frontiers | Research Topics

Computational methods to analyze RNA data for human diseases

Edited by
Min Zeng, Pingjian Ding and Rui Yin

Published in
Frontiers in Genetics
Frontiers in Bioinformatics

ISSN 1664-8714
ISBN 978-2-8325-3419-9
DOI 10.3389/978-2-8325-3419-9

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 to analyze RNA data for human diseases

Topic editors

Min Zeng — Central South University, China

Pingjian Ding — Case Western Reserve University, United States

Rui Yin — University of Florida, United States

Citation

Zeng, M., Ding, P., Yin, R., eds. (2023). *Computational methods to analyze RNA data for human diseases*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-3419-9

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 22 August 2023
DOI 10.3389/fgene.2023.1270334

OPEN ACCESS

EDITED AND REVIEWED BY
Fangqing Zhao,
Beijing Institutes of Life Science (CAS), China

*CORRESPONDENCE
Pingjian Ding,
pxd210@case.edu
Min Zeng,
zengmin@csu.edu.cn
Rui Yin,
ruiyin@ufl.edu

RECEIVED 31 July 2023
ACCEPTED 14 August 2023
PUBLISHED 22 August 2023

CITATION
Ding P, Zeng M and Yin R (2023), Editorial: Computational methods to analyze RNA data for human diseases.
Front. Genet. 14:1270334.
doi: 10.3389/fgene.2023.1270334

Editorial: Computational methods to analyze RNA data for human diseases

Pingjian Ding[1]*, Min Zeng[2]* and Rui Yin[3]*

[1]Center for Artificial Intelligence in Drug Discovery, School of Medicine, Case Western Reserve University, Cleveland, OH, United States, [2]School of Computer Science and Engineering, Central South University, Changsha, China, [3]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States

KEYWORDS

RNA, miRNA, LnRNA, RNA virus, ceRNA network, machine learning/statistics, human disease

Editorial on the Research Topic
Computational methods to analyze RNA data for human diseases

RNA, as a type of nucleic acid, forms one of the four fundamental macromolecules crucial for all known life forms. Unlike DNA (Deoxyribonucleic Acid), which typically serves as the primary genetic material in cells, many viruses use RNA as their genetic material. RNA viruses are known for their ability to mutate rapidly, and the emergence of novel strains and variants (Yin et al., 2020) is potentially responsible for a wide range of diseases, leading to epidemics or pandemics such as swine-origin flu pandemic (Yin et al., 2018) and COVID-19 (V'kovski et al., 2021; Yin et al., 2018; Ding and Xu, 2023). In addition, RNA plays critical roles in various biological processes, including gene expression, protein synthesis (Frye et al., 2018). Understanding the mechanisms and roles of RNA in disease pathogenesis and progression is crucial for advancing our knowledge of human biology and developing optimized therapeutic strategies to combat RNA-related diseases. Computational approaches like machine learning and statistics, have captured much attention in this field due to increasingly available diverse RNA datasets (Yin et al., 2022; Li et al., 2023; Yin et al., 2023). This Research Topic of Frontiers in Genetics features a Research Topic of the latest advances in applying and developing various kinds of computational methods to analyze RNA data towards non-coding RNAs (e.g., miRNA, lncRNA) and RNA viruses (e.g., influenza, coronavirus).

The ncRNAs are crucial for regulating gene expression at both the transcriptional and posttranscriptional levels within the transcriptome, without encoding proteins (Winkle et al., 2021). In particular, miRNAs are a type of small, single-stranded noncoding RNAs, about 19–25 nucleotides long, that have highly conserved sequences and can regulate gene expression at the post-transcriptional level. Through extensive research on miRNA in the context of development and disease, it has emerged as a compelling target for innovative therapeutic approaches (Shen et al., 2020a; Shen et al., 2020b; Li Peng et al., 2022). In this Research Topic, Luo et al. presented a comprehensive perspective of recent progress in miRNA-targeted therapeutics employing machine learning techniques. In addition to discussing resources and preprocessing of pharmacogenomic data, they also presented the main machine learning algorithms employed in identifying miRNA-disease associations. Given the limitations of current methods in constructing negative sample sets, Wei et al.

introduced a clustering-based sampling approach called CSMDA to predict miRNA-disease associations. This method aims to address the Research Topic associated with negative sample selection in the context of miRNA-disease association prediction. Under a five-fold cross-validation, CSMDA computed an impressive Area Under the Curve (AUC) of 0.9610. Additionally, through validation with the dbDEMC database, it was confirmed that all predicted miRNAs, except hsa-mir-34c, were associated with colon cancer.

LncRNAs are a subset of ncRNAs characterized by their length, which exceeds 200 nucleotides. They have important functions in controlling gene expression at various levels, such as translational, transcriptional, and epigenetic processes (Qin et al., 2020). LncRNAs are crucial in controlling genes and proteins related to a range of human diseases like cancer (Xiao et al., 2018), digestive system Research Topic, and heart problems. Their role in disease regulation is well-established and holds promise for future therapies. Yao et al. proposed a computational model called GCHIRFLDA, which utilizes geometric complement heterogeneous information and random forest to predict lncRNA-disease associations. Under five-fold cross-validation, GCHIRFLDA achieved impressive performance metrics with an AUC of 0.9897 and an AUPR of 0. 7040. The study demonstrated that 18 of the predicted lncRNAs were validated through records present in databases or published literature. Meanwhile, the presence of inherent sparsity in known heterogeneous bio-data poses a challenge for computational methods aiming to enhance the accuracy of prediction. Thus, Zhang et al. explored a novel multiple mechanisms to discover underlying lncRNA-disease associations (MM-LDA). By integrating the graph attention network (GAT) and inductive matrix completion (IMC), this approach boosts the prediction accuracy. Firstly, a multiple-operator aggregation was created as part of the n-heads attention mechanism in the GAT. Then, IMC was incorporated into the improved node feature, and subsequently, the LDA network underwent a reconstruction to address the cold start problem caused by insufficient data in either whole rows or columns of a known association matrix. Under 5-fold cross-validation, an AUC of 0.9395 and an AUPR of 0.8057 were computed. The results from MM-LDA suggested a potential link between HOTAIR and HTTAS and gastric cancer.

In recent years, there has been the proposal of a hypothesis about competing endogenous RNA (ceRNA) network (Salmena et al., 2011). Under this hypothesis, lncRNAs possess the capability to function as endogenous molecular sponges for miRNAs, indirectly regulating the expression of messenger RNAs (mRNAs). The intricate nature of the lncRNA-miRNA-mRNA network makes their dysregulation closely linked to the progression and onset of various human diseases. For example, Ye et al. (2019) discovered that the lncRNA MIAT increases the expression of CD47 by acting as a sponge for miR-149-5p, leading to the inhibition of efferocytosis in advanced atherosclerosis. Yang et al. (2021) conducted a study uncovering the role of lncRNA XIST as a ceRNA, promoting atherosclerosis by upregulating TLR4 expression through the mediation of miR-599. Additionally, they identified several putative ceRNA networks, including those associated with implantation failure (Feng et al., 2018), polycystic ovary syndrome (Ma et al., 2021), and epithelial ovarian cancer (Zhao et al., 2019). Chen et al. employed the CIBERSORT algorithm to investigate the potential ceRNA-related mechanism of Peripheral arterial occlusive disease (PAOD) and to identify the associated patterns of immune cell infiltration. They developed an immune-related core ceRNA network that offered valuable insights into the molecular mechanisms underlying Peripheral Arterial Occlusive Disease (PAOD). This network consisting of CREB1, LINC00221, miR-20b-5p, and miR-17-5p, along with the infiltrating immune cells, specifically M1 macrophages and monocytes. Luo et al. introduced a lncRNA–mRNA network based on POI (POILMN) to identify essential lncRNAs. This research yielded a Research Topic of 288 differentially expressed mRNAs and 244 differentially expressed lncRNA. Ultimately, Through the application of topological analysis, POILMN identified four intersecting lncRNAs based on two centralities, namely, degree and betweenness.

CircRNA is a class of ncRNAs that forms a covalently closed loop structures (Li et al., 2020; Xiao et al., 2020; Peng et al., 2022; Peng et al., 2023). CircRNA molecules have been observed or artificially synthesized in various organisms, including mammals (Xu and Zhang, 2021) and viruses (Tan and Lim, 2021). The interactions between miRNAs and circRNAs have been demonstrated to modify gene expression and play a regulatory role in diseases. Therefore, He et al. introduced a novel approach called GCNCMI, which utilizes a graph convolutional neural (GCN) network to uncover latent associations between miRNAs and circRNAs. GCNCMI initially examines the underlying connections between neighboring nodes in the GCN network. Afterward, it iteratively spreads this connection information across the graph convolutional layers. Lastly, the embeddings produced by each layer were combined to output the ultimate prediction results. GCNCMI achieved an AUC of 0.9312 and an AUPR of 0.9412. The results from GCNCMI showed that 8 interactions involving hsa-miR-149-5p and 7 interactions involving hsa-miR-622 were validated.

Additionally, mitochondrial dysfunction could be among the molecular mechanisms implicated in obstructive sleep apnea (OSA) and its concurrent conditions. Despite several studies reporting the involvement of various proteins and miRNAs in OSA (Targa et al., 2020; Pinilla et al., 2021), the impact of OSA on genes and pathways, particularly concerning mitochondrial dysfunction, remains largely unexplored. In a previous study by Li et al. (2017), differentially expressed miRNAs were reported in OSA, but their specific association with mitochondrial dysfunction was not established. Liu et al.developed a novel diagnostic model consisting of a four-gene signature related to mitochondrial dysfunction. Using gene expression related to mitochondrial dysfunction, all samples were categorized into two clusters, with an additional subdivision of three clusters identified specifically among the samples with OSA. In the OSA samples compared to control samples, Significant differences were noted in the levels of M0 and M1 macrophages as well as plasma cells. Additionally, within the clusters associated with mitochondrial dysfunction in OSA samples, various immune cell types, particularly T cells, showed significant differences.

Although multiple databases offer information on virus-host protein interactions, they often lack detailed information about strain-specific virulence factors or the specific protein domains implicated in the interactions (Yin et al., 2017; Yin et al., 2021). Several databases may have incomplete representation coverage of influenza strains of influenza strains due to the challenge of sifting through extensive literature to gather comprehensive information. No existing database has provided complete records of strain-

specific protein-protein interactions for all types of Influenza A viruses. In particular, Ng et al. presented an innovative network that predicts domain-domain interactions between proteins from the mouse host and influenza A virus (IAV). By incorporating vital virulence details like lethal dose, this network facilitates a methodical exploration of disease factors. They created a network of interacting protein domains from both mouse and viral proteins, representing them as nodes and using weighted edges to show their interactions.

In summary, this Research Topic centers on the recent progress in utilizing and refining diverse computational methods, including machine learning and statistical techniques, to analyze RNA data related to RNA viruses and non-coding RNA. As a result, these analyses have delved into the biological disease mechanisms and aided in the understanding of human diseases, leading to improved preventive measures, diagnoses, and treatments.

Author contributions

PD: Conceptualization, Formal Analysis, Writing–original draft, Writing–review and editing. MZ: Conceptualization, Formal Analysis, Writing–original draft, Writing–review and editing. RY: Conceptualization, Funding acquisition, Writing–original draft, Writing–review and editing.

Funding

This study was partially supported by grants from Centers for Disease Control and Prevention (1U18DP006512), National Institute of Environmental Health Sciences (R21ES032762) and the NIH National Center for Advancing Translational Sciences (UL1TR001427).

Conflict of interest

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

Publisher's note

References

Ding, P., and Xu, R. (2023). Causal association of COVID-19 with brain structure changes: findings from a non-overlapping 2-sample mendelian randomization study. medRxiv 2023.07.16.23292735.

Feng, C., Shen, J. M., Lv, P. P., Jin, M., Wang, L. Q., Rao, J. P., et al. (2018). Construction of implantation failure related lncRNA-mRNA network and identification of lncRNA biomarkers for predicting endometrial receptivity. *Int. J. Biol. Sci.* 14, 1361–1377. doi:10.7150/ijbs.25081

Frye, M., Harada, B. T., Behm, M., and He, C. (2018). RNA modifications modulate gene expression during development. *Science* 361, 1346–1349. doi:10.1126/science.aau1646

Li, G., Luo, J., Wang, D., Liang, C., Xiao, Q., Ding, P., et al. (2020). Potential circRNA-disease association prediction using DeepWalk and network consistency projection. *J. Biomed. Inf.* 112, 103624. doi:10.1016/j.jbi.2020.103624

Li, K., Wei, P., Qin, Y., and Wei, Y. (2017). MicroRNA expression profiling and bioinformatics analysis of dysregulated microRNAs in obstructive sleep apnea patients. *Medicine* 96, e7917. doi:10.1097/MD.0000000000007917

Li, M., Zhao, B., Yin, R., Lu, C., Guo, F., and Zeng, M. J. (2023). GraphLncLoc: long non-coding RNA subcellular localization prediction using graph convolutional networks based on sequence to graph transformation. *Briefings Bioinforma.* 24, bbac565. doi:10.1093/bib/bbac565

Li Peng, Y. T., Huang, L., Yang, L., Fu, X., and Chen, X. (2022). Daestb: inferring associations of small molecule–miRNA via a scalable tree boosting model based on deep autoencoder. *Briefings Bioinforma.* 23, bbac478. doi:10.1093/bib/bbac478

Ma, Y., Ma, L., Cao, Y., and Zhai, J. (2021). Construction of a ceRNA-based lncRNA-mRNA network to identify functional lncRNAs in polycystic ovarian syndrome. *Aging (Albany NY)* 13, 8481–8496. doi:10.18632/aging.202659

Peng, L., Yang, C., Chen, Y., and Liu, W. (2023). Predicting CircRNA-disease associations via feature convolution learning with heterogeneous graph attention network. *IEEE J. Biomed. Health Inf.* 27, 3072–3082. doi:10.1109/JBHI.2023.3260863

Peng, L., Yang, C., Huang, L., Chen, X., Fu, X., and Liu, W. (2022). Rnmflp: predicting circRNA–disease associations based on robust nonnegative matrix factorization and label propagation. *Briefings Bioinforma.* 23, bbac155. doi:10.1093/bib/bbac155

Pinilla, L., Barbe, F., and De Gonzalo-Calvo, D. J. (2021). MicroRNAs to guide medical decision-making in obstructive sleep apnea: A review. *Sleep. Med. Rev.* 59, 101458. doi:10.1016/j.smrv.2021.101458

Qin, T., Li, J., and Zhang, K. Q. (2020). Structure, regulation, and function of linear and circular long non-coding RNAs. *Front. Genet.* 11, 150. doi:10.3389/fgene.2020.00150

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the rosetta stone of a hidden RNA language? *Cell* 146, 353–358. doi:10.1016/j.cell.2011.07.014

Shen, C., Luo, J., Lai, Z., and Ding, P. (2020a). Multiview joint learning-based method for identifying small-molecule-associated MiRNAs by integrating pharmacological, genomics, and network knowledge. *J. Chem. Inf. Model.* 60, 4085–4097. doi:10.1021/acs.jcim.0c00244

Shen, C., Luo, J., Ouyang, W., Ding, P., and Wu, H. (2020b). Identification of small molecule–miRNA associations with graph regularization techniques in heterogeneous networks. *J. Chem. Inf. Model.* 60, 6709–6721. doi:10.1021/acs.jcim.0c00975

Tan, K. E., and Lim, Y. (2021). Viruses join the circular RNA world. *FEBS J.* 288, 4488–4502. doi:10.1111/febs.15639

Targa, A., Dakterzada, F., Benítez, I., De Gonzalo-Calvo, D., Moncusí-Moix, A., López, R., et al. (2020). Circulating MicroRNA profile associated with obstructive sleep apnea in alzheimer's disease. *Mol. Neurobiol.* 57, 4363–4372. doi:10.1007/s12035-020-02031-z

V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., and Thiel, V. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. *Nat. Rev. Microbiol.* 19, 155–170. doi:10.1038/s41579-020-00468-6

Winkle, M., El-Daly, S. M., Fabbri, M., and Calin, G. (2021). Noncoding RNA therapeutics—challenges and potential solutions. *Nat. Rev. Drug Discov.* 20, 629–651. doi:10.1038/s41573-021-00219-z

Xiao, Q., Luo, J., Liang, C., Li, G., Cai, J., Ding, P., et al. (2018). Identifying lncRNA and mRNA co-expression modules from matched expression data in ovarian cancer. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17, 623–634. doi:10.1109/TCBB.2018.2864129

Xiao, Q., Yu, H., Zhong, J., Liang, C., Li, G., Ding, P., et al. (2020). An *in-silico* method with graph-based multi-label learning for large-scale prediction of circRNA-disease associations. *Genomics* 112, 3407–3415. doi:10.1016/j.ygeno.2020.06.017

Xu, C., and Zhang, J. (2021). Mammalian circular RNAs result largely from splicing errors. *Cell Rep.* 36, 109439. doi:10.1016/j.celrep.2021.109439

Yang, K., Xue, Y., and Gao, X. (2021). LncRNA XIST promotes atherosclerosis by regulating miR-599/TLR4 axis. *Inflammation* 44, 965–973. doi:10.1007/s10753-020-01391-x

Ye, Z. M., Yang, S., Xia, Y. P., Hu, R. T., Chen, S., Li, B. W., et al. (2019). LncRNA MIAT sponges miR-149-5p to inhibit efferocytosis in advanced atherosclerosis through CD47 upregulation. *Cell death Dis.* 10, 138. doi:10.1038/s41419-019-1409-4

Yin, R., Luo, Z., Zhuang, P., Lin, Z., and Kwoh, C. (2021). VirPreNet: a weighted ensemble convolutional neural network for the virulence prediction of influenza A virus using all eight segments. *Bioinformatics* 37, 737–743. doi:10.1093/bioinformatics/btaa901

Yin, R., Luo, Z., Zhuang, P., Zeng, M., Li, M., Lin, Z., et al. (2023). ViPal: a framework for virulence prediction of influenza viruses with prior viral knowledge using genomic sequences. *J. Biomed. Inf.* 142, 104388. doi:10.1016/j.jbi.2023.104388

Yin, R., Luusua, E., Dabrowski, J., Zhang, Y., and Kwoh, C. (2020). Tempel: time-series mutation prediction of influenza A viruses via attention-based recurrent neural networks. *Bioinformatics* 36, 2697–2704. doi:10.1093/bioinformatics/btaa050

Yin, R., Tran, V. H., Zhou, X., Zheng, J., and Kwoh, C. (2018). Predicting antigenic variants of H1N1 influenza virus based on epidemics and pandemics using a stacking model. *PloS one* 13, e0207777. doi:10.1371/journal.pone.0207777

Yin, R., Zhou, X., Ivan, F. X., Zheng, J., Chow, V. T., and Kwoh, C. K. (2017). "Identification of potential critical virulent sites based on hemagglutinin of influenza a virus in past pandemic strains," in Proceedings of the 6th International Conference on Bioinformatics and Biomedical Science, Singapore, June 22 - 24, 2017, 30–36.

Yin, R., Zhu, X., Zeng, M., Wu, P., Li, M., and Kwoh, C. (2022). A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods. *Briefings Bioinforma.* 23, bbac281. doi:10.1093/bib/bbac281

Zhao, X., Tang, D. Y., Zuo, X., Zhang, T. D., and Wang, C. (2019). Identification of lncRNA–miRNA–mRNA regulatory network associated with epithelial ovarian cancer cisplatin-resistant. *J. Cell. physiology* 234, 19886–19894. doi:10.1002/jcp.28587

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2022
doi: 10.3389/fgene.2022.959701

GCNCMI: A Graph Convolutional Neural Network Approach for Predicting circRNA-miRNA Interactions

Jie He[1], *Pei Xiao*[1], *Chunyu Chen*[1], *Zeqin Zhu*[1], *Jiaxuan Zhang*[2] *and Lei Deng*[1]*

[1]School of Computer Science and Engineering, Central South University, Changsha, China, [2]Department of Electrical Engineering, University of California, San Diego, San Diego, CA, United States

The interactions between circular RNAs (circRNAs) and microRNAs (miRNAs) have been shown to alter gene expression and regulate genes on diseases. Since traditional experimental methods are time-consuming and labor-intensive, most circRNA-miRNA interactions remain largely unknown. Developing computational approaches to large-scale explore the interactions between circRNAs and miRNAs can help bridge this gap. In this paper, we proposed a graph convolutional neural network-based approach named GCNCMI to predict the potential interactions between circRNAs and miRNAs. GCNCMI first mines the potential interactions of adjacent nodes in the graph convolutional neural network and then recursively propagates interaction information on the graph convolutional layers. Finally, it unites the embedded representations generated by each layer to make the final prediction. In the five-fold cross-validation, GCNCMI achieved the highest AUC of 0.9312 and the highest AUPR of 0.9412. In addition, the case studies of two miRNAs, hsa-miR-622 and hsa-miR-149-5p, showed that our model has a good effect on predicting circRNA-miRNA interactions. The code and data are available at https://github.com/csuhjhjhj/GCNCMI.

Keywords: circRNA, miRNA, deep learning, graph convolution neural network, circRNA-miRNA interaction

OPEN ACCESS

Edited by:
Pingjian Ding,
Case Western Reserve University, United States

Reviewed by:
Guanghui Li,
East China Jiaotong University, China
Weidun Xie,
City University of Hong Kong, Hong Kong SAR, China

***Correspondence:**
Lei Deng
leideng@csu.edu.cn

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

Received: *02 June 2022*
Accepted: *23 June 2022*
Published: *05 August 2022*

Citation:
He J, Xiao P, Chen C, Zhu Z, Zhang J and Deng L (2022) GCNCMI: A Graph Convolutional Neural Network Approach for Predicting circRNA-miRNA Interactions. Front. Genet. 13:959701. doi: 10.3389/fgene.2022.959701

1 INTRODUCTION

Non-coding RNA (ncRNA) refers to various RNA molecules that will not translate into a protein. There has been much agreement through numerous studies that ncRNA has monumental biological functions though it only part a small fraction of the genomes. Since the discovery of RNA and ribosomal RNA in the 1950s, non-coding RNA that plays a biological role has been known for 60 years (Palazzo and Lee, 2015). As well as their roles at the transcriptional and post-transcriptional levels, ncRNA plays a critical role in epigenetic regulation of gene expression. The recent finding suggests that some of these RNAs are also involved in translation and splicing (Steitz and Moore, 2003; Butcher and Brow, 2005; Gesteland et al., 2006).

MicroRNA (miRNA) was discovered in 1993 by the Ambros and Ruvkun groups in *Caenorhabditis elegans* (Lee et al., 1993) and brought a revolution to molecular biology. They are small single-stranded molecules that derive from transcripts' unique hairpin structures called pre-miRNA. Most miRNAs are transcribed from DNA sequences into primary miRNAs, then processed into precursor miRNAs and become mature miRNAs finally (O'Brien et al., 2018; Liu et al., 2021). Furthermore, miRNAs have been found to regulate gene expression post-transcriptionally by

affecting mRNA translation, implying that dysregulation of miRNAs may be associated with various diseases by affecting gene expression (Bartel, 2004). For instance, recent studies showed approximately 50% of annotated human miRNAs are located in cancer-associated regions of the genome called fragile sites. This indicated that miRNA plays a crucial role in cancer progression (Calin et al., 2004).

Circular RNA consists of large non-coding RNAs produced by a non-canonical splicing event called back splicing. They are ubiquitous in species ranging from viruses to mammals during post-transcriptional processes. Viroids are the first circRNA to be discovered, though they are not produced by a back splicing mechanism (Sanger et al., 1976). A few years later, most circRNAs are observed in the cytoplasm and some small fractions in the nucleus. Circular forms of RNAs were observed or synthesized in diverse species such as viruses (Kos et al., 1986), prokaryotes (Ford and Ares, 1994), unicellular eukaryotes (Grabowski et al., 1981), and mammals (Capel et al., 1993). Most circRNA are expressed from known encoding proteins, composed of single or multiple exons. With the progress of high-throughput RNA-sequencing and bioinformatics tools, scientists have found the human transcriptome's general feature ubiquitous in many other metazoans.

A diverse set of circRNAs have been identified as having functions such as sponges, decoys, or translatable elements that alter gene or protein expression. Biological functions of circRNAs have only been investigated for a small fraction, while most of which are proposed as miRNA sponges (Hansen et al., 2013; Memczak et al., 2013; Deng et al., 2022). Sponging up miRNA and interacting with RNA-binding proteins (RBP), circRNA plays many pathological functions like regulating miRNA activity. He et al. (He et al., 2022) performed circRNA microarray analysis and found its expression profile in diabetes. By acting as microRNA sponges for miR-7 (ciRS-7) and miR-124-3p and miR-338-3p (circHIPK3), ciRS-7 and circHIPK3 promote insulin secretion. circRNAs were identified in cancers, so it also proposed to play a crucial role in the intimation and development of tumors. (Ashwal-Fluss et al., 2014; Dong et al., 2017; Soslau, 2018). Most studies focus on the role of circRNA in tumors. circRNA was described as oncogenes. Diverse cellular functions of circRNA suggest their potential for cancer treatment as biomarkers and therapeutic targets (Chen and Huang, 2018; Li et al., 2019).

The interactions between circRNA and miRNA have been gradually discovered in recent years, and some related databases have been established. The CircR2Cancer database (Lan et al., 2020) contains 1,439 interactions between 1,135 circRNAs and 82 cancers. In addition, the database also includes basic information such as detection methods and expression patterns of circRNAs. However, there are few datasets on direct circRNA-miRNA interactions. Moreover, the known interactions are only a tiny part. Discovering the interactions between circRNAs and miRNAs is beneficial to understanding the interactions between circRNA and miRNA and disease. Using biological experiments to verify the interactions between circRNA-miRNA is time-consuming and labor-intensive. Computational methods can be used to mine the interactions between circRNA-miRNA more effectively. Still, there is little work to predict the circRNA-miRNA interactions.

As far as we know, GCNCMI is the first method to predict the circRNA-miRNA interactions, but other methods in the field of bioinformatics are still worth reference. Many methods based on computational interactions have recently achieved good results in predicting microbe-disease interactions and ncRNA-disease interactions. AE-RF (Deepthi and Jereesh, 2021) build an autoencoder to mine potential interaction features and then train a random forest model to predict circRNA-disease interactions. The DMFMDA (Liu et al., 2020) uses one-hot encoding of diseases and microorganisms to convert a vector representation in a low-dimensional space by embedding the propagation layer. The obtained vector representation is then input into a multi-layer neural network, and the parameters of the neural network are continuously optimized through Bayesian sorting to achieve accurate prediction. Deng et al. (Deng et al., 2020) constructed a meta-pathway-based circRNA-disease feature vector. This vector representation combines multiple similarities such as circRNA similarity, disease similarity, etc. The prediction is finally achieved using a random forest classifier. KATZHMDA (Chen et al., 2017) predicts the interactions between unknown microbes and disease by the Gaussian kernel similarity between known microbes and disease. NTSHMDA (Luo and Long, 2018) constructs a disease-microbe heterogeneity network based on the known similarity between microorganisms and diseases and assigns equal weights to known disease-microbe interactions according to the different contributions of diseases and microorganisms, which is conducive to reducing prediction error. Liu et al. (Dayun et al., 2021) established a multi-component graph attention network, which first passed a decomposer to identify node-level feature vectors, then combined the feature vectors to obtain a unified embedding vector, which was finally input into a fully connected network to predict microorganisms unknown interactions with the disease. SDLDA (Zeng et al., 2020) extract the linear and nonlinear interactions between lncRNA and diseases through singular value decomposition and neural network and finally unites the linear and nonlinear features into a new feature vector, which is input to the fully connected layer to realize prediction.

Although the above methods have achieved good prediction results, there are still some problems that will affect mining efficiency. Some existing association prediction methods rely on known similarities, but it is difficult to construct such similarities with the increasing number of miRNAs and circRNAs. There are far fewer known associations than unknown associations. Therefore, these methods are unsuitable when the circRNA and miRNA data increase. When the scale of data increases, how to mine the higher-order interactions of circRNA-miRNA is an urgent problem to be solved. In this paper, we construct a bipartite graph to describe the interaction information between circRNA and miRNA using known relationship pairs of them. Then we develop a graph convolutional network method to mine the deep semantic information that carries collaborative signal in the bipartite graph. We propagate the information flow recursively over the graph structure and continuously aggregate the interactive information between nodes to refine the embedding of each

TABLE 1 | The number of circRNAs, miRNAs, and circRNA-miRNA interactions included in the dataset.

circRNA	miRNA	interactions	unlabled interactions
2,115	821	9,589	9,589

node. Finally, We concatenate the embeddings generated by each layer to predict the relationship of unknown circRNA-miRNA pairs. Experimental results show that our GCNCMI model outperforms the other six state-of-the-art methods.

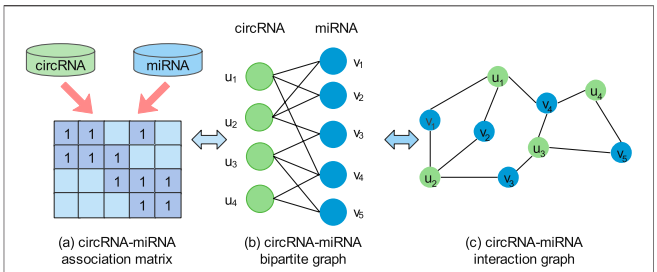


FIGURE 1 | An illustration of the circRNA-miRNA association matrix **(A)**, a bipartite graph **(B)**, and an interaction graph **(C)**.

2 MATERIALS AND METHODS

2.1 Datasets

We built the benchmark dataset from the circBank database (Liu et al., 2019). circBank contains 140,790 circRNAs. Each circRNA collects information such as miRNA binding sites, protein-coding ability, etc. We removed redundant parts of the dataset and extracted 2,115 circRNAs and 821 miRNAs from the circBank database, including 9,589 known circRNA-miRNA interactions. It now can be downloaded on the website http://www.circbank.cn/downloads.html. In addition, we randomly selected 9,589 unlabeled samples from the benchmark dataset. The detailed information can be seen in **Table 1**.

2.2 Problem Description

Our work aims to predict unknown relationships based on known circRNA and miRNA relationships. We use $U = \{u_1, u_2 \ldots u_n\}$ and $V = \{v_1, v_2, \ldots, v_m\}$ to respectively represent the collection of n circRNAs and m miRNAs, and use the interaction matrix $R \in \mathbb{R}^{n \times m}$ to represent the relationship between them. If the circRNA u_i is related to miRNA v_j, then the $R_{ij} = 1$, otherwise $R_{ij} = 0$. It should be noted that $R_{ij} = 0$ can only indicate that the two RNAs have not yet found a relationship, but may actually be related.

2.3 Graph Construction

We use a bipartite graph G(U ∪ V, E) constructed by the interaction matrix $R \in \mathbb{R}^{n \times m}$ to show the relationship between circRNAs and miRNAs, where U, V are the vertex sets denoting the circRNAs and miRNAs, and E is the edge sets constructed from the association matrix $R \in \mathbb{R}^{n \times m}$. This bipartite graph can be expanded into a complex interaction graph as shown in **Figure 1**. This interaction graph contains the higher-order interaction information of circRNA and miRNA, from which we can mine deep semantic information that carry collaborative signal. For example, the path $u_1 - v_1 - u_2$ and $u_1 - v_2 - u_2$ indicate the behavior similarity between u_1 and u_2, as both circRNAs have interacted with v_1 and v_2. Then, the interaction between u_2 and v_3 suggests that u_1 and v_3 are likely to be related.

2.4 GCNCMI

To capture the deep interaction information embedded in the interaction graph, we model the high-order interaction information of circRNA-miRNA in the embedding function. We propagate the information flow recursively over the graph structure and continuously aggregate the information of neighboring nodes to refine the embedding representation of the nodes (Hamilton et al., 2017; Xu et al., 2018; Wang et al., 2019). The architecture of our proposed GCNCMI model is shown in **Figure 2**. There are three parts to the framework: 1) An embedding layer that offers initialized circRNA embeddings and miRNA embeddings from the input data; 2) multiple embedding propagation layers that refine the embeddings by aggregating higher-order interaction information; 3) the prediction layer that concatenates the embeddings from different propagation layers and outputs the prediction score of a circRNA-miRNA pair.

2.4.1 Embedding Layer

We use the embedding vector $\mathbf{e}_{u_i}^k \in \mathbb{R}^s$ ($\mathbf{e}_{v_i}^k \in \mathbb{R}^s$) to describe the circRNA u (miRNA v) in k-th layer, where s is the embedding size. The initial state of circRNA embeddings and miRNA embeddings in embedding layer can be abstracted as:

$$\mathbf{E}_u^0 = \left[\mathbf{e}_{u_1}^0, \mathbf{e}_{u2}^0, \ldots, \mathbf{e}_{u_n}^0\right] \quad (1)$$

$$\mathbf{E}_v^0 = \left[\mathbf{e}_{v_1}^0, \mathbf{e}_{v2}^0, \ldots, \mathbf{e}_{v_m}^0\right] \quad (2)$$

Where $\mathbf{E}_u^0$ is the initial embedding of circRNAs, and $\mathbf{E}_v^0$ is the initial embedding of miRNAs. The initial embedding will be continuously optimized and improved end-to-end, which will be mentioned in the next section.

2.4.2 Embedding Propagation Layers

Next, we continuously aggregate the information of the node itself and its adjacent nodes to refine the embeddings of miRNAs and circRNAs. This is based on the GNN message-passing architecture (Hamilton et al., 2017; Xu et al., 2018). During an embedding update, the message aggregated by each node consists of two parts: the messages from the neighbor nodes of the previous layer and the messages inherited from the node itself.

As shown in **Figure 3**, in the k-th propagation layer, the embedding of circRNA u can be recursively formulated as:

$$e_u^k = \sigma\left(m(u,k) + \sum_{v \in N_u} m(u,v,k)\right) \quad (3)$$

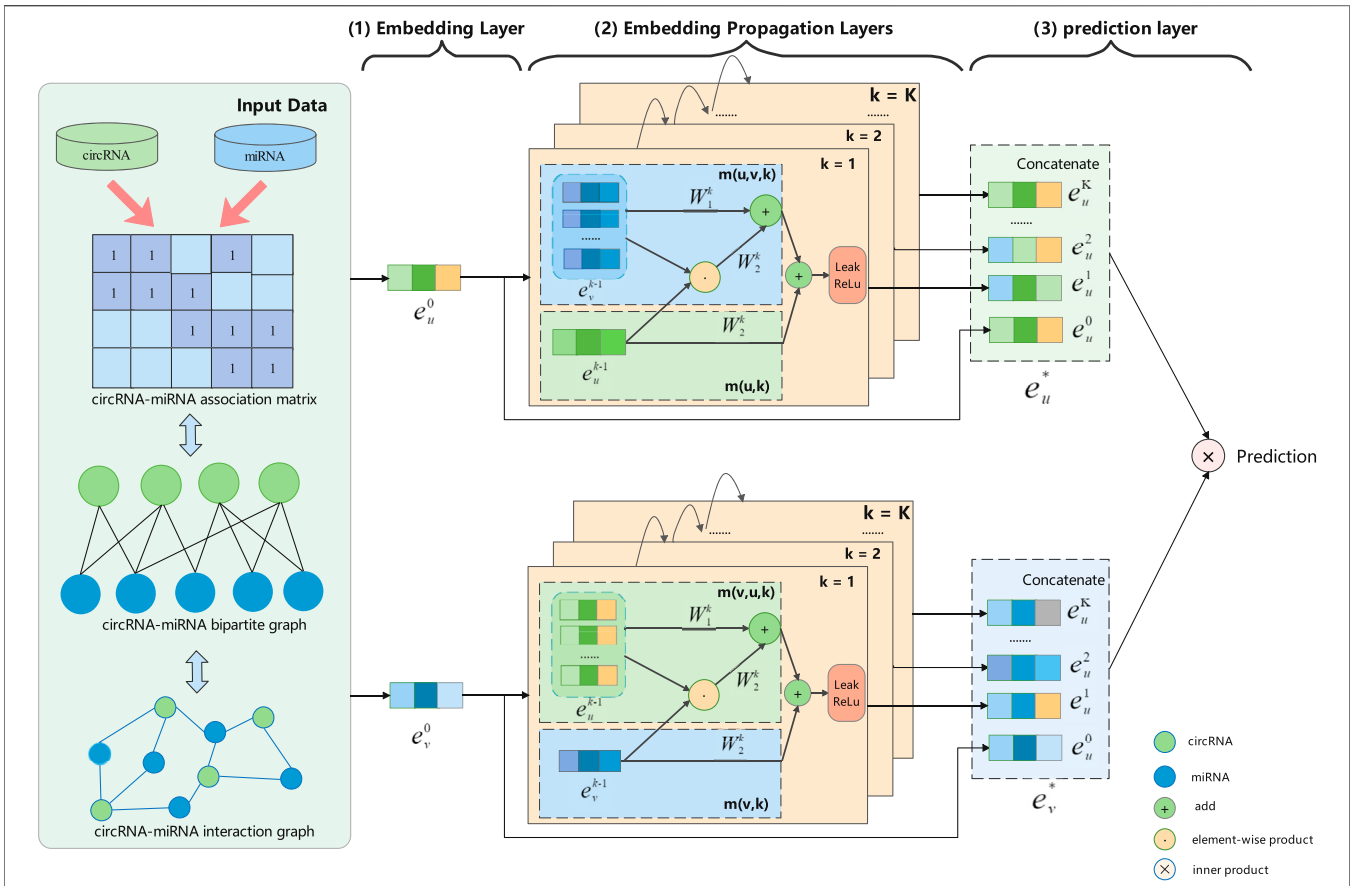


FIGURE 2 | An illustration of GCNCMI model architecture (the arrowed lines present the flow of information). Using GCNCMI to predict the relationship between circRNA u (green) and miRNA v (blue) mainly includes three steps: (1) In the embedding layer, we use input data to initialize circRNA embedding (e_u^0) and miRNA embedding (e_v^0); (2) In embedding propagation layers, the embeddings are continuously refined by recursively aggregating higher-order interaction information; (3) In the prediction layer, we concatenate the embeddings from different propagation layers and make the final prediction.

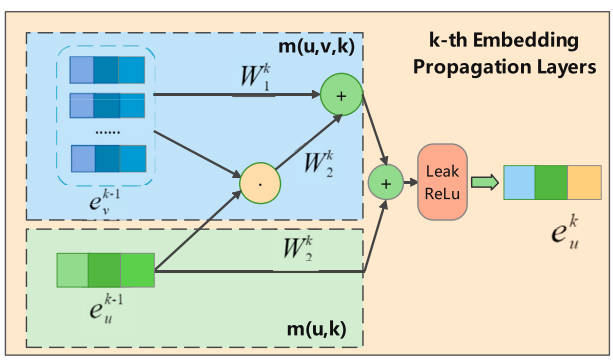


FIGURE 3 | Illustration of message aggregation for circRNA u in k-th embedding propagation layer, where the e_v^{k-1} represents the embedding of the neighbor node v of u in (k-1)-th layer.

Where e_u^k represents the embedding of circRNA obtained in the k-th embedding propagation layer, $\sigma(\cdot)$ is the activation function LeakyReLU (Nikolakopoulos and Karypis, 2019), v denotes the neighbor nodes of u, and $m(u, k)$ represent the messages delivered from the previous layer itself, while $m(u, v, k)$ representing the messages delivered by all neighbor nodes from the previous layer. The $m(u, k)$ and $m(u, v, k)$ can be formulated as follows:

$$m(u,k) = W_1^k e_u^{k-1} \tag{4}$$

$$m(u,v,k) = \frac{W_1^k e_v^{k-1} + W_2^k\left(e_v^{k-1} \odot e_u^{k-1}\right)}{\sqrt{N(u)\|N(v)\|}} \tag{5}$$

Where $\mathbf{W}_1^k, \mathbf{W}_2^k \in \mathbb{R}^{d_k \times d_{k-1}}$ are the trainable transformation matrices used to extract propagation information, and d_k is the transformation size; e_u^{k-1} is the circRNA embedding representation generated from the $(k-1)$-th propagation layer, which will further contribute its information to the circRNA embedding u at layer k. We use the graph Laplacian norm $1/\sqrt{|N_u||N_v|}$ to control how much the propagating message decays as the path length increases, where $N(u)$ represent the first-hop neighbors of circRNA u (miRNA v). In **Eq. 4**, we consider the self-connection of nodes, which can effectively retain the original feature information to avoid information variation when the number of layers increases. For the neighbor nodes of node u, we aggregate

not only the information of node v but also aggregate the interaction information between the u and v. It is encoded *via* $e_v^{k-1} \odot e_u^{k-1}$, where $\odot$ is element-wise product operation. In this way, more information from similar nodes can be passed, which enhances the representation ability of the model and helps to improve the accuracy of prediction results. **Eqs 3–5** represent the calculation process of the embedding circRNA u at the k-th layer. Analogously, the embedded representation of miRNA can be obtained.

2.4.3 Model Prediction

After multi-layer propagation, we can obtain multiple embedding representations of miRNAs and circRNAs. The embeddings obtained by different propagation layers contain different orders of interaction information, so they have different contributions to reflecting the relationship between circRNAs and miRNAs. Therefore, we concatenate all embeddings to express the final embedding. The following formula shows the final embedding representation of circRNA u and miRNA v through K embedding propagation layers:

$$\mathbf{e}_u^* = \mathbf{e}_u^0 \| e_u^1 \| \cdots \| \mathbf{e}_u^K, \quad \mathbf{e}_v^* = \mathbf{e}_v^0 \| e_v^1 \| \cdots \| \mathbf{e}_v^K \tag{6}$$

Where $\|$ denotes concatenation operation, this simple concatenation operation can makes our final embeddings contain richer semantic information without increasing the learning parameters. Finally, we perform an inner product operation on the final embedding to obtain the interaction prediction between circRNA u and miRNA v:

$$\hat{y}_{GCNCMI}(u, v) = \mathbf{e}_u^* \otimes \mathbf{e}_v^* \tag{7}$$

Algorithm 1 shows the pseudocode description for predicting the interaction between circRNA u and miRNA v using GCNCMI.

2.4.4 Model Optimization

Pointwise loss and pairwise loss are two common methods used to update model parameters (He et al., 2016). The pointwise learning emphasizes the loss between the predicted value $\hat{y}_{uv}$ and target value y_{uv}. Still, we prefer to address predicting the interactions between circRNA and miRNA from the perspective of ranking. Therefore, we choose pairwise loss optimization to update model parameters. Bayesian Personalized Ranking (BPR) is a matrix factorization-based pairwise loss function that is often used to optimize recommendation tasks similar to our prediction task (Rendle, 2010). Specifically, it can be formulated as follows:

$$\min_{\Theta} \mathrm{L} = \sum_{(u,i,j) \in D} -\ln s\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \lambda \|\Theta\|_2^2 \tag{8}$$

where $s(\cdots)$ is the sigmoid function; $D = \{(u, i, j) | (u, i) \in \mathrm{R}^+, (u, j) \in \mathrm{R}^-\}$ is the pairwise training sample containing positive samples R^+(i.e., circRNA u has interacted with miRNA v_i) and negative samples R^-(i.e., the interactions between circRNA u and miRNA v_j is unknown). $\hat{y}_{ui}$ denotes the prediction score of u and v_i. $\hat{y}_{uj}$ denotes the prediction score of u and v_j. $\Theta = \left\{\mathbf{E}, \left\{\mathbf{W}_1^k, \mathbf{W}_2^k\right\}_{k=1}^K\right\}$ represents all model parameters that will be trained. λ is a parameter used to control the strength of L_2 regularization. We use Adam as the optimizer to update the model parameters. Additionally, we use message dropout and node dropout to avoid model overfitting during training. Message dropout means that we will drop the message in **Eq. 3** with a certain probability during the propagation, while node dropout randomly drops a specific node and discards all its outgoing messages. Dropout operations can reduce the influence of specific RNAs, making the model more robust.

Algorithm 1. GCNCMI algorithm to predict the interaction between circRNA u and miRNA v

Input: interaction matrix $R \in \mathbb{R}^{n \times m}$; bipartite graph $\mathrm{G}(\mathrm{U} \cup \mathrm{V}, \mathrm{E})$; depth K; dimension D; weight matrices $\mathbf{W}_1^k, \mathbf{W}_2^k, \forall k \in \{1, \ldots, K\}$; neighborhood function N; activation function LeakyReLU σ

Output: interaction prediction between circRNA u and miRNA v

Initialize Embeddings:

$$\mathrm{E}^0 = [\ \underbrace{\mathbf{e}_{u_1}^0, \cdots, \mathbf{e}_{u_n}^0}_{\text{circRNA embeddings}}\ ,\ \underbrace{\mathbf{e}_{v_1}^0, \cdots, \mathbf{e}_{v_m}^0}_{\text{miRNA embeddings}}\ ];$$

for *layer k = 1 to K* **do**
  for *all node embeddings* $e_i^k \in \mathbb{E}^k$ **do**
    $m(i,k) = W_1^k e_i^{k-1}$;
    $m(i,j,k) = \frac{W_1^k e_j^{k-1} + W_2^k \left(e_j^{k-1} \odot e_i^{k-1}\right)}{\sqrt{N(i)\|N(j)|}}$;
    $e_i^k \leftarrow \sigma\left(m(i,k) + \sum_{j \in N_i} m(i,j,k)\right)$;
  end
end
$\mathbf{e}_u^* \leftarrow \mathbf{e}_u^0 \| e_u^1 \| \cdots \| \mathbf{e}_u^K$;
$\mathbf{e}_v^* \leftarrow \mathbf{e}_v^0 \| e_v^1 \| \cdots \| \mathbf{e}_v^K$;
$\hat{y}_{GCNCMI}(u,v) \leftarrow \mathbf{e}_u^* \otimes \mathbf{e}_v^*$;
return $\hat{y}_{GCNCMI}(u,v)$;

3 EXPERIMENT

3.1 Experimental Settings

To evaluate the performance of our model in predicting circRNA-miRNA interactions, we combined the known 9,589 interactions used as positive samples, and 9,589 unlabeled interactions were randomly selected from the benchmark dataset as negative samples. We performed five-fold cross-validation on the constructed dataset. The

TABLE 2 | The five-fold cross-validation results of GCNCMI.

No.	AUPR	AUC	ACC	Pre	Recall	F1
1	0.9293	0.9288	0.8508	0.9390	0.8289	0.8805
2	0.9428	0.9352	0.8531	0.9424	0.8440	0.8905
3	0.9453	0.9372	0.8578	0.9450	0.8357	0.8870
4	0.9396	0.9282	0.8532	0.9392	0.8341	0.8835
5	0.9412	0.9312	0.8503	0.9408	0.8298	0.8818
Average	0.9396	0.9320	0.8530	0.9413	0.8345	0.8847

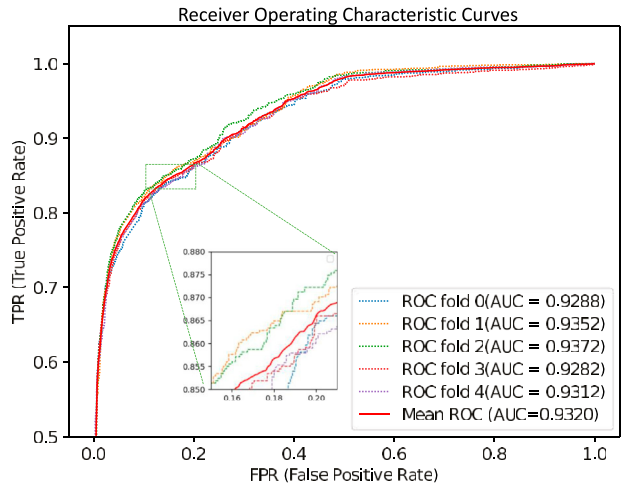


FIGURE 4 | GCNCMI performed the ROC curves of five-fold cross-validation.

TABLE 3 | The performance of GCNCMI on different layers.

K	AUPR	AUC	Acc	Pre	Recall	F1
1	0.9283	0.9198	0.8368	0.9280	0.8319	0.8773
2	0.9412	0.9312	0.8503	0.9408	0.8298	0.8818
3	0.9393	0.9301	0.8480	0.9390	0.8340	0.8834
4	0.9374	0.9272	0.8446	0.9371	0.8444	0.8883
5	0.9361	0.9244	0.8295	0.9358	0.8371	0.8837

validated circRNA-miRNA interactions were randomly divided into five parts. Take each part as a positive sample and an equal number of unlabeled samples from the benchmark data as negative samples to form a test set. At the same time, perform the same operation on the remaining four parts to obtain a training set. This operation is performed until the loop is completed five times.

To measure the performance of GCNCMI more comprehensively, we used AUC, AUPR, Recall, Accuracy (Acc), precision (Pre), and F1 Scores. The definitions of each indicator are as follows:

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

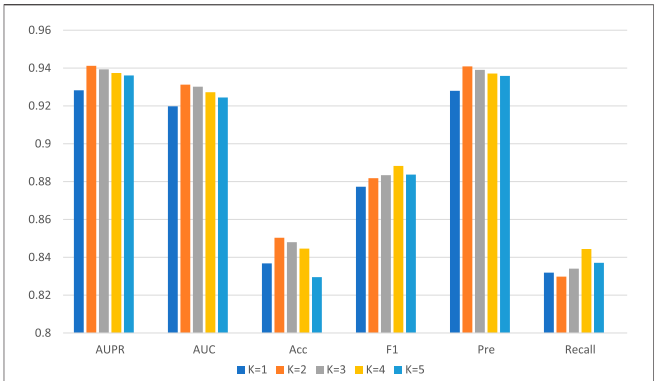


FIGURE 5 | The performance of GCNCMI model on different layers.

TABLE 4 | The performance of GCNCMI model on different embedding sizes.

D	AUPR	AUC	Acc	Pre	Recall	F1
16	0.9260	0.9170	0.8360	0.9257	0.8136	0.8660
32	0.9190	0.9102	0.8316	0.9187	0.8032	0.8571
64	0.9215	0.9110	0.8287	0.9212	0.8105	0.8623
128	0.9361	0.9265	0.8485	0.9357	0.8230	0.8757
256	0.9412	0.9312	0.8503	0.9409	0.8298	0.8819
512	0.9376	0.9268	0.8475	0.9373	0.8303	0.8806

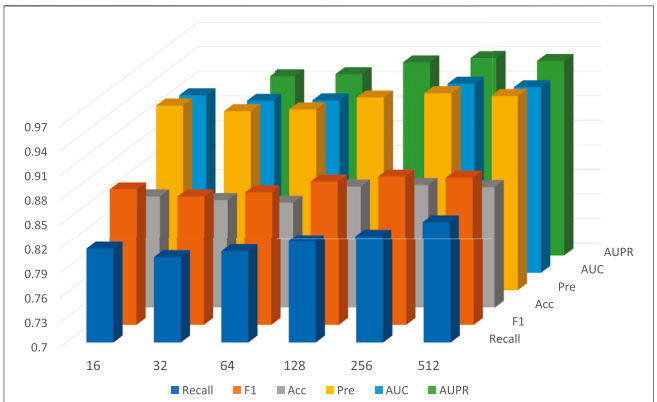


FIGURE 6 | The performance of GCNCMI model on different embedding sizes.

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = \frac{2{*}Precision{*}Recall}{Precision + Recall} \tag{12}$$

Where TP and FP represent the number of correctly classified samples and the number of misclassified samples in known circRNA-miRNA interactions, respectively, TN represents the number of correctly predicted unrelated circRNA-miRNA interactions, and FN represents the

TABLE 5 | Performance comparison of different methods under five-fold cross validation.

Methods	AUC	AUPR	Acc	Pre	Recall	F1
AE-RF	0.7662 ± 0.0050	0.8239 ± 0.0042	0.8333 ± 0.0013	0.8923 ± 0.0019	**0.9279** ± 0.0019	**0.9097** ± 0.0010
DMFCDA	0.7321 ± 0.0240	0.7115 ± 0.0171	0.6975 ± 0.0112	0.8160 ± 0.0265	0.7729 ± 0.1112	0.7938 ± 0.0707
DMFMDA	0.7922 ± 0.0057	0.8230 ± 0.0089	0.7307 ± 0.0049	0.7030 ± 0.0080	0.7246 ± 0.0116	0.7136 ± 0.0065
KATZHMDA	0.8469 ± 0.0017	0.8647 ± 0.0019	0.8073 ± 0.0030	0.8511 ± 0.0055	0.7227 ± 0.0106	0.7816 ± 0.0071
NTSHMDA	0.8526 ± 0.0016	0.8772 ± 0.0018	0.6276 ± 0.0083	0.7556 ± 0.0518	0.4040 ± 0.0531	0.5264 ± 0.0486
SDLDA	0.7875 ± 0.0307	0.8286 ± 0.0189	0.6693 ± 0.0019	0.8287 ± 0.0108	0.7891 ± 0.0809	0.8084 ± 0.0706
GCNCMI	**0.9320** ± 0.0014	**0.9396** ± 0.0406	**0.8530** ± 0.0134	**0.9413** ± 0.0204	0.8345 ± 0.0301	0.8846 ± 0.0068

number of prediction errors in unrelated miRNA-circRNA interactions. *F*1 is a weighted average of model precision and Recall.

3.2 Cross-Validation Results

We performed five-fold cross-validations to evaluate the performance of the GCNCMI model in predicting circRNA-miRNA interactions. The experimental results of the five-fold cross-validation are shown in **Table 2**. As shown in the table, the AUC of the five-fold cross-validations are: 0.9288, 0.9352, 0.9372, 0.9282, 0.9312. On the AUPR, the AUPR of the five-fold cross-validations are 0.9293, 0.9428, 0.9453, 0.9396, 0.9412, respectively. In addition, we also plotted the ROC curve of GCNCMI, as shown in **Figure 4**. The above experimental results show that GCNCMI has good performance in predicting unknown circRNA-miRNA interactions.

3.3 Parameter Influence

For GCNCMI, two essential parameters affect its performance: *K* (the number of layers) and *D* (the dimension of the embedding vector). When *K* is 2, and *D* is 256, our model GCNCMI achieves the best performance under five-fold cross-validation.

The setting of the number of layers *K* indicates that our final embedding model incorporates the information of K-hop neighbor nodes in the bipartite graph, which can learn more hidden interaction information between nodes for the neural network. **Table 3** lists the detailed values, and **Figure 5** shows the trend chart for different layers. We tried from 1 to 5 layers for the number of layers of the model and found that the model's accuracy at the beginning will increase with the increase of the number of layers. The best performance of the model is when the layer is 2. As the number of network layers increases, the hidden feature pairs of nodes tend to converge to the same value, which leads to an over-smoothing problem in the network.

On the other hand, under the framework of five-fold cross-validation, we conducted experiments for *D* in 16, 32, 64, 128, 256, 512, and other 6 cases; the detailed data is shown in **Table 4**. In general, as the dimension of the embedding vector increases, the expressive power of the model increases. But as can be seen from the **Figure 6**, from 16, 32, 64, 128, 256, the model's performance has been increasing at first, but at 256, the commission has reached the maximum value. As *D* continues to grow, it will adversely affect the model's performance.

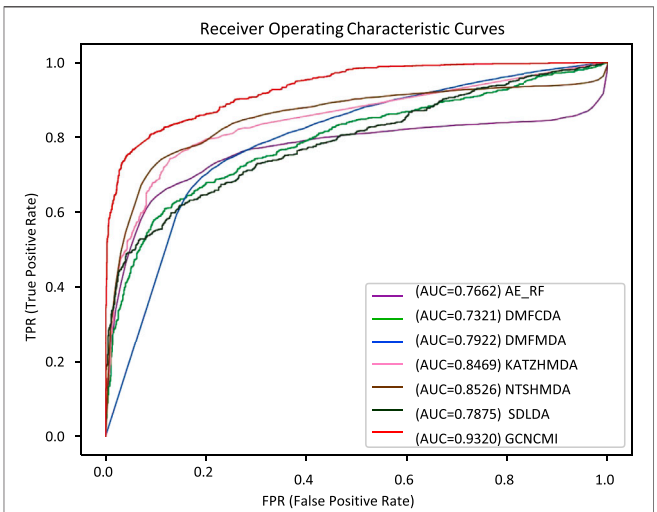


FIGURE 7 | AUC values of different methods under five-fold cross-validation.

TABLE 6 | The number of circRNAs, miRNAs, and circRNA-miRNA interactions included in the independent test dataset.

circRNA	miRNA	interactions	unlabled interactions
1,502	494	9,386	9,386

3.4 Compared With State-Of-The-Art Methods

Since circRNA and miRNA interaction is a relatively new field, GCNCMI is the first method we know to predict the interaction between circRNA and miRNA, but other advanced methods in bioinformatics still provide us with reference. To better verify the performance of GCNCMI in inferring the interaction between circRNA and miRNA. We compare GCNCMI with six other state-of-the-art methods in bioinformatics.

Considering the scarcity of related biological resources, in calculating biological similarity, we only calculated Gaussian interaction profile biological similarity (GIP). In addition, since the adjacency matrix initialized each time is different, it requires us to re-mine the information in the bipartite graph. Strictly speaking, in similarity-based methods [AE-RF (Deepthi and Jereesh, 2021), KATZHMDA (Chen et al., 2017),

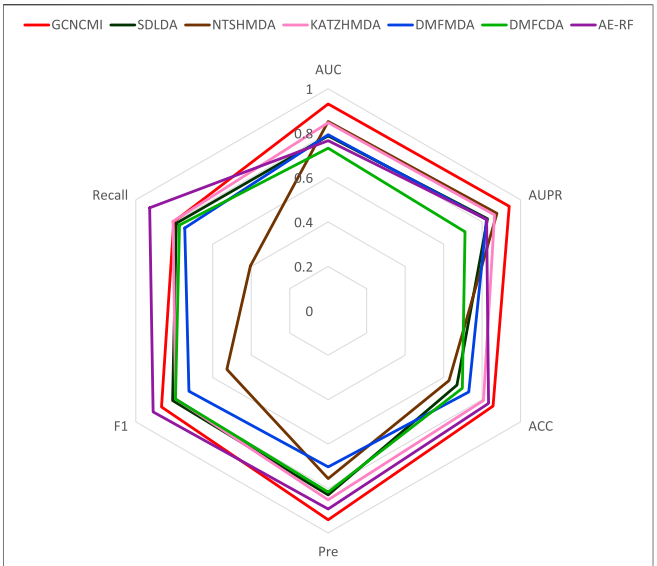


FIGURE 8 | Radar plots of different methods on various performances.

NTSHMDA (Luo and Long, 2018)], the similarity matrix is recalculated each time during the cross-validation process. In the SDLDA method, we used SVD singular value decomposition to obtain linear features of circRNAs and miRNAs. The DMFMDA method chooses a Bayesian loss function over the loss function instead of the mean squared error.

We performed a ten-times, five-fold cross-validation of GCNCMI with six advanced methods, changing the random number seed each time, and calculated the mean and standard deviation of 10 experiments. **Table 5** lists several methods such as AE-RF (Deepthi and Jereesh, 2021), DMFCDA (Liu et al., 2020), DMFMDA (Liu et al., 2020), KATZHMDA (Chen et al., 2017), NTSHMDA (Luo and Long, 2018), SDLDA (Zeng et al., 2020), and compared with the GCNCMI model. **Figure 7** plots the AUC curves to compare the seven methods. As can be seen from **Table 5** and **Figure 7**, GCNCMI mines the high-order interactions between circRNA and miRNA; GCNCMI is higher than other methods in most indicators, among which the AUC value of GCNMCI is 0.9320, and the highest among different methods is NTSHMDA, whose AUC value is 0.8526, which is 7.94% lower than GCNCMI. GCNCMI value of AUPR is 0.9396, which is 6.24% higher than the second-best method, NTSHMDA. The above experimental results show that our model performs well in predicting the relationship between circRNA and miRNA.

The radar **Figure 8** shows the performance of GCNCMI on AUC, AUPR, ACC, Recall, F1, Pre. The evaluation index is set from 0 to 1. As shown from **Figure 8**, the distance between the point and the center of the circle reflects the level of the value. It is evident that GCNCMI is better than other methods in predicting the circRNA-miRNA relationship.

To further verify the accuracy of the GCNCMI model in circRNA-miRNA association prediction, we retrieved the data from the PubMed database, removed the known relationships that overlapped with the training dataset, and established a 9,386 miRNA-circRNA association relationship, 494 miRNAs, an independent test set of 1,502 circRNAs, and 9,386 unlabeled interactions were randomly selected from the benchmark dataset as negative samples. The specific information of the independent test set can be found in **Table 6**. Although there may be a small part of the independent test set and the unknown overlapping relationship in the training set, it can be ignored because it

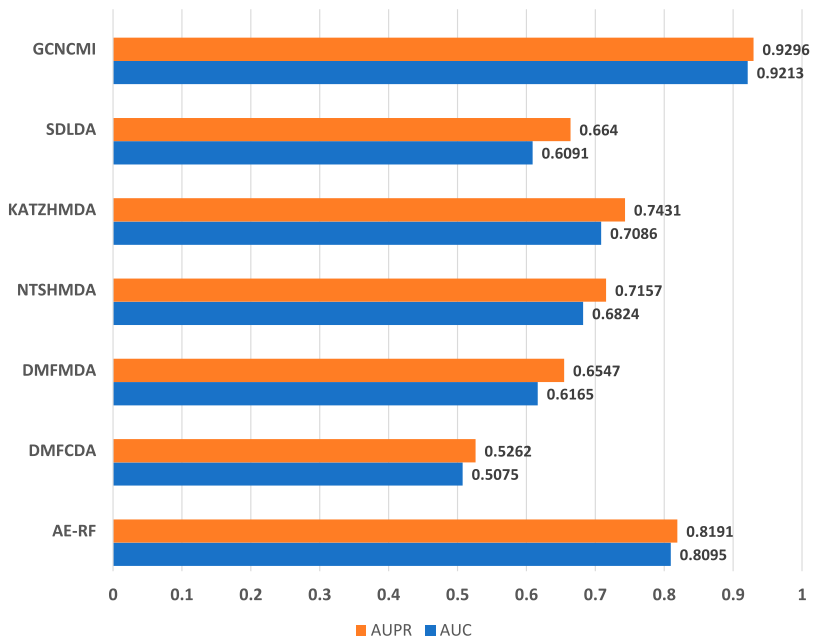


FIGURE 9 | Comparison of AUC and AUPR values of GCNCMI and several other methods on independent test sets.

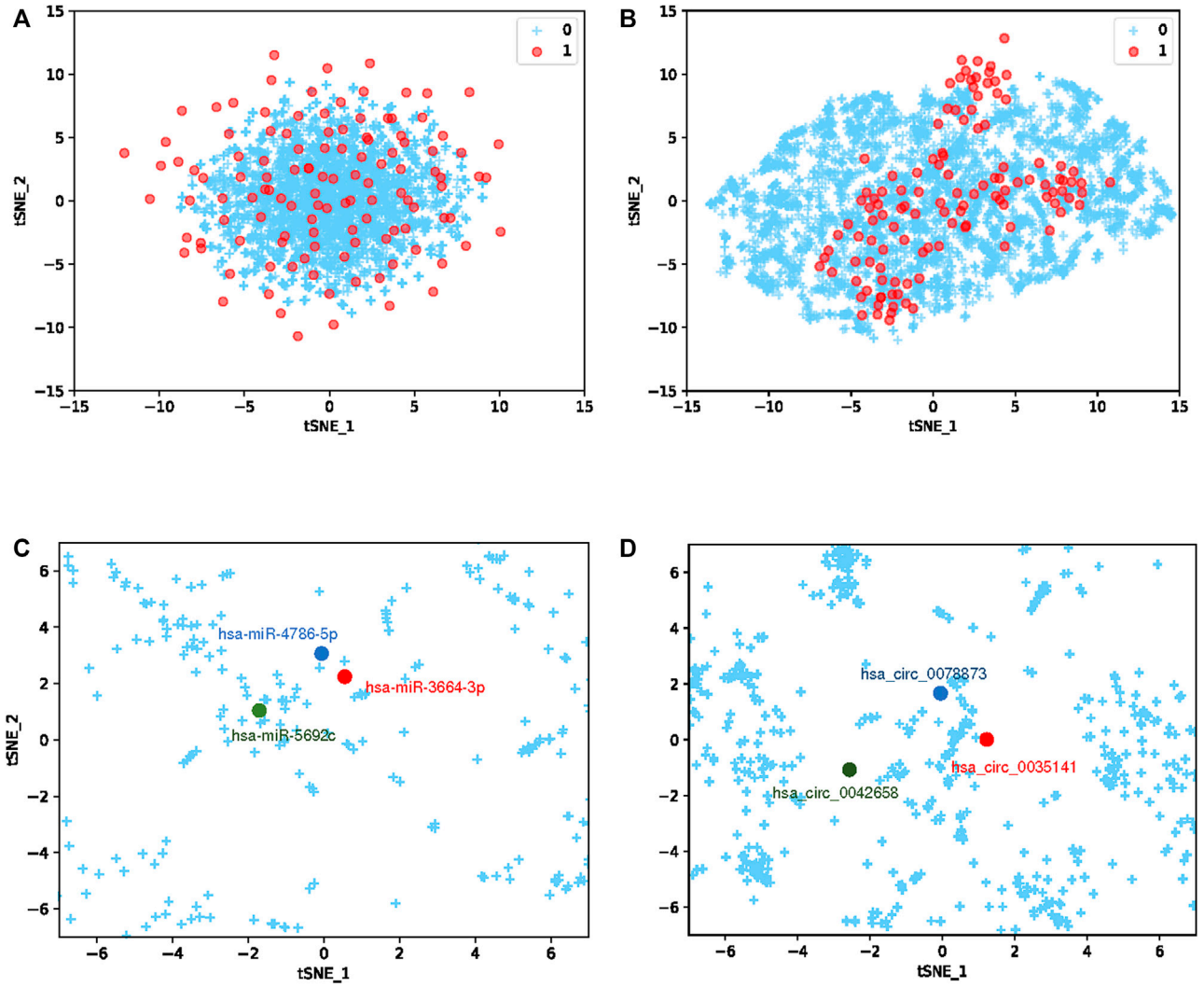


FIGURE 10 | Embedding visualization **(A)** represents the embedding of the initialized circRNA-miRNA interaction pairs, and **(B)** represents the embedding representation of the circRNA-miRNA interaction pairs learned by the GCNCMI model. **(C)** represents the embedding of miRNA after learning by the GCNCMI model, and **(D)** represents the embedding of circRNA after learning by the GCNCMI model.

occupies a small proportion of the entire unvalidated sample set. The basic model for predicting circRNA-miRNA associations was obtained by training on our data set and tested on the independent test set. The test results are as **Figure 9**. The AUC of the GCNCMI model reached 0.9213, and the AUPR value reached 0.9296, which is higher than several other methods of comparison. The independent test results further showed that GCNCMI is an effective tool for inferring miRNA-circRNA associations.

3.5 Embedding Visualization

To more clearly demonstrate the learning ability of the GCNCMI, We use T-SNE (Van der Maaten and Hinton, 2008) to visualize the embedding of circRNA-miRNA interaction pairs. Because the number of unknown relationships is much larger than the number of known associations, and to better visualize the overall mining of higher-order relationships by GCNCMI, we choose to visualize more unlabeled samples than labeled samples. The main goal of T-SNE is to convert multi-dimensional datasets into low-dimensional datasets. Compared with other dimensionality reduction algorithms, T-SNE is the most effective technique in data visualization. Since T-SNE is not a linear dimensionality reduction technique, it can capture the complex manifold structure of high-dimensional data. We initially a 32-dimensional vector to represent miRNA and circRNA. To explore the similarity between vector representations, we used the T-SNE algorithm to reduce the vector to 2-dimensional, as shown in **Figure 10A**. The blue + represents unknown miRNA-circRNA interaction pairs, and the red dots represent the known circRNA-miRNA interaction pairs. **Figure 10B** shows the embedding of the circRNA-miRNA interactions learned by the GCNCMI model. Comparing **Figures 10A,B**, it can be seen that GCNCMI has a good effect on mining

TABLE 7 | The top 10 circRNAs with the closest relationship to hsa-miR-622 predicted by GCNCMI model.

Rank	CircRNA	Evidence(PMID)	Score
1	hsa_circ_0000231	34183076	0.8822
2	hsa_circ_0101432	Unconfirmed	0.8820
3	hsa_circ_0119872	33579337	0.8815
4	hsa_circ_0008574	32616043	0.8798
5	hsa_circ_0000211	31668923	0.8796
6	hsa_circ_0001273	35567340	0.8712
7	hsa_circ_0086902	Unconfirmed	0.8592
8	hsa_circ_KCNQ5	35413218	0.8542
9	hsa_circ_0101432	35297300	0.8498
10	hsa_circ_0006000	Unconfirmed	0.8469

TABLE 8 | The top 10 circRNAs with the closest relationship to hsa-miR-149-5p predicted by GCNCMI model.

Rank	CircRNA	Evidence(PMID)	Score
1	hsa_circ_0061140	32224273	0.8737
2	hsa_circ_0075341	31706100	0.8722
3	hsa_circ_0008956	34153672	0.8702
4	hsa_circ_0000654	31778020	0.8693
5	hsa_circ_0051239	Unconfirmed	0.8689
6	hsa_circ_ROBO2	34649241	0.8673
7	hsa_circ_0011385	34720052	0.8672
8	hsa_circ_0087352	35286916	0.8671
9	hsa_circ_0123996	32707301	0.8661
10	hsa_circ_0031059	Unconfirmed	0.8648

high-order interactions between miRNAs and circRNAs, and the GCNCMI can better use the known interaction pairs to mine potential miRNA-circRNA interaction pairs. In addition, we also visualized the learned circRNA embeddings and miRNA embeddings. **Figure 10C** shows the learned miRNA embeddings. We used the GCNCMI model to predict the top 30 circRNAs most closely associated with each miRNA, and also predicted the top 30 miRNAs most closely associated with each circRNA. The hsa-miR-4786-5p and hsa-miR-3664-3p were associated with nine similar circRNAs, and hsa-miR-4786-5p and hsa-miR-5692c were associated with five similar circRNAs. Therefore, the hsa-miR-4786-5p is more similar to hsa-miR-3664-3p. It can also be seen from **Figure 10C** that the distance between hsa-miR-4786-5p and hsa-miR-3664-3p is closer. **Figure 10D** shows the visualization of the embedding of circRNAs after model learning. The hsa-circ-0078873 and hsa-circ-0042658 were associated with three similar miRNAs, and hsa-circ-0035141 and hsa-circ-0078873 were associated with seven similar miRNAs. Therefore, hsa-circ-0078873 is closer to hsa-circ-0035141, and it can be seen from **Figure 10D** that hsa-circ-0078873 is closer to hsa-circ-0035141. The experimental results show that GCNCMI can effectively learn the potential higher-order interactions between miRNAs and circRNAs.

3.6 Case Studies

It is of great significance to discover unknown associations between circRNAs and miRNAs. We selected two miRNAs, hsa-miR-622 and hsa-miR-149-5p, for case studies. Specifically, we first delete the circRNAs that have been experimentally validated for the selected miRNAs. Then, the remaining circRNAs were sorted in descending order according to the values predicted by the GCNCMI model. The following shows the results of the normalized prediction scores of the GCNCMI model. Finally, we screened the top 10 circRNAs and collected evidence in the published literature for testing.

miR-622 (Lu et al., 2022) is a miRNA of 13q31.3 in the eukaryotic genome, and its expression is mainly in the nucleus. In recent years, studies have found that miR-622 can functionally inhibit the malignant proliferation of cells, which is helpful for cancer treatment. In recent years, miR-149-3p (Yang et al., 2017) can effectively inhibit the proliferation and apoptosis of malignant tumors. Recent studies have found that miR-149-3p can increase the sensitivity of drugs. **Table 7** and **Table 8** list the top 10 candidate circNRAs of hsa-miR-622 and hsa-miR-149-5p. We selected the top 10 candidate circRNAs as our predicted circRNAs, respectively, and finally, we compared the predicted results with the experimentally validated interactions. It can be seen that 7 of hsa-miR-622 were confirmed by existing evidence, and 8 of hsa-miR-149-5p were confirmed by existing evidence. It should be noted that unproven associations may exist and require further experimental verification.

4 CONCLUSION

CircRNAs are circular non-coding RNAs with regulatory functions, most of which exist in eukaryotic excerpts, and most circRNAs are composed of exons. Because circRNAs are less affected by nucleases, circRANs are more stable than linear RNAs. Current studies have shown that circRNAs can competitively adsorb miRNAs, and circRNAs can bind to proteins to inhibit the activity. Therefore, there is an urgent need to explore the relationship between circRNA and miRNA. However, because traditional biological experiments are time-consuming and labor-intensive, a more efficient method is needed to explore the potential relationship between circRNA and miRNA.

In this paper, we proposed a graph convolutional neural network prediction model for circRNA and miRNA interactions. To fully exploit the potential high-order interactions between circRNAs and miRNAs, we designed a graph convolutional neural network method to propagate the interaction's relation recursively without computing the similarity of circRNAs and miRNAs. The experimental results demonstrated the excellent performance of GCNCMI in predicting the interactions between circRNAs and miRNAs. The results of independent tests indicate that the GCNCMI model has good generalization performance in predicting unknown circRNA and miRNA relationships. Finally, a case study compared our predictions with those validated by biological experiments, further demonstrating the model's excellent predictive performance. The above results indicate that GCNCMI is an excellent method for predicting the potential interactions between circRNAs and miRNAs.

While GCNCMI has excellent performance, it also has some limitations. First, due to the scarcity of biological resources,

GCNCMI only uses the association data of circRNAs and miRNAs, and the quality of the data will affect the performance of GCNCMI model training. In the future, using heterogeneous data from multiple perspectives will be considered to improve the model's performance further.

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

JH and LD designed and implemented the prediction method. JH, PX, CC, and JZ analyzed the data and wrote the manuscript. LD reviewed and revised the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant No. 61972422.

ACKNOWLEDGMENTS

We are grateful for resources from the High Performance Computing Center of Central South University. The work was carried out at National Supercomputer Center in Tianjin, and the calculations were performed on TianHe.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.959701/full#supplementary-material

REFERENCES

Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., et al. (2014). Circrna Biogenesis Competes with Pre-mrna Splicing. *Mol. Cell* 56, 55–66. doi:10.1016/j.molcel.2014.08.019

Bartel, D. P. (2004). MicroRNAs. *Cell* 116, 281–297. doi:10.1016/s0092-8674(04)00045-5

Butcher, S. E., and Brow, D. A. (2005). Towards Understanding the Catalytic Core Structure of the Spliceosome. *Biochem. Soc. Trans.* 33, 447–449. doi:10.1042/bst0330447

Calin, G. A., Sevignani, C., Dumitru, C. D., Hyslop, T., Noch, E., Yendamuri, S., et al. (2004). Human Microrna Genes Are Frequently Located at Fragile Sites and Genomic Regions Involved in Cancers. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2999–3004. doi:10.1073/pnas.0307323101

Capel, B., Swain, A., Nicolis, S., Hacker, A., Walter, M., Koopman, P., et al. (1993). Circular Transcripts of the Testis-Determining Gene Sry in Adult Mouse Testis. *Cell* 73, 1019–1030. doi:10.1016/0092-8674(93)90279-y

Chen, B., and Huang, S. (2018). Circular Rna: An Emerging Non-Coding Rna as a Regulator and Biomarker in Cancer. *Cancer Lett.* 418, 41–50. doi:10.1016/j.canlet.2018.01.011

Chen, X., Huang, Y. A., You, Z. H., Yan, G. Y., and Wang, X. S. (2017). A Novel Approach Based on Katz Measure to Predict Associations of Human Microbiota with Non-Infectious Diseases. *Bioinformatics* 33, 733–739. doi:10.1093/bioinformatics/btw715

Dayun, L., Junyi, L., Yi, L., Qihua, H., and Deng, L. (2021). Mgatmda: Predicting Microbe-Disease Associations via Multi-Component Graph Attention Network. *IEEE/ACM Trans. Comput. Biol. Bioinforma*. doi:10.1109/tcbb.2021.3116318

Deepthi, K., and Jereesh, A. S. (2021). Inferring Potential CircRNA-Disease Associations via Deep Autoencoder-Based Classification. *Mol. Diagn. Ther.* 25, 87–97. doi:10.1007/s40291-020-00499-y

Deng, L., Huang, Y., Liu, X., and Liu, H. (2022). Graph2MDA: A Multi-Modal Variational Graph Embedding Model for Predicting Microbe-Drug Associations. *Bioinformatics* 38, 1118–1125. doi:10.1093/bioinformatics/btab792

Deng, L., Yang, J., and Liu, H. (2020). Predicting Circrna-Disease Associations Using Meta Path-Based Representation Learning on Heterogenous Network. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (IEEE), 5–10. doi:10.1109/bibm49941.2020.9313215

Dong, R., Ma, X.-K., Chen, L.-L., and Yang, L. (2017). Increased Complexity of Circrna Expression During Species Evolution. *RNA Biol.* 14, 1064–1074. doi:10.1080/15476286.2016.1269999

Ford, E., and Ares, M. (1994). Synthesis of Circular Rna in Bacteria and Yeast Using Rna Cyclase Ribozymes Derived from a Group I Intron of Phage T4. *Proc. Natl. Acad. Sci. U.S.A.* 91, 3117–3121. doi:10.1073/pnas.91.8.3117

Gesteland, R., Cech, T., and Atkins, J. (2006). *The Rna World*. 3rd edn. Cold Spring Harbor. NY: Cold Spring Harbor Laboratory Press. [Google Scholar].

Grabowski, P. J., Zaug, A. J., and Cech, T. R. (1981). The Intervening Sequence of the Ribosomal Rna Precursor Is Converted to a Circular Rna in Isolated Nuclei of Tetrahymena. *Cell* 23, 467–476. doi:10.1016/0092-8674(81)90142-2

Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive Representation Learning on Large Graphs. *Adv. Neural Inf. Process. Syst.* 30.

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural Rna Circles Function as Efficient Microrna Sponges. *Nature* 495, 384–388. doi:10.1038/nature11993

He, H., Zhang, J., Gong, W., Liu, M., Liu, H., Li, X., et al. (2022). Involvement of Circrna Expression Profile in Diabetic Retinopathy and its Potential Diagnostic Value. *Front. Genet.* 13, 833573. doi:10.3389/fgene.2022.833573

He, X., Zhang, H., Kan, M. Y., and Chua, T. S. (2016). Fast Matrix Factorization for Online Recommendation with Implicit Feedback. *ACM*. doi:10.1145/2911451.2911489

Kos, A., Dijkema, R., Arnberg, A. C., Van der Meide, P. H., and Schellekens, H. (1986). The Hepatitis Delta (δ) Virus Possesses a Circular RNA. *Nature* 323, 558–560. doi:10.1038/323558a0

Lan, W., Zhu, M., Chen, Q., Chen, B., Liu, J., Li, M., et al. (2020). Circr2cancer: A Manually Curated Database of Associations Between Circrnas and Cancers. *Database (Oxford)* 2020. doi:10.1093/database/baaa085

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The c. elegans Heterochronic Gene Lin-4 Encodes Small Rnas with Antisense Complementarity to Lin-14. *Cell* 75, 843–854. doi:10.1016/0092-8674(93)90529-y

Li, Z., Zhou, Y., Yang, G., He, S., Qiu, X., Zhang, L., et al. (2019). Using Circular Rna Smarca5 as a Potential Novel Biomarker for Hepatocellular Carcinoma. *Clin. Chim. Acta* 492, 37–44. doi:10.1016/j.cca.2019.02.001

Liu, D., Huang, Y., Nie, W., Zhang, J., and Deng, L. (2021). Smalf: Mirna-Disease Associations Prediction Based on Stacked Autoencoder and Xgboost. *BMC Bioinforma.* 22, 1–18. doi:10.1186/s12859-021-04135-2

Liu, M., Wang, Q., Shen, J., Yang, B. B., and Ding, X. (2019). Circbank: A Comprehensive Database for Circrna with Standard Nomenclature. *RNA Biol.* 16, 899–905. doi:10.1080/15476286.2019.1600395

Liu, Y., Wang, S., Zhang, J., Zhang, W., Zhou, S., and Li, W. (2020). Dmfmda: Prediction of Microbe-Disease Associations Based on Deep Matrix Factorization Using Bayesian Personalized Ranking. *IEEE/ACM Trans. Comput. Biol. Bioinform* PP, 1763–1772. doi:10.1109/TCBB.2020.3018138

Lu, J., Xie, Z., Xiao, Z., and Zhu, D. (2022). The Expression and Function of Mir-622 in a Variety of Tumors. *Biomed. Pharmacother.* 146, 112544. doi:10.1016/j.biopha.2021.112544

Luo, J., and Long, Y. (2020). Ntshmda: Prediction of Human Microbe-Disease Association Based on Random Walk by Integrating Network Topological Similarity. *IEEE/ACM Trans. Comput. Biol. Bioinform* 17, 1341–1351. doi:10.1109/TCBB.2018.2883041

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). Circular Rnas Are a Large Class of Animal Rnas with Regulatory Potency. *Nature* 495, 333–338. doi:10.1038/nature11928

Nikolakopoulos, A. N., and Karypis, G. (2019). Recwalk: Nearly Uncoupled Random Walks for Top-N Recommendation. *Proc. Twelfth ACM Int. Conf. Web Search Data Min.*, 150–158.

O'Brien, J., Hayder, H., Zayed, Y., and Peng, C. (2018). Overview of Microrna Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol. (Lausanne)* 9, 402. doi:10.3389/fendo.2018.00402

Palazzo, A. F., and Lee, E. S. (2015). Non-Coding Rna: What Is Functional and What Is Junk? *Front. Genet.* 6, 2. doi:10.3389/fgene.2015.00002

Rendle, S. (2010). Factorization Machines. In ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14–17. December 2010. doi:10.1109/icdm.2010.127

Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids Are Single-Stranded Covalently Closed Circular Rna Molecules Existing as Highly Base-Paired Rod-Like Structures. *Proc. Natl. Acad. Sci. U.S.A.* 73, 3852–3856. doi:10.1073/pnas.73.11.3852

Soslau, G. (2018). Circular Rna (Circrna) Was an Important Bridge in the Switch from the Rna World to the Dna World. *J. Theor. Biol.* 447, 32–40. doi:10.1016/j.jtbi.2018.03.021

Steitz, T. A., and Moore, P. B. (2003). Rna, the First Macromolecular Catalyst: The Ribosome Is a Ribozyme. *Trends Biochem. Sci.* 28, 411–418. doi:10.1016/s0968-0004(03)00169-5

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* 9.

Wang, X., He, X., Cao, Y., Liu, M., and Chua, T.-S. (2019). Kgat: Knowledge Graph Attention Network for Recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 950–958.

Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.-i., and Jegelka, S. (2018). Representation Learning on Graphs with Jumping Knowledge Networks. International Conference on Machine Learning. PMLR, 5453–5462.

Yang, D., Du, G., Xu, A., Xi, X., and Li, D. (2017). Expression of Mir-149-3p Inhibits Proliferation, Migration, and Invasion of Bladder Cancer by Targeting S100a4. *Am. J. Cancer Res.* 7, 2209–2219.

Zeng, M., Lu, C., Zhang, F., Li, Y., Wu, F.-X., Li, Y., et al. (2020). Sdlda: Lncrna-Disease Association Prediction Based on Singular Value Decomposition and Deep Learning. *Methods* 179, 73–80. doi:10.1016/j.ymeth.2020.05.002

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.

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

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 24 August 2022
DOI 10.3389/fgene.2022.995532

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Jin-Xing Liu,
Qufu Normal University, China
Cheng Liang,
Shandong Normal University, China
Guoxian Yu,
Shandong University, China
Cunmei Ji,
Qufu Normal University, China

*CORRESPONDENCE
Dengju Yao,
ydkvictory@hrbust.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 16 July 2022
ACCEPTED 01 August 2022
PUBLISHED 24 August 2022

CITATION
Yao D, Zhang T, Zhan X, Zhang S, Zhan X and Zhang C (2022), Geometric complement heterogeneous information and random forest for predicting lncRNA-disease associations. *Front. Genet.* 13:995532. doi: 10.3389/fgene.2022.995532

Geometric complement heterogeneous information and random forest for predicting lncRNA-disease associations

Dengju Yao[1]*, Tao Zhang[1], Xiaojuan Zhan[1,2], Shuli Zhang[1], Xiaorong Zhan[3] and Chao Zhang[4]

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China, [2]College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China, [3]Department of Endocrinology and Metabolism, Hospital of South University of Science and Technology, Shenzhen, China, [4]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China

More and more evidences have showed that the unnatural expression of long non-coding RNA (lncRNA) is relevant to varieties of human diseases. Therefore, accurate identification of disease-related lncRNAs can help to understand lncRNA expression at the molecular level and to explore more effective treatments for diseases. Plenty of lncRNA-disease association prediction models have been raised but it is still a challenge to recognize unknown lncRNA-disease associations. In this work, we have proposed a computational model for predicting lncRNA-disease associations based on geometric complement heterogeneous information and random forest. Firstly, geometric complement heterogeneous information was used to integrate lncRNA-miRNA interactions and miRNA-disease associations verified by experiments. Secondly, lncRNA and disease features consisted of their respective similarity coefficients were fused into input feature space. Thirdly, an autoencoder was adopted to project raw high-dimensional features into low-dimension space to learn representation for lncRNAs and diseases. Finally, the low-dimensional lncRNA and disease features were fused into input feature space to train a random forest classifier for lncRNA-disease association prediction. Under five-fold cross-validation, the AUC (area under the receiver operating characteristic curve) is 0.9897 and the AUPR (area under the precision-recall curve) is 0.7040, indicating that the performance of our model is better than several state-of-the-art lncRNA-disease association prediction models. In addition, case studies on colon and stomach cancer indicate that our model has a good ability to predict disease-related lncRNAs.

KEYWORDS

lncRNA-disease association prediction, geometric complement heterogeneous information, random forest, autoencoder, machine learning

1 Introduction

Long non-coding RNA (lncRNA) is a kind of non-coding RNA with a length of more than 200 nucleotides, which have received increasing attention from researchers. LncRNAs have now been proved to play a key role in transcriptional and posttranslational regulation (Taft et al., 2010; Mathieu et al., 2014; Sun et al., 2018; Xie et al., 2018). The pathogenesis of a series of diseases is significantly associated with mutations and dysregulation of lncRNAs (Washietl et al., 2014; Chen et al., 2017). For example, MALAT1 was discovered to be overexpressed in many entity tumors such as lung cancer (Cheetham et al., 2013). It was shown that clonogenic and anchorage-dependent growth of lung cancer cells would be significantly decreased when H19 was down-regulated (Barsyte-Lovejoy et al., 2006). Confirming the associations between lncRNAs and diseases by biological experiments is time-consuming, labor-intensive and challenging, so using computational method to predict the associations not only provides a more efficient way for biological experiments but also reduces a lot of unnecessary human and material resources. Currently, dozens of computational models have been proposed to identify disease-associated lncRNAs based on various biological data. We can broadly classify the current computational models for lncRNA-disease association (LDA) prediction into three categories.

The first class of LDA prediction models is based on biological networks. Sun et al. implemented random walk and restart on lncRNA functional similarity network (Sun et al., 2014). Zhou et al. integrated the LDA network, disease similarity network and lncRNA-miRNA interaction network into a heterogeneous network and applied random walk on the network (Zhou et al., 2015). Chen et al. integrated the known LDAs, lncRNA expression profiles, lncRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity to predict potential LDAs (Chen, 2015a). Ping et al. (2019) constructed a model based on the known LDA network. However, these models need the known LDA network. Thus, Liu et al. (2014) conceived a model by integrating the known human expression profiles of lncRNA and disease genes, which is the first computational model without relying on the known LDAs. Chen et al. combined miRNA-disease association and lncRNA-miRNA interactions to form a model called HGLDA (Chen, 2015c). Zhou et al. developed a computational method by integrating association among lncRNA, protein, disease, miRNA, drug and high-order proximity preserved embedding for predicting LDAs (Zhou et al., 2021). Sumathipala et al. used the topology of a multi-level network consisting of lncRNA-protein, protein-protein interactions and protein-disease associations to identify LDAs (Sumathipala et al., 2019). Yu et al. used Bi-Random Walks on the lncRNA functional similarity network and disease network to predict LDAs (Yu et al., 2017). Yu et al. (2020) constructed a data fusion model called Attributed Heterogeneous Network Fusion for LDA prediction (AHNF).

The second class of LDA prediction model is based on matrix factorization. Fu et al. proposed a LDA prediction model called MFLDA. MFLDA factored data from heterogeneous data sources into low-rank matrices based on matrix trivialization to discover and explore its intrinsic and shared structure (Fu et al., 2018). Wu et al. constructed a GAMCLDA model by encoding local graph structures and features. The graph convolution network was used to encode the features of this map structure and nodes to learn the potential factorial vectors of lncRNAs and diseases. In addition, the inner product of lncRNA factor vectors and disease factor vectors was used as a decoder to reconstruct the LDA matrix (Wu et al., 2020). Gao et al. (2021) constructed a multi-label fusion collaborative matrix decomposition approach to predict LDAs. Wang et al. (2020) developed a weighted matrix factorization model on multi-relational data to predict LDAs. Liu et al. (2021) introduced a weighted graph regularized collaborative matrix factorization (WGRCMF) method to predict LDAs.

The third class of LDA prediction model is based on machine algorithms. Machine learning methods focus on gaining insights into features and imbalanced labels. Chen et al. formulated Laplace regularized least squares method to predict LDAs (called LRLSLDA) in a semi-supervised learning framework, which is the first machine learning-based methods to predict LDAs (Chen et al., 2015). However, for LRLSLDA, parameter optimization is a challenge. Later, Chen et al. combined lncRNA functional similarity with the LRLSLDA-LNCSIM prediction model and enhanced its performance by introducing similarity scores for predicting gene-disease associations (Huang et al., 2016). In addition, Lan et al. implemented a LDAP model based on SVM bagging by combining disease similarity and lncRNA similarity (Lan et al., 2017). Yao et al. constructed a computational model called RFLDA to identify associations based on feature selection by integrating the experiment-supported associations among lncRNA, miRNA, disease, disease semantic similarity and lncRNA functional similarity (Yao et al., 2020). Xuan et al. have developed a collection of convolutional neural networks-based lncRNA-disease prediction models, including CNNLDA (Xuan et al., 2019a), LDAPred (Xuan et al., 2019b), GCNLDA (Xuan et al., 2019c) and CNNDLP (Xuan et al., 2019d). The CNNLDA developed an analysis of the associations between lncRNA and disease using convolutional neural networks that combined semantic and functional similarity as well as lncRNA-disease associations, miRNA-disease associations and lncRNA-miRNA interactions (Xuan et al., 2019a). The LDAPred integrated a convolutional neural network and information flow propagation, combining associations, interactions, similarity structures and topological structures between lncRNAs, miRNAs and diseases (Xuan et al., 2019b). The GCNLDA is based on the graph convolutional network and convolutional neural network to obtain locally

integrated topological information within the lncRNA-disease-microRNA networks (Xuan et al., 2019c). By combining disease similarity, lncRNA similarity, miRNA-disease association and lncRNA-miRNA interactions, CNNDLP learned the attention and the low-dimensional network representation of the lncRNA-disease pairs (Xuan et al., 2019d). Wei et al. developed a method (LDICDL) that denoised lncRNA and disease features with an autoencoder, and used the matrix decomposition algorithm to test for potential disease-lncRNA association (Lan et al., 2022). Fan et al. proposed an lncRNA-disease prediction method that implemented convolutional matrices with conditional random fields and attention mechanisms for learning the embeddings of nodes for scoring latent associations between lncRNAs and diseases (Fan et al., 2022). Wu et al. proposed a method that combined extra trees with multi-layer graph embedding aggregation to predict LDAs (Wu Q. W. et al., 2021). Cui et al. proposed a novel model based on bipartite local model with nearest profile-based association inferring to predict LDAs (Cui et al., 2020).

These methods described above have achieved good prediction performance, but they also have some limitations. The biological network-based approach was affected by the scarcity of known LDA data; For the matrix factorization-based approach, the combination of model parameters is a very complex and necessary procedure; For the machine learning-based approach, feature processing and the impact of imbalanced data is a challenge. In this paper, we proposed a novel LDA prediction model based on geometric complement heterogeneous information and random forest (GCHIRFLDA in short). Firstly, the geometric complementation of LDA matrix was implemented by integrating the information of lncRNA-miRNA and miRNA-disease association information. Secondly, a low-dimensional feature space was extracted from the obtained LDA matrix by using an autoencoder, which combined Jaccard similarity coefficient and Gaussian interaction profile kernel similarity. Finally, a random forest classifier was trained on the constructed sample set to score potential lncRNA-disease associations. The AUC and AURP under five-fold cross-validation demonstrated that the GCHIRFLDA had a better performance than several state-of-the-art LDA prediction models, and the case studies on stomach cancer and colon cancer indicated that the GCHIRFLDA had excellent ability in identifying disease-associated lncRNAs.

2 Materials and methods

2.1 Representation of lncRNA-disease associations , miRNA-disease associations and lncRNA-miRNA interactions

LncRNA-disease associations (LDA), miRNA-disease associations (MDA) and lncRNA-miRNA interactions (LMI) were obtained from previous reports (Fu et al., 2018). The following l, d and m denote the number of lncRNA, disease and miRNA, respectively. The LDAs are represented by a 240× × 412 adjacency matrix $LD_{i\times j} \in LD^{l\times d}$, l is rows represent lncRNAs and d is columns represent diseases. For each element $LD_{i,j}$, its value is equal to one if lncRNA i is related to disease j; otherwise, its value is equal to 0. Similarly, the MDAs are represented by a 495× × 412 adjacency matrix $MD_{i\times j} \in MD^{m\times d}$, m is rows represent miRNAs and d is columns represent diseases. For each element $MD_{i,j}$, its value is equal to one if miRNA i is related to disease j; otherwise, its value is equal to 0. The LMIs are represented by a 240 × ×495 adjacency matrix $LM_{i\times j} \in LM^{l\times m}$, l is rows represent lncRNAs and m is columns represent diseases. For each element $LM_{i,j}$, its value is equal to one if lncRNA i is related to miRNA j; otherwise, its value is equal to 0.

2.2 Calculation of jaccard similarity of disease and lncRNA

Calculation of similarity of disease and lncRNA is an important step in LDAs predicting process. So far, there are many ways to calculate similarity, such as disease semantic similarity, disease cosine similarity, lncRNA functional similarity and lncRNA cosine similarity. In this work, we combine the Jaccard similarity coefficient which is complementary to the binary matrix and the Gaussian interaction profile kernel similarity which encodes the non-linear vectors in the LDA matrix. By experimental research on different similarity measures, we found that the fusion of these two kinds of similarity can greatly improve the performance of the LDA prediction model. Therefore, we chose Jaccard similarity and Gaussian interaction profile kernel similarity for LDA prediction in this work. Thank you again for your comment. The Jaccard similarity coefficient (Jaccard, 1908) of disease was calculated by LDA matrix by Eq. 1:

$$JDS(i,j) = \frac{LD(:,i) \cap LD(:,j)}{LD(:,i) \cup LD(:,j)} \tag{1}$$

In Eq. 1, $LD(:,i)$ is the i-th column vector of the LDA matrix, which represents the association feature of disease i; $LD(:,i) \cap LD(:,j)$ represents the number of lncRNAs that are associated with both disease i and disease j; $LD(:,i) \cup LD(:,j)$ represents the sum of the number of lncRNAs associated with the disease i and disease j.

Similarly to disease, the Jaccard similarity of lncRNA can be calculated by LDA matrix by Eq. 2:

$$JFS(i,j) = \frac{LD(i,:) \cap LD(j,:)}{LD(i,:) \cup LD(j,:)} \tag{2}$$

In Eq. 2, $LD(i,:)$ is the i-th row vector of the LDA matrix, which represents the association feature of lncRNA i;

$LD(i,:) \cap LD(j,:)$ represents the number of diseases that are associated with both lncRNA i and lncRNA j; $LD(i,:) \cup LD(j,:)$ represents the sum of the number of diseases associated with the lncRNA i and lncRNA j.

2.3 Calculation of Gaussian interaction profile kernel similarity of disease and lncRNA

The Gaussian interaction profile kernel similarity (Chen, 2015b) $GIP_{lnc}(l_i, l_j)$ between lncRNA l_i and lncRNA l_j was calculated by Eq. 3:

$$\begin{cases} GIP_{lnc}(l_i, l_j) = exp(-\lambda \| LD(i,:) - LD(j,:) \|^2) \\ \lambda = \tilde{\lambda} \Big/ \left(\frac{1}{l} \sum_{i=1}^{l} \| l_i \|^2 \right) \end{cases} \tag{3}$$

From the above equation, the Gaussian interaction profile kernel similarity matrix of lncRNA can be obtained. $LD(i,:)$ and $LD(j,:)$ represents i-th and j-th row of LDA matrix respectively, $\tilde{\lambda}$ controls the kernel bandwidth, in this work, we set $\tilde{\lambda}$ to 1.

Similarly, the Gaussian interaction profile kernel similarity matrix of disease $GIP_{dis}(d_i, d_j)$ can be obtained by Eq. 4.

$$\begin{cases} GIP_{dis}(d_i, d_j) = exp(-\lambda \| LD(:,i) - LD(:,j) \|^2) \\ \lambda = \tilde{\lambda} \Big/ \left(\frac{1}{d} \sum_{i=1}^{d} \| d_i \|^2 \right) \end{cases} \tag{4}$$

In Eq. 4, $LD(:,i)$ and $LD(:,j)$ represents i-th and j-th column of LDA matrix respectively, $\tilde{\lambda}$ controls the kernel bandwidth, in this work, we set $\tilde{\lambda}$ to 1.

2.4 Fusing different similarities for lncRNA and disease

In this paper, we used the maximum value method to merge lncRNA Gaussian interaction profile kernel similarity and lncRNA Jaccard similarity into LFJ similarity and fuse disease Gaussian interaction profile kernel similarity and disease Jaccard similarity into DSJ similarity by Eqs. 5, 6, respectively.

$$LFJ\ similarity = \begin{cases} GIP_{lnc}(l_i, l_j) & if\ GIP_{lnc}(l_i, l_j) \geq JFS(i,j) \\ JFS(i,j) & otherwise \end{cases} \tag{5}$$

$$DSJ\ similarity = \begin{cases} GIP_{dis}(d_i, d_j) & if\ GIP_{dis}(d_i, d_j) \geq JDS(i,j) \\ JDS(i,j) & otherwise \end{cases} \tag{6}$$

2.5 Geometric complement for lncRNA-disease associations matrix

The process of constructing the GCHIRFLDA model is divided into three steps (see Figure 1): 1) geometric complement for LDA matrix; 2) feature representation and extraction; 3) random forest classifier training and LDA prediction. Next, we will introduce the process of constructing the GCHIRFLDA model in detail.

Inspired by Francesco et al.'s and Yin et al.'s method (Wang et al., 2021; Yin et al., 2022), from the previous data source, we multiplied the LMI matrix with the MDA matrix and then divided the [i, j]-th element of the result by the i-th row of the LMI matrix and the j-th column of the MDA matrix to represent the potential LDA matrix by Eq. 7:

$$\mathrm{LMD}(\mathrm{i},\mathrm{j}) = \frac{LM(i,:) \cdot MD(:,j)}{\| LM(i,:) \|_1 + \| MD(:,j) \|_1} \tag{7}$$

The fusion matrix of LDA was obtained by taking the maximum value of the potential LDAs computed above and the original LDA matrix in the i-th row and j-th column by Eq. 8.

$$LD_{\mathrm{new}}(i,j) = \max(LD(i,j), \mathrm{LMD}(\mathrm{i},\mathrm{j})) \tag{8}$$

In this way, the original LDA matrix can be geometrically complemented.

2.6 Feature representation and extraction

For the obtained geometric complement matrix, each row represents the feature vector of lncRNA and each column represents the feature vector of disease. We combine the i-th row of the geometric complement matrix and the i-th row of the similarity fusion matrix of lncRNA to form a new feature vector of the i-th lncRNA. Similarly, we combine the j-th column of the geometric complement matrix and the j-th column of the similarity fusion matrix of disease to form a new feature vector of the j-th disease. Finally, each lncRNA and disease is represented as a 652-dimensional feature vector.

Autoencoder is an unsupervised neural network model and has a good performance in data denoising and dimensionality reduction. In the GCHIRFLDA model, we employee autoencoder to compress feature space of lncRNA and disease. We set hidden layer to learn the high-dimensional feature space of the input data so that the hidden layer can reconstruct the original input data (Schmidhuber, 2015; Ji et al., 2021).

In this work, we use an autoencoder with an input layer, a dense layer, an output layer and a fully-connected layer with an activation function sigmoid. The learning process of the noise-reducing encoder is to minimize the error between the reconstructed data and the original data. As a result, each lncRNA, which is originally represented by a 652-dimensional

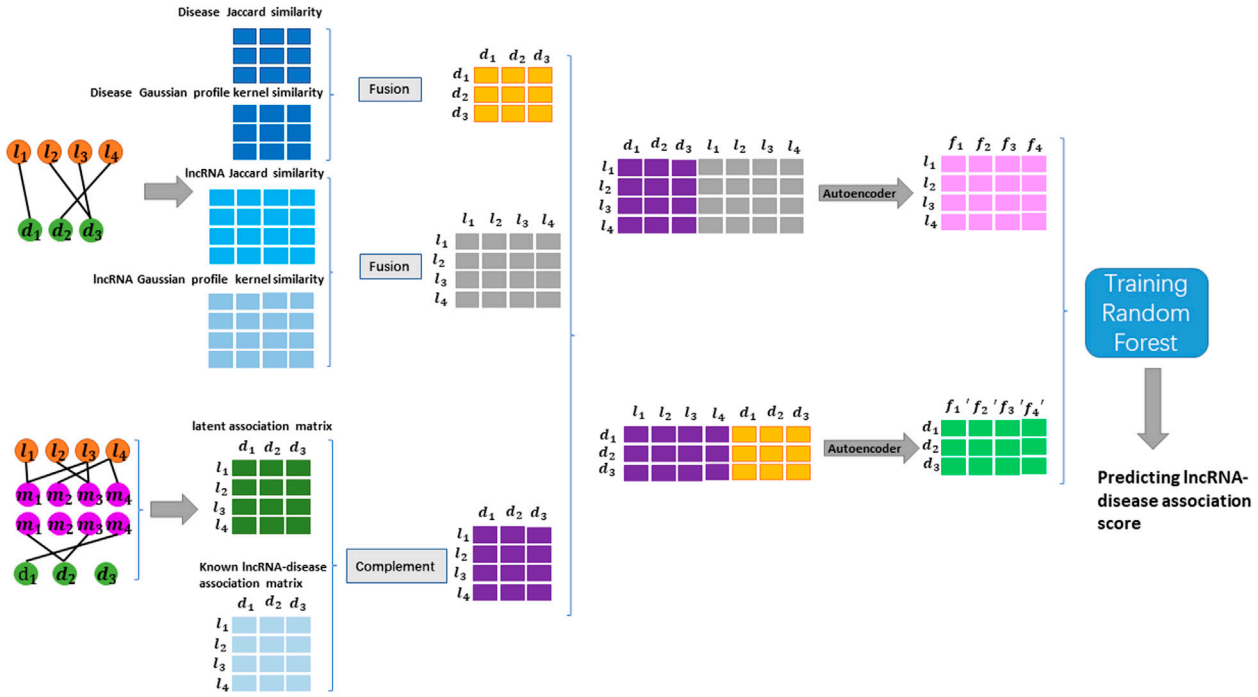


FIGURE 1
The flowchart of constructing the GCHIRFLDA model.

feature vector, is finally compressed into 256-dimensional by autoencoder. Similarly, each disease, which is originally represented by a 652-dimensional feature vector, is finally compressed into 256-dimensional by autoencoder. MSE (mean squared error) is used as model loss evaluation by Eq. 9:

$$\text{loss} = \frac{1}{n}\sum\left(Y_{input} - Y_{output}\right)^2 \tag{9}$$

In Eq. 9, Y_{input} is the original input data, and Y_{output} is the decoded and reconstructed data.

2.7 Random forest classifier training and lncRNA-disease associations prediction

To train the GCHIRFLDA model, the experiment-supported 2697 LDAs in the original LDA matrix were used as positive samples; the remaining lncRNA-disease pairs that were not validated by biological experiments were used as unlabeled samples. To maintain the balance of the training set, an equal number of unlabeled samples were randomly selected from the unlabeled samples as negative samples. The negative samples and the positive samples were combined into the training sample set which consisted of 5394 samples.

For accurately predicting potential LDAs, we employed random forest (RF) for LDA prediction in the GCHIRFLDA model. Random forest is an ensemble machine learning model which combines bagging and random features to add extra diversity of the decision tree model and finally uses a voting method to combine the prediction results of multiple base classifiers (Breiman, 2001). RF has many advantages: 1) it can process a variety of data types, including qualitative data or quantitative data; 2) it has high classification accuracy; 3) it has good robustness for noise data and data with missing values; 4) it has ability to analyze complex interactions between features. In recent years, RF has been widely used in a variety of classification and prediction problems, including differential expression analysis of microarray data, miRNA-disease association prediction, etc. In this work, we have carried out experimental research on six different classifiers, including SVM and Xgboost. Considering AUC, AUPR, Recall and other indicators, the performance of RF classifier is the best. Therefore, RF was chosen as the final classifier in our prediction model. RF has two important parameters, namely the number of randomly selected features (*mtry*) and the number of trees (*ntree*). These parameters have a great impact on the performance of random forest classification model. Here, we set *mtry* and *ntree* by the default value. Then, by the obtained prediction model, all unconfirmed lncRNA-disease pairs are scored, and the closer the score is to 1, the more likely it is that lncRNA is associated with the disease.

TABLE 1 The AUCs under different lncRNA/disease feature dimension.

Dimension	16	32	64	128	256	512
16	0.9576	0.9724	0.9768	0.9782	0.9750	0.9724
32	0.9492	0.9753	0.9775	0.9809	0.9804	0.9788
64	0.9577	0.9760	0.9791	0.9833	0.9842	0.9826
128	0.9561	0.9764	0.9808	0.9872	0.9884	0.9877
256	0.9539	0.9736	0.9804	0.9874	**0.9897**	0.9889
512	0.9109	0.9711	0.9793	0.9880	0.9891	0.9890

TABLE 2 The performance comparison of different classifiers in the GCHIRFLDA model.

Classifier	AUC	AUPR	Recall	Accuracy	F1-score
Xgboost	0.9815	0.4544	0.9523	0.9182	0.9523
RF	**0.9897**	**0.7040**	**0.9673**	**0.9317**	**0.9597**
C50	0.9513	0.1517	0.9340	0.8724	0.9265
GBDT	0.9497	0.2348	0.8942	0.8701	0.9253
SVM	0.9832	0.5826	0.9243	0.9313	0.9595
LightGBM	0.9832	0.5250	0.9428	0.9215	0.9541

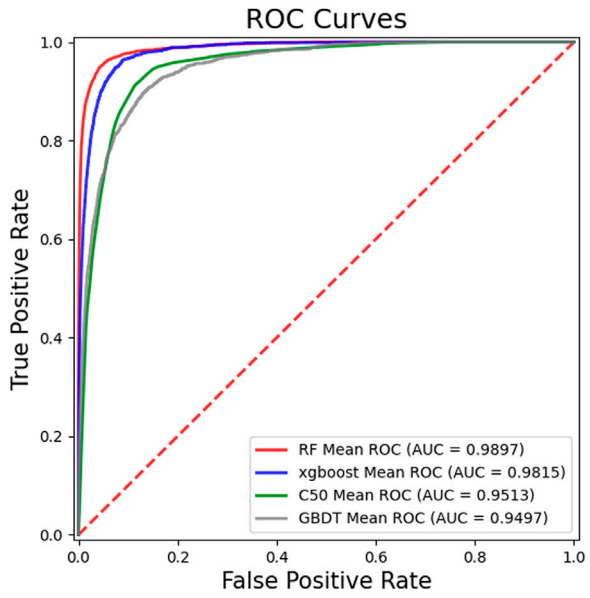


FIGURE 2
The ROC Curves of different classifiers in the GCHIRFLDA model.

3 Results

3.1 Feature dimension analysis of lncRNA and disease

For LDA prediction, the dimensionality of the training sample set has an obvious impact on the accuracy of the prediction model. On the one hand, for a smaller number of features of lncRNAs and diseases, more features are not learned, which leads to under-fitting of the model. On the other hand, for a larger number of features, more time is spent and the model performance will not yet be greatly improved or even over-fitting will occur. Therefore, we used the experimental method to determine the appropriate feature dimension. Specifically, we use autoencoder to compress the dimensions of feature space into 16, 32, 64, 128, 256, and 512 respectively, and the feature dimension that makes the prediction performance of the model the highest is adopted. Table 1 shows the AUC obtained under five-fold cross-validation by different dimensional features, from which one can see that the maximum of AUC is reached when the feature dimension of both lncRNAs and diseases is 256, so we set the feature dimension of extracted lncRNAs and diseases by autoencoder to be 256.

3.2 Performance comparison between random forest and other classifiers

In order to obtain better performance of the GCHIRFLDA model, we compared RF classifier with several classical classifiers, including extreme gradient boosting (Xgboost) (Chen and Guestrin, 2016), C50 (Kuhn, 2013), Gradient Boosting Decision Tree (GBDT) (Ye et al., 2009), SVM (Lan et al., 2017) and LightGBM (Zhang et al., 2021). In this work, we used the average AUC, AUPR, Recall, F1-score and Accuracy based on five-fold cross-validation as evaluation criterion for the six classifiers.

Figure 2 showed the ROC curves and AUCs of different classifiers, from which one can see that the AUC values of RF, Xgboost, C50 and GBDT are 0.9897, 0.9814, 0.98959 and 0.9497, respectively. Figure 3 showed the PR curves and AUPRs of four classifiers, the AUPR values of RF, Xgboost, C50 and GBDT are 0.704, 0.4505, 0.1607 and 0.2336, respectively. Table 2 showed the AUC, AUPR, Recall, F1-score and Accuracy of six classifiers. As one can see from Table 2, all five metrics of RF is the largest among the six classifiers. The results of the experiments suggested that RF outperformed the other five classifiers for LDA prediction. There, RF was finally determined as the final classifier in the GCHIRFLDA model.

3.3 Performance comparison between GCHIRFLDA and other lncRNA-disease associations prediction models

To evaluate the prediction performance of the GCHIRFLDA model, we compared it with seven state-of-the-art LDA

TABLE 3 The AUCs and AUPRs of different LDA prediction models.

Method	AUC	AUPR
GCHIRFLDA	**0.990**	**0.704**
GAERF	0.980	0.491
GCNLDA	0.959	0.223
CNNLDA	0.952	0.251
LDAP	0.863	0.166
MFLDA	0.626	0.066
Ping's Method	0.871	0.219
SIMCLDA	0.746	0.095

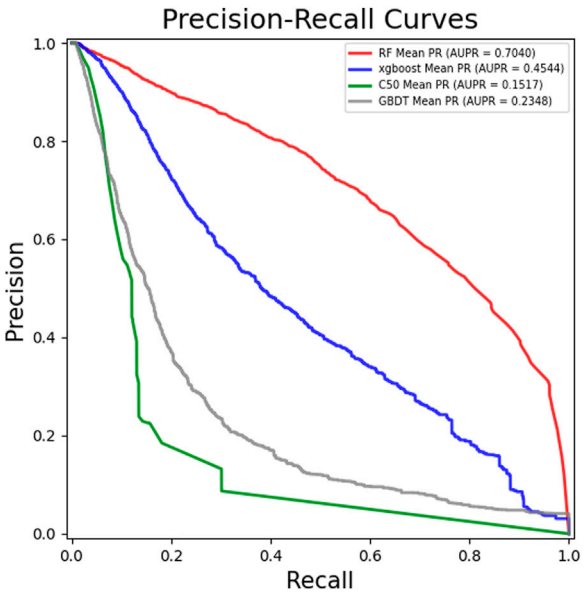


FIGURE 3
The Precision-Recall Curves of different classifiers in the GCHIRFLDA model.

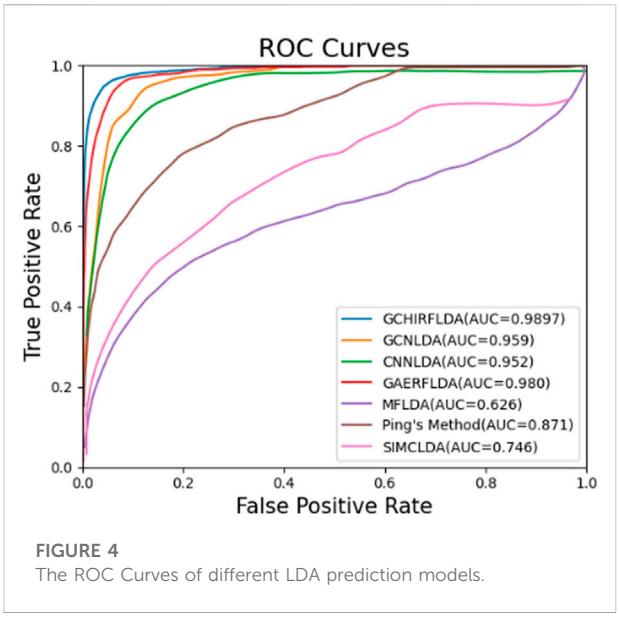


FIGURE 4
The ROC Curves of different LDA prediction models.

prediction models, including GAERF (Wu Q.-W. et al., 2021), CNNLDA (Xuan et al., 2019a), GCNLDA (Xuan et al., 2019c), MFLDA (Fu et al., 2018), Ping's method (Ping et al., 2019) and SIMLDA (Lu et al., 2018). The AUCs and AUPRs of all LDA prediction models are listed in Table 3. Figure 3 showed the ROC curves for these LDA prediction models.

From Table 3 and Figure 4, one can see that the AUC and AUPR of the GCHIRFLDA model are maximal among all LDA prediction models, which achieved 0.990 and 0.704, respectively. In term of AUC, our model achieved 0.990 which was 0.99%, 3.23%, 3.96%, 58.19%, 13.63%, and 32.67% higher than GAERF, GCNLDA, CNNLDA, MFLDA, Ping's method and SIMCLDA, respectively. In term of AUPR, our model achieved 0.704 which was 43.38%, 215.79%, 180.47%, 966.67%, 221.46%, 634.38% higher than GAERF, GCNLDA, CNNLDA, MFLDA, Ping's Method and SIMCLDA, respectively. According to the results of cross validation experiments, our GCHIRFLDA model has better LDA prediction ability.

3.4 Case studies

To further validate the prediction ability of the GCHIRFLDA model, we conducted case studies on two most common cancers, colon cancer and stomach cancer. We used the GCHIRFLDA to score all the unlabeled lncRNA-disease pairs, and selected the top 20 lncRNAs most likely to be associated with stomach cancer and colon cancer respectively according to the score. Finally, the predicted stomach cancer-associated and colon cancer-associated lncRNAs by the GCHIRFLDA model were validated by data from Lnc2Cancer v3.0 (Ning et al., 2016), LncRNADisease v2.0 (Bao et al., 2019) and some published research literature.

Colon cancer is the third most common cancer worldwide and the fourth leading cause of cancer-related death. The incidence of colon cancer has increased dramatically in China because of a shift in our habits as a society (Xue et al., 2015). In this work, we used the GCHIRFLDA to predict colon cance-associated lncRNAs. As a result, the top 20 predicted lncRNAs associated with colon cancer and the provenances of the evidence are shown in Table 4. As one can see from Table 4, 17 predicted lncRNAs have been confirmed by records included in the Lnc2Cancer (v3.0) or LncRNADisease (v2.0) or published literature. For example, Wan et al. showed that the overexpressing of CDKN2B-AS1 exhibited accelerated proliferation in colon cancer (Wan et al., 2013). Xu et al. reported the tumor

TABLE 4 The top 20 colon cancer-related lncRNA candidates predicted by the GCHIRFLDA model.

lncRNA	Rank	Evidence
CDKN2B-AS1	1	Lnc2Cancer 3.0& LncRNADisease v2.0
PVT1	2	Lnc2Cancer 3.0& LncRNADisease v2.0
UCA1	3	Lnc2Cancer 3.0& LncRNADisease v2.0
NEAT1	4	Lnc2Cancer 3.0& LncRNADisease v2.0
KCNQ1OT1	5	Lnc2Cancer 3.0
XIST	6	Lnc2Cancer 3.0& LncRNADisease v2.0
GAS5	7	Lnc2Cancer 3.0& LncRNADisease v2.0
SPRY4-IT1	8	Lnc2Cancer 3.0& LncRNADisease v2.0
MIR17HG	9	Literature (Xu et al., 2019)
TUG1	10	Lnc2Cancer 3.0& LncRNADisease v2.0
BANCR	11	Lnc2Cancer 3.0& LncRNADisease v2.0
HOTTIP	12	Lnc2Cancer 3.0& LncRNADisease v2.0
BCYRN1	13	LncRNADiseasev2.0
HNF1A-AS1	14	Lnc2Cancer 3.0
AFAP1-AS1	15	Lnc2Cancer 3.0
HULC	16	Lnc2Cancer 3.0
TUSC7	17	Lnc2Cancer 3.0
KIRREL3-AS3	18	unknown
LSINCT5	19	unknown
NPTN-IT1	20	unknown

suppressor B-cell linker (BLNK) was reduced in expression *via* MIR17HG, which resulted in an increase in invasion and migration of colorectal cancer cells (Xu et al., 2019).

In the digestive tract, stomach cancer is one of the most prevalent malignancies (Gu et al., 2017). The identification of new biomolecular markers of stomach cancer is essential for treatment and diagnosis. In this work, we used the GCHIRFLDA to predict stomach cancer-associated lncRNAs. As a result, the top 20 predicted lncRNAs associated with colon cancer and the provenances of the evidence are shown in Table 5. As seen in Table 5, 18 predicted lncRNAs have been confirmed by records included in the Lnc2Cancer (v3.0) or LncRNADisease (v2.0) or published literature. For example, Feng et al. revealed that KCNQ1OT1 inhibited stomach cancer cell progression *via* regulating miR-9 and LMX1A expression (Feng et al., 2020); Wu et al. found the high expression of lncRNA-CCAT2 indicated poor prognosis of stomach cancer and promoted cell proliferation and invasion (Wu et al., 2017). Consequently, the case studies on colon cancer and stomach cancer showed that GCHIRFLDA was an excellent predictor.

4 Conclusion

In this work, we proposed a geometric complement heterogeneous information and random forest-based approach for predicting LDAs (named GCHIRFLDA). Firstly, the potential

TABLE 5 The top 20stomach cancer-related lncRNA candidates predicted by the GCHIRFLDA model.

lncRNA	Rank	Evidence
MALAT1	1	Lnc2Cancer 3.0& LncRNADisease v2.0
XIST	2	Lnc2Cancer 3.0& LncRNADisease v2.0
NEAT1	3	Lnc2Cancer 3.0& LncRNADisease v2.0
CCAT2	4	Lnc2Cancer 3.0& LncRNADisease v2.0
TUG1	5	Lnc2Cancer 3.0& LncRNADisease v2.0
KCNQ1OT1	6	Lnc2Cancer 3.0
HOTTIP	7	Lnc2Cancer 3.0& LncRNADisease v2.0
WT1-AS	8	Lnc2Cancer 3.0& LncRNADisease v2.0
HNF1A-AS1	9	Lnc2Cancer 3.0& LncRNADisease v2.0
HULC	10	Lnc2Cancer 3.0& LncRNADisease v2.0
MIR17HG	11	Literature (Bahari et al., 2015)
CRNDE	12	Lnc2Cancer 3.0& LncRNADisease v2.0
NPTN-IT1	13	Lnc2Cancer 3.0& LncRNADisease v2.0
LINC00675	14	Lnc2Cancer 3.0
KIRREL3-AS3	15	unknown
TP53COR1	16	unknown
BCYRN1	17	Lnc2Cancer 3.0
HOTAIRM1	18	Lnc2Cancer 3.0
AFAP1-AS1	19	LncRNADisease v.2.0
LINC01133	20	Lnc2Cancer 3.0

LDA matrix is constructed by integrating the LMIs and MDAs with the original LDA matrix. Then, the Jaccard similarity and the Gaussian interaction profile similarity of lncRNA and disease are combined to represent features of lncRNA and disease. Next, a low-dimensional feature space is extracted by using autoencoder. Finally, RF is employed as the classifier to predict potential LDAs. In conclusion, the AUC and AUPR comparison with other LDA prediction models based on five-fold cross-validation and the case studies show that our model has better LDA prediction performance.

Although the GCHIRFLDA model has a good performance, it still has some limitations. Firstly, the lack of data verified by biological experimental is a big shortcoming for computational models. Secondly, randomly selecting the unknown lncRNA-disease pairs as negative samples may incorrectly classify potential positive samples as negative samples, which may affect the prediction performance. Finally, only the heterogeneous information of miRNAs is introduced in this work, and in the future, more biological information will be fused to improve the performance of the LDA prediction model.

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

TZ conceived and implemented the model, performed the experiments, and wrote the paper. DY directed the research and revised the paper. XjZ and XrZ analyzed the experimental results and revised the paper. CZ performed the experiments. All authors have read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 62172128) and the Postdoctoral Research Start Fund of Heilongjiang Province (LBH-Q20098). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of interest

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

Publisher's note

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

References

Bahari, F., Emadi-Baygi, M., and Nikpour, P. (2015). miR-17-92 host gene, uderexpressed in gastric cancer and its expression was negatively correlated with the metastasis. *Indian J. Cancer* 52, 22–25. doi:10.4103/0019-509X.175605

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

Barsyte-Lovejoy, D., Lau, S. K., Boutros, P. C., Khosravi, F., Jurisica, I., Andrulis, I. L., et al. (2006). The c-Myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis. *Cancer Res.* 66, 5330–5337. doi:10.1158/0008-5472.can-06-0037

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

Cheetham, S. W., Gruhl, F., Mattick, J. S., and Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. *Br. J. Cancer* 108, 2419–2425. doi:10.1038/bjc.2013.233

Chen, T. Q., and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proc. 22nd Acm Sigkdd Int. Conf. Knowl. Discov. Data Min. Kdd'16*, 785–794.

Chen, X., Clarence Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338. doi:10.1038/srep11338

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

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

Chen, X. (2015c). Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* 5, 13186. doi:10.1038/srep13186

Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017). Long non-coding RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform* 18, 558–576. doi:10.1093/bib/bbw060

Cui, Z., Liu, J. X., Gao, Y. L., Zhu, R., and Yuan, S. S. (2020). LncRNA-disease associations prediction using bipartite local model with nearest profile-based association inferring. *IEEE J. Biomed. Health Inf.* 24, 1519–1527. doi:10.1109/jbhi.2019.2937827

Fan, Y., Chen, M., and Pan, X. (2022). Gcrflda: Scoring lncRNA-disease associations using graph convolution matrix completion with conditional random field. *Brief. Bioinform* 23, bbab361. doi:10.1093/bib/bbab361

Feng, L., Li, H., Li, F., Bei, S., and Zhang, X. (2020). LncRNA KCNQ1OT1 regulates microRNA-9-LMX1A expression and inhibits gastric cancer cell progression. *Aging* 12, 707–717. doi:10.18632/aging.102651

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 34, 1529–1537. doi:10.1093/bioinformatics/btx794

Gao, M. M., Cui, Z., Gao, Y. L., Wang, J., and Liu, J. X. (2021). Multi-label fusion collaborative matrix factorization for predicting LncRNA-disease associations. *IEEE J. Biomed. Health Inf.* 25, 881–890. doi:10.1109/jbhi.2020.2988720

Gu, J., Li, Y., Fan, L., Zhao, Q., Tan, B., Hua, K., et al. (2017). Identification of aberrantly expressed long non-coding RNAs in stomach adenocarcinoma. *Oncotarget* 8, 49201–49216. doi:10.18632/oncotarget.17329

Huang, Y. A., Chen, X., You, Z. H., Huang, D. S., and Chan, K. C. (2016). Ilncsim: Improved lncRNA functional similarity calculation model. *Oncotarget* 7, 25902–25914. doi:10.18632/oncotarget.8296

Jaccard, P. (1908). Nouvelles recherches sur la Distribution florale. *Bull. Soc. Vaudoise Sci. Nat.* 44, 223–270.

Ji, C., Gao, Z., Ma, X., Wu, Q., Ni, J., and Zheng, C. (2021). Aemda: Inferring miRNA-disease associations based on deep autoencoder. *Bioinformatics* 37, 66–72. doi:10.1093/bioinformatics/btaa670

Kuhn, M. (2013). *Classification using C5. 0 UseR! 2013*. CT, USA: Pfizer Global R&D: Groton.

Lan, W., Lai, D., Chen, Q., Wu, X., Chen, B., Liu, J., et al. (2022). Ldicdl: LncRNA-disease association identification based on collaborative deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 19, 1715–1723. doi:10.1109/tcbb.2020.3034910

Lan, W., Li, M., Zhao, K., Liu, J., Wu, F. X., Pan, Y., et al. (2017). Ldap: A web server for lncRNA-disease association prediction. *Bioinformatics* 33, 458–460. doi:10.1093/bioinformatics/btw639

Liu, J. X., Cui, Z., Gao, Y. L., and Kong, X. Z. (2021). Wgrcmf: A weighted graph regularized collaborative matrix factorization method for predicting novel LncRNA-disease associations. *IEEE J. Biomed. Health Inf.* 25, 257–265. doi:10.1109/jbhi.2020.2985703

Liu, M. X., Chen, X., Chen, G., Cui, Q. H., and Yan, G. Y. (2014). A computational framework to infer human disease-associated long noncoding RNAs. *Plos One* 9. doi:10.1371/journal.pone.0084408

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

Mathieu, E. L., Belhocine, M., Dao, L. T., Puthier, D., and Spicuglia, S. (2014). Rôle des longs ARN non codants dans le développement normal et pathologique. *Med. Sci. Paris.* 30, 790–796. doi:10.1051/medsci/20143008018

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

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2019). A novel method for LncRNA-disease association prediction based on an lncRNA-disease association network. *IEEE/ACM Trans. Comput. Biol. Bioinf.* 16, 688–693. doi:10.1109/tcbb.2018.2827373

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Netw.* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Sumathipala, M., Maiorino, E., Weiss, S. T., and Sharma, A. (2019). Network diffusion approach to predict LncRNA disease associations using multi-type biological networks: Lion. *Front. Physiol.* 10, 888. doi:10.3389/fphys.2019.00888

Sun, J., Shi, H. B., Wang, Z. Z., Zhang, C. J., Liu, L., Wang, L. T., 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

Sun, W., Shi, Y., Wang, Z., Zhang, J., Cai, H., Zhang, J., et al. (2018). Interaction of long-chain non-coding RNAs and important signaling pathways on human cancers (Review). *Int. J. Oncol.* 53, 2343–2355. doi:10.3892/ijo.2018.4575

Taft, R. J., Pang, K. C., Mercer, T. R., Dinger, M., and Mattick, J. S. (2010). Non-coding RNAs: Regulators of disease. *J. Pathol.* 220, 126–139. doi:10.1002/path.2638

Wan, G., Mathur, R., Hu, X., Liu, Y., Zhang, X., Peng, G., et al. (2013). Long non-coding RNA ANRIL (CDKN2B-AS) is induced by the ATM-E2F1 signaling pathway. *Cell. Signal.* 25, 1086–1095. doi:10.1016/j.cellsig.2013.02.006

Wang, B., Zhang, C., Du, X.-X., and Zhang, J.-F. (2021). lncRNA-disease association prediction based on latent factor model and projection. *Sci. Rep.* 11, 19965. doi:10.1038/s41598-021-99493-5

Wang, Y., Yu, G., Wang, J., Fu, G., Guo, M., and Domeniconi, C. (2020). Weighted matrix factorization on multi-relational data for LncRNA-disease association prediction. *Methods* 173, 32–43. doi:10.1016/j.ymeth.2019.06.015

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res.* 24, 616–628. doi:10.1101/gr.165035.113

Wu, Q.-W., Xia, J.-F., Ni, J.-C., and Zheng, C.-H. (2021a). Gaerf: Predicting lncRNA-disease associations by graph auto-encoder and random forest. *Brief. Bioinform* 22, bbaa391. doi:10.1093/bib/bbaa391

Wu, Q. W., Cao, R. F., Xia, J., Ni, J. C., Zheng, C. H., and Su, Y. (2021b). Extra trees method for predicting LncRNA-disease association based on multi-layer graph embedding aggregation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/tcbb.2021.3113122

Wu, S. W., Hao, Y. P., Qiu, J. H., Zhang, D. B., Yu, C. G., and Li, W. H. (2017). High expression of long non-coding RNA CCAT2 indicates poor prognosis of gastric cancer and promotes cell proliferation and invasion. *Minerva Med.* 108, 317–323. doi:10.23736/S0026-4806.17.04703-6

Wu, X., Lan, W., Chen, Q., Dong, Y., Liu, J., and Peng, W. (2020). Inferring LncRNA-disease associations based on graph autoencoder matrix completion. *Comput. Biol. Chem.* 87, 107282. doi:10.1016/j.compbiolchem.2020.107282

Xie, H., Ma, B., Gao, Q., Zhan, H., Liu, Y., Chen, Z., et al. (2018). Long non-coding RNA CRNDE in cancer prognosis: Review and meta-analysis. *Clin. Chim. Acta* 485, 262–271. doi:10.1016/j.cca.2018.07.003

Xu, J., Meng, Q., Li, X., Yang, H., Xu, J., Gao, N., et al. (2019). Long noncoding RNA MIR17HG promotes colorectal cancer progression via miR-17-5p. *Cancer Res.* 79, 4882–4895. doi:10.1158/0008-5472.can-18-3880

Xuan, P., Cao, Y., Zhang, T., Kong, R., and Zhang, Z. (2019a). Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. *Front. Genet.* 10, 416. doi:10.3389/fgene.2019.00416

Xuan, P., Jia, L., Zhang, T., Sheng, N., Li, X., and Li, J. (2019b). LDAPred: A method based on information flow propagation and a convolutional neural network for the prediction of disease-associated lncRNAs. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20184458

Xuan, P., Pan, S., Zhang, T., Liu, Y., and Sun, H. (2019c). Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells* 8. doi:10.3390/cells8091012

Xuan, P., Sheng, N., Zhang, T., Liu, Y., and Guo, Y. (2019d). Cnndlp: A method based on convolutional autoencoder and convolutional neural network with adjacent edge attention for predicting lncRNA-disease associations. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20174260

Xue, Y., Ma, G. X., Gu, D. Y., Zhu, L. J., Hua, Q. H., Du, M. L., et al. (2015). Genome-wide analysis of long noncoding RNA signature in human colorectal cancer. *Gene* 556, 227–234. doi:10.1016/j.gene.2014.11.060

Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., and Wang, J. (2020). A random forest based computational model for predicting novel lncRNA-disease associations. *BMC Bioinforma.* 21, 126. doi:10.1186/s12859-020-3458-1

Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009). "Stochastic gradient boosted distributed decision trees," in *Proceedings of the 18th ACM conference on Information and knowledge management* (Hong Kong, China: Association for Computing Machinery). doi:10.1145/1645953.1646301

Yin, M. M., Liu, J. X., Gao, Y. L., Kong, X. Z., and Zheng, C. H. (2022). Ncplp: A novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. *IEEE Trans. Cybern.* 52, 5079–5087. doi:10.1109/tcyb.2020.3026652

Yu, G., Fu, G., Lu, C., Ren, Y., and Wang, J. (2017). Brwlda: Bi-random walks for predicting lncRNA-disease associations. *Oncotarget* 8, 60429–60446. doi:10.18632/oncotarget.19588

Yu, G., Wang, Y., Wang, J., Domeniconi, C., Guo, M., and Zhang, X. (2020). Attributed heterogeneous network fusion via collaborative matrix tri-factorization. *Inf. Fusion* 63, 153–165. doi:10.1016/j.inffus.2020.06.012

Zhang, C., Lei, X. J., and Liu, N. (2021). Predicting metabolite-disease associations based on LightGBM model. *Front. Genet.* 12, 660275. doi:10.3389/fgene.2021.660275

Zhou, J. R., You, Z. H., Cheng, L., and Ji, B. Y. (2021). Prediction of lncRNA-disease associations via an embedding learning HOPE in heterogeneous information networks. *Mol. Ther. - Nucleic Acids* 23, 277–285. doi:10.1016/j.omtn.2020.10.040

Zhou, M., Wang, X. J., Li, J. W., Hao, D. P., Wang, Z. Z., Shi, H. B., et al. (2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769. doi:10.1039/c4mb00511b

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 September 2022
DOI 10.3389/fgene.2022.995535

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Jia Qu,
Changzhou University, China
Zhanchao Li,
Guangdong Pharmaceutical University, China
JunLin Xu,
Hunan University, China

*CORRESPONDENCE
Dengju Yao,
ydkvictory@hrbust.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 16 July 2022
ACCEPTED 08 August 2022
PUBLISHED 13 September 2022

CITATION
Wei Z, Yao D, Zhan X and Zhang S (2022), A clustering-based sampling method for miRNA-disease association prediction.
Front. Genet. 13:995535.
doi: 10.3389/fgene.2022.995535

A clustering-based sampling method for miRNA-disease association prediction

Zheng Wei[1], Dengju Yao[1]*, Xiaojuan Zhan[1,2] and Shuli Zhang[1]

[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China, [2]College of Computer Science and Technology, Heilongjiang Institute of Technology, Harbin, China

More and more studies have proved that microRNAs (miRNAs) play a critical role in gene expression regulation, and the irregular expression of miRNAs tends to be associated with a variety of complex human diseases. Because of the high cost and low efficiency of identifying disease-associated miRNAs through biological experiments, scholars have focused on predicting potential disease-associated miRNAs by computational methods. Considering that the existing methods are flawed in constructing negative sample set, we proposed a clustering-based sampling method for miRNA-disease association prediction (CSMDA). Firstly, we integrated multiple similarity information of miRNA and disease to represent miRNA-disease pairs. Secondly, we performed a clustering-based sampling method to avoid introducing potential positive samples when constructing negative sample set. Thirdly, we employed a random forest-based feature selection method to reduce noise and redundant information in the high-dimensional feature space. Finally, we implemented an ensemble learning framework for predicting miRNA-disease associations by soft voting. The Precision, Recall, F1-score, AUROC and AUPR of the CSMDA achieved 0.9676, 0.9545, 0.9610, 0.9928, and 0.9940, respectively, under five-fold cross-validation. Besides, case study on three cancers showed that the top 20 potentially associated miRNAs predicted by the CSMDA were confirmed by the dbDEMC database or literatures. The above results demonstrate that the CSMDA can predict potential disease-associated miRNAs more accurately.

KEYWORDS

miRNA-disease association, ensemble learning, clustering, sampling, computational methods

1 Introduction

MicroRNAs (miRNAs) are a kind of non-coding RNAs with a length of 20–24 nucleotides, which play a critical role in gene expression regulation (Lee et al., 1993; Wightman et al., 1993; He & Hannon, 2004). Accumulating evidences have showed that the dysregulation of miRNA is associated with human complex diseases (Hwang & Mendell, 2006; Mattick & Makunin, 2006; Jonas & Izaurralde, 2015). Wang et al. have proved that the expression level of hsa-mir20b-5p is

associated with the pathogenesis of Alzheimer's disease (Wang et al., 2022). Taverner et al. have proposed that microRNA-425–5p and microRNA-451 can be used as the risk biomarkers of cardiovascular disease (Taverner et al., 2021). Ma et al. have showed that the overexpression of microRNA-10b promotes invasion and metastasis of mammary tumor cells (Ma et al., 2007). Hashimoto et al. have demonstrated that the abnormal expression of miR-1307–3p in human serum is associated with a variety of malignant tumors (Hashimoto et al., 2021). Therefore, accurately identifying disease-associated miRNAs can facilitate the study of the mechanism of miRNA in complex diseases. To guide complex biological experiments, many computational models have been developed for predicting miRNA-disease associations (Chen et al., 2019a).

Thus far, scholars have proposed a series of network-based miRNA-disease association prediction models (Bandyopadhyay et al., 2010). Jiang et al. integrated a human miRNA-phenome network and a miRNA function-related network for predicting disease-associated miRNAs (Jiang et al., 2010). Shi et al. mapped the pathogenic disease genes and miRNA target genes into the protein-protein interaction network, and employed the random walk with restart to identify miRNA-disease associations (Shi et al., 2013). Zeng et al. implemented a structural perturbation approach for miRNA-disease association prediction on a bilayer network which integrated the known miRNA-disease associations and miRNA (disease) similarity network (Zeng et al., 2018). Xiao et al. first calculated the weighted K nearest neighbor profiles of miRNAs and diseases, and then used graph regularized matrix factorization to predict miRNA-disease associations (Xiao et al., 2018). Zhong et al. proposed a global method based on non-negative matrix factorization, which could simultaneously predict all disease-related miRNAs (Zhong et al., 2018). Ma et al. presented a miRNA-disease association prediction model which did not depend on any known miRNA-disease associations (Ma et al., 2019). Li et al. constructed a heterogeneous bilayer network by integrating similarity networks and interaction network, and then utilized the algorithm faster randomized partial matrix completion to infer latent disease-lncRNA associations (Li et al., 2019). Yu et al. proposed a knowledge-driven method to predict disease-miRNA associations (KDFGMDA) (Yu et al., 2022). Based on dynamic neighborhood regularized logistic matrix factorization, Yan et al. proposed a method (DNRLMF-MDA) to predict miRNA-disease associations (Yan et al., 2019). Qu et al. proposed a biased random walk computational method for miRNA-disease association prediction (BRWRMHMDA), which was restarted on multilayer heterogeneous networks (Qu et al., 2021). Jiang and Zhu proposed a model of decision template-based miRNA-disease association prediction (DTMDA) (Jiang & Zhu, 2020).

In recent decades, dozens of miRNA-disease association prediction models based on machine learning have been proposed. One of the major challenges facing these models is how to construct negative samples set. Yao et al. implemented an improved random forest-based model for miRNA-disease association prediction (IRFMDA) which constructed negative samples by randomly combining miRNAs and diseases (Yao et al., 2019). Zhao et al. proposed an adaptive boosting model (ABMDA) which employed the k-means algorithm to cluster the unlabeled samples and selected negative samples randomly from each cluster (Zhao et al., 2019). Zhou et al. designed a miRNA-disease association prediction model based on gradient boosting decision tree and logistic regression (GBDT-LR) which applied the k-means algorithm to cluster the unlabeled samples and extracted negative samples from each cluster by the ratio of the size of each cluster to the entire unlabeled sample set size (Zhou et al., 2020). Li et al. proposed a graph auto-encoder-based miRNA-disease association prediction model (GAEMDA) which randomly selected 5,430 unlabeled samples as negative samples (Li et al., 2021). Chen et al. proposed an anti-noise miRNA-disease association prediction algorithm (ANMDA) which applied the k-means algorithm to cluster the unlabeled samples and selected negative samples equally from each cluster to reduce the noise (Chen et al., 2021). Dai et al. presented a resampling-based ensemble framework (ERMDA) which constructed multiple balanced training subsets by resampling and obtained the final prediction result by soft voting strategy (Dai et al., 2022). Liu et al. proposed a new novel method via deep forest ensemble learning based on autoencoder (DFELMDA) to predict miRNA-disease associations (Liu et al., 2022). Chen et al. presented a model of extreme gradient boosting machine for miRNA-disease association (EGBMMDA), which calculated the statistical measures and matrix factorization results for each miRNA-disease pair to form an information feature vector (Chen et al., 2018). The above methods inevitably introduced potential positive samples into negative sample set, which limited the prediction performance of these models (Rayhan et al., 2017).

In this paper, we proposed a novel clustering-based sampling method for miRNA-disease association prediction (CSMDA) which could construct more reliable negative sample set. Firstly, the CSMDA integrated a variety of similarity information of miRNA and disease to represent the feature vector of miRNA-disease pairs. Secondly, the CSMDA constructed negative sample set based on MiniBatchKMeans clustering to reduce the proportion of potentially positive samples in the negative samples set. Thirdly, the CSMDA generated numerous training subsets through multiple rounds of sampling on the negative sample set to reduce the bias caused by single small-scale sampling. Fourthly, the CSMDA applied a random forest-based feature selection approach to reduce noise and redundant information in the high-dimensional feature

space. Finally, a set of base classifiers were trained on the training subsets after feature selection and the final prediction result was obtained by soft voting. The Precision, Recall, F1-score, AUROC and AUPR of the CSMDA achieved 0.9676, 0.9545, 0.9610, 0.9928 and 0.9940 under 5-fold cross-validation, which was significantly higher than that of the existing methods. Besides, case study on three cancers showed that all the top 20 miRNAs predicted to be most likely associated with these cancers by the CSMDA were confirmed by the dbDEMC database or literatures.

2 Materials and methods

2.1 Experimentally confirmed miRNA-disease associations

Experimentally confirmed 5,430 miRNA-disease associations were obtained from the HMDD (Human microRNA Disease Database) (Li et al., 2014), including 495 miRNAs and 383 diseases. Here, we stored these miRNA-disease associations by a matrix $MD_{N_m \times N_d}$, which was defined as:

$$\mathrm{MD}(m(i), d(j)) = \begin{cases} 1, & miRNA\, m(i) and\, disease\, d(j) are\, verified\, to\, be\, related \\ 0, & miRNA\, m(i) and\, disease\, d(j) are\, not\, verified\, to\, be\, related \end{cases} \quad (1)$$

Here, N_m and N_d represent the number of miRNAs and diseases, respectively.

2.2 Disease semantic similarity

The descriptors of 383 diseases mentioned above were obtained from the MeSH (Medical Subject Headings) database and Directed Acyclic Graphs (DAGs) for each disease were constructed by the previous methods (Wang et al., 2010; Xuan et al., 2013). In a $DAG\ (D)$, the nodes represent disease D and its ancestral nodes, and the directed edges represent the relationship of diseases. The semantic contribution of disease d to disease D in $DAG\ (D)$ was defined as follows:

$$D1_D(d) = \begin{cases} 1, & d = D \\ \max\{\Delta \times D1_D(d') | d' \in children\, of\, d\}, & d \neq D \end{cases} \quad (2)$$

Here, Δ is the semantic contribution factor. As the distance between D and other diseases in $DAG(D)$ increases, the semantic contribution of these diseases will decrease. Then, the semantic value of disease D was defined as follows:

$$DV1(D) = \sum_{d \in T(D)} D1_D(d) \quad (3)$$

Here, $T(D)$ represents the disease D and its all ancestral nodes. For two diseases, $d(k)$ and $d(l)$, the disease semantic similarity between them was defined as follows:

$$SS1(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} \left(D1_{d(i)}(t) + D1_{d(j)}(t) \right)}{DV1(d(i)) + DV1(d(j))} \quad (4)$$

Considering two different diseases in the same layer of a $DAG\ (D)$, if the occurrence rate of one disease is different from another, their semantic contribution to disease D should be different. Inspired by Xuan et al. (Xuan et al., 2013), another way to calculate the semantic contribution of disease d in $DAG\ (D)$ to disease D was defined as follows:

$$D2_D(d) = -log \frac{the\, number\, of\, DAGs\, including\, d}{the\, number\, of\, disease} \quad (5)$$

Similarly, the disease semantic value $DV2(D)$ of disease D was defined as follows:

$$DV2(D) = \sum_{d \in T(D)} D2_D(d) \quad (6)$$

Then, the disease semantic similarity between disease $d(i)$ and disease $d(j)$ was defined as follows:

$$SS2(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} \left(D2_{d(i)}(t) + D2_{d(j)}(t) \right)}{DV2(d(i)) + DV2(d(j))} \quad (7)$$

Finally, we combined the above two methods to calculate the disease semantic similarity of disease $d(i)$ and $d(j)$ as follows:

$$SS(d(i), d(j)) = \frac{SS1(d(i), d(j)) + SS2(d(i), d(j))}{2} \quad (8)$$

2.3 Gaussian interaction profile kernel similarity for diseases

Based on the assumption that miRNAs with similar functions tend to be related to diseases with similar phenotypes (van Laarhoven et al., 2011), Gaussian interaction profile kernel (GIPK) similarity for diseases was introduced to represent the relationship between diseases from another perspective. Here, let $IP\ (d(i))$ represent the i th column vector of the miRNA-disease association matrix MD, which denotes whether there are verified associations between disease $d(i)$ and each miRNA. Then, the GIPK similarity of disease $d(i)$ and $d(j)$ was defined as follows:

$$GD(d(i), d(j)) = \exp\left(-\gamma_d \left\| IP\ (d(i)) - IP\ (d(j)) \right\|^2 \right) \quad (9)$$

In Eq. 9, parameter γ_d controls the kernel bandwidth and was calculated by the following formula:

$$\gamma_d = \frac{\gamma_d'}{\frac{1}{N_d} \sum_{i=1}^{N_d} \left\| IP\ (d(i)) \right\|^2} \quad (10)$$

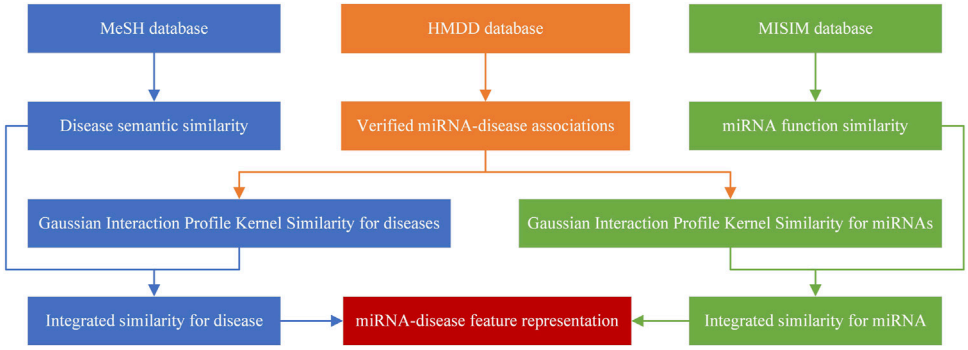


FIGURE 1
The method of sample representation.

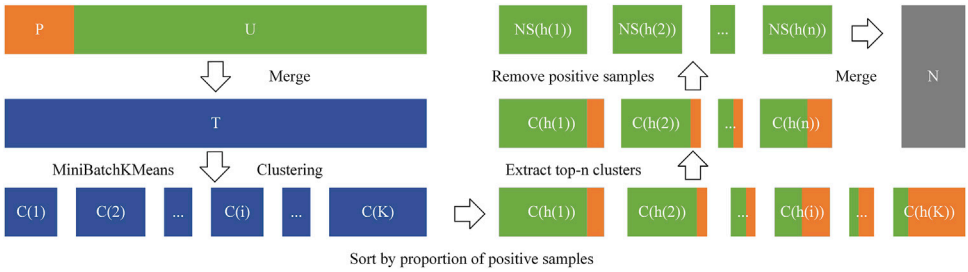


FIGURE 2
The method of constructing a negative sample set.

According to the previous study (Chen & Yan, 2013; Chen et al., 2016), $\gamma_d^{'}$ was set to 1 here.

2.4 Integrated similarity of diseases

Since there may be no semantic similarity between two diseases, we integrated semantic similarity and GIPK similarity of disease here. Inspired by previous works (Dai et al., 2022), the integrated disease similarity between $d(i)$ and $d(j)$ was defined as follows:

$$IDS(d(i),d(j)) = \begin{cases} SS(d(i),d(j)), & SS(d(i),d(j)) \neq 0 \\ GD(d(i),d(j)), & SS(d(i),d(j)) = 0 \end{cases} \tag{11}$$

2.5 MiRNA functional similarity

Based on the hypothesis that miRNAs with similar functions tend to be associated with diseases with similar phenotypes, miRNA functional similarity can be calculated (Wang et al., 2010). Here, we directly obtained miRNA functional similarity from the MISIM database (http://www.cuilab.cn/fi les/images/cuilab/misim.zip) and represented them by $FS(m(i),m(j))$.

2.6 Gaussian interaction profile kernel similarity for miRNAs

Similar to disease, the GIPK similarity between miRNA $m(i)$ and $m(j)$ was defined as follows:

$$GM(m(i),m(j)) = \exp\left(-\gamma_m \left\| IP(m(i)) - IP(m(j)) \right\|^2\right) \tag{12}$$

$$\gamma_m = \frac{\gamma_m^{'}}{\frac{1}{N_m}\sum_{i=1}^{N_m} \left\| IP(m(i)) \right\|^2} \tag{13}$$

Here, $IP(m(i))$ represent the i th row vector of miRNA-disease associations matrix MD, which indicates whether there are verified associations between miRNA $m(i)$ and each disease. Inspired by previous works (Chen & Yan, 2013; Chen et al., 2016), $\gamma_m^{'}$ was set to 1 here.

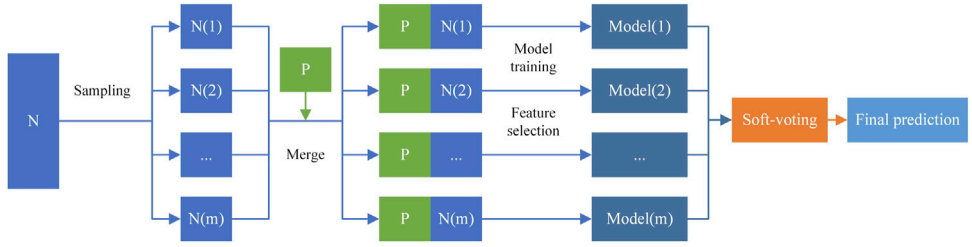


FIGURE 3
Ensemble learning framework.

2.7 Integrated similarity of miRNAs

Since there may be no functional similarity between two miRNAs, we integrated the miRNA functional similarity and the GIPK similarity of miRNA $m(i)$ and $m(j)$. Inspired by previous works (Dai et al., 2022), the integrated miRNA similarity between $m(i)$ and $m(j)$ was defined as follows:

$$IMS(m(i),m(j)) = \begin{cases} FS(m(i),m(j)), & FS(m(i),m(j)) \neq 0 \\ GM(m(i),m(j)), & FS(m(i),m(j)) = 0 \end{cases} \tag{14}$$

2.8 Sample representation

Here, a miRNA-disease pair was taken as a sample. The feature vector of disease $d(i)$ was defined as follow:

$$FD(d(i)) = (IDS(d(i),d(1)), IDS(d(i),d(2)), \ldots, IDS(d(i),d(N_d))) \tag{15}$$

Similarly, the feature vector of miRNA $m(j)$ was defined as follow:

$$FM(m(j)) = (IMS(m(j),m(1)), IMS(m(j),m(2)), \ldots, IMS(m(j),m(N_m))) \tag{16}$$

Then, the feature vector of a sample $(d(i),m(j))$ was defined as follow:

$$F(d(i),m(j)) = (FD(d(i)), FM(m(j))) \tag{17}$$

The method of sample representation is shown in Figure 1.

2.9 Constructing negative sample set

In this work, the 5,430 experimentally confirmed miRNA-disease associations were taken as positive samples and the 184,155 unverified miRNA-disease pairs as unlabeled samples. Most methods (Yao et al., 2019; Zhao et al., 2019; Zhou et al., 2020; Chen et al., 2021; Li et al., 2021; Dai et al., 2022) of constructing negative sample set are to randomly select some unlabeled samples as negative samples, or apply k-means clustering on the unlabeled samples and sample negative examples from the resulted clusters. However, these methods may introduce potential positive samples into negative sample set and lead to the performance degradation of the trained model (Chen et al., 2021). Here, we proposed a novel and effective method to construct negative sample set from the total sample set. Firstly, we defined the positive sample set P, and the unlabeled sample set U:

$$P = \{F(d(i),m(j)) | MD(m(j),d(i)) = 1\} \tag{18}$$

$$U = \{F(d(i),m(j)) | MD(m(j),d(i)) = 0\} \tag{19}$$

And we defined the total sample set T as follows:

$$T = P \cup U \tag{20}$$

Secondly, according to the hypothesis that in the total sample set, the smaller the Minkowski distance between the two samples, the more likely they are to be the same kind of samples (Hartigan & Wong, 1979), we clustered T into K clusters by the MiniBatchKMeans (Pedregosa et al., 2011). The formula for calculating Minkowski distance was as following Eq. 21.

$$D_{mk}(x,y) = \left(\sum_{u=1}^{n} |x_u - y_u|^p \right)^{\frac{1}{p}} \tag{21}$$

MiniBatchkmeans is an optimization of K-Means algorithm. It uses mini-batches to reduce the amount of computation required to converge to a local solution, thereby reducing the computing time required for clustering the large-scale dataset. To ensure the accuracy of clustering results, we repeated clustering ten times. Then, we denoted the K clusters as follows:

$$C(1), C(2), \ldots, C(K) \tag{22}$$

The proportion of positive samples in the i th cluster was defined as follows:

$$p(i) = \frac{|C(i) - U|}{|C(i)|}, \; i \in \{1, 2, \ldots, K\} \tag{23}$$

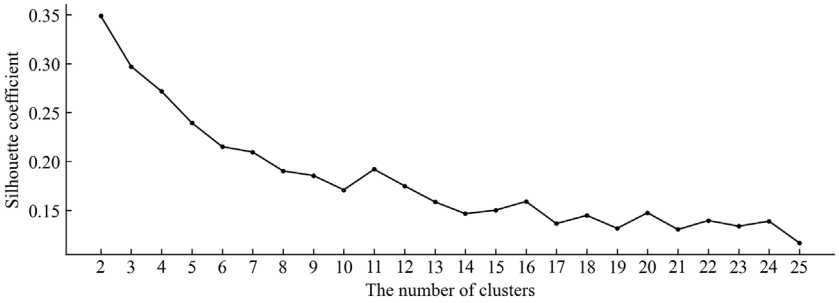


FIGURE 4
The silhouette coefficient of clustering results under different numbers of clusters.

TABLE 1 Performance comparison of the CSMDA using different base classifiers.

Model	Precision	Recall	F1-score	AUROC	AUPR
CSMDA-AB	0.9567	0.9267	0.9414	0.9885	0.9901
CSMDA-ERT	0.9666	0.9514	0.9589	0.9907	0.9926
CSMDA-RF	**0.97**	0.9468	0.9582	0.9912	0.9929
CSMDA-XGB	0.9674	**0.9543**	**0.9608**	**0.9927**	**0.9939**

Thirdly, we ranked all clusters by $p(i)$, and then denoted the top n $(n < K)$ clusters with the fewest $p(i)$ as follows:

$$C(h(1)), C(h(2)), \ldots, C(h(i)), \ldots C(h(n)) \tag{24}$$

Here, $C(h(i))$ represents the cluster with the i th fewest $p(i)$.

Finally, we defined the i th negative sample set $NS(h(i))$ as follows:

$$NS(h(i)) = C(h(i)) - P, i \in \{1, 2, \ldots, n\} \tag{25}$$

Here, $NS(h(i))$ represents the cluster $C(h(i))$ after removing the positive sample.

Then, we constructed the total negative sample set N as follows:

$$N = NS(h(1)) \cup NS(h(2)) \cup \ldots NS(h(n)) \tag{26}$$

The number of samples in the negative sample N set constructed by the above method is 119,659. The method of constructing a negative sample set is shown in Figure 2.

2.10 Ensemble learning framework

In this work, we implemented an ensemble learning framework for miRNA-disease association prediction. Inspired by the previous research (Chen et al., 2019b; Dai et al., 2020; Sherazi et al., 2021; Wang et al., 2021; Zeng et al., 2021), we built the CSMDA through the following three stages: 1) construct multiple training subsets to increase the diversity of base classifiers by randomly sampling from N; 2) perform the random forest-based feature selection to reduce noise and redundant information in the high-dimensional feature space; 3) use soft voting strategy to integrate the prediction results of all base classifiers. The process of constructing the ensemble learning framework is shown in Figure 3.

2.10.1 Constructing training subsets

In this work, we constructed multiple different training subsets and balanced them to improve the prediction performance of the CSMDA. On the one hand, the diversity of subsets makes base classifiers discrepant from each other and improves the generalization ability of the CSMDA. On the other hand, multiple disparate training subsets can make full use of all negative samples. Here, we defined the size of the P as $|P|$. First, all samples in P were regarded as positive samples. Second, the $|P|$ negative samples were randomly sample from N. Finally, the positive and negative samples were combined into each training subset. In this work, we constructed ten training subsets through the above methods for the CSMDA.

2.10.2 Feature selection on each training subset

In the CSMDA, each miRNA-disease feature vector has 878 dimensions, which may contain a large amount of noise and redundant information. Inspired by previous research (Yao et al., 2019; Dai et al., 2022), we performed feature selection based on random forest variable importance score on each training subset. First, we trained a random forest model on each training subset and sorted all features by the variable importance scores which were generated by the random forest. Then, we selected the top X features with the highest variable importance scores to form a new feature space for each subset.

2.10.3 Soft voting strategy

In this work, the Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016) was used as base classifier. Here, let

TABLE 2 Performance comparison of the CSMDA under different dimension training samples.

Model	Precision	Recall	F1-score	AUROC	AUPR
CSMDA-NOFS	0.9674	0.9543	0.9608	0.9927	0.9939
CSMDA-FS75	**0.9676**	0.9545	**0.9610**	**0.9928**	**0.9940**
CSMDA-FS50	0.9667	**0.9551**	0.9608	0.9927	0.9939
CSMDA-FS25	0.9657	0.9540	0.9598	0.9916	0.9930

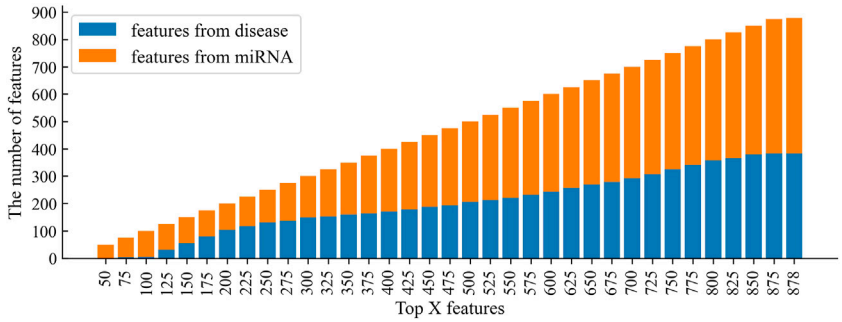


FIGURE 5
The distribution of features from miRNAs and diseases among the top X features.

m represent the number of training subsets. Take an unknown miRNA-disease pair as sample input, m base classifiers could produce m prediction result for the sample, and then the m prediction results were integrated by the soft voting strategy (Sherazi et al., 2021; Wang et al., 2021; Zeng et al., 2021). Specifically, the output of the i th sample by soft voting was defined as follows:

$$O(i) = \frac{1}{m}\sum_{j=1}^{m} O(i,j) \tag{27}$$

Here, $O(i,j)$ represents the prediction scores of the j th classifier for the i th sample. If $O(i) > 0.5$, the miRNA-disease pair were regarded to be associated; otherwise, it was considered to be not associated.

3 Results

3.1 Performance evaluation criteria

In this work, we employed five-fold cross-validation to evaluate the performance of the CSMDA. Firstly, we adopted the known 5,430 miRNA-disease association pairs as positive samples and randomly selected an equal number of samples from the negative sample set N as negative samples. Then, all positive samples and all negative samples were combined into a sample set. Next, the constructed sample set was divided into five parts, and in each cross-validation, one part was taken out and merged with unlabeled samples to make up the test sample set, and the remaining four parts were all used as the training sample set. Here, we evaluated the CSMDA by five metrics: Precision, Recall, F1-score, AUC (Area under the receiver operating characteristic curve) and AUPR (Area under the precision-recall curve). The receiver operating characteristic (ROC) curves were obtained by plotting the true positive rate (TPR) and false-positive rate (FPR) under different levels of thresholds, and then the area under of ROC (AUC) was computed (Hajian-Tilaki, 2013). The higher the turning point of the ROC curve to the upper left, the closer the AUC is to 1, indicating the better performance of the model. The formulae for computing TPR and FPR were as following Eq. 28 and Eq. 29.

$$TPR = \frac{TP}{TP + FN} \tag{28}$$

$$FPR = \frac{FP}{FP + TN} \tag{29}$$

The Precision-Recall (PR) curves were obtained by plotting the Precision and Recall rates under different levels of thresholds, and then the area under of PR curve (AUPR) was computed (Saito & Rehmsmeier, 2015). Similarly, the higher the turning point of the PR curve to the upper right, the closer the AUPR is to 1, indicating that the model has a better performance in predicting. The formulae for computing Precision and Recall were as following Eq. 30 and Eq. 31.

TABLE 3 Performance comparison of the CSMDA with other MDA prediction models.

Model	Precision	Recall	F1-score	AUROC	AUPR
ABMDA [19]	0.8213 ± 0.0033	0.8371 ± 0.0044	0.8290 ± 0.0030	0.9023 ± 0.0021	0.8879 ± 0.0032
ANMDA [22]	0.8561 ± 0.0017	0.8728 ± 0.0020	0.8643 ± 0.0014	0.9373 ± 0.0005	0.9328 ± 0.0008
GAEMDA [21]	0.8146 ± 0.0031	0.9111 ± 0.0028	0.8597 ± 0.0010	0.9352 ± 0.0001	0.8850 ± 0.0010
GBDT-LR [20]	0.8403 ± 0.0026	0.8567 ± 0.0031	0.8484 ± 0.0021	0.9246 ± 0.0010	0.9177 ± 0.0015
IRFMDA [18]	0.8447 ± 0.0021	0.8598 ± 0.0025	0.8521 ± 0.0016	0.9267 ± 0.0009	0.9222 ± 0.0012
ERMDA [23]	0.8740 ± 0.0039	0.9043 ± 0.0019	0.8889 ± 0.0022	0.9561 ± 0.0013	0.9542 ± 0.0020
CSMDA	**0.9676 ± 0.0052**	**0.9545 ± 0.0059**	**0.9610 ± 0.0042**	**0.9928 ± 0.0012**	**0.9940 ± 0.0009**

$$Precision = \frac{TP}{TP + FP} \tag{30}$$

$$Recall = \frac{TP}{TP + FN} \tag{31}$$

Furthermore, F1-Score, as a comprehensive metric, is a toned-down average of precision and recall and is used to balance the effects of precision and recall and evaluate a classifier more comprehensively. In addition, the Accuracy is the result of the correct classification of the response model. The F1-Score and Accuracy can be calculated as Eq. 32 and Eq. 33 as followed.

$$F1\text{-}score = \frac{2*Precision*Recall}{Precision + Recall} \tag{32}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{33}$$

3.2 Performance analysis of clustering

In constructing the negative sample set, the number of clusters K is the key factor affecting the effectiveness of the final clustering. In this work, the silhouette coefficient (SC) (Rousseeuw, 1987) was adopted as the cluster validity index to evaluate the validity of clustering results with different cluster numbers. The silhouette coefficient is a kind of internal index to judge criteria of clustering result and it is calculated as follows:

$$SC(o) = \frac{b(o) - a(o)}{max\{a(o), b(o)\}} \tag{34}$$

Here, $a(o)$ represents the average distance between sample o and other samples in its cluster, and $b(o)$ represents the minimum average distance between sample o and samples in other clusters. The value of $SC(o)$ ranges from -1 to 1, and $SC(o)$ getting closer to 1 indicates that the cluster algorithm works better. First, T was divided into 2, 3 ... 24, and 25 clusters by MiniBatchKMeans clustering. Then, according to each sample and its label obtained through clustering, the silhouette coefficient was calculated in turn. The silhouette coefficient with a different number of clusters is shown in Figure 4. As one can see, the silhouette coefficient decreases gradually with the increase of the number of clusters and achieves a maximum of 0.349 when the number of clusters is 2. Therefore, we set the values of K to 2 in the CSMDA.

3.3 Performance analysis of base classifier

Base classifier plays an importance role in the prediction performance of the ensemble learning framework. In this work, we compared the performance of four base classifiers: AdaBoost, Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Extremely Randomized Trees (ExtRa Trees). For optimal performance, we optimized the hyper-parameters of each model. The prediction performance of the CSMDA using different base classifiers are listed in Table 1. As one can see, the Precision of the XGBoost is 0.9674, the Recall is 0.9543, the F1-score is 0.9608, the AUROC is 0.9927 and the AUPR is 0.9939. The XGBoost is lower than the RF in terms of Precision, but it is higher than other models in all other metrics. Therefore, the XGBoost was employed in the CSMDA.

3.4 Feature dimension analysis of samples

In the feature selection, according to the variable importance scores, 100, 75, 50, and 25% features were selected from the original feature space to construct the training set, denoted as CSMDA-NOFS, CSMDA-FS75, CSMDA-FS50, and CSMDA-FS25, respectively. Then, we evaluated the prediction performance of the CSMDA with different number of features, and the results were listed in Table 2. As one can see, when the dimension of the training sample is 75% of the length of the original feature vector, the effect of feature selection on improving the performance of the CAMDA is optimum. Therefore, we set the feature dimension of the training set to 75% of the length of the original feature vector. We further analyzed the contribution of miRNA and disease to the feature vector, the distribution of features from miRNAs and diseases among the X

TABLE 4 The top 20 miRNAs for three cancers predicted by the CSMDA.

Disease	Rank	miRNA	Evidence
breast cancer	1	hsa-mir-195	dbDEMC
	2	hsa-mir-146a	dbDEMC
	3	hsa-mir-24	dbDEMC
	4	hsa-let-7e	dbDEMC
	5	hsa-mir-9	dbDEMC
	6	hsa-mir-219	dbDEMC
	7	hsa-mir-148a	dbDEMC
	8	hsa-mir-218	dbDEMC
	9	hsa-let-7a	dbDEMC
	10	hsa-mir-29a	dbDEMC
	11	hsa-mir-223	dbDEMC
	12	hsa-mir-30d	dbDEMC
	13	hsa-mir-92a	dbDEMC
	14	hsa-mir-210	dbDEMC
	15	hsa-mir-200c	dbDEMC
	16	hsa-mir-17	dbDEMC
	17	hsa-mir-214	dbDEMC
	18	hsa-mir-372	dbDEMC
	19	hsa-mir-106b	dbDEMC
	20	hsa-mir-221	dbDEMC
colon cancer	1	hsa-mir-24	dbDEMC
	2	hsa-mir-20a	dbDEMC
	3	hsa-mir-125b	dbDEMC
	4	hsa-mir-182	dbDEMC
	5	hsa-mir-29a	dbDEMC
	6	hsa-mir-214	dbDEMC
	7	hsa-mir-17	dbDEMC
	8	hsa-mir-21	dbDEMC
	9	hsa-mir-30b	dbDEMC
	10	hsa-mir-29b	dbDEMC
	11	hsa-mir-19b	dbDEMC
	12	hsa-mir-19a	dbDEMC
	13	hsa-mir-18a	dbDEMC
	14	hsa-mir-141	dbDEMC
	15	hsa-mir-155	dbDEMC
	16	hsa-mir-223	dbDEMC
	17	hsa-mir-127	dbDEMC
	18	hsa-mir-34c	Hiyoshi, Y., et al. [40]
	19	hsa-mir-1	dbDEMC
	20	hsa-mir-126	dbDEMC
lung cancer	1	hsa-mir-29c	dbDEMC
	2	hsa-mir-92a	dbDEMC
	3	hsa-mir-206	dbDEMC
	4	hsa-mir-214	dbDEMC
	5	hsa-mir-183	dbDEMC
	6	hsa-mir-210	dbDEMC
	7	hsa-mir-142	dbDEMC
	8	hsa-mir-221	dbDEMC
	9	hsa-mir-30e	dbDEMC
	10	hsa-mir-24	dbDEMC
	11	hsa-mir-223	dbDEMC
	12	hsa-mir-20b	dbDEMC
	13	hsa-mir-193b	dbDEMC
	14	hsa-mir-191	dbDEMC
	15	hsa-mir-22	dbDEMC
	16	hsa-mir-124	dbDEMC
	17	hsa-mir-18b	dbDEMC
	18	hsa-mir-30a	dbDEMC
	19	hsa-mir-148a	dbDEMC
	20	hsa-mir-15b	dbDEMC

(Continued in next column)

TABLE 4 (*Continued*) The top 20 miRNAs for three cancers predicted by the CSMDA.

features with the highest variable importance scores is shown in Figure 5. As we can see from Figure 5, the number of features from miRNAs is generally greater than that from diseases, which is consistent with the fact that the number of miRNAs is greater than that from the diseases. This indicates that feature selection based on the variable importance score is reasonable.

3.5 Performance comparison between clustering-based sampling method for miRNA-disease association prediction and other miRNA-disease association prediction models

To prove the ability of the CSMDA to predict potential disease-associated miRNAs, we compared it with six state-of-the-art MDA prediction models, including ABMDA (Zhao et al., 2019), ANMDA (Chen et al., 2021), GAEMDA (Li et al., 2021), GBDT-LR (Zhou et al., 2020), IRFMDA (Yao et al., 2019) and ERMDA (Dai et al., 2022). First, the CSMDA and other MDA prediction models constructed negative sample set by their respective methods. Secondly, we used the recommended hyper-parameters for these models. Finally, we performed 500 times five-fold cross-validation for each model. The performance of the above MDA prediction models are shown in Table 3. As one can see, the Precision, Recall, F1-score, AUC and AUPR of the CSMDA is 0.9676 ± 0.0052, 0.9545 ± 0.0059, 0.9610 ± 0.0042, 0.9928 ± 0.0012, and 0.9940 ± 0.0009 respectively, which superior to other methods in all

metrics. The results proved the outstanding prediction performance of the CSMDA.

3.6 Case studies

To prove the application value of the CSMDA in guiding biological experiments, we performed case studies on three common cancers, including breast cancer, colon cancer and lung cancer. Firstly, we combined the 5,430 positive samples verified by the experiment and the 5,430 negative samples randomly selected from the negative sample set *N* into the training set of CSMDA. Secondly, we identified the positive and negative samples to which the three diseases belong. Thirdly, in the case study of current cancer, remove all samples related to current cancer in the training set. Finally, we trained CSMDA on this training set, and scored miRNA-disease pairs related to current cancer by using the CSMDA. We verified the top 20 miRNAs predicted to be associated with each cancer, and the results were listed in Table 4. Here, we validated these predicted miRNAs through the dbDEMC (Database of differentially expressed miRNAs in human cancers) database (Yang et al., 2017) or literatures. As one can see from Table 4, for breast cancer and lung cancer, all predicted miRNAs were confirmed by the dbDEMC database; for colon cancer, all predicted miRNAs except hsa-mir-34c were confirmed by the dbDEMC database. However, Hiyoshi et al. demonstrated that the expression level of Mir-34C in human colon cancer cells was higher than that in non-tumor cells (Hiyoshi et al., 2015). In summary, case study demonstrated that the CSMDA was reliable for predicting disease-associated miRNAs.

4 Conclusion

In this work, we presented a clustering-based sampling method for predicting miRNA-disease associations, named CSMDA. Firstly, the CSMDA integrated similarity of disease and miRNA to represent samples. Secondly, the CSMDA implemented an effective clustering-based sampling method to construct negative sample set. Thirdly, the CSMDA employed a random forest-based feature selection method to reduce noise and redundant information in the high-dimensional feature space. Finally, the CSMDA implemented an ensemble learning framework for predicting miRNA-disease associations by soft voting. The experimental results and case studies on the three cancers demonstrate that the CSMDA is a reliable model to predict disease-associated miRNAs. The main contribution of the CSMDA is to propose a new method to construct a more effective negative sample set, which avoids the possibility of introducing potential positive samples into negative sample set as much as possible. The negative sample set constructed by our method not only makes CSMDA perform well, but also improves the performance of other MDA prediction models. However, it should be noted that there are several limitations to the CSMDA. First, it is still inevitable to introduce potential positive samples in the stage of constructing the negative sample set. Second, the clustering algorithm used in the CSMDA is MiniBatchKMeans which showed good clustering effect, but other clustering algorithms may make the negative sample set purer. We will study the clustering effect of other clustering algorithms on the total sample set in the next work. Finally, in current work, the information associated with miRNA and disease is limited, which may result in the essential features that are helpful to identify miRNA-disease associations not being extracted in the CSDMA. In the future, we will integrate more features related to disease and miRNA into the CSMDA. In summary, we hope that the CSMDA can help researchers make breakthroughs in the treatment of complex human diseases at the miRNA level.

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

ZW and DY designed the experiments and analyzed the data. ZW performed the experiments. ZW and DY wrote the paper. XZ and SZ reviewed and revised the paper. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (Grant No. 62172128), the Postdoctoral Research Start Fund of Heilongjiang Province (LBH-Q20098), and the Innovation and Entrepreneurship Training Program for College Students in Heilongjiang Province (s202110214009). The funding body did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgments

We would like to thank reviewers for their comments and suggestions.

Conflict of interest

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

Publisher's note

References

Bandyopadhyay, S., Mitra, R., Maulik, U., and Zhang, M. Q. (2010). Development of the Human Cancer microRNA Network. *Silence* 1 (1), 6. doi:10.1186/1758-907x-1-6

Chen, T., and Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. doi:10.1145/2939672.2939785

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018). EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association Prediction. *Cell. Death Dis.* 9 (1), 3. doi:10.1038/s41419-017-0003-x

Chen, X. J., Hua, X. Y., and Jiang, Z. R. (2021). ANMDA: Anti-noise Based Computational Model for Predicting Potential miRNA-Disease Associations. *BMC Bioinforma.* 22 (1), 358. doi:10.1186/s12859-021-04266-6

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019a). MicroRNAs and Complex Diseases: from Experimental Results to Computational Models. *Brief. Bioinform.* 20 (2), 515–539. doi:10.1093/bib/bbx130

Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L., Liu, Y., et al. (2016). WBSMDA: Within and between Score for MiRNA-Disease Association Prediction. *Sci. Rep.* 6, 21106. doi:10.1038/srep21106

Chen, X., and Yan, G. Y. (2013). Novel Human lncRNA-Disease Association Inference Based on lncRNA Expression Profiles. *Bioinformatics* 29 (20), 2617–2624. doi:10.1093/bioinformatics/btt426

Chen, X., Zhu, C. C., and Yin, J. (2019b). Ensemble of Decision Tree Reveals Potential miRNA-Disease Associations. *PLoS Comput. Biol.* 15 (7), e1007209. doi:10.1371/journal.pcbi.1007209

Dai, Q., Wang, Z., Liu, Z., Duan, X., Song, J., and Guo, M. (2022). Predicting miRNA-Disease Associations Using an Ensemble Learning Framework with Resampling Method. *Brief. Bioinform.* 23 (1), bbab543. doi:10.1093/bib/bbab543

Dai, Q., Wang, Z., Song, J., Duan, X., Guo, M., and Tian, Z. (2020). "A Stacked Ensemble Learning Framework with Heterogeneous Feature Combinations for Predicting ncRNA-Protein Interaction". in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Casp. J. Intern Med.* 4 (2), 627–635.

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.*, 28(1), 100–108. doi:10.2307/2346830

Hashimoto, K., Inada, M., Yamamoto, Y., and Ochiya, T. (2021). Preliminary Evaluation of miR-1307-3p in Human Serum for Detection of 13 Types of Solid Cancer Using microRNA Chip. *Heliyon* 7 (9), e07919. doi:10.1016/j.heliyon.2021.e07919

He, L., and Hannon, G. J. (2004). MicroRNAs: Small RNAs with a Big Role in Gene Regulation. *Nat. Rev. Genet.* 5 (7), 522–531. doi:10.1038/nrg1379

Hiyoshi, Y., Schetter, A. J., Okayama, H., Inamura, K., Anami, K., Nguyen, G. H., et al. (2015). Increased microRNA-34b and -34c Predominantly Expressed in Stromal Tissues Is Associated with Poor Prognosis in Human Colon Cancer. *PloS one* 10 (4), e0124899. doi:10.1371/journal.pone.0124899

Hwang, H. W., and Mendell, J. T. (2006). MicroRNAs in Cell Proliferation, Cell Death, and Tumorigenesis. *Br. J. Cancer* 94 (6), 776–780. doi:10.1038/sj.bjc.6603023

Jiang, L., and Zhu, J. (2020). Review of MiRNA-Disease Association Prediction. *Curr. Protein Pept. Sci.* 21 (11), 1044–1053. doi:10.2174/1389203721666200210102751

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

Jonas, S., and Izaurralde, E. (2015). Towards a Molecular Understanding of microRNA-Mediated Gene Silencing. *Nat. Rev. Genet.* 16 (7), 421–433. doi:10.1038/nrg3965

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 (5), 843–854. doi:10.1016/0092-8674(93)90529-y

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

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: a Database for Experimentally Supported Human microRNA and Disease Associations. *Nucleic Acids Res.* 42, D1070–D1074. doi:10.1093/nar/gkt1023

Li, Z., Li, J., Nie, R., You, Z. H., and Bao, W. (2021). A Graph Auto-Encoder Model for miRNA-Disease Associations Prediction. *Brief. Bioinform.* 22 (4), bbaa240. doi:10.1093/bib/bbaa240

Liu, W., Lin, H., Huang, L., Peng, L., Tang, T., Zhao, Q., et al. (2022). Identification of miRNA-Disease Associations via Deep Forest Ensemble Learning Based on Autoencoder. *Brief. Bioinform.* 23 (3), bbac104. doi:10.1093/bib/bbac104

Ma, L., Teruya-Feldstein, J., and Weinberg, R. A. (2007). Tumour Invasion and Metastasis Initiated by microRNA-10b in Breast Cancer. *Nature* 449 (7163), 682–688. doi:10.1038/nature06174

Ma, Y., He, T., Ge, L., Zhang, C., and Jiang, X. (2019). MiRNA-disease Interaction Prediction Based on Kernel Neighborhood Similarity and Multi-Network Bidirectional Propagation. *BMC Med. Genomics* 12, 185. doi:10.1186/s12920-019-0622-4

Mattick, J. S., and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.* 1, R17–R29. doi:10.1093/hmg/ddl046

Pedregosa, F., Varoquaux, G., Garmfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Qu, J., Wang, C. C., Cai, S. B., Zhao, W. D., Cheng, X. L., and Ming, Z. (2021). Biased Random Walk with Restart on Multilayer Heterogeneous Networks for MiRNA-Disease Association Prediction. *Front. Genet.* 12, 720327. doi:10.3389/fgene.2021.720327

Rayhan, F., Ahmed, S., Shatabda, S., Farid, D. M., Mousavian, Z., Dehzangi, A., et al. (2017). iDTI-ESBoost: Identification of Drug Target Interaction Using Evolutionary and Structural Features with Boosting. *Sci. Rep.* 7 (1), 17731. doi:10.1038/s41598-017-18025-2

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.*, 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Saito, T., and Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative Than the ROC Plot when Evaluating Binary Classifiers on Imbalanced Datasets. *PloS one* 10 (3), e0118432. doi:10.1371/journal.pone.0118432

Sherazi, S. W. A., Bae, J. W., and Lee, J. Y. (2021). A Soft Voting Ensemble Classifier for Early Prediction and Diagnosis of Occurrences of Major Adverse Cardiovascular Events for STEMI and NSTEMI during 2-year Follow-Up in Patients with Acute Coronary Syndrome. *PloS one* 16 (6), e0249338. doi:10.1371/journal.pone.0249338

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., et al. (2013). Walking the Interactome to Identify Human miRNA-disease Associations through the Functional Link between miRNA Targets and Disease Genes. *BMC Syst. Biol.* 7, 101. doi:10.1186/1752-0509-7-101

Taverner, D., Llop, D., Rosales, R., Ferré, R., Masana, L., Vallvé, J. C., et al. (2021). Plasma Expression of microRNA-425-5p and microRNA-451a as Biomarkers of

Cardiovascular Disease in Rheumatoid Arthritis Patients. *Sci. Rep.* 11 (1), 15670. doi:10.1038/s41598-021-95234-w

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian Interaction Profile Kernels for Predicting Drug-Target Interaction. *Bioinformatics* 27 (21), 3036–3043. doi:10.1093/bioinformatics/btr500

Wang, C., Ju, Y., Zou, Q., and Lin, C. (2021). DeepAc4C: A Convolutional Neural Network Model with Hybrid Features Composed of Physicochemical Patterns and Distributed Representation Information for Identification of N4-Acetylcytidine in mRNA. *Bioinformatics* 38, 52–57. doi:10.1093/bioinformatics/btab611

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 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, R., Chopra, N., Nho, K., Maloney, B., Obukhov, A. G., Nelson, P. T., et al. (2022). Human microRNA (miR-20b-5p) Modulates Alzheimer's Disease Pathways and Neuronal Function, and a Specific Polymorphism Close to the MIR20B Gene Influences Alzheimer's Biomarkers. *Mol. Psychiatry* 27 (2), 1256–1273. doi:10.1038/s41380-021-01351-3

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional Regulation of the Heterochronic Gene Lin-14 by Lin-4 Mediates Temporal Pattern Formation in *C. elegans*. *Cell.* 75 (5), 855–862. doi:10.1016/0092-8674(93)90530-4

Xiao, Q., Luo, J., Liang, C., Cai, J., and Ding, P. (2018). A Graph Regularized Non-negative Matrix Factorization Method for Identifying microRNA-Disease Associations. *Bioinformatics* 34 (2), 239–248. doi:10.1093/bioinformatics/btx545

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 (8), e70204. doi:10.1371/journal.pone.0070204

Yan, C., Wang, J., Ni, P., Lan, W., Wu, F. X., and Pan, Y. (2019). DNRLMF-MDA: Predicting microRNA-Disease Associations Based on Similarities of microRNAs and Diseases. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16 (1), 233–243. doi:10.1109/tcbb.2017.2776101

Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: Updated Database of Differentially Expressed miRNAs in Human Cancers. *Nucleic Acids Res.* 45, D812–D818. doi:10.1093/nar/gkw1079

Yao, D., Zhan, X., and Kwoh, C. K. (2019). An Improved Random Forest-Based Computational Model for Predicting Novel miRNA-Disease Associations. *BMC Bioinforma.* 20 (1), 624. doi:10.1186/s12859-019-3290-7

Yu, S., Wang, H., Liu, T., Liang, C., and Luo, J. (2022). A Knowledge-Driven Network for Fine-Grained Relationship Detection between miRNA and Disease. *Brief. Bioinform.* 23 (3), bbac058. doi:10.1093/bib/bbac058

Zeng, K., Xu, Y., Lin, G., Liang, L., and Hao, T. (2021). Automated Classification of Clinical Trial Eligibility Criteria Text Based on Ensemble Learning and Metric Learning. *BMC Med. Inf. Decis. Mak.* 21, 129. doi:10.1186/s12911-021-01492-z

Zeng, X., Liu, L., Lü, L., and Zou, Q. (2018). Prediction of Potential Disease-Associated microRNAs Using Structural Perturbation Method. *Bioinformatics* 34 (14), 2425–2432. doi:10.1093/bioinformatics/bty112

Zhao, Y., Chen, X., and Yin, J. (2019). Adaptive Boosting-Based Computational Model for Predicting Potential miRNA-Disease Associations. *Bioinformatics* 35 (22), 4730–4738. doi:10.1093/bioinformatics/btz297

Zhong, Y., Xuan, P., Wang, X., Zhang, T., Li, J., Liu, Y., et al. (2018). A Non-negative Matrix Factorization Based Method for Predicting Disease-Associated miRNAs in miRNA-Disease Bilayer Network. *Bioinformatics* 34 (2), 267–277. doi:10.1093/bioinformatics/btx546

Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting Potential miRNA-Disease Associations by Combining Gradient Boosting Decision Tree with Logistic Regression. *Comput. Biol. Chem.* 85, 107200. doi:10.1016/j.compbiolchem.2020.107200

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 15 September 2022
DOI 10.3389/fgene.2022.951537

OPEN ACCESS

EDITED BY
Min Zeng,
Central South University, China

REVIEWED BY
Hang Yin,
Harbin Medical University Cancer Hospital, China
Dan Wang,
Liaoning Cancer Hospital, China

*CORRESPONDENCE
Huan Ouyang,
auyeungfoon1010@163.com
Yongxiang Li,
liyongxiang@ahmu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 24 May 2022
ACCEPTED 19 August 2022
PUBLISHED 15 September 2022

CITATION
Chen Z, Xu J, Zha B, Li J, Li Y and Ouyang H (2022), A construction and comprehensive analysis of the immune-related core ceRNA network and infiltrating immune cells in peripheral arterial occlusive disease. *Front. Genet.* 13:951537. doi: 10.3389/fgene.2022.951537

A construction and comprehensive analysis of the immune-related core ceRNA network and infiltrating immune cells in peripheral arterial occlusive disease

Zhiyong Chen[1†], Jiahui Xu[2†], Binshan Zha[1], Jun Li[1], Yongxiang Li[3*] and Huan Ouyang[1*]

[1]Department of Vascular and Thyroid Surgery, Department of General Surgery, First Affiliated Hospital of Anhui Medical University, Hefei, China, [2]Department of General Medicine, First Affiliated Hospital of Anhui Medical University, Hefei, China, [3]Department of General Surgery, First Affiliated Hospital of Anhui Medical University, Hefei, China

Background: Peripheral arterial occlusive disease (PAOD) is a peripheral artery disorder that increases with age and often leads to an elevated risk of cardiovascular events. The purposes of this study were to explore the underlying competing endogenous RNA (ceRNA)-related mechanism of PAOD and identify the corresponding immune cell infiltration patterns.

Methods: An available gene expression profile (GSE57691 datasets) was downloaded from the GEO database. Differentially expressed (DE) mRNAs and lncRNAs were screened between 9 PAOD and 10 control samples. Then, the lncRNA-miRNA-mRNA ceRNA network was constructed on the basis of the interactions generated from the miRcode, TargetScan, miRDB, and miRTarBase databases. The functional enrichment and protein–protein interaction analyses of mRNAs in the ceRNA network were performed. Immune-related core mRNAs were screened out through the Venn method. The compositional patterns of the 22 types of immune cell fraction in PAOD were estimated through the CIBERSORT algorithm. The final ceRNA network and immune infiltration were validated using clinical tissue samples. Finally, the correlation between immune cells and mRNAs in the final ceRNA network was analyzed.

Results: Totally, 67 DE_lncRNAs and 1197 DE_mRNAs were identified, of which 130 DE_mRNAs (91 downregulated and 39 upregulated) were lncRNA-related. The gene ontology enrichment analysis showed that those down- and upregulated genes were involved in dephosphorylation and regulation of translation, respectively. The final immune-related core ceRNA network included one lncRNA (*LINC00221*), two miRNAs (*miR-17-5p* and *miR-20b-5p*), and one mRNA (*CREB1*). Meanwhile, we found that monocytes and M1 macrophages were the main immune cell subpopulations in PAOD. After verification, these predictions were consistent with experimental results. Moreover, *CREB1* was positively correlated with naive B cells (R = 0.55, p = 0.035) and

monocytes (R = 0.52, p = 0.049) and negatively correlated with M1 macrophages (R = −0.72, p = 0.004), resting mast cells (R = −0.66, p = 0.009), memory B cells (R = −0.55, p = 0.035), and plasma cells (R = −0.52, p = 0.047).

Conclusion: In general, we proposed that the immune-related core ceRNA network (*LINC00221*, *miR-17-5p*, *miR-20b-5p*, and *CREB1*) and infiltrating immune cells (monocytes and M1 macrophages) could help further explore the molecular mechanisms of PAOD.

KEYWORDS

peripheral arterial occlusive disease, PAOD, ceRNA, immune cell infiltration, atherosclerosis

Introduction

Peripheral arterial occlusive disease (PAOD) is an atherosclerotic condition involving non-cardiac and non-cerebral arteries. Nowadays, it has developed into a widespread disease with more than 200 million people affected worldwide and has become the third most common cause of death from cardiovascular disease (Fowkes et al., 2013). The importance of it is growing by virtue of its increasing incidence. Patients with PAOD often suffer from chronic limb ischemia that results in intermittent claudication, resting pain, disability, and even death. As PAOD is one common manifestation of systemic atherosclerosis, it is essential to study peripheral atherosclerosis for exploring the potential pathogenesis and progression of PAOD and effective therapeutic targets.

In recent years, more and more studies have shown that atherosclerosis, as a chronic inflammatory disease, is significantly associated with the infiltration of immune cells such as neutrophils, macrophages, T cells, and B cells into the inner layer of the vessel wall (Hansson and Hermansson, 2011; Baptista et al., 2018). In atherosclerosis, hypercholesterolemia leads to the accumulation of plasma low-density lipoprotein (LDL) in the artery wall, which stimulates and recruits monocytes and elicits local inflammation. Then, monocytes are infiltrated to differentiate locally into macrophages, and the lipid metabolic disorders and efferocytosis of macrophages are reduced, leading to irreversible inflammation (Tabas and Bornfeldt, 2016). Macrophages polarized to M1 and M2 exert pro-inflammatory and anti-inflammatory effects, respectively (Moore and Tabas, 2011). T cells account for about 40% of the total number of immune cells in human atherosclerotic lesions. Among them, regulatory T (Treg) cells produce transforming growth factor β, which inhibits the proliferation of T-helper type 1 (Th1) and T-helper type 17 (Th17) cells (Fantini et al., 2006). Th1 cells and natural killer (NK) cells secrete pro-inflammatory factors, which destroy collagen fibers and promote the transformation of atherosclerotic plaques to vulnerable phenotypes (Konkel et al., 2017). Th17 cells are a subtype of T cells, which can promote the formation of thick collagen fibers and contribute to the stability of plaques (Gisterå et al., 2013). Dendritic cells (DCs) can make the innate and adaptive immune responses to act as important modulators in atherosclerosis (Cybulsky et al., 2016). Several B-lymphocyte subsets contribute to the inflammatory process of atherosclerosis through cellular and humoral responses (Tsiantoulas et al., 2014). However, in PAOD, the landscape of immune cell infiltration has not been fully elucidated. Moreover, the relationship between immune-related genes and immune cells in PAOD is largely unknown.

Noncoding RNAs (ncRNAs) regulate gene expression at transcriptional and posttranscriptional levels without coding proteins in the transcriptome, including microRNAs (miRNAs) and long noncoding RNAs (lncRNAs). miRNAs are a class of highly conserved single-stranded noncoding small RNAs, which contain approximately 19–25 nucleotides and have post-transcriptional regulatory activity. lncRNAs are defined as a type of ncRNAs that are longer than 200 nucleotides in length, with multilevel regulatory functions in gene expression, such as transcription, translation, and epigenetics. Accumulating evidence has shown that functional ncRNAs play an important role in the pathogenesis and progression of many diseases, such as cancer, digestive system diseases, and cardiovascular diseases. In recent years, a competing endogenous RNA (ceRNA) network hypothesis has been proposed (Salmena et al., 2011). In the ceRNA network, lncRNAs can serve as endogenous molecular sponges for miRNAs to regulate the expression of messenger RNAs (mRNAs) indirectly. In this way, the function of ncRNAs can be linked to the function of mRNAs that encode proteins. Given their complexity, the dysregulation of the lncRNA-miRNA-mRNA network is closely related to the pathogenesis and progression of many human diseases, such as cardiovascular diseases. For instance, Ye et al. (2019) found that lncRNA *MIAT* upregulates CD47 expression by sponging *miR-149-5p* and inhibits efferocytosis in advanced atherosclerosis. Yang et al. (2021) discovered that the lncRNA *XIST* serves as a ceRNA and promotes atherosclerosis by increasing *miR-599*-mediated expression of *TLR4*. Nevertheless, few data-based studies have been conducted to analyze the relationship between the immune-related ceRNA regulatory network and infiltrating immune cells in PAOD.

In this study, we compared differentially expressed (DE) mRNAs and lncRNAs between 9 PAOD and 10 control samples downloaded from the Gene Expression Omnibus (GEO) database. Then, the target miRNAs of DE_lncRNAs and DE_mRNAs of target miRNAs were predicted. Subsequently, protein–protein interaction (PPI) analysis among the predicted DE_mRNAs was conducted, and hub DE_mRNAs were identified by the Cytoscape's cytoHubba plugin. The overlapping genes between the hub DE_mRNAs and immune-related genes were identified as core mRNAs to construct the potential immune-related core ceRNA regulatory network of PAOD. Meanwhile, the CIBERSORT method was used to analyze the different patterns of immune cell infiltration in PAOD. The immune-related core ceRNA network and immune infiltration were validated using clinical tissue samples. Finally, we investigated and visualized the correlation between the core mRNAs and infiltrating immune cells in an effort to better understand the molecular immune mechanism during the progression of PAOD.

Materials and methods

Data acquisition

In this study, the microarray dataset GSE57691 (Biros et al., 2015) that assessed the relative gene expression in human abdominal aortic aneurysm (AAA) and PAOD, and GSE137580 that studied the global miRNAs in atherosclerotic models that oxidative LDL treated human aortic endothelial cells (HAEC) were obtained from the Gene Expression Omnibus database (GEO, http://www.ncbi.nlm.nih.gov/geo) (Barrett et al., 2013). The specimens of GSE57691 were obtained from 20 patients with small AAA, 29 patients with large AAA, 9 PAOD patients, and 10 organ donors. Then, the data on PAOD and normal artery were picked out for further analysis. The test platforms of GSE57691 and GSE137580 were GPL10558 Illumina HumanHT-12 V4.0 expression beadchip and GPL24741 Agilent-070156 Human_miRNA_V21.0_ Microarray 046064 (gene name version), respectively. The GSE137580 data were used to validate the relative expression level of miRNAs in the potential ceRNA network. Additionally, immune-related genes were downloaded from the Immunology Database and Analysis Portal (ImmPort) database (http://www.immport.org/) (Bhattacharya et al., 2014).

Differential expression analysis

First, the probe sets were converted into corresponding gene symbols according to the platform profile with annotation information. If multiple probe sets correspond to the same gene, their mean value was calculated by R software (version 4.1.0). Then, based on the gene annotation information included in the ENSEMBL database (https://asia.ensembl.org/Homo_sapiens/Info/Index), the expression profile dataset was divided into lncRNA and mRNA groups. The linear models for the microarray data (LIMMA) package of R software were utilized to normalize raw data and perform DE_RNA analysis between PAOD and normal artery groups (Ritchie et al., 2015). *p*-values were adjusted by the Benjamini–Hochberg (BH) false discovery rate (FDR) method (Glickman et al., 2014). The cut-off value of DE_RNAs was set as adj. *p*-value < 0.01 and |fold change (FC)| > 1.5 (Meng et al., 2019). The heatmap of the DE_lncRNAs and the volcano plot of all RNAs were constructed for data visualization by the "pheatmap" (https://CRAN.R-project.org/package=pheatmap) and "ggplot2" packages in R software, respectively (Ginestet, 2011).

Prediction of lncRNA-miRNA-mRNA interactions

The highly conserved miRNA families of the miRcode database (http://www.mircode.org/) were applied to predict interactions between DE_lncRNAs and potential miRNAs (Jeggari et al., 2012). Subsequently, the TargetScan (http://www.targetscan.org/) (Agarwal et al., 2015), miRDB (http://www.mirdb.org/) (Chen and Wang, 2020), and miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) (Huang et al., 2020) databases were utilized to forecast miRNA–mRNA pairs. Only those genes that concurrently existed in all three databases were considered as candidate targets of mRNAs for further analysis.

Venn method

The Venn method was employed to analyze overlapping genes. Intersections between DE_lncRNA predicted mRNAs and DE_mRNAs, as well as immune-related genes and lncRNA-related DE_mRNAs (the intersections between DE_lncRNA predicted mRNAs and DE_mRNAs), were calculated using an online tool. (http://bioinformatics.psb.ugent.be/webtools/Venn/). Moreover, the tool was also used to identify the hub genes.

Functional enrichment and protein–protein interaction (PPI) analysis

First, the lncRNA-related DE_mRNAs were divided into expression upregulated and downregulated groups. Then, the "clusterProfiler" package (Yu et al., 2012) in R software was used to perform gene ontology (GO) (Gene Ontology Consortium, 2006) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010) enrichment analyses. GO enrichment analysis included three categories: cellular component (CC),

molecular function (MF), and biological process (BP). The "ggplot2" package in R software was utilized to draw the bubble charts for visualization of the results of the GO and KEGG enrichment analyses. p-value < 0.05 was considered significantly enriched when screening. Subsequently, the online Search Tool for the Retrieval of Interacting Genes (STRING, https://string-db.org/) database (Szklarczyk et al., 2019) was used to determine the relationship between the lncRNA-related DE_mRNAs. The minimum required interaction score was 0.7, and the disconnected nodes in the network were hidden. Then, Cytoscape software (version 3.8.2) was used to develop the PPI network. Furthermore, the Maximal Clique Centrality (MCC), Degree, and Maximum Neighborhood Component (MNC) algorithms in the cytoHubba plugin were used to screen out the top 10 mRNAs in the PPI network. The overlapping genes obtained by the three different aforementioned algorithms were identified as hub genes. Finally, we identified the overlapping genes between the hub genes before and the immune-related genes as immune-related core genes.

Construction of the immune-related core ceRNA network

First, the interactions between lncRNAs, miRNAs, and immune-related core mRNAs were confirmed, as described in the aforementioned item. Second, the immune-related lncRNA-miRNA-mRNA ceRNA network was developed, and the "ggalluvial" package (http://corybrunson.github.io/ggalluvial/) in R software was used to draw a Sankey diagram for data visualization. Subsequently, a correlation analysis between lncRNA and immune-related core genes in the ceRNA network was performed by the "LIMMA" package in R software. Ultimately, the relative expression levels of miRNAs in immune-related core ceRNA were validated in the GSE137580 dataset and visualized by GraphPad Prism (version 7.0). p-value < 0.05 was considered statistically significant.

Estimation of immune cell infiltration

CIBERSORT (https://cibersort.stanford.edu/) is a versatile analytical tool that uses gene expression data to quantify the cell fractions from complex tissues and has been confirmed by flow cytometry (Newman et al., 2015). To analyze the proportion of 22 infiltrating immune cells in atherosclerotic plaques of PAOD patients and normal controls, the mRNA expression data were uploaded to the CIBERSORT platform. Only samples that had a CIBERSORT algorithm output of p-value < 0.05 were filtered out, and the immune cell infiltration matrix was obtained for further analysis. Histograms and heatmaps were drawn to show the rate of immune cell infiltration in different samples. Subsequently, the Wilcoxon rank-sum test was performed to assess the differential composition of infiltrating immune cells between PAOD patients and controls. Results were visualized by the "pheatmap" and "vioplot" (https://github.com/TomKellyGenetics/vioplot) packages in R software. Furthermore, Pearson's correlation analysis was adopted to explore the correlation among 22 immune cell subtypes. A correlation heatmap was drawn by the "corrplot" package (https://github.com/taiyun/corrplot) in R software to visualize the correlation analysis results. Finally, the "ggstatsplot" package (https://github.com/IndrajeetPatil/ggstatsplot) was used to perform correlation analysis on the immune-related core DE_mRNAs and infiltrating immune cells, and the "ggplot2" package was used to visualize the results in R software.

Real-time quantitative PCR

From June 2021 to February 2022, we recruited five patients with PAOD who underwent femoral endarterectomy and five organ donors who donated normal iliac arteries in the First Affiliated Hospital of Anhui Medical University. All tissue specimens were divided into two parts and frozen in liquid nitrogen immediately when they were isolated. One part of these samples was pretreated, and total RNA was extracted with TRIzol reagent (Invitrogen Life Technologies, United States) for performing real-time quantitative polymerase chain reaction (RT-qPCR). Spectrophotometry was used to measure the purity of RNA. RNAs were reverse transcribed into complementary DNAs using the Bestar qPCR RT Kit (DBI, Germany), following the instructions of the manufacturer. Subsequently, complementary DNA was amplified by RT-qPCR using an Applied Biosystems SYBR Green mix kit (ABI, United States). GAPDH was used as an internal reference for lncRNAs and mRNAs, while U6 was used as a reference for miRNAs. Primer sequences were obtained from PrimerBank and miRprimer2 databases (Supplementary Table S1). The reactions were measured on the ABI 7900HT Real-Time PCR system (ABI, United States), and the $2^{-\Delta\Delta CT}$ method was used for analysis.

Hematoxylin and eosin (H&E) and immunofluorescence staining

The other part of those samples was demineralized after fixation in 4% paraformaldehyde and then embedded in paraffin. All samples were cut into 4-μm slices for further staining. H&E staining was performed to assess the atherosclerotic lesions. The protein expression levels of CD68 and iNOS were analyzed by immunofluorescence staining. Antibodies (i.e., CD68 and iNOS) were purchased from Proteintech (Chicago, IL, United States).

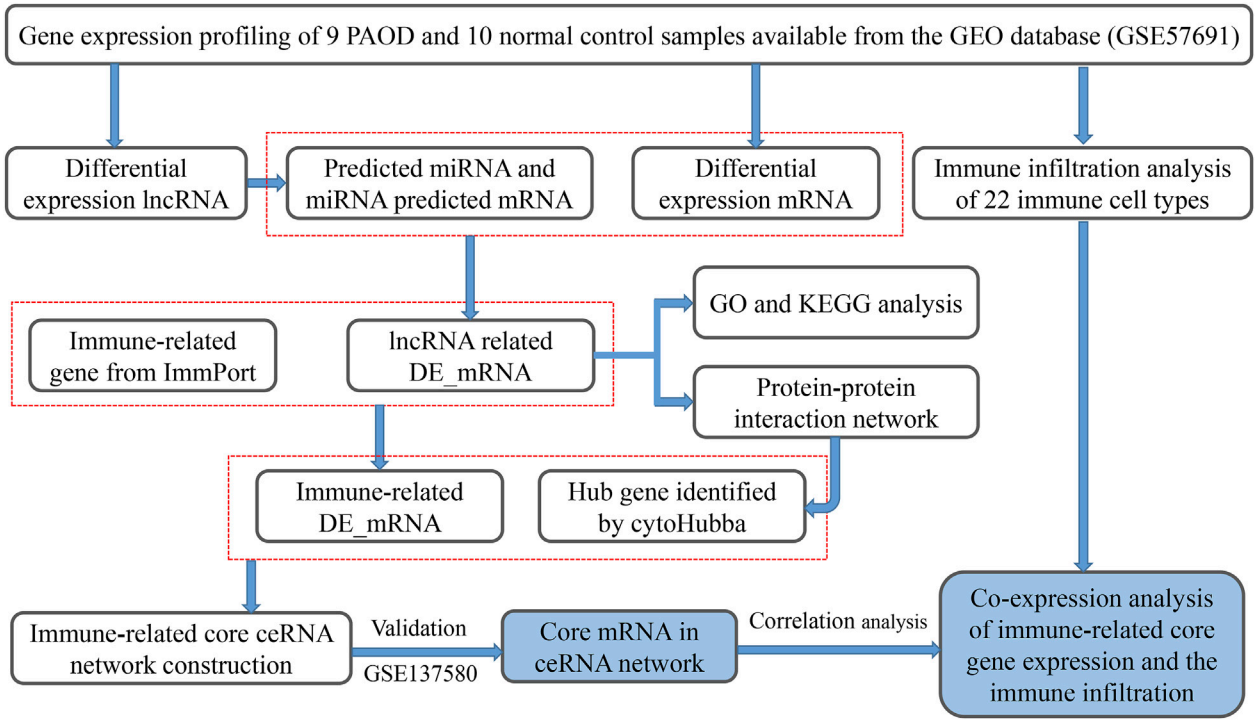


FIGURE 1
Schematic representation of our analytic process.

All procedures were conducted according to the recommendations of the manufacturer. Images were observed with a fluorescence microscope (Leica DMI6000B, Germany) and analyzed by ImageJ software.

Statistical analysis

The data are presented as the mean ± standard deviation (SD). SPSS (SPSS Inc., Chicago, IL, United States, version: 19.0) was used to conduct statistical analysis. Student's t-test was used for comparisons between two groups. A *p*-value of less than 0.05 was considered statistically significant. All experiments were performed at least three times.

Results

Identification of DE mRNAs and lncRNAs

In order to clarify the process of this research and make it easier for readers to read, a schematic representation is provided in Figure 1. Raw data were downloaded from the GSE57691 dataset in the GEO database. In total, RNA-seq data from 9 PAOD and 10 normal tissues were analyzed using criteria of |fold change (FC)| > 1.5 and adj. *p*-value < 0.01. A total of 2,142 lncRNAs and 18,219 mRNAs were re-annotated according to the platform profile with annotation information. Moreover, we identified 1264 DE RNAs, including 67 lncRNAs (27 downregulated and 40 upregulated) and 1197 mRNAs (752 down-regulated and 445 up-regulated) meeting the thresholds. Then, the corresponding volcano plot and heatmap of DE_lncRNAs are shown in Figure 2.

Functional enrichment analysis of lncRNA-related DE_mRNAs

In order to construct the ceRNA network, DE_lncRNAs were further analyzed. The miRcode database was employed to predict potential DE_lncRNA-targeted miRNAs. Then, potential miRNA–mRNA pairs were analyzed using the miRTarBase, TargetScan, and miRDB databases. A total of 34 miRNAs were identified as DE_lncRNA-predicted miRNAs, and 130 mRNAs (91 downregulated and 39 upregulated) were identified as DE_lncRNA-predicted mRNAs. Subsequently, the Venn method was used to analyze the overlapping genes between DE_mRNAs and DE_lncRNA-predicted mRNAs. As shown in Figure 3A, the intersection contains all of the DE_lncRNA-predicted mRNAs.

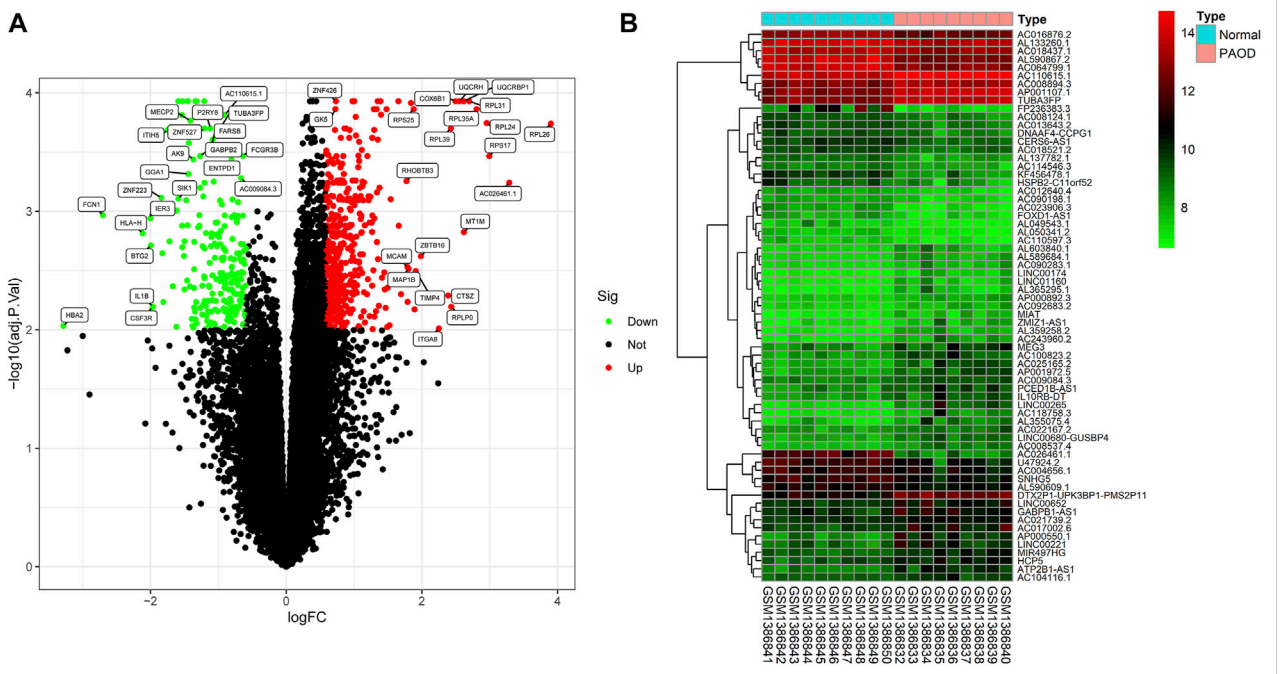


FIGURE 2
Differentially expressed (DE) mRNAs and lncRNAs in tissues between PAOD patients and normal controls. **(A)** Volcano plot of DE_mRNAs. **(B)** Heatmap of DE_lncRNAs.

These mRNAs were also called lncRNA-related DE_mRNAs. To determine the potential mechanisms of lncRNA-related DE_mRNAs, these mRNAs were divided into down- and upregulated groups for further GO and KEGG enrichment analyses (Figures 4A–D). A biological process analysis showed that mRNAs in the downregulated group were significantly enriched in dephosphorylation, regulation of autophagy, and inositol phosphate catabolic processes, while mRNAs in the upregulated group were mostly enriched in regulation of translation, positive regulation of the cellular catabolic process, and positive regulation of fat cell differentiation. A cellular component analysis showed that mRNAs in the downregulated group were significantly enriched in microtubules, while mRNAs in the upregulated group were mostly enriched in the endoplasmic reticulum–Golgi intermediate compartment. A molecular function analysis showed that mRNAs in the downregulated group were significantly enriched in ubiquitin-like protein ligase binding and phosphatase binding, while mRNAs in the upregulated group were mostly enriched in Rho GTPase binding and vinculin binding. The KEGG pathway enrichment analysis showed mRNAs in the downregulated group were significantly enriched in the PI3K-Akt signaling pathway, cellular senescence, and regulation of the actin cytoskeleton. However, mRNAs in the upregulated group were mostly enriched in human cytomegalovirus infection and the cGMP-PKG signaling pathway.

PPI network construction and hub gene identification

The PPI network of lncRNA-related DE_mRNAs containing 130 nodes and 61 edges was constructed based on the STRING online database and visualized by Cytoscape software (Figure 5). Subsequently, Cytoscape's plugin cytoHubba was used to identify the top 10 genes based on three commonly used classification methods (MCC, Degree, and MNC) (Supplementary Table S1). By overlapping these genes, seven hub genes (*YWHAH*, *PTEN*, *PIK3R1*, *FOXO3*, *YWHAE*, *CREB1*, and *FOXO1*) were consequently identified, as shown in Figure 3B.

Construction of the immune-related core ceRNA network

To construct the immune-related core ceRNA network, the Venn method was used to analyze the intersection between lncRNA-related DE_mRNAs and immune-related genes obtained from the ImmPort database. Consequently, a total of seven genes (*TNFAIP3*, *PLXND1*, *CREB1*, *PIK3R1*, *HMGB1*, *PLAU*, and *KRAS*) were identified as lncRNA-related immune DE_mRNAs (Figure 3A). Subsequently, two genes (*PIK3R1* and *CREB1*) were identified as immune-related core genes by overlapping the lncRNA-related immune DE_mRNAs and the

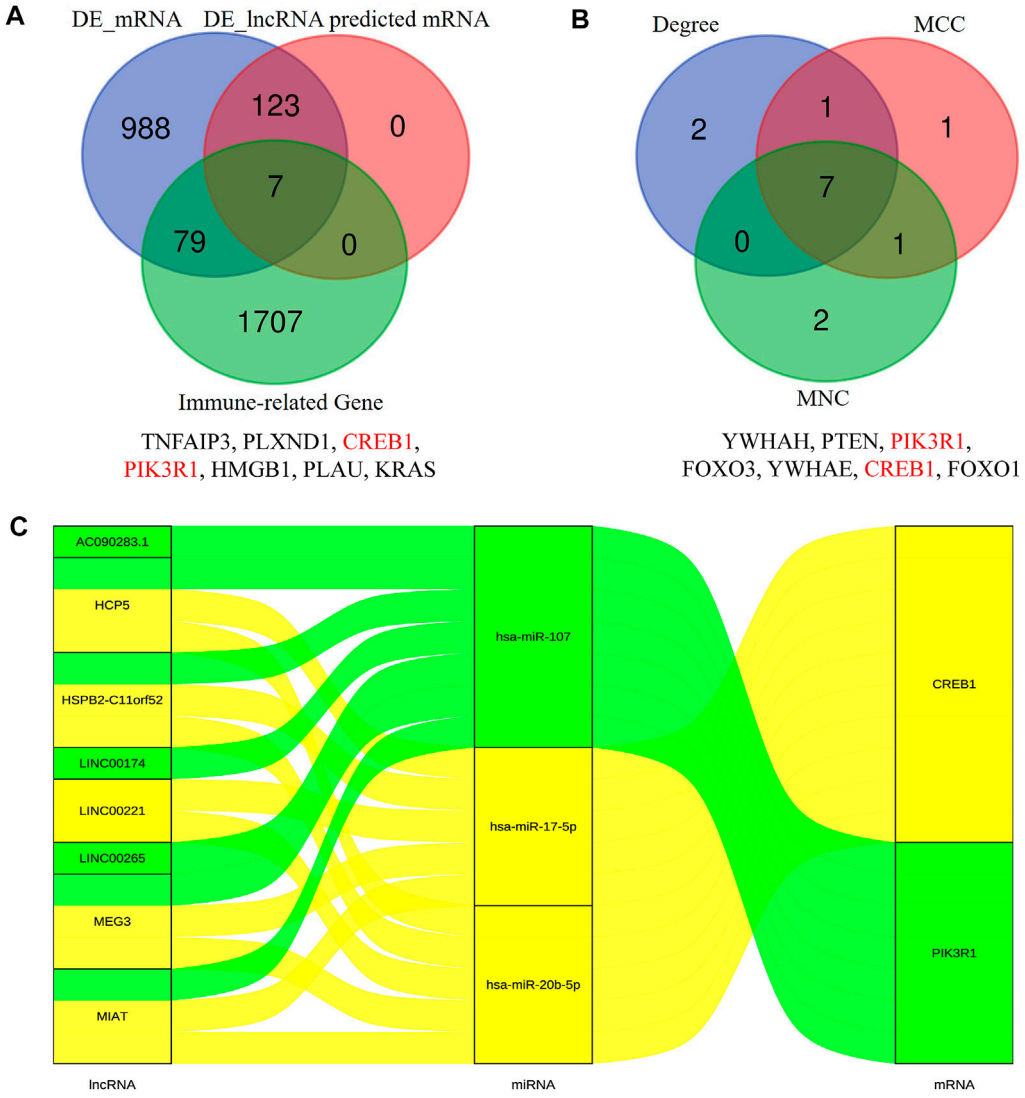


FIGURE 3
Identification of lncRNA-related DE-mRNAs **(A)** and hub genes **(B)**. The details of intersection in the Venn diagram are listed as follows. Construction of the immune-related core ceRNA network in PAOD **(C)**. MCC, maximal clique centrality; MNC, maximum neighborhood component.

seven hub genes acquired in Cytoscape software. After that, immune-related core DE_mRNAs and their paired miRNAs and lncRNAs were chosen to develop the ceRNA regulatory network. In total, the immune-related core ceRNA network contained eight lncRNAs, three miRNAs, and two mRNAs (Figure 3C). Then, the correlation between the expression of immune-related core DE_mRNAs and their paired lncRNAs was analyzed. The results illustrated that the expression of *CREB1* was positively correlated with *LINC00221* ($R = 0.861$, $p < 0.001$) and *MEG3* ($R = 0.492$, $p = 0.033$) (Figures 6A,B). Similarly, *PIK3R1* was positively correlated with that of *HSPB2-C11orf52* ($R = 0.508$, $p = 0.026$), (Figure 6C). Additionally, the relative expression levels of miRNAs in the potential immune-related core ceRNA network were validated in the GSE137580 dataset and visualized by GraphPad Prism 7. As shown in Figures 6D–F, compared with the negative control group, the relative expression levels of *miR-107*, *miR-20b-5p*, and *miR-17-5p* were all low in the atherosclerosis group, due to the expression trend of *miR-20b-5p* and *miR-17-5p* being consistent with that of prediction, while that of *miR-107* was opposite to that of prediction. Moreover, the correlation ship between the expression of *LINC00221* and *CREB1* was very close. Hence, the final potential immune-related core ceRNA network in this study contained one lncRNA (*LINC00221*), two miRNAs (*miR-20b-5p* and *miR-17-5p*), and one mRNA (*CREB1*) in Table 2.

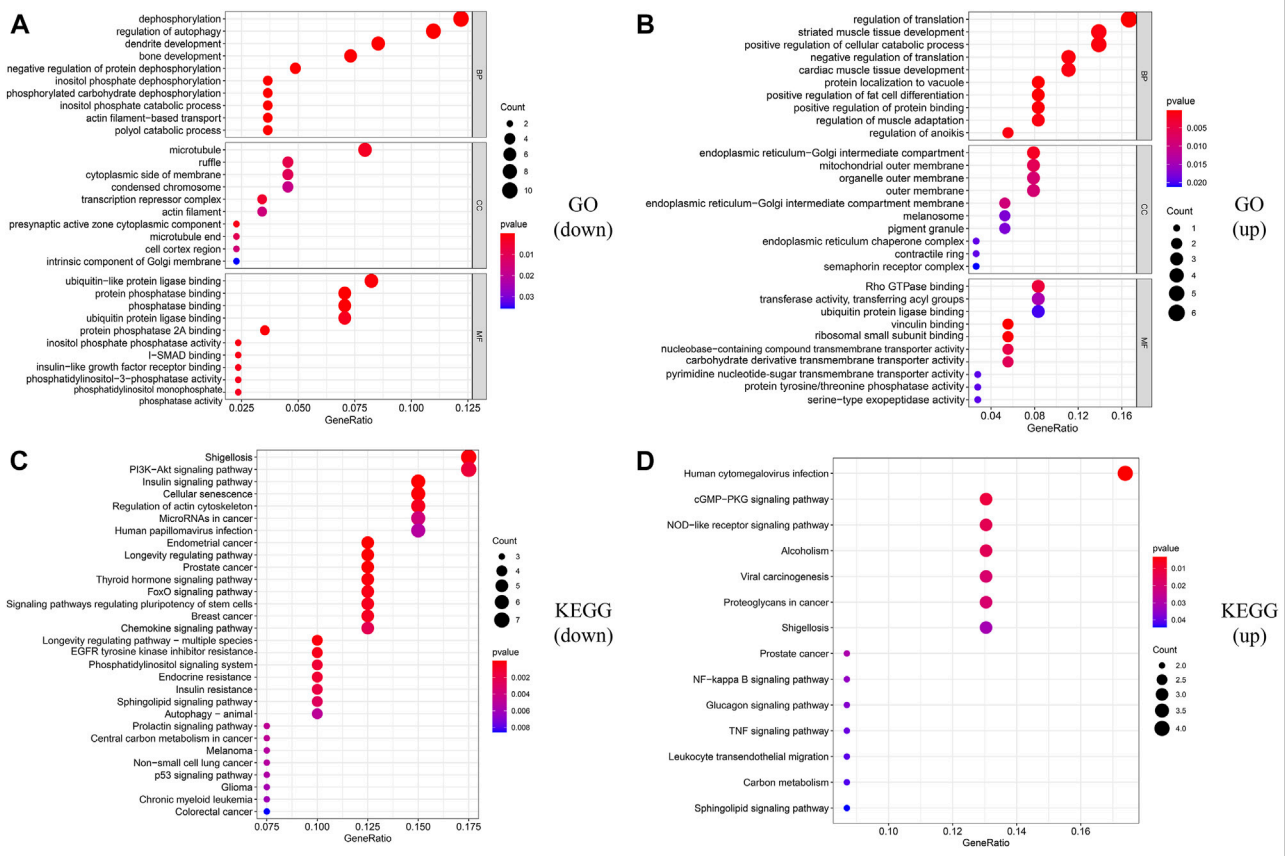


FIGURE 4
Functional enrichment analysis of lncRNA-related DE_mRNAs. **(A,B)** GO enrichment analysis of downregulated and upregulated lncRNA-related DE-mRNAs. **(C,D)** KEGG enrichment analysis of downregulated and upregulated lncRNA-related DE-mRNAs.

Composition of infiltrating immune cells

The composition of 22 infiltrating immune cells in atherosclerosis tissues of PAOD patients and normal controls was estimated using the CIBERSORT algorithm (Figures 7A,B). Since the output *p*-values of GSM1386842, GSM1386843, GSM1386849, and GSE1386836 were greater than 0.05, they were excluded for further analysis. The distribution of 22 immune cell types in each sample varied significantly, among which M2 macrophages accounted for the largest proportion. The relationships among 22 immune cells are presented in Figure 7C. Monocytes were negatively correlated with M1 macrophages (R = −0.75). Activated mast cells were positively correlated with eosinophils (R = 0.75) and activated dendritic cells (R = 0.74). Activated memory CD4 T cells were positively correlated with naive CD4 T cells (R = 0.75) and activated dendritic cells (R = 0.71). Naive CD4 T cells were positively correlated with activated dendritic cells (R = 0.98). Memory B cells were positively correlated with plasma cells (R = 0.88). Other immune cell subpopulations were weakly to moderately correlated. The violin plot of the immune cell infiltration difference showed that, compared with the normal control sample, two types of immune cells, monocytes and M1 macrophages, were differentially expressed. Monocytes were upregulated, while M1 macrophages were downregulated in atherosclerosis tissues of PAOD patients (Figure 7D).

Biological experiments

To validate the immune-related core ceRNA network in PAOD, RT-qPCR was used to detect the expression levels of the core genes. As it is shown in Figures 8A–D, compared with the normal control group, the expression of *LINC00221* and *CREB1* in the PAOD group was increased, while the expression of *miR-20b-5p* and *miR-17-5p* was decreased (all *p*-values < 0.05).

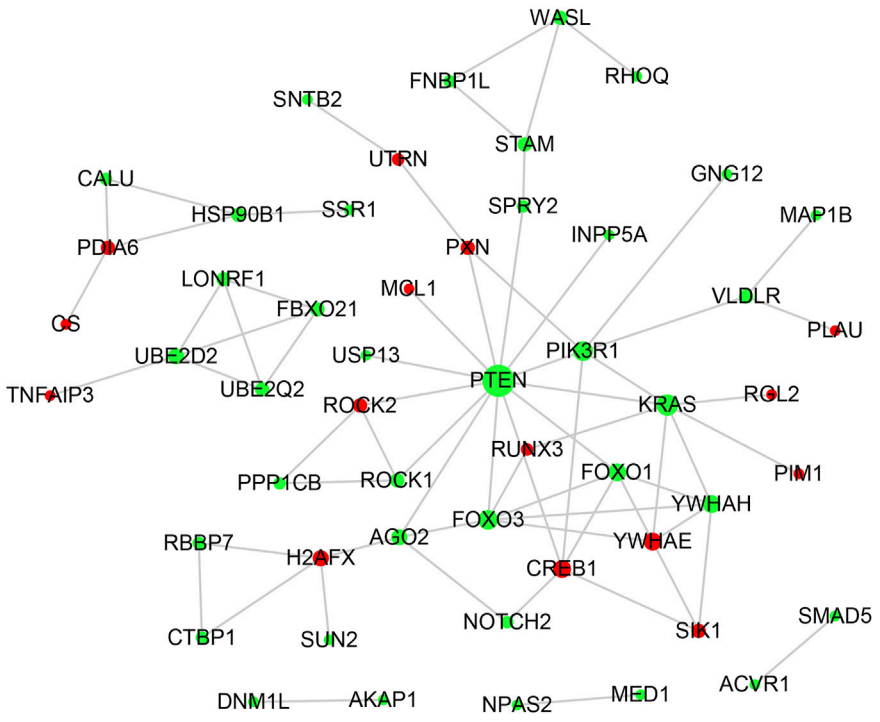


FIGURE 5
PPI network analysis. The PPI network consisting of 130 nodes and 61 edges was visualized in Cytoscape. Red and green notes symbolize the upregulated and downregulated genes. The size of each node is positively correlated with its degree value.

Intimal structure disorder, lipid infiltration, and smooth muscle cell atrophy after endarterectomy of the femoral artery are found in Figure 8E in comparison with normal arterial intima. Subsequently, immunofluorescence staining of monocyte- and macrophage-associated molecules CD68 (monocytes) and iNOS (M1) showed that monocyte and M1 macrophage infiltration were significantly increased and decreased, respectively, in the PAOD group (all p-values < 0.05). The results of positive area analysis of CD68 and iNOS are shown in Figures 8F,G. These results are consistent with the bioinformatics results we predicted.

Correlation analysis

Finally, the correlation between *CREB1* and infiltrating immune cells was estimated. In this analysis, the Wilcoxon test was adopted and significantly correlated pairs with a p-value < 0.05. As shown in Figure 9, *CREB1* was positively correlated with naive B cells ($R = 0.55$, $p = 0.035$) and monocytes ($R = 0.52$, $p = 0.049$) and negatively correlated with M1 macrophages ($R = -0.72$, $p = 0.004$), resting mast cells ($R = -0.66$, $p = 0.009$), memory B cells ($R = -0.55$, $p = 0.035$), and plasma cells ($R = -0.52$, $p = 0.047$).

Discussion

PAOD is a peripheral artery disease increasing with age and often leads to distal limb ischemia that results in reduced quality of life and death. Despite improvements in surgical, interventional, and pharmacological therapy of PAOD, the progression of atherosclerosis is still not prevented or curbed efficiently. Hence, in order to improve the treatment of PAOD, it is of vital significance to explore the underlying molecular mechanisms. For the past few years, the hypothesis of the ceRNA network has greatly raised the interest of researchers, and it makes the link between ncRNAs and mRNAs. On the basis of the ceRNA hypothesis, lncRNAs can act as miRNA sponges to regulate the expression of mRNAs. In this study, for exploring the potential pathogenesis of PAOD, we constructed the immune-related ceRNA regulatory network of PAOD based on the GSE57691 microarray dataset. After comprehensive analysis and careful verification, the final immune-related core ceRNA network, including one lncRNA (*LINC00221*), two miRNAs (*miR-17-5p* and *miR-20b-5p*), and one mRNA (*CREB1*), was identified. The interaction between *LINC00221*, *miR-17-5p*, and *miR-20b-5p* may regulate peripheral atherosclerosis *via CREB1*.

As post-transcriptional regulators, miRNAs play an important role in influencing the expression of downstream

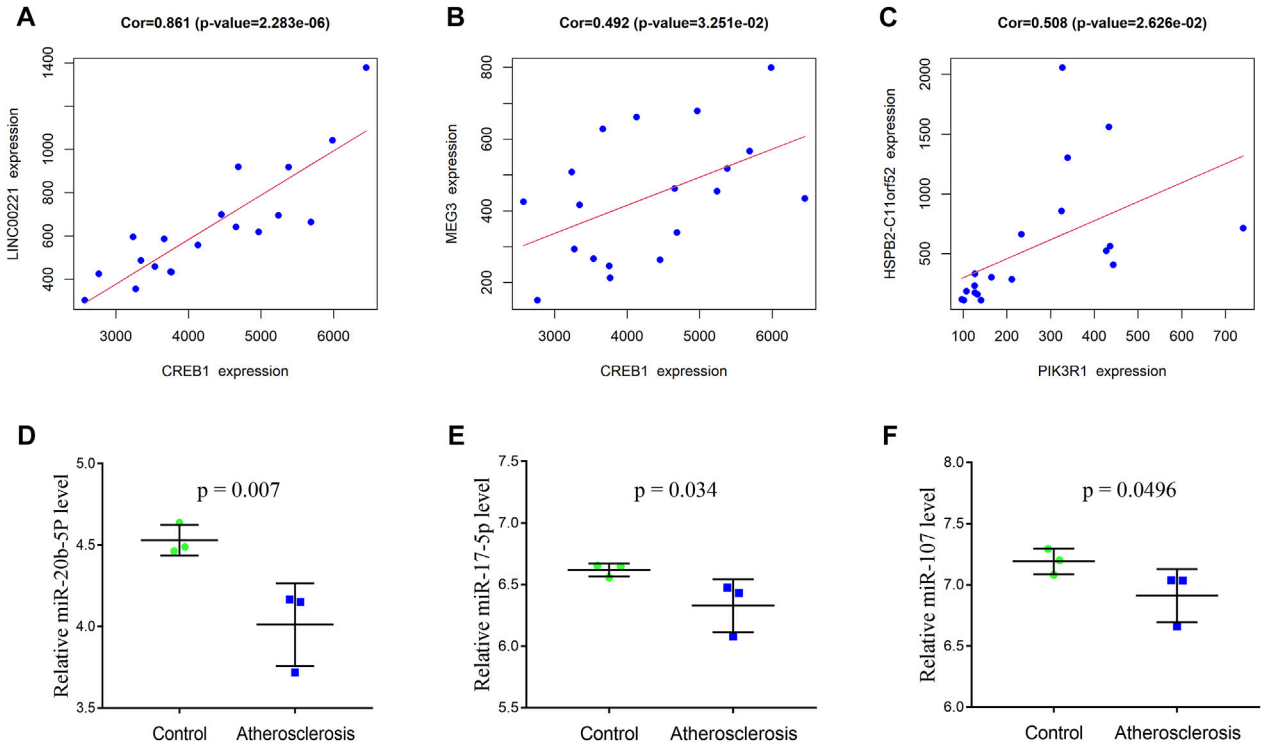


FIGURE 6
Correlation between the expression of immune-related core DE_mRNAs and their paired lncRNAs **(A–C)**. Validation of the miRNAs of the immune-related core ceRNA network in GSE137580 **(D–F)**. The results with statistically significant differences ($p < 0.05$) are shown.

target genes, which are involved in a variety of physiological and pathological processes (Valencia-Sanchez et al., 2006; Bartel, 2009). During recent years, the dysregulation of miRNA expression has been found to be associated with the pathogenesis and progression of many diseases, including atherosclerosis. In fact, several miRNAs have been found to be associated with the pathogenesis of atherosclerosis. Tan et al. (2019) demonstrated that inhibiting the expression of *miR-17-5p* can suppress inflammation and reduce lipid accumulation in atherosclerosis. An et al. (2019) reported that the expression of *miR-17-5p* is significantly decreased, and lncRNA *SNHG16* can promote proliferation and inflammatory response of macrophages through the *miR-17-5p*/NF-κB signaling pathway in patients with atherosclerosis. Shen et al. (2019) identified that the expression of *circRNA0044073* was upregulated, and the expression of *miR-107* was downregulated in atherosclerotic blood cells. Moreover, *circRNA-0044073* can suppress the levels of *miR-107 via* a sponge mechanism and increase the proliferation and invasion of cells in atherosclerosis. As for *miR-20b-5p*, the reports related to its role are mainly in various cancers, but there are few reports on atherosclerosis. More research studies are needed on the role of *miR-20b-5p* in atherosclerosis in the future.

The *CREB1* gene encodes cyclic-AMP response-binding protein 1, which is a transcription factor that is a member of the leucine zipper family of DNA-binding proteins. *CREB1* has been shown to be involved in both positive and negative regulation of atherosclerosis. On the one hand, *CREB1* upregulation is observed in the vessels isolated from normal mice compared to atherosclerotic mice (Schauer et al., 2010). On the other hand, *CREB1* can activate the pro-inflammatory cytokine IL-17 that is directly responsible for macrophage accumulation and the ensuing inflammation in the atherosclerotic plaque in mice (Kotla et al., 2013). Being equally ambiguous is the role of *CREB1* in the endothelial dynamic balance. A large body of data seems to suggest that the deletion of *CREB1* in endothelial cells may result in an enhanced inflammatory response and barrier dysfunction (Chava et al., 2012; Xiong et al., 2020). On the contrary, *CREB1* can promote leukocyte adhesion by directly binding to human umbilical vein endothelial cell *ICAM1* and activating its transcription (Hadad et al., 2011). Similarly, *CREB1* interacts with *BAF47* (*BRG1*-associated factor 47) and recruits *BAF47* to the proximal neogenin 1 promoter, leading to neogenin 1 trans-activation that contributes to endothelial dysfunction (Li et al., 2022). These discrepant roles of *CREB1* may allude to its coupling to various signaling pathways targeting either the stimulation or

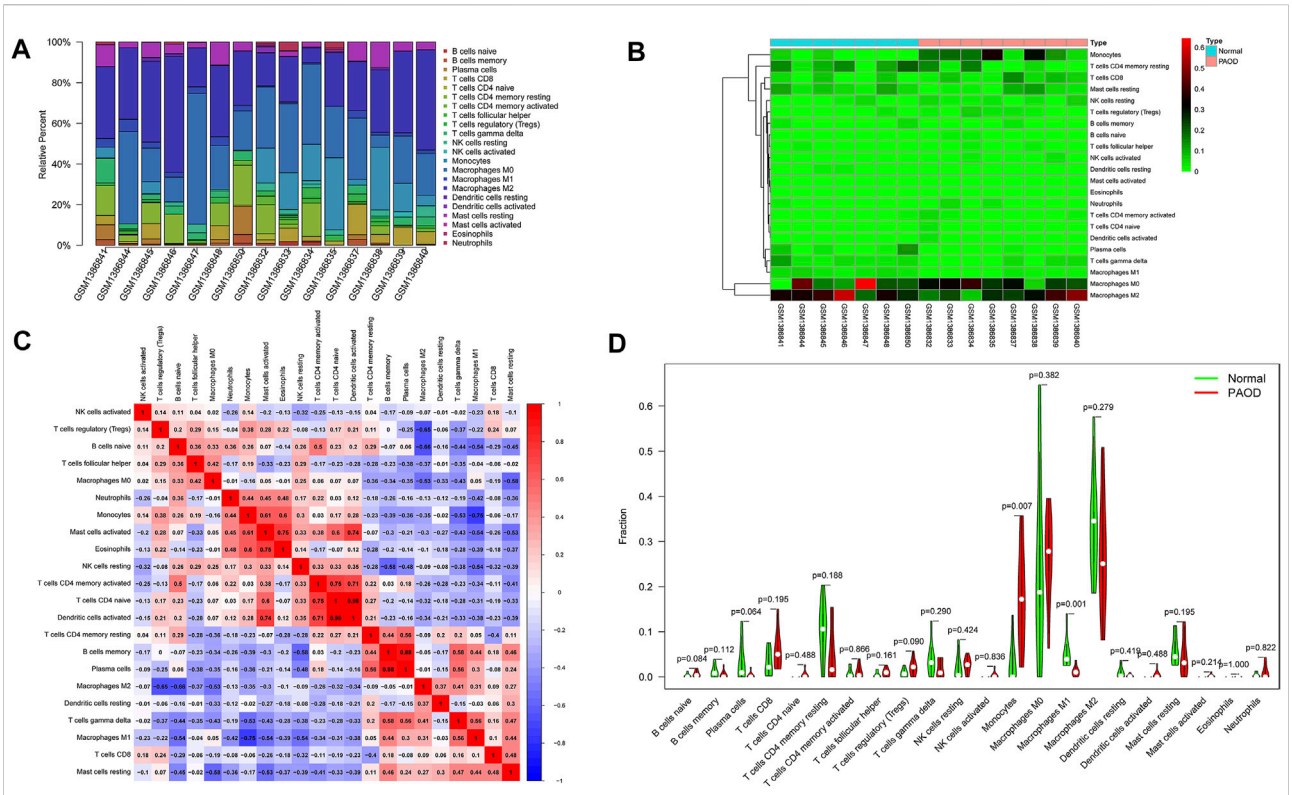


FIGURE 7
Composition of infiltrating immune cells assessed using the CIBERSORT algorithm in PAOD tissues. **(A)** Distribution of immune cell infiltration in each sample. **(B)** Heatmap of immune cell types. **(C)** Correlation among infiltrating immune cells. **(D)** Violin plot of infiltrating immune cells.

suppression of progression in atherosclerosis. Spatiotemporally controlled *CREB1* transgenic animal models should be employed in future studies for delineating the implied role of *CREB1* in atherosclerosis.

Additionally, in this study, we described the infiltrating immune cells in peripheral arterial plaques, analyzed the differences in the abundance of immune cells between the groups, and estimated the correlation between the immune-related core genes (*CREB1*) and infiltrating immune cells. The results demonstrated that monocytes and macrophages were the main immune cell subpopulations in atherosclerosis of peripheral artery tissues. In addition, atherosclerotic plaques had increased infiltration of monocytes while having decreased proportions of M1 macrophages compared to normal artery tissues. The result that M1 macrophages are found at higher levels in normal tissues appears to be in contrast with the current knowledge of proatherogenic cells. The immune cell number is a relative percentage among each group. Atherosclerotic tissues have more immune cell infiltration, so there is a relatively lower content of M1 macrophages in these tissues than in the control tissues. Owing to an increased infiltration of immune cells, the diminution in percentage may not manifest a decreased number of M1 macrophages. If the infiltration score is readjusted, the absolute percentage of M1 macrophages in the PAOD group may not be lower than that in the normal group.

Monocytes are derived from bone marrow-derived progenitor cells, and the early stage of monocytes development may be regulated by the content of cellular cholesterol in a manner that can affect atherogenesis. Recent insight suggests that the key initiating step of incipient atherogenesis in human and animal models indicates subendothelial accumulation of apolipoprotein B-containing lipoproteins (apoB-LPs) (Williams and Tabas, 1995). The pivotal early inflammatory response to accumulated apoB-LPs is the activation of surficial endothelial cells, which leads to the recruitment of circulating monocytes (Mestas and Ley, 2008). Activated endothelial cells secrete chemokines and interact with cognate chemokine receptors on monocytes in a manner that promotes monocytes' migration into the intima, where they differentiate into macrophages and phagocytize lipoproteins, leading to foam cell formation. Importantly, atherogenesis can be prevented or retarded in mouse models of atherosclerosis through preventing monocyte infiltration by blocking chemokines or their receptors (Mestas and Ley, 2008). The functions of macrophages within plaques are shaped largely

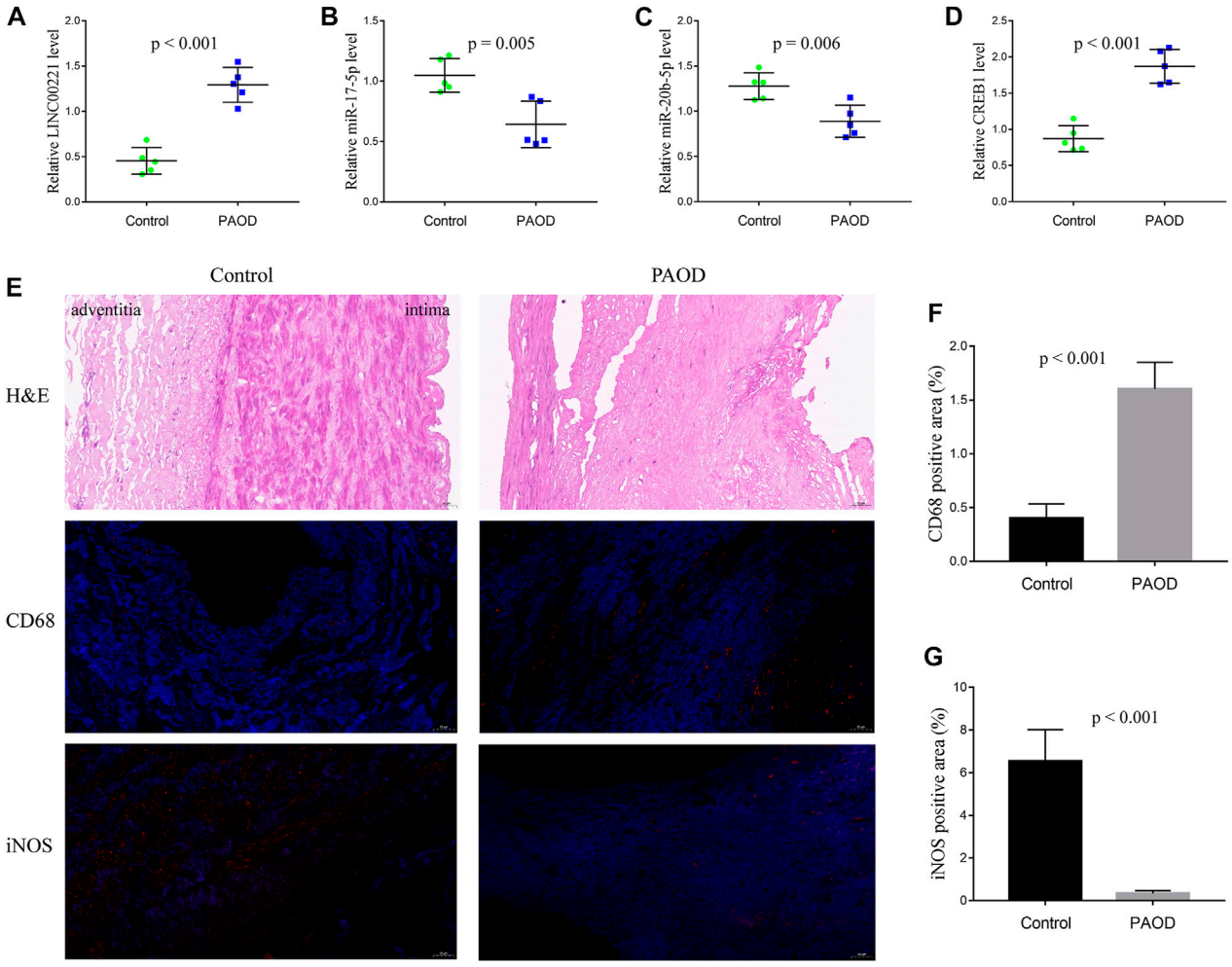


FIGURE 8
RT-qPCR validation of the immune-related core ceRNA network and evaluation of monocytes and M1 macrophages in tissue samples. Quantification of the relative expression levels of *LINC00221* **(A)**, *miR-17-5p* **(B)**, *miR-20b-5p* **(C)**, and *CREB1* **(D)** using RT-qPCR. **(E)** Representative pictures of H&E and immunofluorescence staining (magnification, ×20). Quantification of molecules in CD68 (monocyte) Fig: **(F)** and iNOS (M1 macrophage) Fig: **(G)** positive areas.

by external stimuli such as intracellular energy metabolism (Van den Bossche et al., 2017), gut microbiota metabolites (Wang et al., 2011), and genetic and epigenetic factors including ncRNAs (Erbilgin et al., 2013; Amit et al., 2016). Traditionally, macrophages are classified into pro-inflammatory M1 macrophages (activated by lipopolysaccharide and interferon-γ) and anti-inflammatory M2 macrophages (induced by interleukin-4 and interleukin-13) (Johnson and Newby, 2009; Rath et al., 2014). In general, M1 macrophages perform processes that promote atherosclerosis progression, whereas M2 macrophages carry out functions that can restrain plaque progression and facilitate plaque regression (Peled and Fisher, 2014). M1 macrophages, through secreting cytokines, proteases, and other factors, increase the cellular expansion of lesions and cause changes in plaque morphology that result in plaque rupture and acute lumenal thrombosis. Two key changes in plaque morphology promoted by M1 macrophages are plaque necrosis and fibrous cap thinning. A previous study has confirmed that M1 macrophages can secrete matrix metalloproteinases (MMPs), such as *MMP9* and *MMP2*, which may contribute to plaque rupture, and another study shows that MMPs co-localize with M1 macrophages in advanced plaques (Huang et al., 2012). In this study, we found that the expression of the *CREB1* gene was positively correlated with monocytes ($R = 0.52$, $p = 0.049$) and negatively correlated with M1 macrophages ($R = -0.72$, $p = 0.004$). Potentially, the crosstalk between the ceRNA network and the infiltration of immune cells plays a crucial part in regulating

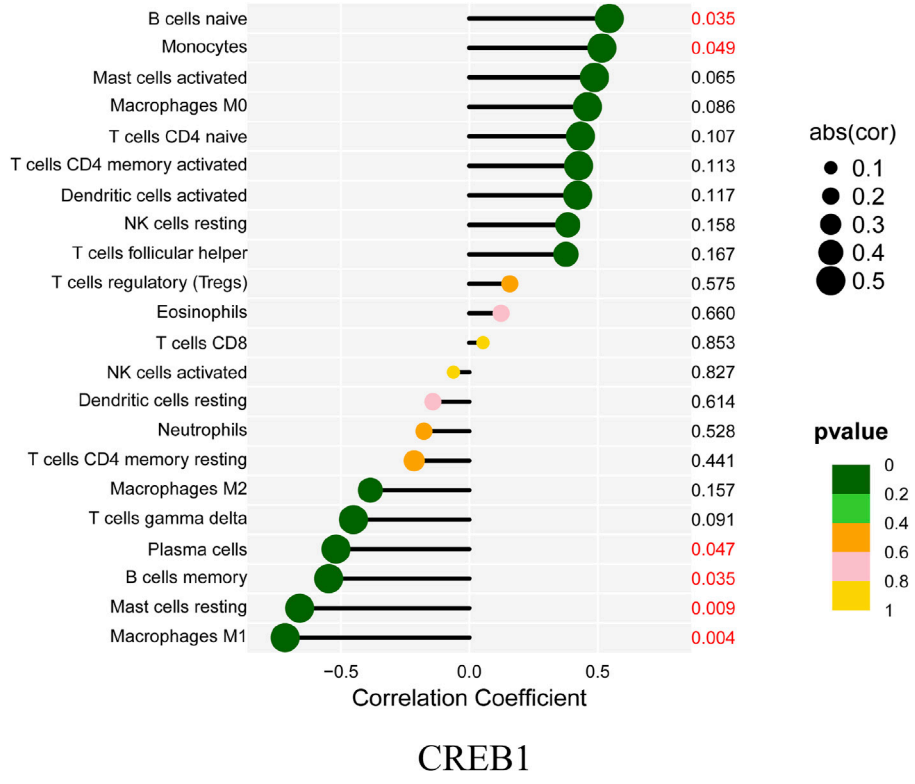


FIGURE 9
Correlation between CREB1 and infiltrating immune cells. The size of each dot is positively correlated with the strength of the correlation between genes and immune cells.

atherosclerosis progression. Further research studies are required to clarify the complex interactions between these genes and immune cells.

However, this article has some limitations. First, only one dataset (GSE57691) that is from the west and lacks ethnic diversity was used as a data source. Moreover, the analysis of immune cell infiltration was based on the CIBERSORT algorithm, the immune cell types of which were not comprehensive. Finally, the underlying regulatory mechanisms of the ceRNA network and immune cells were not elucidated clear enough, and more functional biological experiments with larger sample sizes are needed to further verify this in the future.

Conclusion

Taken together, in this study, the immune-related ceRNA network, by the composition of one lncRNA (*LINC00221*), two miRNAs (*miR-17-5p* and *miR-20b-5p*), and one mRNA (*CREB1*), was first constructed in PAOD. Afterward, the immune cell infiltration analysis was performed to estimate the abundance and differences of different immune cells. The final results show that monocytes and M1 macrophages were considered to be important immune cells associated with PAOD formation. Moreover, the expression of the *CREB1* gene was positively correlated with monocytes ($R = 0.52$, $p = 0.049$) and negatively correlated with M1 macrophages ($R = -0.72$, $p = 0.004$). These findings provide new insights into the pathogenesis and progression of PAOD and novel potential therapeutic targets. Perhaps in the future, new drugs can be developed for these novel potential therapeutic targets to delay the progression of PAOD and improve the long-term patency rate of vascular lumen in PAOD patients undergoing surgery or interventional therapy.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Medical Ethics Committee of Anhui Medical University (No. 20210015).

Author contributions

HO and YL contributed to the conception and design of the study. ZC and JX performed data analysis for the study. BZ contributed to software analysis. JL conducted the biological experiments. ZC and JX prepared the original manuscript. HO and YL reviewed and edited the manuscript. All authors contributed to the manuscript revision, read, and approved the final version.

Funding

This research was supported by the Natural Science Foundation of Anhui Province of China (Grant No. 2108085QH308) and the National Natural Science Foundation of China (Grant No. 81874063).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.951537/full#supplementary-material

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. doi:10.7554/eLife.05005

Amit, I., Winter, D. R., and Jung, S. (2016). The role of the local environment and epigenetics in shaping macrophage identity and their effect on tissue homeostasis. *Nat. Immunol.* 17, 18–25. doi:10.1038/ni.3325

An, J. H., Chen, Z. Y., Ma, Q. L., Wang, H. J., Zhang, J. Q., and Shi, F. W. (2019). LncRNA SNHG16 promoted proliferation and inflammatory response of macrophages through miR-17-5p/NF-κB signaling pathway in patients with atherosclerosis. *Eur. Rev. Med. Pharmacol. Sci.* 23, 8665–8677. doi:10.26355/eurrev_201910_19184

Baptista, D., Mach, F., and Brandt, K. J. (2018). Follicular regulatory T cell in atherosclerosis. *J. Leukoc. Biol.* 104, 925–930. doi:10.1002/JLB.MR1117-469R

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI geo: Archive for functional genomics data sets-update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell* 136, 215–233. doi:10.1016/j.cell.2009.01.002

Bhattacharya, S., Andorf, S., Gomes, L., Dunn, P., Schaefer, H., Pontius, J., et al. (2014). ImmPort: Disseminating data to the public for the future of immunology. *Immunol. Res.* 58, 234–239. doi:10.1007/s12026-014-8516-1

Biros, E., Gäbel, G., Moran, C. S., Schreurs, C., Lindeman, J. H., Walker, P. J., et al. (2015). Differential gene expression in human abdominal aortic aneurysm and aortic occlusive disease. *Oncotarget* 6, 12984–12996. doi:10.18632/oncotarget.3848

Chava, K. R., Tauseef, M., Sharma, T., and Mehta, D. (2012). Cyclic AMP response element-binding protein prevents endothelial permeability increase through transcriptional controlling p190RhoGAP expression. *Blood* 119, 308–319. doi:10.1182/blood-2011-02-339473

Chen, Y., and Wang, X. (2020). miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi:10.1093/nar/gkz757

Cybulsky, M. I., Cheong, C., and Robbins, C. S. (2016). Macrophages and dendritic cells: Partners in atherogenesis. *Circ. Res.* 118, 637–652. doi:10.1161/CIRCRESAHA.115.306542

Erbilgin, A., Civelek, M., Romanoski, C. E., Pan, C., Hagopian, R., Berliner, J. A., et al. (2013). Identification of CAD candidate genes in GWAS loci and their expression in vascular cells. *J. Lipid Res.* 54, 1894–1905. doi:10.1194/jlr.M037085

Fantini, M. C., Becker, C., Tubbe, I., Nikolaev, A., Lehr, H. A., Galle, P., et al. (2006). Transforming growth factor beta induced FoxP3+ regulatory T cells suppress Th1 mediated experimental colitis. *Gut* 55, 671–680. doi:10.1136/gut.2005.072801

Fowkes, F. G., Rudan, D., Rudan, I., Aboyans, V., Denenberg, J. O., McDermott, M. M., et al. (2013). Comparison of global estimates of prevalence and risk factors for peripheral artery disease in 2000 and 2010: A systematic review and analysis. *Lancet* 382, 1329–1340. doi:10.1016/S0140-6736(13)61249-0

Gene Ontology Consortium (2006). The gene ontology (GO) project in 2006. *Nucleic Acids Res.* 34, D322–D326. doi:10.1093/nar/gkj021

Ginestet, C. (2011). ggplot2: Elegant graphics for data analysis. *J. R. Stat. Soc. Ser. A* 174, 245–246. doi:10.1111/j.1467-985x.2010.00676_9.x

Gisterå, A., Robertson, A. K., Andersson, J., Ketelhuth, D. F., Ovchinnikova, O., Nilsson, S. K., et al. (2013). Transforming growth factor-β signaling in T cells promotes stabilization of atherosclerotic plaques through an interleukin-17-dependent pathway. *Sci. Transl. Med.* 5, 196ra100. doi:10.1126/scitranslmed.3006133

Glickman, M. E., Rao, S. R., and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J. Clin. Epidemiol.* 67, 850–857. doi:10.1016/j.jclinepi.2014.03.012

Hadad, N., Tuval, L., Elgazar-Carmom, V., Levy, R., and Levy, R. (2011). Endothelial ICAM-1 protein induction is regulated by cytosolic phospholipase A2α via both NF-κB and CREB transcription factors. *J. Immunol.* 186, 1816–1827. doi:10.4049/jimmunol.1000193

Hansson, G. K., and Hermansson, A. (2011). The immune system in atherosclerosis. *Nat. Immunol.* 12, 204–212. doi:10.1038/ni.2001

Huang, W. C., Sala-Newby, G. B., Susana, A., Johnson, J. L., and Newby, A. C. (2012). Classical macrophage activation up-regulates several matrix metalloproteinases through mitogen activated protein kinases and nuclear factor-κB. *PLoS ONE* 7, e42507. doi:10.1371/journal.pone.0042507

Huang, H. Y., Lin, Y. C., Li, J., Huang, K. Y., Shrestha, S., Hong, H. C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* 48, D148–D154. doi:10.1093/nar/gkz896

Jeggari, A., Marks, D. S., and Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics* 28, 2062–2063. doi:10.1093/bioinformatics/bts344

Johnson, J. L., and Newby, A. C. (2009). Macrophage heterogeneity in atherosclerotic plaques. *Curr. Opin. Lipidol.* 20, 370–378. doi:10.1097/MOL.0b013e3283309848

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi:10.1093/nar/gkp896

Konkel, J. E., Zhang, D., Zanvit, P., Chia, C., Zangarle-Murray, T., Jin, W., et al. (2017). Transforming growth factor-β signaling in regulatory T cells controls T helper-17 cells and tissue-specific immune responses. *Immunity* 46, 660–674. doi:10.1016/j.immuni.2017.03.015

Kotla, S., Singh, N. K., Heckle, M. R., Tigyi, G. J., and Rao, G. N. (2013). The transcription factor CREB enhances interleukin-17A production and inflammation in a mouse model of atherosclerosis. *Sci. Signal.* 6, ra83. doi:10.1126/scisignal.2004214

Li, N., Liu, H., Xue, Y., Chen, J., Kong, X., and Zhang, Y. (2022). Upregulation of neogenin-1 by a CREB1-BAF47 complex in vascular endothelial cells is implicated in atherogenesis. *Front. Cell Dev. Biol.* 10, 803029. doi:10.3389/fcell.2022.803029

Meng, L. B., Shan, M. J., Qiu, Y., Qi, R., Yu, Z. M., Guo, P., et al. (2019). TPM2 as a potential predictive biomarker for atherosclerosis. *Aging (Albany NY)* 11, 6960–6982. doi:10.18632/aging.102231

Mestas, J., and Ley, K. (2008). Monocyte-endothelial cell interactions in the development of atherosclerosis. *Trends cardiovasc. Med.* 18, 228–232. doi:10.1016/j.tcm.2008.11.004

Moore, K. J., and Tabas, I. (2011). Macrophages in the pathogenesis of atherosclerosis. *Cell* 145, 341–355. doi:10.1016/j.cell.2011.04.005

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Peled, M., and Fisher, E. A. (2014). Dynamic aspects of macrophage polarization during atherosclerosis progression and regression. *Front. Immunol.* 5, 579. doi:10.3389/fimmu.2014.00579

Rath, M., Müller, I., Kropf, P., Closs, E. I., and Munder, M. (2014). Metabolism via arginase or nitric oxide synthase: Two competing arginine pathways in macrophages. *Front. Immunol.* 5, 532. doi:10.3389/fimmu.2014.00532

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

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: The rosetta stone of a hidden RNA language. *Cell* 146, 353–358. doi:10.1016/j.cell.2011.07.014

Schauer, I. E., Knaub, L. A., Lloyd, M., Watson, P. A., Gliwa, C., Lewis, K. E., et al. (2010). CREB downregulation in vascular disease: A common response to cardiovascular risk. *Arterioscler. Thromb. Vasc. Biol.* 30, 733–741. doi:10.1161/ATVBAHA.109.199133

Shen, L., Hu, Y., Lou, J., Yin, S., Wang, W., Wang, Y., et al. (2019). CircRNA-0044073 is upregulated in atherosclerosis and increases the proliferation and invasion of cells by targeting miR-107. *Mol. Med. Rep.* 19, 3923–3932. doi:10.3892/mmr.2019.10011

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi:10.1093/nar/gky1131

Tabas, I., and Bornfeldt, K. E. (2016). Macrophage phenotype and function in different stages of atherosclerosis. *Circ. Res.* 118, 653–667. doi:10.1161/CIRCRESAHA.115.306256

Tan, L., Liu, L., Jiang, Z., and Hao, X. (2019). Inhibition of microRNA-17-5p reduces the inflammation and lipid accumulation, and up-regulates ATP-binding cassette transporterA1 in atherosclerosis. *J. Pharmacol. Sci.* 139, 280–288. doi:10.1016/j.jphs.2018.11.012

Tsiantoulas, D., Diehl, C. J., Witztum, J. L., and Binder, C. J. (2014). B cells and humoral immunity in atherosclerosis. *Circ. Res.* 114, 1743–1756. doi:10.1161/CIRCRESAHA.113.301145

Valencia-Sanchez, M. A., Liu, J., Hannon, G. J., and Parker, R. (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.* 20, 515–524. doi:10.1101/gad.1399806

Van den Bossche, J., O'Neill, L. A., and Menon, D. (2017). Macrophage immunometabolism: Where are we (going). *Trends Immunol.* 38, 395–406. doi:10.1016/j.it.2017.03.001

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* 472, 57–63. doi:10.1038/nature09922

Williams, K. J., and Tabas, I. (1995). The response-to-retention hypothesis of early atherogenesis. *Arterioscler. Thromb. Vasc. Biol.* 15, 551–561. doi:10.1161/01.atv.15.5.551

Xiong, S., Hong, Z., Huang, L. S., Tsukasaki, Y., Nepal, S., Di, A., et al. (2020). IL-1β suppression of VE-cadherin transcription underlies sepsis-induced inflammatory lung injury. *J. Clin. Invest.* 130, 3684–3698. doi:10.1172/JCI136908

Yang, K., Xue, Y., and Gao, X. (2021). LncRNA XIST promotes atherosclerosis by regulating miR-599/TLR4 Axis. *Inflammation* 44, 965–973. doi:10.1007/s10753-020-01391-x

Ye, Z. M., Yang, S., Xia, Y. P., Hu, R. T., Chen, S., Li, B. W., et al. (2019). LncRNA MIAT sponges miR-149-5p to inhibit efferocytosis in advanced atherosclerosis through CD47 upregulation. *Cell Death Dis.* 10, 138. doi:10.1038/s41419-019-1409-4

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

Glossary

AAA abdominal aortic aneurysm
apoB-LPs apolipoprotein B-containing lipoproteins
BP biological process
CC cellular component
CD cluster of differentiation
ceRNA competing endogenous RNA
CREB1 cAMP-responsive element-binding protein 1
DCs dendritic cells
DE differentially expressed
FC fold change
FDR false discovery rate
FOXO1 forkhead box O1
FOXO3 forkhead box O3
GEO Gene Expression Omnibus
GO gene ontology
HAEC human aortic endothelial cell
HMGB1 high mobility group box 1
IL interleukin
IMMPORT Immunology Database and Analysis Portal
iNOS inducible nitric oxide synthase
KEGG Kyoto Encyclopedia of Genes and Genomes
KRAS KRAS proto-oncogene GTPase
LDL low-density lipoprotein
LIMMA linear models for microarray data
lncRNA long noncoding RNA
MCC maximal clique centrality
MEG3 maternally expressed 3
MF molecular function
miRNA microRNA
MMPs matrix metalloproteinases
MNC maximum neighborhood component
mRNA messenger RNA
ncRNA noncoding RNA
NK natural killer
PAOD peripheral arterial occlusive disease
PIK3R1 phosphoinositide-3-kinase regulatory subunit 1
PLAU plasminogen activator urokinase
PLXND1 plexin D1
PPI protein–protein interaction
PTEN phosphatase and tensin homolog
RNA ribonucleic acid
RT-qPCR real-time quantitative polymerase chain reaction
STRING Search Tool for the Retrieval of Interacting Genes
Th1 T-helper type 1
Th17 T-helper type 17
TNFAIP3 TNF alpha-induced protein 3
Treg regulatory T
YWHAE tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein epsilon
YWHAH tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein eta

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fgene.2022.956805

OPEN ACCESS

EDITED BY
Pingjian Ding,
Case Western Reserve University,
United States

REVIEWED BY
Hui Jiang,
The First Affiliated Hospital of Anhui University of Chinese Medicine, China
Sayed Haidar Abbas Raza,
Northwest A&F University, China

*CORRESPONDENCE
Caiping Mao,
maocaiping@suda.edu.cn

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

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 30 May 2022
ACCEPTED 23 September 2022
PUBLISHED 13 October 2022

CITATION
Luo C, Zhang J, Bo L, Wei L, Yang G, Gao S and Mao C (2022), Construction of a ceRNA-based lncRNA–mRNA network to identify functional lncRNAs in premature ovarian insufficiency. *Front. Genet.* 13:956805. doi: 10.3389/fgene.2022.956805

Construction of a ceRNA-based lncRNA–mRNA network to identify functional lncRNAs in premature ovarian insufficiency

Chao Luo[1†], Jiakai Zhang[1,2†], Le Bo[1], Lun Wei[1], Guangzhao Yang[1], Shasha Gao[1] and Caiping Mao[1*]

[1]Reproductive Medicine Center, First Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China, [2]Monash University, Caulfield East, Melbourne, VIC, Australia

Premature ovarian insufficiency, characterized by ovarian infertility and low fertility, has become a significant problem in developed countries due to its propensity for late delivery. It has been described that the vital role of lncRNA in the development and progression of POI. The aim of this work was to create a POI-based lncRNA–mRNA network (POILMN) to recognize key lncRNAs. Overall, differently expressed mRNAs (DEGs) and differently expressed lncRNAs (DELs) were achieved by using the AnnoProbe and limma R packages. POI-based lncRNA–mRNA network (POILMN) construction was carried out using the tinyarray R package and hypergeometric distribution. To identify key lncRNAs, we used CentiScaPe plug-in Cytoscape as a screening tool. In total, 244 differentially expressed lncRNAs (DELs) and 288 differentially expressed mRNAs (DEGs) were obtained in this study. Also, 177 lncRNA/mRNA pairs (including 39 lncRNAs and 86 mRNAs) were selected using the hypergeometric test. Finally, we identified four lncRNA (HCP5, NUTM2A-AS1, GABPB1-IT1, and SMIM25) intersections by topological analysis between two centralities (degree and betweenness), and we explored their subnetwork GO and KEGG pathway enrichment analysis. Here, we have provided strong evidence for a relationship with apoptosis, DNA repair damage, and energy metabolism terms and pathways in the key lncRNAs in our POI-based lncRNA–mRNA network. In addition, we evaluated the localization information of genes related to POI and found that genes were more distributed on chromosomes 15, 16, 17, and 19. However, more experiments are needed to confirm the functional significance of such predicted lncRNA/mRNA. In conclusion, our study identified four long non-coding RNA molecules that may be relevant to the progress of premature ovarian insufficiency.

KEYWORDS

Premature ovarian insufficiency, long non-coding RNA, competing endogenous RNA, bioinformatics analysis, lncRNA–mRNA network

Introduction

Premature ovarian insufficiency (POI) is a clinical condition exhibiting symptoms of ovarian hypofunction before the age of 40 (Touraine, 2020). It is characterized by menstrual disturbance (amenorrhea or oligomenorrhea) with raised gonadotropins (FSH>25 U/L) and fluctuating drop in estrogen levels, and is estimated to affect 1% of women (European Society for Human Reproduction and Embryology (ESHRE) Guideline Group on POI et al., 2016). Since the pathogenesis of the disease is still unclear, the lack of effective biomarkers and therapeutic targets poses a challenge for early diagnosis and treatment. Furthermore, China's low fertility rate has also raised a serious question about fertility preservation in women of childbearing age with POI (Liang et al., 2019). Therefore, we must first identify the genes that play an important role in the POI process.

Long non-coding RNAs (lncRNAs) are broadly classified as transcripts longer than 200 nucleotides that are 5′ capped and polyadenylated like most mRNAs, yet this class of transcripts has limited coding potential (Klattenhoff et al., 2013). Because of the development of microarray and RNA-sequencing, we can explore key molecules in disease from multiple perspectives (Raza et al., 2022, 6). Emerging evidence indicates that lncRNAs play critical roles in various biological processes, such as cellular development, differentiation, imprinting control, immune response, and chromatin modification (Rinn et al., 2007; Amaral and Mattick, 2008; Dinger et al., 2008; Ponting et al., 2009; Lee, 2012). Conclusions drawn from different microarray analyses proved that many lncRNAs are unusually expressed in human granulosa cells; this means lncRNAs may be involved in the pathogenesis and progression of POI. For example, the decreased expression of HCP5 is directly related to the apoptosis of granulosa cells and DNA damage repair (Wang et al., 2020, 5). LncRNA DDGC was able to ameliorate the etoposide-induced DNA damage and apoptosis *in vivo* (Li et al., 2021b). Additionally, PVT1 ameliorates granulosa cell apoptosis by promoting SCP4-mediated Foxo3a dephosphorylation (Wang et al., 2021, 1). Therefore, only a few lncRNAs have been further elaborated so far, while the discoveries and confirmations of the vast majority are still an enigma. The competitive endogenous RNA hypothesis is mainly composed of mRNAs and lncRNAs who both share miRNA recognition elements and can compete with each other to occupy miRNAs (Salmena et al., 2011). The ceRNA mechanism in a variety of cancers and gynecological illnesses has been reported; at the same time, they also isolated some putative ceRNA networks, such as epithelial ovarian cancer (Zhao et al., 2019), polycystic ovary syndrome (Ma et al., 2021), and implantation failure (Feng et al., 2018). With insights into the mechanism of ceRNA, in this study, we aim to design a POI-based lncRNA–mRNA network (POILMN) to label key lncRNAs and explore gene location information and functional enrichment that may also point to directions for future research.

Materials and methods

Microarray data

In order to obtain the microarray analysis, the public database Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) was searched using keywords such as "premature ovarian insufficiency," "POI," "granulosa cells," and "Homo sapiens." The GSE135697 dataset was selected for further study. The GSE135697 dataset (platform: GPL21096, Agilent-045997 Arraystar human lncRNA microarray V3) included lncRNA and mRNA expression profiles consisting of 10 POI samples and 10 control samples. The profiling construction and test of the datasets were authorized by the local research ethics committee.

Differential gene expression analysis and probe reannotation

Preprocessed data were acquired from GEO using the R package "GEOquery". After getting the expression matrix, and subsequently, according to the annotation profile recorded in the "AnnoProbe" package (version 0.1.6), probesets were annotated to filter out the duplicate and unannotated probes and then separated into two categories: protein-coding dataset and lncRNA dataset. Log-transformed intensities were quantile normalized using the "normalizeBetweenArrays" function in the "limma" package of R. Then, the limma package was used to identify differentially expressed mRNAs and lncRNAs, respectively. The *p*-value was adjusted using the Benjamini–Hochberg method. Unless stated otherwise, "differentially expressed" (DE) mRNAs and lncRNA were defined as FDR <0.05 and $\log_2|\text{fold change}| > 1$.

The location distribution of differential genes on chromosomes

There is growing evidence that POI was judged to be related to the genetically heterogeneous disorder. The chromosomal location and the starting and ending positions were assessed using the R package "AnnoProbe" (version 0.1.6). The R library "RIdeogram" was used to visualize data along the chromosomes of differentially expressed mRNAs and lncRNAs.

LncRNA–miRNA and miRNA–mRNA interaction data and construction of the POILM network

Starbase v3.0 (http://starbase.sysu.edu.cn) was used to extract lncRNA–miRNA associations from HITS-CLIP and PAR-CLIP

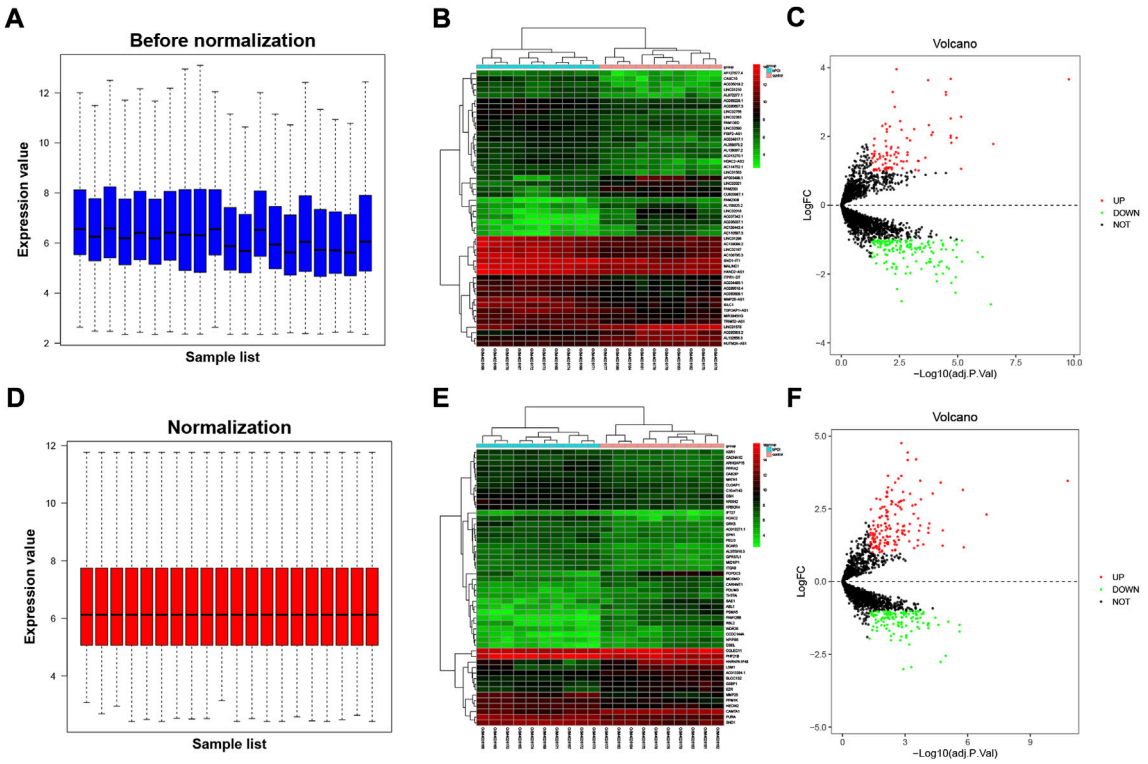


FIGURE 1
Differentially expressed lncRNAs and mRNAs from granulosa cells in patients with premature ovarian insufficiency (POI). The dataset was split into the lncRNA dataset and protein-coding dataset. Microarray data of dataset, **(A)** prior to normalization and **(B)** following normalization. **(C)** Heatmap of top 50 DELs (sorted by adjusted *p*-value); **(D)** heatmap of top 50 DEGs (sorted by adjusted *p*-value). Volcano plot showing differentially expressed lncRNAs **(E)** and mRNAs **(F)** after being screened by FDR <0.05 and log2|fold change| > 1. Red dots and green dots referred to the upregulated and downregulated genes, respectively.

experiments. Sequence-predicted miRNA–mRNA pairs were obtained from miRTarBase (https://mirtarbase.cuhk.edu.cn/). The processed relation pairs are selected as the background relation pairs for the hypergeometric distribution.

To construct the POILMN, the DELs and DEGs were substituted in the background network *via* the R package tinyarray. Then, the lncRNA–miRNA–mRNA network was filtered *via* a hypergeometric test with $p < 0.01$, and counts >3 denote the counts of the number of miRNAs shared between lncRNA and mRNA. The value of p was calculated as

$$P = 1 - \sum_{i=0}^{r-1} \frac{\binom{t}{i}\binom{m-t}{n-i}}{\binom{m}{n}}.$$

In the formula of hypergeometric distribution, m is the total miRNA number in the miRTarBase database, n represents the number of miRNAs interacting with a lncRNA, t is the number of miRNAs interacting with an mRNA, and r indicates the quantities of miRNAs united between the lncRNA/mRNA pair.

Functional and pathway enrichment analysis

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis of the DEGs sifted out in the POILMN and lncRNA subnetwork was performed and visualized using the R package clusterProfiler (V4.2.2) (Wu et al., 2021). KEGG pathways and Gene Ontology (GO) terms were considered statistically significant using $p < 0.05$ as the cut-off value. The GO enrichment analysis consists of three components molecular functions (MFs), biological processes (BPs), and cellular components (CCs).

Topological analysis and selection of key lncRNAs

To explore the central nodes of the POILMN network, we performed a topological analysis of DELs and DEGs and calculated using the CentiScaPe plug-in Cytoscape. We focused on two main topological parameters: “degree” and

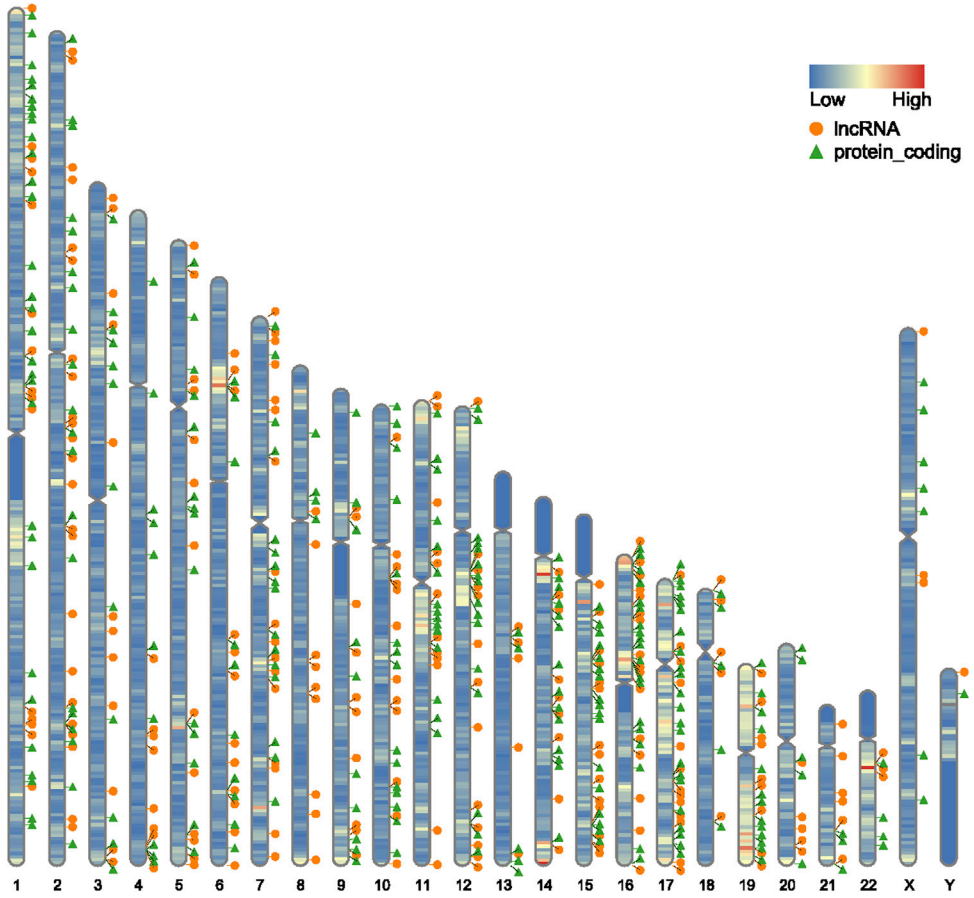


FIGURE 2
Schematic diagram of the distribution of differentially expressed mRNA and lncRNA on chromosomes, orange circles represent lncRNAs, green triangles represent protein-coding, and the color on each chromosome represents gene density.

"betweenness." Retain the topped-eight lncRNA of each parameterization and those overlapped were chosen as hub genes for the follow-up stage.

Construction of ceRNA sub-networks

Through the implementation of the previous approach using a hypergeometric test, we obtained all the key lncRNAs and their adjacent mRNA neighbors. At the same time, we also get the overlapped miRNAs that they commonly shared. Ultimately, based on ceRNA theory, we imported a triple network into Cytoscape software to visualize it.

Subcellular localization analysis

To investigate the intracellular localization of key lncRNAs in topological analysis, we used a web-based public platform lncLocator, for prediction. Target lncRNA sequences were downloaded from NCBI in nucleotide FASTA format. Bar plotting was performed using R with the ggplot2 (V3.3.5) package.

Results

Differentially expressed lncRNAs and mRNAs

Expression estimates were further normalized using quantile normalization; box plots show mean expression level differences before and after normalization in Figures 1A,B. In GSE135697, 244 DELs were identified with log2|fold change| > 1 and $p < 0.05$, including 93 upregulated and 151 downregulated, in the POI granulosa cells in the test compared to the control, as shown in the volcano plot (Figure 1E); At the same time, there were 288 DEGs, with 160 upregulated and 128 downregulated

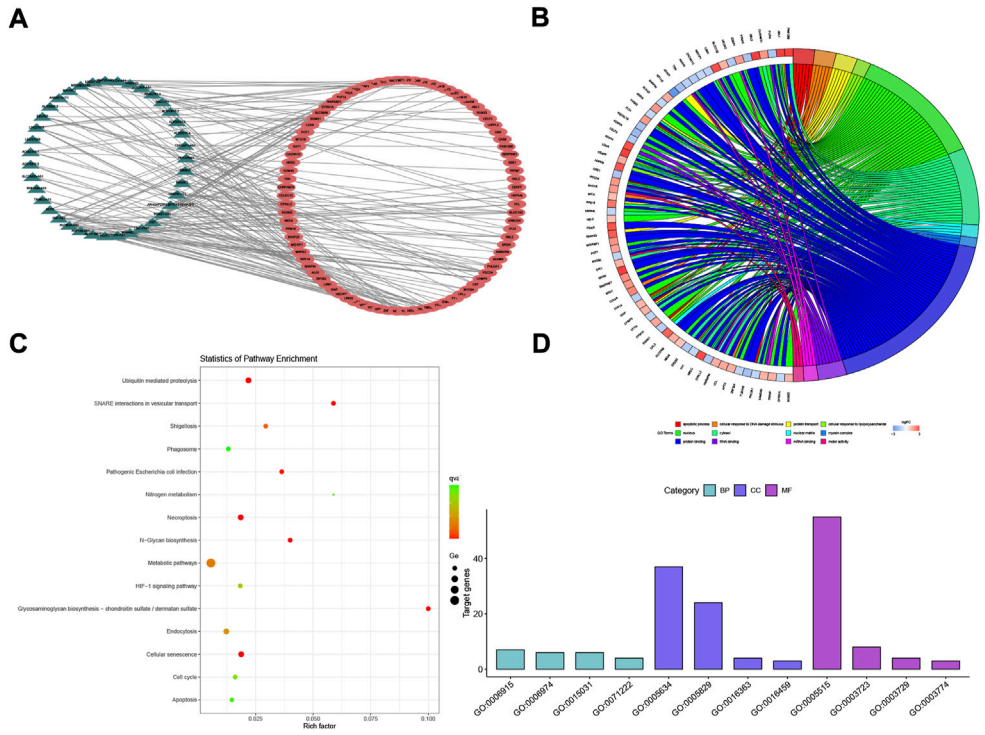


FIGURE 3
POI-based lncRNA–mRNA network (POILMN) and functional enrichment analysis of mRNAs. The dark green triangles represent lncRNA, and the red cycles represent mRNAs. There were 39 lncRNA nodes, 86 mRNA nodes, and 177 edges in the network **(A)**. Distribution of integrated DEGs in premature ovarian insufficiency for different GO-enriched functions **(B)**. KEGG pathway enrichment analysis of the integrated DEGs **(C)**. The GO enrichment bar chart of DEGs presents the number of DEGs enriched in biological processes, cellular components, and molecular functions **(D)**.

(Figure 1F). The list of 288 DEGs and 244 DELs is shown in Supplementary Table S1. Several genome-wide specific genomic features were revealed at the chromosomal level, as shown in Figure 2.

Construction of the POI-related lncRNA–mRNA network

After the stringent filtering, we ended up with 63,698 lncRNA–miRNA pairs (including 642 miRNAs and 3,788 lncRNAs) and 502,653 miRNA–mRNA interaction pairs (including 15,064 mRNA and 2,599 miRNAs). Total DELs/DEMs were matched into those two interaction pairs, and then 177 lncRNA/mRNA pairs (including 39 lncRNAs and 86 mRNAs) were selected using the hypergeometric test with $p < 0.01$ and counts >3 in Figure 3A. The lncRNA/mRNA pairs are shown in Supplementary Table S2.

Functional enrichment analysis of the differentially expressed mRNAs in the ceRNA network

To obtain a better understanding of the role of the DEGs in the lncRNA–mRNA ceRNA network, we performed GO classification and KEGG pathway analysis. Results showed that 16 GO terms ($p < 0.05$) were significantly enriched in the GO analysis. The top four GO terms in BP, MF, and CC are listed in Figure 3D. Biological process (BP) analysis showed that the associated mRNAs were “cellular response to DNA damage stimulus”. Chen et al. (Wang et al., 2020, 5) have proved that reduced expression of long non-coding RNA HCP5 in POI modulates the repair of MSH5 transcription and DNA damage interacting with YB1, leading to GC dysfunction, providing potent evidence for POI pathogenesis in the cellular response to DNA damage stimulus. On the MF dimension, the top four terms associated with gene counts were protein binding,

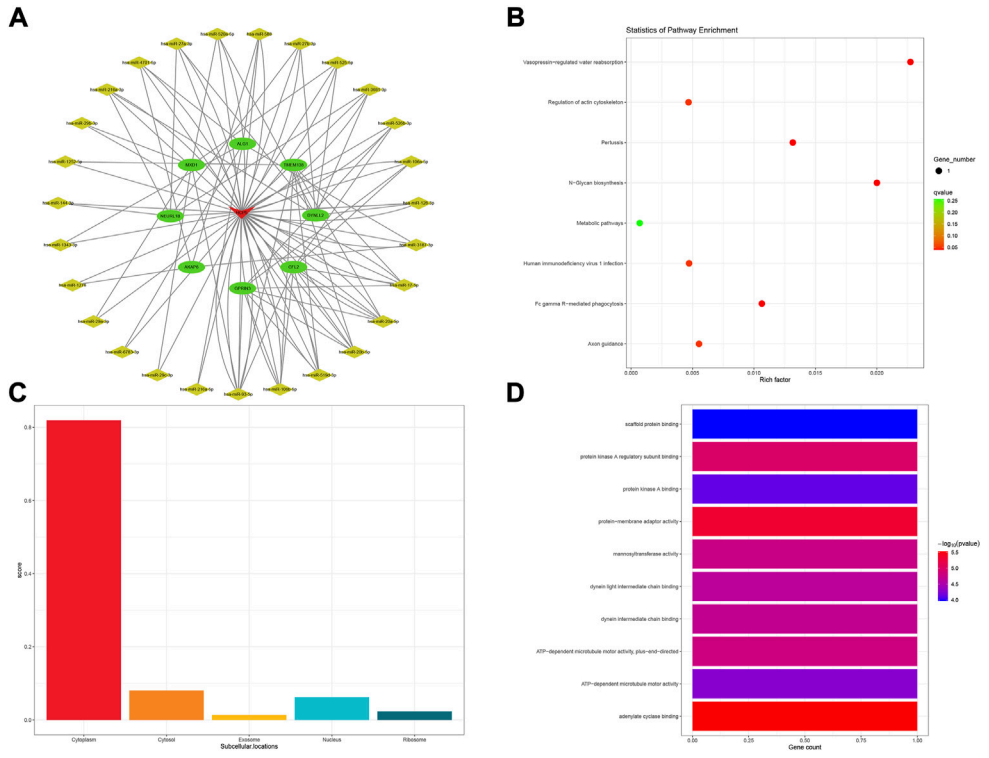


FIGURE 4
HCP5-related ceRNA subnetwork analysis. ceRNA network of HCP5 **(A)**. KEGG pathways enriched in HCP5 **(B)**. Subcellular location analysis for HCP5 **(C)**. GO biological process enrichment results for HCP5 **(D)**.

RNA binding, mRNA binding, and motor activity. The KEGG pathways showing the most significant enrichment were glycosaminoglycan biosynthesis—chondroitin sulfate/dermatan sulfate, SNARE interactions in vesicular transport, ubiquitin-mediated proteolysis, and cellular senescence, as shown in Figure 3C. Herein, these GO terms and KEGG pathways may shed new light on POI pathogenesis and prognosis.

Topological characteristics of the POI-based lncRNA–mRNA network and locations of key lncRNAs

The topological features of the POILMN, including degrees and betweenness, were chosen to forecast the biological functions of the lncRNAs in POILMN. Then, the top eight sorted genes in the POILMN with the highest value were extracted. The intersection of lncRNAs was screened at the degree and betweenness parameter, and four lncRNAs, that is, NUTM2A-AS1, HCP5, SMIM25, and GABPB1-IT1, were jointly identified by the intersection between two features.

Comprehensive rating had eight top results for HCP5, NUTM2A-AS1, SMIM25, and GABPB1-IT1, respectively. For HCP5, a total of eight mRNAs and 36 miRNAs were composed of the subnetwork of the POILMN in Figure 4A. A total of 8 KEGG pathways were significantly enriched ($p < 0.05$) in the KEGG pathways analysis, as shown in Figure 4B. Most of HCP5 is localized in the cytoplasm (score = 0.819), and only a small part of it is localized to the nucleus, ribosome, cytosol, and exosome (Figure 4C). GO classification with $p \leq 0.05$ adjusted by Benjamini–Hochberg found ten enriched GO terms for molecular function, as shown in Figure 4D.

As for the NUTM2A-AS1 subnetwork, 13 mRNA and 43 miRNA comprised the subnetwork (Figure 5A). KEGG pathways analysis showed that nine pathways were enriched in the process, as shown in Figure 5B. Lnclocator analysis revealed that NUTM2A-AS1 is mainly distributed in the cytoplasm and is being distributed in the nucleus, ribosome, cytosol, and exosome at the same time (Figure 5C). In the GO enrichment analysis for the NUTM2A-AS1 subnetwork, we got seven terms, two terms in BP, one in MF, and four in CC (Figure 5D).

The GABPB1-IT1 subnetwork was made up of 10 mRNA and 28 miRNA in all (Figure 6A). In conducting KEGG analysis, 11 KEGG pathways were significant among 14 KEGG pathways

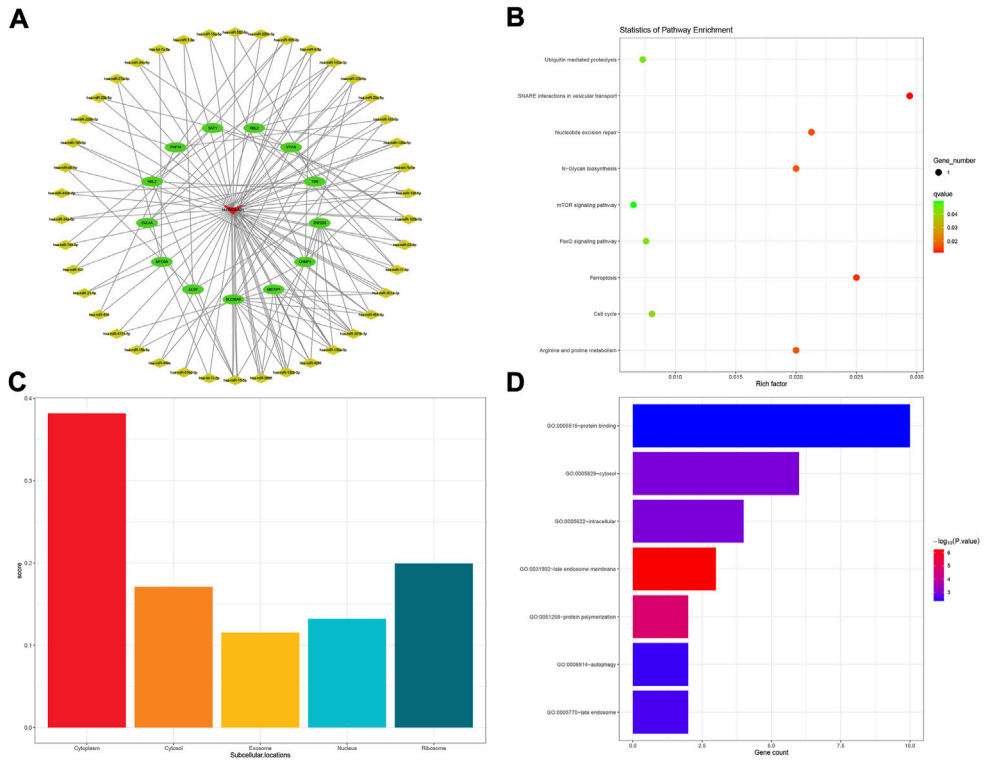


FIGURE 5
NUTM2A-AS1-related ceRNA subnetwork analysis. ceRNA network of NUTM2A-AS1 **(A)**. KEGG pathways enriched in NUTM2A-AS1 **(B)**. Subcellular location analysis for NUTM2A-AS1 **(C)**. GO biological process enrichment results for NUTM2A-AS1 **(D)**.

in total (Figure 6B). GABPB1-IT1 was mostly localized to the cytosol (score = 0.477) (Figure 6C). In contrast, GO term analysis revealed that target genes were involved in 109 GO terms ($p < 0.05$). Only the top ten GO biological process terms are shown in Figure 6D.

The SMIM25 subnetwork was composed of eight mRNA and 19 miRNA (Figure 7A). The enrichment analysis of KEGG pathways included 18 KEGG pathways; among them, the top 10 pathways are visualized in Figure 7B. SMIM25 was mainly localized to the cytoplasm (score = 0.352) and cytosol (score = 0.416), as shown in Figure 7C. Upon GO classification, 2 CC and 8 MF terms were obtained to be enriched in SMIM25 (Figure 7D).

Discussion

POIs are a common disorder of follicular development characterized by hypertrophy of the initial follicle, slowed growth of antral follicles, and selection of the dominant follicle. In our study, we sought to identify these lncRNAs by examining data from POI microchips and performing analyses to determine their potential roles in POILMN.

Chromosomal abnormalities have long been recognized as a frequent cause of POI. We restored the differentially expressed mRNAs and lncRNAs to their chromosomal locations and found that they were not only associated with autosomes but also with sex chromosomes. At the same time, it was found that there are a large number of distributions on the long arm of chromosome 15, the short arm of chromosome 16, and chromosomes 17 and 19.

The lncRNA/miRNA and mRNA/miRNA pairs were obtained from the starBase and miRTarBase databases, respectively, and the lncRNA–mRNA network was constructed by calculating the hypergeometric distribution using differentially expressed mRNAs and lncRNAs. The POI-based lncRNA–mRNA network (POILMN) was then extracted with 86 mRNA nodes, 39 lncRNA nodes, and 177 edges. Results were analyzed to determine putative functional POI biomarkers.

For the GO term enrichment analyses of DEGs in the POILMN, we observed a large proportion in the number of counts in the category of biological process (BP), such as “apoptotic process,” “cellular response to DNA damage stimulus,” and “protein transport.” Considering of its GC-related diseases, some previous studies confirmed that decreasing apoptosis of granulosa cells could improve the function of the ovary in POI mice (Huang et al., 2019, 1; Ling

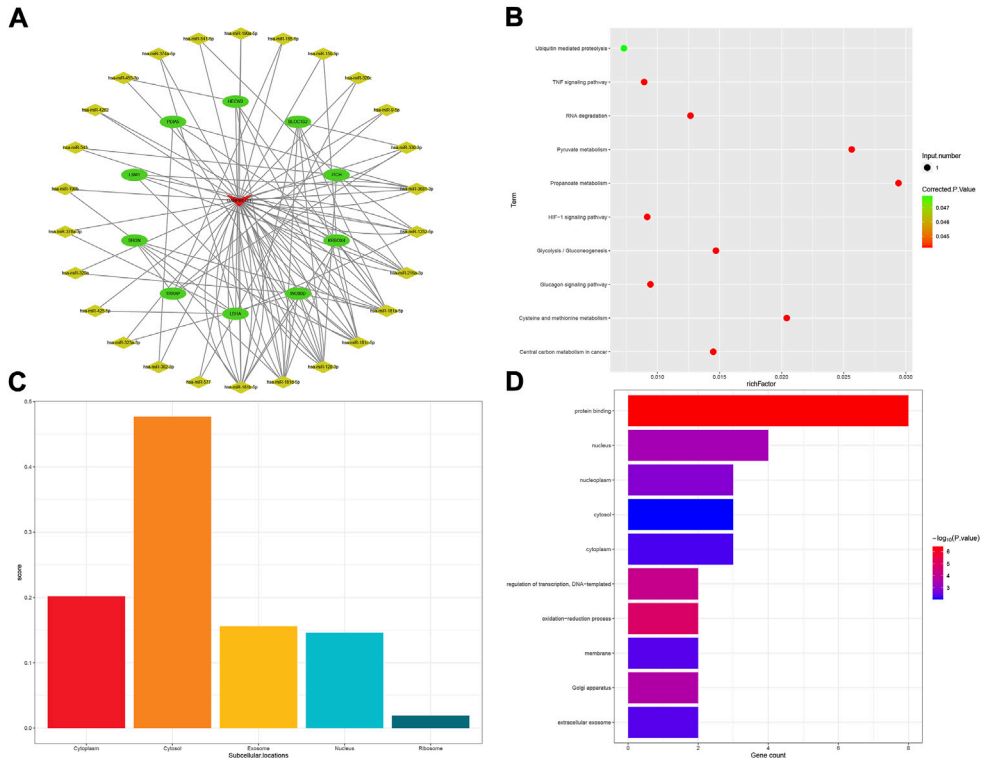


FIGURE 6
GABPB1-IT1-related ceRNA subnetwork analysis. ceRNA network of GABPB1-IT1 **(A)**. KEGG pathways enriched in GABPB1-IT1 **(B)**. Subcellular location analysis for GABPB1-IT1 **(C)**. GO biological process enrichment results for GABPB1-IT1 **(D)**.

et al., 2019; Wang et al., 2019). A study found that low expression of lncRNA HCP5 in granulosa cells from POI patients interfered with GC DNA damage healing, promoting apoptosis of GCs, and mediated the translocation of YB1 protein to the GC nucleus (Wang et al., 2020, 5). As for the molecular functions (MFs) and cellular components (CCs) categories, the primary term was "nucleus" and "protein binding."

Intriguingly, the most enriched KEGG pathway was "metabolic pathways." Aside from lower serum AMH levels and higher FSH levels, we also retrieved some metabolism-related case-control studies of POI. These association studies pointed out that serum concentrations of total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), and low-density lipoprotein cholesterol (LDL-C) were significantly higher in the POI group compared with a healthy matched control group (Ates et al., 2014; Podfigurna et al., 2018). Another cross-sectional case-control study points out women with POI were more likely to exhibit increased serum levels of TG (β, 0.155; 95% CI, 0.086, 0.223) and glucose (0.067; 0.052, 0.083), decreased levels of HDL-C (−0.087; −0.123, −0.051), LDL-C (−0.047; −0.091, −0.003) and uric acid (−0.053; −0.090, −0.015), and impaired kidney function (urea [0.070; 0.033, 0.107]; creatinine [0.277; 0.256, 0.299]; and eGFR [−0.234; −0.252, −0.216]) compared with controls after adjusting for age and BMI (Huang et al., 2021). In conclusion, a higher risk of metabolic syndrome was associated with POI. It also reminds us to adopt lifetime management of metabolic abnormalities that are needed in the early diagnosis of POI.

Several carbohydrate metabolism-related pathways were also enriched, including "glycosaminoglycan biosynthesis – chondroitin sulfate/dermatan sulfate" and "N-glycan biosynthesis." Some studies have reported that glycosaminoglycan chondroitin-4-sulfate may play a role in altering gonadotrophin-stimulated and basal progesterone secretion in follicles during the differentiation of granulosa cells (Ledwitz-Rigby et al., 1987). "HIF-1 signaling pathway" was another KEGG term enriched in our POILMN. According to previous research studies, ROS accumulation induces oxidative damage to ovarian GCs, hence prompting the onset of follicular atresia and relevant anovulatory disorders, such as POI (Agarwal et al., 2012). More importantly, ROS are involved in the hypoxia response through a mechanism that stabilizes hypoxia-inducible factor 1 (Chandel et al., 2000).

We performed a topology analysis of lncRNAs, calculated topological parameters (betweenness and degree), and identified four candidate lncRNAs (HCP5, NUTM2A-AS1, GABPB1-IT1, and SMIM25) that may potentially affect POI susceptibility.

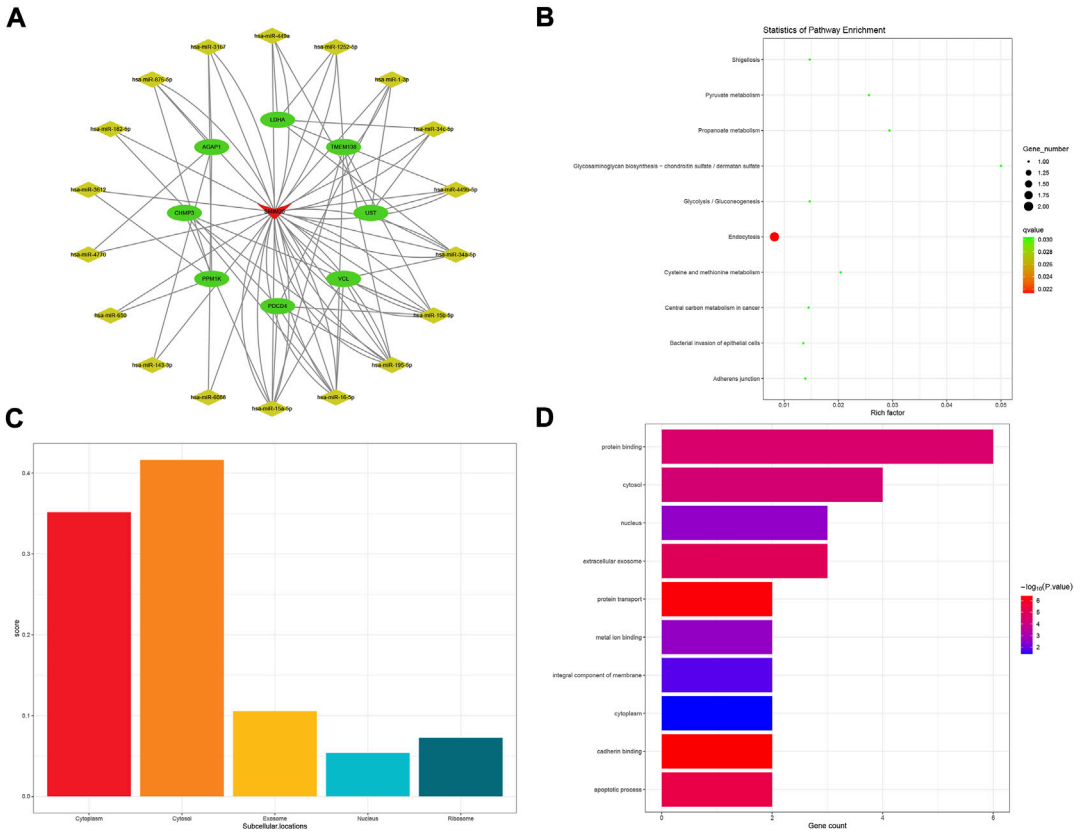


FIGURE 7
SMIM25-related ceRNA subnetwork analysis. ceRNA network of SMIM25 **(A)**. KEGG pathways enriched in SMIM25 **(B)**. Subcellular location analysis for SMIM25 **(C)**. GO biological process enrichment results for SMIM25 **(D)**.

The KEGG pathway of "vasopressin-regulated water reabsorption" was enriched in the ceRNA subnetwork of HCP5. Studies have shown that exposure to high doses of PFOA increases the risk of premature ovarian insufficiency by reducing pituitary expression in the suprachiasmatic nucleus (SCN); this implies a possible connection between POI and vasopressin in some way (Zhang et al., 2020). In addition, we also found that "N-glycan biosynthesis" and "metabolic pathways" exist in the enrichment of pathways; the same pathways were also expounding in the DEGs enrichment as before. Several energy metabolism-related GO terms, including "ATP-dependent microtubule motor activity, plus-end-directed," "ATP-dependent microtubule motor activity," and "adenylate cyclase binding" were enriched. Energy metabolism may play a central role in many physiological and pathological processes when HCP5 is activated and functioning. As for "protein kinase A regulatory subunit binding" and "protein kinase A binding" terms, although a few hundreds of protein kinases regulate key processes in human cells, protein kinases play a pivotal role in health and disease. This study demonstrates that protein kinase A appears to be an important upstream kinase sufficient to initiate complex intracellular signaling pathways and gene expression profiles associated with GC differentiation (Puri et al., 2016).

The KEGG "mTOR signaling pathway" and GO term "autophagy" were enriched in the ceRNA subnetwork of NUTM2A-AS1. Autophagy is an evolutionarily conserved cellular process controlled through a set of essential autophagy genes (Atgs). As an important player in autophagy, mTOR is essential for autophagosome formation and necessary for the closure of isolation membranes of autophagosomes. The activated mTOR pathway stimulates the proliferation of granulosa cells (Kayampilly and Menon, 2007; Yu et al., 2011) and also participates in the regulation of ovarian steroidogenesis. In some cases, researchers unveiled a novel role for mTOR signaling in the maintenance of granulosa cellular homeostasis by regulating autophagy at the transcriptional level (Yin et al., 2020).

For GO term enrichment analysis, the "oxidation-reduction process" was enriched in the ceRNA sub-network of GABPB1-IT1. Oxidative stress-induced granulosa cell (GC) death represents a common reason for follicular atresia, which can

cause amenorrhea beforehand. As for the KEGG pathway "glycolysis/gluconeogenesis," a recent study showed that the energy stress-induced lncRNA ZNF674-AS1 regulates GC proliferation and glycolysis, possibly contributing to follicular dysfunction (Li et al., 2021a).

There are also glucose metabolism pathways enriched in SMIM25 KEGG analysis, such as "glycolysis/gluconeogenesis" and "propanoate metabolism," that further emphasize the relationship between POI and energy metabolism. By comparing the analysis results of GO enrichments, we found that the "extracellular exosome" term was enriched. Exosomes are extracellular vesicles that mediate cellular communication in health and disease. It has also been shown that exosomes contain messenger RNAs (mRNAs) and microRNAs (miRNAs), which can be delivered unidirectionally and functionally between cells (Ratajczak et al., 2006; Valadi et al., 2007). Recent studies have shown that MSC-derived exosomes supplementation can restore ovarian function in premature ovarian insufficiency (Sun et al., 2019; Ding et al., 2020, 7; Yang et al., 2020).

In brief, our research provided a global view of ceRNA, lncRNA, and mRNA with potential implications for the onset and development of POI. Nevertheless, further longitudinal studies are necessary to extend and explore these potential lncRNAs.

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

CL and CM designed the study concept. CL analyzed the data, wrote the original draft, and revised the manuscript. JZ operated R software and visualization. LB and LW operated Cytoscape software. GY and SG obtained the starBase and miRTarBase databases. CM supervised the study process, participated in reviewing, and provided funding. All authors read and approved the final manuscript.

Funding

This study was supported by the Natural Science Foundation of China (81671535), the National Science and Technology support program project (2013BAI04B05), the Jiangsu Key Discipline of Human Assisted Reproduction Medicine Foundation (FXK202149), and the Jiangsu Key Discipline of Medicine Foundation of Commission of Health (ZDB2020007).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.956805/full#supplementary-material

References

Agarwal, A., Aponte-Mellado, A., Premkumar, B. J., Shaman, A., and Gupta, S. (2012). The effects of oxidative stress on female reproduction: A review. *Reprod. Biol. Endocrinol.* 10, 49. doi:10.1186/1477-7827-10-49

Amaral, P. P., and Mattick, J. S. (2008). Noncoding RNA in development. *Mamm. Genome* 19, 454–492. doi:10.1007/s00335-008-9136-7

Ates, S., Yesil, G., Sevket, O., Molla, T., and Yildiz, S. (2014). Comparison of metabolic profile and abdominal fat distribution between karyotypically normal women with premature ovarian insufficiency and age matched controls. *Maturitas* 79, 306–310. doi:10.1016/j.maturitas.2014.07.008

Chandel, N. S., McClintock, D. S., Feliciano, C. E., Wood, T. M., Melendez, J. A., Rodriguez, A. M., et al. (2000). Reactive oxygen species generated at mitochondrial complex III stabilize hypoxia-inducible factor-1alpha during hypoxia: A mechanism of O2 sensing. *J. Biol. Chem.* 275, 25130–25138. doi:10.1074/jbc.M001914200

Ding, C., Zhu, L., Shen, H., Lu, J., Zou, Q., Huang, C., et al. (2020). Exosomal miRNA-17-5p derived from human umbilical cord mesenchymal stem cells improves ovarian function in premature ovarian insufficiency by regulating SIRT7. *Stem Cells* 38, 1137–1148. doi:10.1002/stem.3204

Dinger, M. E., Amaral, P. P., Mercer, T. R., Pang, K. C., Bruce, S. J., Gardiner, B. B., et al. (2008). Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* 18, 1433–1445. doi:10.1101/gr.078378.108

Feng, C., Shen, J.-M., Lv, P.-P., Jin, M., Wang, L.-Q., Rao, J.-P., et al. (2018). Construction of implantation failure related lncRNA-mRNA network and identification of lncRNA biomarkers for predicting endometrial receptivity. *Int. J. Biol. Sci.* 14, 1361–1377. doi:10.7150/ijbs.25081

Huang, B., Qian, C., Ding, C., Meng, Q., Zou, Q., and Li, H. (2019). Fetal liver mesenchymal stem cells restore ovarian function in premature ovarian insufficiency by targeting MT1. *Stem Cell. Res. Ther.* 10, 362. doi:10.1186/s13287-019-1490-8

Huang, Y., Lv, Y., Qi, T., Luo, Z., Meng, X., Ying, Q., et al. (2021). Metabolic profile of women with premature ovarian insufficiency compared with that of age-matched healthy controls. *Maturitas* 148, 33–39. doi:10.1016/j.maturitas.2021.04.003

Kayampilly, P. P., and Menon, K. M. J. (2007). Follicle-stimulating hormone increases tuberin phosphorylation and mammalian target of rapamycin signaling through an extracellular signal-regulated kinase-dependent pathway in rat granulosa cells. *Endocrinology* 148, 3950–3957. doi:10.1210/en.2007-0202

Klattenhoff, C. A., Scheuermann, J. C., Surface, L. E., Bradley, R. K., Fields, P. A., Steinhauser, M. L., et al. (2013). Braveheart, a long noncoding RNA required for cardiovascular lineage commitment. *Cell.* 152, 570–583. doi:10.1016/j.cell.2013.01.003

Ledwitz-Rigby, F., Gross, T. M., Schjeide, O. A., and Rigby, B. W. (1987). The glycosaminoglycan chondroitin-4-sulfate alters progesterone secretion by porcine granulosa cells. *Biol. Reprod.* 36, 320–327. doi:10.1095/biolreprod36.2.320

Lee, J. T. (2012). Epigenetic regulation by long noncoding RNAs. *Science* 338, 1435–1439. doi:10.1126/science.1231776

Li, D., Wang, X., Li, G., Dang, Y., Zhao, S., and Qin, Y. (2021a). LncRNA ZNF674-AS1 regulates granulosa cell glycolysis and proliferation by interacting with ALDOA. *Cell. Death Discov.* 7, 107. doi:10.1038/s41420-021-00493-1

Li, D., Xu, W., Wang, X., Dang, Y., Xu, L., Lu, G., et al. (2021b). lncRNA DDGC participates in premature ovarian insufficiency through regulating RAD51 and WT1. *Mol. Ther. Nucleic Acids* 26, 1092–1106. doi:10.1016/j.omtn.2021.10.015

Liang, J., Li, X., Kang, C., Wang, Y., Kulikoff, X. R., Coates, M. M., et al. (2019). Maternal mortality ratios in 2852 Chinese counties, 1996–2015, and achievement of millennium development goal 5 in China: A subnational analysis of the global burden of disease study 2016. *Lancet* 393, 241–252. doi:10.1016/S0140-6736(18)31712-4

Ling, L., Feng, X., Wei, T., Wang, Y., Wang, Y., Wang, Z., et al. (2019). Human amnion-derived mesenchymal stem cell (hAD-MSC) transplantation improves ovarian function in rats with premature ovarian insufficiency (POI) at least partly through a paracrine mechanism. *Stem Cell. Res. Ther.* 10, 46. doi:10.1186/s13287-019-1136-x

Ma, Y., Ma, L., Cao, Y., and Zhai, J. (2021). Construction of a ceRNA-based lncRNA-mRNA network to identify functional lncRNAs in polycystic ovarian syndrome. *Aging* 13, 8481–8496. doi:10.18632/aging.202659

Podfigurna, A., Stellmach, A., Szeliga, A., Czyzyk, A., and Meczekalski, B. (2018). Metabolic profile of patients with premature ovarian insufficiency. *J. Clin. Med.* 7, 374. doi:10.3390/jcm7100374

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell.* 136, 629–641. doi:10.1016/j.cell.2009.02.006

Puri, P., Little-Ihrig, L., Chandran, U., Law, N. C., Hunzicker-Dunn, M., and Zeleznik, A. J. (2016). Protein kinase A: A master kinase of granulosa cell differentiation. *Sci. Rep.* 6, 28132. doi:10.1038/srep28132

Ratajczak, J., Miekus, K., Kucia, M., Zhang, J., Reca, R., Dvorak, P., et al. (2006). Embryonic stem cell-derived microvesicles reprogram hematopoietic progenitors: Evidence for horizontal transfer of mRNA and protein delivery. *Leukemia* 20, 847–856. doi:10.1038/sj.leu.2404132

Raza, S. H. A., Khan, R., Cheng, G., Long, F., Bing, S., Easa, A. A., et al. (2022). RNA-Seq reveals the potential molecular mechanisms of bovine KLF6 gene in the regulation of adipogenesis. *Int. J. Biol. Macromol.* 195, 198–206. doi:10.1016/j.ijbiomac.2021.11.202

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell.* 129, 1311–1323. doi:10.1016/j.cell.2007.05.022

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: The rosetta stone of a hidden RNA language? *Cell.* 146, 353–358. doi:10.1016/j.cell.2011.07.014

Sun, B., Ma, Y., Wang, F., Hu, L., and Sun, Y. (2019). miR-644-5p carried by bone mesenchymal stem cell-derived exosomes targets regulation of p53 to inhibit ovarian granulosa cell apoptosis. *Stem Cell. Res. Ther.* 10, 360. doi:10.1186/s13287-019-1442-3

Touraine, P. (2020). Premature ovarian insufficiency: Step-by-step genetics bring new insights. *Fertil. Steril.* 113, 767–768. doi:10.1016/j.fertnstert.2019.12.032

Valadi, H., Ekström, K., Bossios, A., Sjöstrand, M., Lee, J. J., and Lötvall, J. O. (2007). Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat. Cell. Biol.* 9, 654–659. doi:10.1038/ncb1596

Wang, F., Chen, X., Sun, B., Ma, Y., Niu, W., Zhai, J., et al. (2021). Hypermethylation-mediated downregulation of lncRNA PVT1 promotes granulosa cell apoptosis in premature ovarian insufficiency via interacting with Foxo3a. *J. Cell. Physiol.* 236, 5162–5175. doi:10.1002/jcp.30222

Wang, S., Lin, S., Zhu, M., Li, C., Chen, S., Pu, L., et al. (2019). Acupuncture reduces apoptosis of granulosa cells in rats with premature ovarian failure via restoring the PI3K/akt signaling pathway. *Int. J. Mol. Sci.* 20, 6311. doi:10.3390/ijms20246311

Wang, X., Zhang, X., Dang, Y., Li, D., Lu, G., Chan, W.-Y., et al. (2020). Long noncoding RNA HCP5 participates in premature ovarian insufficiency by transcriptionally regulating MSH5 and DNA damage repair via YB1. *Nucleic Acids Res.* 48, 4480–4491. doi:10.1093/nar/gkaa127

European Society for Human Reproduction and Embryology (ESHRE) Guideline Group on POIWebber, L., Davies, M., Anderson, R., Bartlett, J., Braat, D., et al. (2016). ESHRE guideline: Management of women with premature ovarian insufficiency. *Hum. Reprod.* 31, 926–937. doi:10.1093/humrep/dew027

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*. 2, 100141. doi:10.1016/j.xinn.2021.100141

Yang, W., Zhang, J., Xu, B., He, Y., Liu, W., Li, J., et al. (2020). HucMSC-derived exosomes mitigate the age-related retardation of fertility in female mice. *Mol. Ther.* 28, 1200–1213. doi:10.1016/j.ymthe.2020.02.003

Yin, N., Wu, C., Qiu, J., Zhang, Y., Bo, L., Xu, Y., et al. (2020). Protective properties of heme oxygenase-1 expressed in umbilical cord mesenchymal stem cells help restore the ovarian function of premature ovarian failure mice through activating the JNK/Bcl-2 signal pathway-regulated autophagy and upregulating the circulating of CD8+CD28− T cells. *Stem Cell. Res. Ther.* 11, 49. doi:10.1186/s13287-019-1537-x

Yu, J., Yaba, A., Kasiman, C., Thomson, T., and Johnson, J. (2011). mTOR controls ovarian follicle growth by regulating granulosa cell proliferation. *PLoS ONE* 6, e21415. doi:10.1371/journal.pone.0021415

Zhang, Y., Cao, X., Chen, L., Qin, Y., Xu, Y., Tian, Y., et al. (2020). Exposure of female mice to perfluorooctanoic acid suppresses hypothalamic kisspeptin-reproductive endocrine system through enhanced hepatic fibroblast growth factor 21 synthesis, leading to ovulation failure and prolonged dioestrus. *J. Neuroendocrinol.* 32, e12848. doi:10.1111/jne.12848

Zhao, X., Tang, D., Zuo, X., Zhang, T., and Wang, C. (2019). Identification of lncRNA–miRNA–mRNA regulatory network associated with epithelial ovarian cancer cisplatin-resistant. *J. Cell. Physiol.* 234, 19886–19894. doi:10.1002/jcp.28587

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 October 2022
DOI 10.3389/fgene.2022.1029300

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Xing Chen,
China University of Mining and Technology, China
Jin-Xing Liu,
Qufu Normal University, China

*CORRESPONDENCE
Yu Wang,
2007002@glut.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 27 August 2022
ACCEPTED 03 October 2022
PUBLISHED 20 October 2022

CITATION
Zhang Y, Wang Y, Li X, Liu Y and Chen M (2022), Identifying lncRNA–disease association based on GAT multiple-operator aggregation and inductive matrix completion.
Front. Genet. 13:1029300.
doi: 10.3389/fgene.2022.1029300

Identifying lncRNA–disease association based on GAT multiple-operator aggregation and inductive matrix completion

Yi Zhang[1,2], Yu Wang[1,2]*, Xin Li[1,2], Yarong Liu[1,2] and Min Chen[3]

[1]Guilin University of Technology, Guilin, China, [2]Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin University of Technology, Guilin, China, [3]School of Computer Science and Technology, Hunan Institute of Technology, Hengyang, China

Computable models as a fundamental candidate for traditional biological experiments have been applied in inferring lncRNA–disease association (LDA) for many years, without time-consuming and laborious limitations. However, sparsity inherently existing in known heterogeneous bio-data is an obstacle to computable models to improve prediction accuracy further. Therefore, a new computational model composed of multiple mechanisms for lncRNA–disease association (MM-LDA) prediction was proposed, based on the fusion of the graph attention network (GAT) and inductive matrix completion (IMC). MM-LDA has two key steps to improve prediction accuracy: first, a multiple-operator aggregation was designed in the n-heads attention mechanism of the GAT. With this step, features of lncRNA nodes and disease nodes were enhanced. Second, IMC was introduced into the enhanced node features obtained in the first step, and then the LDA network was reconstructed to solve the cold start problem when data deficiency of the entire row or column happened in a known association matrix. Our MM-LDA achieved the following progress: first, using the Adam optimizer that adaptively adjusted the model learning rate could increase the convergent speed and not fall into local optima as well. Second, more excellent predictive ability was achieved against other similar models (with an AUC value of 0.9395 and an AUPR value of 0.8057 obtained from 5-fold cross-validation). Third, a 6.45% lower time cost was consumed against the advanced model GAMCLDA. In short, our MM-LDA achieved a more comprehensive prediction performance in terms of prediction accuracy and time cost.

KEYWORDS

graph attention network, inductive matrix completion, association prediction, aggregation, multiple-operator

Abbreviations: ROC, receiver operating characteristic; AUC, area under the ROC curve; FPR, false positive rate; TPR, GAT, IMC, and LDA true positive rate, graph attention network, inductive matrix completion, and LncRNA–disease associations, respectively.

Introduction

Long non-coding RNA, named for its transcription length of over 200 nucleotides, has received extensive attention from biological researchers (Sun et al., 2018). With the in-depth development of biomedicine, many literatures have confirmed that lncRNA plays an important role in the activities of living organisms through dose compensation effect, genetic expression, cell differentiation, and other ways and gradually becomes the focus of bioinformatics. Studies have shown that abnormal lncRNA expression can lead to a variety of complex diseases, especially as both oncogenes and tumor suppressors in the tumorigenesis of diverse cancers (Chen et al., 2020). The exploration of lncRNA leading to disease is helpful in understanding the mechanism of disease generation and provides reference for disease treatment and prognosis (Xia et al., 2013). Therefore, the work on predicting lncRNA–disease associations is significant for human disease diagnostics and prognostics and will improve the development of drug discovery (Chen et al., 2020).

As biological experiments are time-consuming and laborious, numerous computational models are mostly used to replace biological experiments in real life to identify disease-related associations and provide efficient and more accurate candidates for biological experiments in recent years (Chen et al., 2019; Wang et al., 2021; Huang et al., 2022a; Huang et al., 2022b; Huang et al., 2022c). Currently, computational models for predicting lncRNA–disease associations (LDAs) commonly fall into three categories.

The first category of methods is based on constructing biological similarity networks. Label propagation algorithms are used commonly in association-related prediction (Yin et al., 2020), especially as restart random walk and KATZ, whose main difference is applied in different underlying networks. Sun et al, (2014) and Chen et al, (2016) established the global restart random walk algorithm by using the lncRNA functional similarity network so as to predict potential association information. However, these models could not work on isolated diseases (diseases without known association information) or new lncRNAs (lncRNAs without known association information). Based on the gene–disease association and lncRNA–disease similarity network, Ma et al, (2019) introduced the HeteSim algorithm to construct a gene–disease heterogeneous information network, with which the network structure was strengthened by increasing the number of edges in the network. Potential associations can be propagated with more information and with better prediction effects. Chen, 2015; Chen et al, (2019) combined known LDA, lncRNA expression profile information, lncRNA functional similarity, disease semantic similarity, and Gaussian interaction spectrum kernel similarity to establish association prediction models. Although these models could work on isolated diseases or new lncRNAs, the prediction accuracy is still not high enough.

The second category of methods utilizes machine learning with a classifier to identify pathogenic lncRNAs. Chen and Yan, (2013) used lncRNA expression profile information to develop a classic and significant calculation model LRLSLDA for inferring potential lncRNA–disease pair information. This model is the first to use Laplacian regularized least squares in a semi-supervised learning framework, and it could work on new lncRNAs and isolated diseases without needing negative samples. However, its selection of optimal parameters is complicated because of its disease space and lncRNA space belonging to two classifiers. Later, Chen et al, (2015) developed an improved correlation prediction model LNCSIM to further improve the prediction accuracy. However, with its prediction results biased toward those lncRNAs with more known associations, the prediction effect is not good enough for isolated diseases and new lncRNAs with less known information. In addition, selecting attenuation factors of semantic contribution has not been well-solved. Zhao et al, (2015) predicted potentially pathogenic lncRNA by integrating known disease-related lncRNA and a variety of biological data (genomic data, regulatory, and transcriptional biological data) based on the Bayesian algorithm. Although the prediction performance of this model is good, sufficient negative samples of the Bayesian classifier are required to improve the prediction performance.

The third category of methods is based on disease-related genes, for example, mRNA, miRNA, and protein information. Models belonging to the aforementioned two categories all rely on the known LDA, whose number with experimental verification is relatively small. Therefore, researchers have to explore new ideas to infer the potential associations with using third-party data, also known as genetic information. Zhou et al, (2015) selected appropriate thresholds and coefficients to predict lncRNA–disease pairs, using the expression data of three kinds of non-coding RNAs (mRNA, miRNA, and lncRNA). Cheng et al, (2016) introduced mRNA- and miRNA-related data into the prediction of LDA. Compared with other methods, methods within this category are more reliable and stable, but the model performance is highly dependent on coactions found among the three kinds of non-coding RNAs.

Utilizing deep learning technology has gradually become a research hotspot to make up for the deficiencies in the abovementioned three categories. The graph that can abstract the relationship between entities is widely used as a data structure (Wu et al., 2020). Wu et al, (2021) proposed a computational method MLGCNET that applied the graph convolutional network (GCN) to extract the node information with which to feed into an extra tree (ET) classifier for accurately predicting the potential lncRNA–disease associations. The graph attention network (GAT), as a promising graph neural network, has been applied to a number of bioinformatics tasks. Long et al, (2021) proposed a new method GATMDA based on the GAT to identify a microbial–disease association. Bian et al, (2021) proposed a

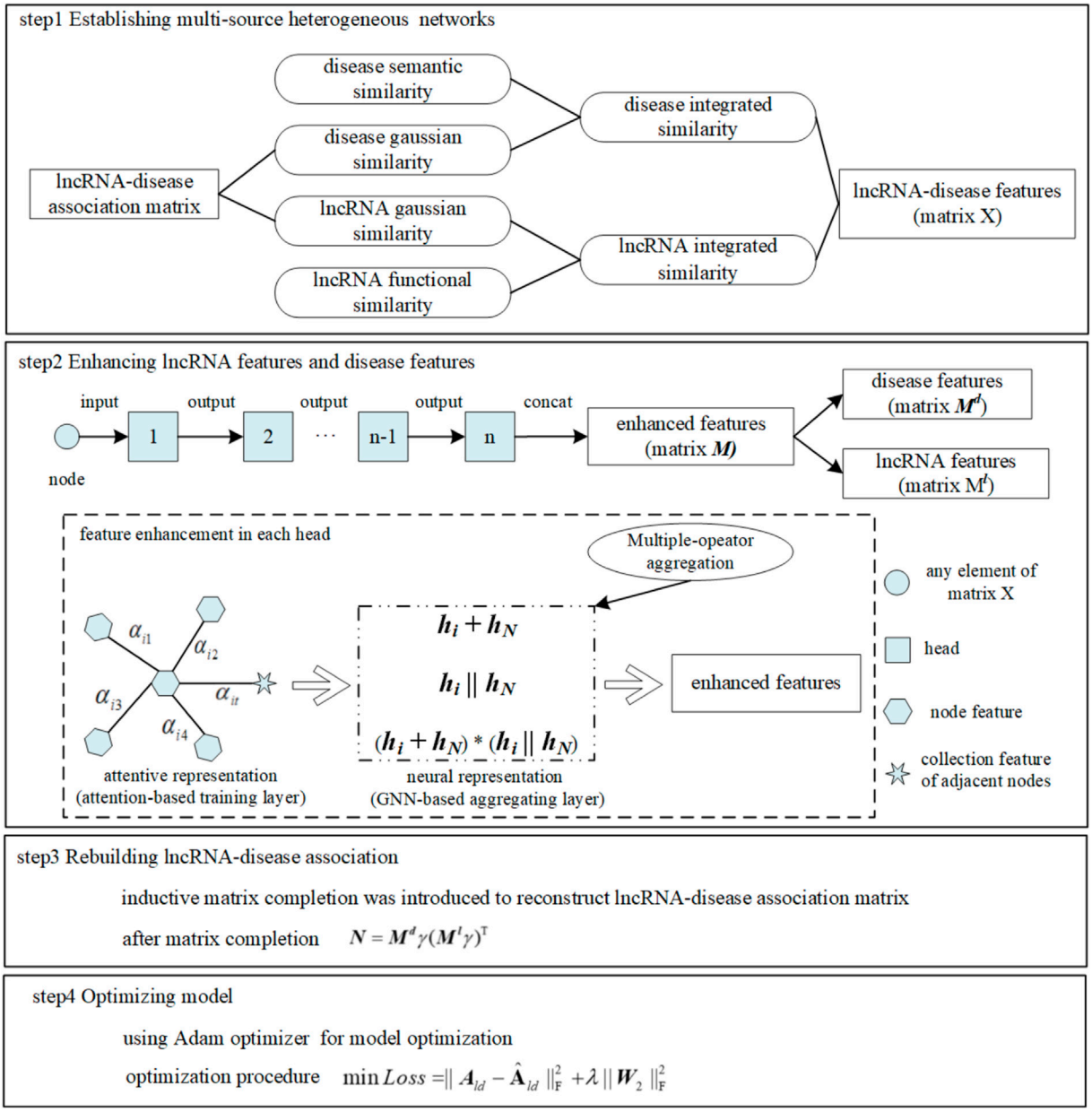


FIGURE 1
MM-LDA workflow.

model GATCDA to predict circRNA–disease associations based on the GAT. Gu et al, (2021) predicted drug ADMET classification based on the GAT. However, this model did not discuss the time complexity consumed for achieving high accuracy. Inductive matrix completion (IMC) that could fill data sparsity existing in the bio-database inherently caused the problem of low prediction accuracy when it was applied in inferring LDA directly and separately (Natarajan and Dhillon, 2014; Huang et al., 2017; Chen et al., 2018; Lu et al., 2018; Fraidouni and Zaruba, 2019; Chen et al., 2021). Therefore, to break through the aforementioned limitations, multiple mechanisms were fused into a new computational model, such as MM-LDA, as shown in Figure 1. On one hand, a multiple-operator aggregation used in the n-heads attention mechanism of the GAT was designed, where it could enhance the features of lncRNA nodes (or disease nodes) to avoid the low prediction accuracy caused by known-data sparsity. On the other hand, with enhanced node features, the LDA network was rebuilt

by IMC that could renew the missing elements in the bio-database. In the end, the Adam optimizer was used to further improve the prediction accuracy.

Materials and methods

Data source

Known lncRNA–disease association: After removing repeated and redundant lncRNAs (diseases) in the original dataset lncRNA disease V2.0 (Bao et al., 2019), a processed dataset composed of associations between human diseases and lncRNAs was used in our model. This dataset contains 352 LDAs verified experimentally, involving 156 lncRNAs and 190 diseases. It is an unbalanced dataset with existing inherent data sparsity because of less known associations against unknown or non-existent associations.

For formal description later, the number of lncRNAs and diseases involved in this dataset (also called association matrix) was denoted by *nl* and *nd*, respectively. In the association matrix ($\mathbf{A}_{ld} \in \mathbb{R}^{nl\times nd}$), any known lncRNA–disease association that relates to disease d_i and lncRNA l_j with experimental verification works as the positive sample, with denotation of $\mathbf{A}_{ld}(l_i, d_j) = 1$. Otherwise, any unknown or non-existent lncRNA–disease association works as the negative sample, with denotation of $\mathbf{A}_{ld}(l_i, d_j) = 0$.

Multi-source heterogeneous networks

Disease–disease semantic similarity network: Directed acyclic graph (DAG) was utilized to calculate the semantic similarity between diseases (Wang et al., 2010). The semantic contribution value of any disease d_t to disease d_i was denoted by $D_{d_i}(d_t)$.

$$D_{d_i}(d_t) = \begin{cases} 1, & d_t = d_i, \\ \max\{\gamma D_{d_i}(d_{t'}) | d_{t'} \in \text{children of } d_t\}, & d_t \neq d_i, \end{cases} \quad (1)$$

where γ is the coefficient regulating semantic contribution (Wang et al., 2010), and it was set to the optimal value of 0.5.

If two diseases have more overlaps in DAG, it implies greater similarity between them (Wang et al., 2010). Matrix$\mathbf{D}_S \in \mathbb{R}^{nd\times nd}$represents the semantic similarity network of diseases, and its element$\mathbf{D}_S(d_i, d_j)$ represents the semantic similarity between diseases d_i and d_j.

$$\mathbf{D}_S(d_i, d_j) = \frac{\sum_{d_m \in (T_{d_i} \cap T_{d_j})} \left(D_{d_i}(d_m) + D_{d_j}(d_m)\right)}{S(d_i) + S(d_j)}, \quad (2)$$

where T_{d_i} represents the DAG of disease d_i and $S(d_i)$ represents the semantic value of disease d_i.

$$S(d_i) = \sum_{d_t \in T_{d_i}} D_{d_i}(d_t). \quad (3)$$

LncRNA–lncRNA functional similarity network: Functionally similar lncRNAs are often associated with diseases in similar phenotypes (Wang et al., 2010). To calculate the functional similarity between two lncRNAs, the semantic similarity of diseases and its correlation to lncRNAs were utilized. Set $D = \{d_1, d_2, \cdots, d_t, \cdots, d_{nd}\}$ represents the disease set, and $\max(d_t, D)$ represents the maximum semantic similarity of any disease d_t in set D:

$$\max(d_t, D) = \max_{1\le i\le nd}(D_S(d_t, d_i)). \quad (4)$$

Matrix $\mathbf{F}_S \in \mathbb{R}^{nl\times nl}$ represents the functional similarity network of lncRNAs, and matrix element $\mathbf{F}_S(l_i, l_j)$ represents the functional similarity between lncRNA l_i and l_j.

$$\mathbf{F}_S(l_i, l_j) = \frac{\sum_{1\le i\le m}\max(d_i, D_1) + \sum_{1\le j\le n}\max(d_j, D_2)}{m+n}, \quad (5)$$

where set D_1 represents the set of diseases associated with lncRNA l_i, set D_2 represents the set of diseases associated with lncRNA l_j, and m and n represent the number of diseases in set D_1 and D_2, respectively.

Gaussian interaction spectrum kernel similarity network: As an efficient and useful method in biological information classification, the Gaussian kernel function (Van Laarhoven et al., 2011) has been applied to the association network when some diseases do not have semantic similarity. Gaussian interaction spectrum kernel similarity of diseases (Gaussian similarity) calculated by the Gaussian kernel function could replace the semantic similarity of disease. If disease d_i has a known experimentally verified association with any lncRNA, $I_P(d_i) = 1$; if disease d_i does not have any known association experimentally verified, $I_P(d_i) = 0$. Matrix $\mathbf{G}_D \in \mathbb{R}^{nd\times nd}$ represents the Gaussian similarity network of diseases, whose element $\mathbf{G}_D(d_i, d_j)$ represents the Gaussian similarity between disease d_i and d_j:

$$\mathbf{G}_D(d_i, d_j) = \exp\left(-\lambda_d \left\| I_P(d_i) - I_P(d_j)\right\|^2\right), \quad (6)$$

where λ_d is the standardized core bandwidth, with detailed calculation as

$$\lambda_d = \frac{1}{\frac{1}{nd}\sum_{i=1}^{nd}\|I_P(d_i)\|^2}. \quad (7)$$

Similarly, matrix $\mathbf{G}_L \in \mathbb{R}^{nl\times nl}$ represents the Gaussian similarity network of lncRNAs, and matrix element $\mathbf{G}_L(l_i, l_j)$ represents the Gaussian similarity between lncRNA l_i and l_j.

$$\mathbf{G}_L(l_i, l_j) = \exp\left(-\lambda_l \left\| I_P(l_i) - I_P(l_j)\right\|^2\right). \quad (8)$$

$$\lambda_l = \frac{1}{\frac{1}{nl}\sum_{i=1}^{nl}\|I_P(l_i)\|^2} \quad (9)$$

.

Integrated similarity network: Since not all diseases involved could calculate the semantic similarity due to the inherent sparsity in the dataset, an integrated similarity network $\mathbf{D}_S^{(I)}$ was constructed to improve the accuracy of disease semantic similarity. The matrix element $\mathbf{D}_S^{(I)}(d_i, d_j)$ was formed as

$$\mathbf{D}_S^{(I)}(d_i, d_j) = \begin{cases} \mathbf{D}_S(d_i, d_j) + \mathbf{G}_D(d_i, d_j), & \mathbf{D}_S(d_i, d_j) \neq 0, \\ \mathbf{G}_D(d_i, d_j), & \mathbf{D}_S(d_i, d_j) = 0. \end{cases} \tag{10}$$

Similarly, matrix $\mathbf{F}_S^{(I)}$ represents the integrated similarity network of lncRNAs, and the matrix element $\mathbf{F}_S^{(I)}(l_i, l_j)$ has the specific form as

$$\mathbf{F}_S^{(I)}(l_i, l_j) = \begin{cases} \mathbf{F}_S(l_i, l_j), & \mathbf{F}_S(l_i, l_j) \neq 0, \\ \mathbf{G}_L(l_i, l_j), & \mathbf{F}_S(l_i, l_j) = 0. \end{cases} \tag{11}$$

Finally, a multi-source heterogeneous network as a diagonal matrix was constructed, preparing for the following calculation in the model:

$$\mathbf{X} = \begin{bmatrix} \mathbf{0} & \mathbf{D}_S^{(I)} \\ \mathbf{F}_S^{(I)} & \mathbf{0} \end{bmatrix}. \tag{12}$$

Node feature enhancement

N-heads attention with multiple-operator aggregation: The original GAT utilizes attention scores to adaptively aggregate information from neighbor nodes during node updating and learns the representation of nodes on the graph by assigning different weights to its neighbor nodes. N-heads attention could stabilize the process of self-attention, with n time iterations (Fraidouni and Zaruba, 2019). However, n-heads attention only uses the "concatenation" operator to aggregate the features coming from each head. The aggregation effect needs to be improved further by adding more operators in each head, and a multiple-operator for n-heads attention was constructed to enhance node features.

Attention-based feature training: Any element in the feature vector matrix $\mathbf{X}$ was considered the node feature. In the kth iteration, attention score e_{ij}^k of node i to neighbor node j in matrix $\mathbf{X}$ was calculated as

$$e_{ij}^k = f\left(\mathbf{h}_i^k \mathbf{W}, \mathbf{h}_j^k \mathbf{W}\right), \tag{13}$$

where $f(\cdot)$ denotes a single-layer neural network; $\mathbf{h}_i^k$ denotes the feature vector of node i in the kth iteration; and $\mathbf{W} \in \mathbb{R}^{(nl+nd)\times 1}$ denotes the weighted matrix.

In order to make the attention score within the interval of [0,1], the softmax function was used for normalization

$$\alpha_{ij}^k = \frac{\exp\left(e_{ij}^k\right)}{\sum_{t\in N_i} \exp\left(e_{it}^k\right)}, \tag{14}$$

where N_i represents the set of all neighbor nodes of node i in matrix $\mathbf{X}$. In the kth iteration, features of all nodes in set N_i were calculated as

$$\mathbf{h}_{N_i}^k = \sum_{t\in N_i} \alpha_{it}^k \mathbf{h}_t^k. \tag{15}$$

GNN-based feature aggregation: In order to enhance node features further, based on a nonlinear graph neural network (GNN), a multiple-operator that aggregated the features coming from the attention-based feature training layer was designed:

$$\begin{aligned} \mathbf{M}^k = &\text{LeakyReLU}\left(\left(\mathbf{h}_i^k + \mathbf{h}_{N_i}^k\right)\mathbf{W}_1\right) + \text{LeakyReLU}\left(\left(\mathbf{h}_i^k \| \mathbf{h}_{N_i}^k\right)\mathbf{W}_1\right) \\ &+\left(\text{LeakyReLU}\left(\left(\mathbf{h}_i^k + \mathbf{h}_{N_i}^k\right)\mathbf{W}_1\right) \times \text{LeakyReLU}\left(\left(\mathbf{h}_i^k \| \mathbf{h}_{N_i}^k\right)\mathbf{W}_1\right)\right), \end{aligned} \tag{16}$$

where $\mathbf{M}^k$ represents the feature vector after aggregating, LeakyReLU$(\cdot)$ is the activating function, "+" denotes the adding operation, "‖" denotes the concatenating operation, and $\mathbf{W}_1 \in \mathbb{R}^{(nl+nd)\times k}$ is a weighted matrix. Finally, the feature vector $\mathbf{M}^k$ *via* the n-heads attention mechanism formed the final feature matrix $\mathbf{M}$:

$$\mathbf{M} = \|_{k=1}^{n} \mathbf{M}^k = \begin{bmatrix} \mathbf{M}^d \\ \mathbf{M}^l \end{bmatrix}, \tag{17}$$

where $\mathbf{M}^d \in \mathbb{R}^{nd\times(nl+nd)}$ represents the feature matrix of diseases and $\mathbf{M}^l \in \mathbb{R}^{nl\times(nl+nd)}$ represents the feature matrix of lncRNAs.

LncRNA–disease association reconstruction

Inductive matrix completion: Known LDA was represented as a low-rank matrix in original matrix completion which recovers missing elements only with less sampling data (Chen and Chen, 2017). However, a cold start phenomenon will occur, when the entire row or column of data is missing. IMC technology introduced could fix the cold start problem and improve prediction accuracy because the number of parameters that was learned in IMC only related to the number of features of lncRNAs (or diseases), not the number of lncRNAs (or diseases).

$$\hat{\mathbf{A}}_{ld} = \boldsymbol{M}^{\boldsymbol{d}} \gamma \left(\boldsymbol{M}^{\boldsymbol{l}} \gamma\right)^{\mathrm{T}}, \tag{18}$$

where $\hat{\mathbf{A}}_{ld}$ represents the reconstruction of association matrix $\mathbf{A}_{ld}$ and γ is the weight decay parameter.

Model optimization: Optimization of MM-LDA mainly focused on parameter training by minimizing the loss function. During parameter training, improper selection of learning rates will cause abnormal loss function. A large learning rate will lead to the non-convergence of the loss function. Otherwise, a small learning rate will make the model trap into local optimization. Therefore, the Adam optimizer (Kingma and Ba, 2014) that combined the advantages of an

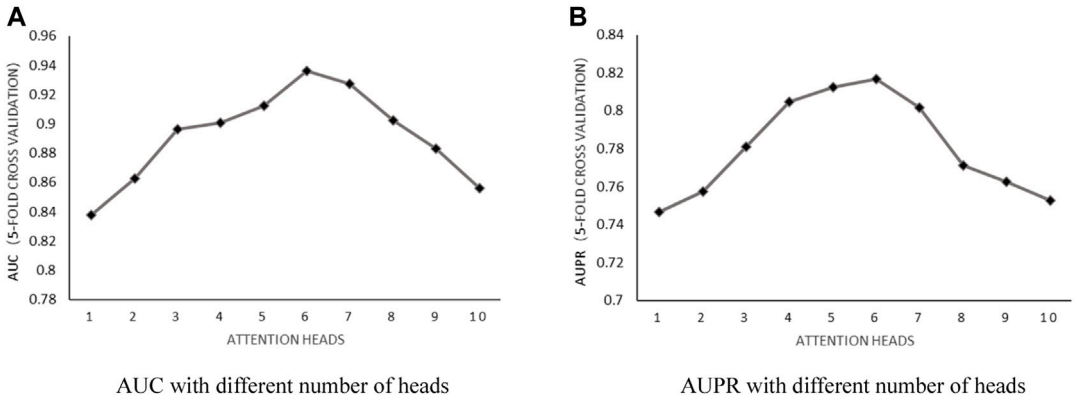


FIGURE 2
(A) AUC with different number of heads. **(B)** AUPR with different number of heads.

AdaGrad (adaptive gradient) optimizer (Lydia and Francis, 2019) and RMSprop (root mean square propagation) optimizer (Xu et al., 2021) was adopted in our model. Only requiring small memory space, the Adam optimizer with a simple and efficient implementation process could adjust the learning rate adaptively without being affected by gradient scaling, thus speeding up the model optimization speed. The optimization process by minimizing the loss function was formalized as

$$\text{minLoss} = \|\mathbf{A}_{ld} - \hat{\mathbf{A}}_{ld}\|_F^2 + \lambda\|\mathbf{W}_2\|_F^2, \quad (19)$$

where λ is the equilibrium factor with the value of 1 and $\mathbf{W}_2 \in \mathbb{R}^{nl\times nd}$ denotes a weighted matrix.

Results

Experimental evaluation

Evaluation metrics: All known LDAs were randomly divided into five groups with which 5-fold cross-validation was carried out to evaluate the predictive performance of our model. Successively selecting one group in five (as negative samples) with a group of unknown lncRNA–disease pairs in the same size (as negative samples) made up the test samples. The remaining four groups in five and the remaining unknown lncRNA–disease pairs were used to train the model. A total of five model evaluation metrics were defined by setting different thresholds, including true positive rate (TPR), false positive rate (FPR), and recall rate. Model performance was measured by an area under the ROC curve (AUC) and an area under the PR curve (AUPR). In order to avoid the influence of grouping randomly, each experiment was repeated 10 times. Finally, an AUC value and AUPR value were calculated according to the average value of the results from the 10 repeated experiments.

Parameter selection: Parameters used in our model could impact the predictive performance in the process of model training. Therefore, this section discussed the selection process of these three parameters in detail.

Number of attention heads: According to the literature (Fraidouni and Zaruba, 2019), the number of heads used in n-heads attention was discussed by setting the weight decay parameter γ as 5E-4 and the number of neurons as 8. After implementing 5-fold cross-validation, the results shown in Figure 2 proved that the number of heads impacted the predictive performance significantly. When the number of heads in n-heads attention was set to 6, the maximum AUC value and AUPR value could be obtained.

Weight decay parameter: According to the previous training, with the number of heads in a fixed value of 6 and the number of neurons in fixed value of 8, the influence of the weight decay parameter γ was discussed. The parameter value of γ was increased from 5E-6 to 5E-1, with a step size of E-1. After implementing 5-fold cross-validation, the results shown in Figure 3 proved that the model achieved the best predictive performance when γ was set to be 5E-2.

Number of neurons: With the number of heads in a fixed value of 6 and the weight decay parameter in fixed value of 5E-2, the influence of the number of neurons on predictive performance was discussed by choosing the value within the set of [4, 8, 16, 32, 64, and 128]. After implementing 5-fold cross-validation, the results shown in Figure 4 proved that AUC and AUPR obtained the best values when the number of neurons was set to 16.

Based on the previously mentioned discussion, by setting the number of heads in a fixed value of 6, the weight decay parameter γ in a fixed value of 5E-2, and the number of neurons in a fixed value of 16, our MM-LDA achieved the best AUC value of 0.9395 and AUPR value of 0.8057.

Ablation experiments: In order to evaluate the role of each kernel part in MM-LDA, such as multiple-operator aggregation

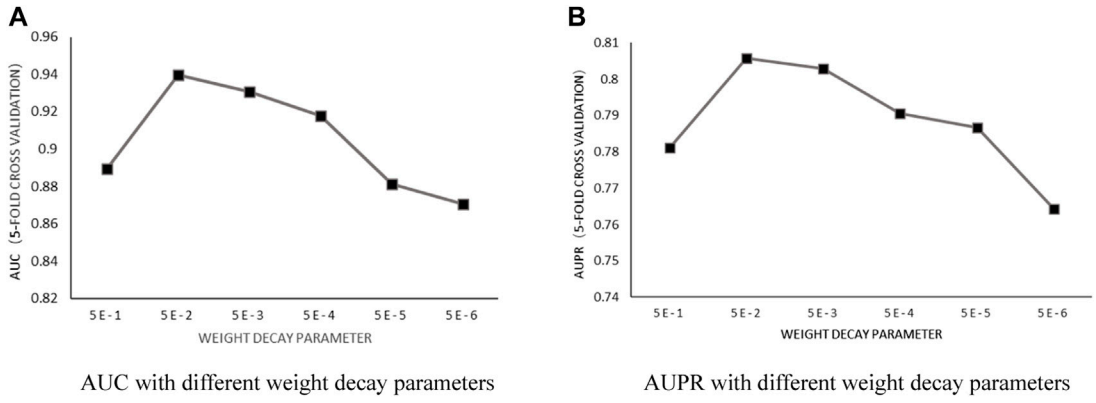


FIGURE 3
(A) AUC with different weight decay parameters. **(B)** AUPR with different weight decay parameters.

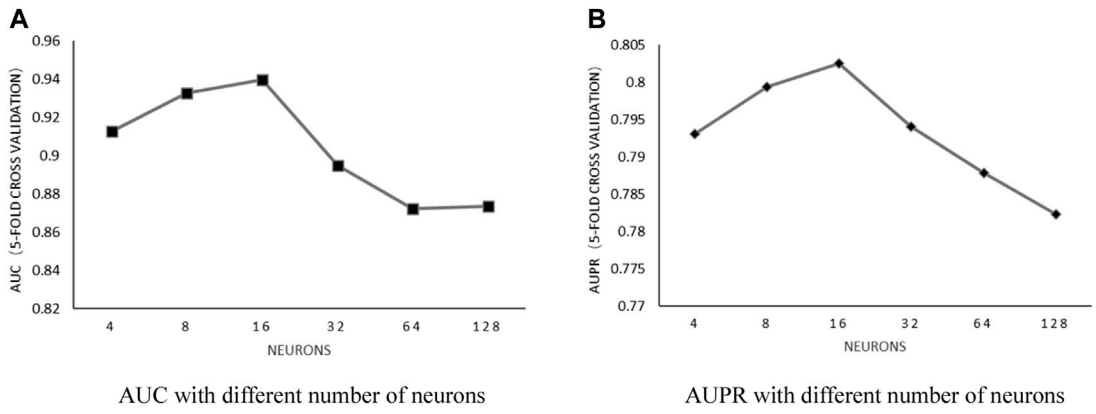


FIGURE 4
(A) AUC with different number of neurons. **(B)** AUPR with different number of neurons.

in n-heads attention, IMC in lncRNA–disease association reconstruction, three ablation experiments that were used to compare with our MM-LDA were set up:

- GAT-NG: A prediction model was constructed without kernel similarity of the Gaussian interaction spectrum as the kernel part.
- GAT-GIMC: A prediction model was constructed only based on a standard multiple-heads graph attention network.
- GAT-GMC: A prediction model was constructed only based on standard matrix completion.

For each ablation experiment, 5-fold cross-validation was repeated 10 times, and the average values of the results are shown in Figure 5.

From the results shown, MM-LDA obtained 5.65%, 3.3%, and 3.1% higher AUC values than GAT-NG, GAT-GMC, and GAT-GIMC, respectively. Furthermore, MM-LDA obtained 14.62%, 9.6%, and 13% higher AUPR values than GAT-NG, GAT-GMC, and GAT-GIMC, respectively. Therefore, it proved that the three kernel parts (integrated Gaussian interaction spectrum kernel similarity, multiple-operator aggregation in n-heads attention, and IMC) of MM-LDA could significantly improve the predictive performance.

Comparison with other models: SDLDA (Zeng et al., 2020b), DMFLDA (Zeng et al., 2020a), and GAMCLDA (Lu et al., 2019), the three computational models based on machine learning and matrix factorization in recent 3 years, were compared with our MM-LDA on the same dataset ($\mathbf{A}_{ld} \in \mathbb{R}^{nl \times nd}$). After 5-fold cross-validation was carried out, the detailed results are shown in Figure 6 and Table 1 to further prove the remarkable performance of MM-LDA.

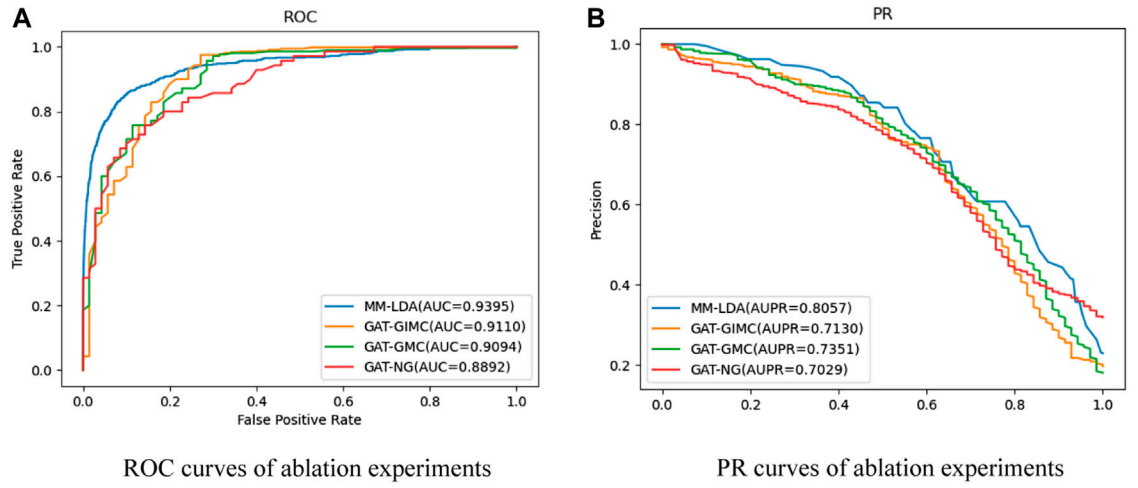


FIGURE 5
(A) ROC curves of ablation experiments. **(B)** PR curves of ablation experiments.

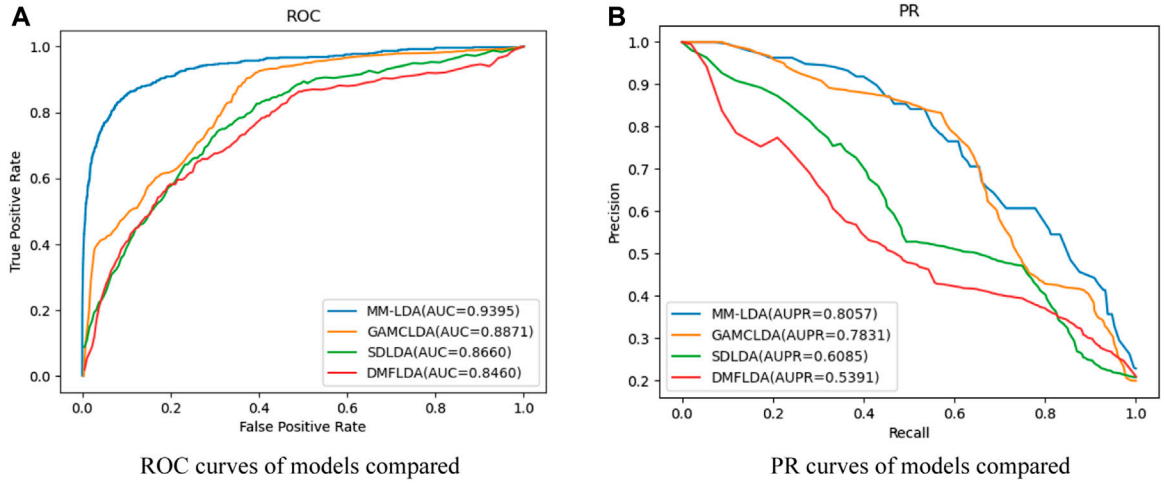


FIGURE 6
(A) ROC curves of models compared. **(B)** PR curves of models compared.

TABLE 1 AUC value (AUPR value) and running time of models compared.

Model	AUC	AUPR	Time (hour)
MM-LDA	0.9395	0.8057	1.24
GAMCLDA	0.8871	0.7831	1.32
SDLDA	0.8660	0.6085	1.15
DMFLDA	0.8460	0.5391	1.18

From the results shown, we could easily find that MM-LDA obtained the best AUC value that is 5.9%, 6.05%, and 11.05% higher than that of GAMCLDA, SDLDA, and DMFLDA, respectively. In addition, MM-LDA also obtained the best AUPR value that is 2.9%, 32.4%, and 49.5% higher than that of GAMCLDA, SDLDA, and DMFLDA, respectively. Though the running time of MM-LDA is 7.82% and 5.08% longer than that of SDLDA and DMFLDA, MM-LDA achieved the highest cost-effective prediction performance comprehensively.

TABLE 2 Top 10 gastric cancer-related lncRNAs.

Rank	LncRNA	Evidence
1	UCA1	LncRNA disease
2	TCL6	Literature [6]
3	PCA3	Literature [6]
4	HOTAIR	LncRNA disease
5	H19	LncRNA disease
6	MALAT1	Unconfirmed
7	BCAR4	LncRNA disease
8	HCP5	LncRNA disease
9	CDKN2B-AS1	LncRNA disease
10	HTTAS	Unconfirmed

Case study

In order to further verify the independent prediction performance of MM-LDA, gastric cancer was selected as the target for the case study. All known associations relating to gastric cancer composed the training set, and unknown associations composed the testing set. Then, gastric cancer-related lncRNAs identified by MM-LDA were sorted by scores. The top 10 lncRNAs with the highest scores were selected to validate the predictive performance of MM-LDA, with the evidence coming from relevant literature and database, as shown in Table 2.

In Table 2, all but two out of 10 lncRNAs predicted by MM-LDA have found evidence from relevant literature and database. Even though, there is no direct evidence showing that HOTAIR and HTTAS relate to gastric cancer so far, some studies found that HOTAIR has stable expression in peripheral blood and can be used as a non-invasive diagnostic marker for gastric cancer (Dong et al., 2019). There is also no published literature which finds the association between HTTAS and gastric cancer. We firmly believe that there will be some researchers to find the experimental evidence for this association inferred by MM-LDA.

Discussion

In this study, a new lncRNA–disease association prediction model, namely, MM-LDA, combining the graph attention network and inductive matrix completion technology was established. MM-LDA designed a multiple-operator aggregation in n-heads attention to enhance the features of nodes. The enhanced features were input into the whole process of induction matrix completion, and the original association matrix was reconstructed by completing the missing elements of the matrix. The results from 5-fold cross-validation showed that MM-LDA obtained the best AUC value and AUPR value compared with the other three state-of-the-art computational models. Comparing with GAMCLDA, 6.45% of training time was saved. In general, MM-LDA deserves to be recommended as the highest cost-effective prediction model. However, there are still some aspects that need to be further improved and studied. First, more biological information relating to lncRNAs and diseases should be effectively integrated. Second, MM-LDA did not predict the associations relating to new lncRNAs and isolated diseases because we could not capture the features of new lncRNAs and isolated diseases without known associations. Third, we should continue to optimize the aggregators by considering the research progress of association prediction in other fields.

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

Conceptualization, YZ; data curation, YW; formal analysis, XL; funding acquisition, YZ; methodology, YZ; software, YW; validation, YL and MC; writing—original draft, YZ; writing—review and editing, YZ and YW.

Funding

This research was funded by the National Natural Science Foundation of China (Grant Nos. 62166014 and 62162019), with funder YZ, and the Natural Science Foundation of Guangxi Province (Grant No. 2020GXNSFAA297255), with funder YZ.

Acknowledgments

The authors thank the reviewers for their suggestions that helped improve the manuscript substantially.

Conflict of interest

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

Publisher's note

References

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

Bian, C., Lei, X., and Wu, F. (2021). Gatcda: Predicting circRNA-disease associations based on graph attention network. *Cancers* 13, 2595. doi:10.3390/cancers13112595

Chen, L., and Chen, S. (2017). Survey on matrix completion models and algorithms. *J. Softw.* 28, 1547–1564.

Chen, L., Shi, H., Wang, Z., Hu, Y., Yang, H., Zhou, C., et al. (2016). IntNetLncSim: An integrative network analysis method to infer human lncRNA functional similarity. *Oncotarget* 7, 47864–47874. doi:10.18632/oncotarget.10012

Chen, X. (2015). Katzlda: KATZ measure for the lncRNA-disease association prediction. *Sci. Rep.* 5, 16840–16850. doi:10.1038/srep16840

Chen, X., Clarence Yan, C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338. doi:10.1038/srep11338

Chen, X., Sun, L., and Zhao, Y. (2021). Ncmcmda: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief. Bioinform.* 22, 485–496. doi:10.1093/bib/bbz159

Chen, X., Sun, Y., Guan, N., Qu, J., Huang, Z., Zhu, Z., et al. (2019). Computational models for lncRNA function prediction and functional similarity calculation. *Brief. Funct. Genomics* 18, 58–82. doi:10.1093/bfgp/ely031

Chen, X., Wang, C., and Guan, N. (2020). Computational models in non-coding RNA and human disease. *Int. J. Mol. Sci.* 21, 1557. doi:10.3390/ijms21051557

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

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

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

Dong, X., He, X., Guan, A., Huang, W., Jia, H., Huang, Y., et al. (2019). Long non-coding RNA Hotair promotes gastric cancer progression via miR-217-GPC5 axis. *Life Sci.* 217, 271–282. doi:10.1016/j.lfs.2018.12.024

Fraidouni, N., and Zaruba, G. (2019). The steering committee of the world congress in computer science. Computer Engineering and Applied Computing (WorldComp), 61–66.A matrix completion approach for predicting lncRNA-disease association, Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), Athens, Greece.

Gu, Y., Zhang, B., Zheng, S., Yang, F., and Li, J. (2021). Building A drug ADMET classification prediction model based on graph attention network. *Data Anal. Knowl. Discov.* 1.

Huang, L., Li, X., Guo, P., Yao, Y., Liao, B., Zhang, W., et al. (2017). Matrix completion with side information and its applications in predicting the antigenicity of influenza viruses. *Bioinformatics* 33, 3195–3201. doi:10.1093/bioinformatics/btx390

Huang, L., Zhang, L., and Chen, X. (2022a). Updated review of advances in microRNAs and complex diseases: Experimental results, databases, webservers and data fusion. *Brief. Bioinform.*, bbac397. doi:10.1093/bib/bbac397

Huang, L., Zhang, L., and Chen, X. (2022b). Updated review of advances in microRNAs and complex diseases: Taxonomy, trends and challenges of computational models. *Brief. Bioinform.* 23, bbac358. doi:10.1093/bib/bbac358

Huang, L., Zhang, L., and Chen, X. (2022c). Updated review of advances in microRNAs and complex diseases: Towards systematic evaluation of computational models. *Brief. Bioinform.*, bbac407. bbac407. doi:10.1093/bib/bbac407

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980.

Long, Y., Luo, J., Zhang, Y., and Xia, Y. (2021). Predicting human microbe–disease associations via graph attention networks with inductive matrix completion. *Brief. Bioinform.* 22, bbaa146. doi:10.1093/bib/bbaa146

Lu, C., Yang, M., Li, M., Li, Y., Wu, F., and Wang, J. (2019). Predicting human lncRNA-disease associations based on geometric matrix completion. *IEEE J. Biomed. Health Inf.* 24, 2420–2429. doi:10.1109/JBHI.2019.2958389

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

Lydia, A., and Francis, S. (2019). Adagrad—An optimizer for stochastic gradient descent. *Int. J. Inf. Comput. Sci.* 6, 566–568.

Ma, Y., Guo, X., and Sun, Y. (2019). Prediction of disease associated long non-coding RNA based on HeteSim. *Comput. Res. Dev.* 56, 1889–1896.

Natarajan, N., and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* 30, i60–i68. doi:10.1093/bioinformatics/btu269

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

Sun, X., Zheng, H., and Sui, N. (2018). Regulation mechanism of long non-coding RNA in plant response to stress. *Biochem. Biophys. Res. Commun.* 503, 402–407. doi:10.1016/j.bbrc.2018.07.072

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, C., Han, C., Zhao, Q., and Chen, X. (2021). Circular RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 22, bbab286. doi:10.1093/bib/bbab286

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

Wu, Q., Cao, R., Xia, J., Ni, J., Zheng, C., and Su, Y. (2021). Extra trees method for predicting LncRNA-disease association based on multi-layer graph embedding aggregation. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1. doi:10.1109/TCBB.2021.3113122

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4–24. doi:10.1109/TNNLS.2020.2978386

Xia, T., Xiao, B., and Guo, J. (2013). Acting mechanisms and research methods of long noncoding RNAs. *Yi Chuan= Hered.* 35, 269–280. doi:10.3724/sp.j.1005.2013.00269

Xu, D., Zhang, S., Zhang, H., and Mandic, D. P. (2021). Convergence of the RMSProp deep learning method with penalty for nonconvex optimization. *Neural Netw.* 139, 17–23. doi:10.1016/j.neunet.2021.02.011

Yin, M., Liu, J., Gao, Y., Kong, X., and Zheng, C. (2020). Ncplp: A novel approach for predicting microbe-associated diseases with network consistency projection and label propagation. *IEEE Trans. Cybern.* 52, 5079–5087. doi:10.1109/TCYB.2020.3026652

Zeng, M., Lu, C., Fei, Z., Wu, F., Li, Y., Wang, J., et al. (2020a). Dmflda: A deep learning framework for predicting lncRNA–disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi:10.1109/TCBB.2020.2983958

Zeng, M., Lu, C., Zhang, F., Li, Y., Wu, F., Li, Y., et al. (2020b). Sdlda: lncRNA-disease association prediction based on singular value decomposition and deep learning. *Methods* 179, 73–80. doi:10.1016/j.ymeth.2020.05.002

Zhao, T., Xu, J., Liu, L., Bai, J., Xu, C., Xiao, Y., et al. (2015). Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Mol. Biosyst.* 11, 126–136. doi:10.1039/c4mb00478g

Zhou, M., Wang, X., Li, J., Hao, D., Wang, Z., Shi, H., et al. (2015). Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. *Mol. Biosyst.* 11, 760–769. doi:10.1039/c4mb00511b

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 21 November 2022
DOI 10.3389/fgene.2022.1056691

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Olivia J. Veatch,
University of Kansas Medical Center, United States
Feng Yuan,
The University of Iowa, United States

*CORRESPONDENCE
Dong Wang,
binyiwangdong@163.com
Xin Zhang,
walterzhangx@139.com

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 29 September 2022
ACCEPTED 31 October 2022
PUBLISHED 21 November 2022

CITATION
Liu Q, Hao T, Li L, Huang D, Lin Z, Fang Y, Wang D and Zhang X (2022), Construction of a mitochondrial dysfunction related signature of diagnosed model to obstructive sleep apnea.
Front. Genet. 13:1056691.
doi: 10.3389/fgene.2022.1056691

Construction of a mitochondrial dysfunction related signature of diagnosed model to obstructive sleep apnea

Qian Liu[1,2], Tao Hao[3], Lei Li[2], Daqi Huang[2], Ze Lin[1,4], Yipeng Fang[4], Dong Wang[2*†] and Xin Zhang[1,4,5*†]

[1]Shantou University Medical College, Shantou, China, [2]Department of Cardiology, The Affiliated Hospital of Binzhou Medical University, Binzhou, Shandong Province, China, [3]Department of General Surgery, The First Affiliated Hospital of Jinan University, Guangzhou, China, [4]Laboratory of Molecular Cardiology, The First Affiliated Hospital of Shantou University Medical College, Shantou, China, [5]Laboratory of Medical Molecular Imaging, The First Affiliated Hospital of Shantou University Medical College, Shantou, China

Background: The molecular mechanisms underlying obstructive sleep apnea (OSA) and its comorbidities may involve mitochondrial dysfunction. However, very little is known about the relationships between mitochondrial dysfunction-related genes and OSA.

Methods: Mitochondrial dysfunction-related differentially expressed genes (DEGs) between OSA and control adipose tissue samples were identified using data from the Gene Expression Omnibus database and information on mitochondrial dysfunction-related genes from the GeneCards database. A mitochondrial dysfunction-related signature of diagnostic model was established using least absolute shrinkage and selection operator Cox regression and then verified. Additionally, consensus clustering algorithms were used to conduct an unsupervised cluster analysis. A protein–protein interaction network of the DEGs between the mitochondrial dysfunction-related clusters was constructed using STRING database and the hub genes were identified. Functional analyses, including Gene Ontology (GO) analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis, gene set enrichment analysis (GSEA), and gene set variation analysis (GSVA), were conducted to explore the mechanisms involved in mitochondrial dysfunction in OSA. Immune cell infiltration analyses were conducted using CIBERSORT and single-sample GSEA (ssGSEA).

Results: we established mitochondrial dysfunction related four-gene signature of diagnostic model consisted of *NPR3, PDIA3, SLPI, ERAP2*, and which could easily distinguish between OSA patients and controls. In addition, based on mitochondrial dysfunction-related gene expression, we identified two clusters among all the samples and three clusters among the OSA samples. A total of 10 hub genes were selected from the PPI network of DEGs between the two mitochondrial dysfunction-related clusters. There were correlations between the 10 hub genes and the 4 diagnostic genes. Enrichment analyses suggested that autophagy, inflammation pathways, and immune pathways are crucial in mitochondrial dysfunction in OSA. Plasma cells and M0 and M1 macrophages

were significantly different between the OSA and control samples, while several immune cell types, especially T cells (γ/δ T cells, natural killer T cells, regulatory T cells, and type 17 T helper cells), were significantly different among mitochondrial dysfunction-related clusters of OSA samples.

Conclusion: A novel mitochondrial dysfunction-related four-gen signature of diagnostic model was built. The genes are potential biomarkers for OSA and may play important roles in the development of OSA complications.

KEYWORDS

mitochondrial dysfunction, obstructive sleep apnea, immunocyte infiltration, bioinformatic analysis, gene signature

1 Introduction

Obstructive sleep apnea syndrome (OSA), a growing health concern that affects nearly one billion people worldwide, is an independent risk factor for cardiovascular and metabolic diseases, but is highly underdiagnosed (Arnaud et al., 2020). Continuous positive airway pressure is highly effective at improving symptoms but cannot reduce the occurrence of comorbidities. The use of biomarkers has been strongly recommended, as the condition often goes undiagnosed because patients remain oblivious to the severity of OSA and its complications (Wang et al., 2022). Therefore, it is an urgent task to find indicators for early diagnosis of OSA and decipher the molecular pathways involved in OSA and its complications in order to ensure earlier treatment and prevent complications.

The physiologic changes in OSA are vast and involve complex mechanisms which play a role in the pathogenesis of cardiovascular and metabolic disorders. Chronic intermittent hypoxia (CIH) is the most deleterious feature of OSA, as it can lead to oxidative damage in every organ (Shan et al., 2007). CIH can suppress mitochondrial function and lead to the generation of reactive oxygen species (ROS) (Wang et al., 2010; Huang et al., 2014; Zhao et al., 2019; Lin et al., 2021; Song et al., 2022). As mitochondrial status is important for the metabolic function of all organs, mitochondrial dysfunction at the cellular level that can affect systemic metabolic balance can significantly contribute to many diseases and have been defined as classical a hallmark of many diseases (Srinivasan et al., 2017; Chapman et al., 2019). Mitochondrial dysfunction is a basic mechanism in inflammation-related non-communicable diseases (Hernandez-Aguilera et al., 2013). The wide acceptance of mitochondrial dysfunction as a correlated factor of Parkinson's disease (Rocha et al., 2018), cardiovascular diseases (Vásquez-Trincado et al., 2016), diabetic kidney disease (Wei and Szeto 2019) and other numerous diseases (Kasapoğlu and Seli 2020; Yapa et al., 2021) has led to the presupposition that mitochondrial dysfunction markers are associated with OSA. Although various microRNAs and proteins (and their genes) have been reported to be involved in OSA (Li et al., 2017; Cao et al., 2021; Shi et al., 2021; Tang et al., 2021), the effects of OSA on genes and pathways remain largely unknown, especially regarding mitochondrial dysfunction.

Previous studies have suggested that mitochondrial dysfunction represents the molecular mechanism underlying OSA and its comorbidities. First, sleep disorders are prevalent in individuals with mitochondrial disorders; the clinical features of the mitochondrial dysfunction affect the type of sleep disturbance (Brunetti et al., 2021). Second, mitochondrial DNA (mtDNA) copy number is significantly reduced in patients with OSA, and it is a reliable biomarker for predicting cardiovascular risk in patients with OSA (Kim et al., 2014). Third, Banxia-Houpu decoction reduced CIH-induced heart damage by regulating mitochondrial function (Song et al., 2022). Fourth, attenuating mitochondria-dependent apoptosis has been suggested as a novel adjunct strategy for ameliorating OSA-induced neurocognitive impairment (Xu et al., 2021). Fifth, mitochondrial dysfunction and the oxidative stress were found to be involved in genioglossus muscle injuries in OSA with obesity, which may provide therapeutic targets for use in OSA with obesity (Chen et al., 2021). Lastly, research has identified certain proteins associated with CIH, and some may serve as novel biomarkers for OSA and related disorders, such as acute coronary syndrome (Shi et al., 2021) and Alzheimer's disease (Wu et al., 2021). In conclusion, patients with OSA exhibit several mitochondrial gene mutations, deletions, and some mitochondrial dysfunction indexes. Unfortunately, little has been reported whether mitochondrial dysfunction related genes and pathways could be used as clinical biomarkers of OSA susceptibility and severity so far.

In the present study, a four-gene (*NPR3, PDIA3, SLPI,* and *ERAP2*) diagnostic model was built to diagnose OSA based on mitochondrial dysfunction-related gene expression. The genes are potential biomarkers and therapeutic targets for use in OSA. Consensus clustering of all the samples (OSA and control) was used to identify two mitochondrial dysfunction-related clusters (A and B). Furthermore, to investigate the underlying biological functions of the clusters, we identified 106 differentially expressed genes (DEGs) between clusters A and B and conducted functional enrichment analyses of these DEGs. A protein–protein interaction (PPI) network of the DEGs was

constructed using the STRING database. Thereafter, immune cell infiltration was evaluated using both CIBERSORT and single-sample gene set enrichment analysis (ssGSEA). In addition, the correlations between the four diagnostic genes and immune cell infiltration were calculated. To our knowledge, this is the first study to integrate bioinformatics analyses in order to identify the key mitochondrial dysfunction-related genes and pathways, and the degree of immune cell infiltration, in OSA. These genes and pathways may facilitate our understanding of the molecular mechanism of OSA and further provide evidence for early diagnosis, prevention, and treatment of this disease.

2 Methods

2.1 Data sources and processing

Two microarray datasets [GSE135917 (Gharib et al., 2020) and GSE38792 (Gharib et al., 2013)] on adipose tissue samples from OSA patients and controls were downloaded from the Gene Expression Omnibus (GEO) database. The sequencing platform was GPL96 (HG-U133A)] Affymetrix (Supplementary Table S1). Data of 58 OSA patients and 8 controls from GSE135917, and 10 OSA patients and 8 controls from GSE38792, were analyzed in our study. The datasets were log2 transformed and normalized using the SVA R package. The expression distribution before and after normalization was visualized using boxplots (Supplementary Figure S1).

2.2 Analysis of differentially expressed genes and mitochondrial dysfunction-related genes

The limma R package (Ritchie et al., 2015) was used to conduct a DEG analysis comparing the OSA and control samples. The genes with |log(fold change [FC])|>1 and $p < 0.05$ were identified as DEGs. The RCircos R package (Zhang et al., 2013) was used to map the chromosomal location of the genes.

A total of 8334 mitochondrial dysfunction-related genes were then downloaded from the GeneCards database (https://www.genecards.org/) (Safran et al., 2010) using the key words *"mitochondrial dysfunction"*. The mitochondrial dysfunction-related DEGs were then identified.

2.3 Correlation analysis among genes

Pearson correlations between pairs of genes were calculated. The GGplot2 R package was used to construct scatter plots of the expression correlations between pairs of genes that met the criteria and to fit correlation curves. The criteria for significant correlation comprised absolute correlation coefficient value >0.5 and $p < 0.05$.

2.4 Establishment of diagnostic model

Least absolute shrinkage and selection operator (LASSO) Cox regression was used for feature selection and dimensionality reduction in order to generate a gene-based classifier [9]. To verify the diagnostic value of the model, ROC curves of the single genes and the four-gene model were plotted using R package pROC (Robin et al., 2011). A nomogram and decision curve analysis (DCA) curves were used for validation.

2.5 Consensus clustering

Using all the OSA and control samples, a consensus clustering analysis of mitochondrial dysfunction-related genes was used to identify distinct mitochondrial dysfunction-related clusters using the k-means clustering algorithm (Sabah et al., 2021). The optimum number of clusters, along with the consistency of clusters, was determined by the consensus clustering algorithm in the ConsensusClusterPlus package (Seiler et al., 2010). A total of 1000 iterations were performed to ensure the stability of the categories. Additionally, using only the OSA samples, consensus clustering was again used to identify distinct mitochondrial dysfunction-related clusters.

2.6 Protein–protein interaction network construction

After determining the DEGs between the mitochondrial dysfunction-related clusters (based on all samples), a PPI network of the DEGs was constructed using STRING network version 11.0 and the default confidence threshold of 0.4. The PPI network was exported and then Cytoscape version 3.8.0 was used to calculate the network attributes of each node. Next, cytoHubba version 1.6 was used to identify hub nodes based on the degree of the nodes.

We predicted the miRNAs and transcription factors related to the hub genes using TarBase (Wang et al., 2013) and miRecords (Fornes et al., 2020). Protein-chemical interactions were obtained from the Comparative Toxicogenomics Database (http://ctdbase.org/) (Davis et al., 2021).

2.7 Functional enrichment analyses of differentially expressed genes

Using the DEGs between the mitochondrial dysfunction-related clusters (based on all samples), Gene Ontology (GO)

enrichment analysis was employed to study the large-scale functional enrichment of the DEGs at three levels: biological process (BP), molecular function (MF) and cellular component (CC) (Ashburner et al., 2000). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was used to identify the biological pathways related to the DEGs (Kanehisa and Goto 2000). The clusterProfiler R package (Wu et al., 2021) was used to perform GO functional annotation for all significant DEGs to identify significantly enriched GO terms. The enrichment results were visualized using the GOplot R package.

Gene set enrichment analysis (GSEA) using data from MSigDB (Liberzon et al., 2015) was employed to identify the significant differences in biological pathways between the high- and low-expression clusters. The "C2.cp.kegg.v7.4.entrez.gmt" (KEGG pathways) gene set was selected as the reference gene set.

Gene set variation analysis (GSVA) using the "c2.cp.kegg.v7.2.symbols.gmt" (KEGG pathways) and "h.all.v7.2.symbols.gmt" (Hallmark pathways) gene sets (Hanzelmann et al., 2013) was employed to identify the significant differences in biological pathways between the mitochondrial dysfunction-related clusters. The criteria for significant enrichment comprised nominal $p < 0.05$, normalized enrichment score (NES) > 1, and false discovery rate (FDR) $q < 0.25$ using the GSVA R package.

2.8 Analysis of immune cell infiltration

The degree of immune cell infiltration was assessed twice, using 1) CIBERSORT and 2) ssGSEA. First, the CIBERSORT R package was used to determine the degree of immune cell infiltration based on the CIBERSORT scores for immune infiltrating cells (Steen et al., 2020). Second, the GSVA R package method based on ssGSEA (Huang et al., 2021) was used to evaluate the degree of immune cell infiltration.

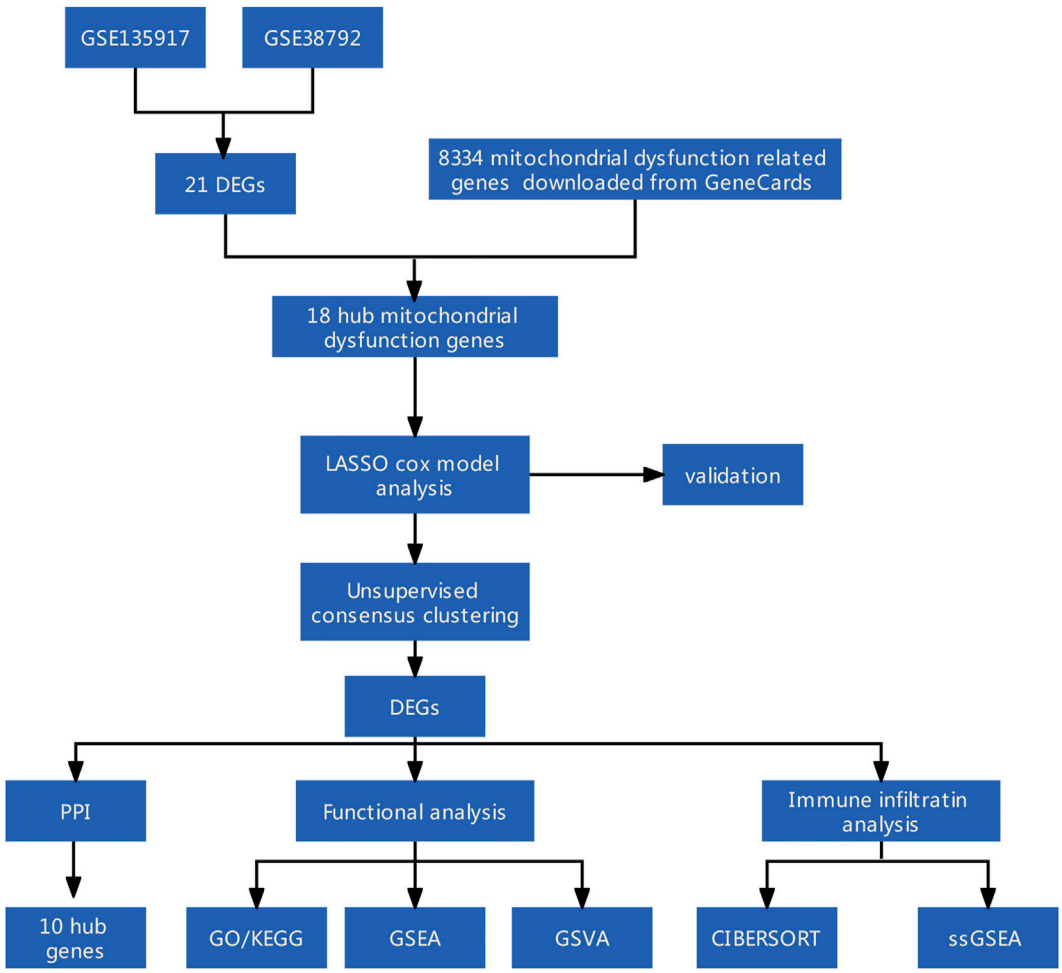


FIGURE 1
Flow diagram of methodologies used in this study.

2.9 Statistical analysis

All data processing and analysis were completed in R software (version 4.0.2). Normally distributed continuous variables were compared using independent-samples Student's t tests, and non-normally distributed continuous variables were compared using Mann–Whitney U tests (i.e., Wilcoxon rank-sum tests). Categorical variables were compared using chi-square tests or Fisher's exact tests. Kruskal–Wallis tests were used for comparison of more than two groups. Two-tailed $p < 0.05$ was considered statistically significant.

3 Results

Figure 1 shows the study flowchart.

3.1 Identification of mitochondrial dysfunction-related differentially expressed genes

First, heatmaps and volcano plots were used to visualize the DEGs between the OSA and control samples in the GSE135917 (Supplementary Figures S2A,B), GSE132651 (Supplementary Figures S2C,D), and combined datasets (Supplementary Figures S2E,F). Next, we analyzed the intersection of DEGs among the GSE135917, GSE38792, and combined datasets, as displayed in a Venn diagram (Figure 2A). A total of 21 overlapping genes were obtained. Moreover, 18 of the overlapping genes were mitochondrial dysfunction-related genes (Figure 2B), which were designated as the mitochondrial dysfunction-related hub genes. Their expression levels in the GSE38792 and GSE135917 datasets are presented in boxplots (Figures 2C,D). Figure 2E shows the chromosomal positions of the mitochondrial dysfunction-related hub genes.

3.2 Diagnostic model based on mitochondrial dysfunction-related hub genes

The 18 mitochondrial dysfunction-related hub genes were subjected to LASSO Cox regression to create a diagnostic model (Figure 3A). Four genes were gathered, the regression model reached the optimal ability (Figure 3B). A plot of the diagnostic genes was used to visualize their differential effectiveness for diagnosing OSA (Figure 3C). The calibration curve regarding the nomogram predictions (Figure 3D) and decision curve analysis curve predicted by irrelevant nomogram (Figure 3E) were constructed. Both showed that 4– gene diagnostic model had good predictive value.

3.3 Verification of diagnostic value of four-gene diagnostic model

Boxplots of the four genes (*NPR3, PDIA3, SLPI,* and *ERAP2*) in the diagnostic model, as OSA-related risk genes, had significant differences in expression between the OSA and control samples in the GSE135917, GSE38792, and combined datasets (Figures 4A–C).

The area under the ROC curve (AUC) was calculated to measure the diagnostic value of the model. Figures 4D–F show the ROC curves of *NPR3, PDIA3, SLPI, and ERAP2* in the GSE135917, GSE38792, and combined datasets, respectively. Figures 4G–I show the ROC curves of the four-gene diagnostic model in the GSE135917, GSE38792 and combined datasets, respectively. The results indicated that the four-gene signature of diagnostic model had high diagnostic value.

3.4 Mitochondrial dysfunction-related clusters

To explore biological characteristics related to the expression of mitochondrial dysfunction-related genes, all the samples were first divided into k (k = 2, 3, 4, 5, 6, 7, and 8) clusters using ConsensusClusterPlus. The optimal categorization occurred when k = 2, based on the cumulative distribution function (CDF) curves of the consensus score. Therefore, the samples were divided into two mitochondrial dysfunction-related clusters: cluster A ($n = 28$) and cluster B ($n = 56$) (Figures 5A–D).

3.5 Transcription factor and miRNA predictions

To explore the interactions related to four-gene diagnostic model at the post-transcriptional level, 41 transcription factors that upregulate the genes and 134 miRNAs that target the genes were identified (Supplementary Figure S3A), along with 104 protein chemical components (Supplementary Figure S3B).

3.6 Protein–protein interaction network

To investigate the underlying biological functions of the mitochondrial dysfunction-related clusters A and B, we identified 106 DEGs between clusters A and B (Figure 6A). We then constructed a PPI network using the STRING database (Figure 6B). The highly connected (hub) genes in the PPI network were identified using the MCODE plug-in in Cytoscape (Figure 6C). The top 10 genes, based on high scores using the cytoHubba plug-in in Cytoscape, were also selected (Figure 6D). A Venn diagram was used to identify the intersection of the results of the MCODE and cytoHubba methods, which led

to 10 hub genes being obtained (Figure 6E). A correlation network diagram (Supplementary Figure S4A), scatter plots (Supplementary Figures S4B–P) and Supplementary Table S2 show the correlations of the 10 hub genes and 4 genes in the diagnostic model, revealing strong correlations between the hub genes and diagnostic genes. This indicated that they may act in synergistic way, contributing to OSA and related complications.

3.7 Functional enrichment analyses

Functional enrichment analyses of the 106 DEGs between clusters A and B were performed (Figure 7). The GO analysis indicated that the genes were significantly enriched in cytoplasmic vesicle lumen, chemokine activity, collagen-containing extracellular matrix, regulation of smooth muscle cell proliferation, DNA binding, and transcription activator activity (Supplementary Table S3). The KEGG analysis showed that the genes were enriched in cytokine and cytokine receptor, interleukin (IL)-17 signaling pathway, tumor necrosis factor (TNF) signaling pathway, pathogenic *Escherichia coli* infection, and complement and coagulation cascades (Supplementary Table S4).

Subsequently, GSEA was performed between the high- and low-expression clusters based on the four diagnostic genes in the GSE135917 and GSE38792 datasets (Supplementary Table S5). The results suggested that the samples in the high-expression cluster were significantly enriched in IL-6 pathway, IL-12 pathway, IL6_7 pathway, DNA repair, IL-1 signaling, nonsense-mediated decay, transcriptional regulation of

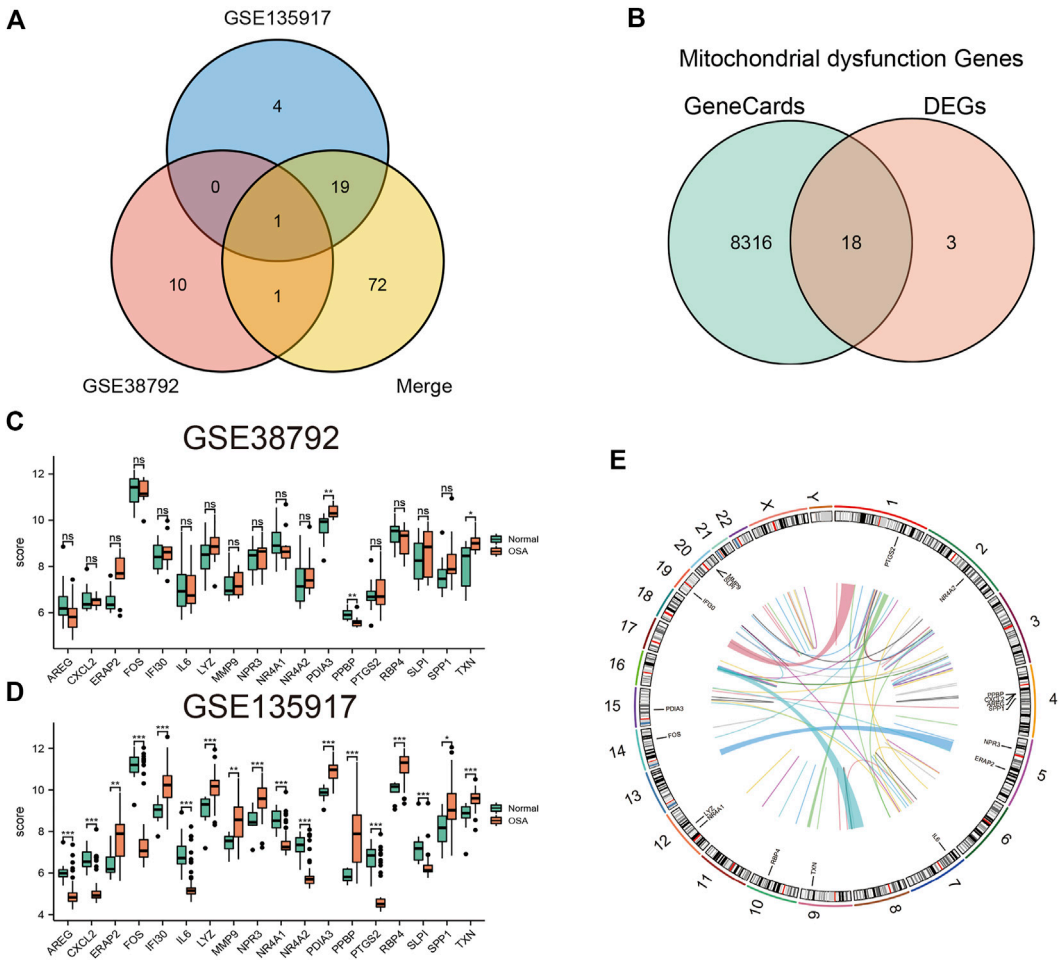


FIGURE 2
Identification of mitochondrial dysfunction-related hub genes. **(A)** Venn diagram of DEGs between the OSA and control samples in the GSE135917, GSE38792, and combined datasets. **(B)** Venn diagram of hub DEGs and mitochondrial dysfunction-related genes. Boxplots of the differences in expression of mitochondrial dysfunction-related hub genes in **(C)** GSE38792 and **(D)** GSE135917 datasets. **(E)** Chromosomal positions and expression of mitochondrial dysfunction-related hub genes. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ns: no significant difference. DEGs: differentially expressed genes.

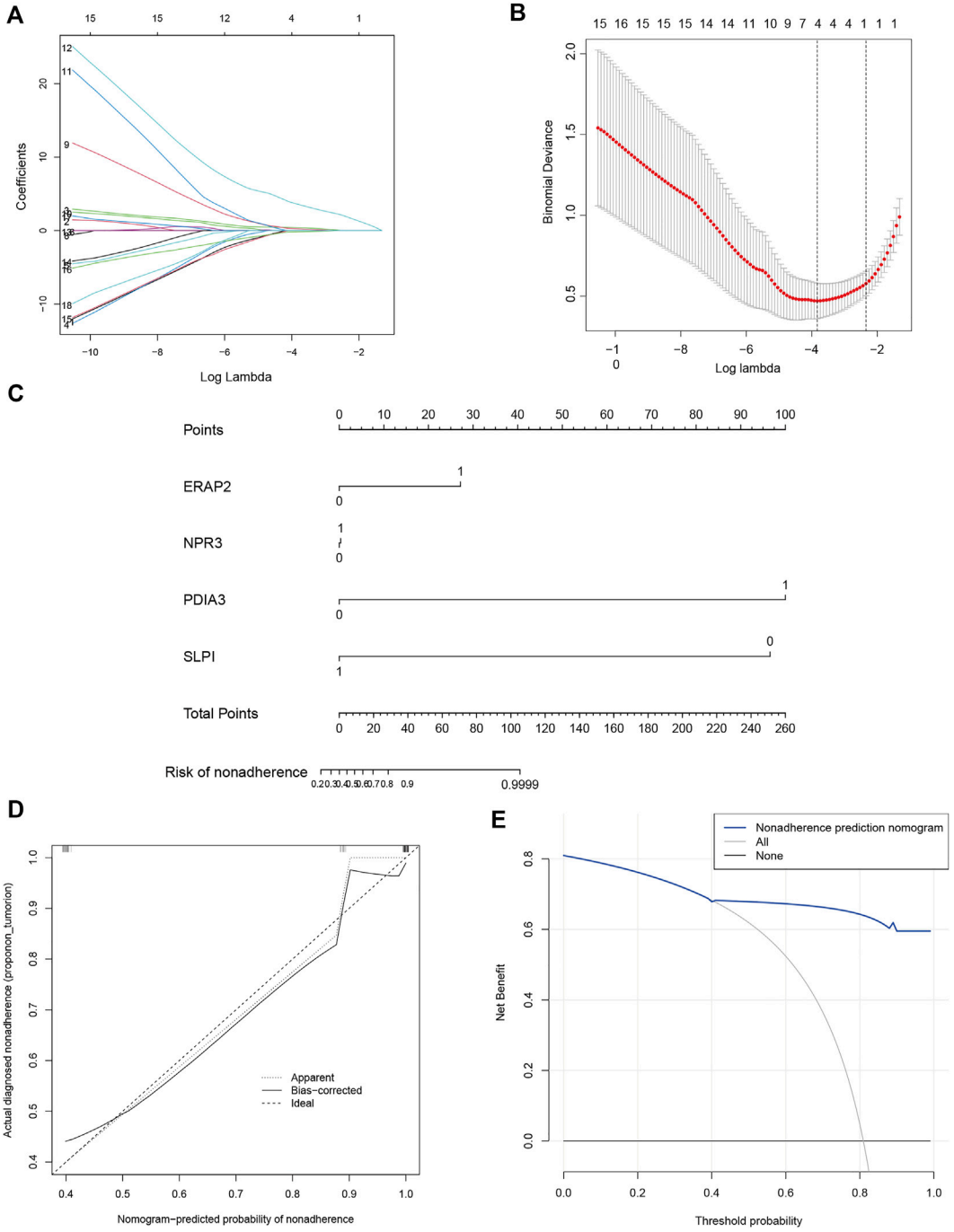


FIGURE 3
Diagnostic model based on mitochondrial dysfunction-related genes. **(A)** Diagnostic model construction using a least absolute shrinkage and selection operator (LASSO) Cox regression model. **(B)** Coefficient distribution plots to select the optimum lambda value. **(C)** Plot of diagnostic genes demonstrating their differential effectiveness for diagnosing OSA. **(D)** Calibration curves based on nomogram predictions and actual observations. **(E)** Decision curve analysis (DCA) of diagnostic model.

pluripotent stem cells, complement activation, and gastrin signaling pathway (Figures 8A–H). The samples in the low-expression cluster were significantly enriched in autophagy, lysosome, proteasome, anaphase promoting complex/cyclosome (APC/C)-mediated degradation of cell cycle proteins, cell cycle check points, metabolism of polyamines,

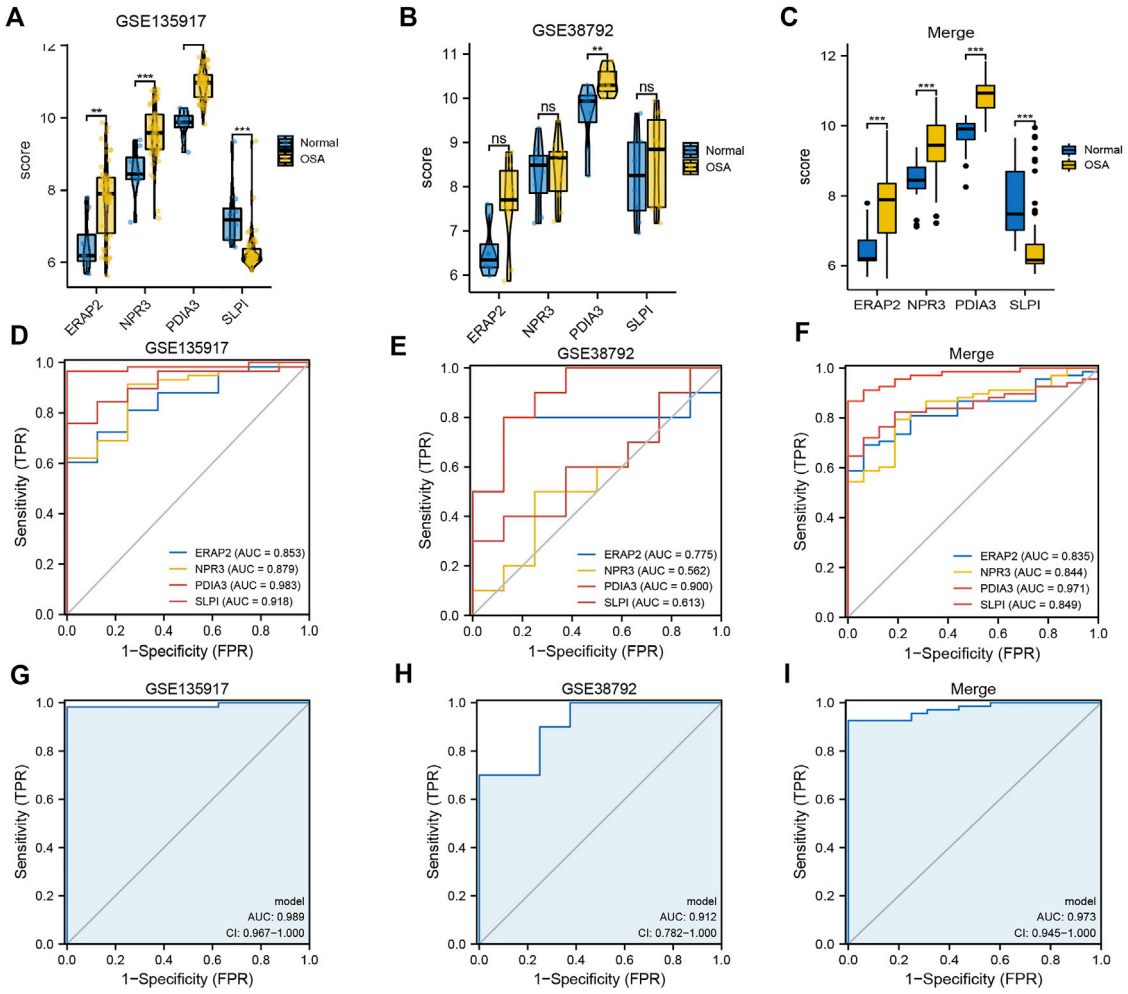


FIGURE 4
Expression differences and diagnostic value of mitochondrial dysfunction-related four-gene diagnostic model. Boxplots of differences in NPR3, PDIA3, SLPI, and ERAP2 expression between OSA and control samples in **(A)** GSE135917, **(B)** GSE38792, and **(C)** combined datasets. ROCcurves of NPR3, PDIA3, SLPI, and ERAP2 in **(D)** GSE135917, **(E)** GSE38792, and **(F)** combined datasets. ROC curves of four-gene diagnostic model in **(G)** GSE135917, **(H)** GSE38792, and **(I)** combined datasets.

and stabilization of P53 (Figures 8I–P). Cytokines exert a vast array of immunoregulatory actions critical to human physiology and disease (Spangler et al., 2015). TNF-α, IL-17, IL-6, IL-12, and IL-1 are inflammatory cytokines. The autophagy–lysosome pathway and ubiquitin–proteasome system are the main mechanisms of intracellular protein degradation and they help to maintain normal cellular functions.

To further investigate the biological pathways that mitochondrial dysfunction may affect, we conducted GSVA between the mitochondrial dysfunction-related clusters A and B to assess pathway enrichment. Regarding the KEGG pathways, most of them, including regulation of pyrimidine metabolism, proteasome, SNARE interactions in vesicular transport, endocytosis, other glycan degradation, amino sugar and nucleotide sugar metabolism, and lysosome, were more enriched in cluster B (Figure 8Q). Regarding the Hallmark pathways, the ROS pathway, heme metabolism, PI3K-AKT-mTOR signaling, mTORC1 signaling, hypoxia, peroxisome, and apoptosis were more enriched in cluster B, whereas myogenesis and KRAS signaling pathways were more enriched in cluster A (Figure 8R). ROS influence metabolic processes such as proteasome function, autophagy, and general inflammatory signaling (Forrester et al., 2018). Heme metabolism influences a wide variety of biological processes relevant to OSA, including redox balance and inflammatory response (Wang et al., 2022). Autophagy, which can be activated by hypoxia, can be beneficial in inflammatory disorders as it eliminates damaged organelles and maintains homeostasis (Yao et al., 2021), and mTOR is a key negative regulator of autophagy. It should be noted that the enriched pathways were mainly autophagy, inflammatory, and

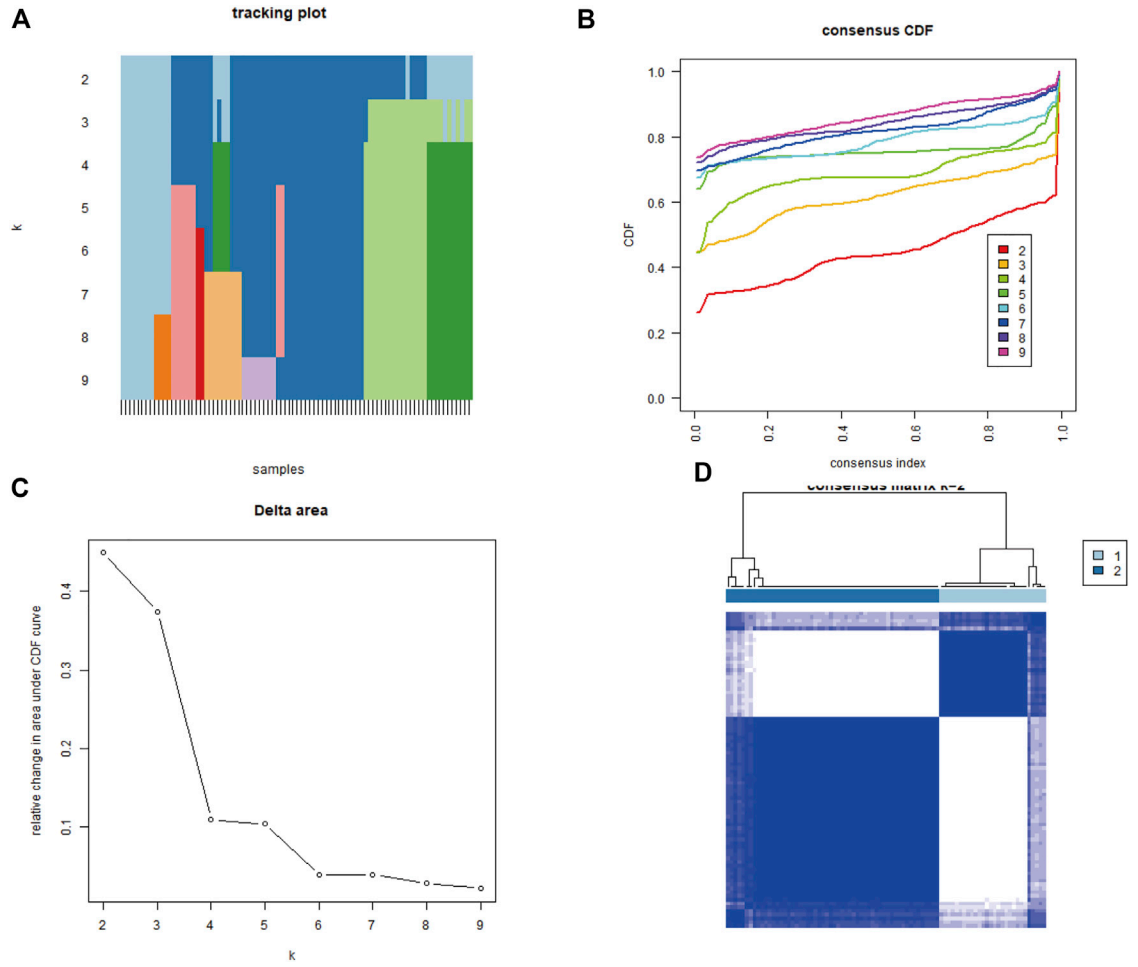


FIGURE 5
Identification of two mitochondrial dysfunction-related clusters using consensus clustering analysis of mitochondrial dysfunction-related genes. **(A)** Tracking plot of consistent clustering. **(B)** Cumulative distribution function (CDF) curves for k = 2–9. **(C)** Elbow plot showing relative change in area under the CDF curve (AUC). **(D)** Consensus clustering matrix for k = 2.

immune related pathways. This indicated that mitochondrial dysfunction might be associated with autophagy, inflammation, and the immune microenvironment in OSA.

3.8 Immune cell infiltration

The CIBERSORT algorithm was used to evaluate the immune microenvironment in OSA. The correlations among immune cells are shown in Figure 9A. A boxplot indicated that the infiltration of several immune cells (plasma cells, M0 macrophages, and M1 macrophages) were significantly different between OSA and control samples (Figure 9B).

To explore the relationship between mitochondrial dysfunction and immune cell infiltration, we compared the immune cell infiltration between clusters A and B. The results showed that several immune cells (activated B cells, CD56^{bright} natural killer cells, eosinophils, macrophages, monocytes, and plasmacytoid dendritic cells) were significantly different between clusters A and B (Figures 9C,D). Figure 9E further visualizes immune cell infiltration differences between clusters.

We then performed consensus clustering of the OSA samples only (n = 68) and identified three mitochondrial dysfunction-related clusters, with 43 samples in cluster A, 8 in cluster B, and 17 in cluster C (Figures 10A–D). We assessed the degree of immune cell infiltration using both CIBERSORT and ssGSEA. Interestingly, the results were consistent with each other. Both methods showed differences among clusters A, B, and C in the infiltration degree of activated B cells, CD56^{bright} natural killer cells, γ/δ T cells, immature dendritic cells, natural killer T cells, regulatory T cells (Tregs), and type 17 T helper cells (Figures 10E,F).

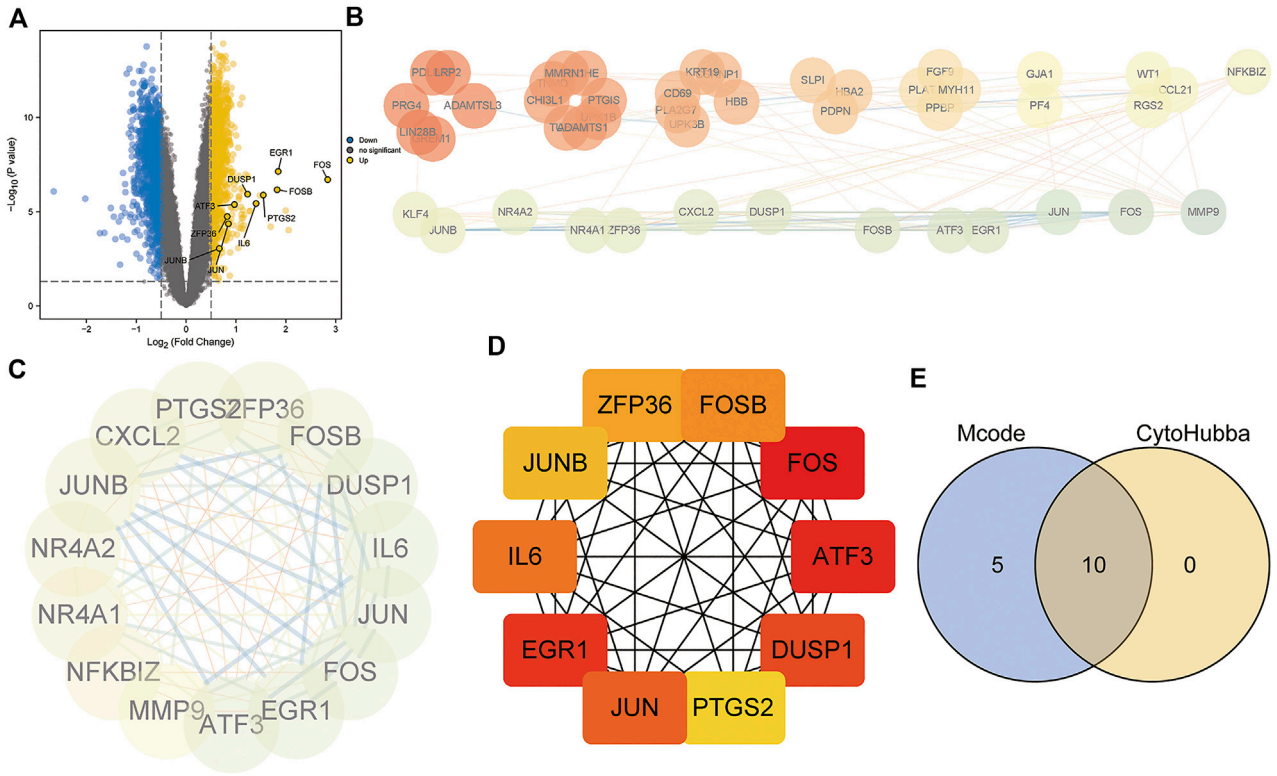


FIGURE 6
Protein–protein interaction (PPI) network construction and hub gene identification. **(A)** Volcano plot of DEGs between mitochondrial dysfunction-related clusters A and B. **(B)** PPI network based on STRING database. **(C)** Hub genes identified by Cytoscape MCODE plug-in. **(D)** Top10 hub genes identified by Cytoscape CytoHubba plug-in. **(E)** Venn diagram showing the intersection of the two methods, identifying 10 hub genes.

To understand the correlations between the genes in the diagnostic model and infiltrating immune cells, we constructed a scatter plot of statistically significant results with correlation coefficient (R) > 0.4. *PDIA3* was correlated with plasma cells, monocytes, M0 macrophages and T cells CD4 memory resting (Supplementary Figures S5A–D). *SLPI* was correlated with M0 macrophages, naive B cells, and plasma cells (Supplementary Figures S5E–G). Taken together, the results indicate that mitochondrial dysfunction plays an important role in immune microenvironment regulation in OSA.

4 Discussion

OSA can cause many complications, such as cardiovascular, metabolic, and neuropsychiatric disorders, pose a major threat to human health (Wang et al., 2022). However, the approaches to the management of OSA are limited due to the incomplete understanding of the underlying molecular mechanisms of OSA.

During respiratory events in OSA patients, intermittent hypoxia together with post-apnea/hypopnea reoxygenation triggers an increase in oxidative stress (Passali et al., 2015). As the major energy-producing organelles, mitochondria, are highly sensitive to hypoxic stress. They can respond dynamically under hypoxia, which can minimize ROS formation and reduce the risk of cell death and tissue damage. However, as a prominent mechanism of mitochondrial dysfunction, abnormal metabolic cues induced by hypoxia can disrupt the dynamic mitochondrial balance. This results in a series of intracellular signaling cascades and apoptosis, followed by the progression of diverse diseases (Wang et al., 2022). Mitochondrial abnormalities may be one of the pathological mechanisms underlying OSA-related cardiac injury, while maintaining the integrity of mitochondria allows the survival of cardiomyocytes under hypoxia. Aged relative to young mouse hearts exhibited maladaptation to CIH because of mitochondrial dysfunction (Wei et al., 2022).

Since the mitochondrial dysfunction appears to be involved in the pathogenesis of OSA and its complications, investigating the role of mitochondrial dysfunction-related genes may provide novel personalized and optimal management strategies for OSA and its comorbidities. In the present study, we identified a mitochondrial dysfunction-related four-gene signature of diagnostic model involving *NPR3, PDIA3, SLPIM,* and *ERAP2*. The model easily

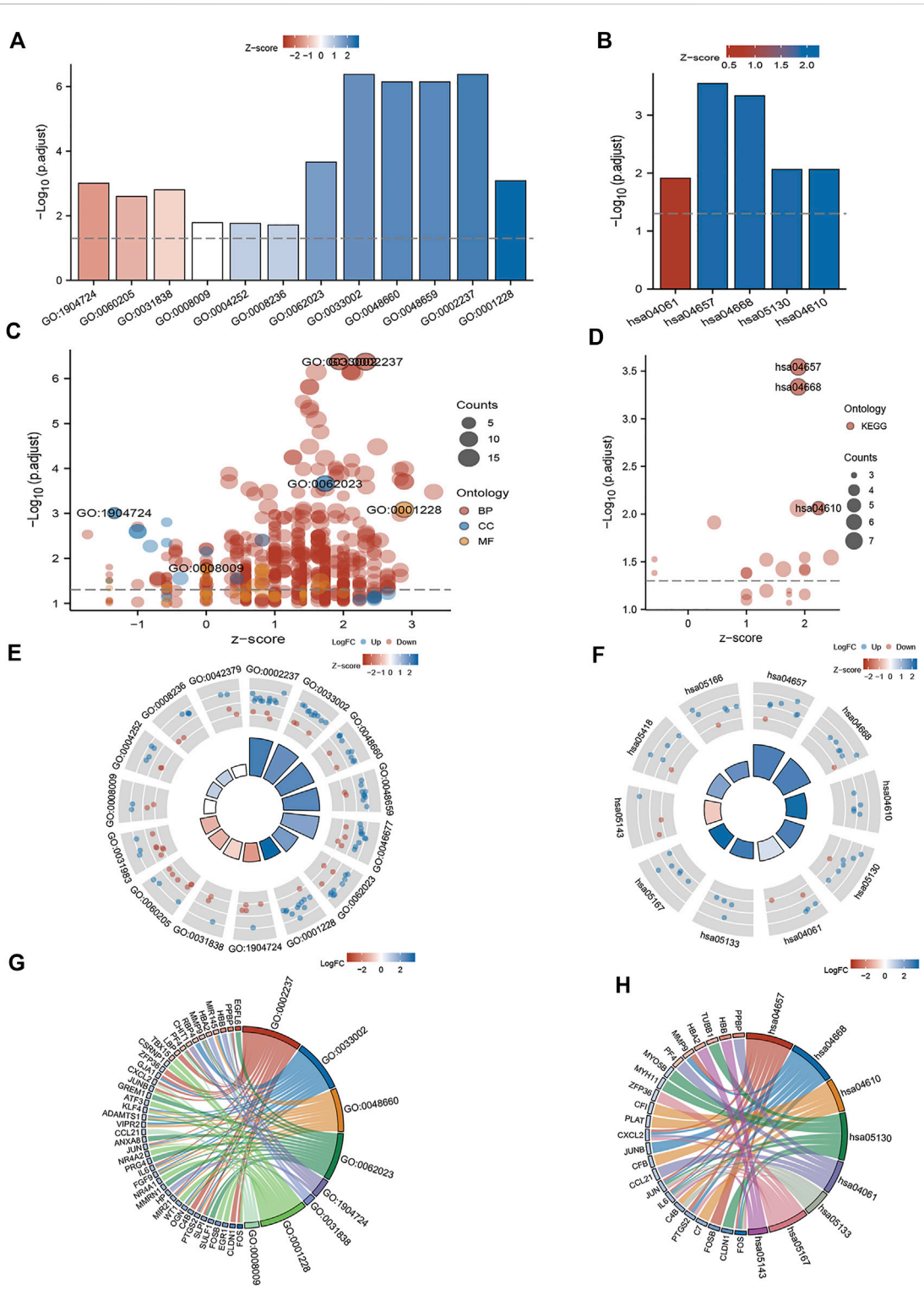


FIGURE 7
GO and KEGG enrichment analyses of DEGs between mitochondrial dysfunction-related clusters A and B. **(A,B)** Histogram, **(C,D)** bubble plot, **(E,F)** circle plot, and **(G,H)** chord diagram of the results of GO and KEGG enrichment analyses.

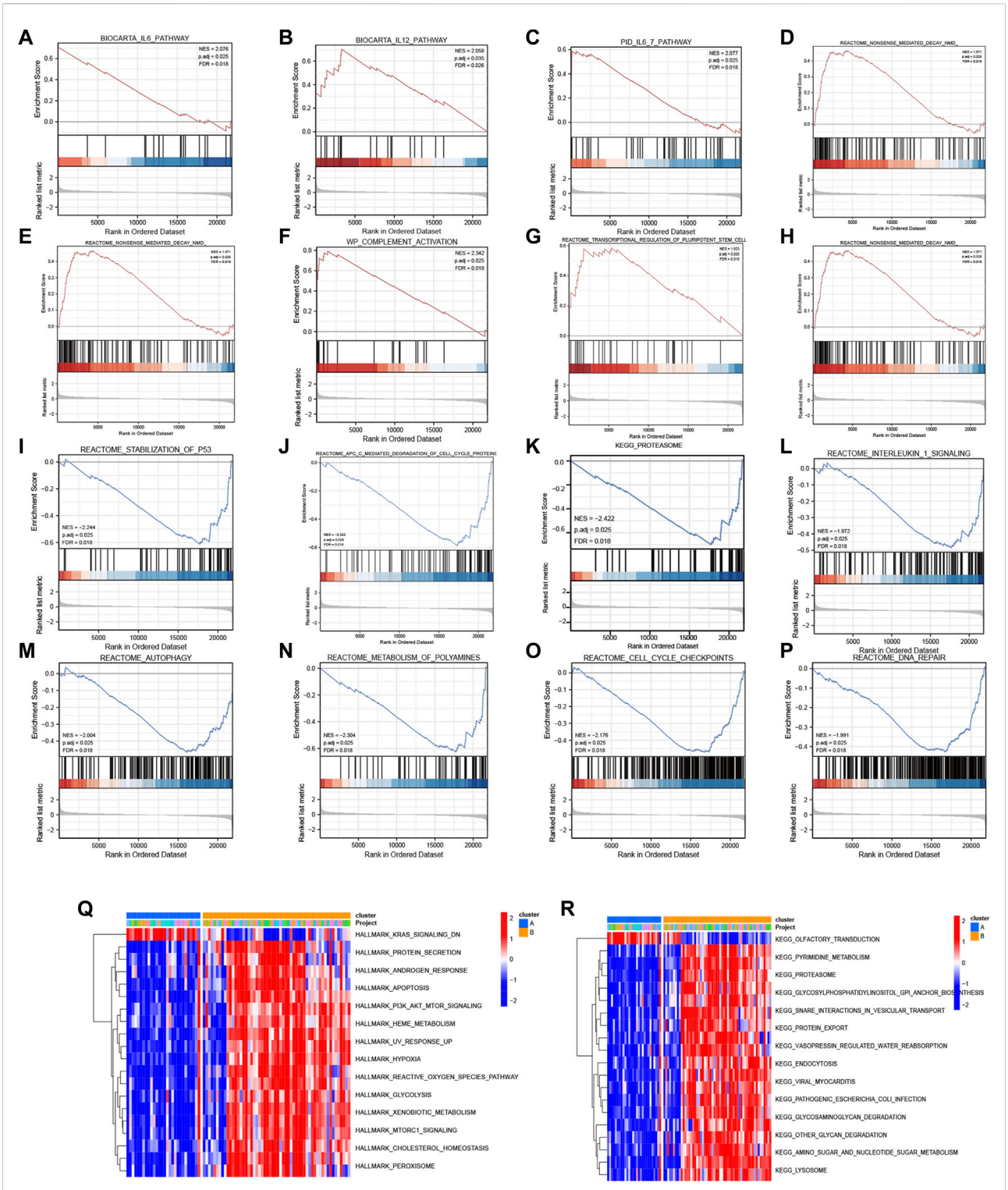


FIGURE 8
GSEA and GSVA. GSEA results: KEGG pathways with significantly differential enrichment between patients with **(A–H)** high and **(I–P)** low expression of the four diagnostic genes. GSVA results: **(Q)** Hallmark and **(R)** KEGG pathways with significantly differential enrichment between mitochondrial dysfunction-related clusters A and B.

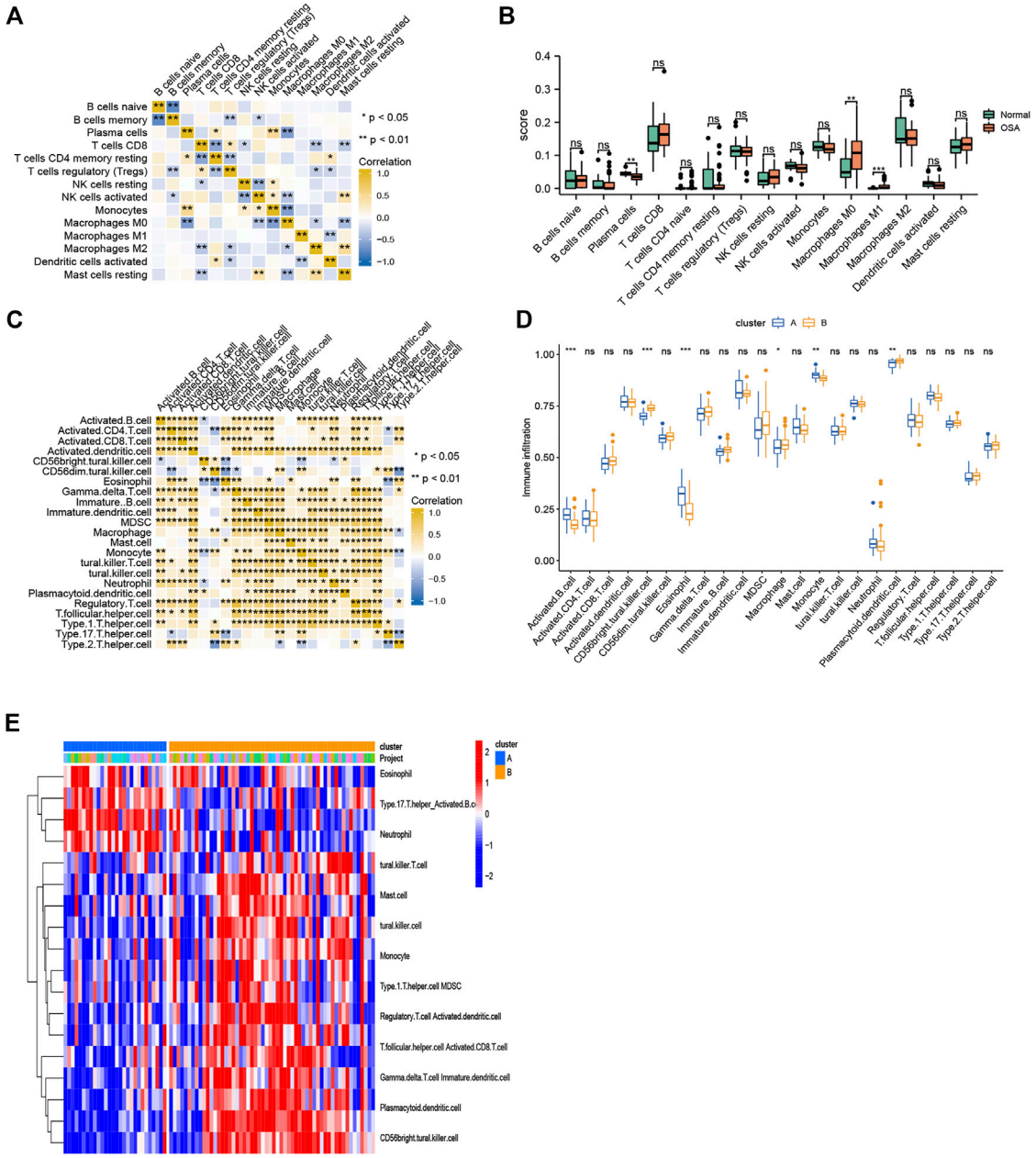


FIGURE 9
Immune cell infiltration. **(A)** Heatmap of correlations among 15 infiltrating immune cells, as analyzed by CIBERSORT. **(B)** Boxplot of differences in 15 infiltrating immune cells between OSA and control samples, as analyzed by CIBERSORT. **(C)** Heatmap of correlations among 23 infiltrating immune cells, as analyzed by ssGSEA. **(D)** Boxplot of differences in 23 infiltrating immune cells between clusters A and B, as analyzed by ssGSEA. **(E)** Heatmap of the differences in immune cell infiltration (based on ssGSEA between clusters A and B.

distinguished between OSA and control samples, which highlights that mitochondrial dysfunction differs between OSA patients and control individuals. Although there have been previous studies on OSA diagnostic genes (Li et al., 2017; Ambati et al., 2020; Bencharit et al., 2021; Cao et al., 2021; Li et al., 2022), we are the first group to establish and validate a mitochondrial dysfunction-related diagnostic model.

Among the mitochondrial dysfunction-related genes, *NPR3* mediates natriuretic peptides degradation and was proved to act as a tumor suppressor in certain types of cancers. Moreover, previous study also showed that it played an important role in modulating intravascular volume and vascular tone and could protect cardiomyocytes from apoptosis. Thus, *NPR3* might be a viable therapeutic target to decrease cancer and cardiovascular

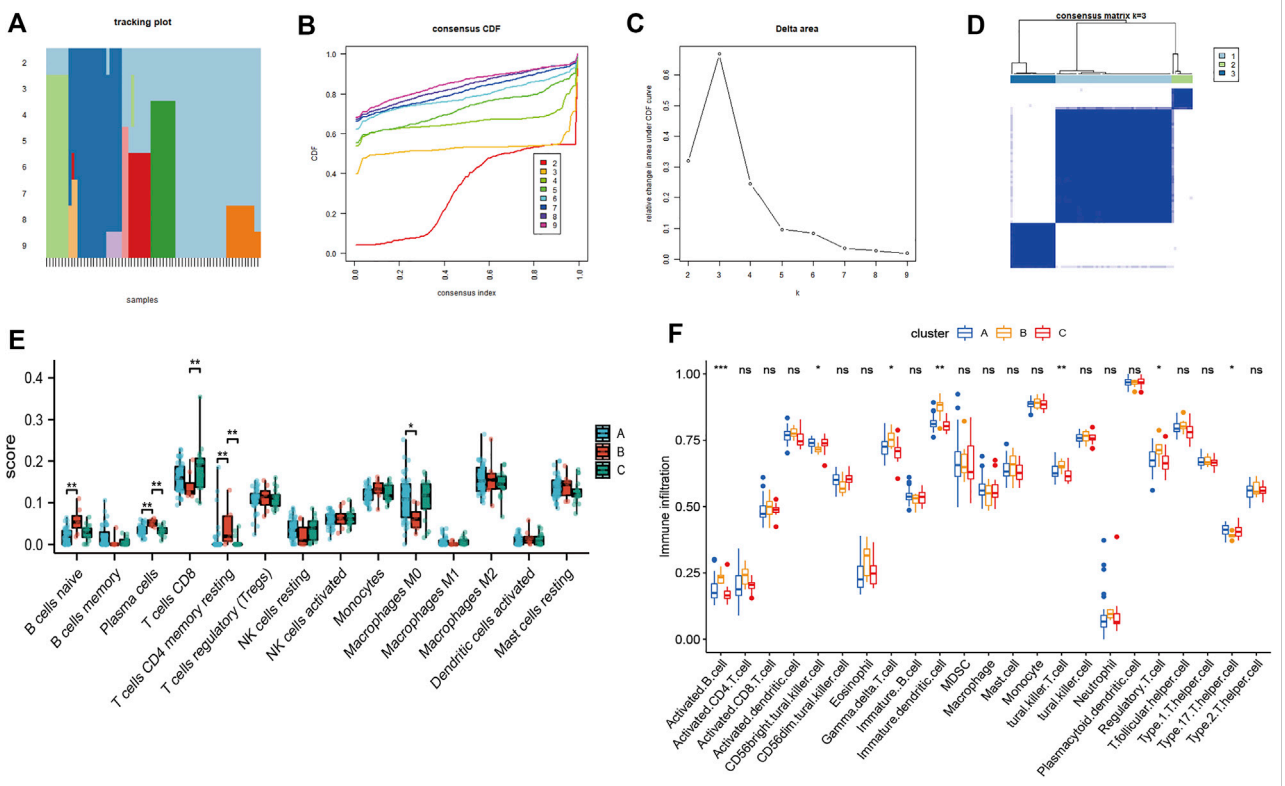


FIGURE 10
Consensus clustering analysis of OSA subjects based on mitochondrial dysfunction-related genes and analysis of immune cell infiltration. **(A)** Tracking plot of consistent clustering. **(B)** Consensus clustering cumulative distribution function (CDF) for k = 2–9. **(C)** Relative change in area under CDF curve for k = 2–7. **(D)** OSA subjects were divided into three clusters when k = 3. Boxplots of degree of immune cell infiltration among clusters A–C, based on **(E)** CIBERSORT and **(F)** ssGSEA methods.

diseases risk in OSA patients (Lin et al., 2016; Li et al., 2021). *PDIA* was reported to be able to intercept the endoplasmic reticulum stress-related apoptotic cellular death and its expression is significantly up-regulated in response to cellular stress (Mahmood et al., 2021). *SLPI* is an important immunity regulator, acts as a component of tissue regenerative programs, and has anti-proteolytic, anti-microbial and immunomodulatory activities (Majchrzak-Gorecka et al., 2016). *ERAP2* plays roles in the processing of antigenic peptides and influences cellular cytotoxic immune responses (de Castro and Stratikos 2019). Obviously, *PDIA3*, *SLPI*, and *ERAP2* are involved in stress and immune response. The experimental models of OSA suggested that the metabolic and inflammatory changes induced by chronic intermittent hypoxia and sleep fragmentation may foster or exacerbate immune alterations (Almendros et al., 2020).

We obtained 134 miRNAs related to the four genes in the mitochondrial dysfunction-related diagnostic model. A previous study reported differentially expressed miRNAs in OSA (Li et al., 2017), but they were not necessarily associated with mitochondrial dysfunction. Additionally, to identify another set of key genes related to OSA, we selected 10 hub genes (*IL-6, FOS, FOSB, JUN, DUSP1, EGR1, PTGS2, ATF3,* and *ZFP36*) from the PPI network of DEGs between the two mitochondrial dysfunction-related clusters. Interestingly, IL-6 receptor levels have been reported to reflect OSA severity (Zheng et al., 2018); *FOS, FOSB,* and *JUN* have been demonstrated to be involved in obesity, osteoporosis, and colorectal cancer (Skrypnik et al., 2017); and *DUSP1* is upregulated in CIH in OSA patients (Hoffmann et al., 2013). Additionally, in this study, we found that *SLP1* expression was positively correlated with *IL-6, FOS, FOSB,* and *JUN*, whereas *PDIA3* expression was negatively correlated with *FOS, FOSB,* and *JUN*. Although most of the 10 hub genes have not been studied in OSA, we speculate that these genes might be involved in the pathogenesis of OSA and its complications and form a regulatory network to coregulate OSA.

Functional enrichment analysis was conducted after reclassifying the microarray according to the mitochondrial dysfunction and our results indicated that DEGs of two clusters were primarily involved in autophagy, inflammation and immune pathways. 1) Regarding autophagy, consistent with our results, OSA is known to induce autophagy as a

result of hypoxia, oxidative stress, and endothelial dysfunction (Ding et al., 2021). Autophagy is related to metabolic disorders, tumors, pulmonary diseases, and neurodegenerative disorders, and mitophagy is an autophagic response that specifically targets mitochondria (Bravo-San Pedro et al., 2017). 2) Regarding inflammation, mitochondrial dysfunction can trigger innate immune responses and inflammation (West 2017). Additionally, inflammatory mediators and infiltrating immune cells can trigger signaling cascades that alter mitochondrial metabolism. Cytokines can inhibit mitochondrial oxidative phosphorylation and induce mitochondrial ROS production, which may alter mitochondrial dynamics and ultimately result in cell death. In particular, it has been reported that OSA may lead to atherosclerosis due to inflammatory processes induced by CIH (Stanke-Labesque et al., 2014). 3) Regarding immune pathways, high levels of IL-6 and TNF-α are predictors of major adverse cardiovascular events in diabetic patients with peripheral artery disease (Biscetti et al., 2019). The IL-1 superfamily of cytokines also plays a vital role in immunity by regulating host defenses, inflammation, and injury. IL-1 inhibition improves glycemic control, and decreases the incidence of cardiovascular disease (Herder et al., 2015; Zheng et al., 2019). Notably, IL-33, a cytokine from the IL-1 family, is an inflammatory mediator, that is, increased in OSA patients compared to controls (Gabryelska et al., 2019). Importantly, pro-inflammatory activation of monocytes activates mTORC1, which enhances the production of chemokines and cytokines (Lin et al., 2014), and the mTORC1 pathway was found to play a key role in mitochondrial dysfunction (Condon et al., 2021). The mTOR pathway was also reported to be the most important DEG-enriched pathway in severe OSA patients with hypertension (Ko et al., 2021). In summary, these results gave a detailed description of the ways and mechanisms how mitochondrial dysfunction participates in OSA's progress, which may benefit future development of precise treatment.

Our data demonstrated the differences in infiltrating immune cells between OSA and control samples, and these cells may also be responsible for OSA comorbidities. Immune cell infiltration may also be of great importance in the remission of OSA (Fan et al., 2021). Consistent with our findings, CIH in OSA was previously found to induce adipose tissue macrophages towards a pro-inflammatory M1 subtype (Ryan 2017), and macrophages are known to contribute to adipose tissue insulin resistance and vascular atherogenesis (Trzepizur et al., 2018). Additionally, imbalanced effector T helper cells were found in patients with OSA and hypertension (Zhang et al., 2022). Moreover, immune cell infiltration in the myocardium adversely affects heart function (Carrillo-Salinas et al., 2019), so OSA may elicit the immunologic alterations that lead to cardiovascular changes. Lastly, there is also a link between OSA and increased cancer incidence and mortality, as intermittent hypoxia induces changes in signaling pathways involved in the regulation of host immunological surveillance that results in tumor formation and invasion (Martínez-García et al., 2016; Picado and Roca-Ferrer 2020). Intermittent hypoxia may lead to a tumor-promoting phenotype among tumor-associated macrophages, leading to more aggressive tumor behavior (Cao et al., 2015). Better understanding of immune infiltration may be of great significance discovering novel therapeutic targets and improving cardiovascular and cancer outcomes in OSA.

Mitochondria are not only important for energy supply during immune activation, but they also induce host immunological surveillance and are involved in immune cell differentiation. We found that several immune cell types, especially T cells (γ/δ T cells, natural killer T cells, Tregs, and type 17 T helper cells), were significantly different among the three mitochondrial dysfunction-related clusters of OSA samples, so T cells are a promising choice for future OSA treatment development. In addition, the diagnostic genes were correlated with immune cell infiltration. Research has shown that mitochondrial processes, along with cytosolic metabolic processes, drive T cell activation, survival, proliferation, and effector functions (Sena et al., 2013). Another study showed that increased γ/δ T cell adhesion occurs in lesion-prone areas of the arterial tree when high cholesterol induces the translocation of ATP synthase β chain from the mitochondria to membrane caveolae in endothelial cells (Fu et al., 2011). Furthermore, research showed that mTORC1 activation induces Treg proliferation while inhibiting the suppressive activity of the Tregs (Procaccini et al., 2010), which suppress immune activation *via* immunosuppressive cytokines (Agita and Alsagaff 2017). Lastly, T cells have been shown to be central to the immune responses contributing to hypertension (Madhur et al., 2021). In summary, our research revealed that mitochondrial dysfunction might influence immune cell infiltration, including T cell infiltration, in OSA and thus promote OSA-related diseases.

However, our study has several limitations. First, the sample size was small. To confirm the diagnostic value for OSA of the hub genes, external validation using a larger sample size would be helpful. Second, not only the identification of mRNA expression level by real time PCR, but also the protein expression levels of these genes would be necessary to examine using western blot to deepen our understanding the molecular mechanisms of OSA. To comprehensively identify the nature of the mitochondrial dysfunctions in OSA, integrated analysis at the molecular, cellular, and organismal levels is warranted, with experimental evidence being required to fully determine the roles of the hub genes and the underlying mechanisms of OSA. Third, the role of each hub gene and the mechanisms underlying OSA were not fully elucidated. Further experimental verifications are necessary to elucidate the biological functions of these genes in OSA. Fourthly, the correlations between the expression of these genes and the clinical parameters of OSA were not explored, so further research on this is required.

In conclusion, we established and validated a mitochondrial dysfunction-related four-gene signature of diagnostic model for OSA. Moreover, we revealed that this model was related to immune cell infiltration. The model could act as a diagnostic biomarker model and might provide therapeutic targets the treatment of OSA. Further studies should be conducted to clarify our findings.

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

QL, DW, and XZ conceived the research theme and supervised the entire study. QL and TH collected the data, analyzed the data, drew the figures, explained the results, and drafted the manuscript. LL, DH, ZL, and YF revised the manuscript and performed reference collection. All the authors approved the final version of the manuscript.

Funding

This work was supported by grants from the National Natural Science Foundation of China (No. 82100244, 31371509), China Postdoctoral Science Foundation (No. 2022M712012), Shandong Province Medical and Health Science and Technology Development Project (No. 202003040648), Binzhou Medical University Scientific Foundation (No. BYFY2020KYQD40), 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (No. 2020LKSFG20B), and Special Fund for Science and technology innovation strategy of Guangdong province (No. 220717237482196).

Acknowledgments

We would like to thank all the contributors to the transcriptome sequencing data used in this study.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.1056691/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Normalization of datasets. Gene expression distribution in **(A)** GSE38792 dataset before normalization, **(B)** GSE38792 dataset after normalization, **(C)** GSE135917 dataset before normalization, **(D)** GSE135917 dataset after normalization, **(E)** combined dataset before normalization, and **(F)** combined dataset after normalization.

SUPPLEMENTARY FIGURE S2
Differentially expressed genes (DEGs) between OSA and control samples. Heatmaps and volcano plots of DEGs between OSA and control samples in **(A,B)** GSE38792, **(C,D)** GSE135917, and **(E,F)** combined datasets. The upregulated, downregulated, and non-significant genes are marked in blue, yellow and grey dots, respectively.

SUPPLEMENTARY FIGURE S3
(A) Prediction of transcription factors and miRNAs related to the genes in the four-gene diagnostic model. Red and green dots in the middle represent genes related to mitochondrial dysfunction and their related transcription factors, respectively. Blue dots in the outer layer represent the related miRNAs. **(B)** Molecular networks of protein chemical interactions of the genes in the four-gene diagnostic model.

SUPPLEMENTARY FIGURE S4
Correlations of 10 hub genes and 4 diagnostic genes. **(A)** Correlation network diagram of correlations among hub gene expression levels. Scatter plots of significant **(B–H)** negative and **(I–P)** positive correlations between hub and diagnostic gene expression.

SUPPLEMENTARY FIGURE S5
Correlations between genes in the diagnostic model and infiltrating immune cells in OSA samples. Scatter plots of significant correlations of **(A–E)** SLPI and **(F–H)** PDIA3 expression with the degree of immune cell infiltration.

References

Agita, A., and Alsagaff, M. T. (2017). Inflammation, immunity, and hypertension. *Acta Med. Indones.* 49 (2), 158–165.

Almendros, I., Martinez-Garcia, M. A., Farré, R., and Gozal, D. (2020). Obesity, sleep apnea, and cancer. *Int. J. Obes.* 44 (8), 1653–1667. doi:10.1038/s41366-020-0549-z

Ambati, A., Ju, Y. E., Lin, L., Olesen, A. N., Koch, H., Hedou, J. J., et al. (2020). Proteomic biomarkers of sleep apnea. *Sleep* 43 (11), zsaa086. doi:10.1093/sleep/zsaa086

Arnaud, C., Bochaton, T., Pépin, J. L., and Belaidi, E. (2020). Obstructive sleep apnoea and cardiovascular consequences: Pathophysiological mechanisms. *Arch. Cardiovasc. Dis.* 113 (5), 350–358. doi:10.1016/j.acvd.2020.01.003

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 (1), 25–29. doi:10.1038/75556

Bencharit, S., Redenz, R. G., Brody, E. R., and Chiang, H. (2021). Salivary biomarkers associated with obstructive sleep apnea: A systematic review. *Expert Rev. Mol. diagn.* 21 (2), 223–233. doi:10.1080/14737159.2021.1873132

Biscetti, F., Ferraro, P. M., Hiatt, W. R., Angelini, F., Nardella, E., Cecchini, A. L., et al. (2019). Inflammatory cytokines associated with failure of lower-extremity endovascular revascularization (ler): A prospective study of a population with diabetes. *Diabetes Care* 42 (10), 1939–1945. doi:10.2337/dc19-0408

Bravo-San Pedro, J. M., Kroemer, G., and Galluzzi, L. (2017). Autophagy and mitophagy in cardiovascular disease. *Circ. Res.* 120 (11), 1812–1824. doi:10.1161/circresaha.117.311082

Brunetti, V., Della Marca, G., Servidei, S., and Primiano, G. (2021). Sleep disorders in mitochondrial diseases. *Curr. Neurol. Neurosci. Rep.* 21 (7), 30. doi:10.1007/s11910-021-01121-2

Cao, J., Feng, J., Li, L., and Chen, B. (2015). Obstructive sleep apnea promotes cancer development and progression: A concise review. *Sleep. Breath.* 19 (2), 453–457. doi:10.1007/s11325-015-1126-x

Cao, Y., Cai, X., Zhu, Q., and Li, N. (2021). Screening and identification of potential biomarkers for obstructive sleep apnea via microarray analysis. *Med. Baltim.* 100 (4), e24435. doi:10.1097/md.0000000000024435

Carrillo-Salinas, F. J., Ngwenyama, N., Anastasiou, M., Kaur, K., and Alcaide, P. (2019). Heart inflammation: Immune cell roles and roads to the heart. *Am. J. Pathol.* 189 (8), 1482–1494. doi:10.1016/j.ajpath.2019.04.009

Chapman, J., Fielder, E., and Passos, J. F. (2019). Mitochondrial dysfunction and cell senescence: Deciphering a complex relationship. *FEBS Lett.* 593 (13), 1566–1579. doi:10.1002/1873-3468.13498

Chen, Q., Han, X., Chen, M., Zhao, B., Sun, B., Sun, L., et al. (2021). High-Fat diet-induced mitochondrial dysfunction promotes genioglossus injury - a potential mechanism for obstructive sleep apnea with obesity. *Nat. Sci. Sleep.* 13, 2203–2219. doi:10.2147/nss.s343721

Condon, K. J., Orozco, J. M., Adelmann, C. H., Spinelli, J. B., van der Helm, P. W., Roberts, J. M., et al. (2021). Genome-wide CRISPR screens reveal multitiered mechanisms through which mTORC1 senses mitochondrial dysfunction. *Proc. Natl. Acad. Sci. U. S. A.* 118 (4), e2022120118. doi:10.1073/pnas.2022120118

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., et al. (2021). Comparative Toxicogenomics database (CTD): Update 2021. *Nucleic Acids Res.* 49 (D1), D1138–d1143. doi:10.1093/nar/gkaa891

de Castro, J. A. L., and Stratikos, E. (2019). Intracellular antigen processing by ERAP2: Molecular mechanism and roles in health and disease. *Hum. Immunol.* 80 (5), 310–317. doi:10.1016/j.humimm.2018.11.001

Ding, H., Guo, H., and Cao, J. (2021). The importance of autophagy regulation in obstructive sleep apnea. *Sleep. Breath.* 25 (3), 1211–1218. doi:10.1007/s11325-020-02261-4

Fan, C., Huang, S., Xiang, C., An, T., and Song, Y. (2021). Identification of key genes and immune infiltration modulated by CPAP in obstructive sleep apnea by integrated bioinformatics analysis. *PLoS One* 16 (9), e0255708. doi:10.1371/journal.pone.0255708

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., et al. (2020). Jaspar 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48 (D1), D87–d92. doi:10.1093/nar/gkz1001

Forrester, S. J., Kikuchi, D. S., Hernandes, M. S., Xu, Q., and Griendling, K. K. (2018). Reactive oxygen species in metabolic and inflammatory signaling. *Circ. Res.* 122 (6), 877–902. doi:10.1161/circresaha.117.311401

Fu, Y., Hou, Y., Fu, C., Gu, M., Li, C., Kong, W., et al. (2011). A novel mechanism of γ/δ T-lymphocyte and endothelial activation by shear stress: The role of ecto-ATP synthase β chain. *Circ. Res.* 108 (4), 410–417. doi:10.1161/CIRCRESAHA.110.230151

Gabryelska, A., Kuna, P., Antczak, A., Bialasiewicz, P., and Panek, M. (2019). IL-33 mediated inflammation in chronic respiratory diseases-understanding the role of the member of IL-1 superfamily. *Front. Immunol.* 10, 692. doi:10.3389/fimmu.2019.00692

Gharib, S. A., Hayes, A. L., Rosen, M. J., and Patel, S. R. (2013). A pathway-based analysis on the effects of obstructive sleep apnea in modulating visceral fat transcriptome. *Sleep* 36 (1), 23–30. doi:10.5665/sleep.2294

Gharib, S. A., Hurley, A. L., Rosen, M. J., Spilsbury, J. C., Schell, A. E., Mehra, R., et al. (2020). Obstructive sleep apnea and CPAP therapy alter distinct transcriptional programs in subcutaneous fat tissue. *Sleep* 43 (6), zsz314. doi:10.1093/sleep/zsz314

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). Gsva: Gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Herder, C., Dalmas, E., Boni-Schnetzler, M., and Donath, M. Y. (2015). The IL-1 pathway in type 2 diabetes and cardiovascular complications. *Trends Endocrinol. Metab.* 26 (10), 551–563. doi:10.1016/j.tem.2015.08.001

Hernandez-Aguilera, A., Rull, A., Rodriguez-Gallego, E., Riera-Borrull, M., Luciano-Mateo, F., Camps, J., et al. (2013). Mitochondrial dysfunction: A basic mechanism in inflammation-related non-communicable diseases and therapeutic opportunities. *Mediat. Inflamm.* 2013, 135698. doi:10.1155/2013/135698

Hoffmann, M. S., Singh, P., Wolk, R., Narkiewicz, K., and Somers, V. K. (2013). Obstructive sleep apnea and intermittent hypoxia increase expression of dual specificity phosphatase 1. *Atherosclerosis* 231 (2), 378–383. doi:10.1016/j.atherosclerosis.2013.09.033

Huang, H., Jiang, X., Dong, Y., Zhang, X., Ding, N., Liu, J., et al. (2014). Adiponectin alleviates genioglossal mitochondrial dysfunction in rats exposed to intermittent hypoxia. *PLoS One* 9 (10), e109284. doi:10.1371/journal.pone.0109284

Huang, L., Wu, C., Xu, D., Cui, Y., and Tang, J. (2021). Screening of important factors in the early sepsis stage based on the evaluation of ssGSEA algorithm and ceRNA regulatory network. *Evol. Bioinform. Online* 17, 11769343211058463. doi:10.1177/11769343211058463

Kanehisa, M., and Goto, S. (2000). Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kasapoğlu, I., and Seli, E. (2020). Mitochondrial dysfunction and ovarian aging. *Endocrinology* 161 (2), bqaa001. doi:10.1210/endocr/bqaa001

Kim, Y. S., Kwak, J. W., Lee, K. E., Cho, H. S., Lim, S. J., Kim, K. S., et al. (2014). Can mitochondrial dysfunction be a predictive factor for oxidative stress in patients with obstructive sleep apnea? *Antioxid. Redox Signal.* 21 (9), 1285–1288. doi:10.1089/ars.2014.5955

Ko, C. Y., Su, H. Z., Zhang, L., and Zeng, Y. M. (2021). Disturbances of the gut microbiota, sleep architecture, and mTOR signaling pathway in patients with severe obstructive sleep apnea-associated hypertension. *Int. J. Hypertens.* 2021, 9877053. doi:10.1155/2021/9877053

Li, K., Wei, P., Qin, Y., and Wei, Y. (2017). MicroRNA expression profiling and bioinformatics analysis of dysregulated microRNAs in obstructive sleep apnea patients. *Med. Baltim.* 96 (34), e7917. doi:10.1097/md.0000000000007917

Li, S., Guo, R., Peng, Z., Quan, B., Hu, Y., Wang, Y., et al. (2021). NPR3, transcriptionally regulated by POU2F1, inhibits osteosarcoma cell growth through blocking the PI3K/AKT pathway. *Cell. Signal.* 86, 110074. doi:10.1016/j.cellsig.2021.110074

Li, Y., Li, L., Zhao, H., Gao, X., and Li, S. (2022). Identifying obstructive sleep apnea syndrome-associated genes and pathways through weighted gene coexpression network analysis. *Comput. Math. Methods Med.* 2022, 3993509. doi:10.1155/2022/3993509

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004

Lin, C. C., Chen, W. J., Sun, Y. K., Chiu, C. H., Lin, M. W., and Tzeng, I. S. (2021). Continuous positive airway pressure affects mitochondrial function and exhaled PGC1-α levels in obstructive sleep apnea. *Exp. Lung Res.* 47 (10), 476–486. doi:10.1080/01902148.2021.2001607

Lin, D., Chai, Y., Izadpanah, R., Braun, S. E., and Alt, E. (2016). NPR3 protects cardiomyocytes from apoptosis through inhibition of cytosolic BRCA1 and TNF-α. *Cell Cycle* 15 (18), 2414–2419. doi:10.1080/15384101.2016.1148843

Lin, H. Y., Chang, K. T., Hung, C. C., Kuo, C. H., Hwang, S. J., Chen, H. C., et al. (2014). Effects of the mTOR inhibitor rapamycin on monocyte-secreted chemokines. *BMC Immunol.* 15, 37. doi:10.1186/s12865-014-0037-0

Madhur, M. S., Elijovich, F., Alexander, M. R., Pitzer, A., Ishimwe, J., Van Beusecum, J. P., et al. (2021). Hypertension: Do inflammation and immunity hold the key to solving this epidemic? *Circ. Res.* 128 (7), 908–933. doi:10.1161/circresaha.121.318052

Mahmood, F., Xu, R., Awan, M. U. N., Song, Y., Han, Q., Xia, X., et al. (2021). PDIA3: Structure, functions and its potential role in viral infections. *Biomed. Pharmacother.* 143, 112110. doi:10.1016/j.biopha.2021.112110

Majchrzak-Gorecka, M., Majewski, P., Grygier, B., Murzyn, K., and Cichy, J. (2016). Secretory leukocyte protease inhibitor (SLPI), a multifunctional protein in

the host defense response. *Cytokine Growth Factor Rev.* 28, 79–93. doi:10.1016/j.cytogfr.2015.12.001

Martínez-García, M., Campos-Rodriguez, F., and Barbé, F. (2016). Cancer and OSA: Current evidence from human studies. *Chest* 150 (2), 451–463. doi:10.1016/j.chest.2016.04.029

Passali, D., Corallo, G., Yaremchuk, S., Longini, M., Proietti, F., Passali, G. C., et al. (2015). Oxidative stress in patients with obstructive sleep apnoea syndrome. *Acta Otorhinolaryngol. Ital.* 35 (6), 420–425. doi:10.14639/0392-100X-895

Picado, C., and Roca-Ferrer, J. (2020). Role of the cyclooxygenase pathway in the association of obstructive sleep apnea and cancer. *J. Clin. Med.* 9 (10), E3237. doi:10.3390/jcm9103237

Procaccini, C., De Rosa, V., Galgani, M., Abanni, L., Cali, G., Porcellini, A., et al. (2010). An oscillatory switch in mTOR kinase activity sets regulatory T cell responsiveness. *Immunity* 33 (6), 929–941. doi:10.1016/j.immuni.2010.11.024

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 (7), e47. doi:10.1093/nar/gkv007

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 Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Rocha, E. M., De Miranda, B., and Sanders, L. H. (2018). Alpha-synuclein: Pathology, mitochondrial dysfunction and neuroinflammation in Parkinson's disease. *Neurobiol. Dis.* 109, 249–257. doi:10.1016/j.nbd.2017.04.004

Ryan, S. (2017). Adipose tissue inflammation by intermittent hypoxia: Mechanistic link between obstructive sleep apnoea and metabolic dysfunction. *J. Physiol.* 595 (8), 2423–2430. doi:10.1113/jp273312

Sabah, A., Tiun, S., Sani, N. S., Ayob, M., and Taha, A. Y. (2021). Enhancing web search result clustering model based on multiview multirepresentation consensus cluster ensemble (mmcc) approach. *PLoS One* 16 (1), e0245264. doi:10.1371/journal.pone.0245264

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., et al. (2010). GeneCards version 3: The human gene integrator. *Database (Oxford)* 2010, baq020. doi:10.1093/database/baq020

Seiler, M., Huang, C. C., Szalma, S., and Bhanot, G. (2010). ConsensusCluster: A software tool for unsupervised cluster discovery in numerical data. *OMICS* 14 (1), 109–113. doi:10.1089/omi.2009.0083

Sena, L. A., Li, S., Jairaman, A., Prakriya, M., Ezponda, T., Hildeman, D. A., et al. (2013). Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. *Immunity* 38 (2), 225–236. doi:10.1016/j.immuni.2012.10.020

Shan, X., Chi, L., Ke, Y., Luo, C., Qian, S., Gozal, D., et al. (2007). Manganese superoxide dismutase protects mouse cortical neurons from chronic intermittent hypoxia-mediated oxidative damage. *Neurobiol. Dis.* 28 (2), 206–215. doi:10.1016/j.nbd.2007.07.013

Shi, Y., Jiang, Z., Jiang, L., and Xu, J. (2021). Integrative analysis of key candidate genes and signaling pathways in acute coronary syndrome related to obstructive sleep apnea by bioinformatics. *Sci. Rep.* 11 (1), 14153. doi:10.1038/s41598-021-93789-2

Skrypnik, K., Suliburska, J., Skrypnik, D., Pilarski, Ł., Reguła, J., and Bogdański, P. (2017). The genetic basis of obesity complications. *Acta Sci. Pol. Technol. Aliment.* 16 (1), 83–91. doi:10.17306/j.afs.2017.0442

Song, J. X., Zhao, Y. S., Zhen, Y. Q., Yang, X. Y., Chen, Q., An, J. R., et al. (2022). Banxia-Houpu decoction diminishes iron toxicity damage in heart induced by chronic intermittent hypoxia. *Pharm. Biol.* 60 (1), 609–620. doi:10.1080/13880209.2022.2043392

Spangler, J. B., Moraga, I., Mendoza, J. L., and Garcia, K. C. (2015). Insights into cytokine-receptor interactions from cytokine engineering. *Annu. Rev. Immunol.* 33, 139–167. doi:10.1146/annurev-immunol-032713-120211

Srinivasan, S., Guha, M., Kashina, A., and Avadhani, N. G. (2017). Mitochondrial dysfunction and mitochondrial dynamics-The cancer connection. *Biochim. Biophys. Acta. Bioenerg.* 1858 (8), 602–614. doi:10.1016/j.bbabio.2017.01.004

Stanke-Labesque, F., Pepin, J. L., Gautier-Veyret, E., Levy, P., and Back, M. (2014). Leukotrienes as a molecular link between obstructive sleep apnoea and atherosclerosis. *Cardiovasc. Res.* 101 (2), 187–193. doi:10.1093/cvr/cvt247

Steen, C. B., Liu, C. L., Alizadeh, A. A., and Newman, A. M. (2020). Profiling cell type Abundance and expression in bulk tissues with CIBERSORTx. *Methods Mol. Biol.* 2117, 135–157. doi:10.1007/978-1-0716-0301-7_7

Tang, X., Li, S., Yang, X., Tang, Q., Zhang, Y., Zeng, S., et al. (2021). Novel proteins associated with chronic intermittent hypoxia and obstructive sleep apnea: From rat model to clinical evidence. *PLoS One* 16 (6), e0253943. doi:10.1371/journal.pone.0253943

Trzepizur, W., Cortese, R., and Gozal, D. (2018). Murine models of sleep apnea: Functional implications of altered macrophage polarity and epigenetic modifications in adipose and vascular tissues. *Metabolism.* 84, 44–55. doi:10.1016/j.metabol.2017.11.008

Vásquez-Trincado, C., García-Carvajal, I., Pennanen, C., Parra, V., Hill, J. A., Rothermel, B. A., et al. (2016). Mitochondrial dynamics, mitophagy and cardiovascular disease. *J. Physiol.* 594 (3), 509–525. doi:10.1113/jp271301

Wang, H., Kurniansyah, N., Cade, B. E., Goodman, M. O., Chen, H., Gottlieb, D. J., et al. (2022). Upregulated heme biosynthesis increases obstructive sleep apnea severity: A pathway-based mendelian randomization study. *Sci. Rep.* 12 (1), 1472. doi:10.1038/s41598-022-05415-4

Wang, S., Sun, H., Ma, J., Zang, C., Wang, C., Wang, J., et al. (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.* 8 (12), 2502–2515. doi:10.1038/nprot.2013.150

Wang, S., Tan, J., Miao, Y., and Zhang, Q. (2022). Mitochondrial dynamics, mitophagy, and mitochondria-endoplasmic reticulum contact sites crosstalk under hypoxia. *Front. Cell Dev. Biol.* 10, 848214. doi:10.3389/fcell.2022.848214

Wang, Y., Zhang, S. X., and Gozal, D. (2010). Reactive oxygen species and the brain in sleep apnea. *Respir. Physiol. Neurobiol.* 174 (3), 307–316. doi:10.1016/j.resp.2010.09.001

Wei, P. Z., and Szeto, C. C. (2019). Mitochondrial dysfunction in diabetic kidney disease. *Clin. Chim. Acta.* 496, 108–116. doi:10.1016/j.cca.2019.07.005

Wei, Q., Xu, X., Chen, L., Wang, T., Xie, L., Yu, F. C., et al. (2022). Effects of chronic intermittent hypoxia on left cardiac function in young and aged mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 322 (3), R241–r252. doi:10.1152/ajpregu.00256.2021

West, A. P. (2017). Mitochondrial dysfunction as a trigger of innate immune responses and inflammation. *Toxicology* 391, 54–63. doi:10.1016/j.tox.2017.07.016

Wu, L., Wang, W., Tian, S., Zheng, H., Liu, P., and Wu, W. (2021). Identification of hub genes in patients with alzheimer disease and obstructive sleep apnea syndrome using integrated bioinformatics analysis. *Int. J. Gen. Med.* 14, 9491–9502. doi:10.2147/ijgm.s341078

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation.* 2 (3), 100141. doi:10.1016/j.xinn.2021.100141

Xu, L., Bi, Y., Xu, Y., Wu, Y., Du, X., Mou, Y., et al. (2021). Suppression of CHOP reduces neuronal apoptosis and rescues cognitive impairment induced by intermittent hypoxia by inhibiting bax and bak activation. *Neural Plast.* 2021, 4090441. doi:10.1155/2021/4090441

Yao, R. Q., Ren, C., Xia, Z. F., and Yao, Y. M. (2021). Organelle-specific autophagy in inflammatory diseases: A potential therapeutic target underlying the quality control of multiple organelles. *Autophagy* 17 (2), 385–401. doi:10.1080/15548627.2020.1725377

Yapa, N. M. B., Lisnyak, V., Reljic, B., and Ryan, M. T. (2021). Mitochondrial dynamics in health and disease. *FEBS Lett.* 595 (8), 1184–1204. doi:10.1002/1873-3468.14077

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track plots. *BMC Bioinforma.* 14, 244. doi:10.1186/1471-2105-14-244

Zhang, L., Ko, C. Y., and Zeng, Y. M. (2022). Immunoregulatory effect of short-chain fatty acids from gut microbiota on obstructive sleep apnea-associated hypertension. *Nat. Sci. Sleep.* 14, 393–405. doi:10.2147/nss.s354742

Zhao, Y. S., An, J. R., Yang, S., Guan, P., Yu, F. Y., Li, W., et al. (2019). Hydrogen and oxygen mixture to improve cardiac dysfunction and myocardial pathological changes induced by intermittent hypoxia in rats. *Oxid. Med. Cell. Longev.* 2019, 7415212. doi:10.1155/2019/7415212

Zheng, M., Wang, X., and Zhang, L. (2018). Association between allergic and nonallergic rhinitis and obstructive sleep apnea. *Curr. Opin. Allergy Clin. Immunol.* 18 (1), 16–25. doi:10.1097/aci.0000000000000414

Zheng, Z. H., Zeng, X., Nie, X. Y., Cheng, Y. J., Liu, J., Lin, X. X., et al. (2019). Interleukin-1 blockade treatment decreasing cardiovascular risk. *Clin. Cardiol.* 42 (10), 942–951. doi:10.1002/clc.23246

frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 04 January 2023
DOI 10.3389/fgene.2022.1088189

OPEN ACCESS

EDITED BY
Rui Yin,
Harvard Medical School, United States

REVIEWED BY
Chu Pan,
University Health Network (UHN), Canada
Guanghui Li,
East China Jiaotong University, China
Zhenxiang Gao,
Case Western Reserve University, United States

*CORRESPONDENCE
Wei Liang,
weiliang99@hnu.edu.cn

SPECIALTY SECTION
This article was submitted to RNA, a section of the journal Frontiers in Genetics

RECEIVED 03 November 2022
ACCEPTED 12 December 2022
PUBLISHED 04 January 2023

CITATION
Luo Y, Peng L, Shan W, Sun M, Luo L and Liang W (2023), Machine learning in the development of targeting microRNAs in human disease.
Front. Genet. 13:1088189.
doi: 10.3389/fgene.2022.1088189

Machine learning in the development of targeting microRNAs in human disease

Yuxun Luo[1,2], Li Peng[1,2], Wenyu Shan[3], Mengyue Sun[4], Lingyun Luo[3] and Wei Liang[1,2]*

[1]School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China, [2]Hunan Key Laboratory for Service computing and Novel Software Technology, Xiangtan, China, [3]School of Computer Science, University of South China, Hengyang, China, [4]School of Polymer Science and Polymer Engineering, The University of Akron, Akron, OH, United States

A microRNA is a small, single-stranded, non-coding ribonucleic acid that plays a crucial role in RNA silencing and can regulate gene expression. With the in-depth study of miRNA in development and disease, miRNA has become an attractive target for novel therapeutic strategies. Exploring miRNA targeting therapy only through experiments is expensive and laborious, so it is essential to develop novel and efficient computational methods to narrow down the search. Recent advances in machine learning applied in biomedical informatics provide opportunities to explore miRNA-targeting drugs, thus promoting miRNA therapeutics. This review provides an overview of recent advancements in miRNA targeting therapeutic using machine learning. First, we mainly describe the basics of predicting miRNA targeting drugs, including pharmacogenomic data resources and data preprocessing. Then we present primary machine learning algorithms and elaborate their application in discovering relationships among miRNAs, drugs, and diseases. Along with the progress of miRNA targeting therapeutics, we finally analyze and discuss the current challenges and opportunities that machine learning confronts.

KEYWORDS

machine learning, mirna therapy, miRNA-disease association, miRNA-drug association, deep learning

1 Introduction

As a kind of non-coding RNA transcript, MicroRNA (miRNA) plays a vital role in cell proliferation, survival and differentiation by modulating the transcription of target messenger RNA (mRNA) and disrupting the translation of mRNA (Rupaimoole and Slack, 2017). The miRNA-mRNA interactions usually lead to translation inhibition or mRNA degradation, which brings about the reduction of the final protein output (Guo et al., 2010). MiRNAs act as novel therapeutic targets and potential diagnostic markers due to they can regulate gene expression involved in the pathogenesis of cancer and other complex diseases (Tay et al., 2008). Just a few years after the first miRNA was discovered by Lee and others in 1993 (Lee et al., 1993), the research of miRNA biology dramatically bloomed. The experimentally validated function of miRNAs laid a solid foundation for

cellular biology, which enables researchers to study associated diseases and drugs at the molecular level (Chen et al., 2020).

The efficacy of various miRNA therapies depends on the accurate relationships between miRNAs and diseases. There are many validated relationships that exist between miRNAs and prevalent diseases, such as lung cancer, pancreatic, ovarian cancer, and so on (Roldo et al., 2006). For example, excisions and downregulation of the miR-15/16 cluster frequently occur in chronic lymphocytic leukemia (Calin et al., 2008), and the significant upregulation of miR-21 is involved in hematological malignancies (Fulci et al., 2007). The decreased expression of human let-7 miRNA family in lung cancer was associated with poor prognosis in patients (Takamizawa et al., 2004). MiRNAs also have been associated with several metabolic pathways (Fernández-Hernando et al., 2013), for example, miR-33 influenced the level of triglyceride and the high-density lipoprotein in serum (Marquart et al., 2010). However, it costs a lot of time, money, and resources to acquire associations verified in experiments, which brought about a widespread interest in the computational discovery of underlying miRNA-disease associations during the last few years. More than 186,000 related articles were available online, and many relevant databases and models were designed (Huang et al., 2022a). For instance, the latest version of Human MicroRNA Disease Database (Huang et al., 2019) records 35,547 entries, and the commonly used database miR2Disease (Jiang et al., 2009) contains 3,273 associations. Meanwhile, based on the conception that similar miRNAs would be associated with similar diseases, various computational models were adopted to identify underlying associations. Usually, the homogeneous network and the heterogeneous network were built to extract desired feature embeddings *via* machine learning methods (Fu and Peng, 2017).

The therapeutic advance of diseases was deeply influenced by the time-consuming and costly process of drug discovery and development. Most drugs generally are small molecules, namely low molecular weight organic compounds, that act as a regulator in a biological process. It was indicated in studies that small molecules could disrupt protein interactions, and also suppress specific functions of a multifunctional protein; hence it may have a positive effect on diseases (Melo et al., 2011). Unlike biologics with which injection and other parenteral administration are usually required, most small-molecule drugs can be taken orally. The urgent request for novel therapeutic alternatives makes the approach of targeting disease-related miRNA with small molecules seem to be promising. Since Gumireddy et al. (2008) developed the first small molecule inhibitor of miRNA for specifically suppressing miR-21, numerous miRNA inhibitors have been discovered *via* a sequence-based computational approach or high throughput screening (Young et al., 2010). For instance, the miR-122 inhibitors were identified to suppress the miR-122 expression and reduce 50% of HCV viral load *in vitro* (Kutay et al., 2006). Besides, streptomycin, neomycin, tobramycin, and amikacin could impede miR-27a function, which plays a role in the regulation of adipogenesis, gastric cancer and so on, by directly interacting with pre-miR-27a (Zhang et al., 2011; Chandrasekhar et al., 2012). Recently, more and more miRNA-drug association research has been launched, such as the Developmental Therapeutics Program funded by the National Cancer Institute of United States, which publicly published related datasets. Similarly, many computational methods based on regression, matrix factorization, neural networks and so on have been proposed.

In this review, we firstly listed several manually curated mainstream databases of miRNA-disease associations and miRNA-drug associations as comprehensive resources for computational approaches. Then, with the rapid bloom of machine learning approaches, we reviewed some representative studies on predicting underlying relationships between miRNAs and diseases or drugs using modified learning models. Due to the length limit of the paper, not all papers related to the above introduction are able to be included. Nevertheless, we collected the commonly used databases and the most representative computational methods to reveal promising development trends for targeting miRNAs in human diseases and drugs.

2 Database

As we all know, miRNA expression deregulation is crucial to the state transition from a physiological to a pathological one. Many studies in recent have suggested that bioactive drugs can act as the regulator of miRNA expression, hence indicating a new therapy that miRNAs targeted with small molecules. Therefore, more and more diversified databases containing various omics data increased dramatically due to the development of system biology and molecular biology. The database of miRNAs-diseases was generated from experimentally validated miRNA-disease associations, and the miRNAs-drugs databases originated from experimentally verified small molecules' impacts on the expression of microRNA. In this section, we concluded data details in the most popularly used and commonly cited databases, most of which were still in maintained status, from aspects of miRNA-diseases and miRNA-drugs. Table 1 listed various information about these mainstream databases.

2.1 miRNA-disease associations

2.1.1 miR2Disease

To date, the latest version of miR2Disease (Jiang et al., 2009) curated 3,273 relationships between 349 human microRNAs and 163 human diseases, one-eighth of which suggested the pathogenic roles of various human diseases related to miRNA deregulation. Resources in the miR2Disease contained various details about microRNA-disease relationships, in which every entry could be retrieved by disease name, miRNA ID, or target gene. Additionally, the literature reference, the detection method

TABLE 1 Main databases for accelerating miRNA therapy based on machine learning.

Database	Published year (latest update)	Data type	Number of data	URL
miR2Disease	2008 (2022)	Relationships between deregulated miRNAs and diverse human diseases	3273 entries, 349 microRNAs, 163 human diseases	http://www.mir2disease.org/
PhenomiR	2009 (2011)	Differential regulation of miRNA expression in diseases	11029 data points and 572 miRNAs	http://mips.helmholtz-muenchen.de/phenomir
miRGen	2007 (2020)	miRNAs related to disease status information	Over 1500 miRNAs and 133 cell lines, primary cells, and tissues	http://www.microrna.gr/mirgenv4
miRmine	2016 (2016)	miRNA expression profiles in tissues, cell lines, and diseases	304 miRNA sequencing datasets for 15 tissues and 24 cell lines	https://guanfiles.dcmb.med.umich.edu/mirmine
miRTarBase	2011 (2022)	miRNA-associated diseases and the relationship between miRNA-target interactions and disease	4630 miRNAs and 30tissues/cell lines from 440CLIP-seq datasets	https://miRTarBase.cuhk.edu.cn/
HMDD	2007 (2022)	miRNA-disease associations could be divided into 6 categories of genetics, target, circulation, tissue, epigenetics, and others	35547 entries of miRNA-disease association between 1206 miRNA genes and 893 diseases	http://www.cuilab.cn/hmdd
Pharmaco-miR Verified Sets	2013 (2013)	miRNA pharmacogenomic sets that were verified in experiments	119 target genes, 72 drugs (whose function depends on the gene), and 105 miRNAs	www.Pharmaco-miR.org
SM2miR	2012 (2015)	The experimentally verified small molecules' effects on miRNA expression	4989 entries of relationships between 1658 miRNAs and 255 small molecules	http://www.jianglab.cn/SM2miR/
DTP NCI-60 dataset	2016 (2022)	A dataset of CellMiner database which screened over chemical compounds by utilizing diverse human cancer cell lines	335 miRNA expressions and half-cell growth inhibition concentration from 18724 drugs	https://discover.nci.nih.gov/cellminer
ncDR	2017 (2017)	miRNA-drug resistance associations for predicting non-coding RNA related to drug resistance	5864 experimentally verified relationships between 145 drug compounds and 877 miRNAs	http://www.jianglab.cn/ncDR

The 1st column gives the database names. The 2nd column presents the published year and the latest update of the database. The 3rd column introduces data type included in the database. The 4th column presents the number of data. The 5th column introduces the URL of the database.

for miRNA expression, the expression pattern of miRNA, and a brief description of a relationship are also included in this database.

2.1.2 PhenomiR

The PhenomiR database (Ruepp et al., 2010) included 11,029 data points and 572 miRNAs, which were collected from 542 related studies focusing on the differential regulation of miRNA expression in diseases. In addition to some usual information, PhenomiR provided in-depth information such as the sample size, the quantitative fold-change of miRNA expression, and the origin analysis of samples (cell culture or patients). Depending on disease type in the PhenomiR dataset, we can contrast conclusions originating from patient studies with independent resources drawn from cell culture studies.

2.1.3 miRGen

The latest version miRGen v4 (Perdikopanis et al., 2021) uniquely integrated annotations for numerous cell-specific miRNA promoters with transcription factor binding sites derived from experiments, which clearly revealed the regulation of miRNA at the transcriptional level. Combined with more than 1,000 cap analyses results from gene expression samples (Shiraki et al., 2003) of 133 cell lines, primary cells, and tissues derived from the FANTOM Consortium (Forrest et al., 2014), miRNA transcription start sites that specific in cell type were provided for more than 1500 miRNAs. Details in this database can be queried through the sample-oriented method or miRNA-oriented method.

2.1.4 miRmine

The miRmine database (Panwar et al., 2017) contained details of different miRNAs and collected expression profiles from various miRNA databases. The miRmine functionality included searches based on miRNA and cell-line/tissue, comparison of multiple miRNAs, normal and human disease information, and so on. For specific tissue or cell-line type, miRmine could retrieve single or multiple miRNAs expression information. Besides, retrieved results could be shown in various graphs and interactive formats.

2.1.5 miRTarBase

The miRTarBase 9.0 (Huang et al., 2022b) released in 2021 documented over 360,000 miRNA-target interactions between 27,172 targets and 4,630 miRNAs collected from 13,389 related studies, which facilitated the research of miRNAs' function in pathology and promoted the improvement of diagnostic and therapeutic tools. Integrating with increasing miRNA expression and biological data, miRTarBase accumulated miRNA-target interactions verified in experiments and satisfied biologists' requirements. Additionally, an optimized scoring system is utilized in the updated version to reinforce the important identification of related articles and relevant disease information.

2.1.6 HMDD

To date, 35,547 entries of miRNA-disease association between 1,206 miRNA genes and 893 diseases curated from 19,280 papers were collected in HMDD (Huang et al., 2019). Disease network analysis modules were applied in the latest HMDD v3.3, which was released in Sep 2022. Covering 20 kinds of detailed evidence code derived from literature, miRNA-disease associations in HMDD were divided into six categories of genetics, target, circulation, tissue, epigenetics, and others. Due to the wide coverage and abundant experimentally verified associations, HMDD became one of the most popular databases regarding association prediction and was widely adopted as the benchmark in training and testing prediction models.

2.2 miRNA-drug associations

2.2.1 Pharmaco-miR Verified Sets

In 2014, Pharmaco-miR Verified Set (Rukov et al., 2014) manually curated 269 miRNA pharmacogenomic data from 149 original literature. It is a dataset of miRNA pharmacogenomic sets that were verified in experiments, containing119 target genes, 72 drugs (whose function depends on the gene), and 105 miRNAs. In Pharmaco-miR Verified Sets, the miRNA directly targeted the gene in a specified context, which was typically exhibited *via* luciferase experiments. Meanwhile, in the same context, the efficacy of drugs was affected by the subsequent suppression of gene expression in this database.

2.2.2 SM2miR

SM2miR (Liu et al., 2013) collected miRNA expression influenced by experimentally verified small molecules' effects in 21 species curated from the published papers. To date, it documented 4,989 entries of relationships between 1,658 miRNAs and 255 small molecules. Various details of each entry encompass species, the miRNA expression pattern, accession number in miRbase and DrugBank, detection conditions, experimental method, PubChem Compound Identifier, PubMed ID, and the related reference information.

2.2.3 DTP NCI-60 dataset

The U.S. National Cancer Institute launched the Developmental Therapeutics Program, which screened over 100,000 chemical compounds by utilizing 60 diverse human cancer cell lines, namely DTP NCI-60 (Blower et al., 2007). In NCI-60 dataset, data consists of 335 miRNA expressions and half-cell growth inhibition concentration (GI50) from 18,724 drugs. The DTP NCI-60 dataset can evaluate the correlations between miRNA expression and drug sensitivity by calculating the Pearson correlation coefficient between miRNA expression level and GI50 value.

2.2.4 ncDR

In 2017, a comprehensive database called ncDR documenting miRNA-drug resistance associations was released to predict non-coding RNA related to drug resistance (Dai et al., 2017). This database contains 5,864 experimentally verified relationships between 145 drug compounds and 877 miRNAs through manually curating from about 3,300 relevant literatures. In addition, 226,109 predicted relationships between drug resistance and miRNA were already provided in this database.

3 Predicting miRNA-disease associations

In past biological experiments, plenty of relationships between diseases and miRNAs have been verified, which laid the foundation for discovering latent miRNA-disease associations *in silico*. At first, both negative and positive samples were included in the training set because the association prediction was usually processed as a binary classification task. Undoubtedly, the known miRNA-disease associations constituted positive training samples; hence, negative ones were randomly sampled from the remaining. The remaining set may contain unknown disease miRNA. As we all know, negative samples should only contain miRNAs and diseases between which the relationship was actually nonexistent; however, there are still many unknown miRNA-disease associations that have not been detected in biological experiments. It is most likely that the current negative samples contained many undiscovered associations. Therefore, to avoid bias brought by the sample, various computational methods only learned from verified associations were proposed to accurately predict miRNA-disease associations. Furthermore, the miRNA-disease association prediction was processed as a triplet classification in machine learning approaches, which could identify the role miRNA played. The main process for predicting miRNA disease associations based on machine learning is presented in Figure 1.

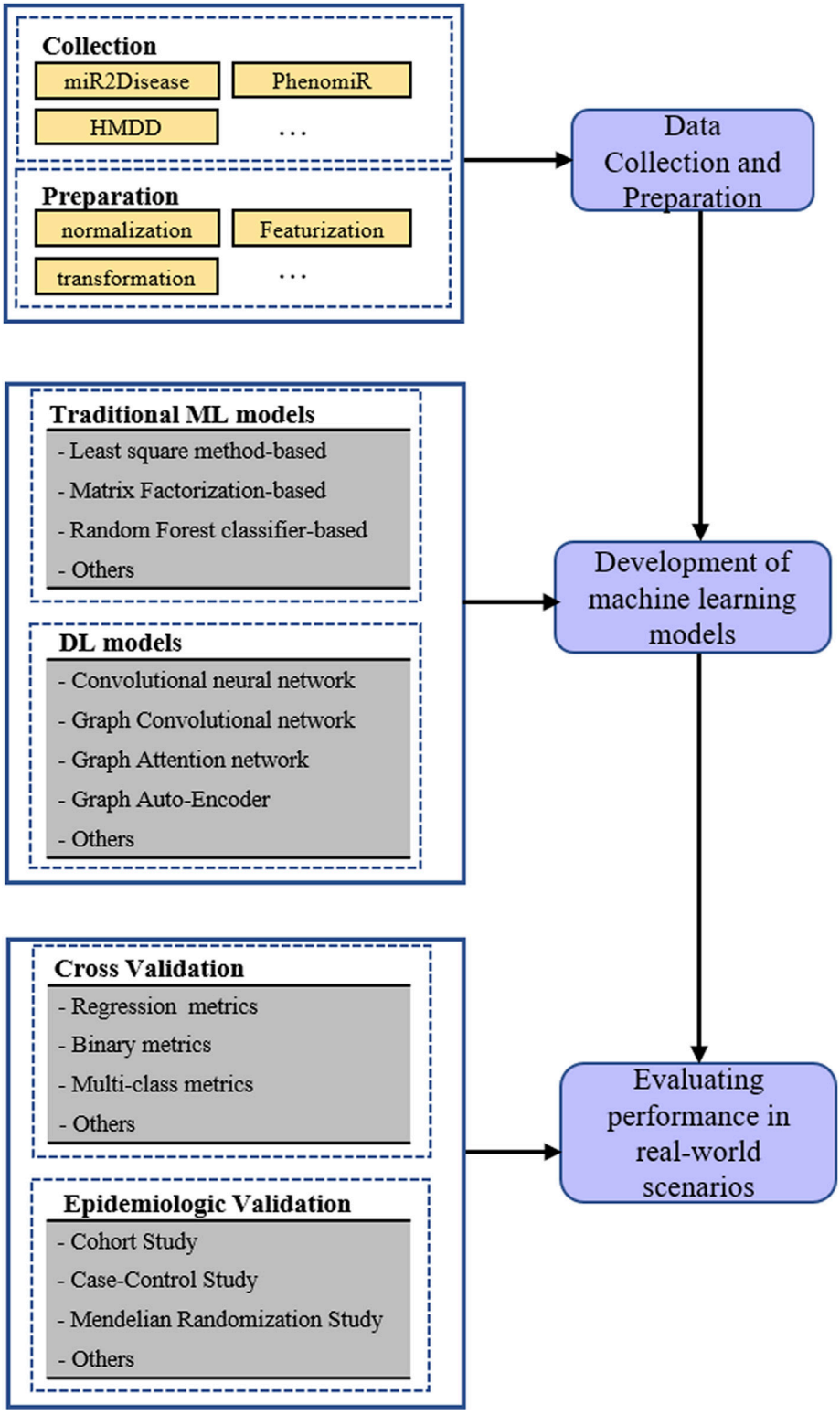


FIGURE 1
The processes of machine learning models for predicting miRNA-disease associations.

3.1 Traditional machine learning models for miRNA disease associations

As an example of using a negative training sample, a previous study (Ji et al., 2020) learned graph representations with global structure knowledge in a heterogeneous network consisting of the known associations among miRNA, disease, drug, and protein. Integrating these embeddings with miRNA sequences, disease semantic similarities and so on, a classifier based on Random Forest was applied to discover underlying relationships between miRNAs and diseases.

Meanwhile, more and more approaches preferred to predict unknown miRNA-disease associations only with known ones, so researchers utilized verified associations, such as miRNA-disease, miRNA-gene, and weighted gene-gene, to construct a regularized framework for inferring the latent miRNA-disease associations (Peng et al., 2017a). Similarly, using the identified disease-associated miRNA information, (Luo et al., 2018) built a semi-supervised classifier to calculate the probability of a miRNA related to a given disease, and also utilized graph regularization to avoid overfitting. Considering the sparsity of known data, have also (Luo et al., 2016) proposed a transductive learning-based collective prediction method in which the relevance score was calculated and updated *via* the disease-miRNA network.

To adequately discover disease-related candidate miRNAs, in (Ding et al., 2018) for example, a heterogeneous disease-gene-miRNA network consisting of three types of nodes and five types of links was built to predict associations *via* a regression-based model. For fully utilizing verified miRNA-disease associations. In (Pan et al., 2019), the miRNA-disease associations were synchronously predicted and updated *via* a multi-label, graph-based model, which firstly introduced a set of kernel matrices and then adaptively obtained two optimal kernel matrices. Considering the inherent noise in current databases, a study in (Liang et al., 2019) adaptively learned an affinity graph from various similarity profiles and simultaneously updated the prediction *via* multi-label learning. According to the latest version of HMDD, a study in (Liang et al., 2018) obtained the semantic similarities of disease and function similarities of miRNA. Then, the similarity matrices and association matrix were iteratively updated to generate the optimized association outcome.

Matrix factorization, a method of multiplying two different entities to generate potential features, is another essential method for predicting miRNA disease associations. As in (Peng et al., 2017b) for example, a matrix recovery approach was utilized to integrate the weight matrix to recover association matrix; hence novel latent associations were accurately inferred without the need for negative samples. Integrated with the label propagation algorithm, a study in (Peng et al., 2022) adopted robust nonnegative matrix factorization to predict underlying associations more precisely. To be specific, using the integrated similarity information, the original adjacency matrix was updated *via* matrix multiplication to reduce the influence of negative samples. For sparse existing associations and new diseases or miRNAs, a previous work (Xiao et al., 2018) developed a preprocessing step that built the interaction score profiles to facilitate prediction, and then utilized graph regularized non-negative matrix factorization based on integrated multisource data to discover underlying associations.

Although most methods *in silico* currently focus on discovering unknown miRNA-disease associations, there are some approaches that could identify the multiple relationship types among various associations as the roles miRNAs played in diseases significantly diverged. For example, the down-regulation of mir-16 and mir-15 could induce chronic lymphocytic leukemia in B cell (Calin et al., 2002), while the different expression of serum miRNAs, such as mir-1307-3p, mir-1246 and so on, could assist researchers in tracing breast cancer early (Shimomura et al., 2016). To this end, a more recent study (Huang et al., 2021) innovatively constructed a tensor composed of miRNA-disease-type triples, and then adopted tensor decomposition that utilized the similarity information as decomposition constraints to detect multi-type of miRNA-disease associations. Another study built a novel model for miRNA-disease-type associations by applying tensor robust principal component analysis (Yu et al., 2021a).

3.2 Deep learning models for miRNA disease associations

Currently, many prediction methods extracted feature embeddings as the input of convolutional neural networks (CNN). Xuan et al. (2018) constructed a dual convolutional neural network, which was divided into the left and right part, to detect underlying associations. The left CNN learned the integrated feature embedding of original information to produce an association score, and the right learned the feature embedding of the network topology to generate the other score. On this basis, a work in (Xuan et al., 2019) firstly projected nodes of miRNAs and diseases into a low dimensional space to obtain feature embeddings, and then utilized network representation learning and two CNN to discover latent disease-associated miRNAs. In (Peng et al., 2019), the low dimensional feature embeddings were selected by an auto-encoder from a three-layer network consisting of multisource data. Then, the association score was calculated by a deep CNN structure, including the fully-connected layer, max-pooling layer, and convolutional layer.

Besides, some Graph Convolutional Network (GCN) based end-to-end models were also implemented to capture candidate associations. In 2020, a work (Li et al., 2020) respectively learned underlying feature embeddings derived from the miRNA function similarity network and the disease semantic

similarity network with GCN encoders. Then an association matrix completion was generated from a novel neural inductive model that adopted learned embeddings as input. As in (Chu et al., 2021), a miRNA-disease pair was regarded as a node in homogeneous graphs, which were easier to learn. Then based on graph sampling, the modified GCN algorithm was implemented on the topology and feature graph to cluster similar nodes. Meanwhile, some other graph neural network methods were also employed in this regard. A graph attention network-based method (Li et al., 2022) aggregated different neighbor information with varying weights to obtain the non-linear features of miRNAs and diseases. Combined with the linear features constructed by correlation profiles, latent miRNA-disease associations were inferred *via* the random forest algorithm. In 2021, Li et al. (2021a) developed an end-to-end framework based on a novel graph auto-encoder model to discover unknown associations. This model aggregated nodes' neighborhood information *via* a graph neural network-based encoder, which consisted of the multi-layer perceptron and aggregator function, to obtain low dimensional embeddings and effectively integrate heterogeneous information.

Some methods aimed at predicting type instead of taking association prediction as a binary task. In (Huang et al., 2021) for example, miRNA-type- disease triples were innovatively regarded as a tensor, and then tensor decomposition with relation constraints was implemented to complete the type prediction task. Similarly, a more recent work (Yu et al., 2022) could identify dysregulation, downregulation, or upregulation relationship between miRNA and disease because a depth graph representation learning model was trained based on a knowledge graph constructed by extracting disease-miRNA-type triples from existing databases and numerous experimental data.

To fully understand the synergistic effect of miRNA-miRNA pairs on the pathogenesis of complicated diseases, a study (Luo et al., 2021) proposed a new tensor decomposition model based on a graph attention network to discover potential miRNA-miRNA pairs related to diseases. The graph attention network aggregated the feature embeddings from the miRNA function similarity graph, disease semantic similarity graph, and miRNA sequence similarity graph. With the aggregated feature embeddings, the deep tensor factorization was implemented to reconstruct the association tensor consisting of miRNA-miRNA-disease triples.

4 Predicting miRNA-drug associations

With the accumulated research on miRNA-small-molecule interactions, computational approaches attract more and more attention because they can efficiently promote miRNA-targeted drug discovery and optimization when compared to conventional routine. Varieties of computational models were proposed to discover latent miRNA-drug candidates. Generally speaking, they can be classified into two kinds of approaches for predicting: the traditional machine learning method and the deep learning method, as shown in Figure 2.

4.1 Traditional machine learning models for miRNA drug associations

Some machine learning methods focused on constructing novel feature engineering with varied features. A random forest prediction model (Wang et al., 2019) adopted similarities of miRNAs and small molecules as features to accurately predict associations. Specifically for cancer, (Li et al., 2021b) innovatively concatenated features extracted from small molecule structures, miRNA sequences, and cancer symptoms to obtain a new feature vector. Then a random forest model was utilized to predict latent cancer-miRNA-small molecule associations. Similarly, Jamal et al. (2012) developed a prediction model by utilizing Naïve Bayes and Random Forest. In 2017, a work (Xie et al., 2017) was proposed to discover the influential miRNA on the drug *via* the support vector machine, in which feature vectors were drug-miRNA pairs extracted from the related literature.

There are some methods based on random walk algorithm to identify latent miRNA-small molecule associations. In (Liu et al., 2020) for example, Random Walk was utilized in a triple-layer heterogeneous network of disease-miRNA-small molecule association after computing similarities and selecting negative samples. Similarly, a restart algorithm-based Random Walk (Lv et al., 2015) was implemented in a comprehensive network, in which miRNA-miRNA associations, small molecule interactions, and verified miRNA-small molecule targeting pairs were integrated. Meanwhile, some other methods are based on regression algorithm. In Chen et al. (2021) for example, a matrix was defined to represent a heterogeneous network consisting of small molecule similarity, miRNA similarity, and verified miRNA-small molecule associations. Then, the model of the Alternating Direction Method of Multipliers was designed to minimize the nuclear norm of the matrix and obtain predicted scores of underlying miRNA-small molecule associations. Likewise, a work (Wang et al., 2022) developed a prediction model based on the Ensemble of Kernel Ridge Regression. They integrated feature dimensionality reduction with ensemble learning to discover latent small molecule-microRNA associations.

It can be seen in various studies that many computational models adopted matrix factorization. In Yin et al. (2019) for example, a sparse learning method (SLM) was proposed to eliminate noises and improve performance. After the small molecule-miRNA adjacency matrix was decomposed by SLM, latent miRNA-small molecule associations would be obtained *via* a heterogeneous graph that integrated the similarities of miRNAs

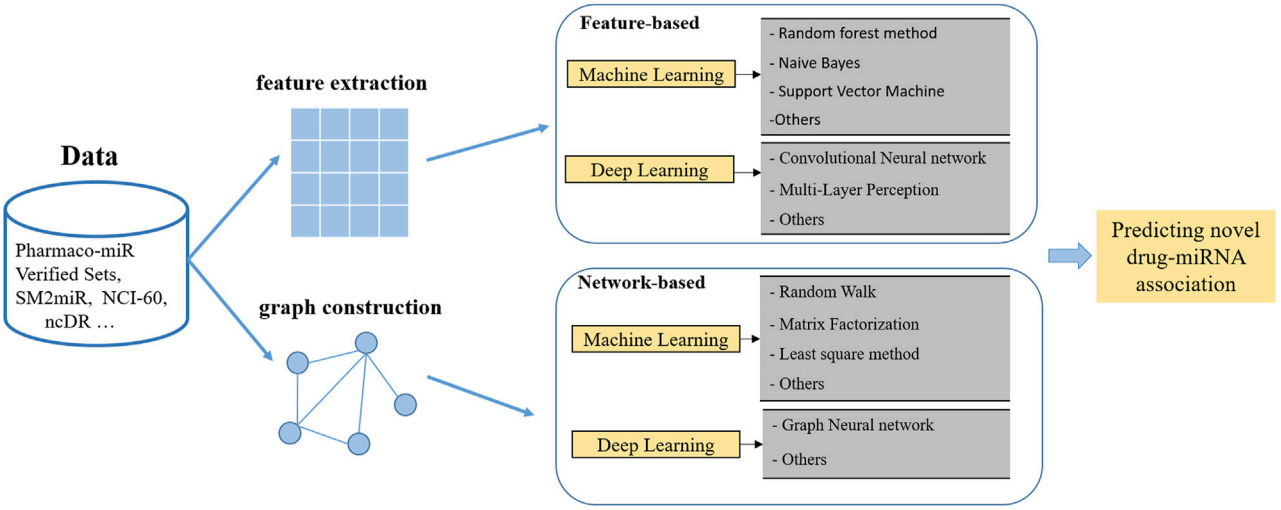


FIGURE 2
The process of computational models for identifying miRNA-drug associations.

and small molecules with the improved association information. At the same time, Wang and Chen (2019) not only adopted similarities of small molecules, miRNAs, and diseases but also integrated with associations between miRNAs and diseases/small molecule. Therefore, a three-layer network was built to obtain potential representations of small molecule-miRNA association *via* in-layer similarities and cross-layer associations. Then cross-layer dependency inference on the three-layer network was utilized to identify unknown miRNA-small molecule associations. In addition, the model adopted a regularized optimization to avoid overfitting. Afterward, a study (Zhao et al., 2020) applied matrix decomposition in integrated similarity matrixes and obtained small molecule-miRNA pair similarity by calculating the Kronecker product. Additionally, regularized least square method was applied to acquire the mapping relationships between associated probabilities and miRNA-small molecule pairs. Considering the functional similarity of two miRNAs, clinical similarity and chemical similarity of small molecules, a work (Luo et al., 2020) adopted a nonnegative matrix decomposition method for discovering the potential miRNA-small molecule associations. Besides, combining small molecule-disease associations with miRNA-disease associations, Shen et al. (2020a) adopted graph regularization techniques and the iterative approach in a heterogeneous network to obtain the prediction scores of miRNA-small molecule pairs. In Shen et al. (2020b), the prediction performance was improved by a Restricted Boltzmann Machine-based joint learning framework, which integrated miRNA sequence, heterogeneous network knowledge, and small molecule structure data.

4.2 Deep learning models for miRNA drug associations

Currently, Graph Convolution Network is commonly used to process node classification tasks in the homogeneous network. In Huang et al. (2020) for example, a three-layer latent factor model based on graph convolution was developed to discover unknown miRNA-drug resistance associations. In this end-to-end learning scheme, they could not only utilize high-dimensional attributes but also learn graph embedding features of miRNAs/drugs. To overcome the problem of over-smoothing in conventional graph convolution networks, a work (Yu et al., 2021b) simplified GCN by constructing the embedding propagation layer utilizing a weighted sum aggregator. Then, the ideal representations were obtained by summing over the embeddings in each layer. At last, they applied the inner product to discover the unknown miRNA-drug sensitivity associations. Wang et al. (2021) firstly extracted drug/miRNA representations *via* a layer attention graph convolution network in the heterogeneous network consisting of known drug similarities, miRNA similarities, and drug-miRNA interactions. Then they obtained the drug/miRNA embedding vectors by concatenating their representations with drug features derived from drug molecular graphs, and the miRNA expression features, respectively. In addition, they utilized compressed tensor network, tensor decomposition, and multi-layer perceptron to extract node-pair embeddings. Eventually, the potential relationship between miRNA and drug resistance was predicted by the completely connected layer with concatenated representations. Similarly focused on prediction for the relationship of miRNA-drug resistance (Zhao et al., 2022), constructed a graph neural network based on positional

encoding to extract embeddings from drug molecular graphs and miRNA-drug heterogeneous networks. Then, these embeddings of different layers were combined with a layer attention mechanism to learn powerful feature representations. Finally, the potential miRNA-drug resistance association could be discovered *via* a multi-channel neural network consisting of tensor network, tensor decomposition, and the multi-layer perceptron.

Besides, there are some other deep learning models based on varied neural networks algorithm. In Deepthi and Jereesh (2021), firstly, the principal component analysis was applied to reduce the dimensions of features extracted from the integrated similarity pairs of drugs and miRNAs. Then, they trained a convolutional neural network to obtain deep retrieved features and adopted the support vector machine classifier to predict latent association. Meanwhile, based on Long Short-Term Memory (LSTM) (Abdelbaky et al., 2021), proposed an encoder-decoder model that could perform on the character level of a sequence. They utilized the LSTM Sequence Auto-Encoders to obtain feature embeddings of miRNAs and small molecules, and sequence-to-sequence learning with an RNN to encode sequences. The input sequence reproduced by the decoder was based on the outcome of the encoder.

5 Conclusion

As the miRNA-related data is explosively growing, developing advanced computational methods for miRNA therapy is not only an opportunity but also a challenge for medical research. Taking advantage of the traditional machine learning method and deep learning method, the discovery of unknown associations among drugs, diseases, and miRNAs could be greatly anticipated. Furthermore, the prediction results of machine learning models could be compared to miRNA-disease/drug associations validated in experimental methods. In this review, we collected commonly used data sources of miRNA-disease and miRNA-drug, which laid a solid foundation for designing feasible prediction models. Various machine learning-based methods were classified into two parts: predicting potential miRNA-disease association and discovering latent miRNA-drug associations, which facilitated exploring miRNA therapy.

Although machine-learning methods have exhibited tremendous potential, it is still a big challenge to accelerate development in miRNA therapy by adopting data-driven computational approaches. This could be improved by utilizing high-quality data resources and integrating domain knowledge when selecting feature to build and verify models. Nevertheless, considering the experimental data might be unavailable for some miRNA, or only a few data points are accessible, reliable models are difficult to construct. Therefore, machine learning approaches like active learning might be a promising strategy to cope with the limitation of available data used to construct reliable prediction models. Meanwhile, generalizability is essential for the widespread application of machine learning approaches, and it could be examined *via* external validation or cross-validation in their proposed model based on machine learning. Recent work adopted anchor regression once a linear shift made training set and test set distributions varied (Rothenhäusler et al., 2018). Different from the "black box" design in which a specific output conducted by a model cannot be explained, machine learning/deep learning models with understandable results or analytical processes are explainable artificial intelligence (Sample, 2017). It is of great importance for domains like miRNA therapy, in which an understandable relationship between outcomes and features is essential. In general, machine learning explainable tools can be mainly divided into two methods: 1) The local model explainability method is helpful to discover which specific features affected a specific decision; 2) The global model explainability method is centered on the features that most affect all decisions or the model's results. Recently, an emerging field as machine learning fairness has been proposed to study the role of data biases and model biases like race, gender, disabilities and so on, played in the prediction performance in miRNA therapy.

Author contributions

YL conceived and wrote the manuscript. LP and WS co-wrote the manuscript. MS, LL, and WL commented on the manuscript. WL supervised YL and polished the manuscript.

Funding

This work has been supported by the National Natural Science Foundation of China (Grant no.61902125) and the Scientific Research Startup Foundation of University of South China (Grant no. 190XQD096).

Conflict of interest

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

Publisher's note

References

Abdelbaky, I., Tayara, H., and Chong, K. T. (2021). Identification of miRNA-small molecule associations by continuous feature representation using auto-encoders. *Pharmaceutics* 14, 3. doi:10.3390/pharmaceutics14010003

Blower, P. E., Verducci, J. S., Lin, S., Zhou, J., Chung, J.-H., Dai, Z., et al. (2007). MicroRNA expression profiles for the NCI-60 cancer cell panel. *Mol. cancer Ther.* 6, 1483–1491. doi:10.1158/1535-7163.MCT-07-0009

Calin, G. A., Cimmino, A., Fabbri, M., Ferracin, M., Wojcik, S. E., Shimizu, M., et al. (2008). MiR-15a and miR-16-1 cluster functions in human leukemia. *Proc. Natl. Acad. Sci.* 105, 5166–5171. doi:10.1073/pnas.0800121105

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., et al. (2002). Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci.* 99, 15524–15529. doi:10.1073/pnas.242606799

Chandrasekhar, S., Pushpavalli, S. N., Chatla, S., Mukhopadhyay, D., Ganganna, B., Vijeender, K., et al. (2012). aza-Flavanones as potent cross-species microRNA inhibitors that arrest cell cycle. *Bioorg. Med. Chem. Lett.* 22, 645–648. doi:10.1016/j.bmcl.2011.10.061

Chen, X., Guan, N.-N., Sun, Y.-Z., Li, J.-Q., and Qu, J. (2020). MicroRNA-small molecule association identification: From experimental results to computational models. *Briefings Bioinforma.* 21, 47–61.

Chen, X., Zhou, C., Wang, C.-C., and Zhao, Y. (2021). Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization. *Briefings Bioinforma.* 22, bbab328. doi:10.1093/bib/bbab328

Chu, Y., Wang, X., Dai, Q., Wang, Y., Wang, Q., Peng, S., et al. (2021). MDA-GCNFTG: Identifying miRNA-disease associations based on graph convolutional networks via graph sampling through the feature and topology graph. *Brief. Bioinform* 22, bbab165. doi:10.1093/bib/bbab165

Dai, E., Yang, F., Wang, J., Zhou, X., Song, Q., An, W., et al. (2017). ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance. *Bioinformatics* 33, 4010–4011. doi:10.1093/bioinformatics/btx523

Deepthi, K., and Jereesh, A. (2021). An ensemble approach based on multi-source information to predict drug-MiRNA associations via convolutional neural networks. *IEEE Access* 9, 38331–38341. doi:10.1109/access.2021.3063885

Ding, P., Luo, J., Liang, C., Xiao, Q., and Cao, B. (2018). Human disease MiRNA inference by combining target information based on heterogeneous manifolds. *J. Biomed. Inf.* 80, 26–36. doi:10.1016/j.jbi.2018.02.013

Fernández-Hernando, C., Ramírez, C. M., Goedeke, L., and Suárez, Y. (2013). MicroRNAs in metabolic disease. *Arteriosclerosis, thrombosis, Vasc. Biol.* 33, 178–185. doi:10.1161/ATVBAHA.112.300144

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J. L., Haberle, V., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi:10.1038/nature13182

Fu, L., and Peng, Q. (2017). A deep ensemble model to predict miRNA-disease association. *Sci. Rep.* 7, 14482–14513. doi:10.1038/s41598-017-15235-6

Fulci, V., Chiaretti, S., Goldoni, M., Azzalin, G., Carucci, N., Tavolaro, S., et al. (2007). Quantitative technologies establish a novel microRNA profile of chronic lymphocytic leukemia. *Blood, J. Am. Soc. Hematol.* 109, 4944–4951. doi:10.1182/blood-2006-12-062398

Gumireddy, K., Young, D. D., Xiong, X., Hogenesch, J. B., Huang, Q., and Deiters, A. (2008). Small-molecule inhibitors of microrna miR-21 function. *Angew. Chem.* 120, 7482–7484. doi:10.1002/anie.200801555

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* 466, 835–840. doi:10.1038/nature09267

Huang, F., Yue, X., Xiong, Z., Yu, Z., Liu, S., and Zhang, W. (2021). Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Briefings Bioinforma.* 22, bbaa140. doi:10.1093/bib/bbaa140

Huang, H.-Y., Lin, Y.-C.-D., Cui, S., Huang, Y., Tang, Y., Xu, J., et al. (2022). miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic acids Res.* 50, D222–D230. doi:10.1093/nar/gkab1079

Huang, L., Zhang, L., and Chen, X. (2022). Updated review of advances in microRNAs and complex diseases: Experimental results, databases, webservers and data fusion. *Briefings Bioinforma.* 23, bbac397. doi:10.1093/bib/bbac397

Huang, Y.-a., Hu, P., Chan, K. C., and You, Z.-H. (2020). Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics* 36, 851–858. doi:10.1093/bioinformatics/btz621

Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3. 0: A database for experimentally supported human microRNA–disease associations. *Nucleic acids Res.* 47, D1013–D1017. doi:10.1093/nar/gky1010

Jamal, S., Periwal, V., and Scaria, V. (2012). Computational analysis and predictive modeling of small molecule modulators of microRNA. *J. cheminformatics* 4, 16–19. doi:10.1186/1758-2946-4-16

Ji, B.-Y., You, Z.-H., Cheng, L., Zhou, J.-R., Alghazzawi, D., and Li, L.-P. (2020). Predicting miRNA-disease association from heterogeneous information network with GraRep embedding model. *Sci. Rep.* 10, 6658–6712. doi:10.1038/s41598-020-63735-9

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. doi:10.1093/nar/gkn714

Kutay, H., Bai, S., Datta, J., Motiwala, T., Pogribny, I., Frankel, W., et al. (2006). Downregulation of miR-122 in the rodent and human hepatocellular carcinomas. *J. Cell. Biochem.* 99, 671–678. doi:10.1002/jcb.20982

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843–854. doi:10.1016/0092-8674(93)90529-y

Li, G., Fang, T., Zhang, Y., Liang, C., Xiao, Q., and Luo, J. (2022). Predicting miRNA-disease associations based on graph attention network with multi-source information. *BMC Bioinforma.* 23, 244–324. doi:10.1186/s12859-022-04796-7

Li, J., Peng, D., Xie, Y., Dai, Z., Zou, X., and Li, Z. (2021). Novel potential small molecule–MiRNA–cancer associations prediction model based on fingerprint, sequence, and clinical symptoms. *J. Chem. Inf. Model.* 61, 2208–2219. doi:10.1021/acs.jcim.0c01458

Li, J., Zhang, S., Liu, T., Ning, C., Zhang, Z., and Zhou, W. (2020). Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 36, 2538–2546. doi:10.1093/bioinformatics/btz965

Li, Z., Li, J., Nie, R., You, Z.-H., and Bao, W. (2021). A graph auto-encoder model for miRNA-disease associations prediction. *Briefings Bioinforma.* 22, bbaa240. doi:10.1093/bib/bbaa240

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

Liang, C., Yu, S., Wong, K.-C., and Luo, J. (2018). A novel semi-supervised model for miRNA-disease association prediction based on $$\ell_{1} $$ ℓ 1-norm graph. *J. Transl. Med.* 16, 357–412. doi:10.1186/s12967-018-1741-y

Liu, F., Peng, L., Tian, G., Yang, J., Chen, H., Hu, Q., et al. (2020). Identifying small molecule-miRNA associations based on credible negative sample selection and random walk. *Front. Bioeng. Biotechnol.* 8, 131. doi:10.3389/fbioe.2020.00131

Liu, X., Wang, S., Meng, F., Wang, J., Zhang, Y., Dai, E., et al. (2013). SM2miR: A database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* 29, 409–411. doi:10.1093/bioinformatics/bts698

Luo, J., Ding, P., Liang, C., Cao, B., and Chen, X. (2016). Collective prediction of disease-associated miRNAs based on transduction learning. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 14, 1468–1475. doi:10.1109/TCBB.2016.2599866

Luo, J., Ding, P., Liang, C., and Chen, X. (2018). Semi-supervised prediction of human miRNA-disease association based on graph regularization framework in heterogeneous networks. *Neurocomputing* 294, 29–38. doi:10.1016/j.neucom.2018.03.003

Luo, J., Lai, Z., Shen, C., Liu, P., and Shi, H. (2021). "Graph attention mechanism-based deep tensor factorization for predicting disease-associated miRNA-miRNA pairs," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, United States, 09-12 December 2021 (IEEE), 189–196.

Luo, J., Shen, C., Lai, Z., Cai, J., and Ding, P. (2020). Incorporating clinical, chemical and biological information for predicting small molecule-microRNA associations based on non-negative matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18, 2535–2545. doi:10.1109/TCBB.2020.2975780

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31, 3638–3644. doi:10.1093/bioinformatics/btv417

Marquart, T. J., Allen, R. M., Ory, D. S., and Baldán, Á. (2010). miR-33 links SREBP-2 induction to repression of sterol transporters. *Proc. Natl. Acad. Sci.* 107, 12228–12232. doi:10.1073/pnas.1005191107

Melo, S., Villanueva, A., Moutinho, C., Davalos, V., Spizzo, R., Ivan, C., et al. (2011). Small molecule enoxacin is a cancer-specific growth inhibitor that acts by

enhancing TAR RNA-binding protein 2-mediated microRNA processing. *Proc. Natl. Acad. Sci.* 108, 4394–4399. doi:10.1073/pnas.1014720108

Pan, Z., Zhang, H., Liang, C., Li, G., Xiao, Q., Ding, P., et al. (2019). Self-weighted multi-kernel multi-label learning for potential miRNA-disease association prediction. *Mol. Therapy-Nucleic Acids* 17, 414–423. doi:10.1016/j.omtn.2019.06.014

Panwar, B., Omenn, G. S., and Guan, Y. (2017). miRmine: a database of human miRNA expression profiles. *Bioinformatics* 33, 1554–1560. doi:10.1093/bioinformatics/btx019

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi:10.1093/bioinformatics/btz254

Peng, L., Peng, M., Liao, B., Huang, G., Liang, W., and Li, K. (2017). Improved low-rank matrix recovery method for predicting miRNA-disease association. *Sci. Rep.* 7, 6007–6010. doi:10.1038/s41598-017-06201-3

Peng, L., Peng, M., Liao, B., Xiao, Q., Liu, W., Huang, G., et al. (2017). A novel information fusion strategy based on a regularized framework for identifying disease-related microRNAs. *RSC Adv.* 7, 44447–44455. doi:10.1039/c7ra08894a

Peng, L., Yang, C., Huang, L., Chen, X., Fu, X., and Liu, W. (2022). Rnmflp: Predicting circRNA–disease associations based on robust nonnegative matrix factorization and label propagation. *Briefings Bioinforma.* 23, bbac155. doi:10.1093/bib/bbac155

Perdikopanis, N., Georgakilas, G. K., Grigoriadis, D., Pierros, V., Kavakiotis, I., Alexiou, P., et al. (2021). DIANA-miRGen v4: Indexing promoters and regulators for more than 1500 microRNAs. *Nucleic acids Res.* 49, D151–D159. doi:10.1093/nar/gkaa1060

Roldo, C., Missiaglia, E., Hagan, J. P., Falconi, M., Capelli, P., Bersani, S., et al. (2006). MicroRNA expression abnormalities in pancreatic endocrine and acinar tumors are associated with distinctive pathologic features and clinical behavior. *J. Clin. Oncol.* 24, 4677–4684. doi:10.1200/JCO.2005.05.5194

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2018). Anchor regression: Heterogeneous data meets causality. Available at http://org.arXiv/abs/1801.06229.

Ruepp, A., Kowarsch, A., Schmidl, D., Buggenthin, F., Brauner, B., Dunger, I., et al. (2010). PhenomiR: A knowledgebase for microRNA expression in diseases and biological processes. *Genome Biol.* 11, R6–R11. doi:10.1186/gb-2010-11-1-r6

Rukov, J. L., Wilentzik, R., Jaffe, I., Vinther, J., and Shomron, N. (2014). Pharmaco-miR: Linking microRNAs and drug effects. *Briefings Bioinforma.* 15, 648–659. doi:10.1093/bib/bbs082

Rupaimoole, R., and Slack, F. J. (2017). MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.* 16, 203–222. doi:10.1038/nrd.2016.246

Sample, I. (2017). Computer says no: Why making AIs fair, accountable and transparent is crucial. *Guard.* 5, 1–15.

Shen, C., Luo, J., Lai, Z., and Ding, P. (2020). Multiview joint learning-based method for identifying small-molecule-associated MiRNAs by integrating pharmacological, genomics, and network knowledge. *J. Chem. Inf. Model.* 60, 4085–4097. doi:10.1021/acs.jcim.0c00244

Shen, C., Luo, J., Ouyang, W., Ding, P., and Wu, H. (2020). Identification of small molecule–miRNA associations with graph regularization techniques in heterogeneous networks. *J. Chem. Inf. Model.* 60, 6709–6721. doi:10.1021/acs.jcim.0c00975

Shimomura, A., Shiino, S., Kawauchi, J., Takizawa, S., Sakamoto, H., Matsuzaki, J., et al. (2016). Novel combination of serum microRNA for detecting breast cancer in the early stage. *Cancer Sci.* 107, 326–334. doi:10.1111/cas.12880

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., et al. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* 100, 15776–15781. doi:10.1073/pnas.2136655100

Takamizawa, J., Konishi, H., Yanagisawa, K., Tomida, S., Osada, H., Endoh, H., et al. (2004). Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Res.* 64, 3753–3756. doi:10.1158/0008-5472.CAN-04-0637

Tay, Y., Zhang, J., Thomson, A. M., Lim, B., and Rigoutsos, I. (2008). MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* 455, 1124–1128. doi:10.1038/nature07299

Wang, C.-C., and Chen, X. (2019). A unified framework for the prediction of small molecule–MicroRNA association based on cross-layer dependency inference on multilayered networks. *J. Chem. Inf. Model.* 59, 5281–5293. doi:10.1021/acs.jcim.9b00667

Wang, C.-C., Chen, X., Qu, J., Sun, Y.-Z., and Li, J.-Q. (2019). Rfsmma: A new computational model to identify and prioritize potential small molecule–mirna associations. *J. Chem. Inf. Model.* 59, 1668–1679. doi:10.1021/acs.jcim.9b00129

Wang, C.-C., Zhu, C.-C., and Chen, X. (2022). Ensemble of kernel ridge regression-based small molecule–miRNA association prediction in human disease. *Briefings Bioinforma.* 23, bbab431. doi:10.1093/bib/bbab431

Wang, H., Khan, S., Liu, S., Zheng, F., and Zhang, W. (2021).“Predicting drug-miRNA resistance with layer attention graph convolution network and multi channel feature extraction,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, United States, 09-12 December 2021. (IEEE), 1083–1089.

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

Xie, W.-B., Yan, H., and Zhao, X.-M. (2017). EmDL: Extracting miRNA-drug interactions from literature. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 16, 1722–1728. doi:10.1109/TCBB.2017.2723394

Xuan, P., Dong, Y., Guo, Y., Zhang, T., and Liu, Y. (2018). Dual convolutional neural network based method for predicting disease-related miRNAs. *Int. J. Mol. Sci.* 19, 3732. doi:10.3390/ijms19123732

Xuan, P., Sun, H., Wang, X., Zhang, T., and Pan, S. (2019). Inferring the disease-associated miRNAs based on network representation learning and convolutional neural networks. *Int. J. Mol. Sci.* 20, 3648. doi:10.3390/ijms20153648

Yin, J., Chen, X., Wang, C.-C., Zhao, Y., and Sun, Y.-Z. (2019). Prediction of small molecule–microRNA associations by sparse learning and heterogeneous graph inference. *Mol. Pharm.* 16, 3157–3166. doi:10.1021/acs.molpharmaceut.9b00384

Young, D. D., Connelly, C. M., Grohmann, C., and Deiters, A. (2010). Small molecule modifiers of microRNA miR-122 function for the treatment of hepatitis C virus infection and hepatocellular carcinoma. *J. Am. Chem. Soc.* 132, 7976–7981. doi:10.1021/ja910275u

Yu, N., Liu, Z.-P., and Gao, R. (2021). “A semi-supervised learning algorithm for predicting MiRNA-disease association,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, United States, 09-12 December 2021 (IEEE), 771–774.

Yu, S., Wang, H., Liu, T., Liang, C., and Luo, J. (2022). A knowledge-driven network for fine-grained relationship detection between miRNA and disease. *Briefings Bioinforma.* 23, bbac058. doi:10.1093/bib/bbac058

Yu, S., Xu, H., Li, Y., Liu, D., and Deng, L. (2021). “Lgcmds: Predicting miRNA-drug sensitivity based on light graph convolution network,” in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, United States, 09-12 December 2021 (IEEE), 217–222.

Zhang, Z., Liu, S., Shi, R., and Zhao, G. (2011). miR-27 promotes human gastric cancer cell metastasis by inducing epithelial-to-mesenchymal transition. *Cancer Genet.* 204, 486–491. doi:10.1016/j.cancergen.2011.07.004

Zhao, C., Wang, H., Qi, W., and Liu, S. (2022). Toward drug-miRNA resistance association prediction by positional encoding graph neural network and multi-channel neural network. *Methods* 207, 81–89. doi:10.1016/j.ymeth.2022.09.005

Zhao, Y., Chen, X., Yin, J., and Qu, J. (2020). Snmfsmma: Using symmetric nonnegative matrix factorization and kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol.* 17, 281–291. doi:10.1080/15476286.2019.1694732

frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 17 February 2023
DOI 10.3389/fbinf.2023.1123993

OPEN ACCESS

EDITED BY
Chen Li,
Monash University, Australia

REVIEWED BY
Jiayuan Huang,
Sun Yat-sen University, China
Jinxin Zhao,
Faculty of Medicine, Nursing and Health Sciences, Monash University, Australia
Quanzhong Liu,
Northwest A&F University, China

*CORRESPONDENCE
Chee Keong Kwoh,
asckkwoh@ntu.edu.sg

SPECIALTY SECTION
This article was submitted to Protein Bioinformatics, a section of the journal Frontiers in Bioinformatics

RECEIVED 22 December 2022
ACCEPTED 03 February 2023
PUBLISHED 17 February 2023

CITATION
Ng TA, Rashid S and Kwoh CK (2023), Virulence network of interacting domains of influenza a and mouse proteins. *Front. Bioinform.* 3:1123993. doi: 10.3389/fbinf.2023.1123993

Virulence network of interacting domains of influenza a and mouse proteins

Teng Ann Ng, Shamima Rashid and Chee Keong Kwoh*

School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

There exist several databases that provide virus-host protein interactions. While most provide curated records of interacting virus-host protein pairs, information on the strain-specific virulence factors or protein domains involved, is lacking. Some databases offer incomplete coverage of influenza strains because of the need to sift through vast amounts of literature (including those of major viruses including HIV and Dengue, besides others). None have offered complete, strain specific protein-protein interaction records for the influenza A group of viruses. In this paper, we present a comprehensive network of predicted domain-domain interaction(s) (DDI) between influenza A virus (IAV) and mouse host proteins, that will allow the systematic study of disease factors by taking the virulence information (lethal dose) into account. From a previously published dataset of lethal dose studies of IAV infection in mice, we constructed an interacting domain network of mouse and viral protein domains as nodes with weighted edges. The edges were scored with the Domain Interaction Statistical Potential (DISPOT) to indicate putative DDI. The virulence network can be easily navigated *via* a web browser, with the associated virulence information (LD_{50} values) prominently displayed. The network will aid influenza A disease modeling by providing strain-specific virulence levels with interacting protein domains. It can possibly contribute to computational methods for uncovering influenza infection mechanisms mediated through protein domain interactions between viral and host proteins. It is available at https://iav-ppi.onrender.com/home.

KEYWORDS

influenza a, lethal dose 50, protein, virulence, mouse model, domain-domain interaction

1 Introduction

Influenza A virus (IAV) is a single stranded, positive ribonucleic acid (RNA) virus that is a respiratory pathogen across many species such as humans, swine, and wild waterfowl. It consists of eight genomic segments which encode at least 11 proteins. The structure and organization of the virus particle is shown in Figure 1, which is reproduced here from the work of Jung and Lee (Jung and Lee, 2020).

Hemagglutinin (HA) and neuraminidase (NA) on the viral particle surface, are the proteins responsible for mediating entry into and cleavage from the host cell, respectively. Matrix protein 1 (M1) is a component of the viral envelop while matrix protein 2 (M2) is found below the lipid bilayer of the viral membrane, strengthening it. Together with the nucleoprotein (NP), they form the ribonucleoprotein complex (indicated as vRNP in Figure 1). The final three proteins are the polymerase basic 1 frame 2 (PB1-F2), and non-structural proteins 1 and 2 (NS1 and NS2), respectively.

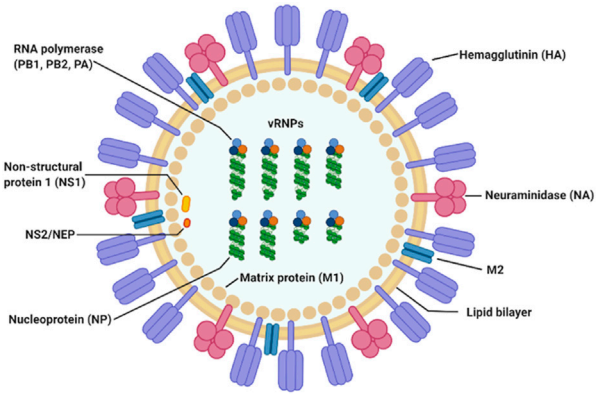


FIGURE 1
IAV particle and its fully assembled constituent proteins. Genomic RNA segments are shown in green, wrapped around nucleoproteins. The hetero-trimeric RNA-dependent RNA polymerase complex comprising of PB1, PB2, and PA is shown in orange, light and dark blue circles. This figure is by Jung and Lee (Jung and Lee, 2020).

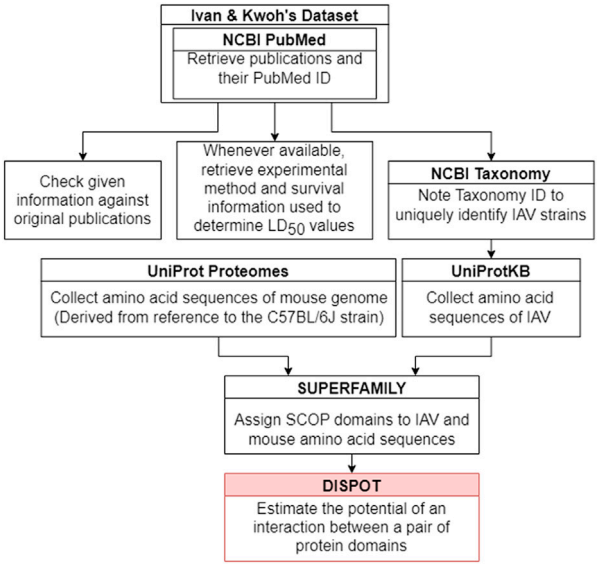


FIGURE 2
Implementation procedure.

IAV can be highly pathogenic in humans and several highly virulent strains have already caused millions of deaths worldwide in multiple pandemic events. Estimated death tolls for the 1918 (H1N1), 1957 (H2N2) and 1968 (H3N2) pandemics are 50 million (Johnson and Mueller, 2002), 1.1 million (Viboud et al., 2016) and 1-4 million (Rogers, 1968), respectively. Further, IAV triggers various respiratory illnesses seasonally, making it endemic in human populations. The yearly number of deaths due to influenza associated respiratory illness from seasonal influenza has been estimated to vary between nearly 300,000 to 646,000 (Iuliano et al., 2018). Since it is endemic in populations, hosts harboring seasonal influenza strains can act as a reservoir for reassortment events, leading to cross-infection with other circulating pathogens such as SARS-CoV-2 to form potentially harmful recombinant strains (Swets et al., 2022). These attributes highlight the complexity of disease factors of respiratory pathogens and indicate the need of wide-scale influenza studies. They also make the continual monitoring of public health outcomes necessary.

Infectious studies of mouse models can help to elucidate host factors responsible for virulence, since they are cost effective, reproducible and allow mechanistic analyses that may not be directly conducted on humans due to ethical reasons (Sarkar and Heise, 2019). One way to measure the pathogenicity of IAV is by obtaining the lethal dose at which 50% of the inoculated animal test population is infected or perishes (abbreviated here as LD_{50}) (Eugene, 2001). By comparing outcomes of influenza infections in different strains of mice, differences due to allelic variations in mice strains could be possibly be established (Lu et al., 1999).

On one hand, databases such as HPIDB (Ammari et al., 2016) and STRING Viruses (Cook et al., 2018) besides several others have already covered the interactions between influenza A and human proteins in an extensive manner. In comparison, the interactions in mouse hosts are lacking. There exist very few database records of IAV-mouse interactions (for both experimental and computational methods).

On the other hand, it is challenging to directly study the effect of influenza A virulence in human hosts owing to ethical considerations. Mice have been used to infer disease pathology of IAV in humans (Lu et al., 1999). While mice contain significant differences in body size and distribution that affect tissue tropism in pathogenesis (Masemann et al., 2020; Perlman, 2016), at present they are widely accepted pre-clinical models for linking virulence levels with IAV-host interactions (Masemann et al., 2020). Collecting IAV-mouse protein interactions provides a practical approach to identifying virulence factors. As an example, after identifying influential mouse host factors from a network of predicted interactions with IAV proteins, the corresponding set of human homologues (target proteins) can be determined from a combination of homology mapping, associated virulence levels and literature evidence. *In-vitro* interactions found to be occurring amongst target proteins and IAV (*via* biochemistry assays or cell cultures) could assist in designing knock-out factors or drug targets that will allow *in-vivo* validation of the interaction in mouse models.

Data records of mouse model infectious studies had been previously collected in an earlier work by F.X. Ivan and C.K. Kwoh (Ivan and Kwoh, 2019). Their study highlighted the role of protein sites of PB2 in influenza virulence by a systematic meta-analysis using rule-based models to predict the virulence. Therefore, a link between macroscopic virulence labels (such as LD_{50} categories) and protein-protein interactions could prove beneficial in understanding the factors contributing to IAV virulence. Domain-domain interaction(s) DDI can be particularly useful because a protein domain is often a discrete functional unit that is modular, and protein-protein interactions rely on combinations of DDI (Itzhaki et al., 2006; Alborzi et al., 2021). Hence here, the network was constructed with domains representing nodes. While 'domain-domain' interactions are by definition a subset of 'protein-protein' interactions, here the quoted terms are used interchangeably, unless specified otherwise.

TABLE 1 Summary statistics of data collected.

# journal publications	57
# IAV subtypes	14
# IAV strains	109
# mouse proteins	2419
# DDIs	1936

TABLE 2 Three-class virulence classification.

Low	Intermediate	High
$LD_{50} > 10^{6.0}$	$LD_{50} \leq 10^{6.0}$ and $LD_{50} > 10^{3.0}$	$LD_{50} \leq 10^{3.0}$

To systematically identify potential interactions between IAV and mouse host-proteins, a protein network consisting of putative DDI between IAV and mouse proteins, scored by the Domain Interaction Statistical Potential (DISPOT) (Narykov et al., 2019) was developed in this work.

The protein network is presented in a clear graphical user interface (GUI) that easily shows the LD_{50} values and interacting protein domains from the C57BL/6J mouse strain as identified by DISPOT. The virulence network of interacting protein domains will assist studies of IAV disease modeling by providing data of putative interacting protein domains that are associated with their LD_{50} values.

The rest of this manuscript is organized as follows. Section 2 details the contents and design and describes the data presented in this database and the data curation procedure. Section 3 details the web server implementation, describing the tools used, graphical user interface layout and functionality. Section 4 details discussion of this data. Section 5 outlines the proposed future work. Section 6 summarizes and concludes this paper.

2 Contents and design

Figure 2 outlines the steps taken to implement the IAV-Mouse protein-protein interaction (PPI) database. The IAV-Mouse PPI web server can be accessed at: https://iav-ppi.onrender.com/home.

Table 1 provides an overview of the data collected. Five out of the eight RNA segments of IAV genome, namely PB1, HA, NA, M1, and NS1 were found to contain the interacting pathogen protein domains. In summary, 31 unique pairs of DDIs were found between seven IAV protein domains and 29 mouse protein domains.

This work built on initial data records of mouse model infectious studies collected in the previous work by F.X. Ivan and C.K. Kwoh (Ivan and Kwoh, 2019). Their same process of assigning virulence levels was followed here. LD_{50} value was the key information needed to identify the virulence class of a specific IAV strain. Virulence was classified as two-class (avirulent/virulent) and three-class (low/intermediate/high) (*shown in* Figure 3). Essentially, the total infection records classified as "virulent" under the two-class problem is the sum of records classified as "intermediate" and "high" under the three-class problem. Likewise, the total infection records classified as "avirulent" is equivalent to the number of records classified as "low". For the three-class virulence classification, LD_{50} thresholds of $10^{3.0}$ and $10^{6.0}$ were applied (*shown in* Table 2), referencing thresholds that are used by World Health Organization (WHO), for classification of influenza virulence in mice (EID_{50} infection unit) (WHO, 2003). LD_{50} infection units include Plaque-forming Unit (PFU), Focus-forming Unit (FFU), 50% Egg Infective Dose (EID_{50}), 50% Tissue Culture Infectious Dose ($TCID_{50}$) and 50% Cell Culture Infectious Dose ($CCID_{50}$), where the equality across all units was assumed.

SUPERFAMILY 2.0 sequence search (https://supfam.org/sequence/search) (Pandurangan et al., 2019) was used to map regions of an amino acid sequence to at least one Structural Classification of Proteins (SCOP) superfamily using the SUPERFAMILY hidden Markov models. SCOP is a

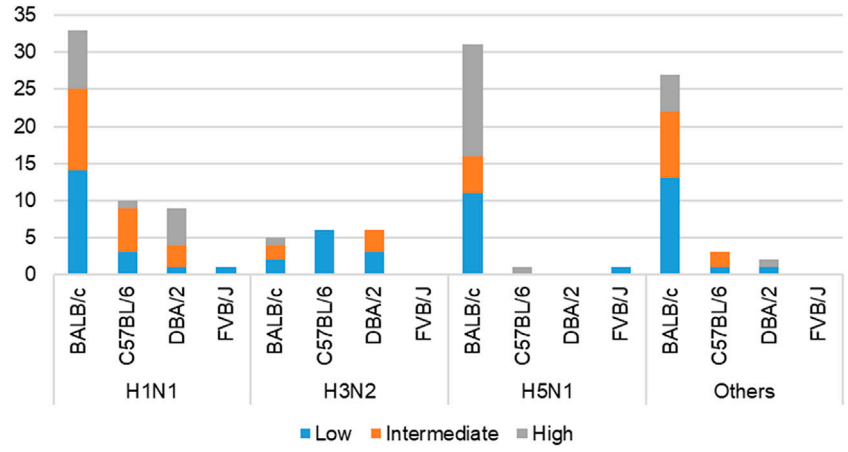


FIGURE 3
Cross-tabulation of IAV subtypes and mouse strains, colored according to three-class virulence classification problem. 'Others' refers to the aggregation of infection records from IAV subtypes—H1N2, H3N8, H5N2, H5N5, H5N6, H5N8, H7N2, H7N3, H7N7, and H7N9.

representation of structure-based hierarchical classification of relationships between protein domains, with "family" being the first level and "superfamily" being the second level. Protein domains from the same SCOP family are strongly related and frequently share the same function (Andreeva et al., 2004).

DISPOT (http://dispot.korkinlab.org/home/pairs) (Narykov et al., 2019) served as the web tool to determine presence of DDIs between pairs of IAV and mouse SCOP superfamily domains. DISPOT uses exclusively DDIs from DOMMINO (Kuang et al., 2012), an in-depth database of structurally resolved macromolecular interactions, where data about DDIs is the amplest, as its source of data. For a given domain pair, DISPOT returns a statistical potential, denoted as the probability P_{ij}. Statistical potentials take values across the entire scale of real numbers. Negative and positive values can be respectively interpreted as having more or less than average number of DDIs in the DOMMINO database. Neutral values are corresponding to the number of DDIs close to the average number. "No information" will be returned instead of a numeric value if the DOMMINO database does not have an entry for the particular domain pair.

The DISPOT calculation of statistical potential formula is given in the equation as follows (Narykov et al., 2019):

$$P_{ij} = \frac{1}{Z_2} \ln \frac{M_{pij}}{M_{mean}}$$

$$\text{where } Z_2 = \sum \sum \ln \frac{M_{pkl}}{M_{mean}}$$

Z_2 is the natural logarithm of observed frequencies of interactions between domains in the DOMMINO database. M_{mean} is the average number of interactions for a pair of domain families, calculated from the non-redundant DOMMINO dataset. Non-redundant refers to two corresponding pairs of domains that do not share 95% or more sequence identity (Narykov et al., 2019).

2.1 Dataset

All 57 journal publications reviewed in this work were retrieved from National Centre for Biotechnology (NCBI) PubMed (Lindberg, 2000), where LD_{50} values were explicitly stated in them. 55 publications referenced the supplementary information given in F.X. Ivan and C.K. Kwoh's publication ("Additional file 5: Supplementary Table S1") (Ivan and Kwoh, 2019), where LD_{50} values were stated as "values given". LD_{50} values reflected in their dataset were checked against the original publications and some missing records were added. Additionally, seven new records from two other papers (Shi et al., 2017) and (Shi et al., 2018) were documented.

2.1.1 Data cleaning

The preliminary dataset presented in this work (https://github.com/tengann/IAV-Host-PPI-Database/blob/main/RawData_2022.xlsm) holds 488 infection records involving wild-type, laboratory, mouse-adapted, recombinant or mutant IAV strains. IAV genomes of wild-type strains are in their natural and non-mutated form while laboratory strains were prepared by means of reverse genetics. Mouse-adapted strains were derived from serial lung-to-lung passages of virus in mice. Genetic amino acid sequences of mutant virus were changed through point mutations *via* single amino acid substitutions. Recombinant strains were formed by the combination of protein segments from at least two different IAV strains.

The initial dataset was manually curated to only include records involving wild-type or laboratory IAV strains, thereby reducing the number of infections to 190 (Supplementary Figure S1). Subsequently, infection records comprising wild-type or laboratory strains where their Taxonomy identification (ID) number (otherwise known as accession number) (Schoch et al., 2020) could not be found were dropped, further reducing the records to 166 (*shown in* Figure 4). In these cases, it was not possible to retrieve the complete protein sequences of IAV gene segments for SCOP domain assignment *via* SUPERFAMILY 2.0.

Lastly, multiple records concerning the same combination of IAV strain and mouse genome were condensed into a single record, adopting the approach from F.X. Ivan and C.K. Kwoh's publication (Ivan and Kwoh, 2019). From this process, the tally of infection records was reduced to 139 (*shown in* Figure 4). Whenever possible, the majority class of the three-class virulence assignment scheme was selected. Otherwise, the class that is more or most virulent was considered. Next, if only the lower bound of the LD_{50} value was presented, the record with the highest lower bound was selected. For cases where the lowest exact or upper bound of LD_{50} value was provided, the record was selected. The final cleaned dataset containing 109 unique IAV strains was used to derive the network of interacting protein domains.

2.2 Data annotation

Firstly, to distinguish between all journal publications referenced, the NCBI PubMed (https://pubmed.ncbi.nlm.nih.gov/) (Lindberg, 2000) ID number uniquely assigned to each publication record was noted. Information collected in F.X. Ivan and C.K. Kwoh's dataset consists of IAV strain, mouse strain, LD_{50} value and infection unit. Then, in this paper, to provide a deeper insight into how the LD_{50} value was determined in each separate experiment, additional evidence, namely, the experimental method, weight loss and/or survival remarks and LD_{50} calculation method were documented. Also, for each IAV strain, the Taxonomy ID number, a unique ten-digit code that designates classification and specialization was retrieved from NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/ taxonomy) (Schoch et al., 2020).

Amino acid sequences of both IAV and mouse proteins were retrieved from the UniProt (release 2021_03) protein knowledgebase (UniProtKB) (https://www.uniprot.org/) (Bairoch et al., 2005). IAV protein sequences were retrieved using strain names and/or matching Taxonomy ID number where available. Mouse protein sequences were retrieved using the Proteome ID number, UP000000589. This reference proteome was derived from the genome sequence of mouse strain C57BL/6J, with Taxonomy ID number, 10090. For this work, among 55, 315 protein records that were available, 17, 120 Swiss-Prot gold star reviewed entries (https://www.uniprot.org/uniprotkb?query=UP000000589) were retrieved. Swiss-Prot reviewed refers to records with information fully and manually extracted from literature or curator-evaluated computational analysis (Bairoch and Apweiler,

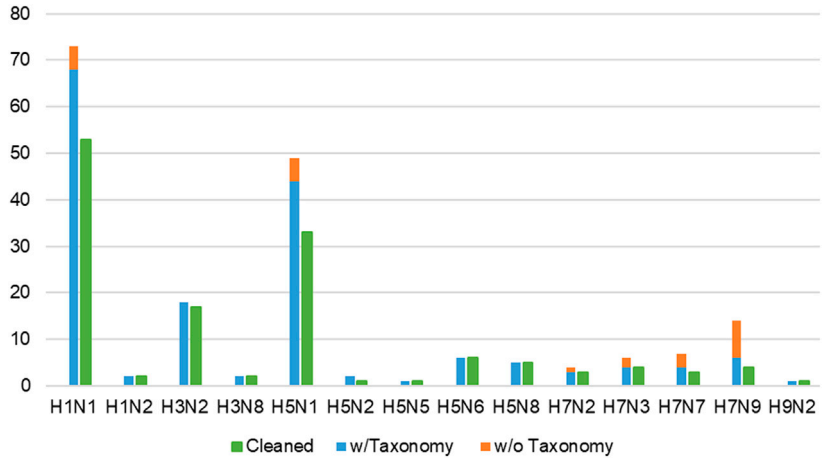


FIGURE 4
Proportion of all wild-type/laboratory IAV strains, separated into with and without taxonomy ID numbers, against remaining records (green bars) in the cleaned dataset. Infection records involving wild-type IAV where taxonomy ID number could not be found (orange bars) were omitted, reducing the number of infection records to 166 (blue bars). "Cleaned" refers to the final dataset of 139 infection records (green bars) available in IAV-Mouse PPI database, where each record corresponds to a unique combination of IAV and mouse strain.

TABLE 3 IAV Domains identified by SCOP Superfamily. Red indicates domains identified as interacting with mouse proteins, while '-' indicates no identified domains.

IAV segment	SCOP superfamily name/Accession number
PB2	PB2 C-terminal domain-like/160453
PB1	DNA/RNA polymerases/56672
PB1-F2	-
HA	Viral protein domain/49818
	Influenza hemagglutinin (stalk)/58064
NP	Flu NP-like/161003
NA	Sialidases/50939
M1	Influenza virus matrix protein M1/48145
	Alpha-catenin/vinculin-like/47220
	Methyl-accepting chemotaxis protein (MCP) signaling domain/58104
M2	-
NS1	NS1 effector domain-like/143021
	S15/NS1 RNA-binding domain/47060
NS2	Nonstructural protein NS2, NEP, M1-binding domain/101156
	Spectrin repeat/46966

1997). It strives to provide high-quality annotations with a minimal level of redundancy and high level of integration with other databases.

As an average protein consists of two or more domains, domain start and end residue numbers, corresponding to regions of protein sequences matching each assigned SCOP domain (by SUPERFAMILY 2.0) were independently noted for every protein sequence retrieved from UniProtKB. Since each domain has its distinct structure and biological function, only a subset of domains constituting each protein are involved in the interaction between a pair of proteins. Thus, this enhances the complexity of host-pathogen protein interaction analysis (Narykov et al., 2019).

Overall, IAV strains in the dataset were found to comprise proteins domains belonging to 13 SCOP superfamilies (Table 3). Then, these domains were paired up individually with 1102 unique

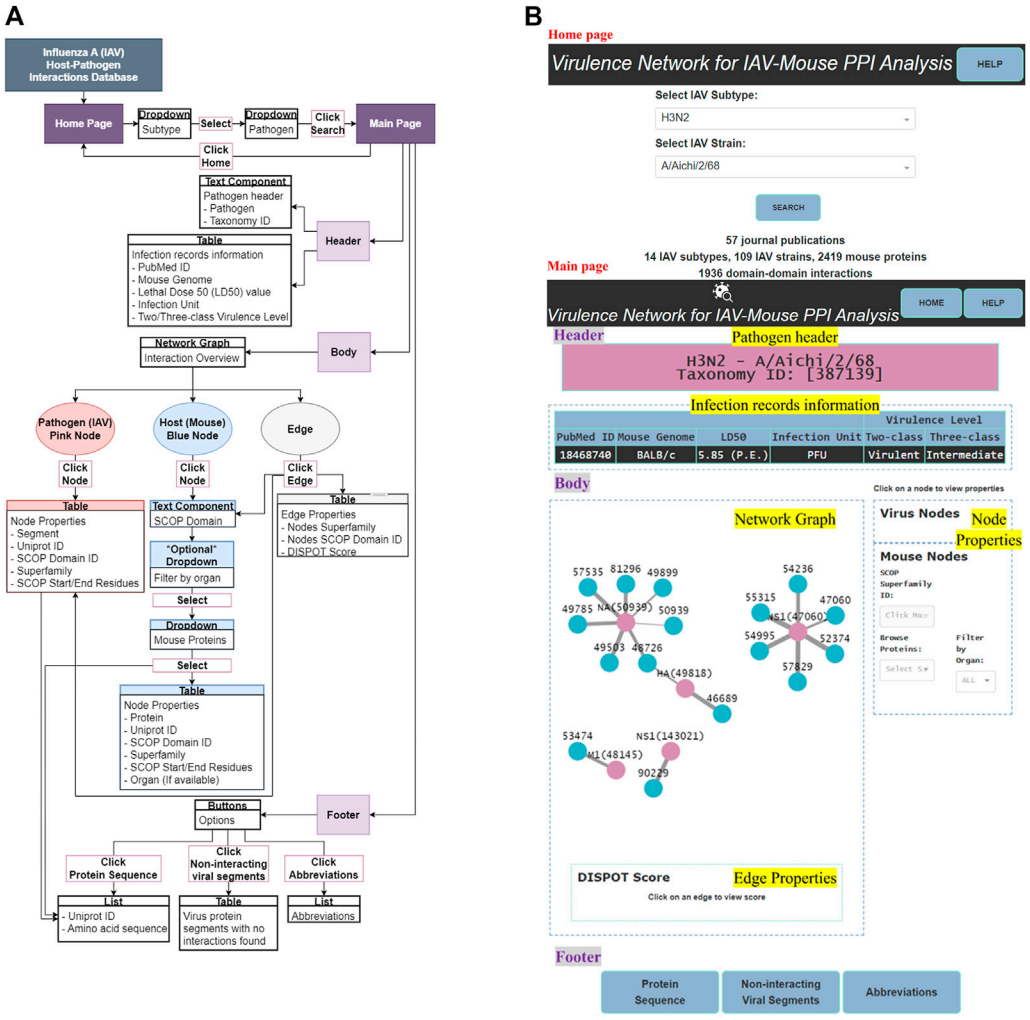


FIGURE 5
(A) GUI functionality navigation. **(B)** Screenshot of GUI layout.

domains found among the mouse protein sequences, forming a total of 14, 326 IAV-mouse protein domain pairs. Subsequently, domain pairs were fed as input into DISPOT, for calculation of statistical potential. Finally, seven domains in IAV proteins (indicated in red in Table 3) and 29 domains in mouse proteins were found to be involved in the host-pathogen PPI. Out of the 17, 120 mouse protein sequences retrieved, 2419 unique proteins were found to contain to at least one of the 29 interacting SCOP domains. In addition, the mouse protein localization in vital organs (lungs, brain, liver, kidney, spleen and heart) or blood was noted, whenever available. (available in Supplementary Figure S2).

3 Web server implementation

IAV-Mouse PPI web server GUI has a comprehensible interface, made up of two pages, with various features, including browsing *via* subtype and strain to view information collected from literature searches, an interactive network graph with accompanying information on node and edge attributes as well as amino acid sequences extracted from UniProt. Figure 5 illustrates navigation and layout of the web server's GUI.

3.1 Tools

Firstly, Microsoft Excel 2016 was used to store and organize data collected from literature. Secondly, the web interface was developed on the code editor Visual Studio Code V1.71.2, with Python V3.7.4 as the programming language. Python libraries used were Pandas V1.3.5, for transforming comma-separated values (csv) from Excel files to dataframes. Dash V2.3.1 was the framework for designing the application's functionalities, described as follows: Dash bootstrap components V0.3.0 for building the application's layout, graph visualization component Dash Cytoscape V0.3.0 for constructing the interactive network graph. Beautiful Soup V4.11.1 was the HTML parser used for pulling protein sequences from the UniProt database. Lastly, cloud hosting application, Render (https://dashboard. render.com/) web service with Python web

server gateway interface HTTP server, Gunicorn V20.1.0 was utilized to build and run the web interface entirely in the cloud.

3.2 Graphical user interface (GUI)

3.2.1 Home page

The home page features a dependent dropdown component to firstly allow the user to search for a specific IAV subtype and subsequently, browse and select the pathogen(s) belonging to the selected subtype group.

3.2.2 Main page

The main page features three main sections—header, body and footer (shown in Figure 5B).

3.2.2.1 Header

The header section consists of two components—pathogen header text component and infection records information table.

3.2.2.1.1 Pathogen header. This text component displays the IAV subtype, pathogen name and Taxonomy ID, based on selections made by the user in the home page.

3.2.2.1.2 Infection records information. This section presents infection records information collected directly from literature and additional information collected from web tools, NCBI Taxonomy and UniProt databases in the form of a table. Information is filtered to present only those relevant to the user's selections in the home page.

3.2.2.2 Body

The body section consists of two panels, where the left panel displays the network graph and edge properties. The right panel is divided into two subsections and displays the virus and mouse node properties, respectively.

3.2.2.2.1 Network graph. The network graph was designed such that the user can clearly differentiate between IAV and mouse nodes by colors, where pink was assigned to IAV nodes and blue to mouse nodes. User can differentiate interaction statistical potentials by edge weights, where a thicker edge line represents a higher possibility of interaction. Also, the edge color will change to blue upon clicking, to highlight the edge selection.

3.2.2.2.2 Node and edge properties. Node properties include either the IAV protein segment or name of mouse protein, UniProt ID, SCOP superfamily ID, superfamily name and SCOP start/end residue(s). Edge properties comprise IAV and mouse SCOP superfamily ID and name with the matching DISPOT statistical potential score. All IAV node properties will be populated upon clicking of any pink IAV node while only the mouse SCOP superfamily ID field will be populated upon clicking of any blue mouse node in the network graph. Similarly, the former, together with its respective edge property will be presented upon click on any edge. To display all mouse node properties, a mouse protein first needs to be selected from the "Browse Proteins" dropdown under the mouse node properties subsection.

3.2.2.3 Footer

The footer provides the user with the following supplementary information—protein sequence, non-interacting viral segments and abbreviations.

3.2.2.3.1 Protein sequences. Web scraping was applied to extract amino acid sequences of IAV and mouse proteins from the UniProt database.

3.2.2.3.2 Non-interacting viral segments. Non-interacting IAV protein segments consist of the following cases: 1) The UniProt ID could not be found. Therefore, the protein sequence for input to SUPERFAMILY 2.0 is unknown and no domain information could be retrieved. In this case, the UniProt ID field was indicated with "Not Found" and remaining information was labelled as "N/A". 2) Some protein sequences retrieved were not mapped to any SCOP superfamily based on the SUPERFAMILY 2.0 database. As such, there was no protein domain information for input to DISPOT. 3) IAV domain information was available but the DOMMINO database did not have any entry between the IAV domain and all of 17, 120 retrieved mouse proteins, hence no interaction information was returned by DISPOT.

3.2.2.3.2 Abbreviations. This section conveys extra information to the user; specifically, the definitions and expansions of abbreviations used as well as notes targeted to help the user better comprehend the network graph.

4 Discussion

This section lists domain pairs identified with high scoring interaction potentials. In section 4.1 the interacting domain pairs are verified against actual protein-protein interactions identified in biochemistry or proteomics literature. It is organized as a series of discussions on the functional role of interacting domains pairs, per paragraph.

4.1 Interacting protein domains

According to DISPOT scores obtained, the top 10 domain pairs with strongest statistical interaction potentials are as listed in Table 4.

Of 109 unique IAV strains presented in this database, the PB1 gene segment of three strains (Table 5) were assigned the DNA/RNA polymerases domain. However, despite the strong interactions, it was not possible to ascertain if the presence of the DNA/RNA polymerases domain has an impact on the pathogenicity, due to variation in virulence levels across the different IAV strains.

The M1 gene segment of eight IAV strains were assigned a Methyl-accepting chemotaxis protein (MCP) signaling domain instead of alpha-catenin/vinculin-like. No interaction was found between the MCP signaling domain with all domains present in mouse proteins. Comparing results from experiments carried out in H3N2 IAV strains (Tables 6, 7), especially on mice strains BALB/c and DBA/2, it is evident that presence of the alpha-catenin/vinculin-like domain is a virulence factor responsible for IAV infection.

TABLE 4 Top 10 Interactions according to statistical potentials returned by DISPOT. A more negative DISPOT score indicates a higher possibility of interaction.

IAV segment	IAV SCOP domain	Mouse SCOP domain	DISPOT score
NS1	S15/NS1 RNA-binding	Nucleotidylyl transferase	-4.301855801
		L30e-like	
PB1	DNA/RNA polymerases	DNA clamp	
		5′ to 3′ exonuclease, C-terminal subdomain	
		PIN domain-like	
		N-acetylmuramoyl-L-alanine amidase-like	
NS1	S15/NS1 RNA-binding	Zn-binding ribosomal proteins	−3.608708621
		Ribosomal protein S6	
M1	Alpha-catenin/vinculin-like	PH domain-like	
		I/LWEQ domain	

TABLE 5 IAV strains with DNA/RNA polymerases domain assigned to PB1 gene segment.

IAV strain	Subtype	Three-class virulence level
A/PuertoRico/8/1934	H1N1	High (BALB/c, DBA/2)
		Intermediate (C57BL/6J)
rA/X-31	H3N2	Low (BALB/c)
A/HongKong/97/98	H5N1	Low (BALB/c)

Alpha-catenins are members of the vinculin family of proteins. Vinculin is an actin-binding protein. Protease treatment revealed that actin present in the interior of influenza virions presumably participates in moving viral components to the assembly site and cytoskeletal reorganization that occurs during bud formation (Peng et al., 2012). Actin is a family of globular multi-functional proteins that form microfilaments in the cytoskeleton. The host cytoskeletal network takes part in transport of viral components in the cell, predominantly during the stages of virus entry and exit (Shaw et al., 2008).

Pleckstrin homology (PH) domain-like is a short peptide module often found in cytoskeletal proteins (Yao et al., 1999). Although cytoskeletal elements are known to be associated with M1, the underlying mechanisms are not clear (Zhao et al., 2017). Ezrin, [*EZRI_MOUSE (UniProt accession number: P26040)*] has a PH-like domain. Based on meta-analysis of IAV interactome studies on the M1 gene segment conducted by (Chua et al., 2022), Ezrin was discovered to be a common interactor and is a positive regulator of virus replication.

Information from UniProtKB indicates that majority of proteins that contain a nucleotidylyl transferase domain possess the tRNA ligase enzyme, otherwise known as aminoacyl-tRNA synthetase (ARSs). ARSs play a crucial role in protein synthesis by attaching amino acids to their cognate transfer RNAs (tRNAs) (Nie et al., 2019). Specifically, Cysteine-tRNA ligase [*SYCC_MOUSE (UniProt accession number: Q9ER72)*], an interactor of NS1, catalyzes the ATP-dependent ligation of cysteine to tRNA (Cys) and plays a role in translation (de Chassey et al., 2013). Furthermore, ARSs plays a vital role in the development of immune cells because of their involvement in maturation, transcription, activation, and recruitment of immune cells. More significantly, ARSs regulate various biological processes and act as signaling molecules in infectious disease (Nie et al., 2019), which supports the high DISPOT score (≈−4.302) predicted for the S15/NS1 RNA-binding domain in NS1 segments.

Ubiquitin-40S ribosomal protein S27a, [*RS27A_MOUSE (UniProt accession number: P62983)*], is a protein with the Zn-binding ribosomal protein domain. It is an NS1-interacting host protein node, classified as belonging to the apoptosis pathway (Thulasi Raman and Zhou, 2016). Although not required for ribosome function, it plays an important role in the life cycle of IAV through regulating viral nucleic acid replication and gene transcription. When interrupted in host cells, the replication and infectivity of IAV is stopped (Li, 2019).

CCCH zinc finger present in mouse proteins is the sole domain that interacts with NS1 effector domain-like instead of the S15/NS1 RNA-binding domain in the NS1 segment of IAV. This domain pair has a DISPOT statistical potential score of -2.916 (*rounded to 3 d.p.*). Mouse protein, cleavage and polyadenylation specificity factor subunit 4 (CPSF4) [*CPSF4_MOUSE (UniProt accession number: Q8BQZ5)*] contains the CCCH zinc finger domain. The interaction between NS1 and CPSF4 controls the alternative splicing of tumor protein p53 (TP53) transcripts, and alters the expression of TP53 isoforms in parallel. As a result, cellular innate response, particularly *via* type I interferon secretion is regulated, leading to efficient viral replication (Dubois et al., 2019).

The Immunoglobulin (Ig) domain, otherwise known as antibodies is the sole SCOP protein domain that interacts with three IAV domains, namely DNA/RNA polymerases, Viral protein domain and Sialidases in the PB1, HA, and NA segments respectively. Immunoglobulin is the most abundant domain found among the 2419 unique mouse proteins containing interacting SCOP domains (Supplementary Figure S2). During natural infection with IAVs, immune response against both HA and NA will be evoked (Creytens et al., 2021). IgM response is dominant in primary infection, while IgG response is dominant in

TABLE 6 IAV strains with a Methyl-accepting chemotaxis protein (MCP) signaling domain assigned to M1 gene segment instead of alpha-catenin/vinculin-like.

IAV strain	Subtype	Three-class virulence level
A/Aichi/2/68	H3N2	Intermediate (BALB/c)
A/Brisbane/10/2007		Low (C75BL/6, DBA/2)
A/Memphis/8/1988		Low (BALB/c)
A/Panama/2007/1999		Low (C57BL/6, DBA/2)
A/Wisconsin/67/2005		Low (C57BL/6, DBA/2)
rA/X-31		Low (BALB/c)
A/duck/Guangxi/53/2002	H5N1	Low (BALB/c)
A/chicken/Shandong/lx1023/2007	H9N2	Low (BALB/c)

TABLE 7 H3N2 strains with alpha-catenin/vinculin-like domain assigned to M1 gene segment.

IAV strain	Three-class virulence level
A/Hong Kong/1/1968	Low (C57BL/6, DBA/2)
A/Philippines/2/1982	High (BALB/c)
A/swine/Spain/54008/2004	Low (C57BL/6)
	Intermediate (DBA/2)
A/swine/Texas/4199-2/1998	Low (C57BL/6)
	Intermediate (DBA/2)
A/Victoria/3/1975	Intermediate (BALB/c)

secondary infection, for Ig secretion. IgA present in nasal secretions can neutralize HA and NA of IAVs (Chen et al., 2018).

4.2 Non-interacting protein domains

The non-interacting influenza hemagglutinin (stalk) SCOP domain present in the HA is an example that not all domains constituting a protein are involved in interaction between a pair of proteins. The stalk evolves slower than the receptor binding head and it is suggested that it has to remain structurally conserved owing to its role in membrane fusion (Kirkpatrick et al., 2018). Studies have also suggested that the stalk domain is not under immune pressure (Wu and Wilson, 2020; Petrova and Russell, 2018). Additionally, mutations in the stalk domain do not drastically impact virus binding or aid in avoiding neutralizing antibody responses from the host (Kirkpatrick et al., 2018). Therefore a potential "true negative" interaction is also identified in the protein domain network, in line with experimental findings.

4.3 Limitations

Generally, DISPOT works as a tool to streamline the PPI prediction problem through providing insight on the possibility of specific DDIs in a given physical PPI. However, it is not a classification method and statistical potentials returned are useful for ranking DDIs but do not directly translate to the probability score. DISPOT which solely uses information about interactions between protein domain should not be used as a standalone PPI prediction tool to identify virulence factors responsible for IAV infections (Narykov et al., 2019). Based on results of this work, IAV genomes across different strains comprise highly similar domains due to their similar structure (i.e., eight segments, encoding at least 11 proteins) and biological function. Furthermore, interactions that involve protein structures are facilitated not only by the protein domains, but also by various non-structured regions, such as interdomain linkers, N and C terminal structures or sequences, protein peptides (Kuang et al., 2012). Therefore, utilizing DISPOT exclusively may produce high number of false negative or false positive PPI predictions.

Mitochondria play an imperative role in antiviral innate immune response through the mitochondiral antiviral-signaling protein (MAVS) [*MAVS_MOUSE (UniProt accession number: Q8VCF0)*] protein, a component of the retinoic acid-inducible gene I (RIG-I) antiviral pathway. This pathway along with multiple others, is essential for combating and resolving viral infection, repair of damaged tissues, and generating adaptive immune response. It has been revealed that PB1-F2 inhibits antiviral cytokines and enhances expression of inflammatory cytokines through direct interaction with MAVS and other components of the RIG-I/MAVS system (Kamal et al., 2017). However, as protein sequences of both PB1-F2 and MAVS were not assigned any SCOP domain by SUPERFAMILY 2.0, it was not possible to verify this interaction *via* DISPOT.

A homeodomain-like domain was identified by DISPOT to be interacting with the viral protein domain present in the HA gene segment of IAV, with a statistical potential score of −3.203 (*rounded to 3 d.p.)*. However a study conducted by (Farooq et al., 2020), which integrates both IAV-Mouse PPIs detected using either small-scale or large-scale researches carried out experimentally or computationally found no evidence for an interaction with HA. In (Farooq et al., 2020), homeobox protein MOX-2 (MEOX2) [*MEOX2_MOUSE, (UniProt accession number: P32443)*], containing the homeodomain-like

domain, was identified to be interacting with IAV gene segments PB1, PA, NA, and M2 but not HA. By comparison, for DISPOT no interaction was detected between the homeodomain-like domain and segments PB1 and NA. Likewise, as protein sequences of segments PA and M2 were not assigned any SCOP domain by SUPERFAMILY 2.0, DISPOT could not be used to ascertain these interactions. Further, the cellular localisation for proteins with homeodomain-like domains was found to be in the nucleus according to UniProt, which indicates its interaction with HA would be unlikely, given that HA, mediates cell-surface recognition and viral entry.

Additionally, the reason for virulence levels to differ across mouse strains infected with the same IAV strain has not been uncovered as protein sequences of mouse strains retrieved from UniProt were derived from referencing the C57BL/6J mouse strain only. This limitation is because currently whole proteome sequences of other mouse strains (i.e., BALB/c, DBA/2 and FVB/J) are not available in any public database. Translation of strain specific genomic sequences to whole proteomes is a challenging task needing extensive experimental effort. In this work, obtaining these proteomes by means of experimental protein sequencing was not possible as the necessary materials and labor were not available.

5 Future work

To bridge the gaps in this work, sequence-based PPI prediction methods can be employed to substantiate DDIs identified by DISPOT. An example is the Human-Virus Protein-Protein Interactions (HVPPI) web server, developed by X.Yang and colleagues (Yang et al., 2020a). HVPPI applied an unsupervised sequence embedding technique (*doc2vec*) to represent protein sequences as low-dimensional rich feature vectors. Then, a random forest classifier was trained using a training dataset that covers known PPIs between human and all viruses to predict human-virus PPIs. Lastly, the HVPPI web server automatically calculates the interaction probability of a query protein pair. The data to be used as input to HVPPI can be constructed as follows: Firstly, human protein sequences can be obtained using mouse and human homologs. Next, all protein sequences with SCOP domain(s) assigned to them can be trimmed, following the collected start and end residue numbers. Subsequently, trimmed protein sequences can be paired corresponding to DDIs recognized by DISPOT. Interaction probabilities provided as predicted outputs by HVPPI for each IAV-human protein pair can then be used to detect false positives. For protein sequences not assigned to any SCOP domain, complete protein sequences can be used, which will in turn aid with the detection of false negatives.

As an extension of this work from the raw dataset, instead of ignoring non-standard strains, the protein sequences of recombinant or mutant IAV strains can be reproduced *via* manually changing the protein sequences of wild-type or laboratory IAV strains that are available in UniProt. This enriches the dataset further.

The DISPOT statistical potentials, HVPPI interaction probabilities and LD_{50} values can be incorporated to represent the PPI network as a weighted undirected graph. Later, graph embedding methods can be applied to this weighted graph to learn low-dimensional node representations (Yue et al., 2020). Structural information of PPI, such as the degree, position and neighbouring nodes in a graph has been recognized to be helpful in PPI prediction (Yang et al., 2020b).

6 Conclusion

As IAV is a significant danger to global human health and life, it is critical to have deeper, accurate as well as reliable insights and knowledge on the virulence factors responsible for IAV infections to counteract potential outbreaks (Ivan and Kwoh, 2019). This work built upon a previously curated dataset of lethal dose studies of IAV infection in mice. Thereafter, superfamily domains involved in DDIs between IAV and mice were discovered, and ranked according to statistical interaction potentials calculated by DISPOT. A one-stop web server integrating information collated from literature and various databases, namely, NCBI Taxonomy, UniProt and SUPERFAMILY 2.0 with the DDI network was constructed. Furthermore, the web server is scalable and can seamlessly accommodate addition of new functions and data when future research is carried out.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://iav-ppi.onrender.com/home https://github.com/tengann/IAV-Host-PPI-Database.

Author contributions

TN and SR conceived and designed the method. TN implemented the method. TN and SR wrote the manuscript. CK supervised throughout the conception, design, implementation and writing stages. All authors reviewed and approved the final manuscript.

Funding

This work was funded by Ministry of Education (MOE) grants MOE2019-T2-2-175 and MOE2020-T1-001-130, Singapore.

Acknowledgments

The authors thank F.X. Ivan for constructing the initial dataset of infection records that was used to develop the virulence network here. They also thank the following consortia of databases and programs: NCBI PubMed and Taxonomy, UniProt, SUPERFAMILY 2.0 and DISPOT.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1123993/full#supplementary-material

References

Alborzi, S. Z., Nacer, A. A., Najjar, H., Ritchie, D. W., and Ppidomainminer, M. D. D. (2021). Inferring domain-domain interactions from multiple sources of protein-protein interactions. *PLOS Comput. Biol.* 17 (8), e1008844. doi:10.1371/journal.pcbi.1008844

Ammari, M. G., Gresham, C. R., McCarthy, F. M., and Nanduri, B. (2016). Hpidb 2.0: A curated database for host–pathogen interactions. *Database* 2016, baw103. doi:10.1093/database/baw103

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C., and Murzin, A. G. (2004). Scop database in 2004: Refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, 226D–229D. doi:10.1093/nar/gkh039

Bairoch, A., and Apweiler, R. (1997). The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res.* 25 (1), 31–36. doi:10.1093/nar/25.1.31

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The universal protein resource (uniprot). *Nucleic Acids Res.* 33, D154–D159. doi:10.1093/nar/gki070

Chen, X., Liu, S., Goraya, M. U., Maarouf, M., Huang, S., and Chen, J. L. (2018). Host immune response to influenza a virus infection. *Front. Immunol.* 9, 320. doi:10.3389/fimmu.2018.00320

Chua, S. C. J. H., Cui, J., Engelberg, D., and Lim, L. H. K. (2022). A review and meta-analysis of influenza interactome studies. *Front. Microbiol.* 13, 869406. doi:10.3389/fmicb.2022.869406

Cook, H. V., Doncheva, N. T., Szklarczyk, D., von Mering, C., and Jensen, L. J. (2018). Viruses.STRING: A virus-host protein-protein interaction database. *Viruses* 10 (10), 519. doi:10.3390/v10100519

Creytens, S., Pascha, M. N., Ballegeer, M., Saelens, X., and de Haan, C. A. M. (2021). Influenza neuraminidase characteristics and potential as a vaccine target. *Front. Immunol.* 12, 786617. doi:10.3389/fimmu.2021.786617

de Chassey, B., Aublin-Gex, A., Ruggieri, A., Meyniel-Schicklin, L., Pradezynski, F., Davoust, N., et al. (2013). The interactomes of influenza virus ns1 and ns2 proteins identify new host factors and provide insights for adar1 playing a supportive role in virus replication. *PLoS Pathog.* 9 (7), e1003440. doi:10.1371/journal.ppat.1003440

Dubois, J., Traversier, A., Julien, T., Padey, B., Lina, B., Bourdon, J. C., et al. (2019). The nonstructural ns1 protein of influenza viruses modulates tp53 splicing through host factor cpsf4. *J. Virol.* 93 (7), 021688–e2218. doi:10.1128/jvi.02168-18

Eugene, W. N. (2001). *Microbiology: A human perspective*. 3rd edition. Dubuque, Iowa: McGraw-Hill.

Farooq, Q. U. A., Shaukat, Z., Aiman, S., Zhou, T., and Li, C. (2020). A systems biology-driven approach to construct a comprehensive protein interaction network of influenza a virus with its host. *BMC Infect. Dis.* 20 (1), 480. doi:10.1186/s12879-020-05214-0

Itzhaki, Z., Akiva, E., Altuvia, Y., and Margalit, H. (2006). Evolutionary conservation of domain-domain interactions. *Genome Biol.* 7 (12), R125. doi:10.1186/gb-2006-7-12-r125

Iuliano, A. D., Roguski, K. M., Chang, H. H., Muscatello, D. J., Palekar, R., Tempia, S., et al. (2018). Estimates of global seasonal influenza-associated respiratory mortality: A modelling study. *Lancet* 391, 101271285–101271300. doi:10.1016/s0140-6736(17)33293-2

Ivan, F. X., and Kwoh, C. K. (2019). Rule-based meta-analysis reveals the major role of pb2 in influencing influenza a virus virulence in mice. *BMC Genomics* 20 (9), 973. doi:10.1186/s12864-019-6295-8

Johnson, N. P. A. S., and Mueller, J. (2002). Updating the accounts: Global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bull. Hist. Med.* 76 (1), 105–115. doi:10.1353/bhm.2002.0022

Jung, H. E., and Lee, H. K. (2020). Host protective immune responses against influenza a virus infection. *Viruses* 12 (5), 504. doi:10.3390/v12050504

Kamal, R. P., Alymova, I. V., and York, I. A. (2017). Evolution and virulence of influenza a virus protein pb1-f2. *Int. J. Mol. Sci.* 19 (1), 96. doi:10.3390/ijms19010096

Kirkpatrick, E., Qiu, X., Wilson, P. C., Bahl, J., and Krammer, F. (2018). The influenza virus hemagglutinin head evolves faster than the stalk domain. *Sci. Rep.* 8 (1), 10432. doi:10.1038/s41598-018-28706-1

Kuang, X., Han, J. G., Zhao, N., Pang, B., Shyu, C. R., and Korkin, D. (2012). Dommino: A database of macromolecular interactions. *Nucleic Acids Res.* 40, D501–D506. doi:10.1093/nar/gkr1128

Li, S. (2019). Regulation of ribosomal proteins on viral infection. *Cells* 8 (5), 508. doi:10.3390/cells8050508

Lindberg, D. A. (2000). Internet access to the national library of medicine. *Eff. Clin. Pract.* 3 (5), 256–260.

Lu, X., Tumpey, T. M., Morken, T., Zaki, S. R., Cox, N. J., and Katz, J. M. (1999). A mouse model for the evaluation of pathogenesis and immunity to influenza a (h5n1) viruses isolated from humans. *J. Virol.* 73 (7), 5903–5911. doi:10.1128/jvi.73.7.5903-5911.1999

Masemann, D., Ludwig, S., and Boergeling, Y. (2020). Advances in transgenic mouse models to study infections by human pathogenic viruses. *Int. J. Mol. Sci.* 21 (23), 9289. doi:10.3390/ijms21239289

Narykov, O., Bogatov, D., and Korkin, D. (2019). Dispot: A simple knowledge-based protein domain interaction statistical potential. *Bioinformatics* 35 (24), 5374–5378. doi:10.1093/bioinformatics/btz587

Nie, A., Sun, B., Fu, Z., and Yu, D. (2019). Roles of aminoacyl-trna synthetases in immune regulation and immune diseases. *Cell. Death Dis.* 10 (12), 901. doi:10.1038/s41419-019-2145-5

Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B., and Gough, J. (2019). The superfamily 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Res.* 47 (D1), D490–D494. doi:10.1093/nar/gky1130

Peng, X., Maiers, J. L., Choudhury, D., Craig, S. W., and DeMali, K. A. (2012). α -catenin uses a novel mechanism to activate vinculin. *J. Biol. Chem.* 287 (10), 7728–7737. doi:10.1074/jbc.m111.297481

Perlman, R. L. (2016). Mouse models of human disease: An evolutionary perspective. *Evol. Med. Public Health* 2016 (1), 170–176. doi:10.1093/emph/eow014

Petrova, V. N., and Russell, C. A. (2018). The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* 16 (1), 47–60. doi:10.1038/nrmicro.2017.118

Rogers, K. (1968). "Flu pandemic," in *Encyclopedia britannica*. 15th edition (Chicago, Illinois, United States: Encyclopaedia Britannica, Inc.). Avaliable At: https://www.britannica.com/event/1968-flu-pandemic.

Sarkar, S., and Heise, M. T. (2019). Mouse models as resources for studying infectious diseases. *Clin. Ther.* 41 (10), 1912–1922. doi:10.1016/j.clinthera.2019.08.010

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., et al. (2020). Ncbi taxonomy: A comprehensive update on curation, resources and tools. *Database (Oxford)* 2020, baaa062. doi:10.1093/database/baaa062

Shaw, M. L., Stone, K. L., Colangelo, C. M., Gulcicek, E. E., and Palese, P. (2008). Cellular proteins in influenza virus particles. *PLOS Pathog.* 4 (6), e1000085. doi:10.1371/journal.ppat.1000085

Shi, J., Deng, G., Kong, H., Gu, C., Ma, S., Yin, X., et al. (2017). H7n9 virulent mutants detected in chickens in China pose an increased threat to humans. *Cell. Res.* 27 (12), 1409–1421. doi:10.1038/cr.2017.129

Shi, J., Deng, G., Ma, S., Zeng, X., Yin, X., Li, M., et al. (2018). Rapid evolution of h7n9 highly pathogenic viruses that emerged in China in 2017. *Cell. Host Microbe* 24 (4), 558–568. doi:10.1016/j.chom.2018.08.006

Swets, M. C., Russell, C. D., Harrison, E. M., Docherty, A. B., Lone, N., Girvan, M., et al. (2022). Sars-cov-2 co-infection with influenza viruses, respiratory syncytial virus, or adenoviruses. *Lancet* 399, 103341463–103341464. doi:10.1016/s0140-6736(22)00383-x

Thulasi Raman, S. N., and Zhou, Y. (2016). Networks of host factors that interact with ns1 protein of influenza a virus. *Front. Microbiol.* 7, 654. doi:10.3389/fmicb.2016.00654

Viboud, C., Simonsen, L., Fuentes, R., Flores, J., Miller, M. A., and Chowell, G. (2016). Global mortality impact of the 1957-1959 influenza pandemic. *J. Infect. Dis.* 213 (5), 738–745. doi:10.1093/infdis/jiv534

WHO (2003). Production of pilot lots of inactivated influenza vaccine in response to a pandemic threat: An interim biosafety risk assessment. *Wkly. Epidemiol. Rec.* 78 (47), 405–408.

Wu, N. C., and Wilson, I. A. (2020). Structural biology of influenza hemagglutinin: An amaranthine adventure. *Viruses* 12 (9), 1053. doi:10.3390/v12091053

Yang, F., Fan, K., Song, D., and Lin, H. (2020a). Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinforma.* 21 (1), 323. doi:10.1186/s12859-020-03646-8

Yang, X., Yang, S., Li, Q., Wuchty, S., and Zhang, Z. (2020b). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput. Struct. Biotechnol. J.* 18, 153–161. doi:10.1016/j.csbj.2019.12.005

Yao, L., Janmey, P., Frigeri, L. G., Han, W., Fujita, J., Kawakami, Y., et al. (1999). Pleckstrin homology domains interact with filamentous actin. *J. Biol. Chem.* 274 (28), 19752–19761. doi:10.1074/jbc.274.28.19752

Yue, X., Wang, Z., Huang, J., Parthasarathy, S., Moosavinasab, S., Huang, Y., et al. (2020). Graph embedding on biomedical networks: Methods, applications and evaluations. *Bioinformatics* 36 (4), 1241–1251. doi:10.1093/bioinformatics/btz718

Zhao, M., Wang, L., and Li, S. (2017). Influenza a virus-host protein interactions control viral pathogenesis. *Int. J. Mol. Sci.* 18 (8), 1673. doi:10.3390/ijms18081673

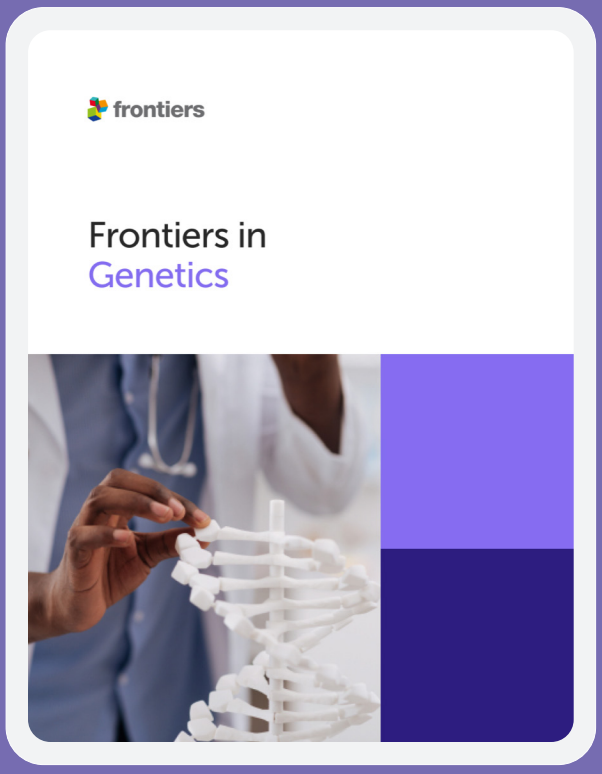
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact