 10.1016/j.neuron.2019.02.019

von Baeyer, C. L. (2009). Numerical rating scale for self-report of pain intensity in children and adolescents: recent progress and further questions. *Eur. J. Pain* 13, 1005–1007. doi: 10.1016/j.ejpain.2009.08.006

Walker, L. S., Dengler-Crish, C. M., Rippel, S., and Bruehl, S. (2010). Functional abdominal pain in childhood and adolescence increases risk for chronic pain in adulthood. *Pain* 150, 568–572. doi: 10.1016/j.pain.2010.06.018

Walker, L. S., and Greene, J. W. (1991). The functional disability inventory: measuring a neglected dimension of child health status. *J. Pediatr. Psychol.* 16, 39–58.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603

Warner, S. C., van Meurs, J. B., Schiphof, D., Bierma-Zeinstra, S. M., Hofman, A., Uitterlinden, A. G., et al. (2017). Genome-wide association scan of neuropathic pain symptoms post total joint replacement highlights a variant in the protein-kinase C gene. *Eur. J. Hum. Genet.* 25, 446–451. doi: 10.1038/ejhg.2016.196

Werner, M. U., and Kongsgaard, U. E. I. (2014). Defining persistent post-surgical pain: is an update required? *Br. J. Anaesth.* 113, 1–4. doi: 10.1093/bja/aeu012

Xu, K., Lichtermann, D., Lipsky, R. H., Franke, P., Liu, X., Hu, Y., et al. (2004). Association of specific haplotypes of D2 dopamine receptor gene with vulnerability to heroin dependence in 2 distinct populations. *Arch. Gen. Psychiatry* 61, 597–606. doi: 10.1001/archpsyc.61.6.597

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: 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