frontiers | Research Topics

Machine learning-based methods for RNA data analysis, volume II

Edited by
Lihong Peng, Jialiang Yang, Liqian Zhou and Minxian Wallace Wang

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-83251-034-6
DOI 10.3389/978-2-83251-034-6

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

Machine learning-based methods for RNA data analysis, volume II

Topic editors

Lihong Peng — Hunan University of Technology, China
Jialiang Yang — Geneis (Beijing) Co. Ltd, China
Liqian Zhou — Hunan University of Technology, China
Minxian Wallace Wang — Beijing Institute of Genomics, Chinese Academy of Sciences (CAS), China

Citation

Peng, L., Yang, J., Zhou, L., Wang, M. W., eds. (2022). *Machine learning-based methods for RNA data analysis, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83251-034-6

Author JY was employed by Geneis (Beijing) Co Ltd.

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

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 29 November 2022
DOI 10.3389/fgene.2022.1010089

OPEN ACCESS

EDITED AND REVIEWED BY
William C. Cho,
QEH, Hong Kong SAR, China

*CORRESPONDENCE
Lihong Peng,
plhhnu@163.com
Liqian Zhou,
zhoulq11@163.com

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

RECEIVED 02 August 2022
ACCEPTED 20 September 2022
PUBLISHED 29 November 2022

CITATION
Peng L, Yang J, Wang M and Zhou L (2022), Editorial: Machine learning-based methods for RNA data analysis—Volume II.
Front. Genet. 13:1010089.
doi: 10.3389/fgene.2022.1010089

Editorial: Machine learning-based methods for RNA data analysis—Volume II

Lihong Peng[1,2]*, Jialiang Yang[3], Minxian Wang[4,5] and Liqian Zhou[1]*

[1]College of Life Sciences and Chemistry, Hunan University of Technology, Zhuzhou, China, [2]School of Computer, Hunan University of Technology, Zhuzhou, China, [3]Geneis (Beijing) Co Ltd., Beijing, China, [4]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China, [5]University of Chinese Academy of Sciences, Beijing, China

KEYWORDS

machine learning, lncRNA, microRNA, circRNA, mRNA, gene expression

Editorial on the Research Topic
Machine learning-based methods for RNA data analysis—Volume II

RNAs regulate multiple biological processes including RNA transcription, splicing, stability, and translation. They play significant roles in cell biology (Connelly et al. (2016); Licatalosi and Darnell (2010); Mukherjee et al. (2022); Chen et al. (2018b)). The Encyclopedia of DNA elements project reported that only 1.5% of human genome is translated into proteins, while approximately 70%–90% is transcribed to RNAs (Falese et al. (2021)). RNAs greatly expand the range of targets from proteins to RNAs by re-targeting mutated targets (Yu et al. (2019); Chen et al. (2020); Li et al. (2022); Yang et al. (2022)). Particularly, noncoding RNAs have dense linkages with human diseases including cancers. Now, RNAs have been diagnostic or prognostic markers of complex diseases (Hui et al. (2011); Xu et al. (2022); Peng et al. (2022a); Shen et al. (2022); Zhang T. et al. (2022); Chai et al. (2022)). In this topic, we aim to analyze diverse RNA data to provide clues for the diagnosis and therapy of various diseases (Dal Molin et al. (2022); Wang S. et al. (2022); Li J. et al. (2019); Liu et al. (2020)). Long noncoding RNAs (lncRNAs) regulate many significant biological processes (such as immune response and embryonic stem cell pluripotency) by linking to RNA-binding proteins (Wapinski and Chang (2011); Chen and Huang (2017); Ping et al. (2018); Wang et al. (2020)), Wang et al. (2021 W.); Peng et al. (2020)). They have been important biomarkers for cancers (Wu et al. (2022a); Banerjee et al. (2020); Zhang S. et al. (2021); Zhou G. et al. (2021); Peng et al. (2022a); Liang et al. (2022b); Peng et al. (2021); Zhou L. et al. (2021)). For example, lncRNAs AFAP1-AS1, CCAT1, CYTOR, GAS5, HOTAIR, and PVT1 are molecular regulators of lung caner (Aftabi et al. (2021)). KCNQ1OT1 may be a prognostic biomarker in colorectal cancer (Lin et al. (2021)). lncRNAs are also oncogenes (such as MKLN1-AS, GHET1, LASP1-AS, MALAT1, HULC, HOTAIR, and PAPAS) and tumor suppressors (such as CASC2, DGCR5, MEG3, GAS5, and NRON) in hepatocellular carcinoma (Guo et al. (2021)). Many machine learning methods have been proposed to

infer new LncRNA-Disease Associations (LDAs). For example, graph convolutional completion with conditional random (Fan et al. (2022)), heterogeneous graph attention network with meta-paths (Zhao et al. (2022)), graph convolutional auto-encoders (Silva and Spinosa (2021)), multi-view attention graph convolutional network and stacking ensemble (Liang et al. (2022b)), and learning to rank-based model (Wu et al. (2022a)) are widely used methods for LDA prediction.

In this research topic, Sun et al. developed a weighted graph-regularized matrix factorization approach (LPI-WGRMF) to identify possible lncRNA-protein interactions (LPIs) based on known biological information and LPI matrix. LPI-WGRMF obtained an AUC of 0.9012 and AUPR of 0.7324 on LPI dataset provided by Zhang et al. (Zhang et al. (2018)) based on 5-fold cross validation. They predicted that lncRNAs SNHG3, SFPQ, and PRPF31 may interact with proteins Q9NUL5, Q9NUL5, and Q9UKV8, respectively. Yao et al. designed a random walk with restart algorithm (MHRWRLDA) to infer LDAs on multiplex and heterogeneous networks. MHRWRLDA computed an AUC of 0.6874 under leave-one-out cross validation, and inferred that lncRNA BCYRN1 may associate with colon cancer and hepatocellular carcinoma. Cheng et al. considered that the recurrence rate of nonfunctioning pituitary adenoma is relatively high after surgical resection and built lncRNA signatures for its prognosis. They obtained microarray sequencing profiles of lncRNA expressions from 66 patients who suffered from nonfunctioning pituitary adenoma. Univariable Cox regression analysis and random survival forests-variable hunting were applied to filter lncRNAs. They found that three lncRNAs, LOC101927765, RP11-23N2.4, and RP4-533D7.4, have dense associations with tumor recurrence and inferred that the three lncRNAs may be potential therapeutic targets of nonfunctioning pituitary adenoma.

MicroRNAs (miRNAs) are a class of endogenous noncoding RNAs with a length of approximately 22 nucleotides (Sun et al. (2022); Chen et al. (2019b, 2018b); Zhang L. et al. (2021)). MiRNAs regulate many biological activities and influence almost all genetic pathways (Chen et al. (2018c); Peng et al. (2017); Chen et al. (2018a)). Thus, miRNAs have been a class of tumor suppressor genes in clinical medicine (Chen et al. (2019a); Peng et al. (2018)). For example, miR-940 is a potential biomarker of prostate cancer (Rajendiran et al. (2021)). Urinary exosome microRNA signatures are noninvasive prognostic markers for prostate cancer (Shin et al. (2021)). Recently, machine learning methods have been widely used to identify possible MicroRNA-Disease Associations (MDAs). For example, tensor decomposition with relational constraints (Huang et al. (2021)), similarity constrained matrix factorization (Li L. et al. (2021)), tensor factorization and label propagation (Yu et al. (2022)), deep attributed network embedding model (Ji et al. (2021)), and multi-view multichannel attention graph convolutional network (Tang et al. (2021)) are popular methods in MDA prediction.

In this topic, Qu et al. explored a computational model (BRWRMHMDA) for MDA inference combining enforcing degree-based biased random walk with restart. BRWRMHMDA computed an AUC of 0.8310 under leave-one-out cross validation. They predicted that hsa-let-7f and hsa-mir-30e may associate with esophageal neoplasms and breast neoplasms, respectively. Zhou et al. proposed a pseudogene-miRNA association identification method (PMGAE) by integrating feature fusion, graph autoencoder, and eXtreme gradient boosting. First, they computed three types of similarities for pseudogenes and miRNAs, that is, Pearson similarity, cosine similarity, and Jaccard similarity. Second, the above similarities were fused to build a similarity profile for each node. Third, the similarity profiles and pseudogene-miRNA associations are further aggregated to depict each node as a low-dimensional vector through a graph autoencoder. Finally, the feature vector was fed into eXtreme gradient boosting for pseudogene-miRNA association prediction. PMGAE computed better AUC of 0.8634 and AUPR of 0.8966. The results from PMGAE showed that miRNAs hsa-miR-34c-5p, hsa-miR-199b-5p, and hsa-miR-103a-3p may associate with pseudogenes RPLP0P2, HLA-H, and HLA-J, respectively.

Circle RNAs (circRNAs) is a class of novel endogenous noncoding RNAs with a covalently closed loop structure (Wang C.-C. et al. (2021); Li G. et al. (2019); Wang et al. (2021b)). circRNAs have more stable expressions due to their resistances to RNA exonuclease degradation (Li et al. (2020); Wang et al. (2021c,b)). They can regulate protein binding, miRNA sponges, alternative splicing and transcription, and generate pseudogenes (Wang C.-C. et al. (2021); Chen (2020)). In addition, they demonstrate close associations with cancers, cardiovascular and nervous system diseases (Wang C.-C. et al. (2021); Li G. et al. (2019, 2020); Wang et al. (2021c,c,b)). Therefore, various computational models have been developed to detect possible CircRNA-Disease Associations (CDAs). For example, network embedding and subspace learning method (Xiao et al. (2021)), knowledge attention network (Lan et al. (2022)), multi-source feature fusion-based machine learning framework (Wang L. et al. (2022)), and robust nonnegative matrix factorization model (Peng et al. (2022c)) are widely used in CDA prediction.

Furthermore, Li et al. developed a computational CDA identification method (GATGCN) based on graph attention network and graph convolutional network. First, they fused several biomedical data from different sources through the centered kernel alignment model. Second, graph attention network was deployed to obtain latent representation of circRNAs and diseases. Finally, graph convolutional network was explored to infer CDAs. GATGCN computed better an AUC of 0.951 under leave-one-out cross validation and an AUC of 0.932 under 5-fold cross-validation. They found that circRNAs hsa_circRNA_404833, hsa_circ_0013509, hsa_circRNA_2149,

circR_284, and circR_284 have the highest association scores with lung cancer, diabetes retinopathy, prostate cancer, cholangiocarcinoma, and clear cell renal cell carcinoma, respectively.

A large quantity of transcriptomic data enable us to investigate complex biological processes at single-cell resolution levels (Peng et al. (2022b); Liang et al. (2022a); Zhang et al. (2022b); Xu et al. (2020). Therefore, Miao et al. (2021) considered specific noises and computing efficiency, and then designed biologically interpretable integration strategies to integrate multi-omics single-cell data. Zhou P. et al. (2021) used multiscale stochastic dynamics to dissect transition cells from transcriptome data. Ye et al. (2022) used combinatorial hybrid sequencing to construct the axolotl cell landscape at single-cell resolution. McKellar et al. (2021) detected transitional progenitor states in mouse skeletal muscle regeneration based on single-cell transcriptomic data. Wu et al. (2022b) exploited a stacking ensemble learning-based model to implement single-cell Hi-C classification.

In particular, Panchy et al. analyzed large-scale transcriptome datasets using non-negative principal component analysis and non-negative matrix factorization. The results showed that the above two methods provided low-dimensional features for the progression of biological processes. They found that gene expression signatures from conserved epithelial-mesenchymal transition can be applied to depict the stages in multiple cell lines. Lang et al. evaluated the performance of two sequencing platforms (Nextseq500 and MGISEQ-2000) using the same capture DNA libraries built by the Illumina protocol. The results demonstrated that a significant loss of fragment occurred in the range of 101–133 bp sizes on MGISEQ-2000 for Illumina libraries while not for the capture DNA libraries. Bao et al. considered that it is crucial to differentiate the transcriptomic and proteomic profiles between unstable and stable atherosclerotic plaques. They obtained 5 unstable and 5 stable human carotid atherosclerotic plaques by carotid endarterectomy to identify lncRNA-targeted genes and circRNA-originated genes. The results indicated that 293 proteins, 488 lncRNAs, 91 circRNAs, and 202 mRNAs are differentially expressed between unstable and stable atherosclerotic plaques. Furthermore, CD5L, S100A12, CKB, CEMIP, and SH3GLB1 may be key genes in regulating the stability of atherosclerotic plaques. In addition, Zheng et al. used a series matrix file search method and obtained data related to breast cancer from the ArrayExpress and Gene Expression Omnibus databases. They found that RSK2 is a possible biomarker in breast cancer.

RNA sequencing data have been broadly applied to screen therapeutic strategies for various diseases (Przybyla and Gilbert (2022); Zhang Y. et al. (2021); Li C.-x. et al. (2021)). Chen et al. (2022) used RNA sequencing to explore the mechanism of oxygen-boosted sonodynamic therapy for the treatment of hepatocellular carcinoma. Zhang et al. (2022c) integrated single-cell and bulk RNA sequencing data to probe a pan-cancer stemness signature. Sammut et al. (2022) combined multi-omics data including DNA and RNA sequencing and machine learning technique to predict breast cancer therapy response. Based on RAN sequencing data, Ma et al. first downloaded RNA sequencing data related to gliomas from the TCGA database. Then they used DESeq2, key driver and weighted gene correlation network to identify differentially expressed genes. They observed that Paclitaxel, Cidofovir, 6-benzyladenine, Erlotinib, Bilirubin, Oxaliplatin, Nutlins, Valproic acid, and Fenofibrate may be potential drugs in inhibiting the recurrence of gliomas. Similarly, Xiang et al. detected gene expression and network differences between limited and advanced stages for the diffuse large B-cell lymphoma (DLBCL) patients to predict potential agents against DLBCL. First, they collected RNA sequencing data from the DLBCL patients at different clinical stages from the TCGA database. Second, they used DESeq2 to identify differentially expressed genes and weighted gene correlation network and differential modules to analyze variations between different stages. Finally, they extracted important genes using key drivers and identified potential agents for DLBCL patients using gene-expression perturbations and the CREEDS database. The results indicated that the thistle1 module had high association with the clinical stage of DLBCL. In addition, MOCOS, RAB6C, ACCSL, MMP1, and RGS21 were highly linked to the occurrence and development of DLBCL.

RNAs are a carrier of genetic information and have broad roles in regulating gene expression and other biological processes. Furthermore, the majority of noncoding RNAs are highly associated with diseases including cancers and nontumorigenic diseases. Thus, RNA data analysis contributes to prioritizing previously unrecognized therapeutic targets. We anticipate that this topic can provide clues for the diagnose and prognosis of complex diseases especially cancers.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

Author JY was employed by Geneis (Beijing) Co Ltd.

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

Publisher's note

References

Aftabi, Y., Ansarin, K., Shanehbandi, D., Khalili, M., Seyedrezazadeh, E., Rahbarnia, L., et al. (2021). Long non-coding rnas as potential biomarkers in the prognosis and diagnosis of lung cancer: A review and target analysis. *IUBMB life* 73, 307–327. doi:10.1002/iub.2430

Banerjee, S., Yabalooru, S. R. K., and Karunagaran, D. (2020). Identification of mrna and non-coding rna hubs using network analysis in organ tropism regulated triple negative breast cancer metastasis. *Comput. Biol. Med.* 127, 104076. doi:10.1016/j.compbiomed.2020.104076

Chai, B., Ma, Z., Wang, X., Xu, L., and Li, Y. (2022). Functions of non-coding rnas in regulating cancer drug targets. *Acta Biochim. Biophys. Sin.* 54, 279–291. doi:10.3724/abbs.2022006

Chen, L.-L. (2020). The expanding regulatory mechanisms and cellular functions of circular rnas. *Nat. Rev. Mol. Cell. Biol.* 21, 475–490. doi:10.1038/s41580-020-0243-y

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., and Huang, L. (2017). Lrsslmda: Laplacian regularized sparse subspace learning for mirna-disease association prediction. *PLoS Comput. Biol.* 13, e1005912. doi:10.1371/journal.pcbi.1005912

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. (2018a). Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265. doi:10.1093/bioinformatics/bty503

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018b). Bnpmda: Bipartite network projection for mirna–disease association prediction. *Bioinformatics* 34, 3178–3186. doi:10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z.-H. (2019a). Micrornas and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 20, 515–539. doi:10.1093/bib/bbx130

Chen, X., Yin, J., Qu, J., and Huang, L. (2018c). Mdhgi: Matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput. Biol.* 14, e1006418. doi:10.1371/journal.pcbi.1006418

Chen, X., Zhu, C.-C., and Yin, J. (2019b). Ensemble of decision tree reveals potential mirna-disease associations. *PLoS Comput. Biol.* 15, e1007209. doi:10.1371/journal.pcbi.1007209

Chen, Y., Shang, H., Wang, C., Zeng, J., Zhang, S., Wu, B., et al. (2022). Rna-seq explores the mechanism of oxygen-boosted sonodynamic therapy based on all-in-one nanobubbles to enhance ferroptosis for the treatment of hcc. *Int. J. Nanomedicine* 17, 105–123. doi:10.2147/IJN.S343361

Connelly, C. M., Moon, M. H., and Schneekloth, J. S., Jr (2016). The emerging role of rna as a therapeutic target for small molecules. *Cell. Chem. Biol.* 23, 1077–1090. doi:10.1016/j.chembiol.2016.05.021

Dal Molin, A., Gaffo, E., Difilippo, V., Buratin, A., Tretti Parenzan, C., Bresolin, S., et al. (2022). Craft: A bioinformatics software for custom prediction of circular rna functions. *Brief. Bioinform.* 23, bbab601. doi:10.1093/bib/bbab601

Falese, J. P., Donlic, A., and Hargrove, A. E. (2021). Targeting rna with small molecules: From fundamental principles towards the clinic. *Chem. Soc. Rev.* 50, 2224–2243. doi:10.1039/d0cs01261k

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

Guo, C., Zhou, S., Yi, W., Yang, P., Li, O., Liu, J., et al. (2021). Long non-coding rna muskelin 1 antisense rna (mkln1-as) is a potential diagnostic and prognostic biomarker and therapeutic target for hepatocellular carcinoma. *Exp. Mol. Pathol.* 120, 104638. doi:10.1016/j.yexmp.2021.104638

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. *Brief. Bioinform.* 22, bbaa140. doi:10.1093/bib/bbaa140

Hui, A., How, C., Ito, E., and Liu, F.-F. (2011). Micro-rnas as diagnostic or prognostic markers in human epithelial malignancies. *BMC cancer* 11, 500–509. doi:10.1186/1471-2407-11-500

Ji, B.-Y., You, Z.-H., Wang, Y., Li, Z.-W., and Wong, L. (2021). Dane-mda: Predicting microrna-disease associations via deep attributed network embedding. *Iscience* 24, 102455. doi:10.1016/j.isci.2021.102455

Lan, W., Dong, Y., Chen, Q., Zheng, R., Liu, J., Pan, Y., et al. (2022). Kgancda: Predicting circrna-disease associations based on knowledge graph attention network. *Brief. Bioinform.* 23, bbab494. doi:10.1093/bib/bbab494

Li, C.-x., Chen, J., Lv, S.-k., Li, J.-h., Li, L.-l., and Hu, X. (2021a). Whole-transcriptome rna sequencing reveals significant differentially expressed mrnas, mirnas, and lncrnas and related regulating biological pathways in the peripheral blood of Covid-19 patients. *Mediat. Inflamm.* 2021, 6635925. doi:10.1155/2021/6635925

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, G., Yue, Y., Liang, C., Xiao, Q., Ding, P., and Luo, J. (2019a). Ncpcda: Network consistency projection for circrna–disease association prediction. *RSC Adv.* 9, 33222–33228. doi:10.1039/c9ra06133a

Li, J., Zhao, H., Xuan, Z., Yu, J., Feng, X., Liao, B., et al. (2019b). A novel approach for potential human lncrna-disease association prediction based on local random walk. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 1049–1059. doi:10.1109/TCBB.2019.2934958

Li, L., Gao, Z., Wang, Y.-T., Zhang, M.-W., Ni, J.-C., Zheng, C.-H., et al. (2021b). Scmfmda: Predicting microrna-disease associations based on similarity constrained matrix factorization. *PLoS Comput. Biol.* 17, e1009165. doi:10.1371/journal.pcbi.1009165

Li, Y., Liang, W., Peng, L., Zhang, D., Yang, C., and Li, K.-C. (2022). Predicting drug-target interactions via dual-stream graph neural network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2022, 1–11. doi:10.1109/TCBB.2022.3204188

Liang, Y., Wu, Y., Zhang, Z., Liu, N., Peng, J., and Tang, J. (2022a). Hyb4mc: A hybrid dna2vec-based model for dna n4-methylcytosine sites prediction. *BMC Bioinforma.* 23, 258. doi:10.1186/s12859-022-04789-6

Liang, Y., Zhang, Z.-Q., Liu, N.-N., Wu, Y.-N., Gu, C.-L., and Wang, Y.-L. (2022b). Magcnse: Predicting lncrna-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinforma.* 23, 189. doi:10.1186/s12859-022-04715-w

Licatalosi, D. D., and Darnell, R. B. (2010). Rna processing and its regulation: Global insights into biological networks. *Nat. Rev. Genet.* 11, 75–87. doi:10.1038/nrg2673

Lin, Z.-b., Long, P., Zhao, Z., Zhang, Y.-r., Chu, X.-d., Zhao, X.-x., et al. (2021). Long noncoding rna kcnq1ot1 is a prognostic biomarker and mediates cd8+ t cell exhaustion by regulating cd155 expression in colorectal cancer. *Int. J. Biol. Sci.* 17, 1757–1768. doi:10.7150/ijbs.59001

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

McKellar, D. W., Walter, L. D., Song, L. T., Mantri, M., Wang, M. F., De Vlaminck, I., et al. (2021). Large-scale integration of single-cell transcriptomic data captures transitional progenitor states in mouse skeletal muscle regeneration. *Commun. Biol.* 4, 1280. doi:10.1038/s42003-021-02810-x

Miao, Z., Humphreys, B. D., McMahon, A. P., and Kim, J. (2021). Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* 17, 710–724. doi:10.1038/s41581-021-00463-x

Mukherjee, D., Maiti, S., Gouda, P. K., Sharma, R., Roy, P., and Bhattacharyya, D. (2022). Rnabpdb: Molecular modeling of rna structure—From base pair analysis in crystals to structure prediction. *Interdiscip. Sci.* 14, 759–774. doi:10.1007/s12539-022-00528-w

Peng, L.-H., Sun, C.-N., Guan, N.-N., Li, J.-Q., and Chen, X. (2018). Hnmda: Heterogeneous network-based mirna–disease association prediction. *Mol. Genet. Genomics* 293, 983–995. doi:10.1007/s00438-018-1438-1

Peng, L., Chen, Y., Ma, N., and Chen, X. (2017). Narrmda: Negative-aware and rating-based recommendation algorithm for mirna–disease association prediction. *Mol. Biosyst.* 13, 2650–2659. doi:10.1039/c7mb00499k

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2020). Probing lncrna–protein interactions: Data repositories, models, and algorithms. *Front. Genet.* 10, 1346. doi:10.3389/fgene.2019.01346

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022a). Enanndeep: An ensemble-based lncrna–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. *Interdiscip. Sci.* 14, 209–232. doi:10.1007/s12539-021-00483-y

Peng, L., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncrna-protein interactions based on deep learning with dual-net neural architecture. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2021, 3116232. doi:10.1109/TCBB.2021.3116232

Peng, L., Wang, F., Wang, Z., Tan, J., Huang, L., Tian, X., et al. (2022b). Cell–cell communication inference and analysis in the tumour microenvironments from single-cell transcriptomics: Data resources and computational strategies. *Brief. Bioinform.* 23, bbac234. doi:10.1093/bib/bbac234

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

Ping, P., Wang, L., Kuang, L., Ye, S., Iqbal, M. F. B., and Pei, T. (2018). A novel method for lncrna-disease association prediction based on an lncrna-disease association network. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 688–693. doi:10.1109/TCBB.2018.2827373

Przybyla, L., and Gilbert, L. A. (2022). A new era in functional genomics screens. *Nat. Rev. Genet.* 23, 89–103. doi:10.1038/s41576-021-00409-w

Rajendiran, S., Maji, S., Haddad, A., Lotan, Y., Nandy, R. R., Vishwanatha, J. K., et al. (2021). Microrna-940 as a potential serum biomarker for prostate cancer. *Front. Oncol.* 11, 628094. doi:10.3389/fonc.2021.628094

Sammut, S.-J., Crispin-Ortuzar, M., Chin, S.-F., Provenzano, E., Bardwell, H. A., Ma, W., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. *Nature* 601, 623–629. doi:10.1038/s41586-021-04278-5

Shen, L., Liu, F., Huang, L., Liu, G., Zhou, L., and Peng, L. (2022). Vda-rwlrls: An anti-sars-cov-2 drug prioritizing framework combining an unbalanced bi-random walk and laplacian regularized least squares. *Comput. Biol. Med.* 140, 105119. doi:10.1016/j.compbiomed.2021.105119

Shin, S., Park, Y. H., Jung, S.-H., Jang, S.-H., Kim, M. Y., Lee, J. Y., et al. (2021). Urinary exosome microrna signatures as a noninvasive prognostic biomarker for prostate cancer. *NPJ Genom. Med.* 6, 45–46. doi:10.1038/s41525-021-00212-w

Silva, A. B. O. V., and Spinosa, E. J. (2021). Graph convolutional auto-encoders for predicting novel lncrna-disease associations. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 2264–2271. doi:10.1109/TCBB.2021.3070910

Sun, F., Sun, J., and Zhao, Q. (2022). A deep learning method for predicting metabolite–disease associations via graph neural network. *Brief. Bioinform.* 23, bbac266. doi:10.1093/bib/bbac266

Tang, X., Luo, J., Shen, C., and Lai, Z. (2021). Multi-view multichannel attention graph convolutional network for mirna–disease association prediction. *Brief. Bioinform.* 22, bbab174. doi:10.1093/bib/bbab174

Wang, C.-C., Han, C.-D., Zhao, Q., and Chen, X. (2021a). Circular rnas and complex diseases: From experimental results to computational models. *Brief. Bioinform.* 22, bbab286. doi:10.1093/bib/bbab286

Wang, L., Wong, L., Li, Z., Huang, Y., Su, X., Zhao, B., et al. (2022a). A machine learning framework based on multi-source feature fusion for circrna-disease association prediction. *Brief. Bioinform.* 23, bbac388. doi:10.1093/bib/bbac388

Wang, L., Yan, X., You, Z.-H., Zhou, X., Li, H.-Y., and Huang, Y.-A. (2021b). Sganrda: Semi-supervised generative adversarial networks for predicting circrna–disease associations. *Brief. Bioinform.* 22, bbab028. doi:10.1093/bib/bbab028

Wang, L., You, Z.-H., Huang, D.-S., and Li, J.-Q. (2021c). Mgrcda: Metagraph recommendation method for predicting circrna-disease association. *IEEE Trans. Cybern.* 2021, 1–9. doi:10.1109/TCYB.2021.3090756

Wang, S., Wang, Y., Cheng, H., Zhang, Q., Fu, C., He, C., et al. (2022b). The networks of noncoding rnas and their direct molecular targets in myocardial infarction. *Int. J. Biol. Sci.* 18, 3194–3208. doi:10.7150/ijbs.69671

Wang, W., Dai, Q., Li, F., Xiong, Y., and Wei, D.-Q. (2021d). Mlcdforest: Multi-label classification with deep forest in disease prediction for long non-coding rnas. *Brief. Bioinform.* 22, bbaa104. doi:10.1093/bib/bbaa104

Wang, W., Guan, X., Khan, M. T., Xiong, Y., and Wei, D.-Q. (2020). Lmi-dforest: A deep forest model towards the prediction of lncrna-mirna interactions. *Comput. Biol. Chem.* 89, 107406. doi:10.1016/j.compbiolchem.2020.107406

Wapinski, O., and Chang, H. Y. (2011). Long noncoding rnas and human disease. *Trends Cell. Biol.* 21, 354–361. doi:10.1016/j.tcb.2011.04.001

Wu, H., Liang, Q., Zhang, W., Zou, Q., Hesham, A. E.-L., and Liu, B. (2022a). ilncda-ltr: Identification of lncrna-disease associations by learning to rank. *Comput. Biol. Med.* 2022, 105605. doi:10.1016/j.compbiomed.2022.105605

Wu, H., Wu, Y., Jiang, Y., Zhou, B., Zhou, H., Chen, Z., et al. (2022b). schicstackl: a stacking ensemble learning-based method for single-cell hi-c classification using cell embedding. *Brief. Bioinform.* 23, bbab396. doi:10.1093/bib/bbab396

Xiao, Q., Fu, Y., Yang, Y., Dai, J., and Luo, J. (2021). Nsl2cd: Identifying potential circrna–disease associations based on network embedding and subspace learning. *Brief. Bioinform.* 22, bbab177. doi:10.1093/bib/bbab177

Xu, H., Hu, X., Yan, X., Zhong, W., Yin, D., and Gai, Y. (2022). Exploring noncoding rnas in thyroid cancer using a graph convolutional network approach. *Comput. Biol. Med.* 145, 105447. doi:10.1016/j.compbiomed.2022.105447

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). Cmf-impute: An accurate imputation tool for single-cell rna-seq data. *Bioinformatics* 36, 3139–3147. doi:10.1093/bioinformatics/btaa109

Yang, J., Ju, J., Guo, L., Ji, B., Shi, S., Yang, Z., et al. (2022). Prediction of her2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning. *Comput. Struct. Biotechnol. J.* 20, 333–342. doi:10.1016/j.csbj.2021.12.028

Ye, F., Zhang, G., Chen, H., Yu, C., Yang, L., Fu, Y., et al. (2022). Construction of the axolotl cell landscape using combinatorial hybridization sequencing at single-cell resolution. *Nat. Commun.* 13, 4228. doi:10.1038/s41467-022-31879-z

Yu, A.-M., Jian, C., Allan, H. Y., and Tu, M.-J. (2019). Rna therapy: Are we using the right molecules? *Pharmacol. Ther.* 196, 91–104. doi:10.1016/j.pharmthera.2018.11.011

Yu, N., Liu, Z.-P., and Gao, R. (2022). Predicting multiple types of microrna-disease associations based on tensor factorization and label propagation. *Comput. Biol. Med.* 146, 105558. doi:10.1016/j.compbiomed.2022.105558

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021a). Using network distance analysis to predict lncrna–mirna interactions. *Interdiscip. Sci.* 13, 535–545. doi:10.1007/s12539-021-00458-z

Zhang, S., He, X., Zhang, R., and Deng, W. (2021b). Lncr2metasta: A manually curated database for experimentally supported lncrnas during various cancer metastatic events. *Brief. Bioinform.* 22, bbaa178. doi:10.1093/bib/bbaa178

Zhang, T., Chen, L., Li, R., Liu, N., Huang, X., and Wong, G. (2022a). Piwi-interacting rnas in human diseases: Databases and computational models. *Brief. Bioinform.* 23, bbac217. doi:10.1093/bib/bbac217

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

Zhang, Y., Wang, D., Peng, M., Tang, L., Ouyang, J., Xiong, F., et al. (2021c). Single-cell rna sequencing in cancer research. *J. Exp. Clin. Cancer Res.* 40, 81–17. doi:10.1186/s13046-021-01874-1

Zhang, Z., Cui, F., Cao, C., Wang, Q., and Zou, Q. (2022b). Single-cell rna analysis reveals the potential risk of organ-specific cell types vulnerable to sars-cov-2 infections. *Comput. Biol. Med.* 140, 105092. doi:10.1016/j.compbiomed.2021.105092

Zhang, Z., Wang, Z.-X., Chen, Y.-X., Wu, H.-X., Yin, L., Zhao, Q., et al. (2022c). Integrated analysis of single-cell and bulk rna sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. *Genome Med.* 14, 45–18. doi:10.1186/s13073-022-01050-w

Zhao, X., Zhao, X., and Yin, M. (2022). Heterogeneous graph attention network based on meta-paths for lncrna–disease association prediction. *Brief. Bioinform.* 23, bbab407. doi:10.1093/bib/bbab407

Zhou, G., Jiang, N., Zhang, W., Guo, S., and Xin, G. (2021a). Biomarker identification in membranous nephropathy using a long non-coding rna-mediated competitive endogenous rna network. *Interdiscip. Sci.* 13, 615–623. doi:10.1007/s12539-021-00466-z

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021b). Lpi-deepgbdt: A multiple-layer deep framework based on gradient boosting decision trees for lncrna–protein interaction identification. *BMC Bioinforma.* 22, 479. doi:10.1186/s12859-021-04399-8

Zhou, P., Wang, S., Li, T., and Nie, Q. (2021c). Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. *Nat. Commun.* 12, 5609–5615. doi:10.1038/s41467-021-25548-w

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fgene.2021.690096

Predicting lncRNA–Protein Interaction With Weighted Graph-Regularized Matrix Factorization

Xibo Sun[1], Leiming Cheng[2], Jinyang Liu[3,4], Cuinan Xie[3,4], Jiasheng Yang[5] and Fu Li[6]**

[1] Yidu Central Hospital of Weifang, Weifang, China, [2] Huaibei Kuanggong Zong Yiyuan, Huaibei, China, [3] Geneis Beijing Co., Ltd., Beijing, China, [4] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [5] Academician Workstation, Changsha Medical University, Changsha, China, [6] Department of Thoracic Surgery, The Second Affiliated Hospital of Hainan Medical University, Haikou, China

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Guanghui Li,
East China Jiaotong University, China
JunLin Xu,
Hunan University, China

***Correspondence:**
Jiasheng Yang
jsyang.mcc@gmail.com
Fu Li
lifu_3251@163.com

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

Received: *02 April 2021*
Accepted: *21 May 2021*
Published: *16 July 2021*

Citation:
Sun X, Cheng L, Liu J, Xie C, Yang J and Li F (2021) Predicting lncRNA–Protein Interaction With Weighted Graph-Regularized Matrix Factorization. Front. Genet. 12:690096. doi: 10.3389/fgene.2021.690096

Long non-coding RNAs (lncRNAs) are widely concerned because of their close associations with many key biological activities. Though precise functions of most lncRNAs are unknown, research works show that lncRNAs usually exert biological function by interacting with the corresponding proteins. The experimental validation of interactions between lncRNAs and proteins is costly and time-consuming. In this study, we developed a weighted graph-regularized matrix factorization (LPI-WGRMF) method to find unobserved lncRNA–protein interactions (LPIs) based on lncRNA similarity matrix, protein similarity matrix, and known LPIs. We compared our proposed LPI-WGRMF method with five classical LPI prediction methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and collaborative filtering (CF). The results demonstrate that the LPI-WGRMF method can produce high-accuracy performance, obtaining an AUC score of 0.9012 and AUPR of 0.7324. The case study showed that SFPQ, SNHG3, and PRPF31 may associate with Q9NUL5, Q9NUL5, and Q9UKV8 with the highest linking probabilities and need to further experimental validation.

Keywords: lncRNA–protein interaction, weighted graph-regularized matrix factorization, lncRNA similarity, protein similarity, SFPQ, SNHG3, PRPF31

INTRODUCTION

Long non-coding RNAs (lncRNAs) are closely associated with many key biological processes, for example, immune response, embryonic stem cell pluripotency, and cell cycle regulation (Chen et al., 2016; Agirre et al., 2019; Gil and Ulitsky, 2020). lncRNAs regulate cellular activities to achieve their biological function through interactions with proteins (Chen and Yan, 2013; Zhang et al., 2018b). Therefore, finding potential lncRNA–protein interactions (LPIs) is important to uncover lncRNA-related biological activities. Wet experiments found a few LPIs; however, experimental methods are costly and time-consuming. Thus, computational methods are developed to identify possible associations between lncRNAs and proteins (Bester et al., 2018; Chen et al., 2018).

LPI prediction methods can be roughly classified into two groups: network-based methods and machine learning-based methods. Network-based LPI identification methods integrated various biological data and network propagation methods (Peng et al., 2019). Li et al. (2015) used random walk with restart on the constructed lncRNA–protein heterogeneous network to find LPI candidates. Zhang et al. (2018a) developed a linear neighborhood propagation method to score for lncRNA–protein pairs. Ge et al. (2016), Zhao et al. (2018a), and Xie et al. (2019) applied bipartite network projection recommended methods to compute the association probabilities between lncRNAs and proteins.

Machine learning-based methods mainly contain matrix factorization-based LPI prediction methods and ensemble learning-based LPI prediction methods. Matrix factorization methods have been widely applied to various association prediction areas (Peng et al., 2020). Liu et al. (2017), Zhang T. et al. (2018), Zhao et al. (2018a), and Shen et al. (2019) used matrix factorization methods to predict possible LPIs. Hu et al. (2018) and Zhang et al. (2018b) utilized ensemble techniques and generated ensemble learning frameworks to discover potential LPIs based on the constructed benchmark datasets. Computational methods effectively revealed the possible associations between lncRNAs and proteins. However, the performance obtained by the above methods is limited and can be further improved.

In this study, we first integrated lncRNA similarity, protein similarity, known LPIs. We then developed a novel LPI prediction method based on weighted graph-regularized matrix factorization (LPI-WGRMF). LPI-WGRMF was compared with five state-of-the-art LPI methods [LPBNI, LPI-IBNRA, LPIHN, RWR, and collaborative filtering (CF)] to measure the performance of the proposed LPI-WGRMF method. LPI-WGRMF obtained the AUC value of 0.9057 and the AUPR value of 0.7324. The results showed that LPI-WGRMF is a useful tool for identifying LPIs. Case study analysis suggests that there are possibly joint links between SFPQ and Q9NUL5, SNHG3 and Q9NUL5, and PRPF31 and Q9UKV8.

MATERIALS AND METHODS

In this manuscript, we developed an LPI prediction model, LPI-WGRMF. The method can be summarized to three steps. First, experimentally validated LPIs from the NPInter 2.0 database were collected. Second, lncRNA similarity matrix and protein similarity matrix are computed based on the assumption that lncRNAs tend to associate with similar proteins and vice versa. Finally, lncRNA similarity, protein similarity, and LPI matrix were integrated to the weight graph-regularized matrix factorization model for computing the association scores for each lncRNA–protein pair.

Materials

LPI Data

We obtained experimentally validated LPI dataset, which was provided by Zhang et al. (2018a). The dataset contains 4158 LPIs between 990 lncRNAs and 27 proteins after preprocessing. The LPI matrix between n lncRNAs and m proteins was denoted as $Y_{n\times m}$.

lncRNA Similarity Matrix

The sequence and expression information of lncRNAs can be downloaded from the NONCODE database. We computed lncRNA similarity matrix by integrating the sequence similarity, expression similarity, and interaction similarity to the similarity kernel fusion technique.

Sequence statistical similarity

Each lncRNA was described a 20-dimensional vector based on the methods provided by Zhang et al. (2018b). Based on the assumption that each vector can be denoted by their k-nearest neighbors, linear neighborhood similarity between two lncRNAs l_i and l_j can be computed and denoted as $s_{l,0}\left(i,j\right)$.

Expression similarity

Suppose that the expression profile of the i^{th} lncRNA can be represented as e_i and thus the expression similarity between two lncRNAs l_i and l_j can be defined as:

$$s_{l,1}\left(i,j\right)=\begin{cases}\frac{1}{2}\left(1+\rho_{i,j}\right) & i\neq j\\ 0 & i=j\end{cases} \tag{1}$$

where $\rho_{i,j}$ is the Pearson's correlation coefficient between two expression profiles e_i and e_j and is defined as:

$$\rho_{i,j}=\frac{cov(e_i,e_j)}{\sigma(e_i)\sigma(e_j)} \tag{2}$$

where $cov()$ denotes the covariance and σ denotes the standard deviation.

Interaction profile similarity

Suppose that the interaction profile of the i^{th} lncRNA can be represented as the i^{th} row $Y_{i.}$ Of the LPI matrix Y, the interaction profile similarity between two lncRNAs l_i and l_j can be defined as:

$$s_{l,2}\left(i,j\right)=\exp\left(-\frac{1}{\gamma_l}||Y_{i.}-Y_{j.}||^2\right) \tag{3}$$

where

$$\gamma_l=\frac{1}{n}\sum_{i\ 1}^{n}||Y_{i.}||^2 \tag{4}$$

where $||\cdot||$ denotes the 2-norm of a matrix.

Protein Similarity Matrix

Sequence alignment similarity

The sequences of proteins were downloaded from the SUPERFAMILY database. The alignment score of the u^{th} protein against the v^{th} protein can be computed by Blast and be denoted as $b_{u,v}$. The sequence similarity between two proteins p_u and p_v can be defined as:

$$s_{p,0}\left(u,v\right)=\begin{cases}\frac{b_{u,v}}{b_{u,u}} & u\neq v\\ 0 & u=v\end{cases} \tag{5}$$

Sequence statistical feature similarity

Each protein can be represented as a 504-dimensional vector based on the method provided by Zhou et al. (2020). Linear neighborhood similarity between two proteins p_u and p_v can be computed and denoted as $s_{p,1}$.

Interaction profile similarity

Suppose that the interaction profile of the u^{th} protein can be represented as the u^{th} column $Y_{.u}$ of the LPI matrix Y, the interaction profile similarity between two proteins p_u and p_v can be defined as:

$$s_{p,2}(u,v) = \exp\left(-\frac{1}{\gamma_l}||Y_{.u} - Y_{.v}||^2\right) \tag{6}$$

where

$$\gamma_l = \frac{1}{n}\sum_{u=1}^{m}||Y_{.u}||^2 \tag{7}$$

Similarity Kernel Fusion

In the above sections, three lncRNA similarity measurements and three protein similarity measurements were proposed. The similarity kernel fusion method provided by Zhou et al. (2020) was applied to integrate this similarity information to compute a more comprehensive similarity.

First, the three lncRNA similarities were normalized as follows:

$$\theta_{l,q}(i,j) = \frac{s_{l,q}(i,j)}{\sum_{t=1}^{n} s_{l,q}(t,j)}, \ (q = 0, 1, 2) \tag{8}$$

The normalized similarity matrix was denoted as:

$$\Theta_{l,q} = \{\theta_{l,q}(i,j)\}_{n\times n} \tag{9}$$

Second, for an lncRNA l_i and $s_{l,q}$, the k most similar lncRNAs were collected as a set $N_{l,q}(i,k)$ and $s_{l,q}$ can be normalized in constraint based on the neighborhood information:

$$\varphi_{l,q}(i,j) = \frac{s_{l,q}(i,j)\, I_{l,q,k}(i,j)}{\sum_{t=1}^{n} s_{l,q}(i,t)\, I_{l,q,k}(i,t)} \tag{10}$$

where

$$I_{l,q,k}(i,j) = \begin{cases} 1 & l_j \in N_{l,q}(u,k) \\ 0 & l_j \notin N_{l,q}(u,k) \end{cases} \tag{11}$$

The neighborhood constrained normalized matrix was denoted as:

$$\phi_{l,q} = \{\varphi_{l,q}(i,j)\}_{n\times n} \tag{12}$$

The above three normalized matrices were integrated based on the following iterative process:

$$\Theta_{l,q}(\lambda+1) = \frac{1}{2}\alpha\left(\phi_{l,q}\sum_{r\neq q}\Theta_{l,r}(\lambda)\phi_{l,q}^{T}\right) + \frac{1}{2}(1-\alpha)\sum_{r\neq q}\Theta_{l,r}(0) \tag{13}$$

where α was a weight parameter with $0 \quad \alpha \quad 1$, T was the transpose of the matrix, λ represented the iterative parameter, and $\Theta_{l,r}(0) \quad \Theta_{l,r}$.

We computed the integrated similarity matrix after z rounds of iteration:

$$\Theta_l = \frac{1}{3}\left(\Theta_{l,0}(z) + \Theta_{l,1}(z) + \Theta_{l,2}(z)\right) \tag{14}$$

By considering data noise, we defined the following indicator function based on the k most similar lncRNAs for each lncRNA:

$$w_{l,k} = \begin{cases} 1 & I_{l,0,k}(i,j) = I_{l,1,k}(i,j) = I_{l,2,k}(i,j) = 1 \\ 0 & I_{l,0,k}(i,j) = I_{l,1,k}(i,j) = I_{l,2,k}(i,j) = 0 \\ 0.5 & otherwise \end{cases} \tag{15}$$

The final lncRNA similarity matrix can be denoted as follows:

$$S_{l,k} = \{\vartheta_l(i,j)\, w_{l,k}(i,j)\}_{n\times n} \tag{16}$$

where $\vartheta_l(i,j)$ is the $(i,j)^{th}$ element in the matrix Θ_l.

Nearest Neighbor Information

Based on the graph regularization theory, similar lncRNAs should tend to interact with similar proteins and vice versa in an LPI network, and thus we first observe the nearest neighbor information for lncRNAs and proteins. Given the lncRNA similarity matrix S^l, we represented a p-nearest neighbor graph N as

$$N_{ij} = \begin{cases} 1 & j \in N_p(i) \ \& \ i \in N_p(j) \\ 0 & j \notin N_p(i) \ \& \ i \notin N_p(j) \\ 0.5 & otherwise \end{cases} \tag{17}$$

where $N_p(i)$ denotes the set of p nearest neighbors of lncRNA l_i. N is applied to increase the sparsify of the lncRNA similarity matrix S^l as

$$\forall i,j \quad \hat{S}_{ij}^{l} = N_{ij}S_{ij}^{l} \tag{18}$$

Thus, the sparse similarity matrix of lncRNAs can be computed. Similarity, the sparse similarity matrix of protein can be done.

Low-Rank Approximation

Based on low-rank approximation idea, the LPI matrix $Y \in \mathbb{R}^{n=m}$ can be decomposed into two low-rank latent feature matrices $A \in \mathbb{R}^{n=k}$ (for lncRNAs) and $B \in \mathbb{R}^{m=k}$ (for proteins) by minimizing the following low-rank approximation objective:

$$\min_{A,B}||Y - AB^T||_F^2 \tag{19}$$

where $||\cdot||_F$ denotes the Fronbenius norm and k is the rank of matrices A and B, that is, the number of features in A and B.

We decomposed $Y \in \mathbb{R}^{n=m}$ into $U \in \mathbb{R}^{n=k}$, $S_k \in \mathbb{R}^{k=k}$, and V$\in \mathbb{R}^{m=k}$ so that US_kV^T is the closest k-rank approximation to Y where U and V are matrices with orthonormal columns, S_k is a diagonal matrix, and $k_{max} = \min(n,m)$. Thus, the feature matrices A and B can be represented as $A = US_k^{1/2}$ and $B = VS_k^{1/2}$.

Graph-Regularized Matrix Factorization

To boost generalization ability and prevent overfitting, we minimize the following GRMF's objective function by adding Tikhonov and graph regularization terms to the above low-rank approximation:

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l \sum_{i,r=1}^{n} \hat{S}_{ij}^l ||a_i - a_r||^2$$
$$+\lambda_p \sum_{j,q=1}^{n} \hat{S}_{ij}^p ||b_j - b_q||^2 \quad (20)$$

where λ_f, λ_l, and λ_p are positive parameters, a_i and b_j are the i^{th} and j^{th} rows of A and B, respectively, and n and m are the numbers of lncRNAs and proteins, respectively. The first term is used to make the model approximate the matrix Y. The second term (Tikhonov regularization) minimizes the norms of A and B. The third and final terms are lncRNA graph regularization and protein graph regularization, respectively. The two terms are applied to minimize the distance between feature vectors of two neighboring lncRNAs or proteins. Based on graph regularization, the above model can be redescribed as

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \mathcal{L}_l A)$$
$$+\lambda_p Tr(B^T \mathcal{L}_p B) \quad (21)$$

where $Tr(\cdot)$ denotes the trace of matrix, $\mathcal{L}_l = D^l - \hat{S}^l$ and $\mathcal{L}_p = D^p - \hat{S}^p$ represent the graph Laplacian terms for $\hat{S}^l$ and $\hat{S}^p$, respectively, and D^l and D^p are diagonal matrices where $D_{ii}^l = \sum_r \hat{S}_{ir}^l$ and $D_{jj}^t = \sum_q \hat{S}_{jq}^p$.

To improve LPI prediction performance, we normalize graph Laplacians $\mathcal{L}_l$ and $\mathcal{L}_p$ by $\widetilde{\mathcal{L}}_l = (D^l)^{-1/2} \widetilde{\mathcal{L}}_l (D^l)^{-1/2}$ and $\mathcal{L}_p = D^p - \hat{S}^p$. Equation (4) can be rewritten as

$$\min_{A,B} ||Y - AB^T||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \widetilde{\mathcal{L}}_l A)$$
$$+\lambda_p Tr(B^T \widetilde{\mathcal{L}}_p B) \quad (22)$$

Weighted Graph-Regularized Matrix Factorization

To prevent unknown lncRNA–protein pairs from affecting the performance of singular value decomposition produced by Y, we add a weight matrix W into the objective function as follows:

$$\min_{A,B} ||W \odot (Y - AB^T)||_F^2 + \lambda_f(||A||_F^2| + |B||_F^2) + \lambda_l Tr(A^T \widetilde{\mathcal{L}}_l A)$$
$$+\lambda_p Tr(B^T \widetilde{\mathcal{L}}_p B) \quad (23)$$

Based on the alternating least square method provided by Ezzat et al. (2016), we can solve the model (6). Let $\frac{\partial L}{\partial a_i} = 0$ and $\frac{\partial L}{\partial b_j} = 0$, run alternatingly the following two update rules until convergence:

$$\forall i = 1, 2, ...n,$$

$$a_i = \left(\sum_{j=1}^{m} W_{ij} Y_{ij} b_j - \lambda_l (\tilde{L}_l)_{i*} A\right)\left(\sum_{j=1}^{m} W_{ij} b_j^T b_j \lambda_f I_k\right)^{-1} \quad (24)$$

$$\forall j = 1, 2, ...m,$$

$$b_i = \left(\sum_{i=1}^{n} W_{ij} Y_{ij} a_i - \lambda_p (\tilde{L}_p)_{j*} B\right)\left(\sum_{j=1}^{n} W_{ij} a_i^T a_i \lambda_f I_k\right)^{-1} \quad (25)$$

where $(\tilde{L}_l)_{i*}$ and $(\tilde{L}_p)_{j*}$ are the i^{th} and j^{th} rows vectors of $\widetilde{\mathcal{L}}_l$ and $\widetilde{\mathcal{L}}_p$, respectively.

We can obtain A and B based on Eqs 7 and 8. Finally, the interaction probability between the i^{th} lncRNA and the j^{th} protein can be computed by

$$Y = AB^T \quad (26)$$

RESULTS

Experimental Settings

We conducted three different fivefold cross validation on the training dataset to set LPI-WGRMF's parameters, that is, k (the rank of matrices A and B), p (the number of nearest neighbors), λ_l, λ_d, *and* λ_t. We set the parameters as $k \in \{50, 100\}$, $p \in \{1, 2, 3, 4, 5, 6, 7\}$, $\lambda_f \in \{2^{-2}, 2^{-1}, 2^0, 2^1\}$, $\lambda_l \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, *and* $\lambda_p \in \{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. And we used grid search and found that the best parameter combination is $k = 50$, $p = 7$, $\lambda_f = 0.5$, $\lambda_l = 0.3$, *and* $\lambda_p = 0.005$.

Evaluation Metrics

Precision, recall, f1 score, accuracy, AUC, and AUPR are widely applied to measure the performance of machine learning methods on association prediction. In this study, we used the six measurements to evaluate the performance of our proposed LPI-WGRMF. AUC is the area under the receiver operating characteristics curve. AUPR is the area under precision–recall curve. The other four criteria are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (27)$$

$$Recall = \frac{TP}{TP + FN} \quad (28)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (29)$$

$$\text{f1 } score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (30)$$

where TP and FP denote the predicted true and false number of positive LPIs, respectively, and TN and FN denote the predicted true and false number of negative LPIs, respectively. The experiments were conducted 20 times. The average precision, recall, accuracy, AUC, and AUPR values for 20 times of experiments were computed as the final performance.

Performance Comparison of LPI-WGRMF and Other Methods

To measure the performance of our proposed LPI-WGRMF method, we compared LPI-WGRMF and five state-of-the-art methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and CF. LPBNI is a bipartite network inference method; LPIHN is a heterogeneous network inference method based on random walk with restart. The two models obtained better prediction performance in the area of LPI identification and are state-of-the-art LPI prediction methods. The experiments were conducted 20 times under fivefold cross validation. The results are shown in **Table 1**. The best performance in each column (measurement metric) is denoted in bold in **Table 1**.

Higher precision, recall, accuracy, and AUC denote better performance. From **Table 1**, we can find that LPI-WGRMF significantly outperformed other five methods in terms of precision, recall, and AUC. Precision computed by LPI-WGRMF was better 59.27, 45.32, 55.74, 61.17, and 67.44% than LPBNI, LPI-IBNRA, LPIHN, RWR, and CF, respectively. Recall computed by LPI-WGRMF was better 36.83, 34.83, 56.19, 44.91, and 53.86%, respectively. F1-score computed by LPI-WGRMF was better 36.83, 30.37, 56.19, 44.91, and 53.86%, respectively. AUC of LPI-WGRMF was higher 5.39, 3.74, 6.69, 10.19, and 15.14%, respectively. AUPR of LPI-WGRMF was higher 54.92, 40.59, 68.61, 61.40, and 67.82%, respectively.

Although accuracy computed by LPI-WGRMF was lower than LPBNI, LPI-WGRMF obtained better precision, recall, and AUC. More importantly, AUC and AUPR are more representative measurement metrics compared with other three evaluation metrics. Thus, AUC and AUPR can be more effectively applied to evaluate the performance of LPI prediction models. LPI-WGRMF is a powerful tool for LPI identification because of its better precision, recall, AUC, and AUPR. **Figures 1**, **2** demonstrate the AUC and AUPR values obtained by the six LPI prediction methods. The results show that LPI-WGRMF obtained the best AUC value, thereby demonstrating LPI-WGRMF's powerful LPI prediction capability.

Case Study

We further conducted four case studies after confirming the performance of LPI-WGRMF. The lncRNAs in the four cases are Splicing Factor Proline and Glutamine Rich (SFPQ), FOrkhead boX protein D2-Adjacent Opposite Strand RNA 1 (FOXD2-AS1), Small Nucleolar RNA Host Gene 3 (SNHG3), and Pre-mRNA-Processing Factor 31 (PRPF31), respectively. We predicted possible LPIs based on lncRNA similarities, protein similarities, known LPIs, and LPI-WGRMF. **Table 2** lists the predicted top five proteins associated with the above four lncRNAs.

SFPQ is a multifunctional nuclear protein participating in a few cellular activities including RNA transport, apoptosis, and DNA repair. SFPQ is densely associated with several diseases including renal cell carcinoma, Xp11-associated tumor, and dyslexia. More importantly, the expression levels of SFPQ impact on the sensitivity of ovarian cancer cells to PT-induced death (Gao et al., 2019; Pellarin et al., 2020). **Table 2** shows that SFPQ has joint connection with Q9NUL5 (ranked as 2). More importantly, the association between SFPQ and Q9NUL5 is ranked as 1 in all other five LPI identification methods. The fact suggests that SFPQ is possibly to link with Q9NUL5.

FOXD2-AS1 is an RNA gene and is abnormally expressed in a variety of malignant tumors. FOXD2-AS1 has close associations with many diseases, for example, nasopharyngeal carcinoma, esophageal cancer, bladder cancer, multiple pterygium syndrome, escobar variant, and ulcerative colitis (Bao et al., 2018; Chen et al., 2018; Su et al., 2018; Huang et al., 2020; Liu et al., 2020). FOXD2-AS1 was predicted to be closely linking with O00425, Q9NZI8, Q9Y6M1, and Q9NUL5, which was ranked as 1, 2, 3, and 4. All these connections were ranked in the top five associations among other five LPI prediction models. Therefore, FOXD2-AS1 is associated with O00425, Q9NZI8, Q9Y6M1, and Q9NUL5.

SNHG3 is a newly found lncRNA and was discovered as a biomarker of malignant cancers, for example, ovarian cancer, hepatocellular carcinoma, colorectal cancer, lung cancer, and

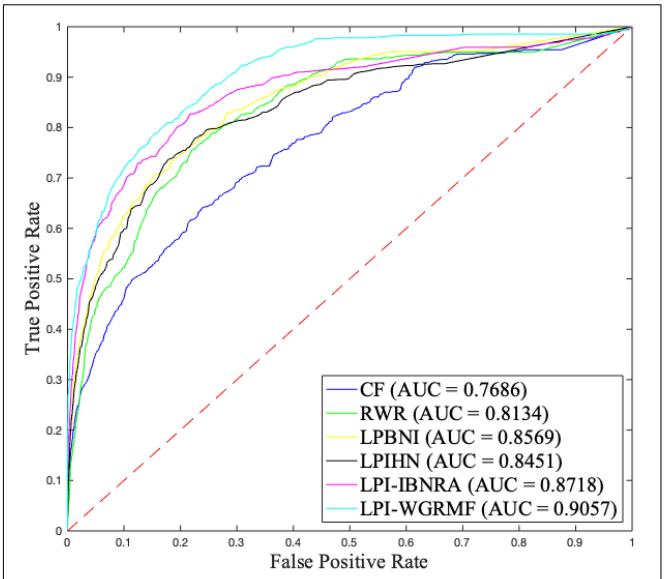


FIGURE 1 | The AUC values of six LPI prediction methods.

TABLE 1 | The performance of five LPI prediction methods.

Methods	Precision	Recall	Accuracy	F1-score	AUC	AUPR
LPBNI	0.3794	0.4037	0.9573	0.3876	0.8569	0.3302
LPI-IBNRA	0.5093	0.4165	**0.9641**	0.4521	0.8718	0.4351
LPIHN	0.4122	0.2800	0.9412	0.3324	0.8451	0.2299
RWR	0.3617	0.3521	0.9531	0.3543	0.8134	0.2827
CF	0.3033	0.2949	0.9488	0.2965	0.7686	0.2357
LPI-WGRMF	**0.9314**	**0.6391**	0.8906	**0.6493**	**0.9057**	**0.7324**

The best performance in each column (measurement metric) is denoted in bold.

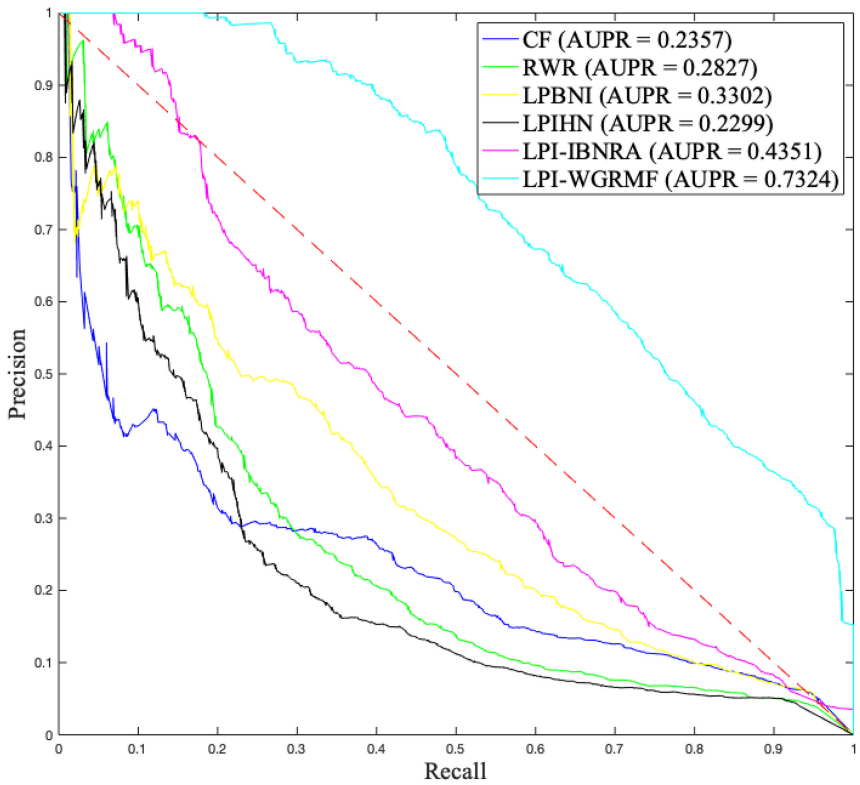


FIGURE 2 | The AUPR values of six LPI prediction methods.

TABLE 2 | The top five proteins associated with the four lncRNAs.

lncRNAs	Proteins	Confirmed	LPI-WGRMF	LPBNI	LPIIBNRA	LPIHN	RWR	CF
MTND2P28	Q9NUL5	NO	1	1	4	2	7	2
	O00425	YES	2	2	2	1	1	1
	P26599	YES	3	8	10	11	4	11
	Q07955	YES	4	16	17	18	5	15
	Q9Y6M1	YES	5	3	1	3	2	3
RPI001_1001892	Q9NUL5	YES	1	1	1	1	1	1
	Q07955	YES	2	9	13	15	8	13
	P35637	YES	3	5	5	5	4	5
	P26599	YES	4	15	17	16	9	16
	Q9NZI8	YES	5	4	4	3	5	3
RPI001_1002045	Q9NUL5	YES	1	1	1	1	1	1
	P35637	YES	2	4	2	5	4	5
	Q01844	YES	3	6	6	6	6	6
	P31483	YES	4	9	10	8	7	9
	Q9Y6M1	YES	5	3	4	3	3	3
RP11-169K16.7	Q9UKV8	YES	1	1	1	1	2	1
	Q9H9G7	YES	2	2	4	2	1	7
	Q9UL18	YES	3	7	3	4	4	10
	Q9HCK5	YES	4	6	2	3	3	9
	Q9NUL5	YES	5	5	5	6	5	2

glioma (Zhang et al., 2016; Huang et al., 2017; Lu et al., 2019; Liu and Tao, 2020). The results from case study analyses showed that SNHG3 tends to link with Q9NUL5 (ranked as 1) and has highest association scores with the protein in LPNI, BPIHN, and CF. Thus, SNHG3 may be possibly linked with Q9NUL5.

PRPF31 is one retinitis pigmentosa-causing gene. Its genetic variants have joint connections with variation in response to metformin in patients with type 2 diabetes (Kiser et al., 2019). In our predicted results, PRPF31 was found to be densely associated with Q9UKV8 (ranked as 1). More importantly, the association between PRPF31 and Q9UKV8 was identified to be ranked as 1, 1, 2, and 1 in LPBNI, LPIHN, RWR, and CF, respectively. PRPF31 obtained the highest association score with Q9UKV8 in five models.

DISCUSSION AND CONCLUSION

In this manuscript, we developed a novel method LPI-WGRMF for identifying possible LPIs, based on lncRNA similarity, protein similarity, known LPIs, and weighted graph regularization-based matrix factorization. We first integrated the similarity information and known LPIs as the initial resource. We then proposed a weighted graph-regularized matrix factorization model to compute the association scores for lncRNA–protein pairs.

LPI-WGRMF was compared with five classical LPI methods, that is, LPBNI, LPI-IBNRA, LPIHN, RWR, and CF. Cross-validation experiments were conducted for 20 times. The results showed the powerful performance of LPI-WGRMF. We conducted four case study analyses after confirming the LPI-WGRMF's accuracy. The results suggest that there are possibly close associations between SFPQ and Q9NUL5, SNHG3 and Q9NUL5, and PRPF31 and Q9UKV8 and need to further experimental validation.

In the future, other sources of LPI-related data may be used to improve the prediction performance, for example, using multiple kernels and designing a multiple kernel learning-based algorithm to effectively integrate the abundant lncRNA and protein information.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

FL and JY conceived, designed, and managed the study. XS and LC designed the LPI-WGRMF method, ran LPI-WGRMF, and wrote the original manuscript. JL and CX revised the original draft. XS, JL, and CX discussed the proposed method and gave further research. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank all authors of the cited references.

REFERENCES

Agirre, X., Meydan, C., Jiang, Y., Garate, L., Doane, A. S., Li, Z., et al. (2019). Long non-coding RNAs discriminate the stages and gene regulatory states of human humoral immune response. *Nat. Commun.* 10:821.

Bester, A. C., Lee, J. D., Chavez, A., Lee, Y.-R., Nachmani, D., Vora, S., et al. (2018). An integrated genome-wide crispra approach to functionalize lncrnas in drug resistance. *Cell* 173, 649–664. doi: 10.1016/j.cell.2018.03.052

Bao, J., Zhou, C., Zhang, J., Mo, J., Ye, Q., He, J., et al. (2018). Upregulation of the long noncoding RNA FOXD2-AS1 predicts poor prognosis in esophageal squamous cell carcinoma. *Cancer Biomark.* 21, 527–533. doi: 10.3233/CBM-170260

Chen, X., Sun, Y.-Z., Guan, N.-N., Qu, J., Huang, Z.-A., Zhu, Z.-X., et al. (2018). Computational models for lncrna function prediction and functional similarity calculation. *Brief. Funct. Genom.* 18, 58–82. doi: 10.1093/bfgp/ely031

Chen, X., Yan, C. C., Zhang, X., and You, Z.-H. (2016). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

Ezzat, A., Zhao, P., Wu, M., Li, X. L., and Kwoh, C. K. (2016). Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 14, 646–656. doi: 10.1109/TCBB.2016.2530062

Gao, Z., Chen, M., Tian, X., Chen, L., Chen, L., Zheng, X., et al. (2019). A novel human lncRNA SANT1 cis-regulates the expression of SLC47A2 by altering SFPQ/E2F1/HDAC1 binding to the promoter region in renal cell carcinoma. *RNA Biol.* 16, 940–949. doi: 10.1080/15476286.2019.1602436

Ge, M., Li, A., and Wang, M. (2016). A bipartite network-based method for prediction of long non-coding rna-protein interactions. *Genomics Proteomics Bioinform.* 14, 62–71. doi: 10.1016/j.gpb.2016.01.004

Gil, N., and Ulitsky, I. (2020). Regulation of gene expression by cis-acting long non-coding RNAs. *Nat. Rev. Genet.* 21, 102–117. doi: 10.1038/s41576-019-0184-5

Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi- ensemble: prediction of human lncrna-protein interactions based on ensemble strategy. *RNA Biol.* 15, 797–806. doi: 10.1080/15476286.2018.1457935

Huang, W., Tian, Y., Dong, S., Cha, Y., Li, J., Guo, X., et al. (2017). The long non-coding RNA SNHG3 functions as a competing endogenous RNA to promote malignant development of colorectal cancer. *Oncol. Rep.* 38, 1402–1410. doi: 10.3892/or.2017.5837

Huang, Y., Yuan, K., Tang, M., Yue, J. M., Bao, L. J., Wu, S., et al. (2020). Melatonin inhibiting the survival of human gastric cancer cells under ER stress involving autophagy and Ras-Raf-MAPK signalling. *J. Cell. Mol. Med.* 2020, 1480–1492. doi: 10.1111/jcmm.16237

Kiser, K., Webb-Jones, K. D., Bowne, S. J., Sullivan, L. S., Daiger, S. P., and Birch, D. G. (2019). Time course of disease progression of PRPF31-mediated retinitis pigmentosa. *Am. J. Ophthalmol.* 200, 76–84.

Li, A., Ge, M., Zhang, Y., Peng, C., and Wang, M. (2015). Predicting long noncoding rna and protein interactions using heterogeneous network model. *BioMed. Res. Int.* 2015:671950. doi: 10.1155/2015/671950

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., Zhao, Q., et al. (2020). Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl. Based Syst.* 191:105261. doi: 10.1016/j.knosys.2019.105261

Liu, H., Ren, G., Hu, H., Zhang, L., Ai, H., Zhang, W., et al. (2017). Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* 8:103975. doi: 10.18632/oncotarget.21934

Liu, Z., and Tao, H. (2020). Small nucleolar RNA host gene 3 facilitates cell proliferation and migration in oral squamous cell carcinoma via targeting nuclear transcription factor Y subunit gamma. *J. Cell. Biochem.* 121, 2150–2158.

Lu, W., Yu, J., Shi, F., Zhang, J., Huang, R., Yin, S., et al. (2019). The long non-coding RNA Snhg3 is essential for mouse embryonic stem cell self-renewal and pluripotency. *Stem Cell Res. Ther.* 10:157. doi: 10.1002/jcb.29421

Pellarin, I., Dall'Acqua, A., Gambelli, A., Pellizzari, I., D'Andrea, S., Sonego, M., et al. (2020). Splicing factor proline-and glutamine-rich (SFPQ) protein regulates platinum response in ovarian cancer-modulating SRSF2 activity. *Oncogene* 39, 4390–4403. doi: 10.1038/s41388-020-1292-6

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2019). Probing lncRNA-protein interactions: data repositories, models, and algorithms. *Front. Genet.* 10:1346. doi: 10.3389/fgene.2019.01346

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020). RNMFMDA: a microbe-disease association identification method based on reliable negative sample selection and logistic matrix factorization with neighborhood regularization. *Front. Microbiol.* 11:592430. doi: 10.3389/fmicb.2020.592430

Su, F., He, W., Chen, C., Liu, M., Liu, H., Xue, F., et al. (2018). The long non-coding RNA *FOXD2-AS1* promotes bladder cancer progression and recurrence through a positive feedback loop with Akt and E2F1. *Cell Death Dis.* 9, 1–17. doi: 10.1038/s41419-018-0275-9

Shen, C., Ding, Y., Tang, J., Jiang, L., and Guo, F. (2019). Lpi-ktaslp: prediction of lncrna-protein interaction by semi-supervised link learning with multivariate information. *IEEE Access* 7, 13486–13496. doi: 10.1109/ACCESS.2019.2894225

Xie, G., Wu, C., Sun, Y., Fan, Z., and Liu, J. (2019). Lpi-ibnra: Long non-coding rna- protein interaction prediction based on improved bipartite network recommender algorithm. *Front. Genet.* 10:343. doi: 10.3389/fgene.2019.00343

Zhang, T., Cao, C., Wu, D., and Liu, L. (2016). *SNHG3* correlates with malignant status and poor prognosis in hepatocellular carcinoma. *Tumor Biol.* 37, 2379–2385. doi: 10.1007/s13277-015-4052-4

Zhang, T., Wang, M., Xi, J., and Li, A. (2018). Lpgnmf: Predicting long non-coding rna and protein interaction using graph regularized nonnegative matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform* 17, 189–197.

Zhang, W., Qu, Q., Zhang, Y., and Wang, W. (2018a). The linear neighborhood propagation method for predicting long non-coding rna-protein interactions. *Neurocomputing* 273, 526–534. doi: 10.1016/j.jpdc.2017.08.009

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

Zhao, Q., Yu, H., Ming, Z., Hu, H., Ren, G., and Liu, H. (2018a). The bipartite network projection-recommended algorithm for predicting long non-coding rna-protein interactions. *Mol. Ther. Nucleic Acids* 13, 464–471.

Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018b). Irwnrlpi: integrating random walk and neighborhood regularized logistic matrix factorization for lncrna-protein interaction prediction. *Front. Genet.* 9:239. doi: 10.3389/fgene.2018.00239

Zhou, Y. K., Hu, J., Shen, Z. A., Zhang, W. Y., and Du, P. F. (2020). LPI-SKF: predicting lncRNA-protein interactions using similarity kernel fusions. *Front. Genet.* 11:615144. doi: 10.3389/fgene.2020.615144

Conflict of Interest: JL and CX were employed by the company Geneis Beijing Co., Ltd.

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

frontiers in Genetics

METHODS
published: 10 August 2021
doi: 10.3389/fgene.2021.720327

Biased Random Walk With Restart on Multilayer Heterogeneous Networks for MiRNA–Disease Association Prediction

Jia Qu[1], Chun-Chun Wang[2], Shu-Bin Cai[3], Wen-Di Zhao[1], Xiao-Long Cheng[1] and Zhong Ming[3]**

[1] School of Computer Science and Artificial Intelligence & Aliyun School of Big Data, Changzhou University, Changzhou, China, [2] Information and Control Engineering, China University of Mining and Technology, Xuzhou, China, [3] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

Numerous experiments have proved that microRNAs (miRNAs) could be used as diagnostic biomarkers for many complex diseases. Thus, it is conceivable that predicting the unobserved associations between miRNAs and diseases is extremely significant for the medical field. Here, based on heterogeneous networks built on the information of known miRNA–disease associations, miRNA function similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases, we developed a computing model of biased random walk with restart on multilayer heterogeneous networks for miRNA–disease association prediction (BRWRMHMDA) through enforcing degree-based biased random walk with restart (BRWR). Assessment results reflected that an AUC of 0.8310 was gained in local leave-one-out cross-validation (LOOCV), which proved the calculation algorithm's good performance. Besides, we carried out BRWRMHMDA to prioritize candidate miRNAs for esophageal neoplasms based on HMDD v2.0. We further prioritize candidate miRNAs for breast neoplasms based on HMDD v1.0. The local LOOCV results and performance analysis of the case study all showed that the proposed model has good and stable performance.

Keywords: microRNA, disease, association prediction, degree, biased random walk with restart

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Wen Zhang,
Huazhong Agricultural University, China
Yi Xiong,
Shanghai Jiao Tong University, China

****Correspondence:***
Jia Qu
TB17060015B4@cumt.edu.cn
Zhong Ming
mingz@szu.edu.cn

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

Received: *04 June 2021*
Accepted: *13 July 2021*
Published: *10 August 2021*

Citation:
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

INTRODUCTION

MicroRNA (miRNA) is a noncoding single-stranded RNA with a length of about 22 nucleotides and pervasive in both animals and plants (Axtell et al., 2011). MiRNAs play their regulator role through binding to imperfect complementary sites within the 3′ untranslated regions (UTRs) of their messenger RNA (mRNA) targets (Reinhart et al., 2000; Ambros, 2004; Bartel, 2009). Nowadays, a large number of experimental studies have proved that miRNAs regulate multiple biological activities and per miRNA can regulate hundreds of gene targets (Lee et al., 1993; Pasquinelli and Ruvkun, 2002; Brennecke et al., 2003; Lin et al., 2003; Cheng et al., 2005; Karp and Ambros, 2005; Miska, 2005; Pillai et al., 2005; Cui et al., 2006; Lu et al., 2008; Bartel, 2009; Alshalalfa and Alhajj, 2013). Moreover, miRNAs have potential influences on almost all genetic pathways, and the upregulation and downregulation of miRNA expression in the human body are correlated to various complex diseases (Liu et al., 2008). It indicates that miRNAs have close

associations with many complex diseases, and miRNAs may be used as a tumor suppressor gene to treat cancer in clinical medicine (Cheng et al., 2005). For example, the abnormal expression of miR-21 could be conducive to the growth and spread of human hepatocellular cancer (HCC) via the regulation of phosphatase and tensin homolog (PTEN) expression and PTEN-dependent pathways (Meng et al., 2007). MiR-10b is expressed in metastatic breast cancer cells highly and has a positive regulatory effect on cell migration and invasion (Ma et al., 2007). Research further suggested that the overexpression of miR-17-92 in lung cancer could enhance cell proliferation (Hayashita et al., 2005). Moreover, the miRNA family of let-7 was reported to downregulate in lung cancers and regulate an oncogene of RAS, so the inhibition of let-7 may help in the treatment of the cancer (Johnson et al., 2005). Also, through targeting an antiapoptotic factor of B-cell lymphoma-2 (BCL2), miR-15 and miRNA-16 were proved to downregulate in chronic lymphocytic leukemias and induce apoptosis (Cimmino et al., 2005). Certainly, identification of potential miRNA–disease associations has become a very significant research goal in the field of biomedical research. Predicting potential miRNAs related to diseases would promote people's understanding of the pathogenesis of diseases at the molecular level and benefit for the diagnosis, treatment, and prevention of diseases. Recently, some reliable databases have been developed to store experimental verified miRNA–disease associations, such as HMDD v2.0 (Li et al., 2014), miR2Disease (Jiang et al., 2009), and dbDEMC (Yang et al., 2010). Using traditional experiment approach to identify potential miRNA-disease associations is usually complex, time consuming and expensive. It is an urgent need for scholars to develop calculation models to predict new miRNA–disease associations. We expect that miRNA–disease pairs with high scores could be selected for experimental verification, which would significantly reduce the time and cost of biological experiments.

Great progress has been made in developing calculation models for the potential miRNA–disease association prediction in recent years. These prediction models are usually proposed by the consideration of complex network-based or machine learning-based methods (Chen et al., 2019a). For the experimentally confirmed miRNA–disease associations that have been collected, a lot of calculation models were put forward for the identification of new miRNA–disease associations on the basis of the hypothesis that functionally similar miRNAs are often associated with phenotypically similar diseases (Perez-Iratxeta et al., 2002; Aerts et al., 2006). In 2013, human disease-related miRNA prediction (HDMP), an effective prediction algorithm based on weighted *k* most similar neighbors, was proposed by Xuan et al. (2015). In the model, functional similarity between each miRNA pair was calculated by combining the information of their related disease terms and disease phenotype similarity. Then the possibility of unobserved miRNA–disease pairs was predicted via the sum of subscores of miRNA's *k* neighbor. The subscore for a neighbor of a miRNA can be calculated based on the weight of the neighbor and the functional similarity between the neighbor and the miRNA. In 2014, based on known miRNA–disease associations, disease similarity, and miRNA similarity, a global method of regularized least squares for miRNA–disease association (RLSMDA) was introduced by Chen and Yan (2014) to uncover novel associations between miRNAs and diseases under the framework of a semisupervised classifier. In 2015, based on the constructed miRNA functional network, another new model of miRNAs associated with diseases prediction (MIDP) was developed by Xuan et al. (2015) to prioritize candidate miRNAs for investigated diseases with known related miRNAs. In the model, for the marked nodes and unmarked nodes, transition matrices are different, and the transition weight of marked nodes was higher than that of unmarked nodes. Moreover, due to the fact that MIDP could not predict potential miRNAs (diseases) associated with new diseases (miRNAs) without any known related miRNAs (diseases), an extension approach of MIDPE was also proposed to predict potential miRNAs (diseases) associated with new diseases (miRNAs). Chen et al. (2017) published a model of ranking-based KNN for miRNA–disease association prediction (RKNNMDA), in which the KNN approach was employed to gain the *k*-nearest-neighbors of each miRNA and disease according to the collected data information. Then, based on the Hamming loss of per disease pair and miRNA pair, a support vector machine (SVM) ranking model was introduced to achieve scores of potential miRNA–disease associations. Furthermore, Chen and Huang (2017) presented a computational model named Laplacian regularized sparse subspace learning for miRNA–disease association prediction (LRSSLMDA), which projected miRNAs' feature and diseases' feature into a common subspace. Then, the local structures of the training data were obtained based on Laplacian regularization, and the final predicted scores would be obtained by carrying out the L1-norm constraint. Chen et al. (2018a) put forward a machine learning-based method of extreme gradient boosting machine for miRNA–disease association prediction (EGBMMDA), in which a feature vector for the miRNA–disease pair was established by merging three matrices of miRNA functional similarity, disease semantic similarity, and known miRNA–disease associations. Then, based on the characteristics and the gradient boosting framework, a regression tree was applied to obtained scores of potential miRNA–disease associations. In the same year, a computational model of ensemble learning and link prediction for miRNA–disease association prediction (ELLPMDA) was brought forward by Chen et al. (2018f); they inferred new miRNA–disease associations via the weight-based integration of three classified results gained from common neighbors, Jaccard index and Katz index. Also, from the angle of reducing the noise of the original collected known miRNA–disease association information, Chen et al. (2018e) further brought up a calculation method of matrix decomposition and heterogeneous graph inference for miRNA–disease association prediction (MDHGI). The sparse learning method was carried out firstly on the initial association information to reduce noise. Then, an iterative formula for propagating miRNA and disease information was established based on the built heterogeneous network to predict potential miRNA–disease associations. Besides, Chen et al. (2018c) presented a novel method of inductive matrix completion for miRNA–disease

association prediction (IMCMDA) through enforcing a low-rank inductive matrix completion approach on the collected datasets. Chen et al. (2018d) also developed a prediction model of bipartite network projection for miRNA–disease association prediction (BNPMDA). In the model, the bias ratings for miRNAs and diseases were built based on agglomerative hierarchical clustering. Then, through assigning transfer weights to resource allocation links between miRNAs and diseases according to the bias ratings, the bipartite network recommendation algorithm was implemented to predict the potential miRNA–disease associations. Chen et al. (2019b) put forward a machine learning-based method named ensemble of decision tree-based miRNA–disease association prediction (EDTMDA), which identifies potential disease–miRNA association by implementing ensemble learning based on decision trees (DTs) and dimensionality reduction based on principal component analysis (PCA). In recent years, Chen et al. (2021) further proposed the neighborhood constraint matrix completion for miRNA–disease association prediction (NCMCMDA), which combined the neighborhood constraint with matrix completion. The prediction problem in NCMCMDA can be transformed into an optimization problem, and a fast iterative shrinkage–thresholding algorithm was implemented to solve it.

Some scholars have also introduced some calculation models on the basis of various types of association networks, rather than limited to the miRNA–disease network. In 2014, through the analysis of miRNA–protein associations and protein–disease associations, Mork et al. (2014) developed a scoring scheme for the potential miRNA–disease association prediction. In 2016, through taking advantage of miRNA–disease associations, miRNA–neighbor associations, miRNA–target associations, miRNA–word associations, and miRNA–family associations, Pasquier and Gardes (2016) expressed the distribution information of miRNAs and diseases in a high-dimensional vector space and then inferred association scores between miRNAs and diseases according to their vector similarity. In 2017, based on the phenome–miRNA network constructed by known miRNA–disease associations, miRNA functional similarity, disease semantic similarity, and phenotypic similarity, a combinatorial prioritization algorithm was proposed by Yu et al. (2017) to predict potential miRNA–disease associations. In 2018, through constructing a three-layer heterogeneous network based on the integration of known miRNA–lncRNA interactions, miRNA–disease associations, miRNA similarity, disease similarity, and lncRNA similarity, Chen et al. (2018b) designed a method of triple-layer heterogeneous network-based inference for miRNA–disease association prediction (TLHNMDA) by establishing two information spreading iterative formulas.

In this manuscript, based on a multilayer heterogeneous network established by known miRNA–disease associations, disease semantic similarity, miRNA functional similarity, and Gaussian interaction profile kernel similarity for diseases and miRNAs, we put forward a calculating model of biased random walk with restart on multilayer heterogeneous networks for miRNA–disease association prediction (BRWRMHMDA). In the model, degree-based biased random walk with restart (BRWR) was implemented to predict potential miRNA–disease associations on the basis of the constructed multilayer heterogeneous network. For evaluating the property of the introduced calculation model, local leave-one-out cross-validation (LOOCV) was presented and the outcome showed that BRWRMHMDA possesses an AUC of 0.8310 in local LOOCV. In the case study, we not only employed BRWRMHMDA to infer candidate miRNAs for esophageal neoplasms in the light of known miRNA–disease associations extracted from HMDD v2.0 (Li et al., 2014) but also implemented the model to predict breast neoplasms-associated miRNAs on the basis of known miRNA–disease associations collected from HMDAD v1.0. From the result of LOOCV and the case study, we can be sure that BRWRMHMDA has better prediction ability, and BRWRMHMDA can be used to predict potential miRNA–disease associations.

MATERIALS AND METHODS

Human miRNA–Disease Association

The dataset of 5,430 experimentally verified associations between 383 diseases and 495 miRNAs came from the HMDD v2.0 database (Li et al., 2014). We used the variables *nm* and *nd* to refer to the number of diseases and miRNAs in the dataset, respectively. Afterward, an adjacency matrix A was established to indicate known miRNA–disease associations. If miRNA $m(i)$ is related to $d(j)$, the value of entity $A(i, j)$ would equal to 1, otherwise 0.

$$A(i,j) = \begin{cases} 1, \textit{ if miRNA } m(j) \textit{ is related to disease } d(i) \\ 0, \qquad\qquad\qquad\qquad \textit{otherwise} \end{cases} \tag{1}$$

MiRNA Functional Similarity

Since functionally similar miRNAs are more likely to be associated with phenotypically similar diseases on the basis of the previous study (Wang et al., 2010), we got the information of miRNA functional similarity from http://www.cuilab.cn/files/images/cuilab/misim.zip. After that, we constructed a miRNA functional similarity matrix *FS* with the row and column of *nm*. It is remarkable that the value of entity $FS(i, j)$ refers to the similarity score between miRNA $m(i)$ and miRNA $m(j)$.

Disease Semantic Similarity Model 1

Each disease can be described as a directed acyclic graph (DAG) according to previous literature (Wang et al., 2010). For example, disease D can be described as $DAG = (D, T(D), E(D))$, where $T(D)$ refers to all disease nodes, and $E(D)$ indicates all edges that connect disease nodes based on $DAG(D)$. Inspired by previous work (Xuan et al., 2013), the contribution value of disease d in $DAG(D)$ to the semantic value of disease D can be defined as follows:

$$\begin{cases} D_D1(d) = 1 & \textit{if } d = D \\ D_D1(d) = \max\left\{\Delta^* D_D1(d') | d' \in \textit{children of } d\right\} & \textit{if } d \neq D \end{cases} \tag{2}$$

where Δ is the semantic contribution decay factor, and the semantic value of disease D can be described as follows:

$$DV1(D) = \sum_{d \in T(D)} D_D1(d) \quad (3)$$

Considering that two diseases would have greater similarity if they share larger part of their *DAGs*, we defined the semantic similarity between disease *d(i)* and *d(j)* in disease semantic similarity model 1 as follows:

$$SS1(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D_{d(i)}1(t) + D_{d(j)}1(t))}{DV1(d(i)) + DV1(d(j))} \quad (4)$$

Disease Semantic Similarity Model 2

Also inspired by previous work (Xuan et al., 2013), we also introduced disease semantic similarity model 2. For two diseases in the same layer of $DAG(D)$, if the first disease occurs more frequently in $DAG(D)$ than the second disease, the second disease would be regarded to be more specific to disease D. By the consideration of the idea that the contribution of different disease terms in the same layer of $DAG(D)$ may be the difference, the contribution of disease d in $DAG(D)$ to the semantic value of disease D could be described as follows:

$$D_D2(d) = -\log[\frac{the\ number\ of\ DAGs\ including\ d}{the\ number\ of\ disease}] \quad (5)$$

The value of semantic similarity in disease semantic similarity model 2 between disease $d(i)$ and $d(j)$ could be calculated as follows:

$$SS2(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D_{d(i)}2(t) + D_{d(j)}2(t))}{DV2(d(i)) + DV2(d(j))} \quad (6)$$

where

$$DV2(D) = \sum_{d \in T(D)} D_D2(d) \quad (7)$$

Gaussian Interaction Profile Kernel Similarity

The calculation of Gaussian interaction profile kernel similarity for diseases and miRNAs depends on the topologic information of known miRNA–disease associations (van Laarhoven et al., 2011). For diseases, we used a binary vector $IP(d(u))$ (i.e., the uth row of the adjacency matrix A) to indicate the interaction profiles of disease $d(u)$. Accordingly, the Gaussian interaction profile kernel similarity between diseases $d(u)$ and $d(v)$ can be described.

$$KD(d(u), d(v)) = \exp(-\gamma_d \left|\left|IP(d(u)) - IP(d(v))\right|\right|^2) \quad (8)$$

The parameter γ_d was used to regulate the kernel bandwidth and could be acquired via the normalization of a new bandwidth $\gamma_d^{'}$ by the average number of associated miRNAs for each disease.

$$\gamma_d = \gamma_d^{'} / \left(\frac{1}{nd} \sum_{u=1}^{nd} \left|\left|IP(d(u))\right|\right|^2\right) \quad (9)$$

For miRNAs, the binary vector $IP(m(i))$ (i.e., the ith column of the adjacency matrix A) was introduced to indicate the interaction profiles of miRNA $m(i)$. At last, the Gaussian interaction profile kernel similarity between miRNA $m(i)$ and $m(j)$ can be constructed as follows:

$$KM(m(i), m(j)) = \exp(-\gamma_m \left|\left|IP(m(i)) - IP(m(j))\right|\right|^2) \quad (10)$$

$$\gamma_m = \gamma_m^{'} / \left(\frac{1}{nm} \sum_{i=1}^{nm} ||IP(m(i))||^2\right) \quad (11)$$

Integrated Similarity for miRNAs and Diseases

Based on past work (Chen et al., 2016), integrated similarity for a pair of diseases ($d(u)$ and $d(v)$) can be defined via the combination of disease semantic similarity and Gaussian interaction profile kernel similarity for diseases. The formula of integrated similarity for diseases is displayed as follows:

$$SD(d(u), d(v)) = \begin{cases} \frac{SS1(d(u) + d(v)) + SS2(d(u), d(v))}{2} & d(u)\ and\ d(v)\ has\ semantic\ similarity \\ KD(m(u), m(v)) & otherwise \end{cases} \quad (12)$$

Also, the integrated similarity for a pair of miRNAs ($m(i)$ and $m(j)$) could be formed by taking miRNA functional similarity with Gaussian interaction profile kernel similarity for miRNA into account (Chen et al., 2016).

$$SM(m(i), m(j)) = \begin{cases} FS(m(i), m(j)) & m(i)\ and\ m(j)\ has\ functional\ similarity \\ KM(m(i), m(j)) & otherwise \end{cases} \quad (13)$$

BRWRMHMDA

Via the integration of known miRNA–disease associations, disease semantic similarity, miRNA functional similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases, we put forward a calculating model of BRWRMHMDA based on the degree for the identification of potential miRNA–disease associations by enforcing BRWR on a constructed multilayer heterogeneous network according to previous work (Bonaventura et al., 2014) (see **Figure 1**).

In the model, based on the constructed multisource dataset, we used W_{dd}, W_{mm}, W_{dm} to represent the initial matrix of integrated disease similarity, integrated miRNA similarity, and known miRNA–disease associations, respectively. Then, the multilayer heterogeneous network was constructed and described as $W = \begin{bmatrix} W_{dd} & W_{dm} \\ W_{md} & W_{mm} \end{bmatrix}$. In BRWR, if we predicted potential miRNAs for disease $d(i)$, the disease $d(i)$ is the seed node in the disease network. If the miRNA $m(j)$ is associated with disease $d(i)$, miRNA $m(j)$ is the seed node for disease $d(i)$ in the miRNA network. If the miRNA $m(j)$ has no known association with disease $d(i)$, miRNA $m(j)$ is the candidate node for disease $d(i)$ in

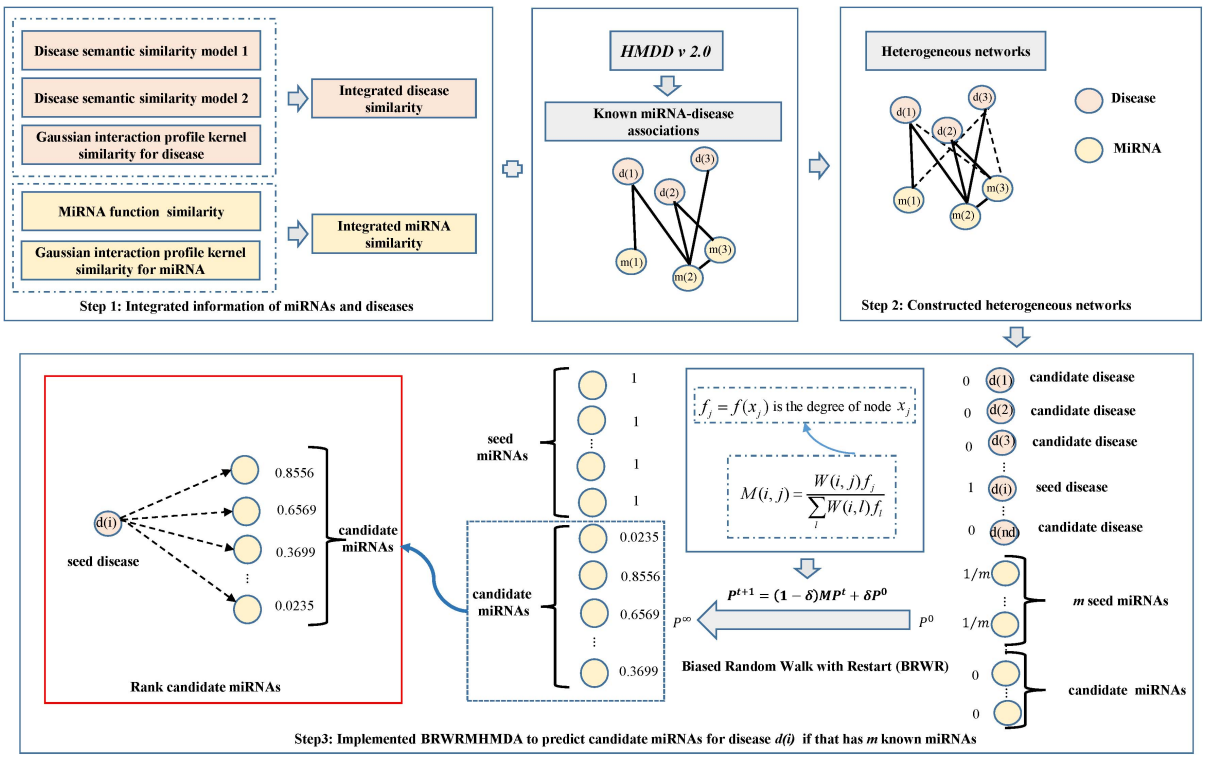


FIGURE 1 | Comparing to the other calculating algorithms (ELLPMDA, IMCMDA, EGBMMDA, MDHGI, TLHNMDA, MaxFlow, RLSMDA, HDMP, WBSMDA, MirAI, and MIDP) in terms of AUCs, BRWRMHMDA gained a better AUC value of 0.8310. It indicates that the proposed model is more suitable for the miRNA–disease association prediction.

the miRNA network. For predicting potential miRNAs for disease $d(i)$, the original probability vector v_0 of the miRNA network is computed through assigning equal probability to the seed nodes in the miRNA network with a total equal to 1. In the disease network, the probability value of 1 was assigned to $d(i)$, and the probability value of 0 was assigned to other diseases to form u_0, where the initial seed node probability $P^0 = \begin{bmatrix} \alpha^* u_0 \\ (1-\alpha)^* v_0 \end{bmatrix}$; α and $(1-\alpha)$ refer to the weight of the disease network and the miRNA network, respectively.

Seed nodes at each step move to their immediate neighbors with a probability $(1-\delta)$ or return to the seed nodes with a restart probability δ ($\delta \in (0,1)$). P^0 was the initial probability vector, and P^{t+1} was a probability vector of node at time $t+1$, which could be defined as follows:

$$P^{t+1} = (1-\delta)MP^t + \delta P^0 \tag{14}$$

where matrix $M = \begin{bmatrix} M_{dd} & M_{dm} \\ M_{md} & M_{mm} \end{bmatrix}$ is the transition matrix of our established network. In random walk with restart (RWR), the transition probability $M(i,j)$ of a walker from node i to node j can be described as follows:

$$M(i,j) = \frac{W(i,j)}{\sum_l W(i,l)} \tag{15}$$

where $W(i,j)$ is the similarity between node i and node j. In this model, BRWR of degree biased random walk was proposed to identify potential miRNA–disease associations. Biases were usually considered to be related to graph topological properties. For example, a walk at node x_i selects it neighbors of x_j with a probability $f_j = f(x_j)$ relying on the node property x_j. Usually, the node property can be described as a function of the vertex properties (the network degree, closeness centrality, etc.) or the edge properties (multiplicity or shortest path), or the combination of them (Gomez-Gardenes and Latora, 2008). There are other related bias choice of maximal entropy (Burda et al., 2009). Thus, the transition probability of a walker from i to j in BRWR can be defined as

$$M(i,j) = \frac{W(i,j)f_j}{\sum_l W(i,l)f_l} \tag{16}$$

Therefore, in the disease similarity network, the transition probability from vertex d_i to d_j can be defined as

$$M_{dd}(i,j) = p(d_j|d_i) = \begin{cases} W_{dd}(i,j)f_j / \sum_j W_{dd}(i,j)f_j & if \sum_j W_{dm}(i,j) = 0 \\ (1-\lambda) W_{dd}(i,j)f_j / \sum_j W_{dd}(i,j)f_j & otherwise \end{cases} \tag{17}$$

In the miRNA similarity network, the transition probability from m_i tom_j can be defined as

$$M_{mm}(i,j) = p(m_j|m_i) = \begin{cases} W_{mm}(i,j)f_j / \sum_j W_{mm}(i,j)f_j & if \sum_j W_{dm}(j,i) = 0 \\ (1-\lambda) W_{mm}(i,j)f_j / \sum_j W_{mm}(i,j)f_j & otherwise \end{cases} \quad (18)$$

In the miRNA–disease association network, the transition probability from vertex d_i to m_j can be defined as

$$M_{dm}(i,j) = p(m_j|d_i) = \begin{cases} \lambda W_{dm}(i,j)f_j / \sum_j W_{dm}(i,j)f_j & if \sum_j W_{dm}(i,j) \neq 0 \\ 0 & otherwise \end{cases} \quad (19)$$

The transition probability from vertex m_i to d_j can be defined as

$$M_{md}(i,j) = p(d_j|m_i) = \begin{cases} \lambda W_{dm}(j,i)f_j / \sum_j W_{dm}(j,i)f_j & if \sum_j W_{dm}(j,i) \neq 0 \\ 0 & otherwise \end{cases} \quad (20)$$

In this paper, we focus on the case of BRWR by considering the degree nodes. Therefore, $f_j = f(x_j)$ in the model is the degree of node x_j in the transition probability. The degree f_i of a disease node i is defined by computing the number of edges involved in the disease node. Therefore, in the disease similarity network, the degree of disease node j can be defined as $f_j = \sum_i W_{dd}(i,j)$. In the miRNA similarity network, the degree of miRNA node j can be defined as $f_j = \sum_i W_{mm}(i,j)$. In the transition probability matrix of the miRNA–disease association network, the degree of miRNA m_j can be described as $f_j = \sum_i W_{dm}(i,j)$. Also, in the transition probability matrix of the miRNA–disease association network, the degree of disease d_j can be described as $f_j = \sum_i W_{dm}(j,i)$. Therefore, based on BRWR of degree nodes, the potential miRNA–disease associations would be obtained.

RESULTS

Performance Evaluation

Since BRWR is a local calculating method, it cannot infer candidate miRNAs for all diseases simultaneously. Therefore, in order to analyze the performance of BRWRMHMDA, the proposed method has been extensively compared with some classic algorithms (ELLPMDA, IMCMDA, EGBMMDA, MDHGI, TLHNMDA, MaxFlow, RLSMDA, HDMP, WBSMDA, MirAI, and MIDP) based on the 5,430 known miRNA–disease associations from the HMDD v2.0 database (Li et al., 2014) via local LOOCV. In local LOOCV, each known miRNA–disease association was considered as a test sample in turn, and the rest of 5,429 known associations were treated as training samples. After enforcing BRWRMHMDA, the score of the test sample would be sorted with the scores of all unobserved pairs between miRNAs and the investigated disease. The proposed approach would be regarded as reliable if the test sample's ranking is higher than a set threshold. Then a receiver operating characteristics (ROC) curve with the true positive rate (TPR, sensitivity) versus the false positive rate (FPR, 1-specificity) at various thresholds would be drawn. Sensitivity refers to the percentage of test samples ranked higher than the given threshold, and specificity refers to the percentage of candidates ranked lower than the threshold. Finally, the area under the ROC curve (AUC) was calculated to accurately evaluate the prediction ability of BRWRMHMDA. The value of the AUC is between 0 and 1, and the higher the value of the AUC, the better the prediction performance of the algorithm. If the value of the AUC is 0.5, the prediction performance of BRWRMHMDA is random. The final assessment results showed that BRWRMHMDA has better prediction performance with an AUC of 0.8310 than those of the other server classical algorithms of ELLPMDA (0.8181), IMCMDA (0.8034), EGBMMDA (0.8221), MDHGI (0.8240), TLHNMDA (0.7756), MaxFlow (0.7774), RLSMDA (0.6953), HDMP (0.7702), WBSMDA (0.8031), MirAI (0.6299), and MIDP (0.8196) (see **Figure 2**). Here, the AUC value of MirAI is lower than that reported in its literature (Pasquier and Gardes, 2016) because MirAI was proposed on the basis of a collaborative filtering algorithm affected by the data sparsity problem. Compared with the training set in the original literature, our dataset is relatively scarce. The training set in the original literature contains 83 diseases and at least 20 known related miRNAs for each disease, while our training set contains 383 diseases and most diseases-related miRNAs are rare.

Case Studies

In order to further analyze the performance of the algorithm effectively, we carried out two types of case studies. The first type of case studies is the prediction of potential miRNAs associated with esophageal neoplasms based on the known miRNA–disease association collect from HMDD v2.0. The second type of case studies is the prediction of potential miRNAs associated with breast neoplasms based on the known miRNA–disease association collect from HMDD v1.0.

Esophageal neoplasm is one of the most lethal cancers in the world; its main nature is highly invasive and of low survival rate (Domper Arnal et al., 2015). The disease contains two main histological types of squamous cell cancer and adenocarcinoma (Zhang et al., 2016). Malnutrition is a main risk factor for esophageal squamous cell carcinoma (ESCC), and obesity is the main risk factor for esophageal adenocarcinoma (Domper Arnal et al., 2015). Accordingly, looking for sensitive molecular biomarkers and individual treatment approach for early diagnosis of esophageal cancer has become the main clinical and basic research direction. Numerous studies suggested that miRNAs play an important role in diseases and can be a biomarker for esophageal neoplasms' treatment. For example, miR-506 was abnormally expressed in a variety of tumors and could be used as a prognostic biomarker for ESCC (Li et al., 2016). Besides, plasma miR-718 was reported to downregulate in ESCC patients and might be treated as a potential diagnostic marker for the disease (Sun et al., 2016). Here, we employed BRWRMHMDA to prioritize candidate miRNAs for esophageal neoplasms according to the dataset of 5,430 known miRNA–disease associations between 383 diseases and 495 miRNAs. As a result, of the first 50 miRNAs predicted

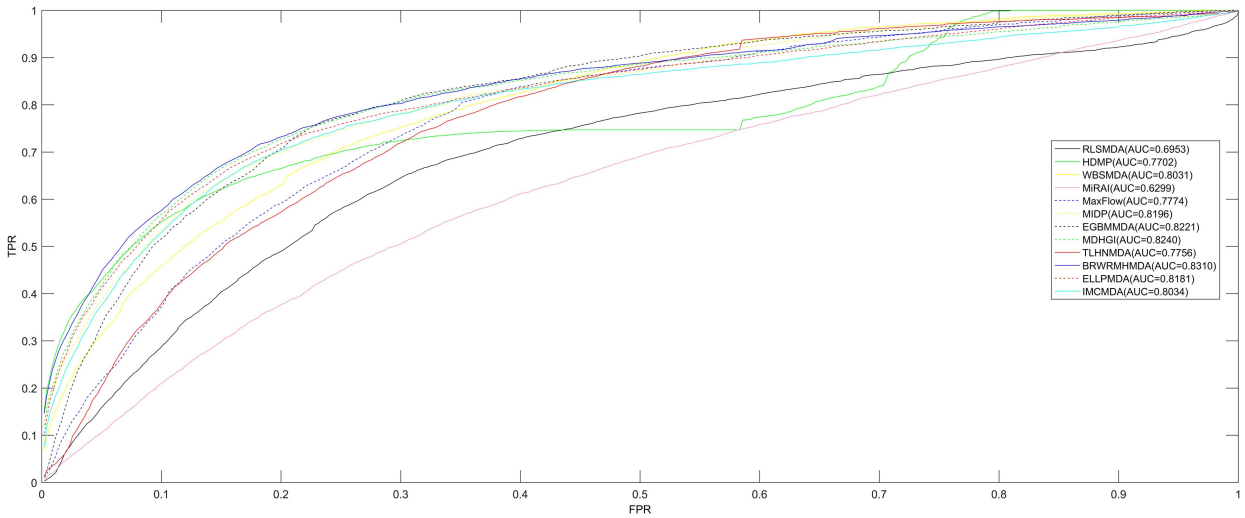


FIGURE 2 | Flowchart of BRWRMHMDA to prioritize candidate miRNAs for diseases. Through employing BRWR on the established heterogeneous networks, final scores p^{∞} of candidate miRNAs predicted for each disease would be gotten after some steps.

for esophageal neoplasms in the ranking, 49 miRNAs have been confirmed by the database of dbDEMC and miR2Disease (see **Table 1**). For example, the predicted association score between hsa-mir-125b and esophageal neoplasms is ranked first. Yu et al. (2020) have found that hsa-mir-125b suppresses cell proliferation and metastasis by targeting HAX-1 in ESCC, which proves that hsa-mir-125b is related to esophageal neoplasms. Moreover, the predicted association score between hsa-mir-200b and esophageal neoplasms is ranked second. Researchers have confirmed that hsa-mir-200b is downregulated in ESCC in the comparison of the respective adjacent benign tissues (Zhang et al., 2014). Therefore, hsa-mir-200b is associated with esophageal neoplasms.

Breast neoplasm is one of the three most common cancers for women (Siegel et al., 2018). In particular, metastatic breast cancer (MBC) is usually incurable, and about 5% of patients have metastases at diagnosis (Torre et al., 2015). With recent research, miR-10b sponge has been shown to effectively inhibit the growth of MDA-MB-231 and MCF-7 cells in breast cancer (Liang et al., 2016). In addition, miR-223 was demonstrated to function as a potential tumor marker for breast neoplasm through suppressing its protein expression of FOXO1 (Wei et al., 2017). Accordingly, identifying breast neoplasm-related miRNAs is meaningful, which could help the medical diagnosis and treatment for MBC (McGuire et al., 2015). Here, we enforced BRWRMHMDA to infer potential miRNAs related to breast neoplasms on the basis of 1,395 known miRNA–disease associations between 137 diseases and 271 miRNAs collected from HMDD v1.0. The results showed that 48 of the first 50 miRNAs predicted for breast neoplasms have been confirmed by the databases of dbDEMC, miR2Disease, and HMDD v2.0 (see **Table 2**). For example, hsa-let-7b was predicted to associate with breast neoplasms, and the predicted score is ranked second. It is worth noting that hsa-let-7b can significantly change oncogenic signaling in breast cancer cells. Consequently, hsa-let-7b may have important roles in breast neoplasm progression and can be considered as potential targets for breast neoplasm therapy and diagnosis (Bozgeyik, 2020). Besides, hsa-mir-16 was predicted to be related to breast neoplasms, and the predicted score is ranked third. Haghi et al. indicated that has-mir-16 and has-mir-34a can collaborate in breast tumor suppression, which proved that hsa-mir-16 has association with breast neoplasms.

At last, we have released the whole prediction results via the implementation of BRWRMHMDA for all miRNA–disease pairs between 383 diseases and 495 miRNAs from HMDD v2.0 (see **Supplementary Table 1**).

DISCUSSION

Through integrating known miRNA–disease associations, disease semantic similarity, miRNA function similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases, BRWRMHMDA was employed in this manuscript to prioritize candidate miRNAs for diseases via the implementation of degree-based BRWR on the established networks. The assessment results of LOOCV showed that the developed algorithm outperforms the other 11 classic prediction algorithms in accuracy. We further enforced the proposed algorithm to infer candidate miRNAs for esophageal neoplasms in the light of known miRNA–disease associations extracted from HMDD v2.0 and infer candidate miRNAs for breast neoplasms in the light of known miRNA–disease associations extracted from HMDD v1.0. The results of the case study fully demonstrated the stability of this introduced algorithm. It is worth mentioning that our research group will keep on studying this issue in depth. Furthermore, we hope more external research groups would select potential associations with high prediction scores and verify them based on biological experiment in the future.

TABLE 1 | The implementation of BRWRMHMDA to prioritize candidate miRNAs for esophageal neoplasms based on experimentally confined miRNA–disease associations collected from HMDD v2.0 and 47 of the first 50 predicted miRNAs were confirmed.

miRNA	Evidence	miRNA	Evidence
hsa-mir-125b	dbDEMC	hsa-mir-429	dbDEMC
hsa-mir-200b	dbDEMC	hsa-mir-106a	dbDEMC
hsa-mir-18a	dbDEMC	hsa-mir-24	dbDEMC
hsa-mir-17	dbDEMC	hsa-mir-30c	dbDEMC
hsa-mir-221	dbDEMC	hsa-mir-218	unconfirmed
hsa-mir-19b	dbDEMC	hsa-mir-93	dbDEMC
hsa-mir-16	dbDEMC	hsa-mir-132	dbDEMC
hsa-mir-1	dbDEMC	hsa-mir-30a	dbDEMC
hsa-mir-222	dbDEMC	hsa-mir-127	dbDEMC
hsa-let-7i	dbDEMC	hsa-mir-195	dbDEMC
hsa-mir-29a	dbDEMC	hsa-mir-199b	dbDEMC
hsa-let-7e	dbDEMC	hsa-mir-10b	dbDEMC
hsa-let-7d	dbDEMC	hsa-mir-15b	dbDEMC
hsa-mir-29b	dbDEMC	hsa-mir-107	dbdemc and miR2Disease
hsa-let-7f	unconfirmed	hsa-mir-7	dbDEMC
hsa-mir-181b	dbDEMC	hsa-mir-224	dbDEMC
hsa-mir-181a	dbDEMC	hsa-mir-18b	dbDEMC
hsa-mir-125a	dbDEMC	hsa-mir-133b	dbDEMC
hsa-let-7g	dbDEMC	hsa-mir-335	dbDEMC
hsa-mir-9	dbDEMC	hsa-mir-194	dbdemc and miR2Disease
hsa-mir-146b	dbDEMC	hsa-mir-302b	dbDEMC
hsa-mir-106b	dbDEMC	hsa-mir-20b	dbDEMC
hsa-mir-182	dbDEMC	hsa-mir-124	dbDEMC
hsa-mir-142	dbDEMC	hsa-mir-373	dbdemc and miR2Disease
hsa-mir-122	unconfirmed	hsa-mir-191	dbDEMC

TABLE 2 | The implementation of BRWRMHMDA to prioritize candidate miRNAs for breast neoplasms based on experimentally confined miRNA–disease associations collected from HMDD v1.0 and 48 of the first 50 predicted miRNAs were confirmed.

miRNA	Evidence	miRNA	Evidence
hsa-let-7i	dbDEMC and miR2Disease and HMDD	hsa-mir-203	dbDEMC and miR2Disease and HMDD
hsa-let-7b	dbDEMC and HMDD	hsa-mir-32	dbDEMC
hsa-mir-16	dbDEMC and HMDD	hsa-mir-30e	unconfirmed
hsa-let-7e	dbDEMC and HMDD	hsa-mir-532	dbDEMC
hsa-let-7g	dbDEMC and HMDD	hsa-mir-335	dbDEMC and miR2Disease and HMDD
hsa-let-7c	dbDEMC and HMDD	hsa-mir-150	dbDEMC
hsa-mir-92a	HMDD	hsa-mir-199b	dbDEMC and HMDD
hsa-mir-126	dbDEMC and miR2Disease and HMDD	hsa-mir-99a	dbDEMC
hsa-mir-223	dbDEMC and HMDD	hsa-mir-98	dbDEMC and miR2Disease
hsa-mir-92b	dbDEMC	hsa-mir-142	unconfirmed
hsa-mir-373	dbDEMC and miR2Disease and HMDD	hsa-mir-128b	miR2Disease
hsa-mir-101	dbDEMC and miR2Disease and HMDD	hsa-mir-107	dbDEMC and HMDD
hsa-mir-191	dbDEMC and miR2Disease and HMDD	hsa-mir-224	dbDEMC and HMDD
hsa-mir-182	dbDEMC and miR2Disease and HMDD	hsa-mir-27a	dbDEMC and miR2Disease and HMDD
hsa-mir-99b	dbDEMC	hsa-mir-195	dbDEMC and miR2Disease and HMDD
hsa-mir-106a	dbDEMC	hsa-mir-124	dbDEMC and HMDD
hsa-mir-181a	dbDEMC and miR2Disease and HMDD	hsa-mir-30a	miR2Disease and HMDD
hsa-mir-29c	dbDEMC and miR2Disease and HMDD	hsa-mir-520b	dbDEMC and HMDD
hsa-mir-100	dbDEMC and HMDD	hsa-mir-95	dbDEMC
hsa-mir-18b	dbDEMC and HMDD	hsa-mir-23b	dbDEMC and HMDD
hsa-mir-372	dbDEMC	hsa-mir-491	dbDEMC
hsa-mir-24	dbDEMC and HMDD	hsa-mir-183	dbDEMC and HMDD
hsa-mir-130a	dbDEMC	hsa-mir-31	dbDEMC and miR2Disease and HMDD
hsa-mir-15b	dbDEMC	hsa-mir-192	dbDEMC
hsa-mir-196b	dbDEMC	hsa-mir-135a	dbDEMC and HMDD

Actually, the method's high accuracy in the miRNA–disease association predictions mainly rely on the following attractive properties. First, the training set of known miRNA–disease associations used in this manuscript was collected from a very reliable database of HMDD v2.0, and the several bioinformatics data (disease semantic similarity, miRNA function similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases) mentioned in the paper were accurately calculated and integrated. All the reliable biological information mentioned above would attribute to the accuracy of BRWRMHMDA. Second, compared with the machine learning-based methods that randomly select negative samples as the training set, the proposed algorithm only uses positive samples as the training set that would provide higher prediction value. At last, BRWRMHMDA, a degree-biased random walk, could fully take advantage of the information about node degree and improve the prediction accuracy. From the preceding discussion, it is no surprise that this algorithm is superior to other comparison algorithms and has good performance.

However, the proposed model still has some weaknesses and shortcomings. For example, despite the biological information collected here being reliable, the number of 5,430 experimentally verified miRNA–disease associations extracted from HMDD v2.0 is still far from enough. If more associations between miRNAs and diseases are validated, the prediction accuracy of the model would be higher. Moreover, except for the fact that miRNA similarity could be calculated via the consideration of miRNA functional similarity and Gaussian interaction profile kernel for

miRNAs, it could also be calculated based on other miRNA features. At the same time, disease similarity could also be calculated based on other disease features. Also, the model could not predict candidate miRNAs for new diseases that have no known related miRNAs. In addition, due to the fact that the proposed algorithm is a local ranking model, it could not infer candidate miRNAs for all diseases simultaneously.

Nowadays, more and more researchers are studying the regulatory interactions between ncRNA classes, as well as the associations between ncRNA and other biological entities including diseases, small molecules, etc. Prediction of ncRNA-related networks will greatly expand our understanding of ncRNA function and its regulatory network. Simultaneously, predictions including miRNA–lncRNA interactions, miRNA–circRNA interactions, drug–target interactions, small molecule–miRNA associations, and disease–lncRNA associations have made great progress. In the field of miRNA–disease association prediction, the number of known miRNA–disease associations is limited, which will affect the prediction performance of the model. In the future, integrating multisource biological data that was mentioned above to build a multilayer heterogeneous network based on machine learning-based method can effectively improve the prediction performance of the 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/s.

AUTHOR CONTRIBUTIONS

JQ implemented the experiments, analyzed the result, and wrote the manuscript. C-CW analyzed the result, revised the manuscript, and supervised the project. S-BC and ZM analyzed the result and revised the manuscript. W-DZ and X-LC contributed to the analysis of the data for the manuscript and revised the manuscript. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.720327/full#supplementary-material

REFERENCES

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., et al. (2006). Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24, 537–544. doi: 10.1038/nbt1203

Alshalalfa, M., and Alhajj, R. (2013). Using context-specific effect of miRNAs to identify functional associations between miRNAs and gene signatures. *BMC Bioinformatics* 14:S1. doi: 10.1186/1471-2105-14-S12-S1

Ambros, V. (2004). The functions of animal microRNAs. *Nature* 431, 350–355. doi: 10.1038/nature02871

Axtell, M. J., Westholm, J. O., and Lai, E. C. (2011). Vive la difference: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.* 12:221. doi: 10.1186/gb-2011-12-4-221

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

Bonaventura, M., Nicosia, V., and Latora, V. (2014). Characteristic times of biased random walks on complex networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* 89:012803.

Bozgeyik, E. (2020). Bioinformatic analysis and in vitro validation of Let-7b and Let-7c in breast cancer. *Comput. Biol. Chem.* 84:107191. doi: 10.1016/j.compbiolchem.2019.107191

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., and Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* 113, 25–36. doi: 10.1016/s0092-8674(03)00231-9

Burda, Z., Duda, J., Luck, J. M., and Waclaw, B. (2009). Localization of the maximal entropy random walk. *Phys. Rev. Lett.* 102:160602.

Chen, X., and Huang, L. (2017). LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* 13:e1005912. doi: 10.1371/journal.pcbi.1005912

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

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018a). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* 9:3.

Chen, X., Qu, J., and Yin, J. (2018b). TLHNMDA: triple layer heterogeneous network based inference for MiRNA-disease association prediction. *Front. Genet.* 9:234. doi: 10.3389/fgene.2018.00234

Chen, X., Sun, L. G., 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., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018c). Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 34, 4256–4265.

Chen, X., Wu, Q. F., and Yan, G. Y. (2017). RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol.* 14, 952–962. doi: 10.1080/15476286.2017.1312226

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z. H., and Liu, H. (2018d). BNPMDA: bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* 34, 3178–3186. doi: 10.1093/bioinformatics/bty333

Chen, X., Xie, D., Zhao, Q., and You, Z. H. (2019a). MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 20, 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.

Chen, X., Yin, J., Qu, J., and Huang, L. (2018e). MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput. Biol.* 14:e1006418. doi: 10.1371/journal.pcbi.1006418

Chen, X., Zhou, Z., and Zhao, Y. (2018f). ELLPMDA: ensemble learning and link prediction for miRNA-disease association prediction. *RNA Biol.* 15, 807–818.

Chen, X., Zhu, C. C., and Yin, J. (2019b). Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput. Biol.* 15:e1007209. doi: 10.1371/journal.pcbi.1007209

Cheng, A. M., Byrom, M. W., Shelton, J., and Ford, L. P. (2005). Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res.* 33, 1290–1297. doi: 10.1093/nar/gki200

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., et al. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc. Natl. Acad. Sci. U.S.A.* 102, 13944–13949. doi: 10.1073/pnas.0506654102

Cui, Q., Yu, Z., Purisima, E. O., and Wang, E. (2006). Principles of microRNA regulation of a human cellular signaling network. *Mol. Syst. Biol.* 2:46. doi: 10.1038/msb4100089

Domper Arnal, M. J., Ferrandez Arenas, A., and Lanas Arbeloa, A. (2015). Esophageal cancer: risk factors, screening and endoscopic treatment in Western

and Eastern countries. *World J. Gastroenterol.* 21, 7933–7943. doi: 10.3748/wjg.v21.i26.7933

Gomez-Gardenes, J., and Latora, V. (2008). Entropy rate of diffusion processes on complex networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* 78:065102.

Hayashita, Y., Osada, H., Tatematsu, Y., Yamada, H., Yanagisawa, K., Tomida, S., et al. (2005). A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. *Cancer Res.* 65, 9628–9632. doi: 10.1158/0008-5472.can-05-2352

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.

Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., et al. (2005). RAS is regulated by the let-7 microRNA family. *Cell* 120, 635–647. doi: 10.1016/j.cell.2005.01.014

Karp, X., and Ambros, V. (2005). Encountering microRNAs in cell fate signaling. *Science* 310, 1288–1289. doi: 10.1126/science.1121566

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, S. P., Su, H. X., Zhao, D., and Guan, Q. L. (2016). Plasma miRNA-506 as a prognostic biomarker for esophageal squamous cell carcinoma. *Med. Sci. Monit.* 22, 2195–2201. doi: 10.12659/msm.899377

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.

Liang, A. L., Zhang, T. T., Zhou, N., Wu, C. Y., Lin, M. H., and Liu, Y. J. (2016). MiRNA-10b sponge: an anti-breast cancer study in vitro. *Oncol. Rep.* 35, 1950–1958. doi: 10.3892/or.2016.4596

Lin, S. Y., Johnson, S. M., Abraham, M., Vella, M. C., Pasquinelli, A., Gamberi, C., et al. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev. Cell* 4, 639–650. doi: 10.1016/s1534-5807(03)00124-2

Liu, Z., Sall, A., and Yang, D. (2008). MicroRNA: an emerging therapeutic target and intervention tool. *Int. J. Mol. Sci.* 9, 978–999. doi: 10.3390/ijms9060978

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An analysis of human microRNA and disease associations. *PLoS One* 3:e3420. doi: 10.1371/journal.pone.0003420

Ma, L., Teruya-Feldstein, J., and Weinberg, R. A. (2007). Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. *Nature* 449, 682–688. doi: 10.1038/nature06174

McGuire, A., Brown, J. A., and Kerin, M. J. (2015). Metastatic breast cancer: the potential of miRNA for diagnosis and treatment monitoring. *Cancer Metastasis Rev.* 34, 145–155. doi: 10.1007/s10555-015-9551-7

Meng, F., Henson, R., Wehbe-Janek, H., Ghoshal, K., Jacob, S. T., and Patel, T. (2007). MicroRNA-21 regulates expression of the PTEN tumor suppressor gene in human hepatocellular cancer. *Gastroenterology* 133, 647–658. doi: 10.1053/j.gastro.2007.05.022

Miska, E. A. (2005). How microRNAs control cell division, differentiation and death. *Curr. Opin. Genet. Dev.* 15, 563–568. doi: 10.1016/j.gde.2005.08.005

Mork, S., Pletscher-Frankild, S., Palleja Caro, A., Gorodkin, J., and Jensen, L. J. (2014). Protein-driven inference of miRNA-disease associations. *Bioinformatics* 30, 392–397. doi: 10.1093/bioinformatics/btt677

Pasquier, C., and Gardes, J. (2016). Prediction of miRNA-disease associations with a vector space model. *Sci. Rep.* 6:27036.

Pasquinelli, A. E., and Ruvkun, G. (2002). Control of developmental timing by micrornas and their targets. *Annu. Rev. Cell Dev. Biol.* 18, 495–513. doi: 10.1146/annurev.cellbio.18.012502.105832

Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* 31, 316–319. doi: 10.1038/ng895

Pillai, R. S., Bhattacharyya, S. N., Artus, C. G., Zoller, T., Cougot, N., Basyuk, E., et al. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* 309, 1573–1576. doi: 10.1126/science.1115079

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., et al. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature* 403, 901–906. doi: 10.1038/35002607

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer statistics, 2018. *CA Cancer J. Clin.* 68, 7–30. doi: 10.3322/caac.21442

Sun, L., Dong, S., Dong, C., Sun, K., Meng, W., Lv, P., et al. (2016). Predictive value of plasma miRNA-718 for esophageal squamous cell carcinoma. *Cancer Biomark.* 16, 265–273. doi: 10.3233/cbm-150564

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. *CA Cancer J. Clin.* 65, 87–108. doi: 10.3322/caac.21262

van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500

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

Wei, Y. T., Guo, D. W., Hou, X. Z., and Jiang, D. Q. (2017). miRNA-223 suppresses FOXO1 and functions as a potential tumor marker in breast cancer. *Cell Mol. Biol. (Noisy-le-grand)* 63, 113–118. doi: 10.14715/cmb/2017.63.5.21

Xuan, P., Han, K., Guo, M., Guo, Y., Li, J., Ding, J., et al. (2013). Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PLoS One* 8:e70204. doi: 10.1371/journal.pone.0070204

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

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 11 Suppl 4:S5. doi: 10.1093/nar/gkw1079

Yu, H., Chen, X., and Lu, L. (2017). Large-scale prediction of microRNA-disease associations by combinatorial prioritization algorithm. *Sci. Rep.* 7:43792.

Yu, Z., Ni, F., Chen, Y., Zhang, J., Cai, J., and Shi, W. (2020). miR-125b suppresses cell proliferation and metastasis by targeting HAX-1 in esophageal squamous cell carcinoma. *Pathol. Res. Pract.* 216:152792. doi: 10.1016/j.prp.2019.152792

Zhang, H. F., Zhang, K., Liao, L. D., Li, L. Y., Du, Z. P., Wu, B. L., et al. (2014). miR-200b suppresses invasiveness and modulates the cytoskeletal and adhesive machinery in esophageal squamous cell carcinoma cells via targeting Kindlin-2. *Carcinogenesis* 35, 292–301. doi: 10.1093/carcin/bgt320

Zhang, L., Ma, J., Han, Y., Liu, J., Zhou, W., Hong, L., et al. (2016). Targeted therapy in esophageal cancer. *Expert. Rev. Gastroenterol. Hepatol.* 10, 595–604.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 August 2021
doi: 10.3389/fgene.2021.712170

Predicting LncRNA–Disease Association by a Random Walk With Restart on Multiplex and Heterogeneous Networks

*Yuhua Yao[1,2,3], Binbin Ji[4], Yaping Lv[1], Ling Li[5], Ju Xiang[6,7,8], Bo Liao[1] and Wei Gao[9]**

[1] School of Mathematics and Statistics, Hainan Normal University, Haikou, China, [2] Key Laboratory of Data Science and Intelligence Education, Ministry of Education, Hainan Normal University, Haikou, China, [3] Key Laboratory of Computational Science and Application of Hainan Province, Hainan Normal University, Haikou, China, [4] Geneis Beijing Co., Ltd., Beijing, China, [5] Basic Courses Department, Zhejiang Shuren University, Hangzhou, China, [6] School of Computer Science and Engineering, Central South University, Changsha, China, [7] Department of Basic Medical Sciences, Changsha Medical University, Changsha, China, [8] Department of Computer Science, Changsha Medical University, Changsha, China, [9] Departments of Internal Medicine-Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fuzhou, China

Studies have found that long non-coding RNAs (lncRNAs) play important roles in many human biological processes, and it is critical to explore potential lncRNA–disease associations, especially cancer-associated lncRNAs. However, traditional biological experiments are costly and time-consuming, so it is of great significance to develop effective computational models. We developed a random walk algorithm with restart on multiplex and heterogeneous networks of lncRNAs and diseases to predict lncRNA–disease associations (MHRWRLDA). First, multiple disease similarity networks are constructed by using different approaches to calculate similarity scores between diseases, and multiple lncRNA similarity networks are also constructed by using different approaches to calculate similarity scores between lncRNAs. Then, a multiplex and heterogeneous network was constructed by integrating multiple disease similarity networks and multiple lncRNA similarity networks with the lncRNA–disease associations, and a random walk with restart on the multiplex and heterogeneous network was performed to predict lncRNA–disease associations. The results of Leave-One-Out cross-validation (LOOCV) showed that the value of Area under the curve (AUC) was 0.68736, which was improved compared with the classical algorithm in recent years. Finally, we confirmed a few novel predicted lncRNAs associated with specific diseases like colon cancer by literature mining. In summary, MHRWRLDA contributes to predict lncRNA–disease associations.

Keywords: lncRNA, disease, association, networks, random walk, predict

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology, China

Reviewed by:
Wei Peng,
Kunming University of Science and Technology, China
Kebo Lv,
Ocean University of China, China

****Correspondence:***
Wei Gao
13960986882@163.com

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

Received: *20 May 2021*
Accepted: *23 July 2021*
Published: *19 August 2021*

Citation:
Yao Y, Ji B, Lv Y, Li L, Xiang J, Liao B and Gao W (2021) Predicting LncRNA–Disease Association by a Random Walk With Restart on Multiplex and Heterogeneous Networks. Front. Genet. 12:712170. doi: 10.3389/fgene.2021.712170

INTRODUCTION

Numerous studies have indicated that protein-coding genes accounted for less than 2% of the human genome (Crick et al., 1961; Yanofsky, 2007). There are many non-translatable RNAs called non-coding RNAs (ncRNAs), which have been considered as transcriptional noise for a long time (Zhang et al., 2017; Xu et al., 2020). Long non-coding RNAs (lncRNAs) whose length are greater than 200 nucleotides are a class of important ncRNAs (Mercer et al., 2009). There are increasing evidence that lncRNAs play key roles in many important biological processes and

diseases (Akerman et al., 2017; Wang et al., 2019; Peng et al., 2020). For example, HOTAIR was considered as a potential biomarker for liver cancer (Yang et al., 2011; Li et al., 2019), lung cancer (Li G. et al., 2014a), and colorectal cancer (Kogo et al., 2011; Maass et al., 2014), and UCA1 was a potential biomarker for bladder cancer diagnosis (Zhang et al., 2012). Li J. et al. (2014b) summarized the important role of lncRNA such as MALAT1, HOTAIR, and other specific lncRNAs for hepatocellular carcinoma. LncRNAs associated with tumor immune invasion in non-small cell lung cancer (NSCLC) have important value in improving clinical efficacy and immunotherapy, compared with normal controls, and the expression of gabpb1-it1 was significantly downregulated in NSCLC. In addition, overexpression of gabpb1-it1 in cancer samples is associated with increased survival in NSCLC patients (Sun et al., 2020). Inferring the association between lncRNA and diseases can better study human diseases and help the diagnosis and treatment of diseases, and accelerate the identification of potential drug response predictors (Liu et al., 2016, 2020). Therefore, the exploration of lncRNA–disease association has attracted more and more attention from biologists. The establishment of an effective computational model to predict the association between lncRNAs and diseases can save time and money spent in biological experiments (Yao et al., 2019; Yan et al., 2020).

At present, many machine learning methods have been proposed to predict the lncRNA–disease association, for example, Laplacian regulated least square method (LRLSLDA; Chen and Yan, 2013), propagation algorithm (Yang et al., 2014), a method based on Bayesian classifier (Zhao et al., 2015), and a method based on induction matrix (Lu C. et al., 2018a). However, these machine learning methods need negative samples, which are difficult to obtain. In order to solve this problem, network-based methods emerge as the times require. With the increasing importance of revealing the molecular basis of human diseases, network-based methods have been widely used in exploring disease-related genes (Yan et al., 2015; Hu et al., 2018; Lu M. et al., 2018b; Yang et al., 2020). For example, Xiang et al. (2021) proposed a multibiological network (NIDM) network pulse dynamics framework and a fast network embedding (Xiang et al., 2020) to predict disease-related genes. Network-based algorithms have also been widely studied in predicting lncRNA–disease association. Bellucci et al. (2011) combined the expression similarity of lncRNA with the Gaussian nuclear interaction spectrum similarity of lncRNA, and proposed a potential protein determination method based on sequence information to predict the function of lncRNA. In the study of Xiao et al. (2015), the function of lncRNA was predicted by constructing the regulatory network between lncRNA and protein coding genes. In the BPLLDA study, the authors estimated the potential relationship between disease and lncRNAs by connecting the length of the disease and lncRNA pathway (Xiao et al., 2018). KATZLDA was a computing method to predict lncRNA–disease association based on the similarity between heterogeneous network nodes (Chen, 2015a). The random walk model is also widely used in the field of data mining and Internet, and many researches use this method to predict potential association (Xing et al., 2012; Yang et al., 2016, 2017; Gu et al., 2017). Zhou et al. (2015) proposed a new method by integrating the related lncRNA–lncRNA network, disease–disease similarity network, and the heterogeneous lncRNA–disease association network, and then realized random walk on the heterogeneous network. Sun et al. (2014) proposed a method for constructing lncRNA–lncRNA functional similarity network and then developed a calculation method based on global network (RWRlncD). Recently, Lei and Bian (2020) used random walk to weight the structural features of circRNA–disease pairs and combined it with k-nearest neighbor algorithm to get the prediction score of each circRNA–disease pair. Although these methods have been proposed to predict lncRNA–disease association successfully, it is still a challenge to make full use of multi-source biological data.

In this study, a random walk algorithm with restart on multiplex and heterogeneous networks was developed. The downloaded known lncRNA–disease association data were used to calculate lncRNA functional similarity, lncRNA Gaussian interaction kernel similarity, disease semantic similarity, and disease Gaussian interaction kernel similarity, respectively. Then, these similarity networks and lncRNA–disease association network were constructed into multiplex and heterogeneous networks. A random walk with restart was carried out on the multiplex and heterogeneous networks, and the potential lncRNA–disease association was predicted using the final stable probability.

MATERIALS AND METHODS

LncRNA–Disease Association

LncRNADisease (Chen, 2015b), Lnc2Cancer (Ning et al., 2016), MNDR (Wang et al., 2013), and other databases stored the known lncRNA–disease association data, which have been of great help in predicting novel association. In this study, 285 lncRNA–disease association was downloaded from lncRNADisease database, including 117 lncRNAs and 159 diseases. We used LD to represent the lncRNA–disease association adjacency matrix. If lncRNA(i) is related to disease(j), then $LD(i, j) = 1$; otherwise, $LD(i, j) = 0$, that is:

$$LD(i,j) = \begin{cases} 1, & if\ lncRNA\ (i)\ is\ associated\ with\ disease\ (j) \\ 0 & otherwise \end{cases} \quad (1)$$

Disease Similarity

Disease Semantic Similarity

Directed acyclic graphs ($DAGs$) were used to calculate disease–disease similarity, for disease d_k, let $DAG(d_k, T(d_k), E(d_k))$ be its directed acyclic graph, where $T(d_k)$ are ancestor nodes of d_k, and $E(d_k)$ represents the corresponding set of edges from parent node to child nodes. Semantic similarity of diseases was calculated by R package called DOSim (Li et al., 2011); for any disease k in $DAG(d_k, T(d_k), E(d_k))$, the semantic contribution of k to d_k was

defined as:

$$D_{d_k}(k) = \begin{cases} 1, & if\ k = d_k \\ max\{0.5 * D_{d_k}(k') | k' \in children\ of\ k\}, & if\ k \neq d_k \end{cases} \quad (2)$$

The above formula indicates that the contribution of the disease to its semantic value is 1. Semantic contribution decreased with the increase of the distance between disease k and other diseases. Then, the semantic similarity between d_i and d_j was defined as:

$$DSS(d_i, d_j) = \frac{\sum_{k \in T_{d_i} \cap T_{d_j}} (D_{d_i}(k) + D_{d_j}(k))}{\sum_{k \in T_{d_i}} D_{d_i}(k) + \sum_{k \in T_{d_j}} D_{d_j}(k)} \quad (3)$$

Gaussian Interaction Profile Kernel Similarity for Diseases

In order to obtain the similarity information between diseases, the Gaussian Interaction Profile kernel similarity between disease was constructed based on the lncRNA–disease association network. First, the Interaction Profile (IP) of each disease represents a binary code in the known lncRNA–disease association network. For example, for given disease d_i, its $IP(d_i)$ represents the ith column of LD. Next, the Gaussian Interaction Profile kernel similarity between d_i and $d_j$was calculated as:

$$DS_{GIP}(d_i, d_j) = exp(-\gamma_d \left|\left|IP(d_i) - IP(d_j)\right|\right|^2) \quad (4)$$

Where γ_d represents the bandwidth that controls the Gaussian Interaction Profile kernel similarity, $\gamma_d = \frac{\gamma'_d}{(\frac{1}{nd}\sum_{i=1}^{nd} \left|\left|IP(d_i)\right|\right|^2)}$; in this study, according to van Laarhoven et al. (2011), we set $\gamma'_d = 1$, and nd represents the number of diseases.

LncRNA Similarity

LncRNA Functional Similarity

Studies have shown that similar lncRNAs are usually associated with similar diseases. Therefore, lncRNA functional similarity can be roughly estimated by their similarity in related diseases (Sun et al., 2014). For any two lncRNAs l_i and l_j, $D_i = \{d_{i_k} | 1 \leq k \leq m\}$ and $D_j = \{d_{j_l} | 1 \leq l \leq n\}$ were disease sets associated with l_i and l_j, respectively. The semantic similarity between disease d and disease set D was firstly defined as:

$$SS(d, D) = \max_{d_l \in D} DSS(d, d_l) \quad (5)$$

Then, the functional similarity between l_i and l_j was defined as:

$$NFS(l_i, l_j) = \frac{\sum_{i=1}^{m} SS(d_{i_a}, D_j) + \sum_{j=1}^{n} SS(d_{j_b}, D_i)}{m + n} \quad (6)$$

Gaussian Interaction Profile Kernel Similarity for LncRNAs

Similar to the disease Gaussian interaction profile kernel similarity. The formula for calculating the Gaussian interaction profile kernel similarity between l_i and l_j was:

$$LS_{GIP}(l_i, l_j) = exp(-\gamma_l \left|\left|IP(l_i) - IP(l_j)\right|\right|^2) \quad (7)$$

Where γ_l represents the bandwidth that controls the property similarity of Gaussian interaction kernel, $\gamma_l = \frac{\gamma'_l}{(\frac{1}{nl}\sum_{i=1}^{nl} \left|\left|IP(l_i)\right|\right|^2)}$; in this study, $\gamma'_l = 1$, nl represents the number of lncRNAs, $IP(l_i)$ and $IP(l_j)$ represent the ith and jth row of the LD, respectively.

A Random Walk With Restart on Multiplex and Heterogeneous Networks

An overview of MHRWRLDA is shown in **Figure 1**. Specifically, we first downloaded the data of known lncRNA–disease association from the LncRNADisease database and got diseased' DO ID from the DO database to calculate disease similarity. After compute disease similarity and lncRNA similarity, a multiplex and heterogeneous network was set up based on these similarity networks and known lncRNA–disease association network. Finally, a random walk algorithm with restart was implemented on networks, and the final stability probability was used to conduct the predictions.

Multiplex and Heterogeneous Network

Based on disease semantic similarity network, disease Gaussian similarity network, lncRNA similarity network, and lncRNA Gaussian similarity network, we constructed a multiplex and heterogeneous network by using lncRNA–disease association. In these networks, the set of lncRNA nodes was defined as: $R_M = \{v_i^\alpha, i = 1, 2, \cdots, n; \alpha = 1, 2, \cdots, L\}$, where v_i^α represents the ith node on the α layer. The set of disease nodes was defined as: $D_M = \{v_j^\beta, j = 1, 2, \cdots, m; \beta = 1, 2, \cdots, K\}$, where v_i^β represents the jth node on the β layer. The adjacency matrix on each layer is:

$$A^{[\alpha]} = A^{[\alpha]}(i, j) = \begin{cases} 1, & if\ the\ i^{th}\ node\ is\ associated\ with \\ & the\ jth\ node\ on\ layer\ \alpha \\ 0, & otherwise \end{cases} \quad (8)$$

A particle can either travel from the previous node v_i^α term to any neighbor node on the same layer, or it can also jump to the same node on a different layer. The matrix A contains different types of jumps that the particle can follow at each step:

$$A = \begin{pmatrix} (1-\delta)A^{[1]} & \frac{\delta}{(L-1)}I & \cdots & \frac{\delta}{(L-1)}I \\ \frac{\delta}{(L-1)}I & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}A \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta}{(L-1)}I & \frac{\delta}{(L-1)}I & \cdots & (1-\delta)A^{[L]} \end{pmatrix} \quad (9)$$

Where I is the $n \times n$ identity matrix, the diagonal element of A represents the particle walking on same layer, the off-diagonal element represents the particle jumping between different layers, and the parameter $\delta \in (0, 1)$ represents the probability of the particle walking on the same layer or jumping between different layers. If $\delta = 0$, the particles will always walk on the same layer.

$A_{RM(nL \times nL)}$, $A_{DM(mK \times mK)}$ is the matrix of lncRNA similarity and disease similarity on multiplex and heterogeneous networks, respectively. n, L, m, and K are the number of

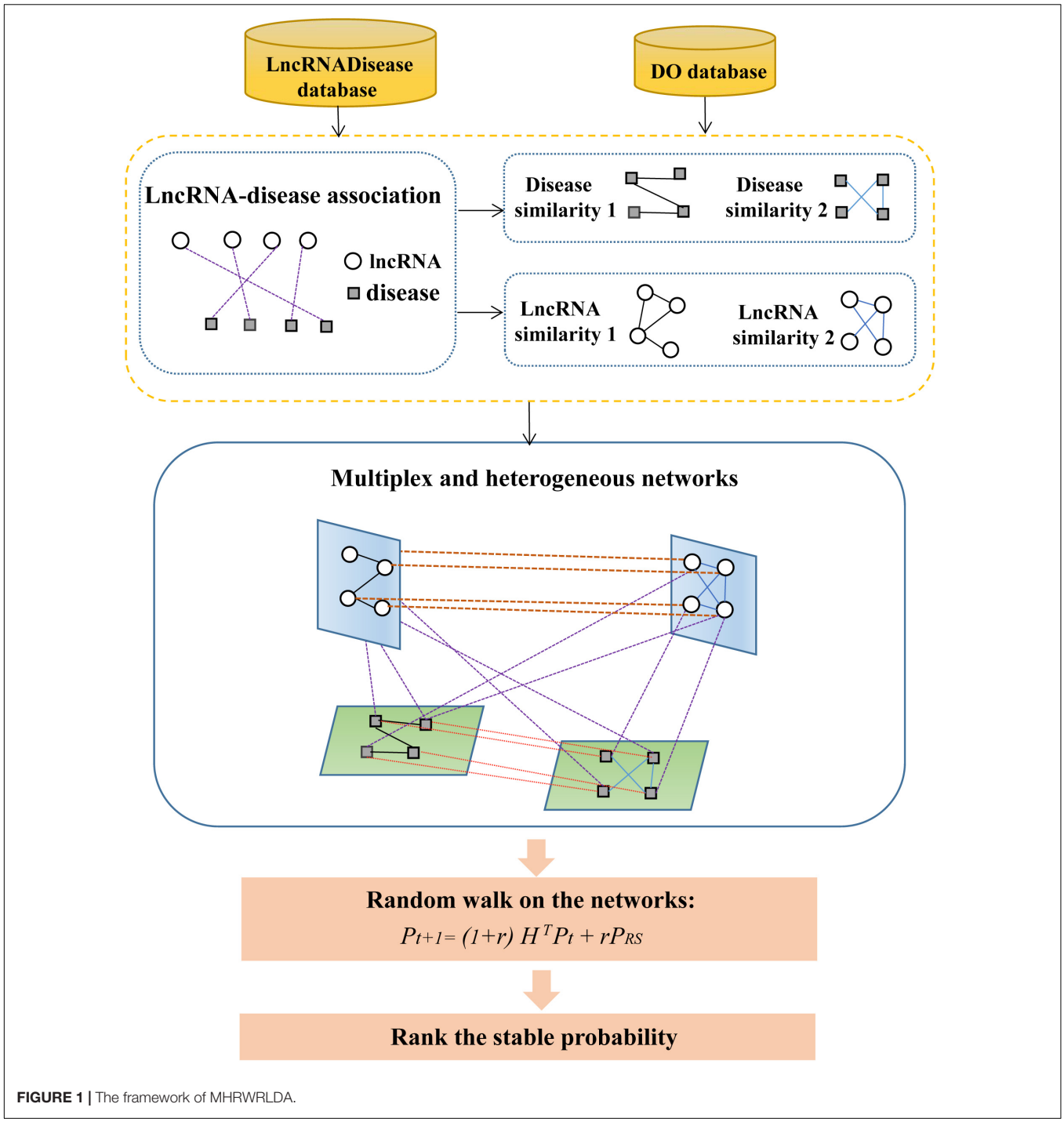


FIGURE 1 | The framework of MHRWRLDA.

lncRNAs, lncRNA similarity networks, diseases, and disease similarity networks, respectively, the adjacency matrix is: $B_{MH} = (B_{n \times m}, B_{n \times m}, \cdots, B_{n \times m})^T$.

The dimension of B_{MH} is $nL \times mK$, which is equivalent to replicating the adjacency matrix $B_{n \times m}$ L^*K times, where $B = LD$. Then, the adjacency matrix of the whole multiplex and heterogeneous networks is: $A = \begin{bmatrix} A_{RM} & B_{MH} \\ B_{MH}^T & A_{DM} \end{bmatrix}$.

Random Walk With Restart on Multiplex and Heterogeneous Networks

A random walk with a restart means that a particle starts at a node and it is faced with two choices at each walk: move to a randomly selected neighbor node, or jump back to the start node. Considering the time is discrete, $t \in \mathbb{N}$, the particle is at node v_t at the tth step. Then, it walks from v_t to v_{t1}. We defined a restart probability $\gamma \in (0, 1)$, and the random walk with restart can be

TABLE 1 | Confusion matrix definitions.

True prediction	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FN)	True negative (TN)

defined as:

$$P_{t+1} = (1-\gamma)H^T P_t + \gamma P_{RS} \tag{10}$$

Where the vectors P_{t1} and P_t represent the probability distribution of v_t and v_{t1}, respectively. P_{RS} is the initial probability distribution and $P_{RS} = \begin{bmatrix} (1-\eta)R_0 \\ \eta D_0 \end{bmatrix}$; the importance of each network is adjusted by adjusting P_{RS}, where R_0 and D_0 represent the initial probability distribution of lncRNA similarity network and disease similarity network, respectively, and the dimensions of the vectors P_{t+1}, P_t, and P_{RS} are $nL \times mK$. The parameter $\eta \in (0, 1)$ controls the probability of each network restarting; if $\eta < 0.5$, the particle is more likely to be restarted in lncRNA similarity networks. $H = \begin{bmatrix} H_{RR} & H_{RD} \\ H_{DR} & H_{DD} \end{bmatrix}$ represents the transition probability matrix of multiplex and heterogeneous networks, where H_{RR} andH_{DD} represent the transition probability of nodes upstream in the same layer, H_{RD} and H_{DR} represent the transition probability of node jump between different layers. For a given node, if dichotomous correlation exists, the particle can jump between layers or stay in the current layer with probability $\lambda \in (0, 1)$, and the closer it is to 1, the higher the probability of jumping between different networks.

We suppose a particle was located at the node $r_i \in R$. In the next step, the particle can walk to the node $r_j \in R$. The transfer probability is:

$$H_{RR} = \begin{cases} \frac{A_R(i,j)}{\sum_{k=1}^{n} A_R(i,k),} & \text{if } \sum_{k=1}^{m} B(i,k) = 0 \\ (1-\lambda)\frac{A_R(i,j)}{\sum_{k=1}^{n} A_R(i,k),} & \text{otherwise} \end{cases} \tag{11}$$

It can also jump to the node $d_b \in D$ through binary correlation, and the transfer probability is:

$$H_{RD} = \begin{cases} \frac{\lambda B(i,b)}{\sum_{k=1}^{m} B(i,k),} & \text{if } \sum_{k=1}^{m} B(i,k) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Similarly, if the particle was located at the node $d_a \in D$, then the transfer probability of the particle walking to the node $d_b \in D$ is:

$$H_{DD} = \begin{cases} \frac{A_D(a,b)}{\sum_{k=1}^{m} A_D(a,k),} & \text{if } \sum_{k=1}^{n} B(k,b) = 0 \\ (1-\lambda)\frac{A_D(a,b)}{\sum_{k=1}^{m} A_D(a,k),} & \text{otherwise} \end{cases} \tag{13}$$

If the particle jumps to the node $r_j \in R$ through binary correlation, then the transfer probability is:

$$H_{DR} = \begin{cases} \frac{\lambda B(j,a)}{\sum_{k=1}^{n} B(k,a),} & \text{if } \sum_{k=1}^{n} B(k,a) \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

When predicting lncRNAs that are potentially associated with the given disease d_i, the node d_i will be used as the seed node in disease similarity networks. The initial probability D_0 is 1 for the given node d_i and 0 for the remaining nodes. If there are known associations among lncRNAs $r_1, r_2 \cdots$ and disease d_i, then the nodes $r_1, r_2 \cdots$ are the seed nodes in lncRNA similarity networks. The initial probability R_0 was assigned to seed node $r_1, r_2 \cdots$, with a probability of 1, and the remaining nodes were 0. P_t converges after some iteration, that is, $P_t - P_{t+1} < 10^{-10}$, and we denoted the stable probability as: $P_\infty = \begin{bmatrix} (1-\eta)R_\infty \\ \eta D_\infty \end{bmatrix}$.

Based on the stabilized R_∞, those seed nodes $r_1, r_2 \cdots$ were removed, and the remaining lncRNAs were ranked. The higher the ranked lncRNA, the more likely it was to be associated with the given disease d_i. Similarly, a lncRNA can also be designated to predict diseases related to it.

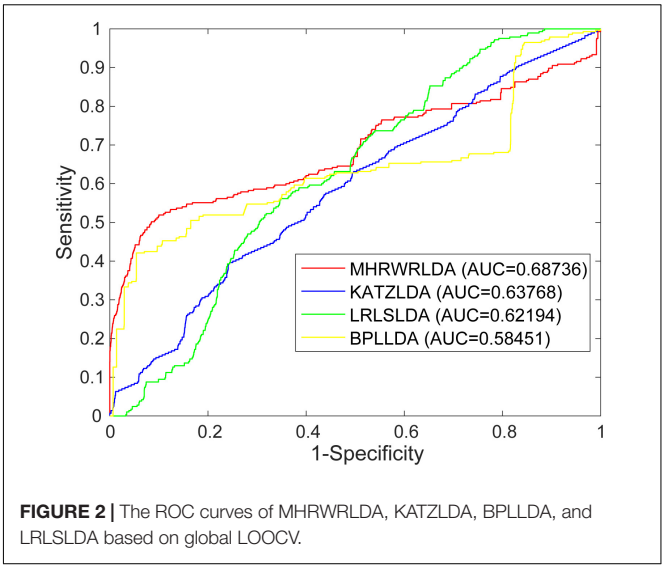


FIGURE 2 | The ROC curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV.

RESULTS

Indicators of Performance Evaluation

For a binary classification problem, the confusion matrix is shown in **Table 1**. Precision, specificity, and sensitivity are evaluation indicators of classification models. They are calculated as:

$$FPR = 1-\text{specificity} = \frac{FP}{TN+FP}$$

$$TPR = \text{sensitivity} = \frac{TP}{TP+FN}$$

To evaluate the performance of MHRWRLDA, the receiver operating characteristic (ROC) curve was drawn by calculating TPR and FPR according to different thresholds. Area under the curve (AUC) is the area under the ROC curve, and this area is less than 1. Since the ROC curve cannot directly indicate which classifier has better effect in many cases, as a value, the larger the AUC is, the better the classifier has an effect.

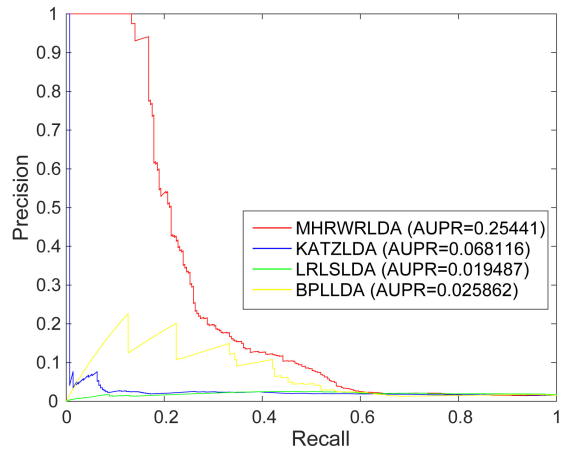


FIGURE 3 | The PR curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV.

TABLE 2 | The predicted top 10 lncRNAs for colon cancer.

Disease	Rank	LncRNA	Evidence
Colon cancer	1	H19	Confirmed
	2	MEG3	Confirmed
	3	CDKN2B-AS1	Confirmed
	4	MALAT1	Confirmed
	5	PVT1	Confirmed
	6	BCYRN1	Unknown
	7	IGF2-AS	Confirmed
	8	Anti-NOS2A	Unknown
	9	WT1-AS	Unknown
	10	UCA1	Confirmed

TABLE 3 | The predicted top 10 lncRNAs for hepatocellular carcinoma.

Disease	Rank	LncRNA	Evidence
Hepatocellular carcinoma	1	H19	Confirmed
	2	MEG3	Confirmed
	3	MALAT1	Confirmed
	4	AIR	Confirmed
	5	HULC	Confirmed
	6	HOTAIR	Confirmed
	7	IGF2-AS	Confirmed
	8	CDKN2B-AS1	Confirmed
	9	PVT1	Confirmed
	10	BCYRN1	Unknown

Performance of MHRWRLDA

In order to evaluate the performance of MHRWRLDA for predicting lncRNA–disease association, we applied the known lncRNA–disease association data to MHRWRLDA, and used Leave-One-Out cross-validation (LOOCV) to verify. For global LOOCV, the scores of all test samples are compared with those of all candidate samples. For local LOOCV, each known lncRNA related to a particular disease is selected as the test sample, and other related lncRNAs are selected as the training samples; the scores of test samples are only compared with those of candidate samples. In this study, there are a total of three parameters, namely, γ, λ, and η, and their range is (0, 1), where γ is the restart probability; λ is the jump probability, reflecting the probability of particles jumping between different networks; and η regulated the probability of each network restarting, When $\eta = \gamma = 0.9$ and $\lambda = 0.9$, the prediction effect is the best; at this point, AUC = 0.68736.

The AUC based on global LOOCV of the KATZLDA (Chen, 2015a), BPLLDA (Xiao et al., 2018), and LRLSLDA (Chen and Yan, 2013) were 0.63768, 0.5845, and 0.6219, respectively. The ROC curves of MHRWRLDA, KATZLDA, BPLLDA, and LRLSLDA based on global LOOCV are shown in **Figure 2**, the PR curves based on global LOOCV are shown in **Figure 3**, and the AUPR values are shown in their legends. Their ROC curves and PR curves based on local LOOCV are shown in **Supplementary Figures 1**, **2**. The results showed that MHRWRLDA performed better than other classical algorithms in predicting lncRNA–disease association.

TABLE 4 | The predicted top 10 lncRNAs for breast cancer.

Disease	Rank	LncRNA	Evidence
Breast cancer	1	H19	Confirmed
	2	CDKN2B-AS1	Confirmed
	3	PVT1	Confirmed
	4	MEG3	Confirmed
	5	BCYRN1	Confirmed
	6	SRA1	Confirmed
	7	XIST	Confirmed
	8	GAS5	Confirmed
	9	HOTAIR	Confirmed
	10	DSCAM-AS1	Confirmed

TABLE 5 | The predicted top five novel disease correlated with MALAT1, PVT1, and MEG3.

LncRNA	Disease	Rank	Evidence
MALAT1	Endometrial stromal sarcoma	1	Confirmed
	Non-small cell lung cancer	2	Confirmed
	Lung adenocarcinoma	3	Confirmed
	Cervical cancer	4	Confirmed
	Osteosarcoma	5	Confirmed
PVT1	Burkitt's lymphomas	1	Confirmed
	Hodgkin's lymphoma	2	Confirmed
	Renal disease	3	Confirmed
	Diabetic nephropathy	4	Confirmed
	Pancreatic cancer	5	Confirmed
MEG3	Pituitary adenomas	1	Confirmed
	Heroin addiction	2	Confirmed
	Nonfunctioning pituitary adenomas	3	Confirmed
	Chronic myeloid leukemia	4	Confirmed
	Myelodysplastic syndrome	5	Confirmed

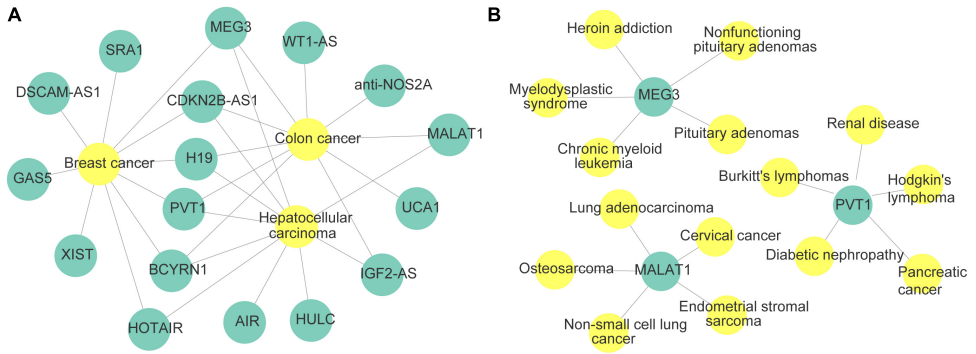


FIGURE 4 | The network of diseases and lncRNAs were made by Cytoscape. **(A)** The network of novel lncRNAs related to colon cancer, hepatocellular carcinoma, and breast cancer. **(B)** The network of novel diseases related to lncRNA MALAT1, PVT1, and MEG3.

Case Study

To further explore the performance of MHRWRLDA in predicting lncRNA–disease association, we selected colon cancer, hepatocellular carcinoma, and breast cancer for the case study. During the experiment, all known associations were considered as the train set, and unknown associations were regarded as the test set. According to LOOCV results, we sorted lncRNAs and selected the top 10 lncRNAs for further verification based on the LncRNADisease database and several recently published studies.

Colon cancer is a malignant tumor, causing nearly 700,000 deaths each year, and has a high incidence rate record in developed countries. We applied MHRWRLDA to colon cancer experiments to predict the top 10 lncRNAs related to colon cancer (**Table 2**). Seven of the top 10 lncRNAs have been confirmed in databases or other literature. Previous studies have found that the third ranked CDKN2B-AS1 up-regulates HCT116, thereby causing cell proliferation (Chiyomaru et al., 2013). In addition, studies have shown that removal of PVT1 (ranked 5) from MCY-driven colon cancer strain HCT116 can reduce carcinogenicity (Tseng et al., 2014).

Hepatocellular carcinoma is one of the most common cancers in the world. Studies have shown that hepatocellular carcinoma is the main component of primary liver cancer. We listed the top 10 lncRNAs related to hepatocellular carcinoma predicted by experiments in **Table 3**. Of the top 10, 9 were all verified in known databases. The overexpression of CDKN2B-AS1, which ranked 8, can inhibit the proliferation and invasion of liver cancer cells (Hua et al., 2015), thereby promoting the apoptosis of liver cancer cells and preventing the occurrence of hepatocellular carcinoma. Ding et al. identified PVT1 (ranked 9) as a novel biomarker for predicting tumor recurrence in patients with hepatocellular carcinoma (Ding et al., 2015).

Breast cancer accounts for 22% of all cancers in women and is the second leading cause of cancer death in women (Donahue and Genetos, 2013; Karagoz et al., 2015). Traditionally, breast cancer has been diagnosed on the basis of histopathological features such as tumor size, grade, and lymph node status. The prediction of breast cancer-related lncRNAs may help diagnose and treat breast cancer (Meng et al., 2014). In order to diagnose and treat breast cancer better, it is necessary to predict lncRNAs associated with breast cancer and identify lncRNA biomarkers (Xu et al., 2015). We implemented MHRWRLDA on breast cancer to predict potentially relevant lncRNAs, and listed the top 10 lncRNAs related to breast cancer in **Table 4**. The downregulation of the top ranked first H19 significantly reduced breast cancer clonal formation and anchored independent growth (Barsytelovejoy et al., 2006). In addition, the incidence of breast cancer is also affected by PVT1 overexpression due to genomic abnormalities (Guan et al., 2007).

Finally, the network of three cases and lncRNAs predicted by MHRWRLDA is shown in **Figure 4A**; it revealed that MEG3, CDKN2B-AS1, H19, PVT1, BCYRN1, HOTAIR, and all three diseases are related. In addition to exploring lncRNAs related to novel diseases, it is also extremely important to predict diseases related to novel lncRNAs. Therefore, taking lncRNA MALAT1, PVT1, and MEG3 as examples, the predicted top five diseases related to them are listed in **Table 5**, and their network is shown in **Figure 4B**. The experimental results proved that MHRWRLDA was useful for predicting the potential lncRNA–disease association.

DISCUSSION

In recent years, the research on the interaction between biomolecules has been growing. Due to the importance of lncRNA, the research on the associations between lncRNAs and diseases has been paid more and more attention. These associations can be characterized by complex networks, so it is urgent to develop network-based computational algorithms to explore functional associations between lncRNAs and diseases. The algorithm of constructing heterogeneous network and implementing random walk on heterogeneous network is widely used in the field of bioinformatics. However, in previous studies, most of them are single heterogeneous networks with a single information source. Therefore, we consider multiple network embedding by integrating different types of edges. Multiplex and heterogeneous networks are the combination of heterogeneous networks connected by multiple interactions; they integrate the framework of multiple information sources, and each layer is

a simplex network with specific types of nodes and edges; when the data set is large, they can produce better results. Multiple heterostructures may provide a richer perspective for the study of the complex relationship between different biological components.

In this study, we extend it to multi-layer heterogeneous networks so as to more effectively predict lncRNA–disease associations. We constitute a multiplex and heterogeneous network by integrating known lncRNA–disease association, lncRNA function similarity, lncRNA Gaussian similarity network, disease semantic similarity network, and disease Gaussian similarity network, and then we generate the final comprehensive predictive scores by the random walk with restart on the multiplex and heterogeneous network, so as to forecast potential lncRNA–disease associations. LOOCV experimental verification results showed that the AUC was 0.68736, which exceeded other algorithms to predict lncRNA–disease association. In novel diseases, the top 10 lncRNAs were verified and predicted by database or literature. In addition, the model can also predict diseases associated with particular lncRNAs.

The network-based approach overcomes the disadvantage of machine learning methods that need to construct negative samples and not only is suitable for predicting lncRNA–disease associations, but also proved to be widely used in exploring disease-related miRNAs, drug repositioning, and prediction of disease–gene associations. Therefore, if the known lncRNA–disease association data are replaced with miRNA-disease association data, MHRWRLDA can be used to predict the potential miRNAs associated with disease; similarly, if it is replaced by drug–disease association data or gene–disease association data, it is possible to make contributions to drug repositioning and the exploration of disease-related genes, respectively. In the future, we will try to apply MHRWRLDA to the above aspects for research.

However, there are some limitations. First, there are only two methods for constructing the similarity network; if the calculation method of the similarity network can be increased, the number of layers in the multi-layer heterogeneous graph can be increased to provide more possibilities for particle migration. Second, the lncRNA–disease association data contain only 117 lncRNAs and 159 diseases, of which there are only 285 pairs of correlations; a small data set may also affect the prediction results. In the future, more association data will be discovered and used to overcome the difficulties caused by the complexity and inconsistency of biological data. In addition, efforts will be made to combine multiple prediction models to achieve more accurate predictions.

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://github.com/jibinbin171222/MHRWRLDA.

AUTHOR CONTRIBUTIONS

WG conceived, designed, and managed the study. YY and BJ designed the method and wrote the original manuscript. YL and LL revised the original draft. JX wrote the code. BL discussed the proposed method and gave further research. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 61762035), the Hainan Provincial Natural Science Foundation of China (Grant No. 119MS037), the National Natural Science Foundation of China (Grant No. 61702054), the Training Program for Excellent Young Innovators of Changsha (Grant Nos. kq2009093 and kq1905045), and the Joint Funds for the Innovation of Science and Technology, Fujian province (Grant No. 2019Y9038).

ACKNOWLEDGMENTS

The data used to support the findings of this study are available from the corresponding author upon request.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.712170/full#supplementary-material

REFERENCES

Akerman, I., Tu, Z., Beucher, A., Rolando, D. M. Y., Sauty-Colace, C., Benazra, M., et al. (2017). Human pancreatic β Cell lncRNAs control cell-specific regulatory networks. *Cell Metab.* 25, 400–411. doi: 10.1016/j.cmet.2016.11.016

Barsytelovejoy, 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

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods* 8, 444–445. doi: 10.1038/nmeth.1611

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

Chen, X. (2015b). Predicting lncRNA–disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* 5:13186.

Chen, X., and Yan, G. Y. (2013). Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 29, 2617–2624. doi: 10.1093/bioinformatics/btt426

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

Crick, F., Barnett, L., Brenner, S., and Wattstobin, R. J. (1961). General nature of the genetic code for proteins. *Nature* 192, 1227–1232. doi: 10.1038/1921227a0

Ding, C., Yang, Z., Lv, Z., Du, C., Xiao, H., Peng, C., et al. (2015). Long non-coding RNA PVT1 is associated with tumor progression and predicts recurrence in hepatocellular carcinoma patients. *Oncol. Lett.* 9, 955–963. doi: 10.3892/ol.2014.2730

Donahue, H. J., and Genetos, D. C. (2013). Genomic approaches in breast cancer research. *Brief. Funct. Genomics* 12, 391–396.

Gu, C., Liao, B., Li, X., Cai, L., Li, Z., Li, K., et al. (2017). Global network random walk for predicting potential human lncRNA-disease associations. *Sci. Rep.* 7:12442.

Guan, Y., Kuo, W. L., Stilwell, J. L., Takano, H., Lapuk, A., Fridlyand, J., et al. (2007). Amplification of PVT1 contributes to the pathophysiology of ovarian and breast cancer. *Clin. Cancer Res.* 13, 5745–5755. doi: 10.1158/1078-0432.ccr-06-2882

Hu, K., Hu, J. B., Tang, L., Xiang, J., Ma, J. L., Gao, Y. Y., et al. (2018). Predicting disease-related genes by path structure and community structure in protein–protein networks. *J. Statal Mech. Theory Exp.* 2018:100001. doi: 10.1088/1742-5468/aae02b

Hua, L., Wang, C. Y., Yao, K. H., Chen, J. T., Zhang, J. J., and Ma, W. L. (2015). High expression of long non-coding RNA ANRIL is associated with poor prognosis in hepatocellular carcinoma. *Int. J. Clin. Exp. Pathol.* 8, 3076–3082.

Karagoz, K., Sinha, R., and Arga, K. Y. (2015). Triple negative breast cancer: a multi-omics network discovery strategy for candidate targets and driving pathways. *OMICS J. Integr. Biol.* 19, 115–130. doi: 10.1089/omi.2014.0135

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., et al. (2011). Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* 71, 6320–6326. doi: 10.1158/0008-5472.can-11-1021

Lei, X., and Bian, C. (2020). Integrating random walk with restart and k-nearest neighbor to identify novel circRNA-disease association. *Sci. Rep.* 10:1943.

Li, G., Zhang, H., Wan, X., Yang, X., Zhu, C., Wang, A., et al. (2014a). Long noncoding RNA plays a key role in metastasis and prognosis of hepatocellular carcinoma. *Biomed Res. Int.* 2014:780521.

Li, J., Gao, C., Wang, Y., Ma, W., Tu, J., Wang, J., et al. (2014b). A bioinformatics method for predicting long noncoding RNAs associated with vascular disease. *Sci. China Life Sci.* 57, 852–857. doi: 10.1007/s11427-014-4692-4

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., et al. (2011). DOSim: an R package for similarity between diseases based on disease ontology. *BMC Bioinformatics* 12:266. doi: 10.1186/1471-2105-12-266

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

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An improved anticancer drug-response prediction based on an ensemble method integrating matrix completion and ridge regression. *Mol. Ther. Nucleic Acids* 21, 676–686. doi: 10.1016/j.omtn.2020.07.003

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A systematic study on drug-response associated genes using baseline gene expressions of the cancer cell line encyclopedia. *Sci. Rep.* 6:22811.

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

Lu, M., Xu, X., Xi, B., Dai, Q., Li, C., Su, L., et al. (2018b). Molecular network-based identification of competing endogenous RNAs in thyroid carcinoma. *Genes (Basel)* 9:44. doi: 10.3390/genes9010044

Maass, P. G., Luft, F. C., and Bähring, S. (2014). Long non-coding RNA in health and disease. *J. Mol. Med.* 92, 337–346.

Meng, J., Li, P., Zhang, Q., Yang, Z., and Fu, S. (2014). A four-long non-coding RNA signature in predicting breast cancer survival. *J. Exp. Clin. Cancer Res.* 33, 84–84.

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159.

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, 980–985.

Peng, L., Liu, F., Yang, J., Liu, X., Meng, Y., Deng, X., et al. (2020). Probing lncRNA–protein interactions: data repositories, models, and algorithms. *Front. Genet.* 10:1358. doi: 10.3389/fgene.2019.01346

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, J., Zhang, Z., Bao, S., Yan, C., Hou, P., Wu, N., et al. (2020). Identification of tumor immune infiltration-associated lncRNAs for improving prognosis and immunotherapy response of patients with non-small cell lung cancer. *J. Immunother. Cancer* 8:e000110. doi: 10.1136/jitc-2019-000110

Tseng, Y. Y., Moriarity, B. S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. *Nature* 512, 82–86. doi: 10.1038/nature13311

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, L., Xiao, Y., Li, J., Feng, X., Li, Q., and Yang, J. (2019). IIRWR: internal inclined random walk with restart for LncRNA-disease association prediction. *IEEE Access* 7, 54034–54041. doi: 10.1109/access.2019.2912945

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

Xiang, J., Zhang, J., Zheng, R., Li, X., and Li, M. (2021). NIDM: network impulsive dynamics on multiplex biological network for disease-gene prediction. *Brief. Bioinform.* doi: 10.1093/bib/bbab080

Xiang, J., Zhang, N. R., Zhang, J. S., Lv, X. Y., and Li, M. (2020). PrGeFNE: predicting disease-related genes by fast network embedding. *Methods* 192, 3–12. doi: 10.1016/j.ymeth.2020.06.015

Xiao, X., Wen, Z., Bo, L., Xu, J., and Gu, C. (2018). BPLLDA: predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* 9:411. doi: 10.3389/fgene.2018.00411

Xiao, Y., Lv, Y., Zhao, H., Gong, Y., Hu, J., Li, F., et al. (2015). Predicting the functions of long noncoding RNAs using RNA-seq based on Bayesian network. *BioMed Res. Int.* 2015:839590.

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

Xu, J., Zhu, W., Cai, L., Liao, B., Meng, Y., Xiang, J., et al. (2020). LRMCMDA: predicting miRNA-disease association by integrating low-rank matrix completion with miRNA and disease similarity information. *IEEE Access* 8, 80728–80738. doi: 10.1109/access.2020.2990533

Xu, N., Wang, F., Lv, M., and Cheng, L. (2015). Microarray expression profile analysis of long non-coding RNAs in human breast cancer: a study of Chinese women. *Biomed. Pharmacother.* 69, 221–227. doi: 10.1016/j.biopha.2014.12.002

Yan, C., Zhang, Z., Bao, S., Hou, P., and Sun, J. (2020). Computational methods and applications for identifying disease-associated lncRNAs as potential biomarkers and therapeutic targets. *Mol. Ther. Nucleic Acids* 21, 156–171. doi: 10.1016/j.omtn.2020.05.018

Yan, X., Bao, M. H., Luo, H. Q., Xiang, J., and Li, J. M. (2015). A meta-analysis of the association between polymorphisms in MicroRNAs and risk of ischemic stroke. *Genes* 6, 1283–1299. doi: 10.3390/genes6041283

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

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42, 353–372. doi: 10.1007/s11357-019-00106-x

Yang, J., Qiu, J., Wang, K., Zhu, L., Fan, J., Zheng, D., et al. (2017). Using molecular functional networks to manifest connections between obesity and obesity-related diseases. *Oncotarget* 8, 85136–85149. doi: 10.18632/oncotarget.19490

Yang, X., Gao, L., Guo, X., Shi, X., Wu, H., Song, F., et al. (2014). A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. *PLoS One* 9:e87797. doi: 10.1371/journal.pone.0087797

Yang, Z., Zhou, L., Wu, L., Lai, M., Xie, H., Zhang, F., et al. (2011). Overexpression of long non-coding RNA HOTAIR predicts tumor recurrence in hepatocellular

carcinoma patients following liver transplantation. *Ann. Surg. Oncol.* 18, 1243–1250. doi: 10.1245/s10434-011-1581-y

Yanofsky, C. (2007). Establishing the triplet nature of the genetic code. *Cell* 128, 815–818. doi: 10.1016/j.cell.2007.02.029

Yao, Y., Ji, B., Shi, S., Xu, J., Xiao, X., Yu, E., et al. (2019). IMDAILM: inferring miRNA-disease association by integrating lncRNA and miRNA data. *IEEE Access* 8, 16517–16527. doi: 10.1109/access.2019.2958055

Zhang, Y., Huang, H., Zhang, D., Qiu, J., Yang, J., Wang, K., et al. (2017). A review on recent computational methods for predicting noncoding RNAs. *Biomed. Res. Int.* 2017:9139504.

Zhang, Z., Hao, H., Zhang, C. J., Yang, X. Y., He, Q., and Lin, J. (2012). Evaluation of novel gene UCA1 as a tumor biomarker for the detection of bladder cancer. *Natl. Med. J. China* 92, 384–387.

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

Conflict of Interest: BJ was employed by Geneis Beijing Co., Ltd.

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

The handling editor declared a past co-authorship with one of the authors JX.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 August 2021
doi: 10.3389/fgene.2021.719099

Interpretable, Scalable, and Transferrable Functional Projection of Large-Scale Transcriptome Data Using Constrained Matrix Decomposition

Nicholas Panchy[1], *Kazuhide Watanabe*[2] *and Tian Hong*[1,3]*

[1] Department of Biochemistry and Cellular and Molecular Biology, The University of Tennessee, Knoxville, Knoxville, TN, United States, [2] RIKEN Center for Integrative Medical Sciences, Yokohama, Japan, [3] National Institute for Mathematical and Biological Synthesis, Knoxville, TN, United States

OPEN ACCESS

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

Reviewed by:
Seth H. Weinberg,
The Ohio State University, United States
Shihua Zhang,
Academy of Mathematics and Systems Science (CAS), China

***Correspondence:**
Tian Hong
hongtian@utk.edu

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

Received: *01 June 2021*
Accepted: *02 August 2021*
Published: *20 August 2021*

Citation:
Panchy N, Watanabe K and Hong T (2021) Interpretable, Scalable, and Transferrable Functional Projection of Large-Scale Transcriptome Data Using Constrained Matrix Decomposition. Front. Genet. 12:719099. doi: 10.3389/fgene.2021.719099

Large-scale transcriptome data, such as single-cell RNA-sequencing data, have provided unprecedented resources for studying biological processes at the systems level. Numerous dimensionality reduction methods have been developed to visualize and analyze these transcriptome data. In addition, several existing methods allow inference of functional variations among samples using gene sets with known biological functions. However, it remains challenging to analyze transcriptomes with reduced dimensions that are interpretable in terms of dimensions' directionalities, transferrable to new data, and directly expose the contribution or association of individual genes. In this study, we used gene set non-negative principal component analysis (gsPCA) and non-negative matrix factorization (gsNMF) to analyze large-scale transcriptome datasets. We found that these methods provide low-dimensional information about the progression of biological processes in a quantitative manner, and their performances are comparable to existing functional variation analysis methods in terms of distinguishing multiple cell states and samples from multiple conditions. Remarkably, upon training with a subset of data, these methods allow predictions of locations in the functional space using data from experimental conditions that are not exposed to the models. Specifically, our models predicted the extent of progression and reversion for cells in the epithelial-mesenchymal transition (EMT) continuum. These methods revealed conserved EMT program among multiple types of single cells and tumor samples. Finally, we demonstrate this approach is broadly applicable to data and gene sets beyond EMT and provide several recommendations on the choice between the two linear methods and the optimal algorithmic parameters. Our methods show that simple constrained matrix decomposition can produce to low-dimensional information in functionally interpretable and transferrable space, and can be widely useful for analyzing large-scale transcriptome data.

Keywords: dimensionality reduction, gene set analysis, EMT, single-cell 'omics, RNA-sequencing data

INTRODUCTION

Recent developments in RNA-sequencing technology have enabled the collection of large-scale transcriptome data at high speed. For example, single-cell RNA-sequencing (scRNA-seq) data of many biological systems have been accumulating rapidly and provide opportunities to gain insights into complex biological processes at both the systems level and the single-cell resolution. Together with the advances in experimental techniques, the recent development of computational methods, including those for dimensionality reduction, allow the visualization and analyses of high-dimensional transcriptome data in low-dimensional space. For example, *t*-distributed stochastic neighbor embedding (tSNE) and Uniform Manifold Approximation and Projection (UMAP) have been instrumental to tackling challenges in transcriptome data visualization and are widely used in biomedical research (Van der Maaten and Hinton, 2008; Stein-O'Brien et al., 2018; Becht et al., 2019; Luecken and Theis, 2019). However, dimensionality reduction methods usually do not provide low-dimensional space that is directly interpretable in terms of biological functions: while these approaches cluster related samples, the positioning of samples along the derived dimension may not correspond to the degree of any biological process even if a predefined gene set with similar functions is chosen before the reduction. In addition, the contribution or significance of individual genes related to the derived dimension cannot be accessed directly with these methods. The lack of interpretability of the dimensions makes it challenging to visualize and analyze the progression of the samples (cells) in known biologically functional space.

Existing methods for functional quantification, such as Z-score and Gene Set Variation Analysis (GSVA; Hänzelmann et al., 2013), are useful for obtaining "functional scores" with the expression levels of multiple genes involved in the same biological process. However, these methods do not have transferability in that the scoring systems obtained with one dataset cannot be used to analyze other datasets directly. This limits the utility of these methods in predicting the progress of new data points, and in studying the relationships between functional spaces in different experimental settings.

One example of cellular processes that contains crucial quantitative information is epithelial-mesenchymal transition (EMT). While extreme changes of cell fate and morphology occur in the classical form of EMT, recent studies with cancer and fibrosis showed that partial EMT involving intermediate states are prevalent, and it may be responsible for pathogenesis (Pastushenko et al., 2018). To quantify the degree of EMT in EMT-induced cell lines and tumor samples, several previous studies analyzed transcriptomic data and their projections onto epithelial (E) and mesenchymal (M) dimensions (Tan et al., 2014; George et al., 2017; Cursons et al., 2018; Chakraborty et al., 2020; Panchy et al., 2020; Hirway et al., 2021). Recently, scRNA-seq analysis has shown that the progression of EMT is highly dependent on inducing signals and cell types (Cook and Vanderhyden, 2020). However, it remains challenging to analyze rapidly accumulating transcriptome information on EMT for obtaining biological insights across multiple conditions. Improvement of methods for reducing dimensions of expression data in a functionally meaningful manner is necessary.

In this study, we used gene set filtered variants of both non-negative principal component analysis (gsPCA) and non-negative matrix factorization (gsNMF) to analyze progression of EMT in single cells at multiple timepoints. We show that these methods describe large-scale transcriptome data of multiple EMT stages in low-dimensional and functionally interpretable space. Taking advantage of the methods' transferability, we constructed dimensionality reduction models that can predict the stages of EMT with data from timepoints that were not used for model construction. We show that these linear methods can be used to compare functional spaces across multiple experimental conditions. Furthermore, we demonstrate the utility of our approach in visualizing drug responses in heterogeneous single cell data. With a validation scheme for rigorous testing, we provide recommendations for the choice of the methods and the parametric settings. Overall, our work provides a new toolbox for analyzing large-scale transcriptome data with efficient visualization and functional quantification.

RESULTS

Overview of Method and Performance Evaluation

The overall goal of our method is to find low-dimensional space of transcriptome data that has both biologically meaningful directionality and the ability to represent data points not used in the procedure to derive the space. This requires one or more preselected functional gene sets, which are readily available in publicly available databases such as Molecular Signature Database (Liberzon et al., 2011), and can be defined manually (**Figure 1**). We propose two linear approaches of matrix decomposition: gsPCA and gsNMF (see "Materials and Methods" section for details). Briefly, gsPCA finds the optimal component (projection) by maximizing the variance of the projected data points under the constraint that each functional gene has a non-negative loading value. For gsNMF, the gene-set-filtered transcriptome matrix is approximated by the product of two non-negative matrices, one of which represents a "meta" expression profile across samples, while the other represents the non-negative coefficients of the functional genes (the procedure for obtaining the number of components is described in **Supplemental Methods**). Following gsNMF, the leading component is selected for subsequent analyses (see "Materials and Methods" section). With either gsPCA or gsNMF, transcriptome data can be projected onto an axis whose direction unambiguously represents expression of the gene set and can be interrogated to reveal the contribution or association of individual genes in the set to scores along the axis.

To test the performance of gsPCA and gsNMF in capturing biological progression through functional space, we first used time-course datasets containing single cells treated with EMT-inducing signals for various periods of time (Cook and Vanderhyden, 2020). In addition to the biological importance of the stepwise progression in EMT (Pastushenko et al., 2018; Kröger et al., 2019), the time labels in the datasets allow

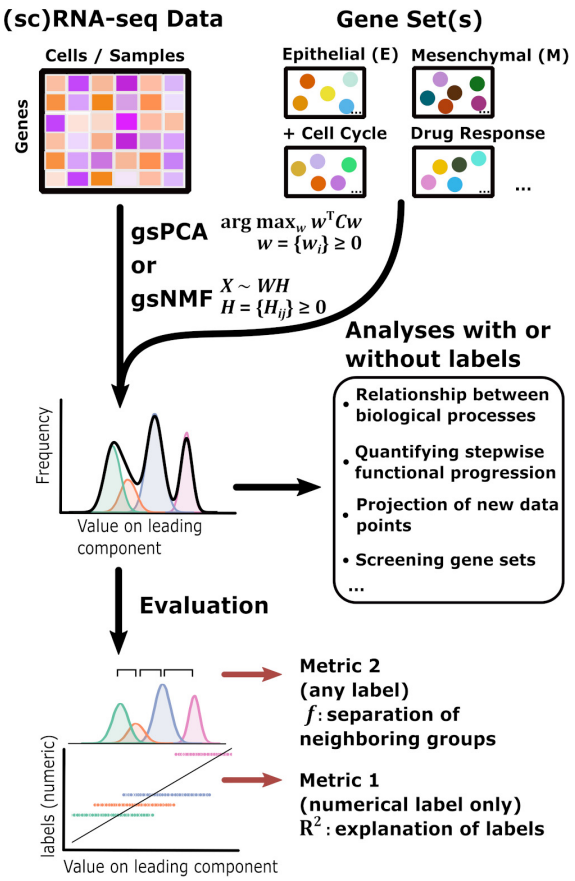


FIGURE 1 | Schematic of the gene set non-negative matrix factorization (gsNMF)/gene set non-negative principal component analysis (gsPCA) analysis process. A diagram of the analysis process used in this study beginning with input data in the form of sequencing data and gene sets. gsNMF/gsPCA is applied to this data to generate a functional scoring or space in the form of component scores (see "Materials and Methods" section for details). These scores can be used in two ways. First, without further data labels, scores can be used to look at relationships between or across biological processes beginning with low-dimensional visualization to identify trends and putative groups. Quantities such as the correlation between different functional scores can be computed for analysis. In addition, the transferable nature of these models means that they can be used to infer the position of new data points and contributions of individual genes, which allows assessment of their importance. Screening can be done both between gene sets and within gene sets. Secondly, when data labels are present, different metrics can be used to assess the performance of a functional score in terms of capturing variance: the common language effect size or *f*-probability can be used to evaluate how well the functional score separates two distinct populations while the variance explained or R^2 can evaluate how much of the variation of a numeric variable representing biological progression, such as time, that the functional score can explain across the data.

us to evaluate the performance of the functional projection. Specifically, we used two metrics for the evaluation: the coefficient of determination (R^2) for quantifying how well the projected values explain the time labels, and the common language effect size (f) for measuring the separation between two neighboring subsets of data with two labels (McGraw and Wong, 1992; See "Materials and Methods" section). The usage of R^2 is only possible when the labels are numerical, while f can be used with any type of label (**Figure 1**). Note that our overall goal is not clustering the data points. Instead, we aim to represent the progression along biologically meaningful axes. In addition, neither gsPCA nor gsNMF requires data labels for analysis. The two metrics are only used for evaluation. In later sections, we will show analyses with additional data sets in which labels are categorical and the biological processes are non-EMT.

gsPCA and gsNMF Capture Cell State Progression in Low Dimensional Functional Space

To show the performance of the proposed methods, we first used two signature gene sets whose high expressions represent the epithelial (E) and mesenchymal (M) states, respectively (Tan et al., 2014; Watanabe et al., 2019; Panchy et al., 2020). With the E and M gene sets, we first performed gsPCA and gsNMF on time-course single-cell transcriptomes of TGF-β-treated A549 cells using two components per model for each gene set (Cook and Vanderhyden, 2020). The two gene sets contain 179 and 114 genes, respectively, in the A549 data set. We then projected the single-cell data from the first five time points, which represent continuous EMT progression, onto the leading dimension for each gene set. This produced two-dimensional plots with dimensions that can be viewed as the progression of cell states in the epithelial and the mesenchymal spectrums (**Figures 2A,B**). We then compared the performance to two widely used approaches: Z-score and GSVA (**Figures 2C,D**). We found that gsPCA and gsNMF both better explained the overall variance of time across the first five time points of EMT progression (Adjusted R^2 = 0.46 and 0.48, respectively) than Z-score (Adjusted R^2 = 0.31) and GSVA (Adjusted R^2 = 0.08). Likewise, when considering neighboring time points, we found that E-scores tended to decrease and M-scores tended to increase with time of TGF-β treatment (**Figure 2E**), with both scores significantly separating all neighboring time points for gsPCA and gsNMF and yielding higher f probabilities than other methods in all but one case (E-scores at 3 vs. 7 days, **Figure 2F**). This suggests that gsPCA and gsNMF not only serve as visualization methods of functional space with defined gene sets, but also describe heterogeneous cell populations containing transitional information in a rigorous fashion. Between the two methods, we found the gsNMF performed better with regard to both overall variance (Adjusted R^2 0.48 vs. 0.46) and separating time points (**Figure 2F**) than gsPCA. However, gsNMF requires selecting the leading dimension based directly on correlation with time of EMT progression, suggesting that gsPCA may be more reliable in a purely unsupervised setting (see "Materials and Methods" section).

In the next few sections, we show various utilities of these linear methods based on their transferability and high-performance features. Because gsNMF gives the best performance with the A549 EMT data set, our discussion will focus on results obtained with gsNMF. The results using gsPCA, which had similar performance in all cases, are included in **Supplementary Materials**.

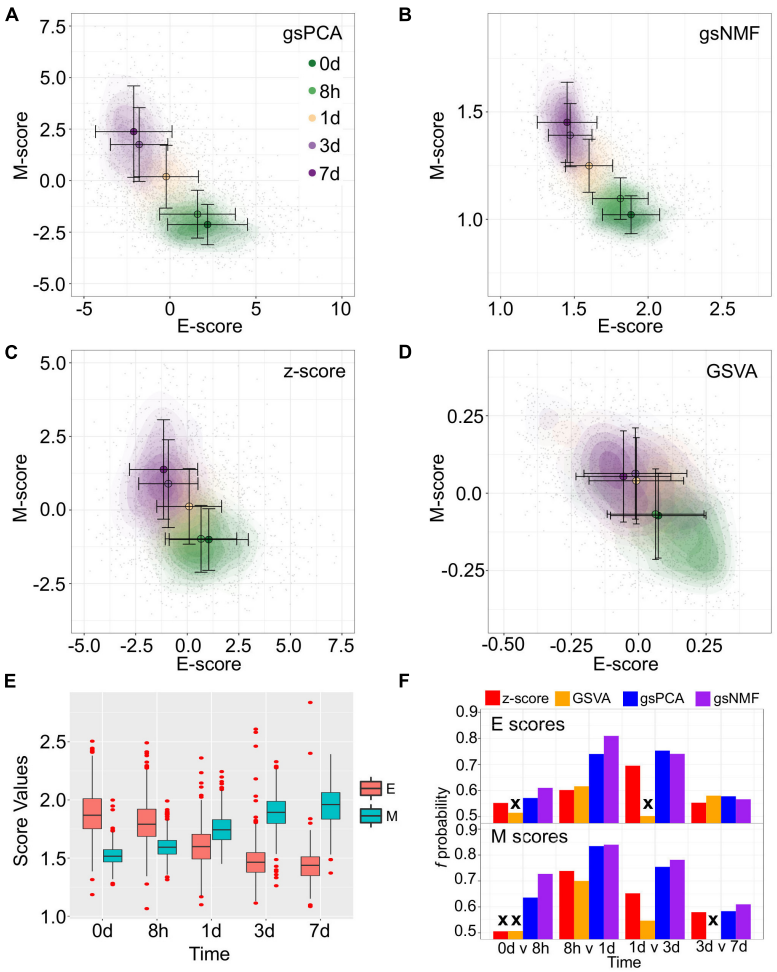


FIGURE 2 | Visualization of epithelial-mesenchymal transition (EMT) progression in TGF-β induced A549 cells by multiple scoring methods. **(A–D)** Contour plots of gene set scores of E (X-axis) and M (Y-axis) genes from four different scoring methods, gsPCA **(A)**, gsNMF **(B)**, z-score **(C)**, and GSVA **(D)**. Color indicates the time of TGF-β induction from 0 days (dark green) to 7 days (dark purple). Circles indicate the mean E- and M-score of samples from each time point and the associated error bars show the standard deviation. **(E)** A box-plot showing the distribution of E (red) and M (blue) scores across all five time points of TGF-β induction from the gsNMF model. Whiskers indicate the 1.5 inter-quartile range of each distribution while the red points indicate outliers beyond this range. **(F)** Bar chart of the *f* probability values for E (top) and M (bottom) scores between all consecutive pairs of time points. Color indicates the method used to produce the score: red is z-score, orange is GSVA, blue is gsPCA, and purple is gsNMF. Bars marked by an "x" indicates that the score did not significantly separate the samples from those time points (Mann–Whitney *U*-test, $p < 0.05$).

Prediction of Cell States With Data From New Conditions

The transferability of gsPCA and gsNMF methods allows the projection of new high-dimensional data points onto previously derived functional dimensions. Similarly, these methods can be used to derive functional dimensions with partial information of the biological process in terms of its stages. To show the predictive power of gsNMF, we removed samples from the 0-, 1-, and 7-day (including revertant) time points in the A549 EMT data (i.e., the start, middle, and the end of the continuous portion of TGF-β induction) and then performed the dimensionality reduction. We found that the low-dimensional functional space was robust with respect to the removal, regardless of whether the missing time point is in the middle of the progress or at the extremes (**Figures 3A–D**), such that when we projected the removed data points onto the space derived from a partial dataset, their positions were highly correlated with their positions when they were included in the data set [Pearson correlation coefficient (PCC > 0.95)]. However, while the inferred 1-day samples were similarly separable from samples in 8-h (f = 0.81 for E, 0.84 for M) and 3-day (f = 0.75 for E, 0.79 for M) time points, we observed reduced separability between the both inferred 0-day vs. 8-h (f = 0.52 for E, 0.63 for M) and 3-day vs. inferred 7-day (f = 0.53 for E, 0.55 for M) time points, with E-scores not significantly separating the first and the last time points (Mann–Whitney *U*-test, p = 0.13 and 0.16, respectively). We also applied the same inference procedure to samples which were exposed to a transient EMT-inducing signal and allowed to revert. However, because the 8- and 24-h reversion samples largely overlap with

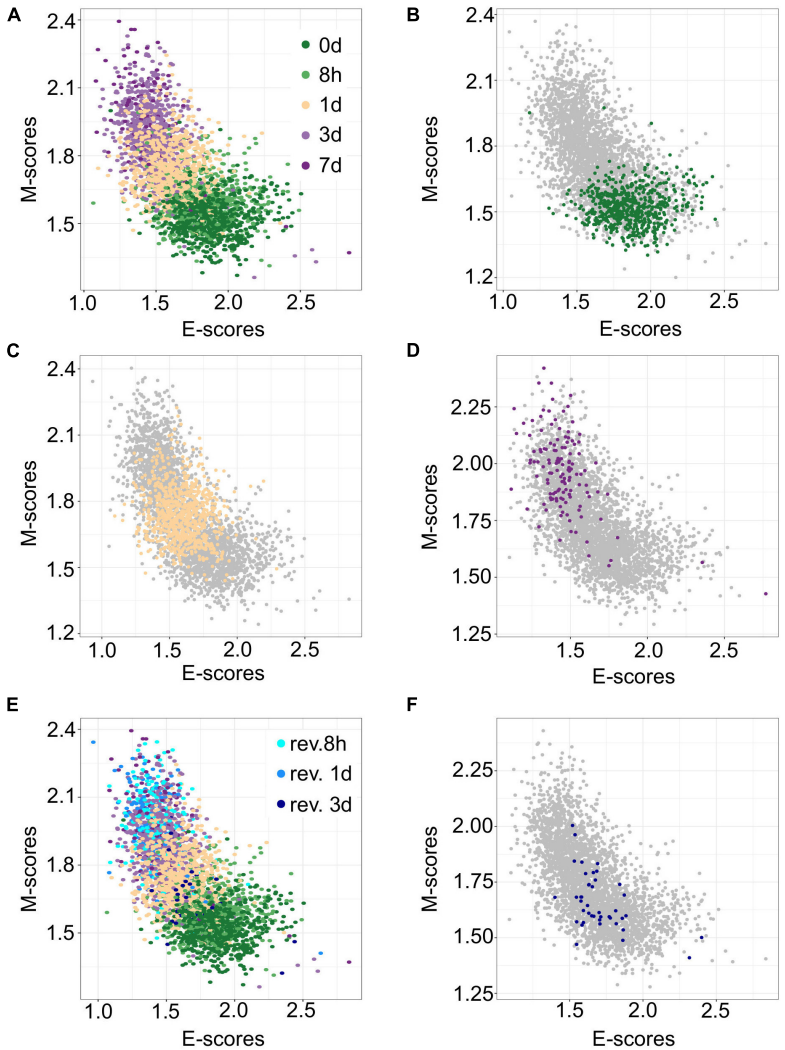


FIGURE 3 | Predicting A549 samples from specific time points using gsNMF. **(A)** Scatter plot of E (X-axis) and M (Y-axis) scores for all TGF-β induction samples using gsNMF. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(B–D)** Scatter plot of 0-day (green, **B**), 1-day (yellow, **C**), and 7-day samples (purple, **D**) inferred using a gsNMF model built with all other time points (gray). **(E)** A scatter plot of TGF-β induction samples with TGF-β reversion samples (i.e., 7 days induction followed by removal from TGF-β). Induction samples are labeled as in panel **(A)**, while reversion samples are colored blue, with darker shade indicating longer time since removal. **(F)** Scatter plot of 3-day reversion samples (dark blue) inferred using a gsNMF model built with all non-reversion time points (gray).

7-day (hence their removal for 7-day inference, **Figure 3E**), we focused on inferring 3-day reversion samples after performing dimensionality reduction on the data set without any reversion samples. We found that 3-day reversion samples were positioned in the middle of the EMT spectrum, consistent with when they were included in functional space construction (PCC = 0.99 for E and 0.98 for M, **Figure 3F**). Additionally, the inferred 3-day reversion samples were similarly separable from the 7-day samples (f = 0.91 for E, 0.91 for M) as when they were when included in functional space construction (f = 0.90 for M, 0.91 for M). We obtained similar results using gsPCA when inferring the position of samples from missing time points (**Supplementary Figure 1**), but neither E- nor M-scores significantly separated the end points (0-day vs. 8-h and 3- vs. 7 day). These results suggest that gsPCA and gsNMF can predict cell states of new data without retraining the model, and that these methods can be used to predict new cell states that have not been observed directly, though it may be difficult to separate these samples when they are positioned the edge of the spectrum and/or when the new samples are closely related to existing samples.

Using Functional Space Across Cell Lines

The transferability of gsNMF can be extended to data from different cell lines. We performed gsNMF on single-cell transcriptomes of TGF-β-treated DU145 from Cook and Vanderhyden (2020) using the same procedure as A549 and

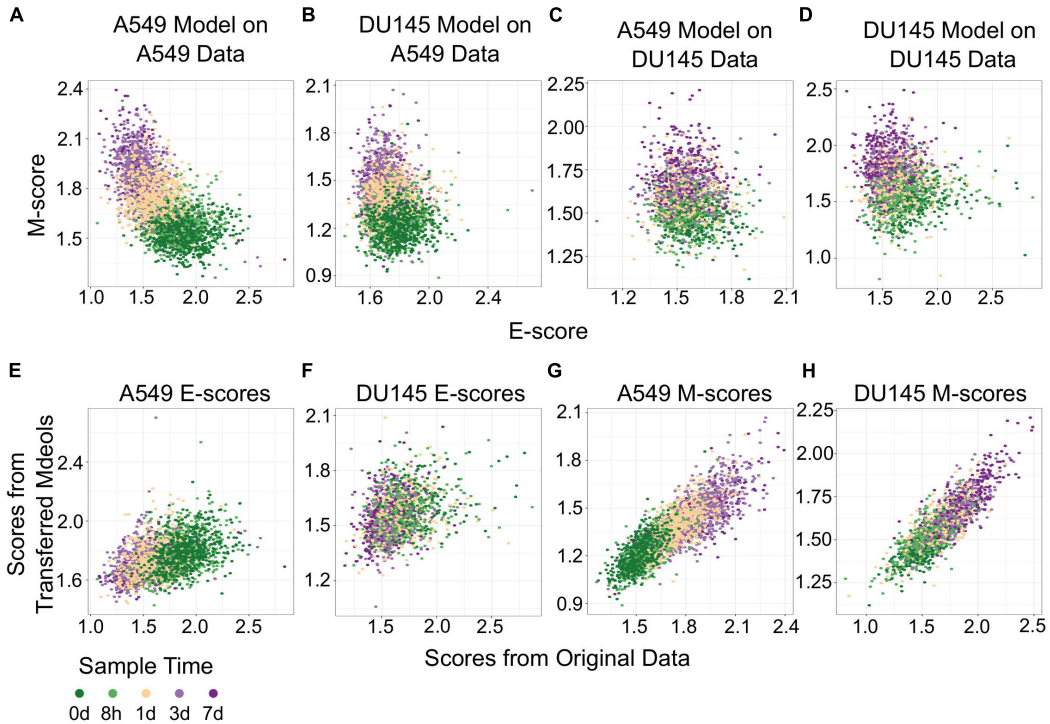


FIGURE 4 | Transferring gsNMF models between A549 and DU145 TGF-β induced samples. **(A–D)** Scatter plot of E (X-axis) and M (Y-axis) scores for different combinations of data and gsNMF model: **(A)** A549 model on A549 data, **(B)** DU145 model on A549 data, **(C)** A549 model on DU145 data, and **(D)** DU145 model on DU145 data. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(E,F)** Comparison of E-scores of samples from A549 **(E)** and DU145 **(F)** data. The X-axis is the E-score from using the model from the same data set (A549 on A549 and DU145 by DU145), while the Y-axis is the E-score from the opposite model (DU145 on A549 and A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(G,H)** Comparison of M-scores of samples from A549 **(G)** and DU145 **(H)** data. The X-axis is the M-score from using the model from the same data set (A549 on A549 and DU145 by DU145), while the Y-axis is the M-score from the opposite model (DU145 on A549 and A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple).

obtained a moderate explanation of variance in time of EMT progression using the E and M dimensions (Adjusted $R^2 = 0.31$). We then inferred the position of the five continuous time points in the A549 data set using the DU145 model and vice versa (**Figures 4A–D**). Transferred models (DU145 on A549 and A549 on DU145) were able to separate the individual time points, but overall performance decreased as they can explain only part of the variance seen in the original models (Adjusted $R^2 = 0.30$ for DU145 on A549 and 0.25 for A549 on DU145). Therefore, it was expected that the individual sample scores would be positively correlated between models along both the E (**Figures 4E,F**) and M dimensions (**Figures 4G,H**). However, while the correlations between all pairs of scores were significant (minimum $p = 2.7e{-}73$), the correlation between E-scores was weaker overall and worse for models of DU145 (PCC = 0.31) than models of A549 (PCC = 0.46). Comparably, the M-scores for both models of A549 (PCC = 0.84) and models of DU145 (PCC = 0.84) were more highly correlated and consistent between models. However, none of the sample scores between A549 and DU145 models were as correlated as inferred sample scores from missing point and the complete A549 model (PCC > 0.95). This suggests a reduced transferability across cells lines compared to within cells lines. In addition, across the data sets we used, changes along the M dimension were more consistent than the E dimension. We observed similar results using gsPCA, including M-scores being more correlated (PCC, A549 = 0.92, DU145 = 0.94) than E-scores (PCC, A549 = 0.76, DU145 = 0.72; **Supplementary Figure 2**). This is consistent with the fact that the same inducing agent was used across all cell lines, and also implies that inducing EMT in different cell types may yield more consistent changes in M genes compared to E genes.

Using Functional Space Across Experimental Conditions

In addition to predicting the locations in the functional space across cell lines, gsPCA and gsNMF can be used across both experimental conditions and cell types. To test the cross-condition transferability, we first used our low-dimensional functional EMT space for A549 and DU145 cells to analyze tumor transcriptomes measured with bulk RNA-seq (The Cancer Genome Atlas, TCGA). To perform the most comparable transfer, we used lung adenocarcinoma (LUAD) and prostate adenocarcinoma (PRAD) data, which correspond to A549 and DU145 in terms of tissue type. We considered transfers between both similar (projecting LUAD data by a A549-trained model,

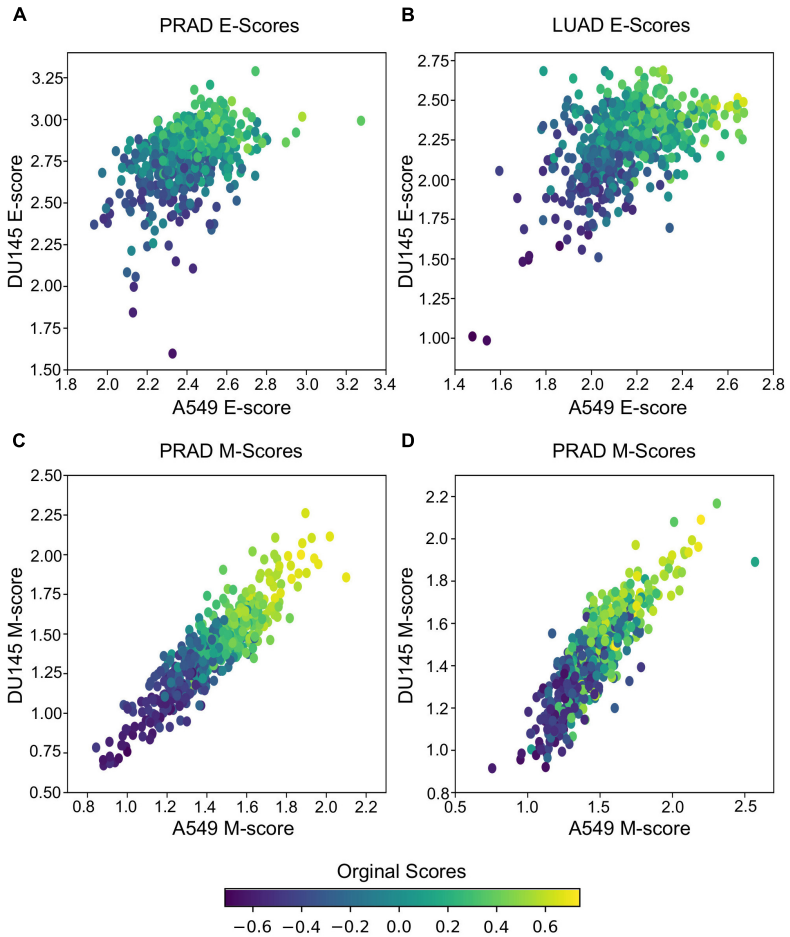


FIGURE 5 | Transferring gsNMF models to TCGA data. **(A,B)** Scatter plots of E-scores for PRAD **(A)** and LUAD **(B)** from transferring gsNMF models built on A549 (X-axis) and DU145 (Y-axis) data. The color of individual points indicates the original GSVA based E-score of the TCGA data set. **(C,D)** Scatter plots of M-scores for PRAD **(C)** and LUAD **(D)** from transferring gsNMF models built on A549 (X-axis) and DU145 (Y-axis) data. The color of individual points indicates the original GSVA based M-score of the TCGA data set.

and PRAD by a DU145-trained model) and dissimilar (LUAD by DU145 and PRAD by A549) cell types. We found that the low-dimensional functional space obtained with in-vitro data captured tumor sample heterogeneity in the EMT spectrum when compared to our previous GSVA analysis of the same data (**Figure 5**). Overall, the original E- and M-scores were significantly correlated with the A549 models in all cases (smallest p = 2.8e−34). Models from both cell lines showed similar correlation with the original GSVA scores, except in the case of PRAD scores, where the DU145 model was better correlated than the one built on A549 data (**Table 1**). We also observed that M-models built on A549 and DU145 data were more similar to each other than E-models, and we obtained similar results with gsPCA (**Supplementary Figure 3**), which showed greater overall correlation with GSVA scores, but the same pattern of reduced correlation for the A549 model of PRAD E-scores (**Supplementary Table 1**).

It should be noted that these results are partly due to higher average correlation of expression of EMT genes in bulk RNA-seq data (average PCC = 0.28 LUAD, 0.38 PRAD for all pairs of M-genes, average PCC = 0.18 LUAD, 0.10 PRAD for E-genes), compared to the scRNA-seq data (average PCC = 0.01 in all cases). This is expected given that bulk RNA-seq is derived from populations rather than individual cells, but as a result, the effect of differentially weighing individual genes across models and components within models is reduced. This would explain the stronger correlation of M-scores between A549 and DU145 derived models, as well as the reduced performance of A549 on PRAD E-gene data, which is the most variable bulk RNA-seq data set. Yet, as the same time, this would suggest the variance present in PRAD bulk RNA-seq data is more similar to the model built on DU145 scRNA-seq data, than scRNA-seq data from a more dissimilar background. This also has implication for comparing multiple model components as they tend to be more similar in the bulk RNA-seq model despite if they were anti-correlated (gsNMF) or relatively uncorrelated (gsPCA) in the original scRNA-seq model or other scRNA-seq data (see **Supplementary Table 2**). Nonetheless, we have shown that the transferred models are, overall, consistent with the prior analysis of TCGA data and

detected the expected variance in bulk RNA-seq data when it is present.

Using Functional Space Across Spatial and Temporal Progression

We next examined if gsNMF can produce transferrable models that reveal both spatial and temporal progression of EMT. Using single-cell RNA-seq, McFaline-Figueroa et al. (2019) previously found that epithelial cells exhibit an E to M spectrum from the inner position of a colony to the outer position. This dataset that contains binarized identities (inner and outer) obtained with macro-dissection (defined as spatial EMT data) from two experiments, one in which cells were allowed to migrate without external induction of EMT (Mock), and one in which EMT was induced with TGF-β (TGF-β). Since there are only two populations in this data set, the leading dimension for E- and M-scores was chosen to maximize the separation based on the f probability. Overall, three analyses were performed for each data set: spatial data with its own gsNMF model, spatial data with the model from the other spatial data set (TGF-β on Mock and Mock on TGF-β), and spatial data with A549 time series model (**Figure 6**). As with our previous results, the best separation of inner and outer data points was observed when Mock (f = 0.61 for E, 0.73 for M) and TGF-β (f = 0.77 for E, 0.82 for M) data sets had their own model applied to them. However, for Mock data, the TGF-β model (f = 0.64 for E, 0.69 for M) outperformed the A549 model (f = 0.45 for E, 0.60 for M) on both dimensions and, in fact, the E dimension of the A549 model did not effectively separate inner and outer points in the Mock data (p = 0.99). In comparison, the Mock model better separated TGF-β inner and out points in the E direction (f = 0.68 for E, 0.63 for M), while the A549 model better separated them in the M direction (f = 0.59 for E, 0.76 for M). gsPCA models gave similar results, including the A549 model yielding better performance along the M-dimension (f = 0.76) for TGF-β data than Mock data (f = 0.68; **Supplementary Figure 4**).

The fact the A549 model better separated TGF-β spatial points along the M dimension than the Mock model, but did not outperform TGF-β on the Mock model suggests that there is conserved TGF-β induced M-gene expression regardless of context. To explore the basis of this similarity in M-scores, we compared the coefficient matrices (H, see "Materials and Methods" section) between Mock, TGF-β, and A549 gsNMF models, which represent the weights of individual genes along the components. We found little correlation between A549 and spatial E-gene coefficient values for the lead dimension (PCC = −0.02. p = 0.82 for Mock; PCC = 0.06, p = 0.57 for TGF-β), however, while there was also little correlation between A549 and spatial M-gene coefficient values for the Mock model (PCC = 0.02, p = 0.83) there was significantly positive correlation for the TGF-β model (PCC = 0.48, p = 8.8e−7). Additionally, we examined which genes were in the top 10th percentile of coefficient values across models and found that the A549 and TGF-β models share six M-genes (FN1, LGALS1, SERPINE1, TAGLN, TPM2, and VIM), compared to three E-genes (ELF3, PERP, and SLPI). Furthermore, another four E-genes (AREG, KRT18, KRT8, and NQO1) were in the top 10th percentile of A549 E-gene coefficient values, but the bottom 10th percentile of TGF-β E-gene coefficient values. We observed similar results from gsPCA, finding significant correlation of loading values only between A549 and TGF-β M-models (PCC = 0.61, p = 5.8e−11) with many of the same genes in the top 10th percentiles of both models (FN1, TPM2, VIM, TAGLN, GLIPR1, and LGALS1). Notably, the M-genes with high coefficient values in both A549 and TGF-β models across both A549 and TGF-β models are key regulators/inducers of EMT (FN1, LGALS1, and VIM; Mendez et al., 2010; Griggs et al., 2017; Zhu et al., 2019) or specific activators of migratory behavior in epithelial/cancer cells (TAGLN, TPM2; Lee et al., 2010; Shin et al., 2017). Conversely, while KRT8 and KRT18 are considered epithelial cytokeratins (Tomaskovic-Crook et al., 2009), both of these genes undergo an initial increase in expression in the A549 time-course (**Supplementary Figure 5**), compared to largely unaltered distributions across the inner and outer samples of migration data. This is consistent with previous observations that, both KRT8 (Wang et al., 2020) and KRT18 (Zhang et al., 2019) are over-expressed/aberrantly expressed in certain human cancers and such expression is associated with cancer progression/poor-prognosis. This potentially reflects intermediate EMT states caused by full or partial arrest of the process at an early timepoint, independent of the resulting migratory potential of the cells. Coefficient values for all genes in each model can be found in **Supplementary Table 3**.

Together, these results suggest a coherence of the progression of the EMT program in both the spatial and temporal context with regard to M-genes, while E-gene progression appears to be more sensitive to context, being only transferable between the two spatial data sets. The coefficient values of genes across both contexts offers insight into the difference in transferability between E and M models: high scoring M-genes across both contexts constitute important drivers of EMT/migration, suggesting common regulatory mechanisms, while the differential expression of KRT8 and KRT18 across time, but not space, suggests E-gene expression can be sensitive to biological context. Finally, these results highlight the usefulness of the transferability of gsPCA and gsNMF outside of a time series context, where performance may need to be evaluated on discrete groups.

TABLE 1 | Pearson correlation coefficients of E and M scores between GSVA, A549, and DU145 models of TCGA data.

TCGA data set	GSVA vs. A549	GSVA vs. DU145	A549 vs. DU145
PRAD E-genes	0.49	0.72	0.43
LUAD E-genes	0.72	0.71	0.52
PRAD M-genes	0.85	0.84	0.90
LUAD M-genes	0.62	0.65	0.86

Characterizing Relationships Among Multiple Functional Spectrums

To test the capacity of gsNMF to infer functional spaces across a broader range of gene sets and data, we first returned to

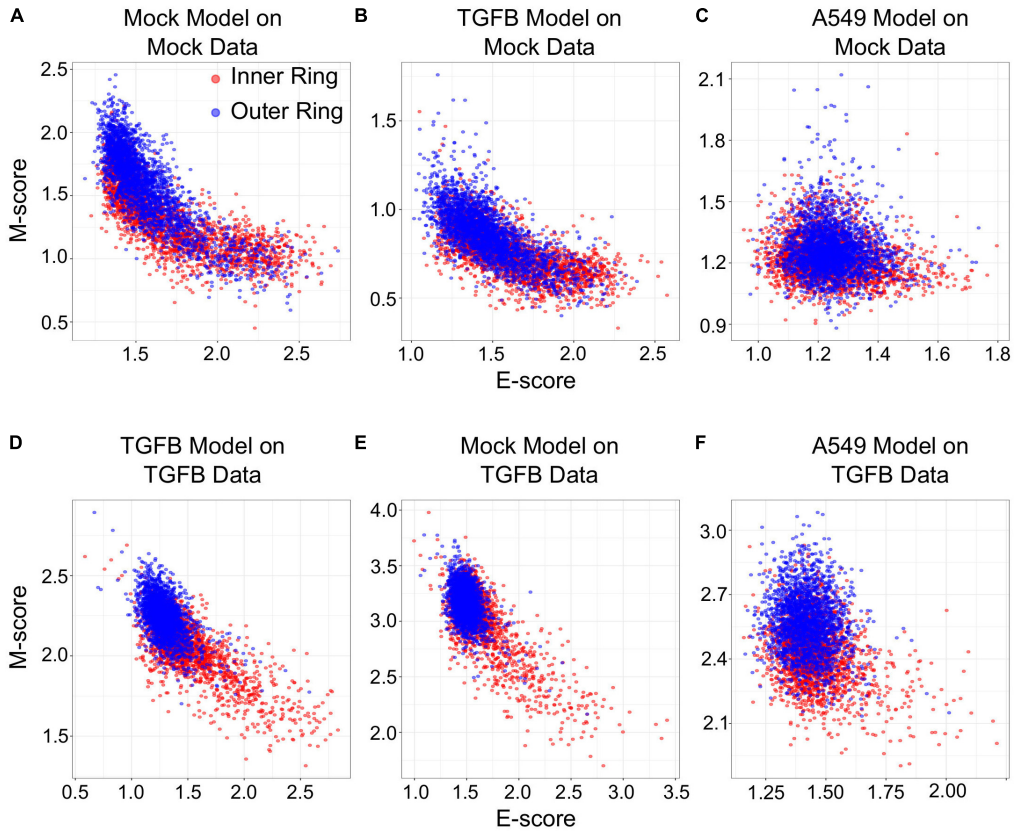


FIGURE 6 | Transferring gsNMF models between temporal and spatial data sets. **(A–C)** Scatter plots of E (X-axis) and M (Y-axis) scores for Mock spatial data from gsNMF models built on different data sets: Mock spatial data **(A)**, TGF-β induced spatial data **(B)**, and TGF-β induced A549 temporal data **(C)**. The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue). **(D–F)** Scatter plots of E (X-axis) and M (Y-axis) scores for TGF-β spatial data from gsNMF models built on different data sets: TGF-β induced spatial data **(D)**, Mock spatial data **(E)**, and TGF-β induced A549 temporal data **(F)**. The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue).

the A549 data set and examined the expression changes of multiple gene sets across EMT progression. Taking advantage of the high-efficiency of this method, we began with 5455 C2 curated gene sets from the Molecular Signature Database (see "Materials and Methods" section) and applied a gsNMF model to A549 data for each. For simplicity, we used a two-component model, but we applied stricter convergence and selection criteria because of the diversity of gene set size and coverage by the data set (see "Materials and Methods" section). Overall, 867 gene sets (15.9%) had a leading dimension whose magnitude of correlation (PCC) was > 0.5 (**Supplementary Table 4**). As such, we expected that functional spaces constructed from highly correlated gene sets should show similar results to our original E vs. M functional space.

To construct unambiguous functional spaces, we initially focused on pairs of up/down regulated gene sets where the leading dimensions had a high magnitude of correlation (PCC), but opposite sign, in order to emulate our original E/M model of EMT progression for A549 (**Figure 7A**). For example, two pairs of gene sets, up regulation or down regulation in response to KRAS knockdown (SWEET_KRAS_TARGETS, **Figure 7B**) and up regulation or down regulation in low-malignancy ovarian cancer relative to control (WAMUNYOKOLI_OVARIAN_CANCER_LMP, **Figure 7C**), yielded functional spaces similar to E and M genes (**Figure 7A**) and captured a similar amount of variance explained among non-revertant cells (R^2 = 0.48 and 0.49, respectively). Furthermore, the results suggest that EMT progression is correlated with expression of genes normally repressed by KRAS, a pro-proliferation signal, and anti-correlated with the expression of genes associated with tumorigenic, but non-metastatic ovarian cancer, consistent with the idea of the E state of EMT being pro-proliferative and the M state being pro-migratory. However, not all pairs of gene sets provide well defined functional spaces: for example, the gene set down-regulated in metastatic vs. non-metastatic head and neck tumors (RICKMAN_METASTASIS_DN) produced a strong anti-correlated leading dimension (PCC = −0.63), but the leading dimension of the up-regulated variant has a far smaller magnitude of correlation (PCC = 0.34). However, combining the metastatic down-regulated gene set with another correlated gene set, genes silenced during angiogenesis (HELLEBREKERS_SILENCED_DURING_TUMOR_ANGIOGENESIS, PCC = 0.66), generated a functional space of EMT

progression competitive with E and M genes (**Figure 7D**, R^2 = 0.48). As such, functional space constructs need not be confined to reciprocal or connected gene sets, though this does not excluded the possibility of an underlying, common genetic basis between these functional spaces. Nevertheless, the divergent origins of the gene sets in terms of the biological processes they represent demonstrates the breadth over which the functional significance of variation can be explored using this methodology.

To move beyond EMT associated data and gene sets, we next used gsNMF to analyze data from McFarland et al. (2020) which is composed of 7,245 cells with heterogeneous origins treated with trametinib for 3, 6, 12, 24, or 48 h as well as an untreated control (0 h). Because this data set mixes 24 cell lines from several different origin tissues and focuses specifically on the response to a cancer drug, we focused our exploration of functional spaces on 1,022 gene sets derived from the C6 database from Molecular Signature Database as well as the drug resistant genes identified by Wang et al. (2017) and their overlapping KEGG pathways and GO terms (see "Materials and Methods" section). Overall, 57 gene sets (5.6%) had a leading dimension whose magnitude of correlation (PCC) was >0.5 and relaxing this threshold to >0.4 yielded only 200 (19.6%) gene sets, suggesting that the explained temporal variance in this data set is lower than that obtained with A549 (**Supplementary Table 5**). Nevertheless, using positive regulation of gene expression (GO:0010628) and negative regulation of gene expression (GO:0010629), we were able to a functional space of trametinib response with similar performance (R^2 = 0.30, **Figure 8A**) to our model of EMT progression in DU145 data (R^2 = 0.31). Additionally, a number of oncogenic signatures which were positively correlated with trametinib response, though there were no up/down regulated pairs that with leading dimensions in opposed directions. Instead, we selected two oncogenic signatures, down regulation in response to KRAS over-expression (KRAS.600_UP.V1_DN) and down regulation in response to LEF over-expression (LEF1_UP.V1), whose leading dimension were strongly correlated with trametinib response (PCC = 0.54). We then took the negatively correlated component of the corresponding up regulation gene sets models (KRAS.600_UP.V1_UP and LEF_UP.V1_UP), even though the magnitude of the positively and negatively correlated components was similar (difference the absolute value of PCC $\leq$ 0.005). This process gave functional spaces which improved variance explained over the previous gene regulation model (R^2 = 0.35 and 0.36, respectively, **Figures 8B,C**). Together, these results suggest suppression of gene expression in general and of oncogenes specifically in response to trametinib treatment, consistent with the results in McFarland et al. which observed greater enrichment of KRAS responsive genes among down-regulated genes in later time points relative to earlier ones. As with A549 data, we were also able to combine distinct functional sets, response to drug (GO:0042493, PCC = 0.58) and positive regulation of cell cycle (GO:0045787, PCC = −0.49) to explain an comparable amount of variance in expression as the reciprocal onco-gene sets (R^2 = 0.36, **Figure 8D**). As such, while the amount of variance we can capture with our models is dependent on the data set, our approach overall is capable of producing functional spaces that broadly characterize variance in expression across diverse data and gene sets.

DISCUSSION

Previous methods that aimed to address the challenges of visualizing single-cell data in functional space were primarily based on weighted sum of expression values or Kolmogorov–Smirnov test with full datasets (Hänzelmann et al., 2013; DeTomaso and Yosef, 2016). These methods are useful to analyzing samples with functional gene sets, they do not provide transferability which is essential for predicting cell states with existing models and new data. We showed that constrained linear transformation enables good performance in depicting cell states with straightforward interpretation in functional space and satisfactory efficiency. While more sophisticated methods such as deep generative models have potentials to address similar problems, current methods primarily focus on the interpretability in terms of inter-sample distances in low dimensions rather than the dimensions themselves (Ding et al., 2018; Lopez et al., 2018), and we expect that the gsPCA and gsNMF methods are more efficient than models based on non-linear connectivity.

Factorization approaches like PCA and NMF have previously been applied to the problem of gene expression, with NMF in particular having been used to deconvolute expression patterns scRNA-seq data sets (Chen and Zhang, 2018; Fujita et al., 2018; Min et al., 2018; Kotliar et al., 2019; Zhang and Zhang, 2019), but these approaches have primarily focused on the unsupervised clustering of samples and/or for *de-novo* module discoveries at relatively high dimensionality (n > 10). In contrast, our approach suggests there is a utility in applying these factorization approaches to interrogating the relationship between known gene modules and data with implicit structure and/or separable populations of samples, particularly when assessing a single biological process (EMT) across multiple contexts (e.g., cell line, time and space), such that the simplicity of low-dimension space (n = 2) can be leveraged for visualization and analysis.

In this work we have found that conserved EMT gene expression signatures can be used to describe stages of EMT in multiple cell lines (e.g., A549 and DU145), and these signatures not only capture the subpopulation heterogeneity resulting from differential times of treatment with EMT-inducing signals such as TGF-β, but also reflect the EMT program driven by spatial heterogeneity with cell populations (McFaline-Figueroa et al., 2019). These results are consistent with the existence of conversed EMT program across cell lines (Cook and Vanderhyden, 2020), but do not contradict the idea of context specific expression as models trained and applied to the same data set always explained more variance in EMT progression. The coexistence of a common EMT signature and context specific expression is further supported by the observation that M-scores were more consistent and better separate data across different contexts of EMT than E-scores, and the related observation that M-gene component values were correlated across spatial and temporal models, while E-genes were not. This suggests that M-gene induction by TGF-β is consistent across cellular contexts, while

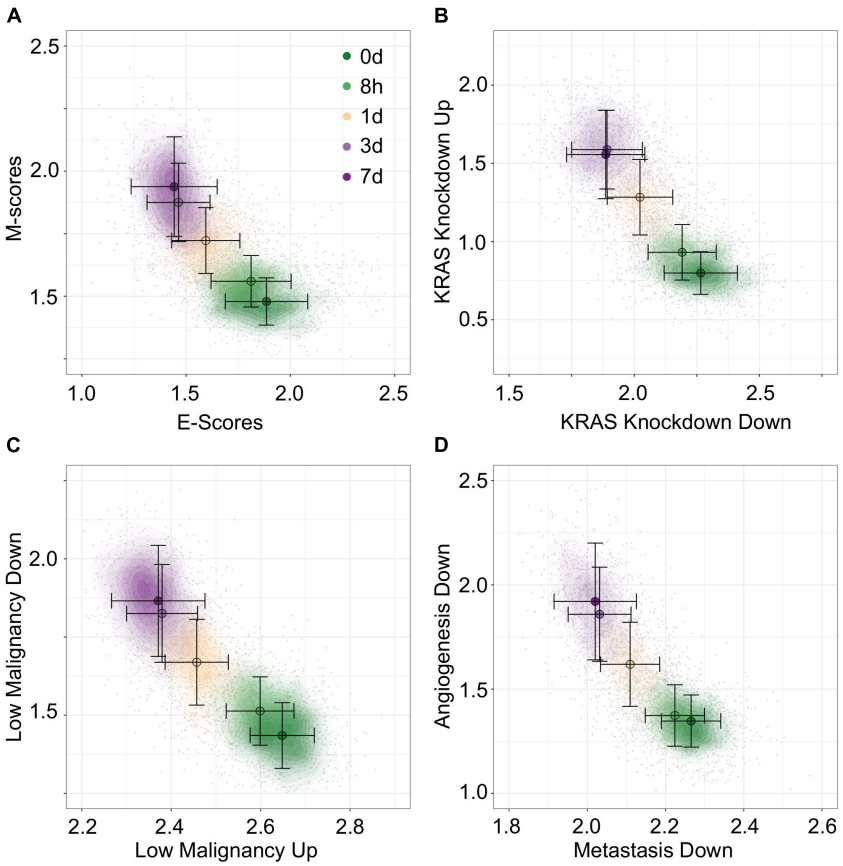


FIGURE 7 | Visualization of EMT progression in TGF-β induced A549 cells by multiple gene sets. **(A–D)** Contour plots of A549 functional space generated using gsNMF with different gene sets: E vs. M **(A)**, KRAS knockdown up and down **(B)**, non-malignant ovarian cancer up and down **(C)**, and metastasis downregulation vs. angiogenesis downregulation **(D)**. Color indicates the time of TGF-β induction from 0 days (dark green) to 7 days (dark purple). Circles indicate the mean gene set score of samples from each time point and the associated error bars show the standard deviation.

changes in E-gene expression are more variable, possibly due to greater sensitivity to cell line, environmental context, or other initial conditions effecting the cell prior to induction.

The transferability of models across EMT context indicates the synergy between spatial arrangement of cells and external signals (e.g., TGF-β) in determining the stages of EMT. In addition, we found that the functional dimensions obtained with TGF-β can serve as reasonable approximations for the positioning of tumor transcriptomes in the EMT spectrum. Similar to the EMT spectrum, many biological processes involve stepwise changes of gene expression programs. A possible mechanism underlying these non-binary programs is the feedback-driven formation of stable intermediate cell states (Yui and Rothenberg, 2014; Ye et al., 2019). With the rapid advances of the single-cell technology, transcriptome-wide gene expression data will become available for more biological systems. We expect that our functional projection methods can be widely useful for visualizing and analyzing these data. In particular, the transferability of the models can be a powerful feature for interrogating the relationships among different experimental conditions and cell types.

MATERIALS AND METHODS

Gene Expression Data Sources

Single-cell RNA-sequencing data and meta data for A549 and DU145 cell lines were obtained from Cook and Vanderhyden (2020). In brief, we obtained pre-processed SeuratObjects for A549 and DU145 TGF-β as .rds data files and extracted expression data for E, M, all genes using the ScaleData function from Seurat to regress out mitochondrial gene expression, total unique reads in a sample, cell cycle gene expression, and batch effects as well as center and scale each data set across genes (Stuart et al., 2019). For McFaline-Figueroa et al. (2019) spatial data we obtained aggregated count data from GEO in the form of a pre-processed .cds file (GSE114687). We then dropped genes expressed in less than 50 cells (∼1% of each data set) from Mock and TGFB1 and split samples into Mock and TGFB1 subset for subsequent steps. Because we planned to compare models from these data to those from A549 and DU145, we followed the preprocessing procedure from Cook and Vanderhyden: we normalized the Mock and TGFB1 data sets independently in Seurat using the NormalizeData function

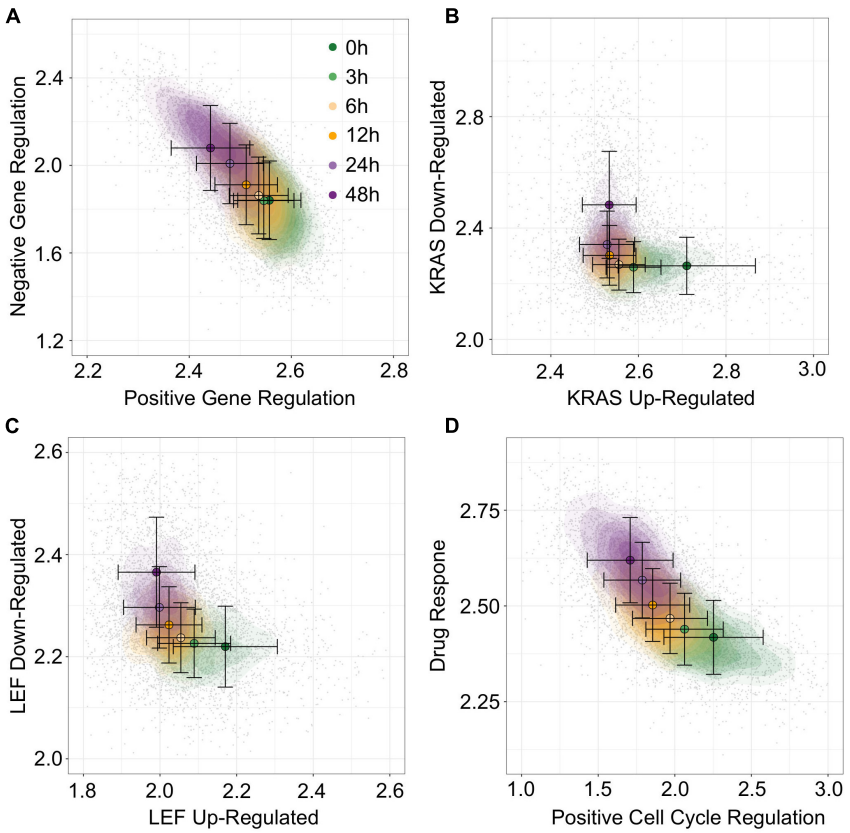


FIGURE 8 | Visualization of trametinib treatment data by multiple gene sets. **(A–D)** Contour plots of trametinib treatment functional space generated using gsNMF with different gene sets: positive vs. negative gene regulation **(A)**, KRAS overexpression up and down regulation **(B)**, LEF overexpression up and down regulation **(C)**, and positive cell-cycle regulation vs. drug response **(D)**. Color indicates the time of trametinib treatment from 0 h (dark green) to 48 h (dark purple). Circles indicate the mean gene set score of samples from each time point and the associated error bars show the standard deviation.

and then used ScaleData to regress out mitochondrial gene expression, total unique reads in a sample, and cell cycle gene expression as well as scale each data set across genes. Finally, we obtained Cell Ranger output for trametinib time-course data from McFarland et al. (2020) and processed it in R using the Read10X function. We dropped the DMSO time course, and used the Untreated samples as time 0 as well as annotations from the original manuscript to eliminate low quality cells and then filtered genes expressed in less than 73 cells (∼1% of the data set). Pre-processing was done in Seurat as with using NormalizeData and ScaleData as previously described, except that we additionally regressed out the effect of each different cell line used in the experiment, but did not regress out cell-cycle gene expression as the original manuscript suggested that cell cycle disruption may be induced by trametinib treatment.

TCGA bulk RNA-seq data was obtained from TCGAbiolinks (Colaprico et al., 2016; McFaline-Figueroa et al., 2019). Raw counts were transformed to $\log_2$TPM with a pseudo-count of 1 using gene models for the hg38 annotation of the human genome obtained from RefSeq (O'Leary et al., 2016).

Non-negative PCA and NMF

Gene set non-negative principal component analysis uses the non-negative approach to PCA pioneered by Sigg and Buhmann (2008). In brief, the vector of weights, w, used to define the first principal component of PCA is defined such that it maximizes the variance of the first component, i.e.:

$$\arg\max_w \; w^{\mathrm{T}} C w$$

Where C is the covariance matrix of the original data set X and w is unit vector ($||w||^2 = 1$). In our case, X is an m by n matrix of expression values where m is the number of samples and n is the number of genes in the selected gene set. This method for determining w can be treated as an expectation maximization problem where the original data is projected using the current estimate of w ($y = Xw_t$) and this projection is used to re-estimate w using the following minimization step:

$$w_{t+1} \quad = \arg\min_w \sum_{n=1}^{N} ||x_n - y_n w||_2^2$$

Where x_n are the rows of the original data and y_n are the rows of the projected data (Sigg and Buhmann, 2008). This expectation-maximization formulation allows additional constraints on w, including forcing the component values to be non-negative. Note that the non-negativity constraint applies only to the weight components such that negative scores can still exist if there are negative values in underlying data, such as those produced

by centering expression data to zero which we did for all gsPCA inputs. Subsequent components are calculated in the same way, under the constraint that they are orthogonal to the preceding ones.

NMF involves factorizing the original data matrix of non-negative values into two matrices whose product estimates the original data, i.e.:

$$X \cong WH$$

Where X is the original matrix (m by n), W is the basis or features matrix (m by p), and H is the coefficient matrix (p by n), such that m is the number of rows in the original matrix (samples in our case), n is the columns (genes in our case), and p is the number of components used in the factorization. In addition to factorizing X, NMF naturally clusters the elements of the original data: W represents the "centers" of column clusters whose memberships is determined by the relative coefficient values in H, and vice versa with H representing the centers of row clusters determined by W (Brunet et al., 2004). Because the original matrix is constrained to being non-negative, we subtracted the minimum of value of the scaled expression matrix from all values to create a non-negative input matrix. As a consequence, the values of the W and H matrices must likewise be non-negative such that product is non-negative.

Implementation of Dimension Reduction Approaches

We implemented non-negative PCA in R using the *nsprcomp* function (with the option nneg = TRUE) from the package of the same name (Sigg and Buhmann, 2008). We used the standard convergence parameters for the algorithm as these produced consistent principal components across multiple runs and different number of components. This is to be expected as *nsprcomp* greedily maximizes the variation explained by each component in order. For gsNMF, we used the Scikit-learn implementation of NMF (Pedregosa et al., 2011). To optimize convergence criteria, we performed a cross-validation analysis of A549 data and found that a two-component model fit with a tolerance of 1e–6 and a max of 500 iterations gave the best results (see **Supplemental Methods** and **Supplementary Figure 6**). We also tested ten random seeds of the two component A549 model on the full data to confirm that consistent results were given (average PCC of dimensions > 0.99). We tested ten random seeds against the other data sets to tune the convergence parameters, raising maximum iterations to 2,500 and tolerance to 1e–9 if the initial parameters did not yield consistent results (i.e., average PCC of dimensions > 0.99). GSVA and Z-score methods were implemented using the GSVA package in R (Hänzelmann et al., 2013).

Unlike GSVA and Z-scores methods, which produce a single score per gene set, gsPCA and gsNMF both produce multiple sample level scores in the form of principal component scores (wX, gsPCA) or the columns of the features matrix (W, gsNMF), while the corresponding loading values/weights (w, gsPCA) or coefficient matrix (H, gsNMF) represent gene level scores (i.e., gene importance). Therefore, we need to choose one of these sample level scores as a "leading dimension" to represent each gene set in functional space. For gsPCA, we used the first principal component as this represents the direction of greatest variance for gene expression in that gene set. For gsNMF, we used the magnitude of correlation between the columns of the transform matrix and the sample metric that best represented progression in EMT (i.e., time for A549 and DU145 data). For E and M genes, we also required the sign of correlation to match the expected change in E and M genes during EMT (i.e., picking the greatest negative PCC for E and the greatest positive PCC for M). For our spatial EMT data, where there were only two populations, f probability was used instead (see below), but with the same constraint on the direction of E and M dimensions (i.e., higher M scores for outer samples and higher E scores for inner samples).

Evaluating Functional Spaces

To evaluate a functional space, we used two metrics. First, if the sample data had an associated time variable, we created a model of time as a linear function of the two axes of the functional space (time $\sim X + Y$) and calculated the coefficient of determination, which is the percent of overall variance in the dependent variable explained by the independent variables (Adjusted R^2). Second, to evaluate the ability of functional space to separate distinct populations, such as neighboring time points or spatial locations, we used the common language effect size (f), which is the probability that a value or score randomly sampled from one population will be larger than a random score from the other. This metric is advantageous because we can calculate it from the test statistic of Mann–Whitney U-test, which also provide a measure of significance, and is related to the area under of the receiver operating curve (AUC-ROC), which is commonly used to asses classification algorithms. Additionally, since the f probability is reciprocal, the choice which population we want to be larger is arbitrary, so for EMT we can chose to calculate the f probability such that the larger population is the more progressed for M and less progressed for E. Therefore, a higher probability of f always indicates better correspondence with EMT progression in our results.

Inference and Model Transfer

To infer the position of new data in functional space for gsPCA, we multiplied the new data directly by loading vectors (also known as weights, w) of the E- and M-scores. For gsNMF, we used the Scikit-learn "transform" method which transforms the input data according to the fitted model (i.e., it fixes the coefficient matrix, H, and generates a new feature matrix, W). In both cases, we used the same leading dimension for inference as in the original model. For inferring missing data points, no further steps were required as the new data always had the same coverage of the E and M gene sets as the original. However, for transferring models across cell-line, TCGA, and spatial data, we first had to determine the common set of genes between the two data sets. Common genes were then used to filter the weight vectors for gsPCA and to refit the model on the original data using the common subset of genes for gsNMF. The data set that was the target of the transferred model was then subset by the same common set of genes and inference was done as described previously. Transferred models were assessed against the new

data set using the same approaches as the original models, but relationship between E/M-scores and gene loading/coefficient values between models were assessed by PCC.

Multi-Gene Set Evaluation

C2 gene sets were obtained from the Molecular Signature Database (version 7.1)[1] (Subramanian et al., 2005; Liberzon et al., 2011, 2015). gsNMF was performed as described for EMT gene sets expect that we increased the iteration (2,500) and convergence threshold (1e–9) of the NMF algorithm to ensure consistent results across the gene sets which varied widely in size (2–1,581 genes present in the data set) and coverage by the A549 data set due to the sparsity of scRNA-seq data. To test the robustness of this approach, we looked at the correlation of PCC scores along the leading axis for each gene set across 10 random seeds and found they were highly similar (average PCC between seeds = 0.998).

We used the same iteration and convergence threshold for analysis of the C6 (Molecular Signature Database) and the GEAR drug resistance gene sets (Wang et al., 2017) which were used to project the trametinib data. Gene sets, KEGG pathways and GO terms associated with the GEAR drug genes were obtained using KEGGREST package in R for KEGG pathways and http://geneontology.org/ for GO terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2017).

DATA AVAILABILITY STATEMENT

Code and data for generating the primarily results of this study can be found at https://github.com/panchyni/gsNMF.

AUTHOR CONTRIBUTIONS

TH designed the research. NP and TH performed the research and wrote the manuscript. NP, KW, and TH analyzed the data. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM140462 to TH. The funder had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.719099/full#supplementary-material

[1]http://www.gsea-msigdb.org/gsea/msigdb/index.jsp

Supplementary Figure 1 | Predicting A549 samples from specific time points using gsPCA. **(A)** Scatter plot of E (X-axis) and M (Y-axis) scores for all TGF-β induction samples using gsPCA. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(B–D)** Scatter plot of 0-day (green, **B**), 1-day (yellow, **C**), and 7-day samples (purple, **D**) inferred using a gsPCA model built using all other time points (gray). **(E)** A scatter plot of TGF-β induction samples with TGF-β reversion samples (i.e., 7 days induction followed by removal from TGF-β). Induction samples are labeled as in panel **(A)**, while reversion samples are colored blue, with darker shade indicating longer time since removal. **(F)** Scatter plot of 3-day reversion samples (dark blue) inferred using a gsPCA model built using all non-reversion time points (gray).

Supplementary Figure 2 | Transferring gsPCA models between A549 and DU145 TGF-β induced samples. **(A–D)** Scatter plot of E (X-axis) and M (Y-axis) scores for different combinations of data and gsPCA model: **(A)** A549 model on A549 data, **(B)** DU145 model on A549 data, **(C)** A549 model on DU145 data, and **(D)** DU145 model on DU145 data. Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(E,F)** Comparison of E-scores of samples from A549 **(E)** and DU145 **(F)** data. The X-axis is the E-score from using the model from the same data set (A549 on A549 and DU145 by DU145), while the Y-axis is the E-score from the opposite model (DU145 on A549 and A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple). **(G,H)** Comparison of M-scores of samples from A549 **(G)** and DU145 **(H)** data. The X-axis is the M-score from using the model from the same data set (A549 on A549 and DU145 by DU145), while the Y-axis is the M-score from the opposite model (DU145 on A549 and A549 on DU145). Samples from different time points are indicated by color going from 0 days (dark green) to 7 days (dark purple).

Supplementary Figure 3 | Transferring gsPCA models to TCGA data. **(A,B)** Scatter plots of E-scores for PRAD **(A)** and LUAD **(B)** from transferring gsPCA models built on A549 (X-axis) and DU145 (Y-axis) data. The color of individual points indicates the original GSVA based E-score of the TCGA data set. **(C,D)** Scatter plots of M-scores for PRAD **(C)** and LUAD **(D)** from transferring gsPCA models built on A549 (X-axis) and DU145 (Y-axis) data. The color of individual points indicates the original GSVA based M-score of the TCGA data set.

Supplementary Figure 4 | Transferring gsPCA models between temporal and spatial data sets. **(A–C)** Scatter plots of E (X-axis) and M (Y-axis) scores for Mock spatial data from gsPCA models built on different data sets: Mock spatial data **(A)**, TGF-β induced spatial data **(B)**, and TGF-β induced A549 temporal data **(C)**. The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue). **(D–F)** Scatter plots of E (X-axis) and M (Y-axis) scores for TGF-β spatial data from gsPCA models built on different data sets: TGF-β induced spatial data **(D)**, Mock spatial data **(E)**, and TGF-β induced A549 temporal data **(F)**. The color of the sample indicates whether it originates from a cell in the inner-ring (non-motile, red) or the outer ring (motile, blue).

Supplementary Figure 5 | Normalized expression of KRT8 and KRT18 across A549 and migration data sets. Boxplots showing the normalized expression of KRT8 **(top)** and KRT18 **(bottom)** across A549 data **(left)** and migration data **(right)**. The central black line indicates the average of each distribution while the whiskers show 1.5 times the interquartile range. For A549 data, color of the boxplot indicates the time from 0 day (dark green) to 7 days (dark purple) for TGF-β treatment, followed by removal of TGF-β for 8 h, 1 days, and 3 days (darkening shades of blue). For migration data, color differentiates samples in the inner (red) vs. outer (blue) rings of the assay.

Supplementary Figure 6 | Performance of gsPCA models across cross-validation data sets. Boxplots showing Adjusted R^2 of the linear model of gsNMF leading E and M dimensions across different number of model components (X-axis) and different convergence criteria: basic (algorithm standard, **A**), strong (1e–6 tolerance, 500 iterations, **B**), very strong (1e–9 tolerance, 2,500 iterations, **C**), and nndsvd (initialization with non-negative singular value decomposition, 1e–6 tolerance, 500 iterations, **D**). Left and right panels separate the result for all validation folds and the mean performance of independent folds within each training data set (see **Supplemental Methods**). The yellow line indicates the average of each distribution while the whiskers show 1.5 times the interquartile range. Red dots show the individual Adjusted R^2 values in each distribution.

REFERENCES

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

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4164–4169. doi: 10.1073/pnas.0308531101

Chakraborty, P., George, J. T., Tripathi, S., Levine, H., and Jolly, M. K. (2020). Comparative study of transcriptomics-based scoring metrics for the epithelial-hybrid-mesenchymal spectrum. *Front. Bioeng. Biotechnol.* 8:220.

Chen, J., and Zhang, S. (2018). Discovery of two-level modular organization from matched genomic data via joint matrix tri-factorization. *Nucleic Acids Res.* 46, 5967–5976. doi: 10.1093/nar/gky440

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Cook, D. P., and Vanderhyden, B. C. (2020). Context specificity of the EMT transcriptional response. *Nat. Commun.* 11:2142.

Cursons, J., Pillman, K. A., Scheer, K. G., Gregory, P. A., Foroutan, M., Hediyeh-Zadeh, S., et al. (2018). Combinatorial targeting by MicroRNAs Co-ordinates post-transcriptional control of EMT. *Cell Syst.* 7, 77–91.e7. doi: 10.1016/j.cels.2018.05.019

DeTomaso, D., and Yosef, N. (2016). FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* 17:1–12.

Ding, J., Condon, A., and Shah, S. P. (2018). Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nat. Commun.* 9:2002.

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep.* 8:9743.

George, J. T., Jolly, M. K., Xu, S., Somarelli, J. A., and Levine, H. (2017). Survival outcomes in cancer patients predicted by a partial EMT gene expression scoring metric. *Cancer Res.* 77, 6415–6428. doi: 10.1158/0008-5472.can-16-3521

Griggs, L. A., Hassan, N. T., Malik, R. S., Griffin, B. P., Martinez, B. A., Elmore, L. W., et al. (2017). Fibronectin fibrils regulate TGF-β1-induced epithelial-mesenchymal transition. *Matrix Biol.* 60, 157–175. doi: 10.1016/j.matbio.2017.01.001

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Hirway, S. U., Hassan, N. T., Sofroniou, M., Lemmon, C. A., and Weinberg, S. H. (2021). Immunofluorescence image feature analysis and phenotype scoring pipeline for distinguishing epithelial–mesenchymal transition. *Microsc Microanal* 27, 849–859. doi: 10.1017/s1431927621000428

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., et al. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* 8:e43803.

Kröger, C., Afeyan, A., Mraz, J., Eaton, E. N., Reinhardt, F., Khodor, Y. L., et al. (2019). Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. *Proc. Natl. Acad. Sci. U.S.A.* 2019:201812876.

Lee, E.-K., Han, G.-Y., Park, H. W., Song, Y.-J., and Kim, C.-W. (2010). Transgelin promotes migration and invasion of cancer stem cells. *J. Proteome Res.* 9, 5108–5117. doi: 10.1021/pr100378z

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi: 10.1038/s41592-018-0229-2

Luecken, M. D., and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15, e8746.

McFaline-Figueroa, J. L., Hill, A. J., Qiu, X., Jackson, D., Shendure, J., and Trapnell, C. (2019). A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat. Genet.* 51, 1389–1398. doi: 10.1038/s41588-019-0489-5

McFarland, J. M., Paolella, B. R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* 11:4296.

McGraw, K. O., and Wong, S. P. (1992). A common language effect size statistic. *Psychol. Bull.* 111:361. doi: 10.1037/0033-2909.111.2.361

Mendez, M. G., Kojima, S. I., and Goldman, R. D. (2010). Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. *FASEB J.* 24, 1838–1851. doi: 10.1096/fj.09-151639

Min, W., Liu, J., and Zhang, S. (2018). Edge-group sparse PCA for network-guided high dimensional data analysis. *Bioinformatics* 34, 3479–3487. doi: 10.1093/bioinformatics/bty362

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745.

Panchy, N., Azeredo-Tseng, C., Luo, M., Randall, N., and Hong, T. (2020). Integrative transcriptomic analysis reveals a multiphasic epithelial–mesenchymal spectrum in cancer and non-tumorigenic Cells. *Front. Oncol.* 9:1479.

Pastushenko, I., Brisebarre, A., Sifrim, A., Fioramonti, M., Revenco, T., Boumahdi, S., et al. (2018). Identification of the tumour transition states occurring during EMT. *Nature* 556, 463–468. doi: 10.1038/s41586-018-0040-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research* 12, 2825–2830.

Shin, H., Kim, D., and Helfman, D. M. (2017). Tropomyosin isoform Tpm2.1 regulates collective and amoeboid cell migration and cell aggregation in breast epithelial cells. *Oncotarget* 8:95192. doi: 10.18632/oncotarget.19182

Sigg, C. D., and Buhmann, J. M. (2008). "Expectation-maximization for sparse and non-negative PCA," in *Proceedings of the 25th international conference on Machine learning; 2008 2008*, (Zurich).

Stein-O'Brien, G. L., Arora, R., Culhane, A. C., Favorov, A. V., Garmire, L. X., Greene, C. S., et al. (2018). Enter the matrix: factorization uncovers knowledge from omics. *Trends Genet.* 34, 790–805. doi: 10.1016/j.tig.2018.07.003

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck Iii, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902. doi: 10.1016/j.cell.2019.05.031

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

Tan, T. Z., Miow, Q. H., Miki, Y., Noda, T., Mori, S., Huang, R. Y. J., et al. (2014). Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* 6, 1279–1293. doi: 10.15252/emmm.201404208

The Gene Ontology Consortium (2017). Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338.

Tomaskovic-Crook, E., Thompson, E. W., and Thiery, J. P. (2009). Epithelial to mesenchymal transition and breast cancer. *Breast Cancer Res.* 11, 1–10.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, W., He, J., Lu, H., Kong, Q., and Lin, S. (2020). KRT8 and KRT19, associated with EMT, are hypomethylated and overexpressed in lung adenocarcinoma and link to unfavorable prognosis. *Biosci. Rep.* 40:BSR20193468.

Wang, Y.-Y., Chen, W.-H., Xiao, P.-P., Xie, W.-B., Luo, Q., Bork, P., et al. (2017). GEAR: A database of G enomic E lements A ssociated with drug R esistance. *Sci. Rep.* 7:44085.

Watanabe, K., Panchy, N., Noguchi, S., Suzuki, H., and Hong, T. (2019). Combinatorial perturbation analysis reveals divergent regulations of mesenchymal genes during epithelial-to-mesenchymal transition. *NPJ. Syst. Biol. Appl.* 5:21.

Ye, Y., Kang, X., Bailey, J., Li, C., and Hong, T. (2019). An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. *PLoS Comput. Biol.* 15:e1006855. doi: 10.1371/journal.pcbi.1006855

Yui, M. A., and Rothenberg, E. V. (2014). Developmental gene networks: a triathlon on the course to T cell identity. *Nat. Rev. Immunol.* 14, 529–545. doi: 10.1038/nri3702

Zhang, J., Hu, S., and Li, Y. (2019). KRT18 is correlated with the malignant status and acts as an oncogene in colorectal cancer. *Biosci. Rep.* 39:BSR20190884.

Zhang, L., and Zhang, S. (2019). Learning common and specific patterns from data of multiple interrelated biological scenarios with matrix factorization. *Nucleic Acids Res.* 47, 6606–6617. doi: 10.1093/nar/gkz488

Zhu, J., Zheng, Y., Zhang, H., Liu, Y., Sun, H., and Zhang, P. (2019). Galectin-1 induces metastasis and epithelial-mesenchymal transition (EMT) in human ovarian cancer cells via activation of the MAPK JNK/p38 signalling pathway. *Am. J. Transl Res.* 11:3862.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fgene.2021.756784

Agent Repurposing for the Treatment of Advanced Stage Diffuse Large B-Cell Lymphoma Based on Gene Expression and Network Perturbation Analysis

*Chenxi Xiang[1†], Huimin Ni[2†], Zhina Wang[3†], Binbin Ji[4,5], Bo Wang[4,5], Xiaoli Shi[4,5], Wanna Wu[2], Nian Liu[2], Ying Gu[2], Dongshen Ma[1] and Hui Liu[2]**

[1]Department of Pathology, The Affiliated Hospital of Xuzhou Medical University, Xuzhou, China, [2]Department of Pathology, Xuzhou Medical University, Xuzhou, China, [3]Department of Oncology, Emergency General Hospital, Beijing, China, [4]Genies Beijing Co., Ltd., Beijing, China, [5]Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology, China

Reviewed by:
Lina Zhao,
Chinese Academy of Medical Sciences, China
Kebo LV,
Ocean University of China, China

***Correspondence:**
Hui Liu
hliu@xzhmu.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: *11 August 2021*
Accepted: *24 September 2021*
Published: *14 October 2021*

Citation:
Xiang C, Ni H, Wang Z, Ji B, Wang B, Shi X, Wu W, Liu N, Gu Y, Ma D and Liu H (2021) Agent Repurposing for the Treatment of Advanced Stage Diffuse Large B-Cell Lymphoma Based on Gene Expression and Network Perturbation Analysis. Front. Genet. 12:756784. doi: 10.3389/fgene.2021.756784

Over 50% of diffuse large B-cell lymphoma (DLBCL) patients are diagnosed at an advanced stage. Although there are a few therapeutic strategies for DLBCL, most of them are more effective in limited-stage cancer patients. The prognosis of patients with advanced-stage DLBCL is usually poor with frequent recurrence and metastasis. In this study, we aimed to identify gene expression and network differences between limited- and advanced-stage DLBCL patients, with the goal of identifying potential agents that could be used to relieve the severity of DLBCL. Specifically, RNA sequencing data of DLBCL patients at different clinical stages were collected from the cancer genome atlas (TCGA). Differentially expressed genes were identified using DESeq2, and then, weighted gene correlation network analysis (WGCNA) and differential module analysis were performed to find variations between different stages. In addition, important genes were extracted by key driver analysis, and potential agents for DLBCL were identified according to gene-expression perturbations and the Crowd Extracted Expression of Differential Signatures (CREEDS) drug signature database. As a result, 20 up-regulated and 73 down-regulated genes were identified and 79 gene co-expression modules were found using WGCNA, among which, the thistle1 module was highly related to the clinical stage of DLBCL. KEGG pathway and GO enrichment analyses of genes in the thistle1 module indicated that DLBCL progression was mainly related to the NOD-like receptor signaling pathway, neutrophil activation, secretory granule membrane, and carboxylic acid binding. A total of 47 key drivers were identified through key driver analysis with 11 up-regulated key driver genes and 36 down-regulated key diver genes in advanced-stage DLBCL patients. Five genes (*MMP1*, *RAB6C*, *ACCSL*, *RGS21* and *MOCOS*) appeared as hub genes, being closely related to the occurrence and development of DLBCL. Finally, both differentially expressed genes and key driver genes were subjected to CREEDS analysis, and 10 potential agents were predicted to have the potential for application in advanced-stage DLBCL patients. In conclusion, we propose a novel pipeline to utilize perturbed

gene-expression signatures during DLBCL progression for identifying agents, and we successfully utilized this approach to generate a list of promising compounds.

Keywords: diffuse large B-cell lymphoma, drug repurposing, differentially expressed genes, differential module analysis, key driver analysis

INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL) is the most commonly diagnosed non-Hodgkin lymphoma (NHL), representing approximately 25% of new NHL cases each year in the United States (Liu and Barta, 2019). In practice, about one half of DLBCL patients presented with advanced-stage disease (Prakash et al., 2012), featuring bulky tumor burden and poor patient response to treatment. According to published data, advanced-stage DLBCL (stage I/II and stage III/IV) may be both biologically and clinically different from limited-stage DLBCL cases (stage I and II). For example, advanced-stage DLBCL patients were more likely to express higher levels of CD30 (Rodrigues-Fernandes et al., 2021) and CD25 (Oka et al., 2020), both of which are biomarkers of B-cell activation. In addition, advanced-stage DLBCL was also shown to be associated with a higher immune-inflammation index (Wang et al., 2021) and an increased level of lymphopenia at diagnosis (Shin et al., 2020), highlighting its deteriorating immune regulation. Green and Johnson et al. reported there were a few biological factors known to adversely impact the prognosis of DLBCL patients, including the cell-of-origin, co-expression of MYC/BCL2 and co-occurrence of *MYC* and *BCL2/BCL6* rearrangements failed to predict poorer prognosis in limited stage DLBCL(Green et al., 2012; Johnson et al., 2012). Ajay, Major *et al* reported that stage I and II DLBCL cases had a slightly increased risk of secondary primary malignancies after DLBCL treatment in long-term follow-up (>20 years) (Major et al., 2020). Comparing with limited stage DLBCL, advanced-stage DLBCL patients were more likely to benefit from intensified radiotherapy (Hoiland et al., 2020; Freeman et al., 2021). Also, the pattern of late disease relapses observed in advanced stage DLBCL cases was different from that of limited-stage cases, further corroborating that limited and advanced stage DLBCL were biologically heterogeneous (Hoiland et al., 2020). All of these observations prompted us to treat advanced- and limited-stage DLBCL with different strategies, better tailoring for their specific biological and clinical characteristics.

However, there is limited knowledge regarding the genomic and transcriptomic differences between limited- and advanced-stage DLBCL. Two previous large analyses exploring the genetic landscape of DLBCL were not intended to compare the limited and advanced stages of the disease (Reddy et al., 2017; Schmitz et al., 2018). Moreover, at the single gene or single locus level, advanced- and limited-stage DLBCL may also be different in terms of their altered gene regulation and regulatory/co-expression networks, which was confirmed in other clinical comparisons such as cancer vs normal tissue (Zhang et al., 2018; Xu et al., 2019) and young vs old (Yang et al., 2015; Yang et al., 2016b).

Although frontline chemoimmunotherapies have been shown to cure up to 60% of patients with advanced-stage disease, with a clear plateau in progression-free survival (PFS) and rare relapses beyond 5 years (Coiffier et al., 2010), there still is a fraction of patients who are subject to relapse and have tumors that are refractory to treatment (Coiffier et al., 2010), highlighting the heterogeneity of advanced DLBCL. Thus, it is critical to develop new drugs for improving the treatment of advanced-stage DLBCL, so that it might be effectively treated by using existing treatment strategies as limited-stage DLBCL patients are. However, the development of a novel drug is usually costly and time-consuming (Liu et al., 2020; Yang et al., 2020) and highlights the need for effective drug repositioning strategies. There are many computer-based drug repositioning methods that have been used for cancers (Xu et al., 2019; Liu et al., 2020) and other diseases, such as Coronavirus disease 2019 (COVID-19) (Tang et al., 2020; Li et al., 2021).

In this study, we propose a new strategy for identifying new agents that have the potential to specifically target advanced-stage DLBCL. In general, we retrieved advanced-stage DLBCL-specific expressed genes by comparing the transcriptome of advanced-stage disease with that of limited-stage DLBCL. These differentially expressed genes (DEGs) were then subjected to weighted gene correlation network analysis (WGCNA) to discover the co-expression modules that may contribute to the progression of this disease. Finally, potential personal agents were obtained from the Crowd Extracted Expression of Differential Signatures (CREEDS) based on the down-regulation and up-regulation of genes (see Materials and methods for details). We aimed to specifically reveal the transcriptomic scenario occurring in advanced-stage DLBCL and to elucidate the genes that were most likely contributing to disease progression. Based on this knowledge, we then identified some potential agents for the treatment of advanced-stage DLBCL in future clinical practice.

MATERIALS AND METHODS

Data Collection

RNA sequencing data from patients with DLBCL cancer were collected from the cancer genome atlas (TCGA). Based on the imaging results, including computed tomography (CT) scans, magnetic resonance imaging (MRI) or positron emission tomography (PET) scanning, patients were divided into four stages (I–IV) according to the Ann Arbor system (Heidelberg, 2020).

Differential Gene Expression Analysis Between Samples at Different Stages

An expression matrix of 42 patients and their group information (stage I/II or III/IV) were used as the input for DEG discovery.

DEGs between samples at stage I/II and stage III/IV were obtained using DESeq2 (Love et al., 2014) using $\log_2$ |fold change| ≧ 1 and a p value ≦ 0.05.

Survival Analysis

After identifying DEGs, we performed survival analysis on these genes for all of the patients. Next, Kaplan-Meier (Bland and Altman, 1998) survival estimation was used for all differentially expressed genes to identify genes correlated with overall survival. Kaplan- Meier arranged the survival time in descending order, at each death node, it estimated the proportion of the observed values that survived for a certain period of time under the same circumstances, which could intuitively show the survival and mortality rates of two or more groups. The R packages survival and survminer were used for survival analysis and curve plotting, respectively.

Weighted Gene Correlation Network Analysis

The WGCNA package in R (Peter and Horvath, 2008) was used to construct a co-expression network. For this step, we randomly picked 400 genes from the stage III/IV patients to generate a topological overlap matrix since the gene number was too large to perform this analysis using all of the genes. For the constructed gene network to conform to a scale-free distribution, a soft threshold was used to select the appropriate β after removing outliers. Finally, the soft threshold was set to 10. Then, genes were clustered by hierarchical clustering, and the tree was cut into different modules using a dynamic cutting algorithm, in which genes were highly correlated. Furthermore, we calculated the Pearson correlation coefficient between different modules and clinical stage and used this Pearson correlation coefficient to judge the relationship between the module and clinical stage. Finally, significant modules closely related to the occurrence and development of DLBCL were identified for follow-up analysis.

Functional and Pathway Enrichment Analyses

KEGG pathway (Ogata et al., 1999) analysis and Gene Ontology (GO) analysis (Botstein et al., 2000), including biological process (BP), cellular composition (CC) and molecular function (MF), were performed on the genes in the module identified by WGCNA to understand the biological significance of the progression of DLBCL. The R package clusterProfiler (Yu et al., 2012) was used in the process of enrichment analysis to analyze the functions of the genes from these modules.

Key Driver Analysis

For key driver analysis, we used up- or down-regulated genes separately as inputs to identify key drivers. Key driver analysis (Yang et al., 2016a) (KDA) was used to identify hub genes, and protein actions v11.0 was used as a reference protein–protein interaction network (Szklarczyk et al., 2021). Parameters were set as follows: nlayerExpansion was set to 1, nlayerSearch was set to 6 and enrichedNodesPercentCut was set to −1. A p value_whole ≦ 0.05 was used to filter out key drivers. The hub genes were of great significance in terms of the occurrence and development of DLBCL.

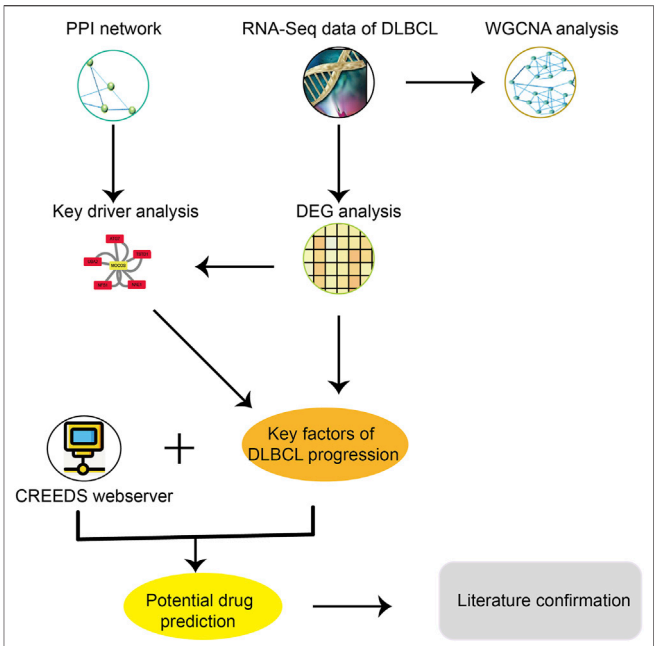


FIGURE 1 | A brief study design for drug repurposing, including these major steps: 1) Download and organize the RNA-seq data and clinical information of DLBCL from TCGA; 2) Got key factors of DLBCL progression through DEG analysis, key driver analysis and WGCNA analysis; 3) Potential drug prediction through CREEDs; 4) Literature confirmation.

Drug Discovery

CREEDS includes single gene perturbation signatures, as well as disease and drug perturbation signatures, and it can be used to identify the relationship between gene, disease and drug (Gillies et al., 2016). CREEDS is composed of single-drug perturbation-induced gene expression signatures. Utilizing this database, agents that can reverse the behavior of up/down-regulated genes can be discovered, and the best matched agents are reported. We used this tool for drug discovery for advanced-stage DLBCL. In this work, we combined differentially expressed genes and key driver genes as a new gene set to discover new agents related to advanced-stage DLBCL.

RESULTS

A Brief Study Design of Drug Repurposing

For the purpose of specifically developing new agents that could be utilized in combination with R- CHOP backbones to treat advanced stage DLBCL patients, we proposed a new method of drug repurposing based on gene expression and network perturbation (**Figure 1**). In order to identify key factors for DLBCL progression, WGCNA and DEG, differential module (DM) and key driver (KD) analyses were performed. Then, the key factors of DLBCL progression and drug perturbation signature

TABLE 1 | Summary of general clinical information of DLBCL cases in TCGA.

		Limited stage	Advanced stage	χ^2	*P*
Gender	Male	9	10	0.006	0.938
	Female	16	7		
Age	≥60	6	10	5.203	0.023
	<60	19	7		
Extranodal disease	Yes	8	11	4.369	0.037
	No	17	6		
B symptoms	Yes	1	9	13.36	0.000
	No	24	8		

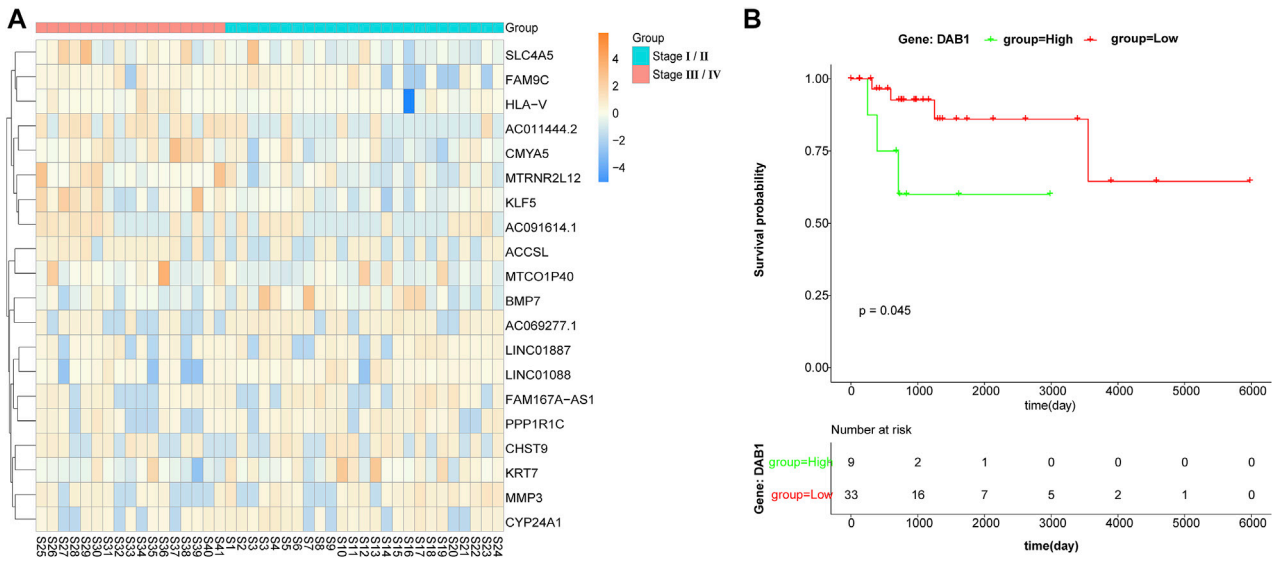


FIGURE 2 | Analysis of differentially expressed genes. **(A)** Heat map of the top 10 differentially expressed genes. The *x*-axis represents different samples from TCGA, blue indicates samples at limited stage (stage I/II) and red indicates samples at advanced stage (stage III/IV). The *y*-axis represents differentially expressed genes. **(B)** Survival curve of the association between the expression levels of *DAB1* and survival time after diagnosis with DLBCL.

were used to predict potential agents for the treatment of advanced stage DLBCL. Finally, some previous studies were reviewed to demonstrate the effectiveness of the newly identified agents.

Patient Characteristics

The clinical characteristics of DLBCL cancer patients collected from TCGA are presented in **Table 1**, including 25 patients at clinical stage I/II and 17 patients at clinical stage III/IV. It was more likely to occur in elder patients and involve extranodal sites or organs. Patients of advanced stage disease also tended to have B symptoms. No gender preference was observed in this group of patients and all patients received no treatment before resection of tumors.

Identification of DEGs and Survival Analysis

After collecting data from TCGA, DEGs were obtained using DESeq2, by comparing the transcriptome of advanced stage DLBCL with limited stage DLBCL. Of the 93 DEGs that were identified with a $\log_2$ |fold change| ≧ 1 and a *p* value ≦ 0.05, 20 genes were up-regulated and 73 genes were down-regulated in advanced DLBCL. The top 10 genes that were differently expressed between advanced and limited stage DLBCL are shown in **Figure 2A**.

We aimed to evaluate whether this set of differentially expressed genes could define a group of patients with poorer prognosis. We dichotomized 42 DLBCL cases into either the high expression group or the low expression group as per the mean expression level of each DEG. In addition, the Kaplan-Meier survival estimation method was used to evaluate all DEGs to study the relationship between gene expression and overall survival. Through this Kaplan-Meier survival estimation analysis, we found that *DAB1* was negatively correlated with overall survival, while other DEGs were not correlated with overall survival.

Weighted Gene Correlation Network Analysis and Differential Model Analysis

WGCNA, based on a scale-free network to analyze genes according to their expression patterns, was used to cluster highly related genes into one module. As can be seen from **Figure 3A**, the soft threshold value was set at 10 to build this scale-free network. Next, 79 gene modules were identified by

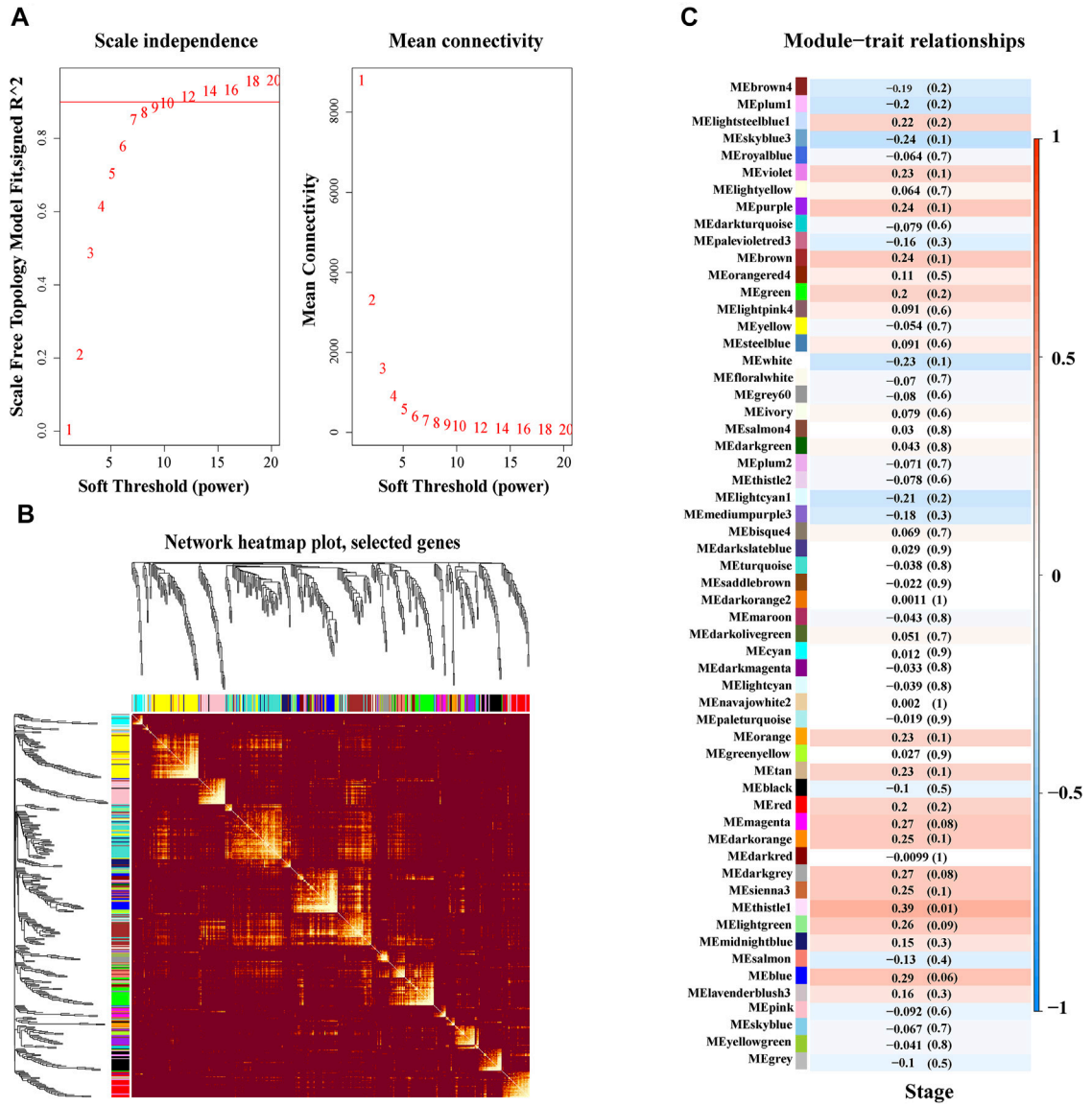


FIGURE 3 | Weighted co-expression and key module identification associated with clinical DLBCL stage. **(A)** Determination of soft threshold in WGCNA. **(B)** Topological overlapping WGCNA heat map. **(C)** The relationship between modules and clinical traits. Pearson correlation coefficient was used to calculate the correlation degree between each module and trait.

hierarchical clustering and dynamic branch cutting, and each module was assigned a unique color identifier (**Supplementary Figure S4**). We then selected a portion of these genes to construct a topological overlapping heat map, shown in **Figure 3B**. Through differential module analysis, we found that the thistle1 module was most relevant to advance stage of DLBCL in this dataset (**Figure 3C**).

Functional and Pathway Enrichment Analysis of the thistle1 Module

In order to understand the causes of DLBCL deterioration from the biological level, we analysed the genes in the thistle1 module using KEGG pathway and GO enrichment analysis. KEGG pathway analysis results indicated that the development of DLBCL was very strongly correlated to the NOD-like receptor signalling pathway, osteoclast differentiation, leishmaniasis, *Staphylococcus aureus* infection and viral protein interaction with cytokine and cytokine receptor (**Figure 4A**). Furthermore, GO enrichment was performed based on three aspects: BP, CC and MF. In the BP analysis, we found that the genes in the thistle1 module were mainly related to neutrophil activation, positive regulation of response to external stimulus and response to interferon−gamma (**Figure 4B**). In addition, in the CC analysis, the genes in the thistle1 module were related to secretory granule membrane, endocytic vesicle and apical part of

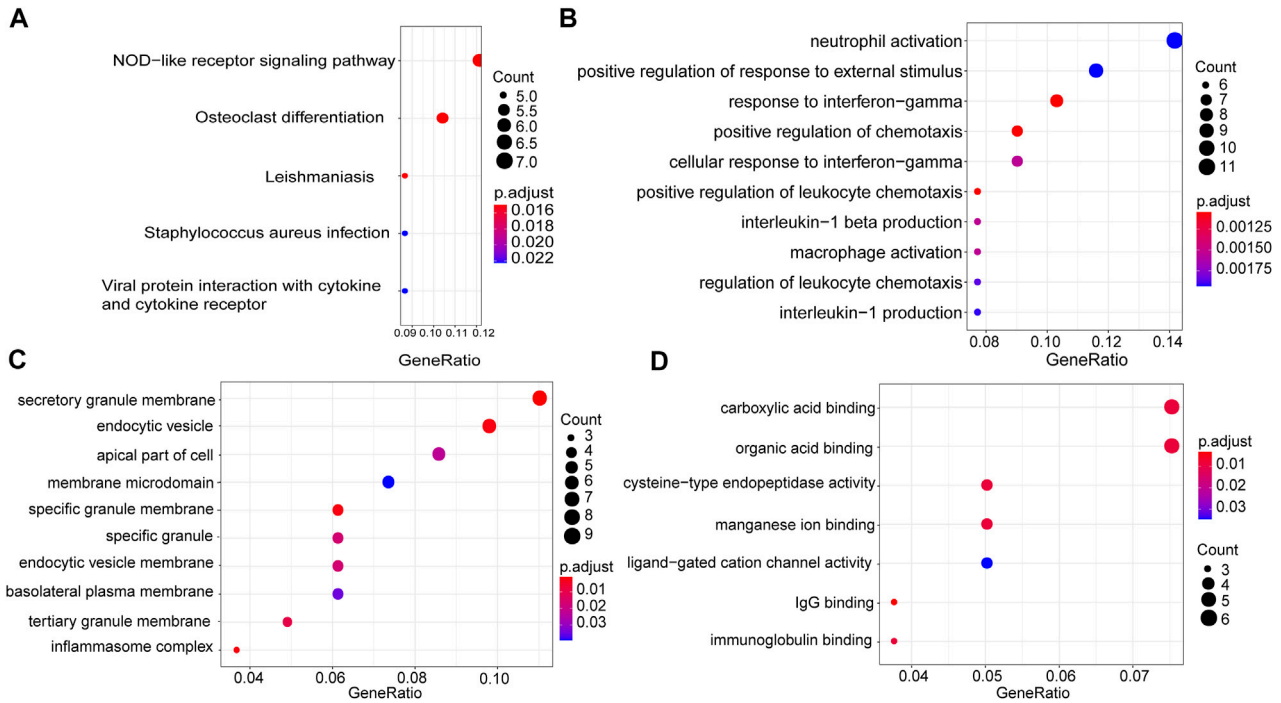


FIGURE 4 | Pathway and functional enrichment analysis of genes in the thistle1 module. **(A)** KEGG pathway analysis. **(B)** GO enrichment for biological process. **(C)** GO enrichment for cellular composition. **(D)** GO enrichment for molecular function. The *x*-axes are the ratio of genes, and the *y*-axes are the GO terms.

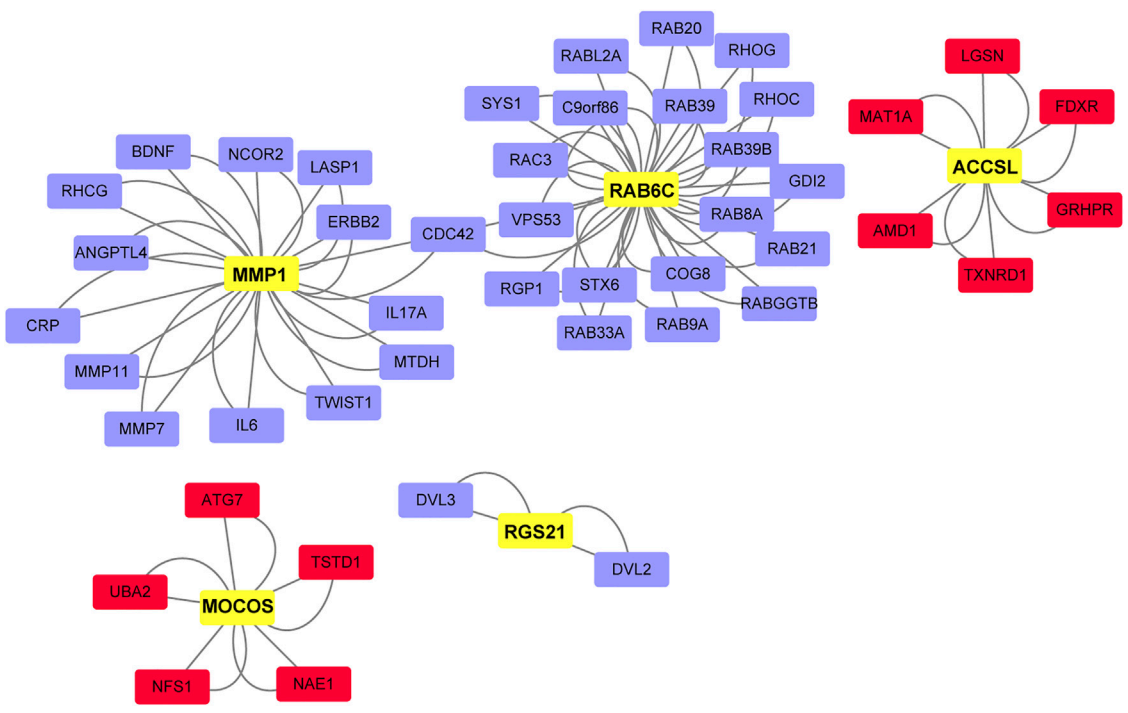


FIGURE 5 | Network of key DLBCL drivers and hub genes. Red, key drivers from up-regulated gene set in advanced-stage samples. Blue, key drivers from down-regulated gene set in advanced-stage samples. Yellow, hub genes.

TABLE 2 | Potential DLBCL treatment agents.

Gene type	Drug/Small molecule	Possible effect	Evidence
Down	Formaldehyde	A metabolite of vitamin A that plays important roles in cell growth, differentiation and organogenesis acts as an inhibitor of the transcription factor Nrf2 through the activation of retinoic acid receptor alpha	DOI:10.14423/SMJ.0000000000000545
Down	Ethanol	Similar to pharmacological mTOR inhibitors, which can inhibit the mTOR signaling pathway	DOI: 10.1186/s12964-015-0091-0
Down	Dibutyl phthalate	Is expected to cause severe side effects to the central nervous system of animals and humans	DOI:10.1016/S0145-2126 (96)00033-1
Down	Paclitaxel	A synthetic macrocyclic ketone analog of the marine sponge natural product halichondrin B, which leads to the inhibition of microtubule growth in the absence of effects on microtubule shortening at microtubule plus ends	Unknown
Down	Prednisolone	Belongs to the adrenal corticotropic hormone and adrenal corticotropic hormone class and has strong anti-inflammatory effects	DOI:10.3109/10428194.2011.588761 DOI:10.5045/kjh. 2012.47.4.293
Up	Oxaliplatin	It selectively inhibits the synthesis of deoxyribonucleic acid (DNA). The guanine and cytosine contents correlate with the degree of oxaliplatin-induced cross-linking	DOI: 10.1016/S2352-3026 (18)30054-1
Up	Eribulin	Is a microtubule inhibitor indicated for the treatment of patients with metastatic breast cancer who have previously received at least two chemotherapeutic regimens for the treatment of metastatic disease. Also being investigated for use in the treatment of advanced solid tumors	DOI: 10.1007/s00280-012-1976-x. Epub 2012 Sep 26
Up	NC1153	Specifically inhibits JAK3 via NC1153-induced apoptosis of certain leukemia/lymphoma cell lines	DOI: 10.1016/j.febslet. 2010.02.071
Up	EPZ-6438	Selectively inhibits intracellular histone H3 lysine 27 (H3K27) methylation in a concentration- and time-dependent manner in both EZH2 wild-type and mutant lymphoma cells	DOI: 10.1158/1535-7163.MCT-13-0773
Up	R547	A potent CDK inhibitor with a potent anti-proliferative effect at pharmacologically relevant doses	DOI: 10.1158/1535-7163.MCT-09-0083

cell (**Figure 4C**). Moreover, the genes in the thistle1 module were mainly enriched in 7 MFs, including carboxylic acid binding, organic acid binding, cysteine–type endopeptidase activity, manganese ion binding, ligand–gated cation channel activity, immunoglobulin G (IgG) binding and immunoglobulin binding (**Figure 4D**).

Hub Genes Identified Through Key Driver Analysis

A total of 47 key drivers were identified through key driver analysis, with 11 up-regulated key driver genes and 36 down-regulated key diver genes being diagnostic of advanced-stage DLBCL relative to limited-stage DLBCL. Then, five hub genes were identified from key drivers as shown in **Figure 5**, which were most related to the occurrence and development of DLBCL. *MMP1* (Rosas et al., 2008), also known as matrix metalloproteinase-1, encodes a protein of 469 amino acid residues and is a kind of photolytic enzyme closely related to tumor genesis, invasion and metastasis. *Rab6c* (Young et al., 2010) is a member of the RAS family. Its mutation can affect the balance of Ras-GTP and cause malignant transformation of cells. Gene ontology annotations for 1-Aminocyclopropane-1-Carboxylate Synthase Homolog (Inactive) Like (*ACCSL*) (Chen and Karampinos, 2020) include pyridoxal phosphate binding. Dysregulation of gene levels of molybdenum cofactor sulfurase (*MOCOS*) (Kurzawski et al., 2012) can lead to cell disorders. Studies have demonstrated that this gene can be used as a key detection gene for kidney genetic diseases. *RGS21* (Von Buchholtz et al., 2004), a new member of the regulator of G protein signaling (RGS) protein family. It can inhibit signal transduction by increasing GTPase activity.

Agent Screening

Potential personal agents associated with DLBCL were identified according to the differences between differential genes and drug signaling. Approximately 10 potential agents were selected according to their drug perturbation-induced gene expression signatures, and detailed information on these agents is presented in **Table 2**, including the type, drug/small molecule, possible effect and evidence for activity. The top five agents could reverse the expression of down-regulated genes, and the remaining agents could reverse the expression of up-regulated genes. In other words, after treatment with these drugs, gene expression levels may return to normal. The top five agents that may reverse down-regulated gene expression are formaldehyde, ethanol, dibutyl phthalate, paclitaxel, and prednisolone. Ethanol (EtOH) is similar to pharmacological mTOR inhibitors and has been shown to inhibit the mTOR signaling pathway. Mazan et al. studied the influence of EtOH on the mTOR signaling pathway and explored the translational group analysis of downstream effects of EtOH in DLBCL, and the results showed that EtOH partially inhibited mTOR signaling and protein translation (Mazan-Mamczarz et al., 2015). In a previous study, newly diagnosed DLBCL patients treated with rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisolone (R-CHOP) were evaluated for their clinical characteristics, therapeutic efficacy and patient survival, and DLBCL patients treated with R-CHOP had better survival than other patients (Hong et al., 2011). Ohe et al. also

reported a case of DLBCL successfully treated with prednisolone (Ohe et al., 2012). The top five agents that may reverse up-regulated gene expression are oxaliplatin, eribulin, NC1153, EPZ-6438 and R547. Oxaliplatin selectively inhibits the synthesis of deoxyribonucleic acid (DNA). Shen et al. studied the efficacy, safety and feasibility of the combination of rituximab, gemcitabine, and oxaliplatin (R-GemOx) as a first-line treatment in elderly patients with DLBCL. They found that R-GemOx might be a therapeutic option for the management of DLBCL (Shen et al., 2018).

DISCUSSION

DLBCL remains a highly heterogenous disease, with the frontline R-CHOP modality achieving only a 40–60% complete response (CR) rate in unselected patients. The prognosis of patients with DLBCL with refractory tumors or relapse remains dismal. As a result, designing more sophisticated personal treatment modalities has the potential to improve the outcomes in high-risk DLBCL patients. Although a wealth of studies has focused on targeted therapies based on the molecular classification of DLBCL, the clinical stage of DLBCL remains an important factor for choosing an appropriate treatment regime. DLBCL patients with advanced- and limited-stage disease have different responses to standard chemoimmunotherapies, due to the different genomic profiles of advanced-stage disease relative to limited-stage disease (Miao et al., 2019). In this study, we propose a new approach to gain insights into the intrinsic heterogeneity of DLBCL, which focused on comparing the transcriptomic profile of advanced- and limited-stage DLBCL and distilling the disease to a few distinctly expressed genes and hub genes that might contribute to disease progression. In general, 20 genes were up-regulated and 73 genes were down-regulated in advanced-stage samples compared to limited-stage samples. We also found that *DAB1* was negatively correlated with overall survival through survival analysis of all identified DEGs (**Figure 2B**, $p = 0.045$). Due to the limitations of differential expression analysis, it is impossible to group genes with the same function together. Therefore, we carried out weighted gene co-expression network analysis and analysis on different modules. During these analyses, 79 similar gene expression modules were found using WGCNA, among which, the thistle1 module was highly related to disease stage. KEGG pathway and GO enrichment analyses of the genes in the thistle1 module indicated that DLBCL progression was mainly related to the NOD-like receptor signaling pathway, neutrophil activation, secretory granule membrane and carboxylic acid binding. There is evidence that tumors and their mesenchymal cells produce many cytokines and chemokines to stimulate the differentiation of N2 neutrophils (Valerius et al., 1993; Souto et al., 2014). However, neutrophils can cause DNA damage through reactive oxygen species and related products of myeloperoxidase (MPO), and N2 cells secrete VEGF, TNF and other cytokines to promote tumor angiogenesis and, at the same time, synthesize and secrete MMP and NE to the tumor stroma to participate in the tumor reconstruction of the extracellular matrix to promote tumor growth and metastasis (Zvi et al., 2009; Mishalian et al., 2013; Zhou et al., 2016). During key driver analysis, 47 key drivers were identified and five hub genes were extracted from these key drivers, including *MMP1*. *MMP1* (Rosas et al., 2008) can alter the microenvironment of cells. When *MMP1* is out of balance, it accelerates the degradation of the matrix barrier and promotes the formation and growth of tumors by releasing matrix-related growth factors. Studies have shown that *MMP1* is associated with lung squamous cell carcinoma, colon cancer and adenocarcinoma.

Based on gene expression and network perturbations, 10 potential agents for the treatment of DLBCL were obtained. For instance, NC1153 can inhibit JAK3 specifically and induce the apoptosis of certain leukemia/lymphoma cell lines. Using Affymetrix microarray profiling following NC1153 treatment, Nagy et al. reported that JAK3-dependent survival modulating pathways (p53, TGF-beta, TNFR and ER stress) were altered in Kit225 cells (Nagy et al., 2010). EPZ-6438 selectively inhibited intracellular H3K27 methylation in a concentration- and time-dependent manner in both EZH2 wild-type and mutant lymphoma cells. Inhibition of H3K27 trimethylation (H3K27Me3) leads to selective cell killing of human lymphoma cell lines bearing EZH2 catalytic domain point mutations (Knutson et al., 2014).

In summary, we proposed a novel pipeline to utilize perturbed gene-expression signatures during DLBCL progression for identifying agents, and we successfully utilized this approach to generate a list of promising compounds. Whether this can be used clinically needs further research. We will continue to follow the latest developments of these agents in the treatment of DLBCL and explore its pharmaco-mechanisms under the aid of stage-of-art technologies in the future.

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

HL and CX designed the project, CX, HN, and ZW wrote the manuscript, BJ and XS collected data, BW carried out data analysis, NL and WW analyzed experimental results. YG and DM researched literatures. All authors read and gave their approval for the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.756784/full#supplementary-material

REFERENCES

Bland, J. M., and Altman, D. G. (1998). Statistics Notes: Survival Probabilities (The Kaplan-Meier Method). *Bmj* 317, 1572–1580. doi:10.1136/bmj.317.7172.1572

Botstein, D., Cherry, J. M., Ashburner, M., Ball, C., Blake, J., Butler, H., et al. (2000). *Gene Ontology: tool unification Biol.* 25, 25–29. doi:10.1038/75556

Chen, W., and Karampinos, D. C. (2020). Chemical-Shift Encoding-Based Water-Fat Separation With Multifrequency Fat Spectrum Modeling in Spin-Lock MRI. *Magn. Reson. Med.* 83, 1608–1624. doi:10.1002/mrm.28026

Coiffier, B., Thieblemont, C., Van Den Neste, E., Lepeu, G., Plantier, I., Castaigne, S., et al. (2010). Long-Term Outcome of Patients in the LNH-98.5 Trial, the First Randomized Study Comparing Rituximab-CHOP to Standard CHOP Chemotherapy in DLBCL Patients: a Study by the Groupe d'Etudes des Lymphomes de l'Adulte. *Blood.* 116, 2040–2045. doi:10.1182/blood-2010-03-276246

Freeman, C. L., Savage, K. J., Villa, D. R., Scott, D. W., Srour, L., Gerrie, A. S., et al. (2021). Long-Term Results of PET-Guided Radiation in Patients With Advanced-Stage Diffuse Large B-Cell Lymphoma Treated With R-CHOP. *Blood.* 137, 929–938. doi:10.1182/blood.2020005846

Gillies, R. J., Kinahan, P. E., and Hricak, H. (2016). Radiomics: Images Are More Than Pictures, They Are Data. *Radiology.* 278, 563–577. doi:10.1148/radiol.2015151169

Green, T. M., Young, K. H., Visco, C., Xu-Monette, Z. Y., Orazi, A., Go, R. S., et al. (2012). Immunohistochemical Double-Hit Score Is a Strong Predictor of Outcome in Patients With Diffuse Large B-Cell Lymphoma Treated With Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone. *J. Clin. Oncol.* 30, 3460–3467. doi:10.1200/jco.2011.41.4342

Heidelberg, S. B. (2020). *Ann Arbor Staging System*. Springer Berlin Heidelberg. doi:10.1007/978-3-540-47648-1_287

Hoiland, R. L., Fergusson, N. A., Mitra, A. R., Griesdale, D. E. G., Devine, D. V., Stukas, S., et al. (2020). The Association of ABO Blood Group With Indices of Disease Severity and Multiorgan Dysfunction in COVID-19. *Blood Adv.* 4, 4981–4989. doi:10.1182/bloodadvances.2020002623

Hong, J., Park, S., Park, J., Kim, H. S., Kim, K.-H., Ahn, J. Y., et al. (2011). Evaluation of Prognostic Values of Clinical and Histopathologic Characteristics in Diffuse Large B-Cell Lymphoma Treated with Rituximab, Cyclophosphamide, Doxorubicin, Vincristine, and Prednisolone Therapy. *Leuk. Lymphoma.* 52, 1904–1912. doi:10.3109/10428194.2011.588761

Johnson, N. A., Slack, G. W., Savage, K. J., Connors, J. M., Ben-Neriah, S., Rogic, S., et al. (2012). Concurrent Expression of MYC and BCL2 in Diffuse Large B-Cell Lymphoma Treated With Rituximab Plus Cyclophosphamide, Doxorubicin, Vincristine, and Prednisone. *J. Clin. Oncol.* 30, 3452–3459. doi:10.1200/jco.2011.41.0985

Knutson, S. K., Kawano, S., Minoshima, Y., Warholic, N. M., Huang, K.-C., Xiao, Y., et al. (2014). Selective Inhibition of EZH2 by EPZ-6438 Leads to Potent Antitumor Activity in EZH2-Mutant Non-Hodgkin Lymphoma. *Mol. Cancer Ther.* 13, 842–854. doi:10.1158/1535-7163.mct-13-0773

Kurzawski, M., Dziewanowski, K., Safranow, K., and Drozdzik, M. (2012). Polymorphism of Genes Involved in Purine Metabolism (XDH, AOX1, MOCOS) in Kidney Transplant Recipients Receiving Azathioprine. *Ther. Drug Monit.* 34, 266–274. doi:10.1097/ftd.0b013e31824aa681

Li, T., Huang, T., Guo, C., Wang, A., Shi, X., Mo, X., et al. (2021). Genomic Variation, Origin Tracing, and Vaccine Development of SARS-CoV-2: A Systematic Review. *The Innovation.* 2, 100116. doi:10.1016/j.xinn.2021.100116

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids.* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, Y., and Barta, S. K. (2019). Diffuse Large B-cell Lymphoma: 2019 Update on Diagnosis, Risk Stratification, and Treatment. *Am. J. Hematol.* 94, 604–616. doi:10.1002/ajh.25460

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8

Major, A., Smith, D. E., Ghosh, D., Rabinovitch, R., and Kamdar, M. (2020). Risk and Subtypes of Secondary Primary Malignancies in Diffuse Large B-cell Lymphoma Survivors Change Over Time Based on Stage at Diagnosis. *Cancer.* 126, 189–201. doi:10.1002/cncr.32513

Mazan-Mamczarz, K., Peroutka, R. J., Steinhardt, J. J., Gidoni, M., Zhang, Y., Lehrmann, E., et al. (2015). Distinct Inhibitory Effects on mTOR Signaling by Ethanol and INK128 in Diffuse Large B-Cell Lymphoma. *Cell Commun Signal.* 13, 15. doi:10.1186/s12964-015-0091-0

Miao, Y., Medeiros, L. J., Li, Y., Li, J., and Young, K. H. (2019). Genetic Alterations and Their Clinical Implications in DLBCL. *Nat. Rev. Clin. Oncol.* 16, 634–652. doi:10.1038/s41571-019-0225-1

Mishalian, I., Bayuh, R., Levy, L., Zolotarov, L., Michaeli, J., and Fridlender, Z. G. (2013). Tumor-Associated Neutrophils (TAN) Develop Pro-Tumorigenic Properties During Tumor Progression. *Cancer Immunol. Immunother.* 62, 1745–1756. doi:10.1007/s00262-013-1476-9

Nagy, Z. S., Ross, J. A., Rodriguez, G., Bader, J., Dimmock, J., and Kirken, R. A. (2010). Uncoupling JAK3 Activation Induces Apoptosis in Human Lymphoid Cancer Cells via Regulating Critical Survival Pathways. *FEBS Lett.* 584, 1515–1520. doi:10.1016/j.febslet.2010.02.071

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27, 29–34. doi:10.1093/nar/27.1.29

Ohe, M., Hashino, S., and Hattori, A. (2012). Successful Treatment of Diffuse Large B-Cell Lymphoma With Clarithromycin and Prednisolone. *Korean J. Hematol.* 47, 293–297. doi:10.5045/kjh.2012.47.4.293

Oka, S., Ono, K., and Nohgawa, M. (2020). Clinical Effect of CD25 on the Prognosis of Diffuse Large B Cell Lymphoma With Secondary Central Nervous System Relapse. *Pathol. Oncol. Res.* 26, 1843–1850. doi:10.1007/s12253-019-00778-y

Peter, L., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *Bmc Bioinformatics.* 9, 559. doi:10.1186/1471-2105-9-559

Prakash, G., Sharma, A., Raina, V., Kumar, L., Sharma, M. C., and Mohanti, B. K. (2012). B Cell Non-Hodgkin's Lymphoma: Experience From a Tertiary Care Cancer center. *Ann. Hematol.* 91, 1603–1611. doi:10.1007/s00277-012-1491-5

Reddy, A., Zhang, J., Davis, N. S., Moffitt, A. B., Love, C. L., Waldrop, A., et al. (2017). Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell.* 171, 481–494. doi:10.1016/j.cell.2017.09.027

Rodrigues-Fernandes, C. I., Abreu, L. G., Radhakrishnan, R., Perez, D. E. D. C., Amaral-Silva, G. K., Gondak, R. D. O., et al. (2021). Prognostic Significance of CD30 Expression in Diffuse Large B-Cell Lymphoma: A Systematic Review with Meta-Analysis. *J. Oral Pathol. Med.* 50, 587. doi:10.1111/jop.13208

Rosas, I. O., Richards, T. J., Konishi, K., Zhang, Y., Gibson, K., Lokshin, A. E., et al. (2008). MMP1 and MMP7 as Potential Peripheral Blood Biomarkers in Idiopathic Pulmonary Fibrosis. *Plos Med.* 5, e93. doi:10.1371/journal.pmed.0050093

Schmitz, R., Wright, G. W., Huang, D. W., Johnson, C. A., Phelan, J. D., Wang, J. Q., et al. (2018). Genetics and Pathogenesis of Diffuse Large B-Cell Lymphoma. *N. Engl. J. Med.* 378, 1396–1407. doi:10.1056/nejmoa1801445

Shen, Q.-D., Zhu, H.-Y., Wang, L., Fan, L., Liang, J.-H., Cao, L., et al. (2018). Gemcitabine-Oxaliplatin Plus Rituximab (R-GemOx) as First-Line Treatment in Elderly Patients With Diffuse Large B-Cell Lymphoma: a Single-Arm, Open-Label, Phase 2 Trial. *Lancet Haematol.* 5, e261–e269. doi:10.1016/s2352-3026(18)30054-1

Shin, H. J., Kim, D. Y., Chung, J., Shin, K. H., and Lee, H. (2020). Prognostic Impact of Peripheral Blood T-Cell Subsets at the Time of Diagnosis on Survival in Patients With Diffuse Large B-Cell Lymphoma. *Acta Haematol.* 144 (4), 427–437. doi:10.1159/000510912

Souto, J. C., Alcolea, S., R, A., Remacha, A., Camacho, M., Soler, M., et al. (2014). Tumour Cell Lines HT-29 and FaDu Produce Proinflammatory Cytokines and Activate Neutrophils *In Vitro*: Possible Applications for Neutrophil-Based Antitumour Treatment. *Mediators Inflamm.* 2009, 817498. doi:10.1007/978-3-540-47648-1_287

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2020). Indicator Regularized Non-Negative Matrix Factorization Method-Based Drug Repurposing for COVID-19. *Front. Immunol.* 11, 603615. doi:10.3389/fimmu.2020.603615

Valerius, T., Repp, R., de Wit, T., Berthold, S., Platzer, E., Kalden, J., et al. (1993). Involvement of the High-Affinity Receptor for IgG (Fc Gamma RI; CD64) in Enhanced Tumor Cell Cytotoxicity of Neutrophils during Granulocyte Colony-Stimulating Factor Therapy. *Blood.* 82, 931–939. doi:10.1182/blood.v82.3.931.931

Von Buchholtz, L., Elischer, A., Tareilus, E., Gouka, R., Kaiser, C., Breer, H., et al. (2004). RGS21 Is a Novel Regulator of G Protein Signalling Selectively Expressed in Subpopulations of Taste Bud Cells. *Eur. J. Neurosci.* 19, 1535–1544. doi:10.1111/j.1460-9568.2004.03257.x

Wang, Z., Zhang, J., Luo, S., and Zhao, X. (2021). Prognostic Significance of Systemic Immune-Inflammation Index in Patients With Diffuse Large B-Cell Lymphoma. *Front. Oncol.* 11, 655259. doi:10.3389/fonc.2021.655259

Xu, X., Long, H., Xi, B., Ji, B., Li, Z., Dang, Y., et al. (2019). Molecular Network-Based Drug Prediction in Thyroid Cancer. *Int. J. Mol. Sci.* 20, 263. doi:10.3390/ijms20020263

Yang, J., Huang, T., Song, W. M., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016a). Discover the Network Underlying the Connections Between Aging and Age-Related Diseases. *Sci. Rep.* 6, 32566–32612. doi:10.1038/srep32566

Yang, J., Huang, T., Song, W.-m., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016b). Discover the Network Mechanisms Underlying the Connections between Aging and Age-Related Diseases. *Sci. Rep.* 6, 32566. doi:10.1038/srep32566

Yang, J., Huang, T., Huang, T., Petralia, F., Long, Q., Zhang, B., et al. (2015). Synchronized Age-Related Gene Expression Changes Across Multiple Tissues in Human and the Link to Complex Diseases. *Sci. Rep.* 5, 15145. doi:10.1038/srep15145

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human Geroprotector Discovery by Targeting the Converging Subnetworks of Aging and Age-Related Diseases. *Geroscience.* 42, 353–372. doi:10.1007/s11357-019-00106-x

Young, J., Ménétrey, J., and Goud, B. (2010). RAB6C Is a Retrogene that Encodes a Centrosomal Protein Involved in Cell Cycle Progression. *J. Mol. Biol.* 397, 69–88. doi:10.1016/j.jmb.2010.01.009

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, W., Long, H., He, B., and Yang, J. (2018). DECtp: Calling Differential Gene Expression Between Cancer and Normal Samples by Integrating Tumor Purity Information. *Front. Genet.* 9, 321. doi:10.3389/fgene.2018.00321

Zhou, S.-L., Zhou, Z.-J., Hu, Z.-Q., Huang, X.-W., Wang, Z., Chen, E.-B., et al. (2016). Tumor-Associated Neutrophils Recruit Macrophages and T-Regulatory Cells to Promote Progression of Hepatocellular Carcinoma and Resistance to Sorafenib. *Gastroenterology* 150, 1646–1658. e1617. doi:10.1053/j.gastro.2016.02.040

Zvi, G., Jing, S., and Samuel, K. (2009). Polarization of Tumor-Associated Neutrophil Phenotype by TGF-β: "N1" versus "N2" TAN. *Cancer Cell.* 16, 183–194. doi:10.1016/j.ccr.2009.06.017

Conflict of Interest: Authors BJ, BW and XS were employed by Geneis Beijing Co. Ltd.

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

The handling Editor declared a past co-authorship/collaboration with one of the authors BW.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fgene.2021.754503

A Novel Three-LncRNA Signature Predicting Tumor Recurrence in Nonfunctioning Pituitary Adenomas

*Sen Cheng[1†], Jing Guo[2†], Dawei Wang[2], Qiuyue Fang[2], Yulou Liu[2], Weiyan Xie[2], Yazhuo Zhang[1,2,3,4] and Chuzhong Li[1,2,3,4]**

[1]Department of Neurosurgery, Beijing Tiantan Hospital Affiliated to Capital Medical University, Beijing, China, [2]Beijing Neurosurgical Institute, Capital Medical University, Beijing, China, [3]Beijing Institute for Brain Disorders Brain Tumor Center, Beijing, China, [4]China National Clinical Research Center for Neurological Diseases, Beijing, China

The nonfunctioning pituitary adenoma (NFPA) recurrence rate is relatively high after surgical resection. Here, we constructed effective long noncoding RNA (lncRNA) signatures to predict NFPA prognosis. LncRNAs expression microarray sequencing profiles were obtained from 66 NFPAs. Sixty-six patients were randomly separated into a training (n = 33) and test group (n = 33). Univariable Cox regression and a machine learning algorithm was used to filter lncRNAs. Time-dependent receiver operating characteristic (ROC) analysis was performed to improve the prediction signature. Three lncRNAs (LOC101927765, RP11-23N2.4 and RP4-533D7.4) were included in a prognostic signature with high prediction accuracy for tumor recurrence, which had the largest area under ROC curve (AUC) value in the training/test group (AUC = 0.87/0.73). The predictive ability of the signature was validated by Kaplan-Meier survival analysis. A signature-based risk score model divied patients into two risk group, and the recurrence-free survival rates of the groups were significantly different (log-rank p < 0.001). In addition, the ROC analysis showed that the lncRNA signature predictive ability was significantly better than that of age in the training/testing/entire group (AUC = 0.87/0.726/0.798 *vs.* AUC = 0.683/0.676/0.679). We constructed and verified a three-lncRNA signature predictive of recurrence, suggesting potential therapeutic targets for NFPA.

Keywords: non-functioning pituitary adenoma (NFPA), recurrence, long noncoding RNAs, signature, machine learning

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Zhitao Jing,
The First Affiliated Hospital of China Medical University, China
Jian Zhang,
Harbin Medical University, China

****Correspondence:***
Chuzhong Li
lichuzhong@163.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: *06 August 2021*
Accepted: *04 October 2021*
Published: *20 October 2021*

Citation:
Cheng S, Guo J, Wang D, Fang Q, Liu Y, Xie W, Zhang Y and Li C (2021) A Novel Three-LncRNA Signature Predicting Tumor Recurrence in Nonfunctioning Pituitary Adenomas. Front. Genet. 12:754503. doi: 10.3389/fgene.2021.754503

INTRODUCTION

Pituitary adenoma (PA) is a common and benign intracranial tumor that occurs in the pituitary gland (Fernandez et al., 2010; Ostrom et al., 2015). It can be divided into functioning and nonfunctioning pituitary adenoma (FPA and NFPA, respectively) according to the presence or absence of hormone oversecretion and/or related clinical symptoms, like hyperthyroidism, acromegalic features, and hyperprolactinemia (Moreno et al., 2005). NFPAs account for 14–54% of PAs, and the annual incidence is 0.65–2.34 cases/100,000 (Raappana et al., 2010; Tjörnstrand et al., 2014; Al-Dahmani et al., 2016; Day et al., 2016). Due to the lack of typical symptoms related to hormone hypersecretion, NFPA is usually detected based on symptoms caused by tumor pressure on surrounding structures, such as headaches or visual impairment, or found incidentally on imaging

tests (Chen et al., 2011; Ntali and Wass, 2018). Surgical treatment is effective for NFPAs; however, total resection is not achievable for some tumors because they can invade the cavernous sinus or the area around the internal carotid artery (Meij et al., 2002; Shomali and Katznelson, 2002). Moreover, the recurrence rate of residual tumors reches 40 and 50% at 5 and 10 years, respectively, and even tumors that are completely resected have a recurrence rate of 10–20% after 5–10 years (Brochier et al., 2010; Chen et al., 2012; Sadik et al., 2017). Therefore, addressing the recurrence of NFPA is warranted. Currently, radiotherapy is considered to be effective in treating patients with residual or recurrent NFPA, although it may lead to progressive hypopituitarism and other long-term complications (Brada and Jankowska, 2008; Pollock et al., 2008). However, many questions remain about which subsets of NFPA patients are more likely to have recurrence and which subsets of residual tumors need to be further treated to prevent regrowth. Therefore, a method for predicting tumor recurrence after initial surgery is needed for early intervention.

Long noncoding RNAs (lncRNAs) are greater than 200 nt in length and have limited protein-coding ability (Moran et al., 2012). Emerging evidence suggests that lncRNAs regulate gene expression at the transcriptional and posttransciptional levels and that the dysfunction of lncRNAs contributes to the progression of many cancers, including PA (Poliseno et al., 2010; Wang and Chang, 2011; Huarte, 2015; Beylerli et al., 2020). Zhao et al. (2021) showed that downregulation of lncRNA PCAT6 could inhibit the proliferation, migration, viability, and invasion of PA cells by modulating the miR-139-3p/BRD4 axis . A study by D'Angelo et al. (2019) found that the lncRNA RPSAP52 promotes PA cell growth by acting as a microRNA (miRNA) sponge for HMGA proteins. The above studies verify that lncRNAs play a critical role in PA progression. Moreover, recent studies suggest that lncRNAs can be used to predict cancer prognosis and can as a signature in several cancers, such as oesophageal squamous cell carcinoma, gastric cancer, and hepatocellular carcinoma (Li et al., 2014; Zhu et al., 2016; Hong et al., 2020). However, the mechanism and prognostic value of lncRNAs in NFPA are still unclear. Therefore, it is necessary to find an appropriate lncRNA signature to accurately predict the recurrence of NFPA patients after surgery to provide early intervention.

In this study, tumor recurrence refers to regrowth of residual tumor cells and tumor relapse after total resection. We analyzed the expression of lncRNAs in 66 NFPA patients through microarray sequencing and identified genes associated with tumor recurrence. We aimed to develop and validate a useful multi-lncRNA prediction model that may be used to evaluate recurrence and guide treatment after surgical resection in patients with NFPA.

METHODS

Patients and Samples

From October 2007 to July 2014, patients who were diagnosed with NFPA and underwent surgical resection at Beijing Tiantan Hospital were included in this study (n = 66). The mean age of these 66 patients was 51.5 years (range, 25–73), there were 34 males and 32 females, and the median follow-up was 76.5 months (range, 5–122). The clinical and pathological characteristics of all the patients are shown in **Supplementary Table S1**. Cavernous sinus (CS) invasion was defined by the Knosp grading scale (grade 3 or 4) on preoperative enhanced magnetic resonance imaging (MRI) (Knosp et al., 1993). Postoperative tumor recurrence was defined as recurrence identified from any direction on enhanced MRI from the day of surgery to the end of the follow up; the maximum tumor diameter needed to increase by > 2 mm. According to tumor size, NFPA were divided into microadenoma (<10 mm in diameter), macroadenoma (≥10 mm) and giant adenoma (≥40 mm). The local Ethics Committee approved this study, and informed consent was obtained from each subject.

Total RNA Extraction

According to the instructions provided, total RNA was extracted and purified from collected samples using the phenol-free mirVana™ miRNA Isolation Kit (Cat # AM1561; Ambion; Thermo Fisher Scientific, Inc.). A Thermo Scientific™ NanoDrop 2000 was used to quantify and assess purity of the extracted RNA.

RNA Microarray Analysis

RNA samples were used to generate fluorescence-labeled cRNA targets for the SBC human ceRNA array V1.0 (4 × 180 K) and were subsequently hybridized with slides and scanned in an Agilent Microarray Scanner (Agilent Technologies, Santa Clara, CA, United States) to obtain the data. The raw data was extracted using feature extraction software 10.7 (Agilent Technologies, Inc.). Then, the quantile algorithm provided by the "limma" package (http://bioconductor.org/packages/limma/) of the R program was used to normalize the data.

Identification of Prognostic LncRNAs

The "sample" function of R progrom (www.r-project.org/) was used to randomly divided 66 NPFA patients into a training set (n = 33) and a testing set (n = 33). In the training group, univariable Cox proportional hazards regression analysis was performed to determine the association between recurrence-free survival (RFS) and lncRNA expression in each patient. We used a machine learning approach, random survival forests-variable hunting (RSFVH) algorithm, to narrow the scope of the gene set through an iteration procedure, discarding the bottom quarter of lncRNAs (the least important lncRNAs) at each step. In total, nine lncRNAs were selected (Mogensen et al., 2012; Li et al., 2014; Ishwaran and Lu, 2019).

Construction of Prognostic LncRNA Signature

The selected lncRNAs was used to construct a risk prediction score model as follows (Ritchie et al., 2015; Guo et al., 2016).

$$\text{Risk Score (RS)} = \sum_{i=1}^{N} (\text{Explg} * \text{Coef})$$

In this formula, N represent the number of prognostic lncRNA, Explg represents the expression value of lncRNA, and Coef

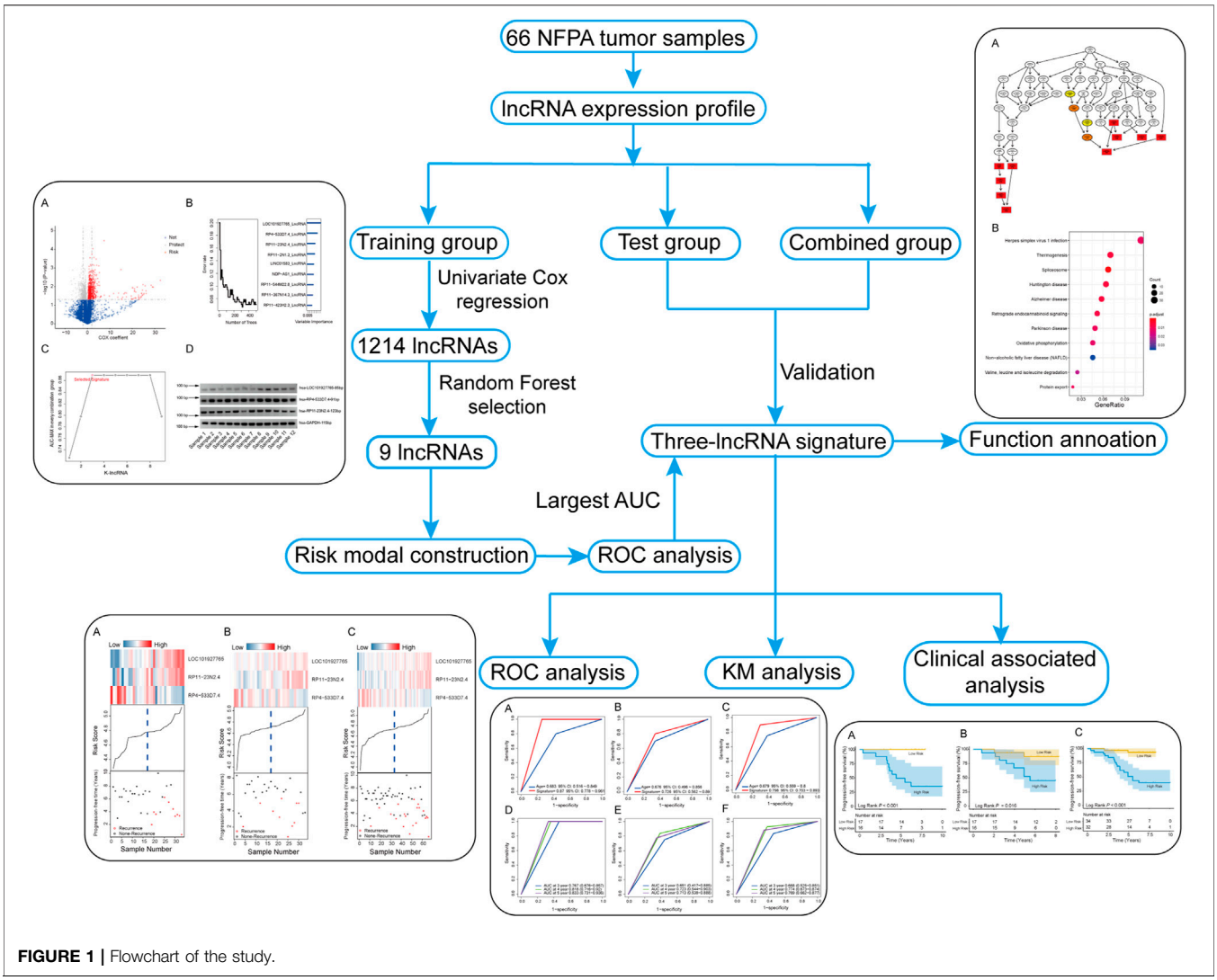


FIGURE 1 | Flowchart of the study.

represents the estimated regression coefficient of the lncRNA in the univariable Cox regression analysis.

Since the nine selected lncRNAs could form $2^9-1 = 511$ combinations or signature, each patient received 511 risk scores. Then, in the training dataset, the sensitivity and specificity of the 511 signatures were analyzed by the time-dependent receiver operating characteristic (ROC) curves. The prognostic signature was obtained by comparing the area under the ROC curve (AUC) values.

Validation the Reliability of Microarray Data by RT-PCR

To verify the existence of the lncRNA signature, twelve samples were randomly selected from the entire group for RT-PCR and agarose gel electrophoresis. LncRNA reverse transcription was performed using a High Capacity cDNA Reverse Transcription Kit (0049472, Thermo Fisher). Next, PCR was performed using I-5TM High-Fidelity Master Mix (I5HM, 200MCLAB). PCR was conducted as follows: 2 min of initial denaturation at 98°C, 32 cycles of 10 s at 98°C, 58°C for 10 s and 72°C for 10s, and final extension step for 5 min at 72°C. GAPDH was used as an internal control gene. The PCR products were run on 2% agarose gel and visualized using a UV transilluminator. The primer sequences are presented in **Supplementary Table S2**.

Statistical Analysis

The survival distribution of different groups was evaluated and compared using Kaplan-Meier survival analyses and two-sided log-rank tests. The chi-square test was used to analyzed the associations with clinical signatures. $p < 0.05$ was considered to indicate statistical significance. All analyses were performed using R program 3.6.1. The packages were downloaded from Bioconductor, including the survival, ROC, and randomForestSRC packages.

Functional Enrichment Analysis of LncRNAs With Prognostic Value

To investigate the potential function of the lncRNAs in the signature, Pearson correlation tests were used to identified

TABLE 1 | Clinical Data of the included tumors.

	Entire set (n)	Training set (n)	Test set (n)
Gender			
Male	32	18	14
Female	34	15	19
Age (years)			
≤52	38	19	19
>52	28	14	14
Tumor size classification			
Macro	47	24	23
Giant	19	9	10
CS Invasion			
Yes	38	20	18
No	28	13	15
Headache			
Yes	31	14	17
No	35	19	16
Vision and visual field disorders			
Yes	50	26	24
No	16	7	9
Recurrence			
Yes	20	10	10
No	46	23	23

CS, cavernous sinus; Giant, giant adenoma; Macro, macroadenoma.

protein-coding genes (PCGs) coexpressed with the prognostic lncRNAs. The genes with a $p < 0.05$ and an absolute value of the Pearson coefficient > 0.6 were selected for Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology, 2017) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa et al., 2016; Kanehisa et al., 2017) enrichment analyses. The GO and KEGG analyses were performed with the clusterProfiler package (Yu et al., 2012) of the R program.

RESULTS

Identification of LncRNA Signatures for the Prediction of NFPA Recurrence

A total of 19,741 lncRNAs were extracted from the 66 NFPA expression profiles. The flow chart of this study is shown in **Figure 1**. The patient information of all patients is summarized in **Table 1**.

Initially, in the training set, univariate Cox proportional hazards regression analysis was used to obtain RFS-related lncRNAs. The 1,214-lncRNA set was identified using recurrence as the dependent variable, and the signature was significantly associated with patient recurrence (**Supplementary Table S3**, p value < 0.05, **Figure 2A**).

Secondly, to further reduce the number of prognostic lncRNAs, the random forest supervised cassification (RFSC) algorithm was employed to analyze the 1,214-lncRNA set, and the nine lncRNAs most related to recurrence were obtained according to the permutation important score calculated with the RFSC algorithm (**Figure 2B**; **Supplementary Figure S1**).

Thirdly, based on the nine types of lncRNA, we constructed a risk-score model of 2^9–1 (511) types of lncRNA set combinations, which contained different lncRNA numbers from 1 to 9. To screen for a better prediction signature, we conducted a time-dependent ROC analysis that used recurrence status as a lable and signature risk scores as a variable in the training group and compared the sensitivities and specificities (**Supplementary Table S4**).

According to the AUC values of all 511 signatures (**Supplementary Table S4**), we identified the lncRNA combination composed of LOC101927765, RP11-23N2.4, and RP4-533D7.4 as the most promising, as it had strong ability to predict recurrence and the smallest node and the largest AUC value of 0.87 (**Figure 2C**; **Table 2**). RT-PCR was used to confirm the reliability of microarray sequencing. Consisting with the microarray data, the three lncRNAs were detected in 12 tumor tissues (**Figure 2D**), which revealed that the lncRNA are stable and can be used as prognostic maker.

The risk score of the signature was calculated as follows: risk score = (3.41 × expression value of LOC101927765) + (1.90 × expression value of RP11-23N2.4) + (−3.43 × expression value of RP4-533D7.4).

Validation the Prediction Ability of the Three LncRNA Signature

Each patient obtains a risk score according to the risk score model. Then, the patients from the training group were divided into a high-risk group (n = 16) and a low-risk group (n = 17) based on the cutoff point, which was the median risk score. Kaplan-Meier survival analysis was performed to determined the difference in RFS between the two risk groups. The median RFS time was significantly shorter in the high-risk group (4.44 years) than in the low-risk group (6.74 years) ($p < 0.001$; log-rank test, **Figure 3A**). Moreover, the recurrence rate of the high-risk group was higher than that of the low-risk group (>60% *vs.* < 1%). In a similar manner, patients from the test group were also divided into two risk groups. The results of Kaplan-Meier analyses for the high-risk (n = 16) and low-risk (n = 17) groups in the test dataset were plotted and are shown in **Figure 3B** (median RFS time: 5.51 *vs.* 6.82 years; log-rank test, p = 0.016), and the RFS rates were approximately 52.25 and 87.40%, respectively. In addition, patients in the entire group were similary divided into high-risk (n = 32) and low-risk (n = 34) groups, and Kaplan-Meier analysis further confirmed the ability of the lncRNA signature to predict recurrence (median PFS time: 4.97 *vs.* 6.79 years; log-rank test, $p < 0.001$, **Figure 3C**).

Figures 4A–C intuitively shows the risk score, survival status and expression pattern of lncRNAs in the training, testing, and independent datasets. For patients with low risk scores in the three datasets, RP4−533D7.4 was highly expressed, while LOC101927765 and RP11-23N2.4 was expressed at low levels; the opposite patterns for each lncRNA were seen in patients with high risk scores.

The Value of the LncRNA Signature is Independent of Traditional Clinical Features

After proving the recurrence prediction ability of the lncRNA signature, we explored the correlation between the signature and

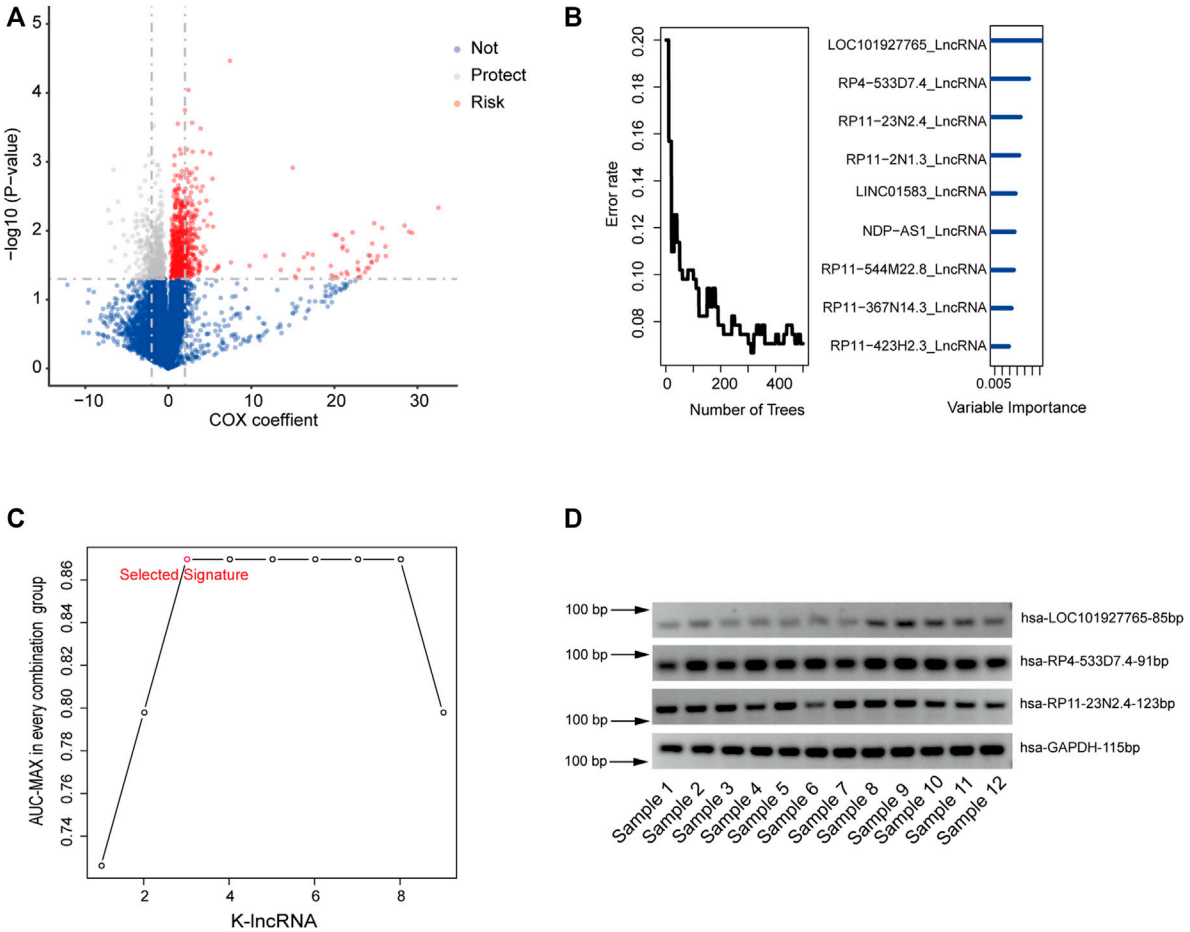


FIGURE 2 | Identifying the LncRNA signature in the training dataset. **(A)**, Univariate Cox proportional hazards regression analysis of the lncRNA expression profiling data in the training dataset. **(B)**, Identifying the lncRNAs by RFSC algorithm. **(C)**. The AUC of all 511 signatures were calculated and the nine highest AUC for k = 1, 2. . .9 is shown in the plot by ROC analysis for the lncRNA signatures predicting model in the training dataset. **(D)**, Agarose electrophoresis of selected lncRNAs PCR products.

TABLE 2 | Identities of PCG and lncRNAs in the prognostic expression signature and their univariable cox association with prognosis.

Gene symbol	Coefficient[a]	*p* Value[a]	Gene expression level association with poor prognosis
LOC101927765	3.406	0.001	high
RP11-23N2.4	1.895	0.007	high
RP4-533D7.4	−3.440	0.002	low

[a]*Derived from the univariable Cox regression analysis in the training set.*

clinical characteristics in the entire dataset ($n = 66$) to understand the clinical significance of the lncRNA signature.

Table 3 shows that there was an association between the lncRNA signature and age in the entire group (chi-suqre test, $p = 0.03$, **Table 3**). In addition, we further assessed whether the prognostic value of the three-lncRNA signature was independent of other clinical factors. Univariate and multivariate Cox regression analyses of factors including age, sex, tumor size classification, CS invasion, and the signature were performed. In the entire dataset, age (HR = 0.33, 95% CI = 0.12–0.93, $p = 0.04$) and the signature risk score (HR = 1.50, 95% CI = 1.24–1.82, $p < 0.001$) were significantly associated with the RFS of patients (**Table 4**). Moreover, the three-signature score was also an indenpent prognostic factor associated with RFS in the training (HR = 2.06, 95% CI = 1.36–3.12, $p < 0.001$) and test set (HR = 6.96, 95% CI = 1.21–40.16, $p = 0.03$). Hence, the results indicate that the three-lncRNA signature is an independent prognostic factor for NFPA RFS.

Comparison of the Predictive Power of the LncRNA Signature and Age

It has been reported that age is associated with a risk of tumor recurrence (Losa et al., 2008). ROC analysis was performed to determine the predictive power of the lncRNA signature and age. The results showed that in the training/testing/entire group, the AUC values of the lncRNA signature were lager those of age (AUC = 0.87/0.726/0.798 *vs.* AUC = 0.683/0.676/0.679, **Figures**

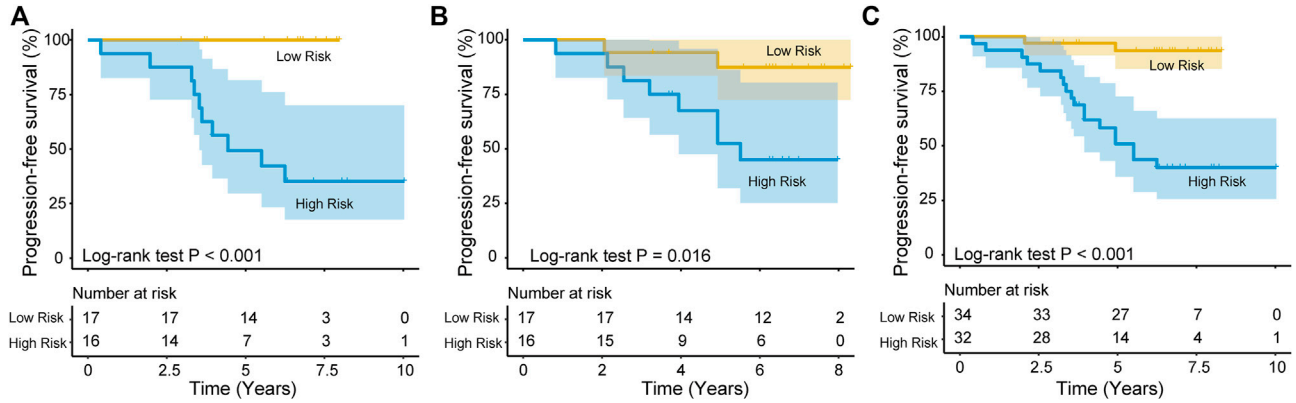


FIGURE 3 | The lncRNA signature for predicting recurrence PFS of patients with NFPA. Kaplan–Meier survival curves of patients classified into high- and low-risk groups using the lncRNA signature in the training **(A)**, test **(B)** and entire dataset **(C)**.

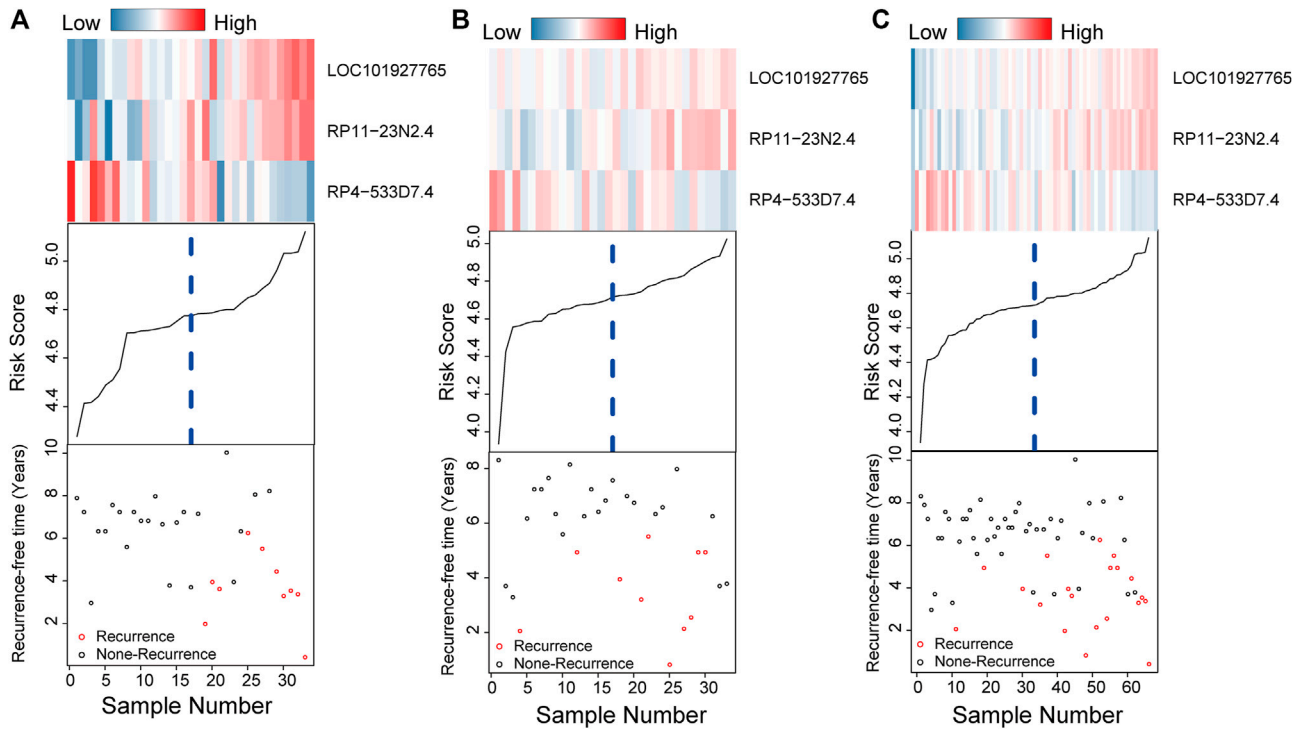


FIGURE 4 | Risk score distribution, survival status and gene expression patterns for patients in high- and low-risk cluster grouped by the three-lncRNA signature in the training **(A)**, testing **(B)**, and independent datasets **(C)**.

5A–B), indicating that the signature had high accuracy and important clinical significance. In addition, time-dependent ROC analysis was performed on the three datasets to further understand the signature prediction capabilities for 3-, 4- and 5 year RFS. The signature AUC values in the training/test/entire group at 3, 4, and 5 years, as shown in **Figures 5D–F**, indicated a strong predictive power of the signature for RFS (AUC = 0.767/0.818/0.833, 0.651/0.723/0.713, and 0.688/0.774/0.769, respectively).

Functional Enrichment Analysis of Genes Associated with the Prognostic LncRNAs in the Signature

The PCGs correlated with the lncRNAs in our prognostic signature were obtained by Pearson correlation analysis in all 66 patients, and their potential biological function were explored. The expression of 1,056 PCGs was highly correlated with that of at least one of the LncRNAs (Pearson correlation coefficient >

TABLE 3 | Association of the signature with clinicopathological characteristics in Pituitary adenoma patients.

Variables	Training			Test			Entire		
	Low risk	High risk	P	Low risk	High risk	P	Low risk	High risk	P
Sex			1.00			0.36			0.62
Female	8	7		8	11		16	18	
Male	9	9		9	5		18	14	
Age			0.21			0.12			0.03
≤52	7	11		5	10		12	21	
>52	10	5		12	6		22	11	
Tumor size classification			0.50			0.40			0.70
Giant	6	3		5	5		11	8	
Macro	11	13		12	11		23	24	
Invasion			1.00			0.21			0.43
No	6	6		10	5		16	11	
Yes	11	10		7	11		18	21	

Data were analyzed using the Chi-squared test; p-*value < 0.05 was considered to indicate a statistically significant difference.*

TABLE 4 | Univariable and multivariable Cox regression analysis of the signature and survival of NFPA patients in the training, test group and entire group.

Variables		Univariable analysis				Multivariable analysis			
		HR	95% CI of HR		P	HR	95% CI of HR		P
			Lower	Upper			Lower	Upper	
Training set (*n* = 33)									
Age	>52 *vs.*≤52	0.23	0.05	1.09	0.06	0.18	0.03	1.08	0.06
Sex	Male *vs*. Female	1.05	0.29	3.74	0.94	0.73	0.17	3.11	0.68
Tumor size classification	Macro *vs*. Giant	1.10	0.23	5.18	0.91	1.19	0.18	7.72	0.85
CS invasion	Yes *vs*. No	1.40	0.36	5.40	0.63	1.37	0.28	6.59	0.70
Signature	High risk *vs*. low risk	2.03	1.40	2.94	<0.001	2.06	1.36	3.12	<0.001
Test set (*n* = 33)									
Age	>52 *vs.*≤52	0.35	0.09	1.36	0.13	0.62	0.15	2.67	0.52
Sex	Male *vs*. Female	0.53	0.14	2.03	0.35	0.78	0.17	3.72	0.76
Tumor size classification	Macro *vs*. Giant	0.31	0.09	1.08	0.07	0.15	0.02	0.88	0.04
CS invasion	Yes *vs*. No	2.05	0.53	7.92	0.30	0.43	0.06	3.29	0.42
Signature	High risk *vs*. low risk	5.49	1.16	14.92	0.03	6.96	1.21	40.16	0.03
Entire set (*n* = 66)									
Age	>52 *vs.*≤52	0.29	0.10	0.79	0.02	0.33	0.12	0.93	0.04
Sex	Male *vs*. Female	0.75	0.31	1.81	0.52	0.89	0.36	2.18	0.80
Tumor size classification	Macro *vs*. Giant	1.21	0.82	1.78	0.33	0.59	0.21	1.67	0.32
CS invasion	Yes *vs*. No	1.72	0.66	4.48	0.26	1.23	0.42	3.54	0.71
Signature	High risk *vs*. low risk	1.49	1.24	1.80	<0.001	1.50	1.24	1.82	<0.001

CS, cavernous sinus; Giant, giant adenoma; Macro, macroadenoma.

0.60, $p < 0.05$, **Supplementary Table S5**). Next, we performed GO and KEGG analyses and found these the genes were enriched in 99 different terms (**Supplementary Table S6**), such as mRNA processing, RNA splicing and oxidative phosphorylation (**Figure 6**).

DISCUSSION

The prevalence of NFPA ranges from 7 to 41.3 cases per 100,000 population, and it is the second most common type of adenomas after prolactinomas (Ntali and Wass, 2018). Despite NFPA being a histologically benign tumor and advances in endosopic techniques, the recueence rate of NFPA is relatively high (Batista et al., 2018). Therefore, it is necessary to accurately predict tumor recurrence after NFPA surgery to obtain the most effective and accurate treatment plan. Herein, we constructed a three-lncRNA signature to predict the prognostic of NFPAs and verifies its predictive power.

First, we obtained 19,741 lncRNA expression profiles by sequencing 66 NFPAs and identified 1,214 lncRNAs that were significantly related to RFS in NFPA in the training set. RSFVH algorithm, a machine learning method, was used to narrow down the number of RFS-related lncRNAs to 9. A three-lncRNAs (LOC101927765, RP11-23N2.4, and RP4-533D7.4) signature with the highest AUC value of 511signatures, which contained combinations of 1–9 different lncRNAs, was identified. The risk model of the signature was constructed basde on the three lncRNAs. Second, patients were divided into two risk group in the training and testing sets, and the recurrence prediction power

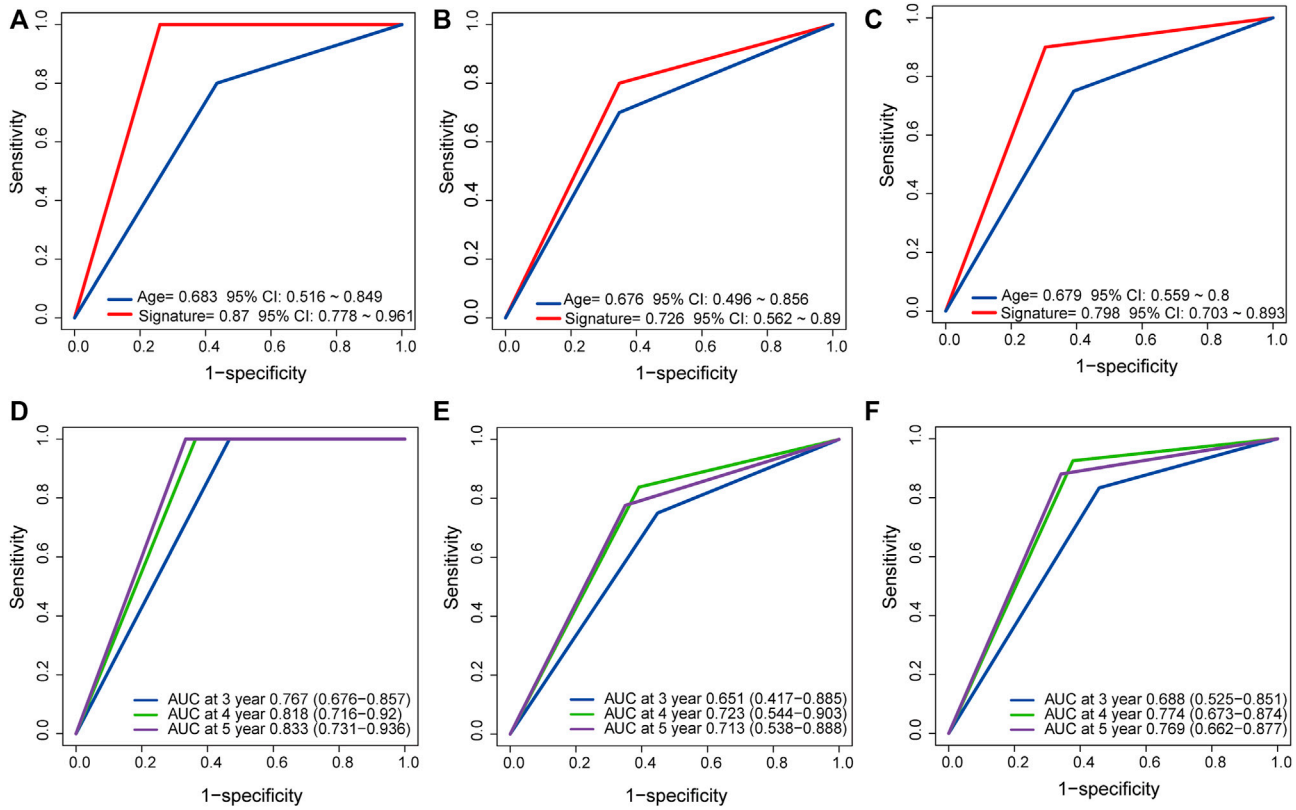


FIGURE 5 | Compare prediction power of the lncRNA signature to that of Age by ROC in the training, test, and entire dataset **(A, B, C)** and TimeROC analysis for the signature at 3,4, and 5 years in the three sets **(D, E, F)**.

was validated by Kaplan-Meier analysis. Third, the three-lncRNA signature-based risk score was identified as a prognostic factor independent of clinical features like sex, tumor size classification, CS invasion. Age is a controversial factor related to recurrence in NFPA. Batista et al. (2018) showed that recurrence of NFPA was not associated with age while Subramanian and indicated that older age at surgery was related to a lower risk of recurrence (Lyu et al., 2021; Subramanian et al., 2021). Even so, the ROC analysis showed that the predictive ability of the three-lncRNA signature was better than that of age. Finally, we explore the potential biological function of the three lncRNAs through functional enrichment analysis of coexpressed PCGs, which were identified as related to the three lncRNAs by Perason correlation analysis.

In recent years, lncRNAs has been considered potential prognostic markers and therapectic targets for cancers (Sanchez Calle et al., 2018; Zhang et al., 2021). Liu et al. (2020) found that lncCSMD1-1 is overexpressed in hepatocellular carcinoma (HCC) and interacts with the MYC protein to promote tumor progression, suggesting that it may serve as a prognostic marker for HCC. The lncRNA PiHL (RP11-382A18.2) is upregulated in colorectal cancer (CRC), and its upregulation is an independent predictor of poor CRC prognosis (Deng et al., 2020). In addition, lncRNA also play a crucial role in PA progression. Wang et al. (2019) demonstrated that the lncRNA clarin 1 antisense RNA 1 (CLRNA-AS1) was expressed at low levels in prolactinoma and inhibited cell proliferation and autophagy Moreover, lncRNA-H19 is downregulated in PA and negatively correlated with tumor progression (Wu et al., 2018). Therefore, lncRNAs may be developed into a prognostic makers of PA. Recently, an increasing number of studies have identified several lncRNAs that can be studied to predict cancer prognostic. Meng et al. (2014) identified four lncRNA genes (U79277, AK024118, BC040204, AK000974) that can be used to predict breast cancer survival. Jiang et al. (2020) found that three-lncRNA (LINC02434, AL139327.2, and AC126175.1) could be used to predict prognosis in head and neck squamous cell cancer However, these studies did not confirm the reliability of the lncRNAs in tumor samples. In the present study, to avoid false positives in sequencing data, RT-PCR was performed to verify the reliability of the three lncRNA.

There are some limitations in this study that need to be acknowledged. First, potential lncRNAs may have been overlooked because the study only included 19,741 lncRNAs, which is only a small fraction of human lncRNAs. Second, the construction and evaluation of the model were based on the limited NFPAs samples, and more external samples are needed to verify the prediction power. Third, further *in vivo* and *in vitro* experiment need to be performed to elucidate the mechanisms and potential functions of the three lncRNAs.

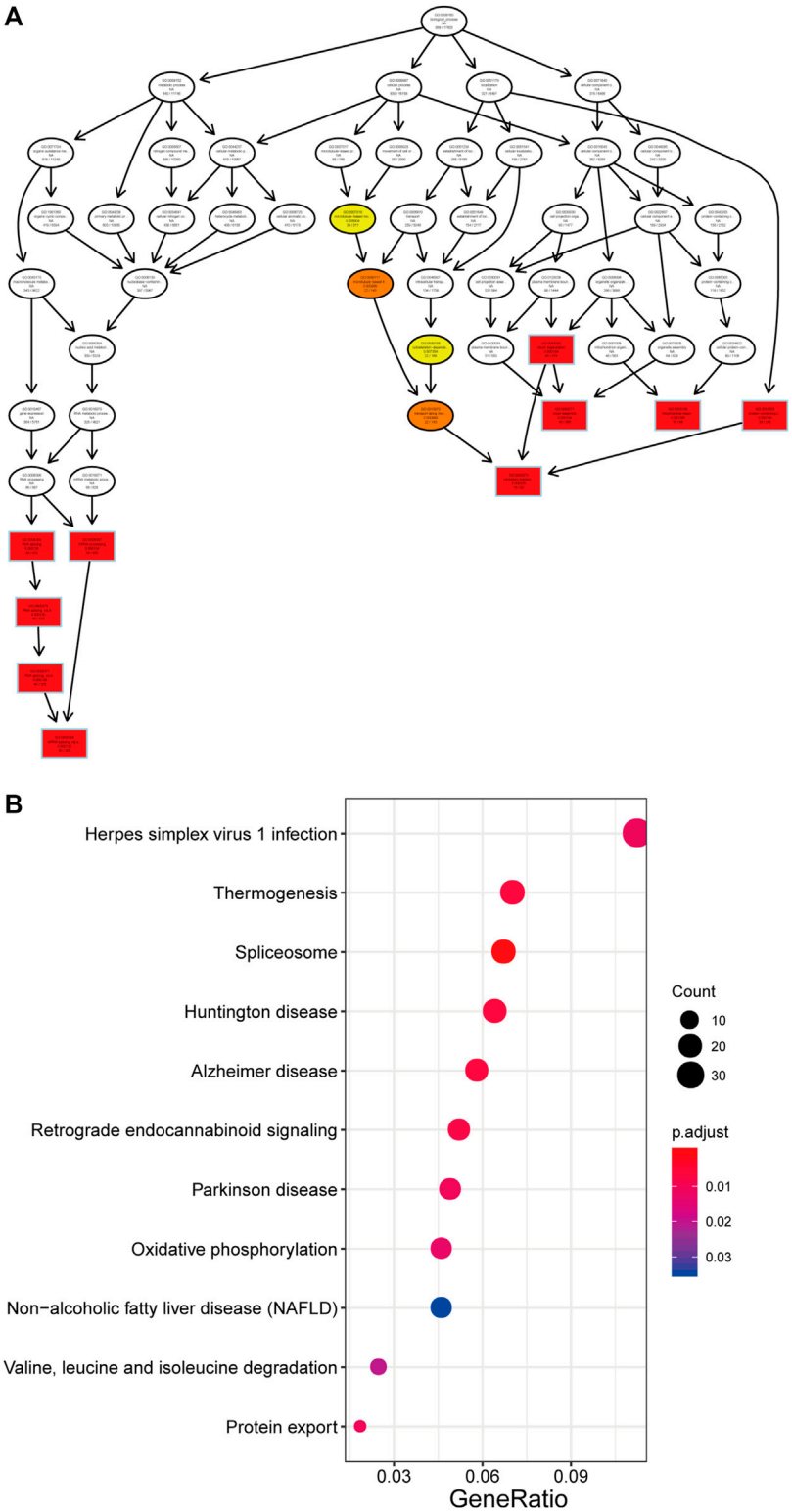


FIGURE 6 | Function of the three lncRNA for GO **(A)** and KEGG **(B)** analysis by clusterProfiler.

In summary, we constructed a three-lncRNAs signature that could serve as a precise predictive biomarker for NFPAs. In addition, patients identified by the 3-lncRNA signature to be at high risk of NFPA after surgery could benefit from early and accurate intervention.

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 Ethics Committee of Beijing Tiantan Hospital, Capital Medical University . The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WX, YZ, and CL worked on the conception and designed the research. DW, QF, and YL were involved in the collection and analysis of patients' clinical data. SC and JG were dedicated to data analysis, interpretation, and drafting. All authors read and approved the final manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (Grant codes: 81672495, 81771489, 82072804, 82071559, and 82071558).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.754503/full#supplementary-material

REFERENCES

Al-Dahmani, K., Mohammad, S., Imran, F., Theriault, C., Doucette, S., Zwicker, D., et al. (2016). Sellar Masses: An Epidemiological Study. *Can. J. Neurol. Sci.* 43, 291–297. doi:10.1017/cjn.2015.301

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Batista, R. L., Trarbach, E. B., Marques, M. D., Cescato, V. A., Da Silva, G. O., Herkenhoff, C. G. B., et al. (2018). Nonfunctioning Pituitary Adenoma Recurrence and its Relationship with Sex, Size, and Hormonal Immunohistochemical Profile. *World Neurosurg.* 120, e241–e246. doi:10.1016/j.wneu.2018.08.043

Beylerli, O., Gareev, I., Pavlov, V., Chen, X., and Zhao, S. (2020). The Role of Long Noncoding RNAs in the Biology of Pituitary Adenomas. *World Neurosurg.* 137, 252–256. doi:10.1016/j.wneu.2019.10.137

Brada, M., and Jankowska, P. (2008). Radiotherapy for Pituitary Adenomas. *Endocrinol. Metab. Clin. North America* 37, 263–275. doi:10.1016/j.ecl.2007.10.005

Brochier, S., Galland, F., Kujas, M., Parker, F., Gaillard, S., Raftopoulos, C., et al. (2010). Factors Predicting Relapse of Nonfunctioning Pituitary Macroadenomas after Neurosurgery: A Study of 142 Patients. *Eur. J. Endocrinol.* 163, 193–200. doi:10.1530/eje-10-0255

Chen, L., White, W. L., Spetzler, R. F., and Xu, B. (2011). A Prospective Study of Nonfunctioning Pituitary Adenomas: Presentation, Management, and Clinical Outcome. *J. Neurooncol.* 102, 129–138. doi:10.1007/s11060-010-0302-x

Chen, Y., Wang, C. D., Su, Z. P., Chen, Y. X., Cai, L., Zhuge, Q. C., et al. (2012). Natural History of Postoperative Nonfunctioning Pituitary Adenomas: A Systematic Review and Meta-Analysis. *Neuroendocrinology* 96, 333–342. doi:10.1159/000339823

D'angelo, D., Mussnich, P., Sepe, R., Raia, M., Del Vecchio, L., Cappabianca, P., et al. (2019). RPSAP52 lncRNA Is Overexpressed in Pituitary Tumors and Promotes Cell Proliferation by Acting as miRNA Sponge for HMGA Proteins. *J. Mol. Med. (Berl)* 97, 1019–1032. doi:10.1007/s00109-019-01789-7

Day, P. F., Loto, M. G., Glerean, M., Picasso, M. F. R., Lovazzano, S., and Giunta, D. H. (2016). Incidence and Prevalence of Clinically Relevant Pituitary Adenomas: Retrospective Cohort Study in a Health Management Organization in Buenos Aires, Argentina. *Arch. Endocrinol. Metab.* 60, 554–561. doi:10.1590/2359-3997000000195

Deng, X., Li, S., Kong, F., Ruan, H., Xu, X., Zhang, X., et al. (2020). Long Noncoding RNA PiHL Regulates P53 Protein Stability through GRWD1/RPL11/MDM2 Axis in Colorectal Cancer. *Theranostics* 10, 265–280. doi:10.7150/thno.36045

Fernandez, A., Karavitaki, N., and Wass, J. A. H. (2010). Prevalence of Pituitary Adenomas: A Community-Based, Cross-Sectional Study in Banbury (Oxfordshire, UK). *Clin. Endocrinol. (Oxf)* 72, 377–382. doi:10.1111/j.1365-2265.2009.03667.x

Guo, J.-C., Li, C.-Q., Wang, Q.-Y., Zhao, J.-M., Ding, J.-Y., Li, E.-M., et al. (2016). Protein-coding Genes Combined with Long Non-coding RNAs Predict Prognosis in Esophageal Squamous Cell Carcinoma Patients as a Novel Clinical Multi-Dimensional Signature. *Mol. Biosyst.* 12, 3467–3477. doi:10.1039/c6mb00585c

Hong, W., Liang, L., Gu, Y., Qi, Z., Qiu, H., Yang, X., et al. (2020). Immune-Related lncRNA to Construct Novel Signature and Predict the Immune Landscape of Human Hepatocellular Carcinoma. *Mol. Ther. - Nucleic Acids* 22, 937–947. doi:10.1016/j.omtn.2020.10.002

Huarte, M. (2015). The Emerging Role of lncRNAs in Cancer. *Nat. Med.* 21, 1253–1261. doi:10.1038/nm.3981

Ishwaran, H., and Lu, M. (2019). Standard Errors and Confidence Intervals for Variable Importance in Random forest Regression, Classification, and Survival. *Stat. Med.* 38, 558–582. doi:10.1002/sim.7803

Jiang, H., Ma, B., Xu, W., Luo, Y., Wang, X., Wen, S., et al. (2020). A Novel Three-lncRNA Signature Predicts the Overall Survival of HNSCC Patients. *Ann. Surg. Oncol.* doi:10.1245/s10434-020-09210-1

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* 45, D353–d361. doi:10.1093/nar/gkw1092

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070

Knosp, E., Steiner, E., Kitz, K., and Matula, C. (1993). Pituitary Adenomas with Invasion of the Cavernous Sinus Space. *Neurosurgery* 33, 610–618. doi:10.1227/00006123-199310000-00008

Li, J., Chen, Z., Tian, L., Zhou, C., He, M. Y., Gao, Y., et al. (2014). LncRNA Profile Study Reveals a Three-lncRNA Signature Associated with the Survival of Patients with Oesophageal Squamous Cell Carcinoma. *Gut* 63, 1700–1710. doi:10.1136/gutjnl-2013-305806

Liu, J., Xu, R., Mai, S.-J., Ma, Y.-S., Zhang, M.-Y., Cao, P.-S., et al. (2020). LncRNA CSMD1-1 Promotes the Progression of Hepatocellular Carcinoma by Activating MYC Signaling. *Theranostics* 10, 7527–7544. doi:10.7150/thno.45989

Losa, M., Mortini, P., Barzaghi, R., Ribotto, P., Terreni, M. R., Marzoli, S. B., et al. (2008). Early Results of Surgery in Patients with Nonfunctioning Pituitary Adenoma and Analysis of the Risk of Tumor Recurrence. *J. Neurosurg.* 108, 525–532. doi:10.3171/jns/2008/108/3/0525

Lyu, W., Fei, X., Chen, C., and Tang, Y. (2021). Nomogram Predictive Model of post-operative Recurrence in Non-Functioning Pituitary Adenoma. *Gland Surg.* 10, 807–815. doi:10.21037/gs-21-47

Meij, B. P., Lopes, M.-B. S., Ellegala, D. B., Alden, T. D., and Laws, E. R., Jr. (2002). The Long-Term Significance of Microscopic Dural Invasion in 354 Patients with Pituitary Adenomas Treated with Transsphenoidal Surgery. *J. Neurosurg.* 96, 195–208. doi:10.3171/jns.2002.96.2.0195

Meng, J., Li, P., Zhang, Q., Yang, Z., and Fu, S. (2014). A Four-Long Non-Coding RNA Signature in Predicting Breast Cancer Survival. *J. Exp. Clin. Cancer Res.* 33, 84. doi:10.1186/s13046-014-0084-7

Mogensen, U. B., Ishwaran, H., and Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *J. Stat. Softw.* 50, 1–23. doi:10.18637/jss.v050.i11

Moran, V. A., Perera, R. J., and Khalil, A. M. (2012). Emerging Functional and Mechanistic Paradigms of Mammalian Long Non-Coding RNAs. *Nucleic Acids Res.* 40, 6391–6400. doi:10.1093/nar/gks296

Moreno, C. S., Evans, C.-O., Zhan, X., Okor, M., Desiderio, D. M., and Oyesiku, N. M. (2005). Novel Molecular Signaling and Classification of Human Clinically Nonfunctional Pituitary Adenomas Identified by Gene Expression Profiling and Proteomic Analyses. *Cancer Res.* 65, 10214–10222. doi:10.1158/0008-5472.can-05-0884

Ntali, G., and Wass, J. A. (2018). Epidemiology, Clinical Presentation and Diagnosis of Non-Functioning Pituitary Adenomas. *Pituitary* 21, 111–118. doi:10.1007/s11102-018-0869-3

Ostrom, Q. T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., et al. (2015). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro Oncol.* 17 (Suppl. 4), iv1–iv62. doi:10.1093/neuonc/nov189

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A Coding-independent Function of Gene and Pseudogene mRNAs Regulates Tumour Biology. *Nature* 465, 1033–1038. doi:10.1038/nature09144

Pollock, B. E., Cochran, J., Natt, N., Brown, P. D., Erickson, D., Link, M. J., et al. (2008). Gamma Knife Radiosurgery for Patients with Nonfunctioning Pituitary Adenomas: Results from a 15-year Experience. *Int. J. Radiat. Oncol. Biol. Phys.* 70, 1325–1329. doi:10.1016/j.ijrobp.2007.08.018

Raappana, A., Koivukangas, J., Ebeling, T., and Pirilä, T. (2010). Incidence of Pituitary Adenomas in Northern Finland in 1992-2007. *J. Clin. Endocrinol. Metab.* 95, 4268–4275. doi:10.1210/jc.2010-0537

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

Sadik, Z. H. A., Voormolen, E. H. J., Depauw, P. R. A. M., Burhani, B., Nieuwlaat, W. A., Verheul, J., et al. (2017). Treatment of Nonfunctional Pituitary Adenoma Postoperative Remnants: Adjuvant or Delayed Gamma Knife Radiosurgery? *World Neurosurg.* 100, 361–368. doi:10.1016/j.wneu.2017.01.028

Sanchez Calle, A., Kawamura, Y., Yamamoto, Y., Takeshita, F., and Ochiya, T. (2018). Emerging Roles of Long Non-coding RNA in Cancer. *Cancer Sci.* 109, 2093–2100. doi:10.1111/cas.13642

Shomali, M. E., and Katznelson, L. (2002). Medical Therapy of Gonadotropin-Producing and Nonfunctioning Pituitary Adenomas. *Pituitary* 5, 89–98. doi:10.1023/a:1022312530900

Subramanian, V., Lee, R. S. M., Howell, S., Gregson, S., Lahart, I. M., Kaushal, K., et al. (2021). Non-Functioning Pituitary Macroadenomas: Factors Affecting Postoperative Recurrence, and Pre- and Post-Surgical Endocrine and Visual Function. *Endocrine* 73 (2), 407–415. doi:10.1007/s12020-021-02713-1

The Gene Ontology (2017). Expansion of the Gene Ontology Knowledgebase and Resources. *Nucleic Acids Res.* 45, D331–D338. doi:10.1093/nar/gkw1108

Tjörnstrand, A., Gunnarsson, K., Evert, M., Holmberg, E., Ragnarsson, O., Rosén, T., et al. (2014). The Incidence Rate of Pituitary Adenomas in Western Sweden for the Period 2001-2011. *Eur. J. Endocrinol.* 171, 519–526. doi:10.1530/eje-14-0144

Wang, C., Tan, C., Wen, Y., Zhang, D., Li, G., Chang, L., et al. (2019). FOXP1-Induced lncRNA CLRN1-AS1 Acts as a Tumor Suppressor in Pituitary Prolactinoma by Repressing the Autophagy via Inactivating Wnt/β-Catenin Signaling Pathway. *Cell Death Dis* 10, 499. doi:10.1038/s41419-019-1694-y

Wang, K. C., and Chang, H. Y. (2011). Molecular Mechanisms of Long Noncoding RNAs. *Mol. Cel* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wu, Z. R., Yan, L., Liu, Y. T., Cao, L., Guo, Y. H., Zhang, Y., et al. (2018). Inhibition of mTORC1 by lncRNA H19 via Disrupting 4E-BP1/Raptor Interaction in Pituitary Tumours. *Nat. Commun.* 9, 4624. doi:10.1038/s41467-018-06853-3

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zhang, F., Wang, H., Yu, J., Yao, X., Yang, S., Li, W., et al. (2021). LncRNA CRNDE Attenuates Chemoresistance in Gastric Cancer via SRSF6-Regulated Alternative Splicing of PICALM. *Mol. Cancer* 20, 6. doi:10.1186/s12943-020-01299-y

Zhao, P., Cheng, J., Li, B., Nie, D., Wang, H., Li, C., et al. (2021). LncRNA PCAT6 Regulates the Progression of Pituitary Adenomas by Regulating the miR-139-3p/BRD4 Axis. *Cancer Cel Int* 21, 14. doi:10.1186/s12935-020-01698-7

Zhu, X., Tian, X., Yu, C., Shen, C., Yan, T., Hong, J., et al. (2016). A Long Non-Coding RNA Signature to Improve Prognosis Prediction of Gastric Cancer. *Mol. Cancer* 15, 60. doi:10.1186/s12943-016-0544-0

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 October 2021
doi: 10.3389/fgene.2021.730519

Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries

Jidong Lang[1,2,3†], Rongrong Zhu[4†], Xue Sun[1], Siyu Zhu[5], Tianbao Li[1,2], Xiaoli Shi[1], Yanqi Sun[1], Zhou Yang[1], Weiwei Wang[1,2], Pingping Bing[3], Binsheng He[3*] and Geng Tian[1,2*]*

[1]Bioinformatics and R and D Department, Geneis (Beijing) Co. Ltd., Beijing, China, [2]Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [3]Academician Workstation, Changsha Medical University, Changsha, China, [4]Vascular Surgery Department, Tsinghua University Affiliated Beijing Tsinghua Changgung Hospital, Beijing, China, [5]Department of Medicine, School of Medicine, University of California at San Diego, La Jolla, CA, United States

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology, China

Reviewed by:
Bing Wang,
Anhui University of Technology, China
Attila Patocs,
Semmelweis University, Hungary
Li Peng,
Hunan University of Science and Technology, China

***Correspondence:**
Jidong Lang
langjd@geneis.cn
Binsheng He
hbscsmu@163.com
Geng Tian
tiang@geneis.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: *25 June 2021*
Accepted: *24 September 2021*
Published: *27 October 2021*

Citation:
Lang J, Zhu R, Sun X, Zhu S, Li T, Shi X, Sun Y, Yang Z, Wang W, Bing P, He B and Tian G (2021) Evaluation of the MGISEQ-2000 Sequencing Platform for Illumina Target Capture Sequencing Libraries. Front. Genet. 12:730519. doi: 10.3389/fgene.2021.730519

Illumina is the leading sequencing platform in the next-generation sequencing (NGS) market globally. In recent years, MGI Tech has presented a series of new sequencers, including DNBSEQ-T7, MGISEQ-2000 and MGISEQ-200. As a complex application of NGS, cancer-detecting panels pose increasing demands for the high accuracy and sensitivity of sequencing and data analysis. In this study, we used the same capture DNA libraries constructed based on the Illumina protocol to evaluate the performance of the Illumina Nextseq500 and MGISEQ-2000 sequencing platforms. We found that the two platforms had high consistency in the results of hotspot mutation analysis; more importantly, we found that there was a significant loss of fragments in the 101–133 bp size range on the MGISEQ-2000 sequencing platform for Illumina libraries, but not for the capture DNA libraries prepared based on the MGISEQ protocol. This phenomenon may indicate fragment selection or low fragment ligation efficiency during the DNA circularization step, which is a unique step of the MGISEQ-2000 sequence platform. In conclusion, these different sequencing libraries and corresponding sequencing platforms are compatible with each other, but protocol and platform selection need to be carefully evaluated in combination with research purpose.

Keywords: illumina sequencing platform, MGISEQ-2000 sequencing platform, next generation sequencing, DNA nanoball, target capture library

INTRODUCTION

With the launch of the Human Genome Project, next-generation sequencing (NGS) technology has had a huge impact on the biological field in the past 20 years (Consortium, 2015; Yang et al., 2015; Goodwin et al., 2016). Different companies and research institutions have developed various sequencing approaches and platforms, such as Roche's 454 sequencing platform, Illumina's sequencing by synthesis (SBS) technology, and PacBio's single-molecule nanopore sequencing technology (Rivas et al., 2015; Goodwin et al., 2016). Among them, the sequencers or sequencing platforms developed by the Illumina Company have a dominant position in the sequencing market due to their high throughput and high sequencing accuracy. Over time, the development of machine hardware and the diversification of bioinformatics analysis software tools have led to drastic reductions in sequencing costs and increases in convenience and usability, even for new developed techniques like single cell sequencing (Yang et al., 2020a; Xu et al., 2020). For

example, NGS technology plays a vital role in analyzing somatic mutations that occur in multiple tumor types. The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) and International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) have sequenced thousands of tumors from more than 50 cancer types and summarized the significant genetic somatic mutations that occur during the process of tumorigenesis (Alexandrov et al., 2013). These data have played an extremely important role in promoting cancer genome research and development (He et al., 2020a; He et al., 2020b; Liu et al., 2021).

Recently, MGI Tech Co., Ltd (referred to MGI) launched a series of NGS sequencers and platforms based on DNA nanoball (DNB) and probe-anchor synthesis (cPAS) technology, such as MGISEQ-200, MGISEQ-2000, and DNBSEQ-T7 (Fehlmann et al., 2016). They have gradually achieved a certain sales volume and have become another option for high-throughput sequencing. For example, MGISEQ-2000 can generate approximately 1.44 TB sequencing data per run with a running cost of only 10 USD/GB. Several studies have compared the performance between MGI and the Illumina sequencing platform, and the results showed that they were highly consistent for different types of sequencing libraries, including whole-exome sequencing (WES) (Xu et al., 2019), whole-genome sequencing (WGS) (Patch et al., 2018), transcriptome sequencing (Zhu et al., 2018; Jeon et al., 2019; Patterson et al., 2019; Zeng et al., 2020), single-cell transcriptome sequencing (Natarajan et al., 2019; Peng et al., 2020a; Senabouth et al., 2020; Zhuang et al., 2021), metagenome sequencing (Fang et al., 2018) and small RNA sequencing (Huang et al., 2017) libraries.

When MGI launched their sequencers, they indicated that they were compatible with the sequencing libraries constructed based on Illumina protocols, that is, that the MGISEQ platform could sequence the Illumina libraries. In our study, we used the same capture DNA libraries constructed based on the Illumina protocol for sequencing with the Illumina NextSeq 500 and MGISEQ-2000 sequencing platforms. We found that the two platforms had high consistency in the hotspot mutation analysis and that there was a significant loss of the 101–133 bp fragments on the MGISEQ-2000 sequencing platform but not in the capture DNA libraries based on the MGISEQ protocol. We hypothesized that this might be related to fragment selection or low ligation efficiency during the DNA circularization step, a step that is unique to the MGISEQ-2000 sequence platform. Hence, although the selection of sequencers and platforms is becoming increasingly diversified and all theoretically compatible and applicable to each other, the choice of platform for practical applications may need to be further evaluated according to the research purpose and library characteristics.

MATERIALS AND METHODS

Sample Collection and Experimental Groups

Our research was approved by the Qingdao Geneis Institute of Big Data Mining and Precision Medicine in November 2019, and the research ID was Ethics-QD-[2020] No. 001. A total of 272 samples (patient age: 29–91 years old) were collected at Qingdao Geneis Institute of Big Data Mining and Precision Medicine from December 2019 to March 2020, including 79 plasma samples, 21 white blood cell samples and 172 formalin-fixed and paraffin-embedded (FFPE) samples. Informed written consent forms were obtained from patients, and identifying information was removed. The clinical information of the samples is shown in **Table 1**.

TABLE 1 | Clinical information for collected samples.

Clinical characteristics		All samples (*n* = 272)
Unknown		46
Age, Median (Range)-yrs		62.5 (29.0–91.0)
Age groups-No.%	15–49 years	24/226 (10.62)
	50–64 years	97/226 (42.92)
	≥65 years	105/226 (46.46)
Sex-No.%	Female	103/226 (45.58)
	Male	123/226 (54.42)
Disease-No.%	Lung cancer	166/226 (73.45)
	Colon cancer	13/226 (5.75)
	Rectal cancer	11/226 (4.87)
	Gastric cancer	6/226 (2.65)
	Breast cancer	5/226 (2.21)
	Esophageal cancer	5/226 (2.21)
	Colorectal cancer	4/226 (1.77)
	Nasopharyngeal carcinoma	2/226 (0.88)
	Liver cancer	1/226 (0.44)
	Ovarian cancer	1/226 (0.44)
	Tongue cancer	1/226 (0.44)
	Unknown	11/226 (4.87)

We randomly selected 204 (75%: 204/272) samples to construct capture libraries based on the Illumina protocol and performed data analysis. The remaining samples were divided into two groups of 34 samples (12.5%: 34/272) using different capture panels and constructing capture libraries based on the MGISEQ protocol for sequencing and data analysis, respectively.

Library Preparation Based on Illumina Platform and Sequencing

DNA for NGS-based analysis was extracted using the GeneRead Kit (Qiagen, Hilden, Germany) for FFPE tissue and the QIAamp DNA Blood Mini Kit (Qiagen, Hilden, Germany) for white blood cell samples. DNA (200 ng) was used to build the library by using the NEBNext Ultra II DNA library Prep Kit for Illumina (96 reactions) (NEB, Ipswich, MA, United States). Cell-free DNA was extracted using a QIAamp Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. The extracted DNA (20 ng/sample) was then used to build libraries using Accel-NGS® 2S Plus DNA Library Kits (96 reactions; Swift BioSciences, Ann Arbor, MI, United States). Integrated DNA Technologies (IDT, Skokie, IL, United States) or Agilent Technologies (Santa Clara, CA, United States) custom probes were used for hybridization capture. We used the IDT 38-hotspot gene panel or Agilent 519 gene panel (**Supplementary Table S5**) for all 272 libraries.

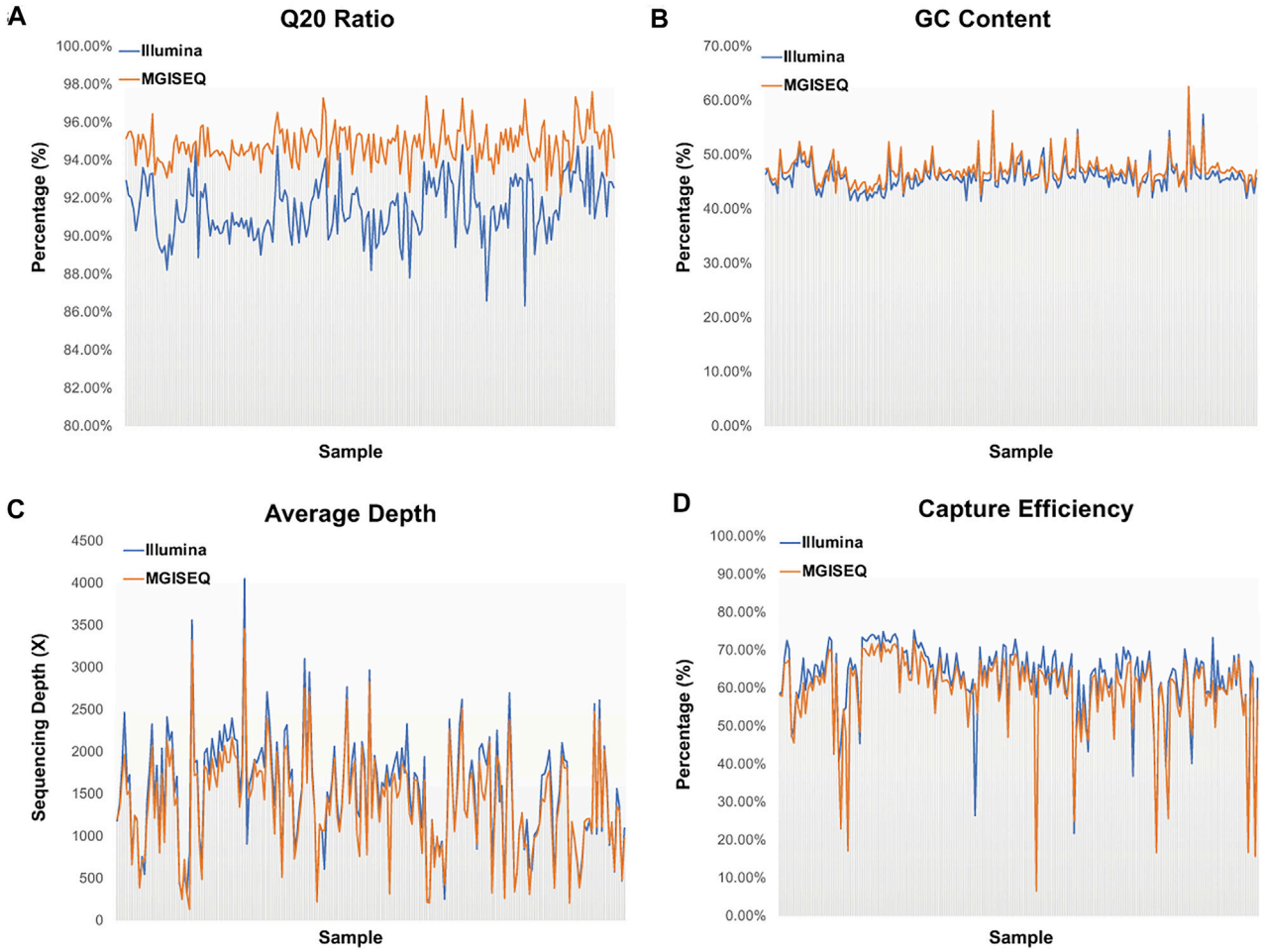


FIGURE 1 | Comparison of Sequencing Data Quality Control Parameters between Illumina and MGISEQ-2000 Platforms. **(A)** Distribution of Q20 ratio by each sample. **(B)** Distribution of GC content by each sample. **(C)** Distribution of average depth by each sample. **(D)** Distribution of probe capture efficiency by each sample.

Quantification was performed with an Illumina/Universal Library Quantification Kit (Kapa Biosystems, Wilmington, MA, United States) on an ABI 7500 Real Time Polymerase Chain Reaction (PCR) System (Applied Biosystems, Waltham, MA, United States). The quality control for Agilent 2,100 Bioanalyzer used a High-Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA, United States). Next-generation sequencing-based analysis was performed on a NextSeq500 or MiSeqDX instrument according to the manufacturer's instructions (Illumina, San Diego, CA, United States). With the NextSeq500/550 High Output V2 Kit or MiSeqTMDX Reagent V3 Kit, Illumina NextSeq500 or MiSeqDX (Illumina, San Diego, CA, United States) was used for DNA sequencing in 302 cycles for 151 bp paired-end sequencing. All 272 libraries were also analyzed on a MGISEQ2000 instrument according to the manufacturer's instructions (BGI, Shenzhen, Guangdong, China). With the MGISEQ-2000RS High Output kit (BGI, Shenzhen, Guangdong, China), MGISEQ-2000 (BGI, Shenzhen, Guangdong, China) was used for DNA sequencing in 200 cycles and 300 cycles for 100 bp and 150 bp paired-end sequencing, respectively.

Library Preparation Based on the MGISEQ Platform and Sequencing

DNA libraries were prepared with the MGIEasy FS DNA Library Prep Set (BGI, Shenzhen, Guangdong, China). DNA (50–200 ng) was fragmented physically with a Covaris S220 instrument (Covaris, Woburn, MA, United States), followed by A-tailing, adapter ligation and PCR amplification. DNA library quality was assessed using a Qubit and Agilent 2,100 Bioanalyzer with a High Sensitivity DNA Kit. Cot-1 DNA blocking reagent (Thermo Fisher Scientific, Waltham, MA, United States), IDT universal blocking oligonucleotides and IDT adapter-specific blocking oligonucleotides were added to the pooled libraries and dried in a SpeedVac. The dried mixture was redissolved in mixed liquids of IDT hybridization buffer, IDT hybridization enhancer and BOKE capture probes (BOKE bioscience, Bejing, China). After hybridization at 65°C for 4 h, the target regions were captured with M270 streptavidin

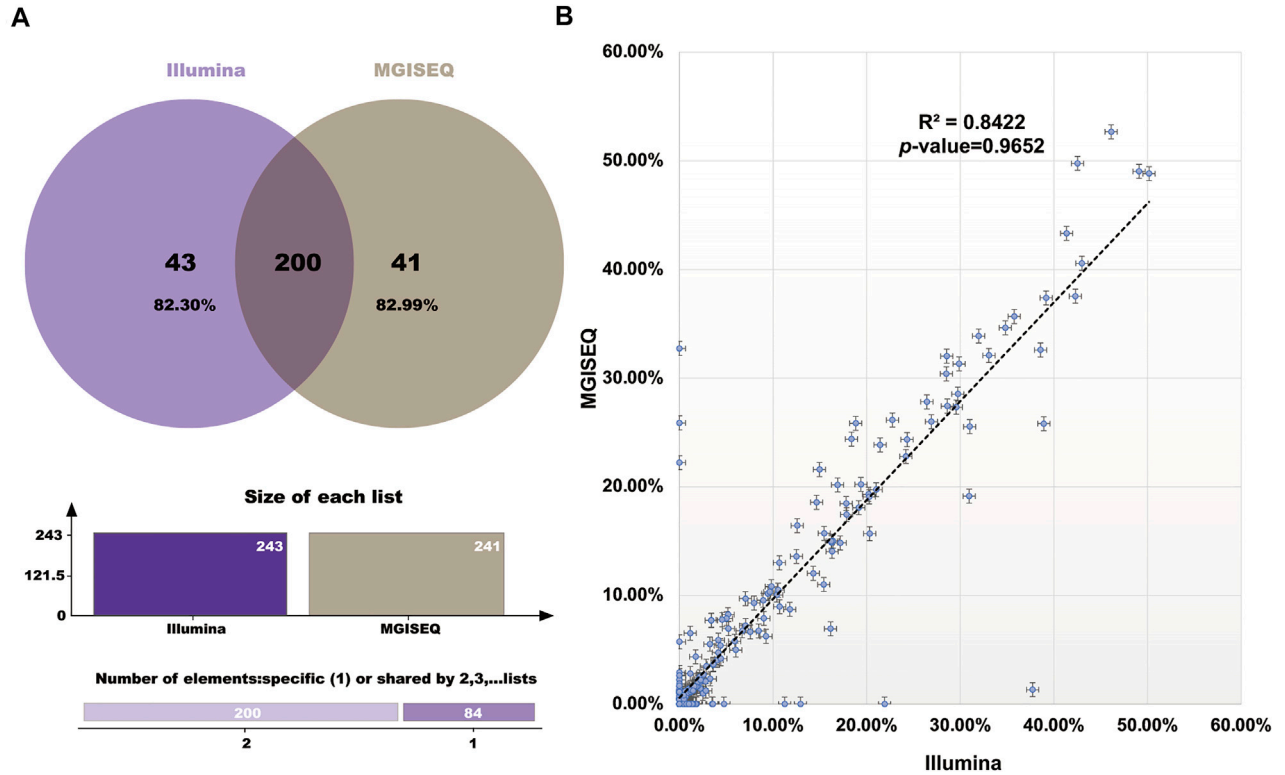


FIGURE 2 | Hotspot Mutation Results Comparison between the Illumina and MGISEQ-2000 Platforms. **(A)** The Venn digram of the detected hotspot mutations comparision. **(B)** The correlation comparison of the detected hotspot mutation frequency values.

beads by incubation at 65°C for 45 min and then washed 3 times at 65°C and another 3 times at room temperature with IDT xGen lockdown reagents. Then, 15 postcapture amplification cycles were performed to obtain the captured libraries. Final libraries were pooled and sequenced using the MGISEQ-2000 sequencing platform with a 150 bp paired-end cycle kit.

Data Normalization and Statistics

As the volume of sequencing data and read length of the Illumina and MGISEQ-2000 platforms were different (**Supplementary Table S1**), we "normalized" all 272 sample sequencing datasets, that is, each sample had the same read length and read number. We used seqtk (version: 1.0-r73-dirty) (https://github.com/lh3/seqtk) to "normalize" the raw sequencing data. We used a in-house perl program to caculate the number of reads, Q20 ratio and GC content (**Supplementary Table S2**).

Data Preprocessing and Analysis

The normalized data were cleaned by Trimmomatic (version: 0.39) (Bolger et al., 2014), which filtered out the adapter contamination reads and low-quality reads and the parameter's setting was ILLUMINACLIP:adapter sequence:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 (adapter sequences for Illumina Nextseq 500 and MGISEQ-2000 were AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC/AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA and AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA/AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG, respectively). BWA-ALN algorithm (version: 0.7.12) (Li and Durbin, 2009) was applied for alignment with the reference genome hg19 (parameters: -o 1 -e 50 -t 4 -i 15 -q 10). The output SAM file was sorted and deduplicated with Samtools (version: 0.1.19) (Li et al., 2009), and the BAM format file was obtained. We used FreeBayes (version: 1.0.2) (Garrison and Marth, 2012) to detect SNP/InDel mutations (parameters: -j -m 10 -q 20 -F 0.001 -C 1). The mutations were annotated from the ANNOVAR database (Wang et al., 2010). Fragment size distribution was summarized from the paired-end alignment information (column ninth) in the BAM format file. Statistical analysis used the statistical functions in Microsoft Excel 2019 and R software (version 3.2.5).

RESULTS

Data Quality Control Parameters Were Significantly Different Between the Illumina and MGISEQ-2000 Sequencing Platforms

We compared the Q20 rate, GC content, mean depth and capture efficiency of 204 samples generated based on the Illumina library protocol, which were captured by the IDT 38-hotspot gene panel and sequenced on the Illumina and MGISEQ-2000 sequencing

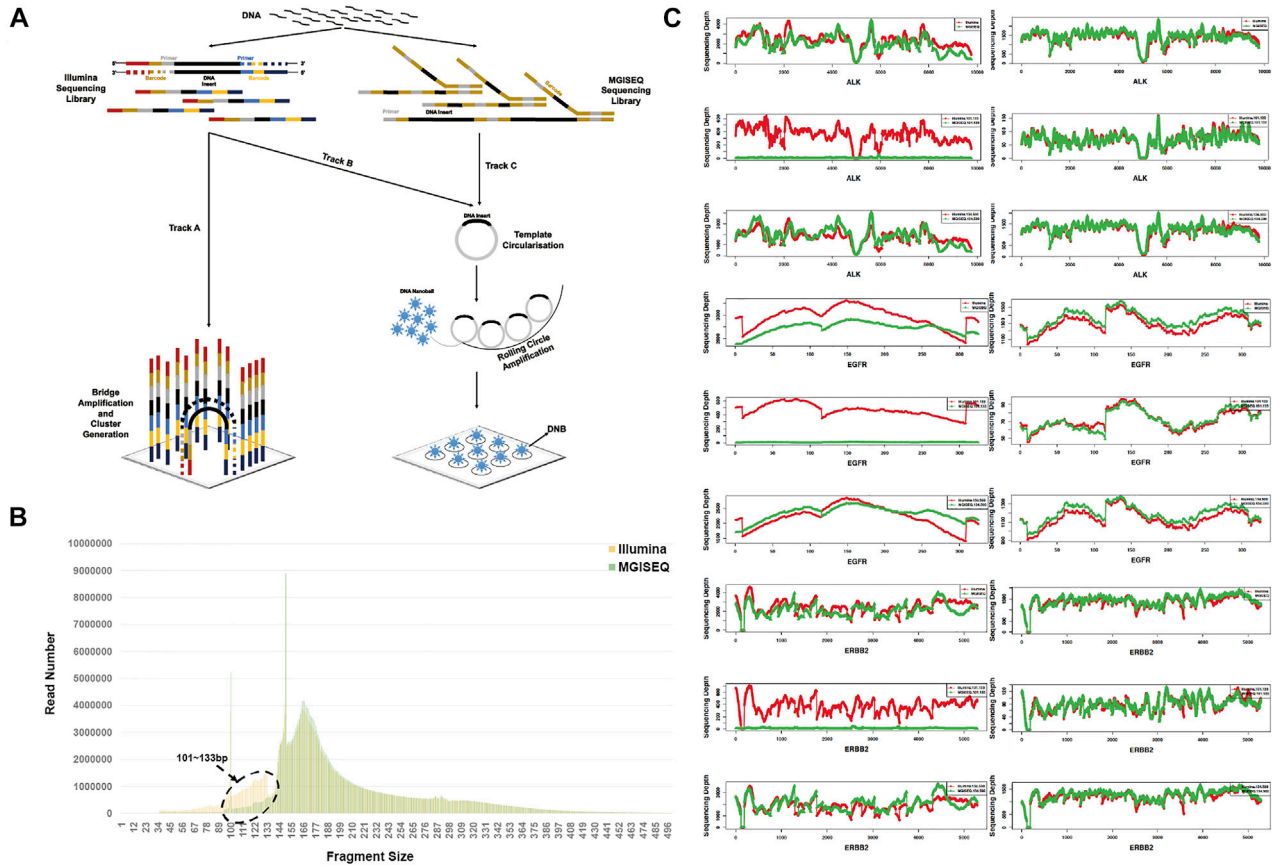


FIGURE 3 | The Insert Fragment Size Analysis of Illumina-based Capture Library On the Illumina and MGISEQ-2000 Platforms. **(A)** The principles of Illumina and MGISEQ sequencers. Track A represents the Illumina library combined with bridge PCR amplification of the Illumina platform. Track B represents the Illumina library combined with DNA circularization for DNB of the MGISEQ-2000 platform. Track C represents the MGISEQ library combined with DNA circularization for DNB of the MGISEQ-2000 platform. **(B)** Compared with the Illumina platform data, the MGISEQ-2000 platform data had a significant loss of 101–133 bp fragments. **(C)** Statistical analysis of sequencing depth distribution in ALK, EGFR and ERBB2 with one sample. Sequencing depth distribution of 101–133 bp (left panel) and 134–500 bp (right panel) in ALK (top), EGFR (middle), and ERRBB2 (bottom). In each figure panel, the top panel shows the sample total sequencing depth distribution, the middle panel shows the sequencing depth of 101–133 bp fragment size, and the bottom panel shows the 134–500 bp fragment size sequencing depth distribution.

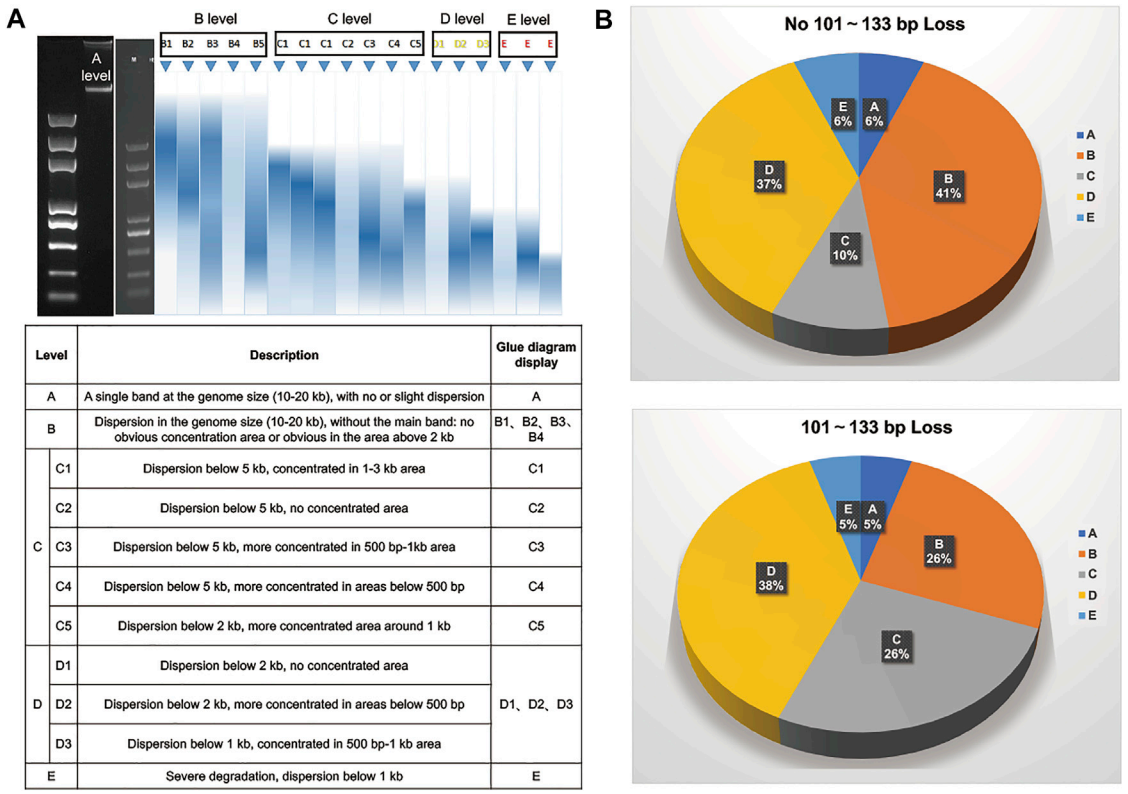
platforms (**Figure 1**, details in **Supplementary Table S3**), respectively. We found that all of the quality control parameters had significant differences, with *p*-values of 4.87e-85, 1.15e-4, 0.0326 and 0.0035, respectively, in the two-tailed heteroscedasticity *t*-test analysis. We thought that these differences could be due to the sequencing principles, the algorithm used for base recognition or the sequencing platform characteristics. For example, the Nextseq500 platform treated all unrecognized bases as G, while HiSeq-2000, MGISEQ-2000 and other previous four-color imaging sequencers treated these bases as N. Therefore, the GC content tended to be higher in the Illumina NextSeq500 results than in the others.

Hotspot Mutations Showed High Consistency Between the Illumina and MGISEQ-2000 Sequencing Platforms.

The hotspot mutations (SNPs and InDels) detected in 204 sample datasets were compared between the Illumina and MGISEQ-2000 platforms (**Supplementary Table S4**). We defined a positive detection filter condition as mutation frequency ≥ 0.4% for plasma samples and mutation frequency ≥ 1% for FFPE samples. We found that the hotspot mutation detection results had high consistency rates of 82.30% (Illumina: 200/243) and 82.99% (MGISEQ-2000: 200/241) (**Figure 2A**). Furthermore, no significant difference (R^2 = 0.8422, *p*-value = 0.9652) in mutation frequency was observed between the Illumina and MGISEQ-2000 platform data. (**Figure 2B**).

MGISEQ-2000 sequencing platform data based on Illumina libraries showed a significant loss of the 101–133 bp fragment.

Insert fragment size and distribution were evaluated and analyzed for all 204 samples. As we used the same sample library for sequencing, the theoretical difference only existed in Illumina's bridge PCR amplification and MGISEQ-2000s DNB circularization. (**Figure 3A**) (Goodwin et al., 2016; Chen et al., 2019; Korostin et al., 2020). Combining all 204 sample data for fragment size analysis, our results revealed a significant loss of 101–133 bp fragments in the MGISEQ-2000 platform data, with a



Level		Description	Glue diagram display
A		A single band at the genome size (10-20 kb), with no or slight dispersion	A
B		Dispersion in the genome size (10-20 kb), without the main band: no obvious concentration area or obvious in the area above 2 kb	B1、B2、B3、B4
C	C1	Dispersion below 5 kb, concentrated in 1-3 kb area	C1
	C2	Dispersion below 5 kb, no concentrated area	C2
	C3	Dispersion below 5 kb, more concentrated in 500 bp-1kb area	C3
	C4	Dispersion below 5 kb, more concentrated in areas below 500 bp	C4
	C5	Dispersion below 2 kb, more concentrated area around 1 kb	C5
D	D1	Dispersion below 2 kb, no concentrated area	D1、D2、D3
	D2	Dispersion below 2 kb, more concentrated in areas below 500 bp	
	D3	Dispersion below 1 kb, concentrated in 500 bp-1 kb area	
E		Severe degradation, dispersion below 1 kb	E

FIGURE 4 | Statistical Analysis On The Quality of 204 Samples. **(A)** Sample quality grading table of gDNA agarose gel electrophoresis. **(B)** The distribution of different sample quality levels in samples with and without loss of 101–133 bp fragment size. The top figure represented sample quality grade distribution of samples without 101–133 bp fragment size loss. The bottom figure represented sample quality grade distribution of samples with 101–133 bp fragment size loss.

t-test *p*-value of 3.3072e-17 (**Figure 3B**), while other fragment sizes, such as 134–500 bp (*t*-test *p*-value = 0.7264), did not show a difference. Although significant differences were found in the Q20 rate, GC content and other quality control statistics, these should be attributable to the sequencer system characteristics and should not have a great impact on the fragment size distribution. Therefore, the loss of the 101–133 bp fragment size may be related to the DNA cyclization step, that is, there may be fragment size selection in the circularization step or enrichment bias for longer DNA molecules and low ligation efficiency for shorter DNA molecules.

Then, we extracted 101–133 bp and 134–500 bp fragment size information from BAM files for each sample and analyzed the sequencing depth distribution of three common cancer genes, ALK receptor tyrosine kinase (*ALK*), epidermal growth factor receptor (*EGFR*) and erb-b2 receptor tyrosine kinase 2 (*ERBB2*). The results showed that 69.12% (141/204) of samples had 101–133 bp fragment size loss, while the sequencing depth distribution of 134–500 bp fragments was consistent with the overall total sequencing depth, indicating that the phenomenon was not due to stochasticity in specific genes (**Figure 3C**). The sequencing depth distribution of all samples was in the Supplementary Figures by each sample.

As we know, the use of FFPE or hemolyzed samples may have a great influence on the distribution of DNA fragment size. Therefore, we performed statistical analysis on the quality of 204 samples with and without 101–133 bp loss. First, we defined the sample quality levels with DNA agarose gel electrophoresis as A, B, C, D or E (**Figure 4A**). Then, all samples in each grade were subgrouped according to whether the 101–133 bp fragment size was lost. We found that the sample proportions of A, D and E levels were consistent in the two groups, while B and C levels were quite different. The proportions of B [C] level samples in the 101–133 bp loss group and 101–133 bp nonloss group were 25.53% (36/141) [26.24% (37/141)] and 41.27% (26/63: 6) [9.52% (6/63)], respectively (**Figure 4B**). Therefore, our results showed that the circularization step of MGISEQ-2000 not only biased the selection of DNA fragment size but also may have a greater impact on samples with quality grade B or C.

Fragment Size Loss had no Probe Preference and was not Obvious in the Database of MGISEQ-2000 Libraries.

To verify whether the phenomenon was related to capture-probe preference, we analyzed the fragment size distribution of the sequencing data from 34 samples that were captured with an Agilent 519 gene panel and sequenced separately by Illumina Nextseq500 and MGISEQ-2000. As shown in **Figure 5A**, the same 101–133 bp fragment size loss was found. In addition, we

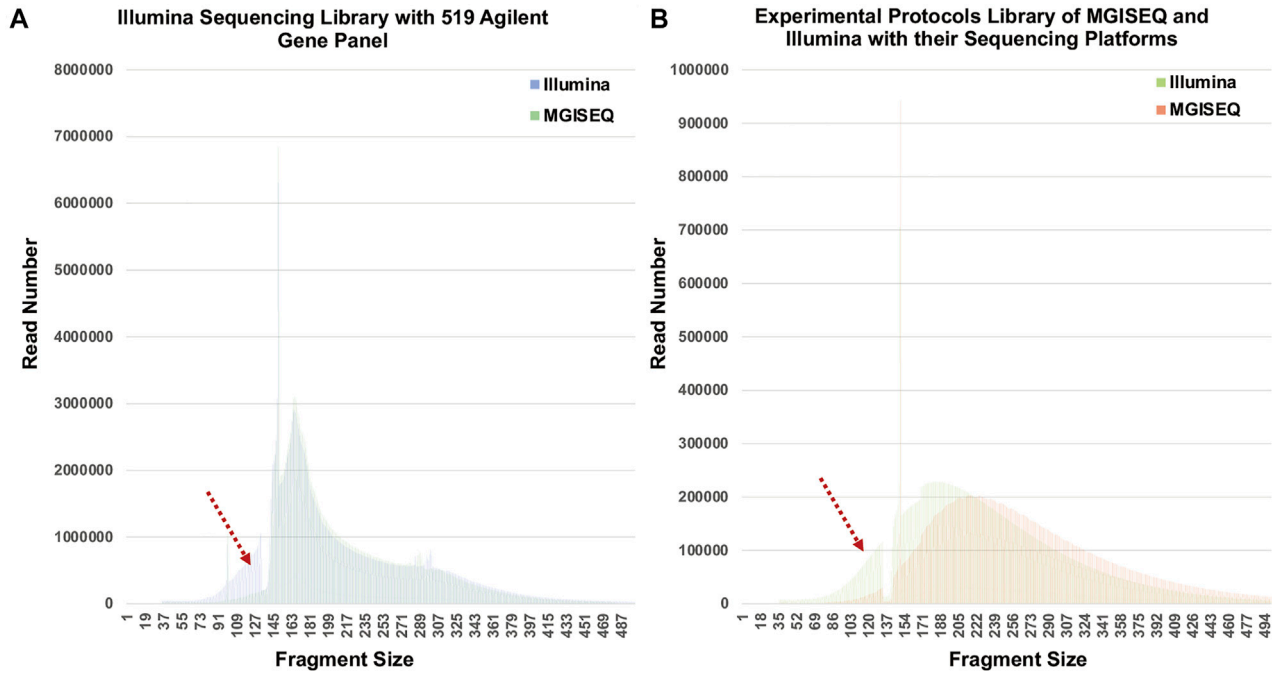


FIGURE 5 | Fragment Distribution in Illumina Nextseq500 and MGISEQ-2000. **(A)** The libraries were constructed following the instructions of Illumina and captured with an Agilent 519 gene panel. (**B)** Fragment distribution when experiments were performed according to the experimental systems and kits recommended by Illumina and MGISEQ, respectively.

constructed 34 other libraries according to the experimental protocols of MGISEQ and Illumina and generated data on their sequencing platforms. We also analyzed the fragment size distribution and found that the fragment size (peak 183 bp) distribution on the Illumina platform had a "left offset" compared to that (peak 214 bp) on the MGISEQ-2000 platform. The fragment size distribution curve of the MGISEQ data was smooth, and there was no obvious 101–133 bp fragment size loss (**Figure 5B**).

DISCUSSION

In recent decades, next-generation sequencing technology has undergone rapid development. With the greatly reduced sequencing cost, increasing scientific research and technical product development are being applied to NGS. In particular, to meet the needs of precision medicine and big data mining, the number and scale of cancer omics research and clinical projects are constantly increasing (Yang et al., 2020b; Zeng et al., 2020). For a large number of samples, the expenses and costs borne are unaffordable; thus, sequencing costs are still the bottleneck for large-scale NGS applications. At present, Illumina sequencers dominate the high-throughput sequencing market, but MGI sequencers based on DNB technology have gradually become more popular worldwide. Recently, several studies have compared the performance of BGI-500 and the Illumina HiSeq machine and showed that both of them could produce high-quality data in various applications. However, a comparison of their quality for capture panel sequencing (except WES), which is widely used in tumor research, has not been published.

In this study, we compared the data produced from the same library by different sequencing platforms. For the library preparation step, Illumina used bridge PCR technology, while MGI achieved single-molecule template amplification by DNB circularization amplification. We applied both the Illumina (Nextseq500 and MiSeqDx) platform and MGISEQ (MGISEQ-2000) platform to the same library constructed by the Illumina protocol. Theoretically, any difference in sequencing data should have been caused by the differences between bridge PCR and circularization amplification or the consequent sequencing system differences. Comparison of the data analysis results revealed the disadvantage of fragment size selection and short fragment size ligation efficiency in the circularization step. These results suggest that the sequencing data based on Illumina library preparations and in which sample types with shorter fragment sizes (such as hemolyzed plasma samples) or a more complex distribution of DNA fragment sizes (such as FFPE samples with longer storage times) are used may encounter short DNA fragment size loss on the MGISEQ sequencing platform. Therefore, we should evaluate the compatibility of sequencing libraries and sequencing platforms for scientific research that focuses on the distribution of fragment size, especially for small RNA (Fehlmann et al., 2016), cell-free DNA (cfDNA) and circulating tumor DNA (ctDNA) research (Underhill et al., 2016; Liu et al., 2020). Although the sequencing library is basically compatible with different sequencing platforms, appropriate experimental systems and sequencing platforms

should be selected based on the research purpose and sample type. Otherwise, there may be an unexpected impact on the sequencing results. Our data showed the results of only target capture panel sequencing; the assessment of other sequencing applications requires further investigation.

Considering that the alignment algorithm may also have an impact on the fragment size distribution analysis, we replaced the BWA "aln" algorithm mentioned in the article with the BWA "mem" algorithm. The "mem" algorithm is much looser than the "aln" algorithm, and it can perform local alignment and splicing. The "mem" algorithm allows multiple different parts of the sequencing reads to have their own optimal matches, resulting in multiple optimal alignment positions for the reads and greatly improving the alignment rate. After comparing and analyzing the combined data with 204 samples of the IDT 38-hotspot gene panel and 34 samples of the Agilent 519 gene panel by using the "mem" algorithm, we found that the number of reads in the 101–133 bp fragment size from the MGISEQ-2000 platform data was significantly improved (**Supplementary Figure S1**), but there were still significant differences, with *t*-test *p*-values of 0.0277 and 0.0252, respectively. The conclusion was consistent with that based on the "aln" algorithm.

We also found that the data without the 101–133 bp fragment size loss were derived from different sequencing read lengths of the Illumina Nextseq500 and MGISEQ-2000 platforms, while the data with the same sequencing read length showed the 101–133 bp fragment size loss. To investigate whether the data with or without the phenomenon were related to the sequencing read length, we reanalyzed and compared data with the same number of sequencing reads but not read length, and found that the results were consistent with the previous conclusion. Since the 101–133 bp fragment size loss was concentrated in the data with long read length (150 bp) but not in the data with short read length (100 bp), we hypothesized that the phenomenon may also be related to the sequencing read length. We will conduct more in-depth research on this point in our future work.

In summary, the MGISEQ-2000 platform has good compatibility with Illumina sequencing libraries, but the DNB circularization step may cause fragment size selection or have low ligation efficiency for short DNA fragment sizes. For the accuracy of downstream data analysis, we recommend that different sequencing platforms should be used with their official experimental systems and kits. If the experiment needs to change between different platforms, for cost considerations or other reasons, the selected platform should be evaluated carefully with respect to the purpose of the research or actual needs, as it may have a significant impact on outcomes. In the future, it would be interesting to compare the performances of two platforms in specific applications like cancer diagnosis (He et al., 2020b; Peng L.-H. et al., 2020), prognosis (Peng et al., 2020c; Song et al., 2020; Zhou et al., 2020), evolution inference (Yang et al., 2013; Yang et al., 2014), drug repositioning (Peng et al., 2015; Zhou et al., 2019; Liu et al., 2020), and so on. However, it is out of the scope of this study.

DATA AVAILABILITY STATEMENT

The data has been uploaded to NCBI - BioProject 744584.

AUTHOR CONTRIBUTIONS

GT, JL and BH designed the study, collected, analyzed and interpreted the data, and wrote the article. XuS and ZY performed the experiment. RZ, SZ, TL, XiS, YS, WW and PB reviewed and modified the article. All authors approved the final version of the article.

ACKNOWLEDGMENTS

We thank Tingting Hui in Geneis (Beijing) Co. Ltd. for modifying and adjusting the figures.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.730519/full#supplementary-material

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013). Deciphering signatures of mutational processes operative in human cancer. *Cel Rep.* 3, 246–259. doi:10.1016/j.celrep.2012.12.008

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Chen, J., Li, X., Zhong, H., Meng, Y., and Du, H. (2019). Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. *Sci. Rep.* 9, 9345. doi:10.1038/s41598-019-45835-3

Consortium, G. T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi:10.1126/science.1262110

Fang, C., Zhong, H., Lin, Y., Chen, B., Han, M., Ren, H., et al. (2018). Assessment of the cPAS-based BGISEQ-500 platform for metagenomic sequencing. *Gigascience* 7, 1–8. doi:10.1093/gigascience/gix133

Fehlmann, T., Reinheimer, S., Geng, C., Su, X., Drmanac, S., Alexeev, A., et al. (2016). cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin. Epigenet* 8, 123. doi:10.1186/s13148-016-0287-1

Garrison, E., and Marth, G. (2012). Haplotype-Based Variant Detection from Short-Read Sequencing. *Quantitative Biol.* arXiv:1207.3907v2.

Goodwin, S., Mcpherson, J. D., and Mccombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi:10.1038/nrg.2016.49

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020a). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1866, 165916. doi:10.1016/j.bbadis.2020.165916

He, B., Lang, J., Wang, B., Liu, X., Lu, Q., He, J., et al. (2020b). TOOme: A Novel Computational Framework to Infer Cancer Tissue-of-Origin by Integrating Both Gene Mutation and Expression. *Front. Bioeng. Biotechnol.* 8, 394. doi:10.3389/fbioe.2020.00394

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., et al. (2017). A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* 6, 1–9. doi:10.1093/gigascience/gix024

Hudson, T. J., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi:10.1038/nature08987

Jeon, S. A., Park, J. L., Kim, J.-H., Kim, J. H., Kim, Y. S., Kim, J. C., et al. (2019). Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. *Genomics Inform.* 17, e32. doi:10.5808/gi.2019.17.3.e32

Korostin, D., Kulemin, N., Naumov, V., Belova, V., Kwon, D., and Gorbachev, A. (2020). Comparative analysis of novel MGISEQ-2000 sequencing platform vs Illumina HiSeq 2500 for whole-genome sequencing. *PLoS One* 15, e0230301. doi:10.1371/journal.pone.0230301

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

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, H., Qiu, C., Wang, B., Bing, P., Tian, G., Zhang, X., et al. (2021). Evaluating DNA Methylation, Gene Expression, Somatic Mutation, and Their Combinations in Inferring Tumor Tissue-of-Origin. *Front. Cel Dev. Biol.* 9, 619330. doi:10.3389/fcell.2021.619330

Liu, X., Lang, J., Li, S., Wang, Y., Peng, L., Wang, W., et al. (2020). Fragment Enrichment of Circulating Tumor DNA With Low-Frequency Mutations. *Front. Genet.* 11, 147. doi:10.3389/fgene.2020.00147

Natarajan, K. N., Miao, Z., Jiang, M., Huang, X., Zhou, H., Xie, J., et al. (2019). Comparative analysis of sequencing technologies for single-cell transcriptomics. *Genome Biol.* 20, 70. doi:10.1186/s13059-019-1676-5

Patch, A.-M., Nones, K., Kazakoff, S. H., Newell, F., Wood, S., Leonard, C., et al. (2018). Germline and somatic variant identification using BGISEQ-500 and HiSeq X Ten whole genome sequencing. *PLoS One* 13, e0190264. doi:10.1371/journal.pone.0190264

Patterson, J., Carpenter, E. J., Zhu, Z., An, D., Liang, X., Geng, C., et al. (2019). Impact of sequencing depth and technology on de novo RNA-Seq assembly. *BMC Genomics* 20, 604. doi:10.1186/s12864-019-5965-x

Peng, L., Liao, B., Zhu, W., Li, Z., and Li, K. (2017). Predicting Drug-Target Interactions With Multi-Information Fusion. *IEEE J. Biomed. Health Inform.* 21 (2), 561–572. doi:10.1109/JBHI.2015.2513200

Peng, L.-H., Zhou, L.-Q., Chen, X., and Piao, X. (2020b). A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front. Bioeng. Biotechnol.* 8, 40. doi:10.3389/fbioe.2020.00040

Peng, L., Tian, X., and Shen, L. (2020c). Identifying effective antiviral drugs against SARS-CoV-2 by drug repositioning through virus-drug association prediction. *Front. Genet.* 11, 1072. doi:10.3389/fgene.2020.577387

Peng, L., Tian, X., Tian, G., Xu, J., Huang, X., Weng, Y., et al. (2020a). Single-cell RNA-seq clustering: datasets, models, and algorithms. *RNA Biol.* 17 (6), 765–783. doi:10.1080/15476286.2020.1728961

Rivas, M. A., Pirinen, M., Conrad, D. F., Lek, M., Tsang, E. K., Karczewski, K. J., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348, 666–669. doi:10.1126/science.1261877

Senabouth, A., Andersen, S., Shi, Q., Shi, L., Jiang, F., Zhang, W., et al. (2020). Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genom Bioinform* 2, lqaa034, 2020 . lqaa034. doi:10.1093/nargab/lqaa034

Song, Z., Chen, X., Shi, Y., Huang, R., Wang, W., Zhu, K., et al. (2020). Evaluating the Potential of T Cell Receptor Repertoires in Predicting the Prognosis of Resectable Non-Small Cell Lung Cancers. *Mol. Ther. - Methods Clin. Dev.* 18, 73–83. doi:10.1016/j.omtm.2020.05.020

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., et al. (2016). Fragment Length of Circulating Tumor DNA. *Plos Genet.* 12, e1006162. doi:10.1371/journal.pgen.1006162

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Weinstein, J. N., Collisson, E. A., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Xu, J., Cai, L., Liao, B., Zhu, W., and Yang, J. (2020). CMF-Impute: an accurate imputation tool for single-cell RNA-seq data. *Bioinformatics* 36, 3139–3147. doi:10.1093/bioinformatics/btaa109

Xu, Y., Lin, Z., Tang, C., Tang, Y., Cai, Y., Zhong, H., et al. (2019). A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinformatics* 20, 153. doi:10.1186/s12859-019-2751-3

Yang, J., Grünewald, S., and Wan, X.-F. (2013). Quartet-net: a quartet-based method to reconstruct phylogenetic networks. *Mol. Biol. Evol.* 30, 1206–1217. doi:10.1093/molbev/mst040

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

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

Yang, J., Liao, B., Zhang, T., and Xu, Y. (2020a). Editorial: Bioinformatics Analysis of Single Cell Sequencing Data and Applications in Precision Medicine. *Front. Genet.* 10, 1358. doi:10.3389/fgene.2019.01358

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020b). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42, 353–372. doi:10.1007/s11357-019-00106-x

Zeng, L., Yang, J., Peng, S., Zhu, J., Zhang, B., Suh, Y., et al. (2020). Transcriptome analysis reveals the difference between "healthy" and "common" aging and their connection with age-related diseases. *Aging Cell* 19, e13121. doi:10.1111/acel.13121

Zhou, L., Li, Z., Yang, J., Tian, G., Liu, F., Wen, H., et al. (2019). Revealing drug-target interactions with computational models and algorithms. *Molecules* 24 (9), 1714. doi:10.3390/molecules24091714

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method. *Genomics* 112 (6), 4427–4434. doi:10.1016/j.ygeno.2020.07.044

Zhu, F.-Y., Chen, M.-X., Ye, N.-H., Qiao, W.-M., Gao, B., Law, W.-K., et al. (2018). Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods* 14, 69. doi:10.1186/s13007-018-0337-0

Zhuang, J., Cui, L., Qu, T., Ren, C., and Yang, J. (2021). A streamlined scRNA-Seq data analysis framework based on improved sparse subspace clustering. *IEEE Access*, 1. doi:10.1109/access.2021.3049807

Conflict of Interest: JL, XuS, TL, XiS, YS, ZY, WW, and GT were employed by Geneis (Beijing) Co. Ltd.

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

The handling Editor declared a past co-authorship/collaboration with several of the authors.

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

frontiers in Genetics

SYSTEMATIC REVIEW
published: 01 November 2021
doi: 10.3389/fgene.2021.770134

Association Between *RSK2* and Clinical Indexes of Primary Breast Cancer: A Meta-Analysis Based on mRNA Microarray Data

Kun Zheng[1†], Shuo Yao[1†], Wei Yao[1], Qianxia Li[1], Yali Wang[1], Lili Zhang[1], Xiuqiong Chen[1], Huihua Xiong[1], Xianglin Yuan[1], Yihua Wang[2,3], Yanmei Zou[1] and Hua Xiong[1]**

[1]Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, United Kingdom, [3]Institute for Life Sciences, University of Southampton, Southampton, United Kingdom

OPEN ACCESS

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

Reviewed by:
Feng Zhu,
Affiliated Hospital of Guilin Medical University, China
Zhen Guo,
Changsha Medical University, China

***Correspondence:**
Yanmei Zou
whtjzym@tjh.tjmu.edu.cn
Hua Xiong
cnhxiong@tjh.tjmu.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: 03 September 2021
Accepted: 30 September 2021
Published: 01 November 2021

Citation:
Zheng K, Yao S, Yao W, Li Q, Wang Y, Zhang L, Chen X, Xiong H, Yuan X, Wang Y, Zou Y and Xiong H (2021) Association Between RSK2 and Clinical Indexes of Primary Breast Cancer: A Meta-Analysis Based on mRNA Microarray Data. Front. Genet. 12:770134. doi: 10.3389/fgene.2021.770134

Background: Although ribosomal protein S6 kinases, 90 kDa, polypeptide 3 (*RSK2*, *RPS6KA3*) has been reported to play an important role in cancer cell proliferation, invasion, and migration, including breast cancer, its clinical implication in primary breast cancer patients is not well understood, and there were not many studies to explore the relationship between *RSK2* and breast cancer on a clinical level.

Methods: A systematic series matrix file search uploaded from January 1, 2008 to November 31, 2017 was undertaken using ArrayExpress and Gene Expression Omnibus (GEO) databases. Search filters were breast cancer, RNA assay, and array assay. Files eligible for inclusion met the following criteria: a) sample capacity is over 100, b) tumor sample comes from unselected patient's primary breast tumor tissue, and c) expression of *RSK2* and any clinical parameters of patients were available from the files. We use median as the cutoff value to assess the association between the expression of *RSK2* and the clinical indexes of breast cancer patients.

Finding: The meta-analysis identified 13 series matrix files from GEO database involving 3,122 samples that come from patients' primary breast cancer tissue or normal tissue. The expression of *RSK2* in tumor tissues is lower than that in normal tissues [odds ratio (OR), 0.54; 95% credible interval (CI), 0.44–0.67; Cochran's *Q* test $p = 0.14$; $I^2 = 41.7\%$]. Patients with a high expression of *RSK2* showed more favorable overall survival [hazard ratio (HR), 0.71; 95% CI, 0.49–0.94; Cochran's *Q* test $p = 0.95$; $I^2 = 0.0\%$] and less potential of distant metastasis (OR, 0.59; 95% CI, 0.41–0.87; Cochran's *Q* test $p = 0.88$; $I^2 = 0.0\%$) and lymph node infiltration (OR, 0.81; 95% CI, 0.65–0.998; Cochran's *Q* test $p = 0.09$; $I^2 = 42.8\%$). Besides, the expression of *RSK2* in luminal breast cancer is lower than Cochran's *Q* test $p = 0.06$; $I^2 = 63.5\%$). *RSK2* overexpression corresponded with higher

Abbreviations: CI, credible interval; DFS, disease-free survival time; DMFS, distant metastasis-free survival; ER, estrogen receptor; FGFR2, fibroblast growth factor receptor 2; GEO, gene expression omnibus; HER2, human epidermal growth factor receptor 2; HR, hazard ratio; MAPK, mitogen-activated protein kinase; NOS, newcastle-ottawa quality assessment scale; OR, odds ratio; OS, overall survival; PR, progesterone receptor; RFS, relapse-free survival time; RSK2/RPS6KA3, ribosomal protein S6 kinases, 90 kDa, polypeptide 3; TNBC, triple-negative breast cancer; YB-1, Y-box binding protein-1.

histological grade (OR, 1.329; 95% CI, 1.03–1.721; Cochran's *Q* test $p = 0.69$; $I^2 = 0.0\%$). *RSK2* expression is also associated with estrogen receptor (ER) and age.

Conclusion: The meta-analysis provides evidence that *RSK2* is a potential biomarker in breast cancer patients. The expression of *RSK2* is distinctive in different intrinsic subtypes of breast cancer, indicating that it may play an important role in specific breast cancer. Further study is needed to uncover the mechanism of *RSK2* in breast cancer.

Systematic Review Registration: (website), identifier (registration number).

Keywords: ribosomal protein S6 kinase, 90kDa, polypeptide 3 (RSK2), breast cancer, molecular subtype, microarray, prognostic value, biomarkers

INTRODUCTION

Breast cancer is the most frequent cancer among women. Data from the World Health Organization shows that breast cancer impacts over 1.5 million women each year and also causes the greatest number of cancer-related deaths among women. In 2015, 570,000 women died from breast cancer—that is, approximately 15% of all cancer deaths among women. Although there has been a breakthrough in the treatment and prevention of breast cancer in the past few years, leading to the 5-years relative survival rate rising to 90%, the majority of breast cancer patients with distant metastasis succumb to cancer progression within 5 years (Siegel et al., 2018). Also, breast cancer is more than one single disease. Several molecular subtypes of breast cancer have been classified depending on their molecular characteristics (Perou et al., 2000), and each individual subtype corresponds to a different underlying biology, survival rate, and response to therapy (Prat et al., 2015; Nielsen et al., 2017). Therefore, the identification of biomarkers to screen high-risk patients, predict breast prognostic outcomes, and provide new therapeutic targets for specific breast cancer is urgently needed.

RSK2, ribosomal protein S6 kinase, 90 kDa, polypeptide 3, belongs to *RSK* serine/threonine kinase family and is a downstream of the mitogen-activated protein kinase (MAPK) pathway (Zhao et al., 2016). *RSK* is unique among serine–threonine kinases in that it contains two functional kinase domains: an N-terminal kinase that phosphorylates the substrates of *RSK* and a C-terminal kinase involved in the activation mechanism of *RSK* (Frödin and Gammeltoft, 1999). *RSK* isoforms are activated by virtually all extracellular signaling molecules including growth factors, peptide hormones, neurotransmitters, and environmental stresses (Arul and Cho, 2013). It has been demonstrated that *RSK2* plays an important role in cancer cell proliferation, invasion, and migration, including breast cancer (Yoo et al., 2019; Guo and Kong, 2021).

In previous studies, it has been reported that *RSK2* expression is different between breast cancer tissue and normal breast tissue and varies among different subtypes or histological grades of breast cancer. Some studies suggested that *RSK2* overexpression is correlated with basal-like breast cancer and higher histological grade, and *RSK2* mRNA is associated with poor survival in breast cancer patients who had not received chemotherapy (Stratford et al., 2012; Zhao et al., 2016). A protein downstream of *RSK2* named Y-box binding protein-1(*YB-1*) was reported to transform human mammary epithelial cells in the development of basal-like breast cancer (Davies et al., 2014). Another study indicated that *RSK2* activation status positively correlates with patient response to anti-estrogen hormonal therapies and is required for estrogen receptor+ (ER+) breast cancer tumorigenesis (Clark et al., 2001). Several drug trials illustrated that by suppressing *RSK2* expression, the metastasis of human epidermal growth factor receptor 2+ (*HER2*+) breast cancer was repressed (Mao et al., 2016), the ability of migration and invasion of lung cancer cell was inhibited (Lee et al., 2015), and the carcinogenesis of ultraviolet radiation-induced skin cancer was prevented (Yao et al., 2014). These prompt us to investigate whether *RSK2* might be a potential biomarker that can act as a promising biomarker of breast cancer or a novel therapy target for a specific subtype of breast cancer. However, *RSK2* expression is rarely associated with clinical practice, which drives us to investigate the association between *RSK2* expression and clinical parameters and prognosis of breast cancer patients.

In the case of very limited clinical studies of *RSK2*, we performed a meta-analysis using public electronic databases ArrayExpress (Parkinson et al., 2007) and Gene Expression Omnibus (GEO) (Clough and Barrett, 2016) to summarize and evaluate the clinical significance of *RSK2* in breast cancer patients and in order to explore the possibility of *RSK2* expression as a predictive marker of clinicopathological parameters and prognosis in primary breast cancer, so as to screen high-risk patients or to provide new targets and directions for the treatment of breast cancer patients with specific molecular subtypes.

METHODS

Literature Search

We conducted a search of a series matrix files in the electronic database ArrayExpress (ArrayExpress, 2017) uploaded from January 1, 2008 to November 1, 2017 using the search filter "breast cancer," "*Homo sapiens*," "RNA assay," "array assay," and "all assay." We also conducted a search of a series matrix files in the GEO database (NCBI, 2017) uploaded from January 1, 2008 to November 1, 2017 using the search filter "breast cancer," "*Homo sapiens*," "series," and "expression profiling by array." A total of 207 and 227 expressions by array dataset were listed in

TABLE 1 | Studies included in the meta-analysis.

First author	GSE ID	Platform	Sample (N)	Year	Country or area	Duration (months)	Quality score
Calabrò et al. (2009)	GSE10510	GPL6486	152	2009	Germany	36	9
Haakensen et al. (2010)	GSE18672	GPL6848	143	2010	Norway	24	7
Enerly et al. (2011)	GSE19783	GPL6480	115	2011	Norway	44	8
Kao et al. (2011)	GSE20685	GPL570	327	2011	Taiwan	168	9
Terunuma et al. (2014)	GSE39004	GPL6244	108	2014	USA	127	8
Clarke et al. (2013)	GSE42568	GPL570	121	2013	Ireland	60	7
Gruosso et al. (2016)	GSE45827	GPL570	155	2016	France	NA	7
Lu et al. (2008)	GSE5460	GPL570	129	2008	USA	NA	8
Tofigh et al. (2014)	GSE58644	GPL6244	317	2014	Switzerland	NA	9
Huang et al. (2015)	GSE59595	GPL17581	175	2015	Italy	12	9
Chanrion et al. (2008)	GSE9893	GPL5049	155	2008	France	156	9
Li et al. (2010)	GSE19615	GPL570	115	2010	USA	NA	7
Wang et al. (2015)	GSE93601	GPL22920	1,110	2015	USA	324	9

the ArrayExpress and GEO databases, respectively. Relevant literatures were found in GEO database using the GSE ID.

Inclusion and Exclusion Criteria

This meta-analysis collects data from primary breast cancer patient's tumor tissue to assess the relationship of *RSK2* expression and the clinical parameters of breast cancer patients, such as clinicopathological features and prognostic factors. Inclusion criteria are as follows: a) patients in the study have not been selected, or the selection had no effect on clinical indicators given the *RSK2* expression might be different among different subtypes of breast cancer and different clinical status of patients; b) the sample size in each file is greater than 100 and comes from the same study; c) sample comes from breast cancer patients' primary tumor tissue or normal tissue; d) the expression of *RSK2* was efficient and available from the series matrix file; and e) any of a patient's clinicopathological features and prognostic factors can be extracted from the file. When samples in a file are duplicated with samples in another file, we selected the larger or qualified one. The exclusion criteria are as follows: a) samples were gathered from different studies; b) the original study could not be found; c) the sample size in the document is inconsistent with the number of patients in the study, and the replicated samples could not be found in the file; and d) the file was ineligible after invalid samples were removed. Files were filtered by two researchers separately; the disagreement was resolved through discussion. The eligible files are listed on **Table 1**.

Data Extraction

For observational studies, the Newcastle-Ottawa Quality Assessment Scale (NOS) was employed for assessing the quality of these studies. All data was abstracted by using a standardized data collection form, with information recorded as follows: first author's name, publication year, country of origin, number of cases and controls, detection method, GSE ID, and platform of detection. Each sample's *RSK2* expression and corresponding clinicopathological features and prognostic factors were extracted from the series matrix file, including age, tissue, ER status, progesterone receptor (PR) status, *HER2* status, lymph node infiltration, histological grade, TNM stage, tumor size, tumor type, metastasis, intrinsic subtype (by PAM50), overall survival time (OS), disease-free survival time (DFS), relapse-free survival time (RFS), and relevant status of the patient. Samples with incomplete information or data described above were removed.

Data extraction is conducted by two researchers, respectively, and disagreements were resolved by discussion.

Statistical Analysis

Statistical analysis was conducted by the guidelines proposed by the Meta-Analysis of Observational Studies group (Stroup et al., 2000). Median was used as the cutoff value to determine the level of *RSK2* expression because there was no suitable cutoff value to help us distinguish the expression status of *RSK2*. Odds ratio (OR) was employed for evaluating the association between *RSK2* expression and clinicopathological features. Hazard ratio (HR) and 95% credible interval (CI) were appraised to assess the association between *RSK2* expression and prognostic indicators, including OS, DFS, and RFS by using IBM SPSS Statistics 24. Heterogeneity of the OR and HR was calculated by using the Cochran's Q and I^2 test. A random-effect model was applied when $p < 0.1$ or $I^2 > 50\%$. When heterogeneity was absent, a fixed-effect model was employed. Begg's rank correlation method and Egger's weighted regression methods were employed to assess publication bias. STATA software package (version 12.0) was used to calculate pooled ORs, HRs, and corresponding 95% CI; all p values were two tailed.

Given the limited prognostic information of specific breast cancer patients in those GSE files, we used the Kaplan–Meier Plotter (Hou et al., 2017; Kaplan-Meier plotter), an online database including gene expression data and clinical data, to assess the prognostic value of *RSK2* in breast cancer. The patient samples were divided into two cohorts according to the median expression of the gene (high vs. low expression).

RESULTS

Search Results

The flow diagram for the recognition of eligible studies is presented in **Figure 1**. There were 207 and 227 GSE files

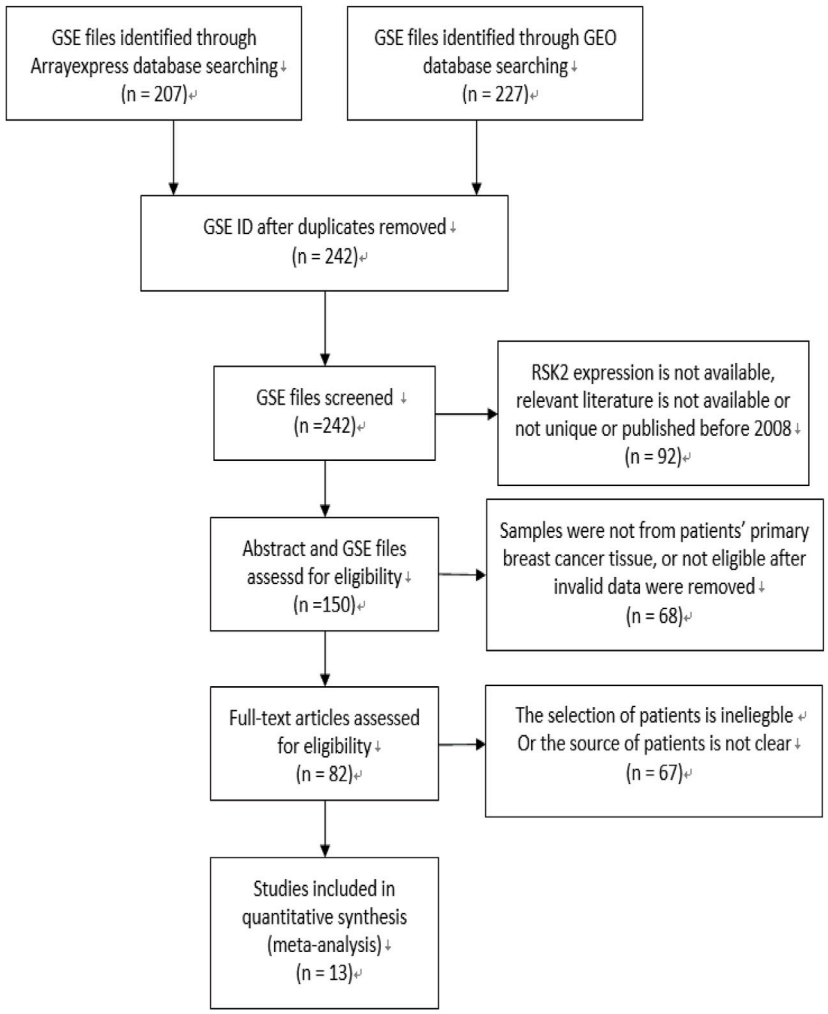


FIGURE 1 | Flow diagram of literature selection. The search and selection process for the expression of *RSK2* of breast cancer patients and the number of eligible studies.

identified from the ArrayExpress database and GEO database, respectively. After duplicates were removed, *RSK2* expression, abstract, and full text were checked, and 13 GSE files from 13 independent studies involving 3,122 patients were identified by our search strategy. The features of the 13 studies are listed in **Table 1**. Ineligible samples and data were removed, such as samples from cell, blood, or distant metastasis. When sample was detected multiple times, the one with the highest *RSK2* expression was selected. Histological grades I and II were grouped as low-grade disease, and III was grouped as high-grade disease. Clinical stages I and II were grouped as early-stage disease, and III and IV were grouped as late-stage disease. Tumors larger than 2 cm were grouped as large tumors, and the rest were grouped as small tumors. Patients were divided into high-age group and low-age group, with 55 years old as the cutoff value. Clinical stage of GSE20685 was not available, which we estimated depending on the T, N, and M stage shown in the GSE file using the NCCN guidelines of breast cancer (version 2. 2011). Data in GSE39004 were only used for comparing *RSK2* expression between normal tissues and tumor tissues, because its tumor sample size is not large enough.

RSK2 Expression in Breast Tumor Tissues was Lower Than That in Normal Breast Tissue and Enriched in Basal-Like Breast Cancer

Our meta-analysis demonstrated that *RSK2* expression in breast cancer tissue was lower than that in normal tissue (pooled OR = 0.54, 95% CI: 0.44–0.67, Cochran's Q test p = 0.14, I^2 = 41.7%) (**Figure 2A**). There was no statistically significant difference between ductal carcinoma and lobular carcinoma (pooled OR = 0.75, 95% CI: 0.35–1.60, Cochran's Q test p = 0.104, I^2 = 51.3%) (**Figure 2B**). The relationship of *RSK2* expression and molecular subtype was analyzed in our meta-analysis. The expression of *RSK2* was obviously different between the luminal subtype and basal subtype of breast cancer (pooled OR = 0.25, 95% CI: 0.08–0.80, Cochran's Q test p = 0.06, I^2 = 63.5%) (**Figure 2C**), and no

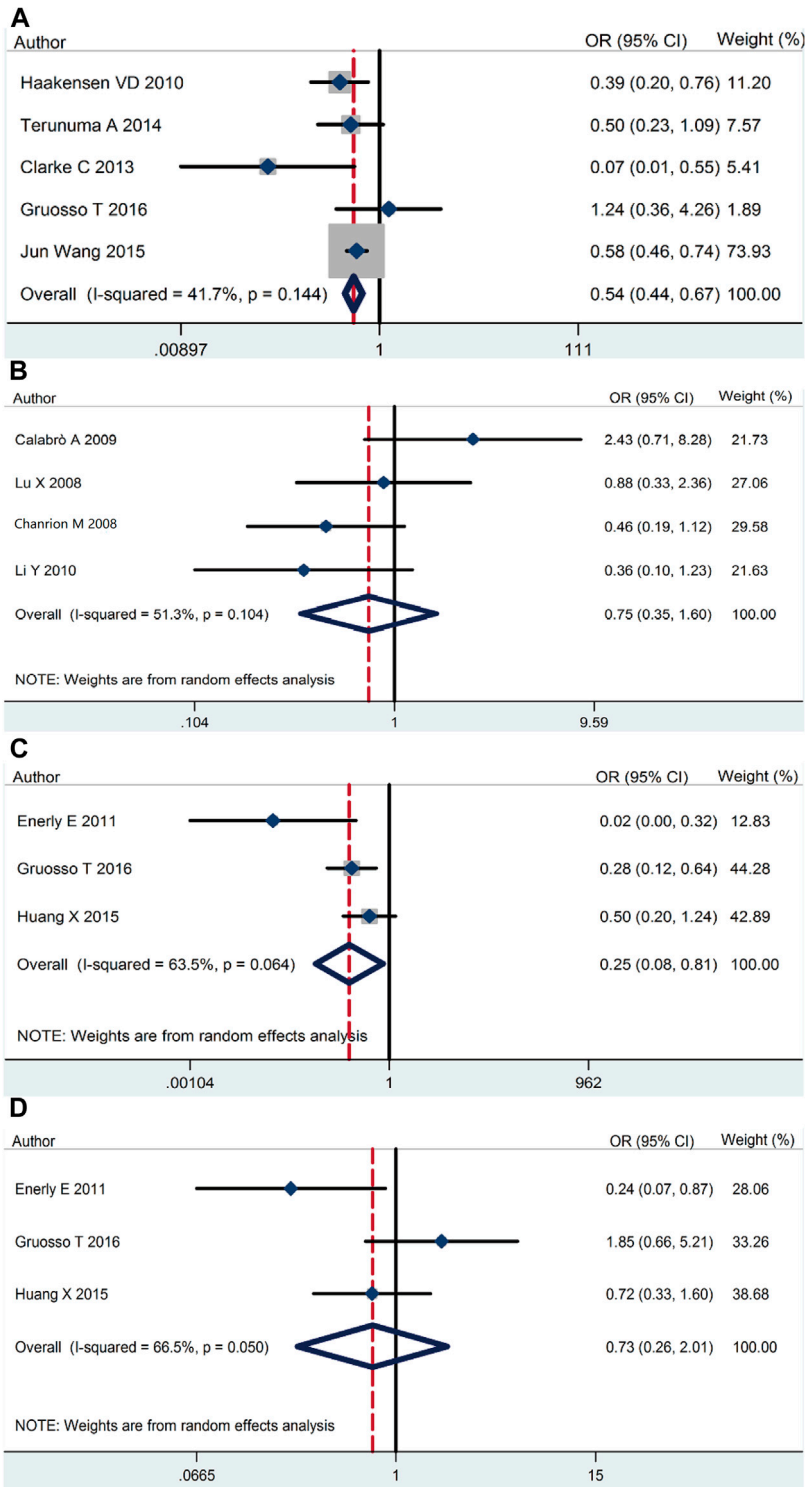


FIGURE 2 | *RSK2* expression among different types of breast tumor tissues as well as normal breast tissues. **(A)** *RSK2* expression in breast cancer tissue compared with normal tissue. **(B)** The association between *RSK2* expression and the ductal breast cancer relative to the lobular breast cancer. **(C)** *RSK2* expression in luminal breast cancer compared with basal-like breast cancer. **(D)** *RSK2* expression in luminal A breast cancer compared with luminal B breast cancer.

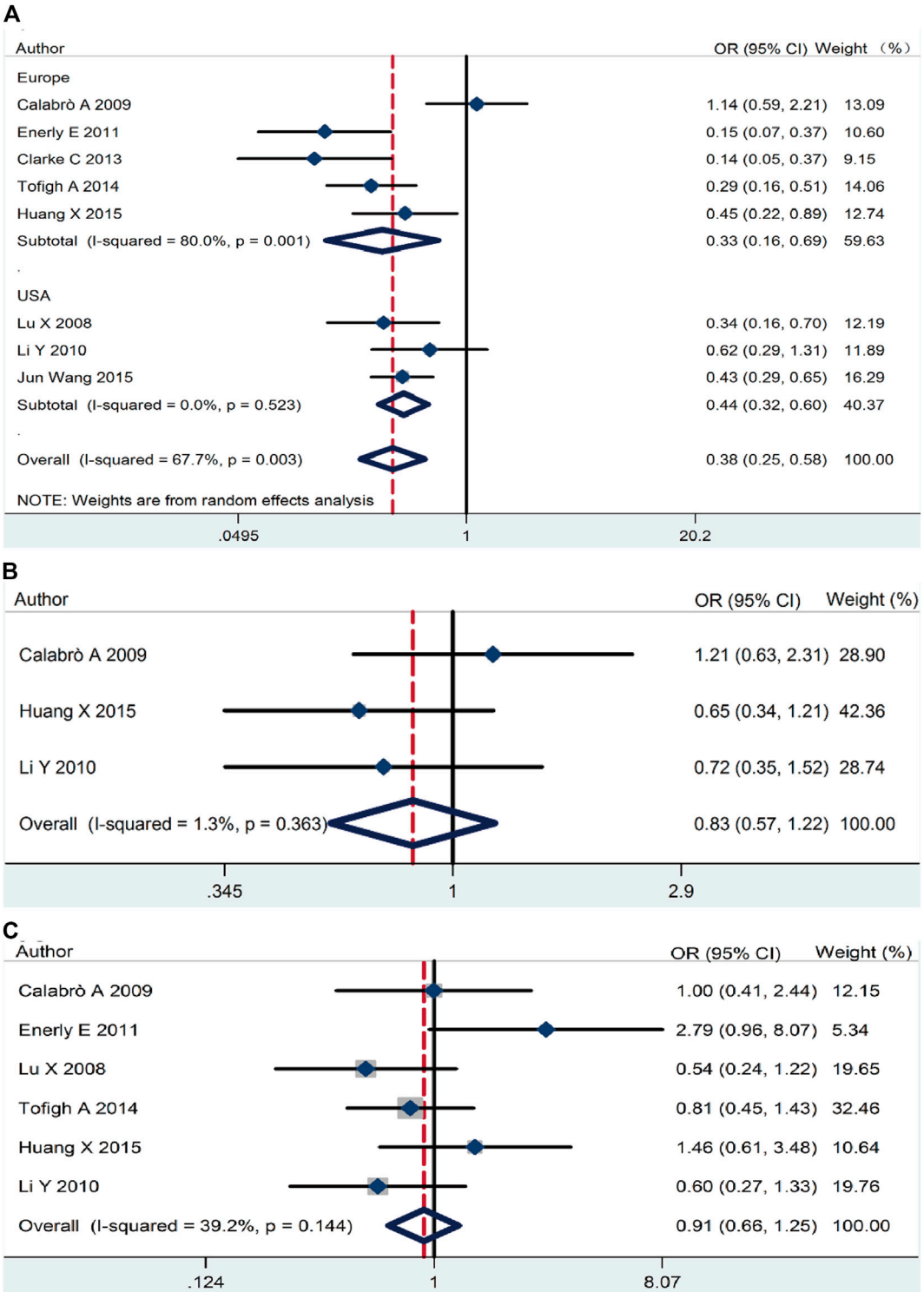


FIGURE 3 | The association between *RSK2* expression and the three main breast cancer biomarkers. **(A)** The association between *RSK2* expression and estrogen receptor (ER) status. **(B)** The association between *RSK2* expression and progesterone receptor (PR) status. **(C)** The association between *RSK2* expression and *HER2* status.

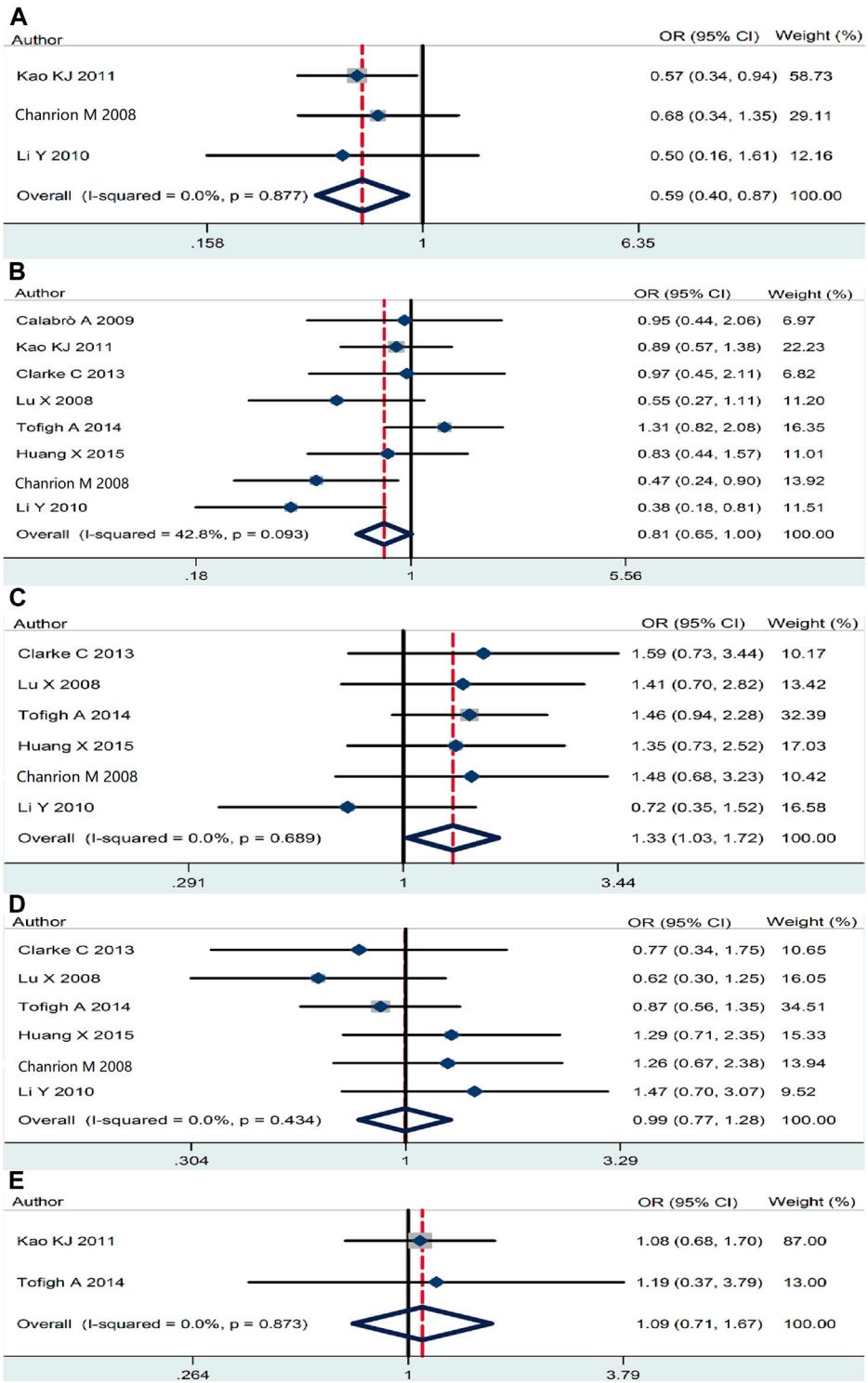


FIGURE 4 | The association between *RSK2* expression and other clinicopathological characters. **(A)** The association between *RSK2* expression and distant metastasis. **(B)** The association between *RSK2* expression and lymph node infiltration. **(C)** The association between *RSK2* expression and histological grade. **(D)** The association between *RSK2* expression and tumor size. **(E)** The association between *RSK2* expression and clinical stage of breast cancer.

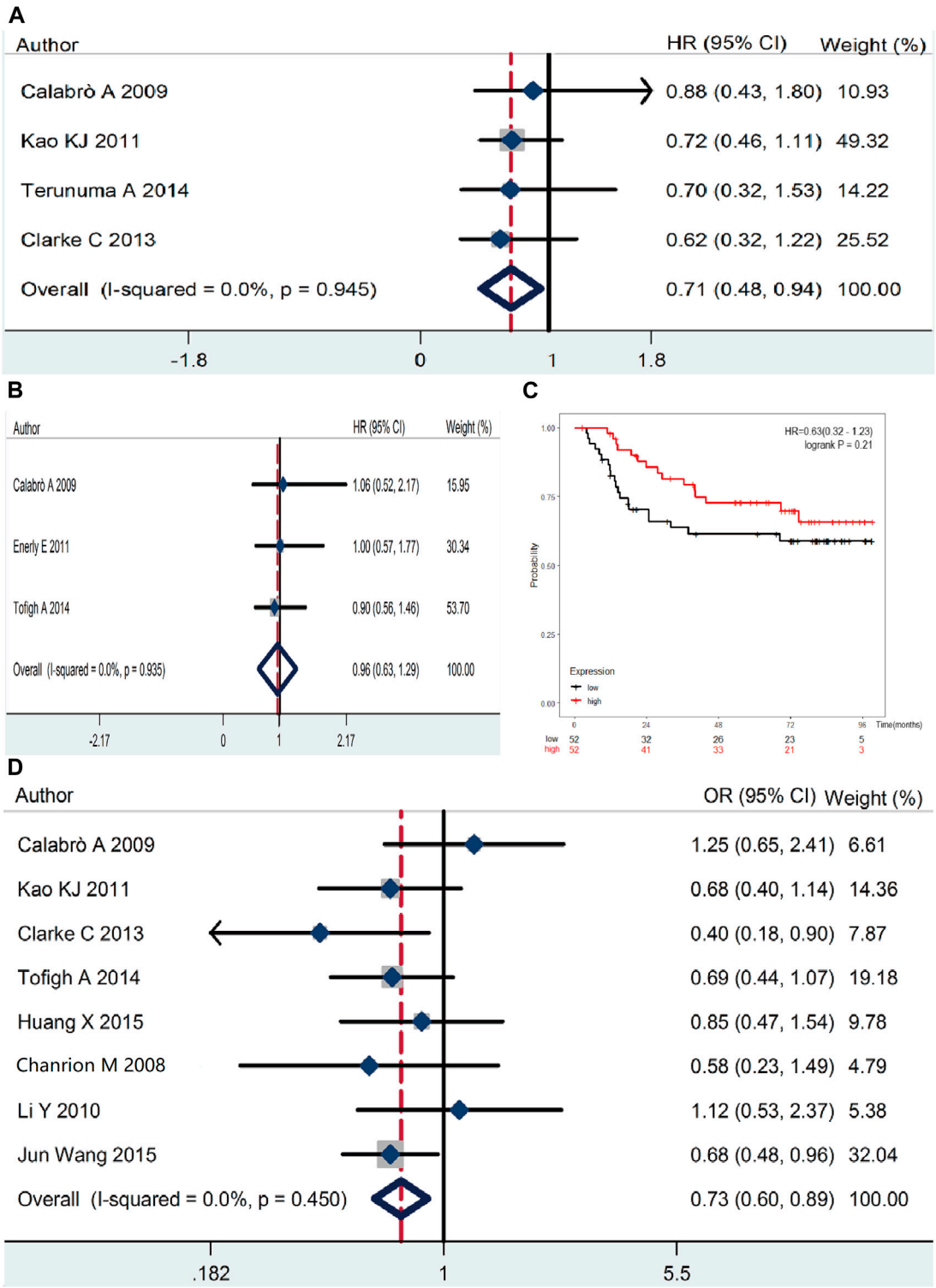


FIGURE 5 | The association between *RSK2* expression and survival outcome of breast cancer patients. **(A)** The association between *RSK2* expression and breast cancer overall survival (OS). **(B)** The association between *RSK2* expression and breast cancer disease-free survival (DFS). **(C)** The association between *RSK2* expression and age of breast cancer patients.

distinctive *RSK2* expression was found between luminal A and luminal B subtypes of breast cancer (pooled OR = 0.73, 95% CI: 0.26–2.01, Cochran's *Q* test $p = 0.05$, $I^2 = 66.5\%$) (**Figure 2D**).

RSK2 Expression is Negatively Correlated With ER Status

On the basic data obtained, we analyzed the relationship between *RSK2* expression and the three main breast cancer biomarkers. *RSK2* expression was inversely correlated with ER expression (pooled OR = 0.38, 95% CI: 0.25–0.58, Cochran's *Q* test $p = 0.009$, $I^2 = 67.7\%$) (**Figure 3A**). There was no statistically significant relationship between PR expression (pooled OR = 0.83, 95% CI: 0.57–1.22, Cochran's *Q* test $p = 0.36$, $I^2 = 1.3\%$) (**Figure 3B**) and *HER2* expression (pooled OR = 0.91, 95% CI: 0.66–1.25, Cochran's *Q* test $p = 0.14$, $I^2 = 39.2\%$) (**Figure 3C**).

RSK2 Expression Effects on Breast Cancer Progression

The association of *RSK2* expression with breast cancer clinicopathological features was also analyzed. High *RSK2* expression reduced the possibility of distant metastasis (pooled OR = 0.59, 95% CI: 0.41–0.87, Cochran's *Q* test $p = 0.88$, $I^2 = 0.0\%$) (**Figure 4A**) and lymph node metastasis (pooled OR = 0.81, 95% CI: 0.65–0.998, Cochran's *Q* test $p = 0.09$, $I^2 = 42.8\%$) (**Figure 4B**). The overexpression of *RSK2* positively correlated with histological grade (pooled OR = 1.33, 95% CI: 1.03–1.72, Cochran's *Q* test $p = 0.69$, $I^2 = 0.0\%$) (**Figure 4C**). However, *RSK2* expression has no statistically significant relationship with other biological characters of breast cancer, including tumor size (pooled OR = 0.995, 95% CI: 0.77–1.28, Cochran's *Q* test $p = 0.43$, $I^2 = 0.0\%$) (**Figure 4D**) and clinical stage (pooled OR = 1.09, 95% CI: 0.71–1.67, Cochran's *Q* test $p = 0.87$, $I^2 = 0.0\%$) (**Figure 4E**).

High *RSK2* Expression is Indicative of Longer OS in Breast Cancer Patients

We evaluated the association between *RSK2* expression level and survival outcome of breast cancer patients. The results indicate that *RSK2* overexpression was statistically associated with the OS rate of breast cancer patients (pooled HR = 0.71, 95% CI: 0.48–0.94, Cochran's *Q* test $p = 0.95$, $I^2 = 0.0\%$) (**Figure 5A**), while there was no significant relationship between *RSK2* expression and DFS (pooled HR = 0.96, 95% CI: 0.63–1.29, Cochran's *Q* test $p = 0.94$, $I^2 = 0.0\%$) (**Figure 5B**). Only one GSE file (GSE42568) involving 104 patients has the data of RFS, and there was no statistical significance between them (HR = 0.62, 95% CI: 0.32–1.22, log rank test $p = 0.21$) (**Figure 5C**). We did not find a significant difference in survival outcome between basal-like breast cancer and luminal breast cancer for the limited sample capacity and accessible data.

RSK2 Expression is Negatively Correlated With the Age of Breast Cancer Patients

We also recorded the corresponding age of every sample and grouped them into low-age group and high-age group with 55 years old as the cutoff value, which was randomly selected. Interestingly, we found that *RSK2* expression is lower in the high-age group (pooled OR = 0.73, 95% CI: 0.60–0.89, Cochran's *Q* test $p = 0.45$, $I^2 = 0.0\%$) (**Figure 5D**). In order to investigate whether *RSK2* expression decreased with age, we calculated the relationship between *RSK2* expression and age involving 508 normal tissue samples in GSE93601, and no statistically significant correlation was found (OR = 0.95, 95% CI: 0.66–1.36). The status of p53 was available in GSE19783, which involves 110 patients, indicating there was a positive relationship between *RSK2* and P53 mutation (OR = 4.09, 95% CI: 1.74–9.22, Cochran's *Q* test $p = 0.0007$).

Prognostic Value of Various *RSK2* Among Different Molecular Subtypes of Breast Cancer

The online database Kaplan–Meier was employed to evaluate the impact of *RSK2* expression on the prognostic outcome in different molecular subtypes of breast cancer, indicating that elevated *RSK2* expression predicts a favorable OS (luminal A breast cancer: HR = 1.04, 95% CI: 0.74–1.48, log rank $p = 0.81$; luminal B breast cancer: HR = 0.67, 95% CI: 0.46–0.97, log rank $p = 0.034$; basal-like breast cancer: HR = 0.48, 95% CI: 0.28–0.82, log rank $p = 0.006$; *HER2*+ breast cancer: HR = 0.7, 95% CI: 0.37–1.35, log rank $p = 0.29$) (**Figure 6**) and distant metastasis-free survival (DMFS) (luminal A breast cancer: HR = 0.96, 95% CI: 0.72–1.28, log rank $p = 0.78$; luminal B breast cancer: HR = 0.63, 95% CI: 0.44–0.9, log rank $p = 0.009$; basal-like breast cancer: HR = 0.54, 95% CI: 0.32–0.92, log rank $p = 0.021$; *HER2*+ breast cancer: HR = 1, 95% CI: 0.54–1.87, log rank $p = 0.99$) (**Figure 7**) in basal-like and luminal B breast cancer, but not in luminal A and *HER2*+ breast cancer. The overexpression of *RSK2* predicts a favorable prognostic value of RFS (**Figure 8**) in all those subtypes of breast cancer (luminal A breast cancer: HR = 0.78, 95% CI: 0.65–0.92, log rank $p = 0.004$; luminal B breast cancer: HR = 0.68, 95% CI: 0.56–0.82, log rank $p < 0.001$; basal-like breast cancer: HR = 0.67, 95% CI: 0.52–0.87, log rank $p = 0.002$; *HER2*+ breast cancer: HR = 0.51, 95% CI: 0.35–0.76, log rank $p < 0.001$).

The sample capacity is very large in GSE93601, which may have a great impact on the statistical results. We reanalyzed the data after removing the data from GSE93061 and got the same result as the previous one.

Publication Bias

Publication bias statistics were obtained using the Begg's test and Egger's test, which did not indicate any significant publication bias (**Table 2**).

DISCUSSION

RSK2 is an X-linked dominant gene and acts as a modulator of craniofacial development, and the mutation of *RSK2* was responsible for Coffin–Lowry syndrome (Laugel-Haushalter

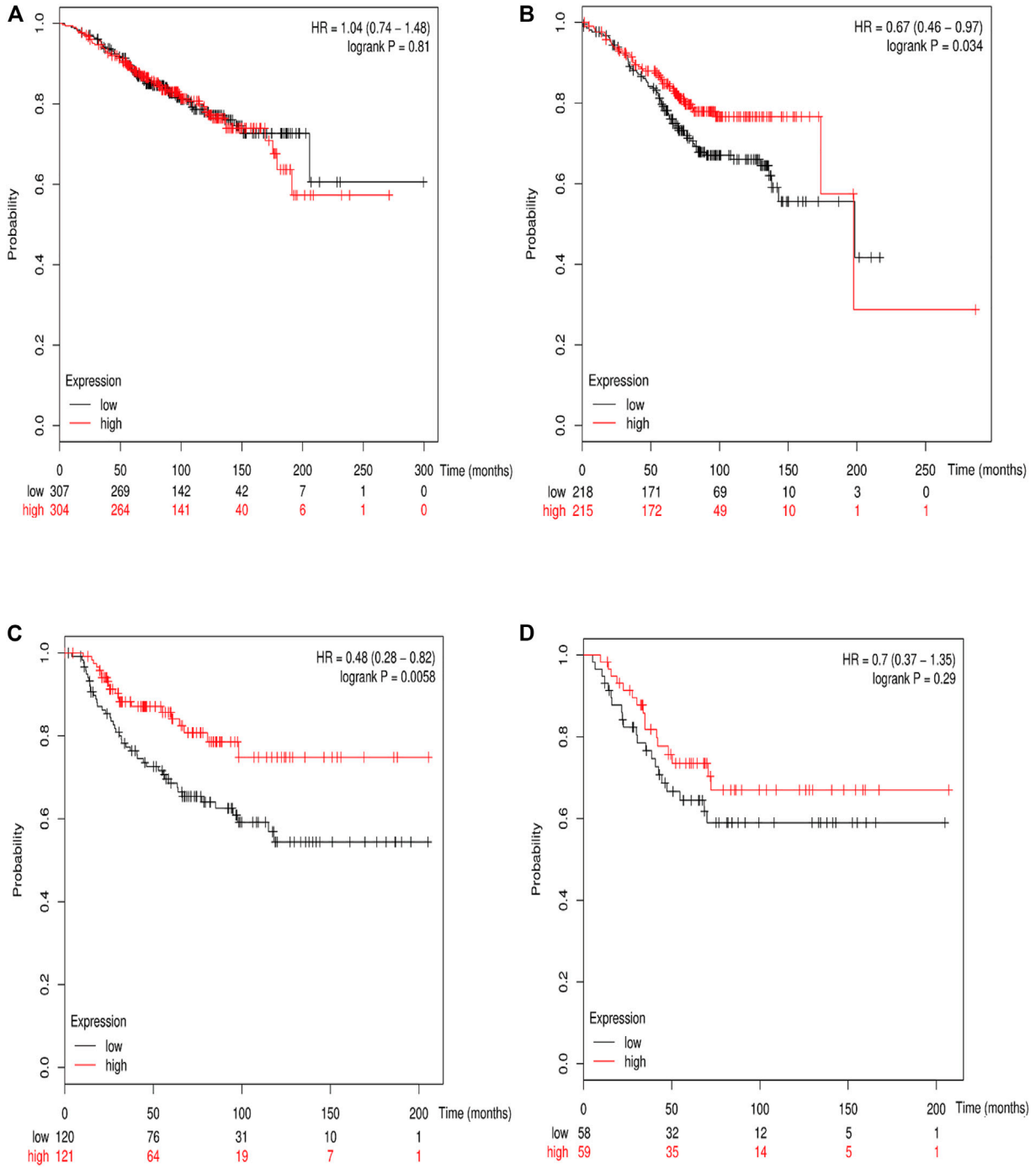


FIGURE 6 | The association between *RSK2* expression and OS in different molecular subtypes of breast cancer. **(A)** The association between *RSK2* expression and OS in luminal A breast cancer. **(B)** The association between *RSK2* expression and OS in luminal B breast cancer. **(C)** The association between *RSK2* expression and OS in basal-like breast cancer. **(D)** The association between *RSK2* expression and OS in *HER2+* breast cancer.

et al., 2014). It is generally believed that *RSK2* plays an important role in the tumorigenesis, migration, invasion, cell proliferation, and response to stress (Sulzmaier and Ramos, 2013; Laugel-Haushalter et al., 2014; Alesi et al., 2016). Precisely measuring the prognostic value of *RSK2* may help to guide individual therapies for breast cancer patients. Our meta-analysis takes advantage of a public electronic database to evaluate the association between the abundance of *RSK2* mRNA and the clinical parameters of breast cancer patients for the first time. Although it is not

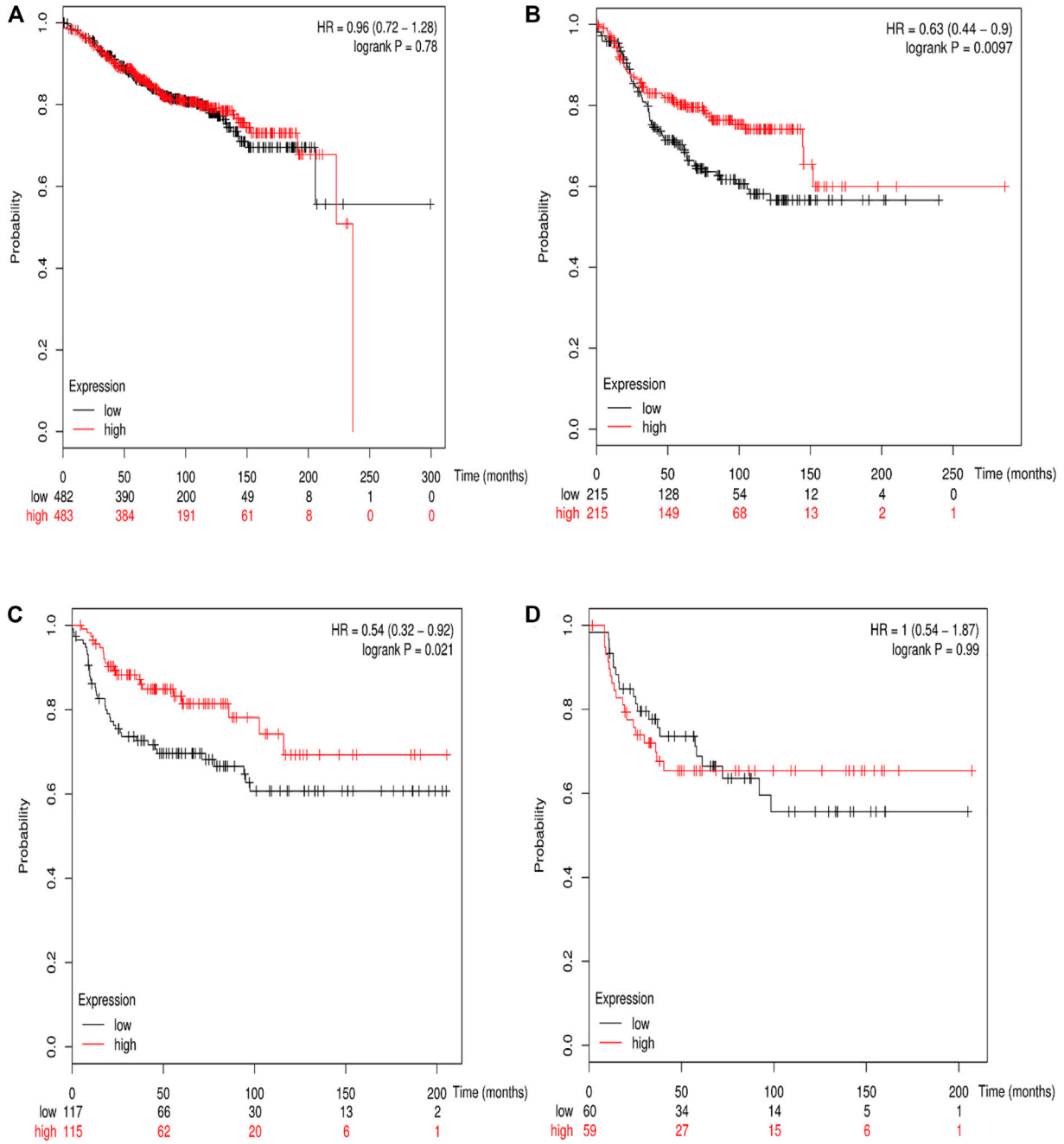


FIGURE 7 | The association between *RSK2* expression and distant metastasis-free survival (DMFS) in different molecular subtypes of breast cancer. **(A)** The association between *RSK2* expression and DMFS in luminal A breast cancer. **(B)** The association between *RSK2* expression and DMFS in luminal B breast cancer. **(C)** The association between *RSK2* expression and DMFS in basal-like breast cancer. **(D)** The association between *RSK2* expression and DMFS in *HER2+* breast cancer.

possible to draw conclusions about causality, these findings suggest that *RSK2* is a potential biomarker in breast cancer, especially in a specific subtype of breast cancer, and might provide new perspective of the interaction between *RSK2* and breast cancer.

From our research, *RSK2* expression was overexpressed in basal-like breast cancer and higher histological grade breast cancer and negatively correlated with estrogen receptor. These results corresponded with previous studies that suggest that *RSK2* expression is highest in basal-like breast cancer and those with the highest histological grade (Stratford et al., 2012; Zhao et al., 2016). A protein downstream of *RSK2*, namely *YB-1*, transforms human mammary epithelial cells through chromatin remodeling leading to the development of basal-like breast cancer (Davies et al., 2014). Inactivating *YB-1* can depress tumor-initiating cells of basal-like breast cancer. A study suggested that ER-α physically

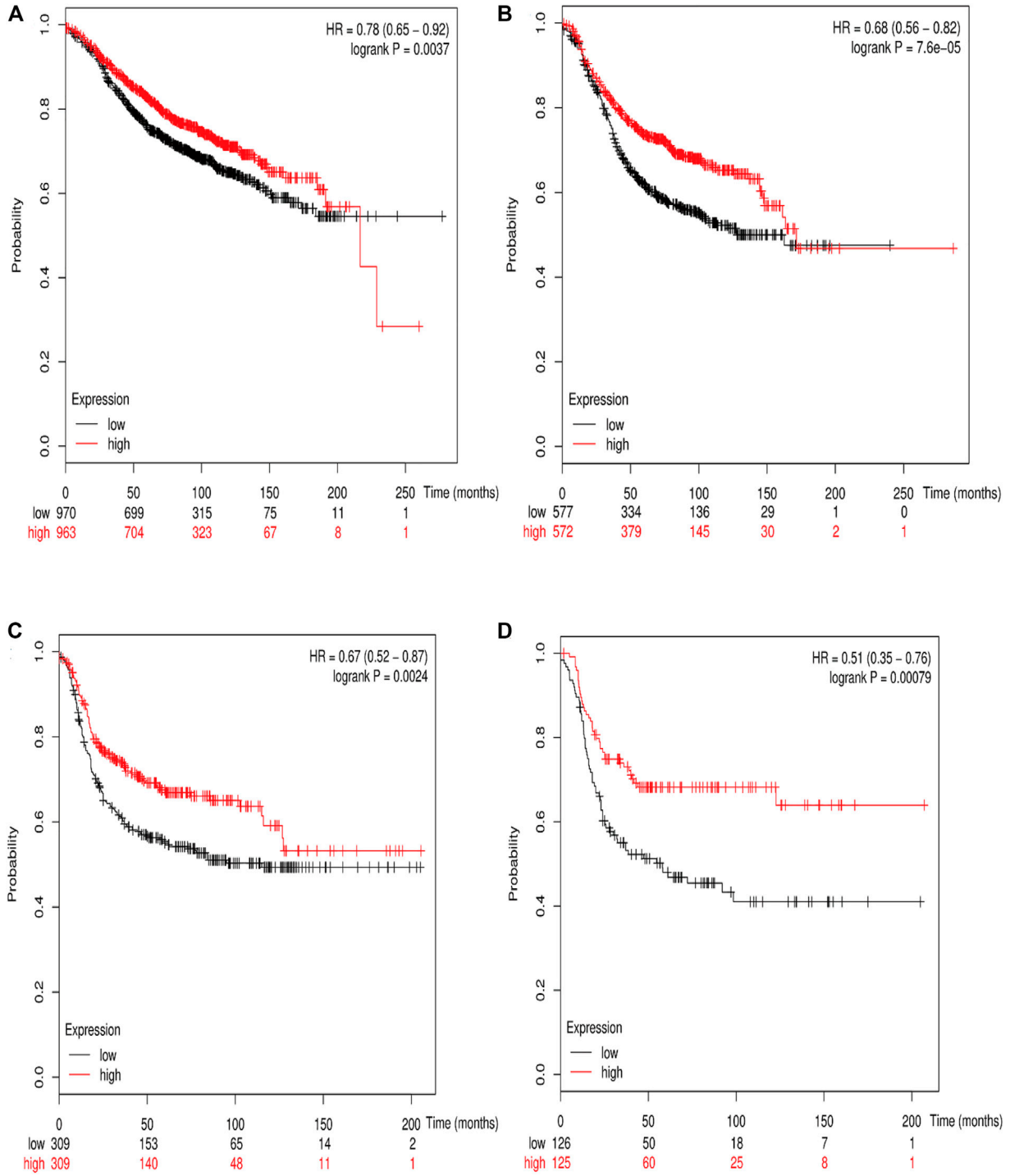


FIGURE 8 | The association between *RSK2* expression and relapse-free survival (RFS) in different molecular subtypes of breast cancer. **(A)** The association between *RSK2* expression and RFS in luminal A breast cancer. **(B)** The association between *RSK2* expression and RFS in luminal B breast cancer. **(C)** The association between *RSK2* expression and RFS in basal-like breast cancer. **(D)** The association between *RSK2* expression and RFS in *HER2+* breast cancer.

interacts with *RKS2*, resulting in the accumulation of *RSK2* in nuclear sequestration, and *RSK2* can promote neoplastic transformation and facilitate metastatic tumor growth of ER+ breast cancer (Ludwik et al., 2018), but there was no explanation for *RSK2* expression negatively correlating with ER status. However, there is no obvious statistical significance between *RSK2* expression and progesterone receptor on the basic data. Luminal A breast cancer is an ER-positive breast cancer with a

TABLE 2 | Publication bias tested by Egger's test and Begg's test.

	Egger's test *P*	Begg's test *P*
Normal tissue and tumor tissue	0.476	0.806
ER status	0.412	0.266
PR status	0.936	1
HER2 status	0.264	0.260
Lymph node infiltration	0.149	0.536
Tumor size	0.890	1
Histological grade	0.582	1
Basal-like and luminal breast cancer	0.325	1
Luminal A and B breast cancer	0.764	1
Distant metastasis	0.866	1
Lobular and ductal breast cancer	0.615	1
Age of patient	0.776	0.902
Overall survival	0.892	1
Disease-free survival	0.222	0.296

ER, estrogen receptor; PR, progesterone receptor.

lower histological grade, while luminal B breast cancer is an ER-positive breast cancer with a higher histological grade. Although *RSK2* is overexpressed in higher histological grade breast tumor, there is no obvious distinction between luminal A and luminal B breast cancer from our data. It may be partly due to the limited sample size of luminal A and luminal B breast cancer, and the distinction of *RSK2* expression was not large enough between them.

There were some unexpected results based on our data. It is generally believed that the expression of *RSK2* in cancer tissue is higher than that in normal tissue, and reducing the expression of *RSK2* can prevent tumorigenesis, tumor cell growth, and ability of migration and invasion (Lee et al., 2013; Yao et al., 2014; Mao et al., 2016; Zhao et al., 2016). However, based on our data, the contrary result was obtained. Furthermore, no evidence indicates that the immunohistochemistry outcome of *RSK2* was different from that of mRNA microarray assay. The overexpression of *RSK2* alone or *RSK2* combined with other biomarkers indicates a poor prognostic outcome was reported. For example, it has been shown that targeting *RSK2* with specific inhibitors or small interfering RNAs remarkably inhibits the growth and renewal of tumor-initiating cells in triple-negative breast cancer (TNBC) and that *RSK2* promotes migration through the ERK/MEK pathway (Stratford et al., 2012). In addition, Czaplinska et al. found that fibroblast growth factor receptor 2 (*FGFR2*) can form an indirect complex with *RSK2*, which may be involved in the progression of breast cancer and lead to poor prognosis in breast cancer patients (Czaplinska et al., 2016).

Based on the data collected from microarray, the OS of breast cancer patients is higher in *RSK2* high-expression patients than that in *RSK2* low-expression patients, and with the increase of *RSK2* expression, the potential of distant metastasis and lymph node infiltration decreased. Moreover, it is strange that the expression of *RSK2* is highest in basal-like breast cancer (TNBC), which is defined by the absence of the three main breast cancer biomarkers—i.e., a lack of expression of ER and PR and a lack of amplification or overexpression of *HER2*—and cooperates with poor prognosis and high risk of distant metastasis (Carey et al., 2010), but the OS, the potential of distant metastasis, and the lymph node metastases were more favorable in the *RSK2* high-expression group in the basic data from microarray. No relevant study was available to help us understand the mechanism under the paradoxical phenomenon. We hypothesize that *RSK2* plays a different role in different subtypes of breast cancer. We did not find a significant difference in survival outcomes between basal-like breast cancer and luminal breast cancer for the limited sample capacity and accessible data. In reference to the result from the online Kaplan–Meier Plotter, the overexpression of *RSK2* predicts more favorable prognostic value of RFS in all subtypes of breast cancer. As for the OS and DMFS, only basal-like and luminal B breast cancer patients were able to benefit from *RSK2* overexpression.

We found that *RSK2* expression is negatively correlated with the age of breast cancer patients for the first time, but there is no such relationship in normal tissue. There is also no statistically significant difference in *RSK2* expression between the early-stage group and late-stage group of breast cancer patients. It was reported that *RSK2* is sequestered in stress granules, which can aid cell survival in response to environmental stress by acting as sites of translational repression, and facilitates stress granule assembly to repress translation and to enhance cell survival (Eisinger-Mathason et al., 2008). The body's response to stress decreases with age and may provide a possible explanation for the phenomenon.

Heterogeneity tests are an essential part of a meta-analysis. In this study, minor heterogeneities were observed with respect to OS, DFS, tumor size, clinical stage, and histological grade; however, there were substantial heterogeneities with respect to ER status, *HER2* status, lymph node infiltration, and different subtypes of breast cancer. This unbalanced phenomenon could partly result from the detection method and accuracy of ER status, PR status, and *HER2* status being different from each other, and the data completeness obtained from GSE files was not identical. Three GSE files have efficient PR status, and the heterogeneity was not obvious among them, while the other three GSE files have identified the molecular subtype of breast cancer, which shows a significant heterogeneity. Patients from different areas may respond to the heterogeneities. There were no heterogeneities in United States patients, while the main heterogeneity of ER status and *HER2* status originates from different European countries when we conducted a subgroup analysis. Another significant heterogeneity was likely due to the detection platform. Publication bias is worth considering in a meta-analysis. In this study, there was no significant publication bias based on the Egger's and Begg's test.

There are still some limitations in this meta-analysis. First of all, the relevant studies and complete available data were limited, and the available clinical parameter is not homogenous among those matrix files. Secondly, the detection platform, method, and accuracy of hormone receptors are different among these studies. Thirdly, the therapy level and method are distinctive, and we cannot eliminate their effect. Lastly, we cannot ignore the publication bias. Some data are still unavailable.

For further verification, we could download the mRNA expression data and corresponding clinical information of breast cancer patients from other databases as a validation

cohort to verify the relationship between the *RSK2* expression level and the clinicopathological features as well as the prognosis of patients. In terms of experimental validation, future researchers could modify the expression of *RSK2* in different breast cancer cell lines and then perform various *in vitro* and mice xenografts *in vivo* trials to observe the effects of altering *RSK2* expression on the proliferation, apoptosis, cell cycle, metastasis, and invasion capabilities of breast cancer cells. Furthermore, the prognostic significance of *RSK2* could also be verified by measuring the protein expression level of *RSK2* by flow cytometry, western blotting, and immunohistochemistry staining on tumor and paracancerous normal tissues of breast cancer patients in conjunction with clinical information analysis. The strategies above could help to confirm the reliability of our meta-analysis findings based on *RSK2* mRNA expression and prognosis.

In conclusion, our meta-analysis was the first study that used microarray assay to research the association between *RSK2* expression and clinicopathological features and prognostic factors of primary breast cancer patients. Although some results corresponded with previous studies and some results were opposite to previous studies, both of them indicated *RSK2* is a promising biomarker of breast cancer. This study provides a new research direction and area of *RSK2*, while more experimental studies and elaborate research are needed to uncover the sealed mechanism of *RSK2* in breast cancer.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

YZ and HX contributed to conception and design of the study. KZ, SY, WY, and QL organized the database. YW and LZ performed the statistical analysis. KZ and SY wrote the first draft of the manuscript. XC, HX, XY, and YW wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant number 81772827).

REFERENCES

Alesi, G. N., Jin, L., Li, D., Magliocca, K. R., Kang, Y., Chen, Z. G., et al. (2016). RSK2 Signals through Stathmin to Promote Microtubule Dynamics and Tumor Metastasis. *Oncogene* 35 (41), 5412–5421. doi:10.1038/onc.2016.79

ArrayExpress (2017). ArrayExpress. Available at: https://www.ebi.ac.uk/arrayexpress/ArrayExpress (Accessed November 1, 2017).

Arul, N., and Cho, Y.-Y. (2013). A Rising Cancer Prevention Target of RSK2 in Human Skin Cancer. *Front. Oncol.* 3, 201. doi:10.3389/fonc.2013.00201

Calabrò, A., Beissbarth, T., Kuner, R., Stojanov, M., Benner, A., Asslaber, M., et al. (2009). Effects of Infiltrating Lymphocytes and Estrogen Receptor on Gene Expression and Prognosis in Breast Cancer. *Breast Cancer Res. Treat.* 116 (1), 69–77. doi:10.1007/s10549-008-0105-3

Carey, L., Winer, E., Viale, G., Cameron, D., and Gianni, L. (2010). Triple-negative Breast Cancer: Disease Entity or Title of Convenience. *Nat. Rev. Clin. Oncol.* 7 (12), 683–692. doi:10.1038/nrclinonc.2010.154

Chanrion, M., Negre, V., Fontaine, H., Salvetat, N., Bibeau, F., Grogan, G. M., et al. (2008). A Gene Expression Signature that Can Predict the Recurrence of Tamoxifen-Treated Primary Breast Cancer. *Clin. Cancer Res.* 14 (6), 1744–1752. doi:10.1158/1078-0432.CCR-07-1833

Clark, D. E., Poteet-Smith, C. E., Smith, J. A., and Lannigan, D. A. (2001). Rsk2 Allosterically Activates Estrogen Receptor Alpha by Docking to the Hormone-Binding Domain. *EMBO J.* 20 (13), 3484–3494. doi:10.1093/emboj/20.13.3484

Clarke, C., Madden, S. F., Doolan, P., Aherne, S. T., Joyce, H., O'Driscoll, L., et al. (2013). Correlating Transcriptional Networks to Breast Cancer Survival: a Large-Scale Coexpression Analysis. *Carcinogenesis* 34 (10), 2300–2308. doi:10.1093/carcin/bgt208

Clough, E., and Barrett, T. (2016). The Gene Expression Omnibus Database. *Methods Mol. Biol.* 1418, 93–110. doi:10.1007/978-1-4939-3578-9_5

Czaplinska, D., Mieczkowski, K., Supernat, A., Skladanowski, A. C., Kordek, R., Biernat, W., et al. (2016). Interactions between FGFR2 and RSK2-Implications for Breast Cancer Prognosis. *Tumor Biol.* 37 (10), 13721–13731. doi:10.1007/s13277-016-5266-9

Davies, A. H., Reipas, K. M., Pambid, M. R., Berns, R., Stratford, A. L., Fotovati, A., et al. (2014). YB-1 Transforms Human Mammary Epithelial Cells through Chromatin Remodeling Leading to the Development of Basal-like Breast Cancer. *Stem Cells* 32 (6), 1437–1450. doi:10.1002/stem.1707

Eisinger-Mathason, T. S. K., Andrade, J., Groehler, A. L., Clark, D. E., Muratore-Schroeder, T. L., Pasic, L., et al. (2008). Codependent Functions of RSK2 and the Apoptosis-Promoting Factor TIA-1 in Stress Granule Assembly and Cell Survival. *Mol. Cel.* 31 (5), 722–736. doi:10.1016/j.molcel.2008.06.025

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., et al. (2011). miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS One* 6 (2), e16915. doi:10.1371/journal.pone.0016915

Frödin, M., and Gammeltoft, S. (1999). Role and Regulation of 90 kDa Ribosomal S6 Kinase (RSK) in Signal Transduction. *Mol. Cel Endocrinol.* 151 (1-2), 65–77. doi:10.1016/s0303-7207(99)00061-1

Gruosso, T., Mieulet, V., Cardon, M., Bourachot, B., Kieffer, Y., Devun, F., et al. (2016). Chronic Oxidative Stress Promotes H2 AX Protein Degradation and Enhances Chemosensitivity in Breast Cancer Patients. *EMBO Mol. Med.* 8 (5), 527–549. doi:10.15252/emmm.201505891

Guo, Z.-F., and Kong, F.-L. (2021). Akt Regulates RSK2 to Alter Phosphorylation Level of H2A.X in Breast Cancer. *Oncol. Lett.* 21 (3), 187. doi:10.3892/ol.2021.12448

Haakensen, V. D., Biong, M., Lingjærde, O. C., Holmen, M. M., Frantzen, J. O., Chen, Y., et al. (2010). Expression Levels of Uridine 5'-Diphospho-Glucuronosyltransferase Genes in Breast Tissue from Healthy Women Are Associated with Mammographic Density. *Breast Cancer Res.* 12 (4), R65. doi:10.1186/bcr2632

Hou, G.-X., Liu, P., Yang, J., and Wen, S. (2017). Mining Expression and Prognosis of Topoisomerase Isoforms in Non-small-cell Lung Cancer by Using Oncomine and Kaplan-Meier Plotter. *PLoS One* 12 (3), e0174515. doi:10.1371/journal.pone.0174515

Huang, X., Dugo, M., Callari, M., Sandri, M., De Cecco, L., Valeri, B., et al. (2015). Molecular Portrait of Breast Cancer in China Reveals Comprehensive Transcriptomic Likeness to Caucasian Breast Cancer and Low Prevalence of Luminal A Subtype. *Cancer Med.* 4 (7), 1016–1030. doi:10.1002/cam4.442

Kao, K.-J., Chang, K.-M., Hsu, H.-C., and Huang, A. T. (2011). Correlation of Microarray-Based Breast Cancer Molecular Subtypes and Clinical Outcomes:

Implications for Treatment Optimization. *BMC Cancer* 11, 143. doi:10.1186/1471-2407-11-143

Kaplan-Meier plotter Kaplan-meier Plotter. Available at: http://kmplot.com/private/ (Accessed November 1, 2017).

Laugel-Haushalter, V., Paschaki, M., Marangoni, P., Pilgram, C., Langer, A., Kuntz, T., et al. (2014). RSK2 Is a Modulator of Craniofacial Development. *PLoS One* 9 (1), e84343. doi:10.1371/journal.pone.0084343

Lee, C.-J., Lee, M.-H., Yoo, S.-M., Choi, K.-I., Song, J.-H., Jang, J.-H., et al. (2015). Magnolin Inhibits Cell Migration and Invasion by Targeting the ERKs/RSK2 Signaling Pathway. *BMC Cancer* 15, 576. doi:10.1186/s12885-015-1580-7

Lee, M.-H., Huang, Z., Kim, D. J., Kim, S.-H., Kim, M. O., Lee, S.-Y., et al. (2013). Direct Targeting of MEK1/2 and RSK2 by Silybin Induces Cell-Cycle Arrest and Inhibits Melanoma Cell Growth. *Cancer Prev. Res.* 6 (5), 455–465. doi:10.1158/1940-6207.CAPR-12-0425

Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., et al. (2010). Amplification of LAPTM4B and YWHAZ Contributes to Chemotherapy Resistance and Recurrence of Breast Cancer. *Nat. Med.* 16 (2), 214–218. doi:10.1038/nm.2090

Lu, X., Lu, X., Wang, Z. C., Iglehart, J. D., Zhang, X., and Richardson, A. L. (2008). Predicting Features of Breast Cancer with Gene Expression Patterns. *Breast Cancer Res. Treat.* 108 (2), 191–201. doi:10.1007/s10549-007-9596-6

Ludwik, K. A., McDonald, O. G., Brenin, D. R., and Lannigan, D. A. (2018). ERα-Mediated Nuclear Sequestration of RSK2 Is Required for ER+ Breast Cancer Tumorigenesis. *Cancer Res.* 78 (8), 2014–2025. doi:10.1158/0008-5472.CAN-17-2063

Mao, L., Summers, W., Xiang, S., Yuan, L., Dauchy, R. T., Reynolds, A., et al. (2016). Melatonin Represses Metastasis in Her2-Postive Human Breast Cancer Cells by Suppressing RSK2 Expression. *Mol. Cancer Res.* 14 (11), 1159–1169. doi:10.1158/1541-7786.MCR-16-0158

NCBI (2017). Gene Expression Omnibus. Available at: https://www.ncbi.nlm.nih.gov/gds (Accessed November 1, 2017).

Nielsen, T. O., Jensen, M.-B., Burugu, S., Gao, D., Jørgensen, C. L. T., Balslev, E., et al. (2017). High-Risk Premenopausal Luminal A Breast Cancer Patients Derive No Benefit from Adjuvant Cyclophosphamide-Based Chemotherapy: Results from the DBCG77B Clinical Trial. *Clin. Cancer Res.* 23 (4), 946–953. doi:10.1158/1078-0432.CCR-16-1278

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., et al. (2007). ArrayExpress--a Public Database of Microarray Experiments and Gene Expression Profiles. *Nucleic Acids Res.* 35 (Database issue), D747–D750. doi:10.1093/nar/gkl995

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular Portraits of Human Breast Tumours. *Nature* 406 (6797), 747–752. doi:10.1038/35021093

Prat, A., Fan, C., Fernández, A., Hoadley, K. A., Martinello, R., Vidal, M., et al. (2015). Response and Survival of Breast Cancer Intrinsic Subtypes Following Multi-Agent Neoadjuvant Chemotherapy. *BMC Med.* 13, 303. doi:10.1186/s12916-015-0540-z

Siegel, R. L., Miller, K. D., and Jemal, A. (2018). Cancer Statistics, 2018. *CA: A Cancer J. Clinic.* 68 (1), 7–30. doi:10.3322/caac.21442

Stratford, A. L., Reipas, K., Hu, K., Fotovati, A., Brough, R., Frankum, J., et al. (2012). Targeting P90 Ribosomal S6 Kinase Eliminates Tumor-Initiating Cells by Inactivating Y-Box Binding Protein-1 in Triple-Negative Breast Cancers. *Stem Cells* 30 (7), 1338–1348. doi:10.1002/stem.1128

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., et al. (2000). Meta-analysis of Observational Studies in EpidemiologyA Proposal for Reporting. *JAMA* 283 (15), 2008–2012. doi:10.1001/jama.283.15.2008

Sulzmaier, F. J., and Ramos, J. W. (2013). RSK Isoforms in Cancer Cell Invasion and Metastasis. *Cancer Res.* 73 (20), 6099–6105. doi:10.1158/0008-5472.CAN-13-1087

Terunuma, A., Putluri, N., Mishra, P., Mathé, E. A., Dorsey, T. H., Yi, M., et al. (2014). MYC-driven Accumulation of 2-hydroxyglutarate Is Associated with Breast Cancer Prognosis. *J. Clin. Invest.* 124 (1), 398–412. doi:10.1172/JCI71180

Tofigh, A., Suderman, M., Paquet, E. R., Livingstone, J., Bertos, N., Saleh, S. M., et al. (2014). The Prognostic Ease and Difficulty of Invasive Breast Carcinoma. *Cel Rep.* 9 (1), 129–142. doi:10.1016/j.celrep.2014.08.073

Wang, J., Zhang, X., Beck, A. H., Collins, L. C., Chen, W. Y., Tamimi, R. M., et al. (2015). Alcohol Consumption and Risk of Breast Cancer by Tumor Receptor Expression. *Horm. Canc* 6 (5-6), 237–246. doi:10.1007/s12672-015-0235-0

Yao, K., Chen, H., Liu, K., Langfald, A., Yang, G., Zhang, Y., et al. (2014). Kaempferol Targets RSK2 and MSK1 to Suppress UV Radiation-Induced Skin Cancer. *Cancer Prev. Res.* 7 (9), 958–967. doi:10.1158/1940-6207.CAPR-14-0126

Yoo, S.-M., Lee, C.-J., An, H.-J., Lee, J. Y., Lee, H. S., Kang, H. C., et al. (2019). RSK2-Mediated ELK3 Activation Enhances Cell Transformation and Breast Cancer Cell Growth by Regulation of C-Fos Promoter Activity. *Int. J. Mol. Sci.* 20 (8), 1994. doi:10.3390/ijms20081994

Zhao, H., Martin, T. A., Davies, E. L., Ruge, F., Yu, H., Zhang, Y., et al. (2016). The Clinical Implications of RSK1-3 in Human Breast Cancer. *Anticancer Res.* 36 (3), 1267–1274.

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

The reviewer (FZ) declared a shared parent affiliation, with several of the authors (KZ, SY, WY, QL, YW, LZ, XC, HX, XY, YZ, and HX), to the handling editor at the time of the review.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 November 2021
doi: 10.3389/fgene.2021.755507

Transcriptomic and Proteomic Profiling of Human Stable and Unstable Carotid Atherosclerotic Plaques

Mei-hua Bao[1,2], Ruo-qi Zhang[2], Xiao-shan Huang[3], Ji Zhou[1], Zhen Guo[1], Bao-feng Xu[1,4]* and Rui Liu[1,5]**

[1]Academician Workstation, Changsha, China, [2]School of Stomatology, Changsha Medical University, Changsha, China, [3]Department of Pharmacology, Changsha Health Vocational College, Changsha, China, [4]First Hospital of Jilin University, Changchun, Jilin, China, [5]Department of VIP Unit, China-Japan Union Hospital of Jilin University, Changchun, China

Atherosclerosis is a chronic inflammatory disease with high prevalence and mortality. The rupture of atherosclerotic plaque is the main reason for the clinical events caused by atherosclerosis. Making clear the transcriptomic and proteomic profiles between the stabe and unstable atherosclerotic plaques is crucial to prevent the clinical manifestations. In the present study, 5 stable and 5 unstable human carotid atherosclerotic plaques were obtained by carotid endarterectomy. The samples were used for the whole transcriptome sequencing (RNA-Seq) by the Next-Generation Sequencing using the Illumina HiSeq, and for proteome analysis by HPLC-MS/MS. The lncRNA-targeted genes and circRNA-originated genes were identified by analyzing their location and sequence. Gene Ontology and KEGG enrichment was carried out to analyze the functions of differentially expressed RNAs and proteins. The protein-protein interactions (PPI) network was constructed by the online tool STRING. The consistency of transcriptome and proteome were analyzed, and the lncRNA/circRNA-miRNA-mRNA interactions were predicted. As a result, 202 mRNAs, 488 lncRNAs, 91 circRNAs, and 293 proteins were identified to be differentially expressed between stable and unstable atherosclerotic plaques. The 488 lncRNAs might target 381 protein-coding genes by *cis*-acting mechanisms. Sequence analysis indicated the 91 differentially expressed circRNAs were originated from 97 protein-coding genes. These differentially expressed RNAs and proteins were mainly enriched in the terms of the cellular response to stress or stimulus, the regulation of gene transcription, the immune response, the nervous system functions, the hematologic activities, and the endocrine system. These results were consistent with the previous reported data in the dataset GSE41571. Further analysis identified CD5L, S100A12, CKB (target gene of lncRNA MSTRG.11455.17), CEMIP (target gene of lncRNA MSTRG.12845), and SH3GLB1 (originated gene of hsacirc_000411) to be critical genes in regulating the stability of atherosclerotic plaques. Our results provided a comprehensive transcriptomic and proteomic knowledge on the stability of atherosclerotic plaques.

Keywords: atherosclerosis, unstable plaques, RNA-seq, proteome, transcriptome

OPEN ACCESS

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

Reviewed by:
Qiang Shi,
National Center for Toxicological Research (FDA), United States
Hong Zheng,
Stanford University, United States

****Correspondence:***
Mei-hua Bao
mhbao78@163.com
Bao-feng Xu
xubf@jlu.edu.cn
Rui Liu
liur@jlu.edu.cn

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

Received: *09 August 2021*
Accepted: *12 October 2021*
Published: *04 November 2021*

Citation:
Bao M-h, Zhang R-q, Huang X-s, Zhou J, Guo Z, Xu B-f and Liu R (2021) Transcriptomic and Proteomic Profiling of Human Stable and Unstable Carotid Atherosclerotic Plaques. Front. Genet. 12:755507. doi: 10.3389/fgene.2021.755507

INTRODUCTION

Atherosclerosis is a chronic inflammatory disease with high prevalence and mortality. The stability of atherosclerotic plaques is the main reason for its clinical manifestations. Unstable plaques, also known as vulnerable plaques, are characterized by a large lipid core, a thin fibrotic cap, less smooth muscle cells, less collagen, and elevated inflammation. The broken of unstable plaques blocks capillaries, forms thrombosis, and eventually triggers clinical manifestations, such as ischemic stroke and myocardial infarction. Figuring out the genes and proteins which play critical roles in the stability of atherosclerotic plaques is important.

Recent researches have demonstrated the functions of lncRNAs and circRNAs in atherosclerosis (Xiao et al., 2018; Fasolo et al., 2019; L.; Wang et al., 2019; Cao et al., 2020). For example, lncRNA FENDRR, LincRNA-p21, ANRIL, MIAT, CDK2B-AS1, PELATON participated in the formation and stability of atherosclerotic plaques. They interfered with the phagocytosis, lipid uptake, and reactive oxygen species formation during atherogenesis (Çekin et al., 2018; Hung et al., 2020; Ou et al., 2020). CircRNA is a plentiful, stable, diversified, and conserved class of non-coding RNA molecules that circularized from head to tail with a covalent bond of 5–3 (Jeck et al., 2013). It is involved in a wide range of biological and pathological processes, such as carcinogenesis and cardiovascular diseases. CircRNAs act as miRNA sponges, decoys, or scaffolds of gene expression (Poller et al., 2018). Studies reported the participation of cirRNA00044073, circRNA-PTPRA, circRNA_0003204, and circHIPK3 in atherosclerosis through affecting the autophagy, proliferation and invasion, and tube formation (Shen et al., 2019; Wei et al., 2020; Zhang, 2020; Zhang et al., 2020).

Recently, some studies have investigated the transcriptome profiles in atherosclerosis. In a microarray analysis, 236 lncRNAs and 488 mRNAs were identified to be differentially expressed in human advanced atherosclerotic plaques compared to the normal arterial intimae (Bai et al., 2019). An RNA-Seq identified 1,259 annotated and 381 new RNAs in coronary artery disease (Pan et al., 2019). A weighted gene correlation network analysis identified several key genes in ruptured atherosclerotic plaques and other aging diseases (Yang et al., 2016; Xu et al., 2020; Yang et al., 2020). However, no comprehensive study and analysis was performed on the whole transcriptome and proteome in stable and unstable plaques.

In our present study, we obtained the stable and unstable plaques from patients conducting carotid endarterectomy (CEA), measured the transcription profiles and protein profiles by RNA-Seq and HPLC-MS/MS, identified the differentially expressed (DE) mRNAs, lncRNAs, circRNAs, and proteins, analyzed the functions of these differentially expressed genes. The present study provided a comprehensive knowledge of the gene and protein expression profiles responsible for the stability of atherosclerotic plaques, and identified several essential RNAs and proteins.

METHODS AND MATERIALS

Patients and Samples

The atherosclerotic plaques were obtained from 10 patients undergoing CEA operation in the First Hospital of Jilin University (Changchun, Jilin, China) from July 2019 to November 2019. The plaques were fast-frozen in liquid nitrogen and stored at −80°C. The classification of unstable or stable plaques was carried out according to the criteria of the American Heart Association (AHA) (Hetterich H et al., 2016). Briefly, type I/II: near-normal wall thickness, no calcification; Type III: diffuse intimal thickening or small eccentric plaque, no calcification; Type IV/V: plaque with lipid or necrotic core surrounded by fibrous tissue with possible calcification; Type VI: complex plaque with possible surface defect, hemorrhage, or thrombus; Type VII: calcified plaque; Type VIII: fibrotic plaque without lipid core and with possible small calcification. Type I-II, III, VII, VIII were considered stable, while type IV-V, VI to be unstable plaques. Two independent investigators conducted the plaque classification.

The procedures were approved by the Ethics Committee of the First Hospital of Jilin University (No. 2019-272, Changchun, Jilin). Written informed consent was obtained from every participant. Eventually, 5 stable plaques and 5 unstable plaques were obtained for further analysis. The stable or unstable plaque from each patient was divided into two parts evenly. One part was used for whole transcriptome sequencing (RNA-Seq), while the other part was for LC-MS/MS detection.

RNA Extraction, Library Preparation, and RNA-Sequencing

The Next-Generation Sequencing (NGS) analysis was performed in the Shanghai Personalbio Technology Co., Ltd. (Shanghai, China). Total RNA was isolated using the Trizol reagent (Invitrogen, Carlsbad, CA, United States). The qualities and quantities of the RNA were measured using NanoDrop spectrophotometer (Thermo Scientific, Waltham, Massachusetts, United States). The integrity of the total RNA was determined by Bioanalyzer 2,100 (Agilent Technologies, Santa Clare, CA, United States) and 1% agarose gel electrophoresis.

Sequencing libraries were generated according to the following steps: poly-T oligo-attached magnetic beads were used to purify mRNA from total RNA. The mRNA was fragmented by divalent cations under elevated temperature in an Illumina proprietary fragmentation buffer. Then, the first strand cDNA was synthesized using random oligonucleotides and SuperScript II, followed by the second strand cDNA synthesis using DNA Polymerase I and RNase H. After adenylation of the 3' ends of the DNA fragments, Illumina PE adapter oligonucleotides were ligated to prepare for hybridization. The cDNA fragments with a length of 400–500 bp were selected. Then the library fragments were purified using the AMPure XP system (Beckman Coulter, Beverly, CA, United States). DNA fragments with ligated adaptor molecules on both ends were amplified using Illumina PCR Primer Cocktail in a 15 cycle PCR reaction. The products

were purified (AMPure XP system) and quantified using the Agilent high sensitivity DNA assay on a Bioanalyzer 2,100 system (Agilent). The sequencing library was then sequenced on NovaSeq 6,000 platform (Illumina) by Shanghai Personal Biotechnology Co. Ltd.

Samples are sequenced on the NovaSeq 6,000 platform to get image files, which were then transformed to the original data in FASTQ format (Raw Data). These Raw Data were filtered to get high-quality sequence (Clean Data) by Cutadapt (v1.15) software, which removed low-quality Reads and connectors. The reference genome and gene annotation files were downloaded from the genome website. The filtered reads were mapped to the reference genome (Homo-sapiens.GRCh38. dna.primary_ assembly. fa) using HISAT2 v2.0.5.

Identification of Differentially Expressed mRNA, lncRNA, and circRNA

The expression of mRNAs was identified by the HTSeq (0.9.1) statistics. The Read Count values on each gene were considered to be the original expression level, and then the FPKM method was used to standardize them. The expression difference of mRNA between stable and unstable plaque groups was analyzed by DESeq (1.30.0).

The lncRNAs were identified by the Stringtie software. Briefly, the transcripts with length >200 bp and exon number≥2 were identified first. Transcriptions with the class code of x/u/i (x stand for antisense lncRNA, u stand for intergenic lncRNA, i stand for intronic lncRNA) were identified secondly. LncRNAs with coverage >3 were identified as expressed lncRNAs. The newly identified lncRNA was nominated by the Stringtie automatically with a title of "MSTRG".

The remaining unmapped reads were considered to be candidates of circRNAs. On the candidate transcripts, 20 bp at each end was used as 5′ Anchor or 3' Anchor. The Anchors were then mapped to the reference sequence. If the sequences of the Anchors were reverse to the mapped sequence and the junction was consistent with the splicing pattern of AG-GT, it was determined to be a circRNA. The expression level of circRNAs was calculated by the method of Transcripts Per kilobase of exon model per Million mapped reads (TPM).

The DESeq was used to analyze the DE mRNAs, lncRNAs, circRNAs. RNAs with $|\log2FoldChange| > 1.0$ and p-value < 0.05 were identified as differentially expressed.

The MeV 4.9.0 software was used to perform clustering and visualizing the expression pattern of 20 DE mRNAs, lncRNAs, and circRNAs.

Verification of Differentially Expressed mRNA, lncRNA, and circRNA

The expression level of 4 DE mRNA, 4 DE lncRNA, 4 DE circRNA were verified by qPCR. The total RNAs from 5 stable and 5 unstable plaques were extracted by Trizol (Takara, Dalian, China). After concentration and quality evaluation, the total RNAs were reverse transcripted to cDNA by the PrimeScript RT reagent Kit with gDNA Eraser (perfect real time) kits (Takara, Dalian, China). The PCR reactions were conducted by Applied Biosystems Quantstudio 5 system with the following program: 95 °C 30 s, followed by 40 cycles of 95°C 5 s, 60°C 30 s. The primers were provided by Shanghai Sangon Biological Engineering Co. Ltd. (Shanghai, China). The primers used in the present study were as following: CD163: forward 5′-GGA TCT GCT GAC TTC AGA AG -3′, reverse 5′-CTC CTT GTC TGT TCC TCC AA-3'; (antisense); S100A8: Forward 5′-ATG CCG TCT ACA GGG ATG ACC T-3′, reverse 5′- AGA ATG AGG AAC TCC TGG AAG TTA-3'; FGF14: forward 5′-TAT TGC AGG CAA GGC TAC TAC TTG-3′, reverse 5′-GTT TTC ACT CCC TGG ATG GCA AC-3'; CDH19: forward 5′-ATT GGT CAG CCA GGA GCG TTG T-3′, reverse 5′- GCA GAT TCA GAG ACA GTC AAG CG-3'. lncRNA ENST00000430222 forward 5′- TCT CAA GTC GCT GAC ACC TCC TC-3′, reverse 5′- GGG TTG CCG AGT GAA GCT AAG AC-3'; lncRNA ENST0000062895 forward 5′-GCA AGG CGT CCG AAG TAT GAG TC-3′, reverse 5′-CGT CAG TAG AAG TTA GGC GAT CAG C -3'; lncRNA ENST00000631338 forward 5′-AGT TCA TCA CGG CTG CTG CTA AC-3′, reverse 5′- CTT GGC TTG GAG GGA GAA GAA TCA C-3'; lncRNA MSTRG18183 forward 5′-CCA GAG AGG AGG AAG AGG GGA ATC-3′, reverse 5′-TTA GGT GGG TGG AAG GCA GAG ATC-3'; hsacirc_013041 forward 5′-TGG TGT ATG CAA GTG GCC-3′, reverse 5′-TGC TGA AAA GCC AAC TGC TGG GTA G-3'; hsacirc_025902 forward 5′-AGA CCG TGG TGG TCA TCC-3′, reverse 5′-CCT GAG CCT TGA GAT AGT T-3'; hsacirc_054182 forward 5′-CAG AGC CAG CAT TCT TTC C-3′, reverse 5′-GAG CCT GTG GAT GAA GTG AG-3'; hsacirc_037511 forward 5′- CCC TAA AGA AAA TTG CTA-3′, reverse 5′- TTA TCA CAA ATC TCA GCC -3'. GAPDH forward: 5′-CTC TGC TCC TCC TGT TCG AC-3′, reverse: 5′-GCG CCC AAT ACG ACC AAA TC-3'. All samples were run in triplicate, and the results were analyzed using the $2^{(-\Delta\Delta Ct)}$ Method.

The lncRNA-Targeted Genes and circRNA-Originated Genes Identification

As reported previously, the *cis*-acting regulation is an important mechanism of lncRNA. Through this, the lncRNAs activate, repress, or modulate the expression of neighboring target genes (Lam et al., 2014). These lncRNAs are treated as *cis*-acting lncRNAs. Usually, the *cis*-acting lncRNAs regulate the expression of their neighboring genes in a manner dependent on the location of their own sites of transcription. Therefore, the protein-coding genes neighboring the *cis*-acting lncRNAs might be their targets. Here, we searched the protein-coding genes within a distance of 1 kb ~ −1 kb to the potential *cis*-acting lncRNAs. These protein-coding genes were considered to be the lncRNA-target genes of the corresponding lncRNA.

To predict the functions or mechanisms of circRNAs, we identified the circRNA-originated genes based on the theory that many circRNAs are originated from protein-coding genes and contain exonic sequences (Guo et al., 2014; You et al., 2015). Therefore, we analyzed the sequence of the differentially expressed circRNAs, identified the circRNAs with exons of protein-coding genes. These protein-coding genes were considered to be the originate genes of the corresponding circRNAs.

Proteomic Analysis of Differentially Expressed Proteins in Stable and Unstable Plaques

Extraction and digestion of proteins: The samples were lysed by SDT solution (4% SDS, 100 mM Tris/HCl pH 7.6, 0.1 M DTT). The protein concentration was determined by BCA method. The extracted total proteins were hydrolyzed by trypsin using filter aided proteome preparation (FASP). The desalination was conducted on the C_{18} cartridge. After freeze-drying, the peptide was dissolved in 40 μl Dissolution buffer and quantified by OD280.

LC-MS/MS detection: Trypsin-digested peptides were separated using the Easy nLC nanoHPLC system. 2 μg of the sample were loaded with a constant flow of 4 μl/min on a Thermo Scientific EASY column C_{18} column. After trap enrichment, the peptides were eluted in the Easy C_{18} nanocolumn (75 μm*10 cm) by a linear gradient of solvent A (0.1% formic acid solution) and solvent B [0.1% formic acid in acetonitrile (84%)] with a constant flow of 250 nL/min. The solvent B changed from 0 to 35% in 0–50 min, from 35 to 100% in 50–58 min, and 100% in 58–60 min.

The HPLC system was coupled to an OrbiTrap QExactive mass spectrometer (Thermo Fisher Scientific Inc.) via an EasySpray source. The full scan MS survey spectra was m/z 300 to 1800 in positive mode. The first-grade mass spectrometry resolution was 70,000 at m/z 200. The automatic gain control (AGC) target was 3e6. The maximum IT was 10 ms. The dynamic exclusion was 40.0 s. The charge to mass ratio of peptides was collected under the following conditions: after a full scan, 10 MS2 scan was obtained; The MS2 activation type was HCD, with an isolation window of 2 m/z; The resolution of MS2 was 17,500 at 200 m/z; The normalized collision energy was 30 eV, and the underfill ratio was 0.1%.

Annotation and quantification of proteins: The raw file was annotated by Maxquant software (version 1.5.5.1). The parameters used for the analysis was as following: main search ppm was 6; max missed cleavages was 2; De-isotopic was True; enzyme was trypsin; fixed modifications was carbamidomethyl; variable modifications were oxidation, and the database was uniport_Homo_sapiens_186616_20191202; the decoy database pattern was Reverse; the label-free quantification (LFQ) was True; the peptide mass tolerance was ±20 ppm; the peptide FDR was≤0.01, and the protein FDR was≤0.01.

The differentially expressed proteins (DEPs) between stable and unstable plaques were identified by the criteria of $|\log2FoldChange| > 1.0$ and p-value < 0.05.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes pathway analysis of Differentially Expressed mRNAs, lncRNA-targeted genes, circRNA-originated genes, and Differentially Expressed Proteins.

GO enrichment analysis and KEGG pathway analysis were used to analyze the functions of DE mRNAs, lncRNA-targeted genes, circRNAs-originated genes, and DEPs. The DAVID online tool (https://david.ncifcrf.gov/) was used for GO and KEGG enrichment. The GO enriched genes into three annotations: biological process (BP), cell components (CC), and molecular function (MF). A p-value < 0.05 was considered to be significantly related GO terms or KEGG pathways.

Protein-Protein Interaction Analysis of the Differentially Expressed Proteins

To identified the interactions between DEPs, the STRING online tool (website: https://string-db.org/) was used. The STRING database covers 9′643′763 proteins from 2′031 organisms. It provides direct (physical) interactions and indirect (functional) associations between proteins based on the computational prediction, knowledge transfer between organisms, and interactions aggregated from other (primary) databases. After the PPI analysis by STRING, the key clusters in the PPI network were analyzed by the tool of MCODE in Cytoscape software 3.8.3.

Comparison Analysis of Transcriptome and Proteome, and Previously Reported Data Profile GSE41571.

A previous study reported a low correlation between transcripts and proteins (Ghazalpour et al., 2011). To analyze the consistency of expression pattern between transcriptome and proteome in the present study, and to identify the key genes which may play critical roles in the stability of atherosclerotic plaques, we analyzed the overlapped genes between DE mRNAs and DEPs. We also analyzed the relationship between lncRNA-targeted genes (identified in *The lncRNA-Targeted Genes and circRNA-Originated Genes Identification*) and the DEPs, the circRNA-originated genes (identified in *The lncRNA-Targeted Genes and circRNA-Originated Genes Identification*) and DEPs. The overlapped genes and the expression levels were summarized.

To verify our transcriptomic and proteomic results, we also downloaded the gene expression profile GSE41571 from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/). The dataset GSE41571 contains 5 unstable and 6 stable atherosclerotic plaques obtained from CEA. These plaques were carried out genome-wide gene expression profiling using microarrays. The GEO2R tool on the GEO website was used to screen out the differentially expressed genes in GSE41571. Then we analyzed the overlapped genes between DE mRNAs from GSE41571 and our transcriptomic and proteomic profiles.

LncRNA/circRNA-miRNA-mRNA Network Analysis

The lncRNAs and cirRNAs are involved in the gene regulation by many methods, such as binding to target genes, affecting the histone modification, activating transcriptional factors, and binding to miRNA as competitive endogenous RNA (ceRNA). To further analyze the functions of specific lncRNAs or circRNA, the miRNAs which interacted with the lncRNA/circRNA were analyzed by miRDB (http://mirdb.org/). Then the miRNA targeted mRNA was analyzed by miRDB. Cytoscape software

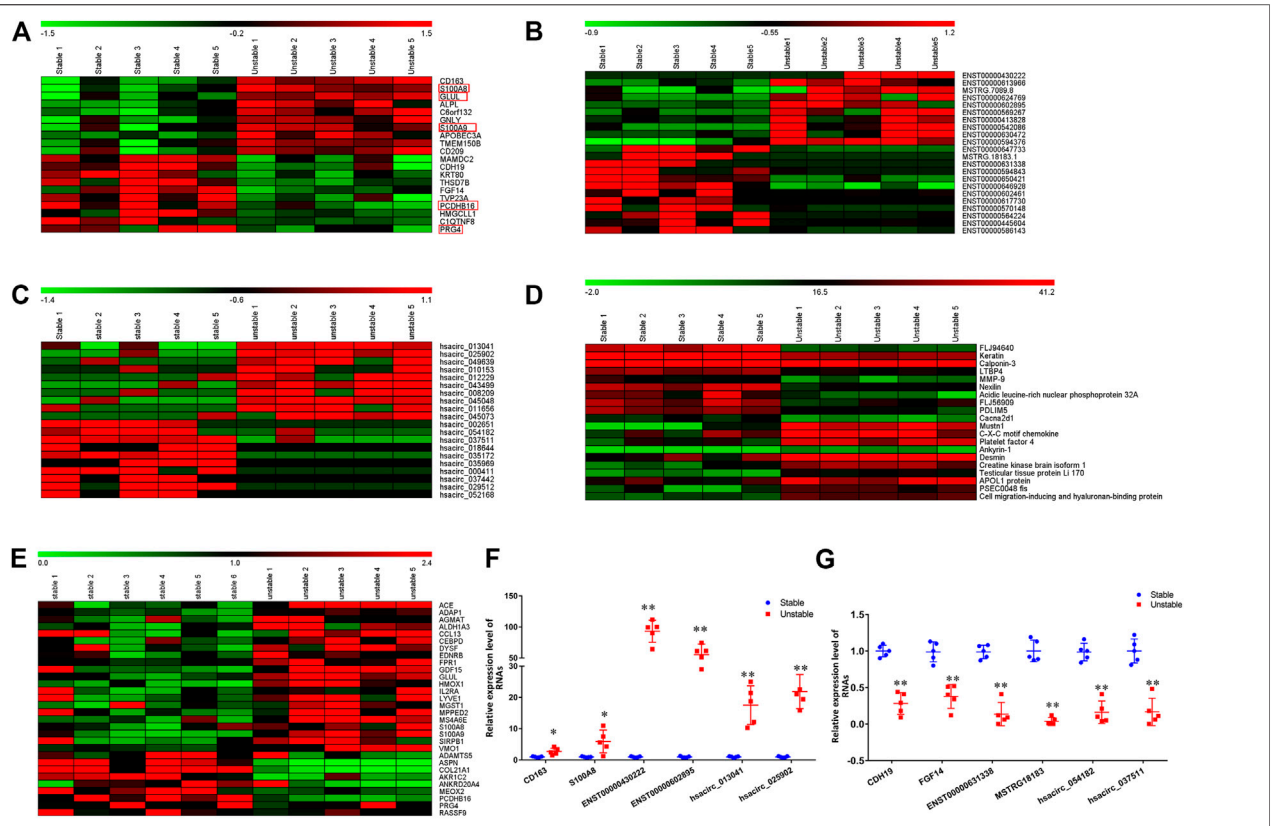


FIGURE 1 | The hierarchical clustering heatmap of 20 DE mRNAs, lncRNAs, circRNAs, and proteins. A–D: The hierarchical clustering heatmap of mRNAs **(A)**, lncRNAs **(B)**, circRNA **(C)**, proteins **(D)**. **(E)**: The heatmap of overlapping genes between current study and GSE41571; **(F)**: the qPCR verification of 6 randomly selected upregulated mRNAs, lncRNAs, and circRNAs. **(G)**: the qPCR verification of 6 randomly selected downregulated mRNAs, lncRNAs, and circRNAs. The values (mean ± S.D. from 5 independent experiments) are relative to Stable group, which was set as 1. $^{*}p < 0.05$, $^{**}p < 0.01$. Red box: the overlapping genes between current study (A) and GSE41571 (E).

3.8.3 was used to visualize the lncRNA/circRNA-miRNA-mRNA network.

Statistical Analysis

The data were presented in the form of mean ± S.D. p-value < 0.05 is considered statistically significant.

RESULTS

Patients and Samples Information

A total of 5 patients with stable plaques and 5 patients with unstable plaques were involved in the present study. The characteristic information of these patients were shown in **Supplementary Table S1**. There was no significant difference between the two groups on age, sex, body weight index, smoke, alcohol, and lipid profiles. In the unstable plaque group, 40% of the patients are taking antihypertensives, antihyperlipidemic drugs, or antiplatelet drugs. In the stable plaque group, 40% of the patients are taking antihypertensives.

The DE mRNAs, lncRNAs, circRNAs, and Proteins in Stable and Unstable Plaques.

In the RNA-Seq analysis, more than 168 466 008 clean reads were identified with more than 96.57% been mapped to the reference genome. Totally, 20025 mRNAs, 31751 lncRNAs, and 12131 circRNAs were identified. Among these genes, 202 mRNAs, 488 lncRNAs, and 91 circRNAs were differentially expressed. In the 202 DE mRNAs, 125 were upregulated and 77 were downregulated in unstable plaques. The heatmap of 20 DE mRNAs were shown in **Figure 1A**. In unstable atherosclerotic plaques, 207 upregulated and 281 downregulated lncRNAs, 61 upregulated and 30 downregulated circRNAs were also been identified. The heatmaps of part of these differentially expressed lncRNAs and circRNAs were shown in **Figures 1B,C**. To verify the RNA-Seq results, the qPCR analysis was conducted for several DE mRNAs, lncRNAs, and circRNAs. The qPCR results were consistent with that of the RNA-Seq results (**Figures 1F,G**).

In the HPLC-MS/MS analysis, a total of 3,082 proteins in 23494 peptides were identified. Among these proteins, 148 were upregulated

and 145 were downregulated in unstable plaques. The hierarchical clustering heatmap of 20 differentially expressed proteins were shown in **Figure 1D**.

Comparison Analysis of the Present Transcriptome and Previously Reported Data from GSE41571.

To further verify our results and to figure out the key genes in the stability of atherosclerotic plaques, we analyzed the dataset GSE41571 downloaded from the GEO database. 1,595 downregulated and 750 upregulated genes were found in GSE41571. Among them, 30 genes were differentially expressed in both dataset GSE41571 and our RNA-Seq data, 42 genes were overlapped with our DEPs (**Supplementary Table S2**, **Figure 1E**). These overlapped genes were mainly related to cell adhesion, immune response, and inflammatory responses.

The lncRNA-Targeted mRNA and circRNA-Originated Genes Identification

We screened the neighboring protein-coding genes of the 488 differentially expressed lncRNAs. 381 protein-coding genes were at the distance of 1 kb ~ −1 kb to the corresponding 488 lncRNAs (**Supplementary Table S3**). These protein-coding genes were treated as the potential lncRNA-targeted genes. The corresponding lncRNAs were treated as *cis*-acting lncRNAs. The expression of these targeted genes might be regulated by the *cis*-acting lncRNAs.

We also analyzed the sequence of 91 differentially expressed circRNAs. All of these 91 circRNAs contained exonic sequences. These exonic sequences belong to 97 protein-coding genes. These 97 genes were considered to be the circRNA-originated genes (**Supplementary Table S4**).

The GO enrichment and KEGG pathway analysis of the DE mRNAs, lncRNA-targeted genes, circRNA-originated genes, and DEPs.

The GO and KEGG analysis was conducted to predict the functions of differentially expressed mRNA, the lncRNA-targeted genes, circRNA-originated genes, and the DEPs. The results were shown in **Supplementary Table S5**. The DE mRNAs were mainly enriched in the GO terms of extracellular region, RAGE receptor binding, defense response, and KEGG terms of neuron ligand-receptor interaction, and cytokine-receptor interaction. The lncRNA-targeted genes were enriched in the GO terms of intracellular membrane-bounded organelle, transcription regulator activity, transcription from RNA polymerase II promoter, and KEGG terms of the TNF signaling pathway. The circRNA-originated genes were enriched in the GO terms of basal cortex, malonyl-CoA decarboxylase activity, negative regulation of stress fiber assembly, and KEGG terms of cellular senescence and focal adhesion. The proteins were mainly related to ECM-receptor interaction, hematopoietic cell lineage, and phagosome.

Protein-Protein Interaction Network and Clusters Analysis of Differentially Expressed Proteins

The protein-protein interaction network between the DEPs was analyzed by the online tool STRING and shown in **Figure 2**. In total, 195 nodes and 532 edges were identified. Four clusters with 24 genes were identified by the MCODE tool in the Cytoscape software (**Figure 2** and **Supplementary Table S6**). The DEPs were clustered into 4 clusters which may be involved in the functions of smooth muscle contraction, metabolism and transportation of lipoproteins, immune system function, and mRNA splicing.

Comparison Analysis of Transcriptome and Proteome

To analyze the consistency of transcripts and proteins, and to identify the key genes playing critical roles in the stability of atherosclerotic plaques, we compared the expression of differentially expressed mRNA and DEPs, the lncRNA-target genes and DEPs, as well as the circRNA-originated genes and DEPs. Surprisingly, only two DEPs (CD5L, S100A12) were overlapped with mRNA. Two proteins (CKB, CEMIP) were overlapped with the lncRNA-targeted genes, one protein (SH3GLB1) was overlapped with the cirRNA-originated gene. The expression levels of these mRNAs, related lncRNAs, related circRNA, and DEPs were shown in **Supplementary Table S7**. Both CD5L and S100A12 mRNAs and proteins were upregulated in the unstable plaques. The CKB and CEMIP proteins, as well as their related lncRNA, MSTRG.11455.17 and MSTRG.12845 were upregulated in unstable plaques. While the SH3GLB1 protein was upregulated, but its related circRNA, hsacirc_ooo411 was downregulated. These genes may play critical roles in the stability of atherosclerotic plaques.

LncRNA (circRNA)-miRNA-mRNA Network Analysis

In *LncRNA (circRNA)-miRNA-mRNA Network Analysis*, two lncRNAs (MSTRG.11455.17, MSTRG.12845) and one circRNA (hsacirc_000411) were identified to interact with DEPs. To further explore the functions of MSTRG.11455.17, MSTRG.12845, and hascirc_000411, we analyzed the related miRNA, and the subsequent miRNA targeted mRNAs. The lncRNA (circRNA)-miRNA-mRNA network was shown in **Figure 3**. The lncRNA MSTRG.11455.17 was predicted to bind miR-7849, miR-7856, and miR-4760, which may affect the functions of subsequent 33 genes. LncRNA MSTRG.12845 was predicted to bind miR-4797, miR-3915, miR-5009, miR-6873, and miR-6817, which may affect subsequent 26 genes. Hsacirc-000411 was predicted to bind miR-647 and miR-4433b, which may affect subsequent 7 genes (**Figure 3**).

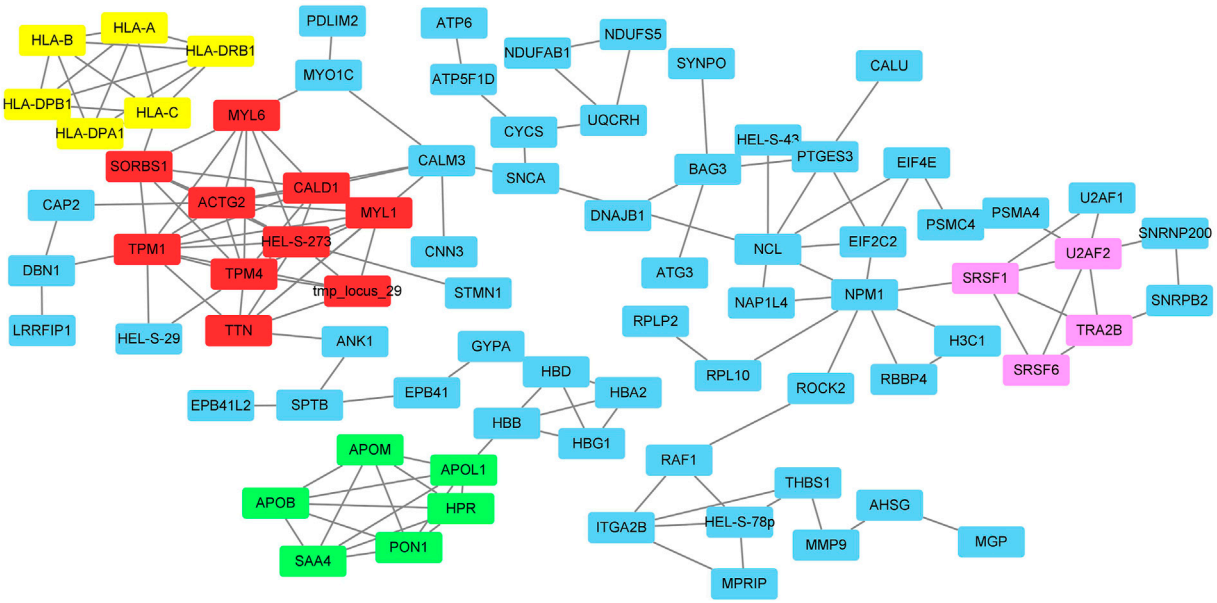


FIGURE 2 | The protein-protein interactions of DEPs and the clusters identified by MCODE. Red box: cluster1; Green box: cluster 2; Yellow box: cluster 3; Pink box: cluster 4.

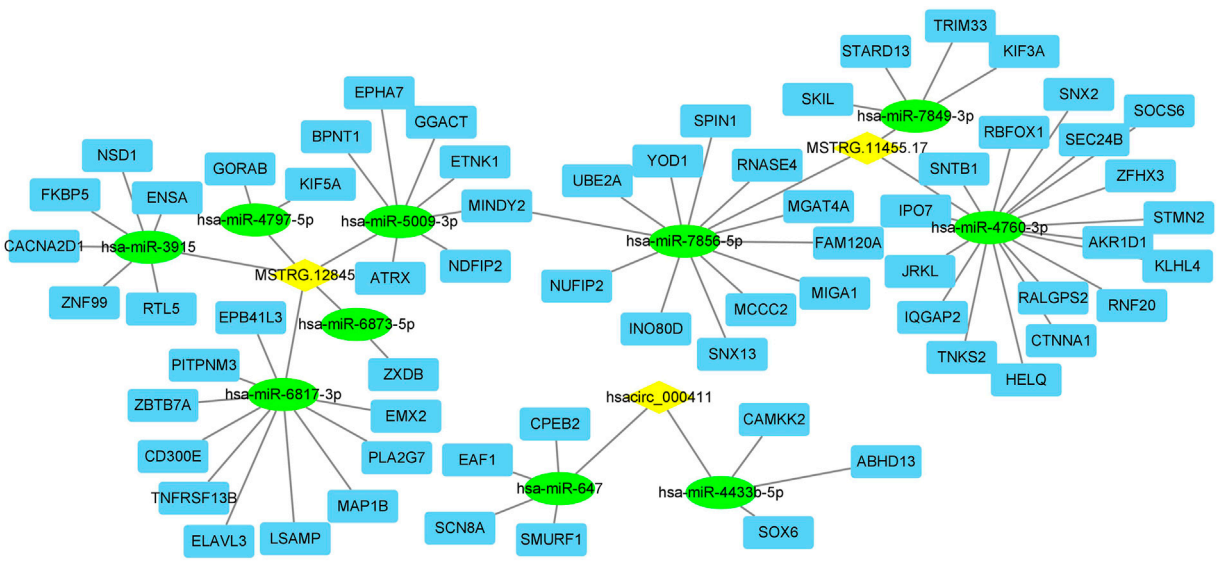


FIGURE 3 | The lncRNA (or circRNA)-miRNA-mRNA interactions network. Yellow diamond: lncRNA or circRNA; green circle: miRNA; blue square: mRNA.

DISCUSSION

The present study screened the transcripts and proteins in human stable and unstable atherosclerotic plaques. A total of 202 mRNAs, 488 lncRNAs, 91 circRNAs, and 293 DEPs were identified to be differentially expressed between stable and unstable atherosclerotic plaques. The 488 lncRNAs might affect the production of 381 protein-coding genes by *cis*-acting regulations. Sequence analysis indicated the 91 circRNAs were originated from 97 protein-coding genes.

To figure out the functions of these DE mRNAs, lncRNAs, circRNAs, and proteins, we conducted the GO enrichment and KEGG pathway analysis, as well as the PPI network. The enrichment and pathway analysis indicated these differentially expressed RNAs and proteins were involved in the cellular response to stress or stimulus, regulated the gene transcription

through ways such as histone acetyltransferase binding, spliceosome, and dihydropteridine reductase activity. In the atherosclerotic plaque stability, the following functions may play critical roles: the immune response (RAGE receptor binding, cytokine-cytokine receptor interactions, ECM-receptor interaction, phagosome, B cell receptor signaling pathway, cGMP-PKG signaling pathway, antigen processing and presentation), nervous system functions (neuroactive ligand-receptor interactions, cholinergic synapse, neurotrophin signaling pathway), hematologic activities (hematopoietic cell lineage, coagulation cascades), and endocrine system (cortisol synthesis and secretion, insulin secretion). The PPI and MCODE analysis of DEPs discovered 4 clusters relating to the function of smooth muscle contraction, insulin function, lipid metabolism, immune system, and gene expression (**Supplementary Table S6**). These results indicated that the immune response, endocrine system, metabolisms are major functional alterations between stable and unstable atherosclerotic plaques.

Since the previous microarray dataset GSE41571 analyzed the gene expression profiles in the macrophage-rich regions of stable and unstable atherosclerotic plaques, we compared our data with the GSE41571 data. 30 genes and 42 proteins were found to be differentially expressed in both our data and GSE41571. These 72 genes and proteins are mainly related to cell adhesion, immune response, and inflammatory responses, which were consistent with our GO and KEGG analysis.

To further screen out the key genes which may play critical roles in the stability of atherosclerosis, we analyzed the consistency of transcriptome and proteome in the present study. Surprisingly, only a few genes were screened out. They are CD5L, S100A12, CKB (target gene of lncRNA MSTRG.11455.17), CEMIP (target gene of lncRNA MSTRG.12845), and SH3GLB1 (originated gene of hsacirc_000411). CD5L and S100A12 were upregulated in unstable plaques at both mRNA and protein levels. CD5L encodes the secreted glycoprotein antigen protein CD5, which is involved in the inflammatory response. It is primarily expressed in macrophages and promotes M2 macrophage polarization, promotes anti-inflammation in response to TLR activation (Sanjurjo et al., 2015, 2018). S100A12 is a member of the S100 protein family. It binds to RAGE and activates the downstream pro-inflammatory signals, such as NF-kB and ROS (Xiao et al., 2020). S100A12 is involved in the pathogenesis of atherosclerosis through the S100A12-CD36 axis (Farokhzadian et al., 2019). CKB and CEMIP are the targets of two novel lncRNAs MSTRG.11455.17 and MSTRG.12845, respectively. CKB encodes protein creatine kinase B, which plays a role in energetic hemostasis in ischemic and inflammatory disorders (Kitzenberg et al., 2016). CEMIP encodes the cell migration-inducing and hyaluronan-binding protein, which regulates epithelial-mesenchymal transition (EMT), tumor cell growth and migration (Li et al., 2017). In atherosclerosis, CEMIP was reported to regulate the proliferation and migration of vascular smooth muscle cells (Xue et al., 2020). SH3GLB1 gene encodes the endophilin-B1 or Bif-1 protein, which is implicated in the apoptotic and autophagic pathways (Takahashi et al., 2013). However, its effects on atherosclerosis are still unknown. Therefore, the above five genes, and their correlated lncRNAs (MSTRG.11455.17, MSTRG.12845) and circRNA (circ_000411) may play critical roles in the stability of atherosclerotic plaques through inflammation, cell growth or migration.

Competitively endogenous RNA (ceRNA) is another important mechanism of the functions of non-coding RNAs. ceRNAs regulate other RNA transcripts by competing for shared microRNAs (miRNAs) (Salmena et al., 2011). Based on the sequence of MSTRG.11455.17, MSTRG.12845, and circ_000411, we found 10 miRNAs that may bind to them, indicating their ceRNA potential.

In our present study, a whole atherosclerotic plaque was collected through method of CEA, and was used for detection. Atherosclerotic plaques are comprised of many different kinds of cells, such as foam cells, macrophages, smooth muscle cells. Different cell types are with different gene expression patterns. This is a major limitation in our present analysis, which might decrease the efficacy of screening out more differentially expressed genes. Single-cell RNA-Seq is needed in future studies to distinguish the gene expression pattern in different cells types between stable and unstable plaques.

In summary, our study screened the transcription and protein profiles in human stable and unstable atherosclerotic plaques by RNA-Seq and LC-MS/MS, analyzed the functions and pathways of differentially expressed RNAs and proteins, identified a few key genes and noncoding RNAs. The results may provide new knowledge on understanding the stability of atherosclerotic plaques.

DATA AVAILABILITY STATEMENT

The transcriptome datasets PRJNA752896 for this study can be found in the Sequence Read Archive (SRA) of National Library of Medicine [https://submit.ncbi.nlm.nih.gov/subs/bioproject/]. The proteome datasets were submitted to the iProX (https://www.iprox.cn/), with the Project ID of IPX0003457000.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the First Hospital of Jilin University (No. 2019-272, Changchun, Jilin). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

M-hB and RL planned and designed the experiments; M-hB, R-qZ, and X-sH performed experiments; JZ and ZG analyzed

data; B-fX and M-hB wrote the paper. All authors discussed the article and gave comments.

FUNDING

The present study was supported by the Hunan Key Laboratory Cultivation Base of the Research and Development of Novel Pharmaceutical Preparations (No. 2016TP1029); The Application Characteristic Discipline of Hunan Province; The Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105); The Hunan Provincial Key Laboratory of Fundamental and Clinical Research on Functional Nucleic Acid; The National Science Foundation of China (81900739); Hunan Provincial Natural Science Foundation (No. 2019JJ40330, 2018JJ3569); Foundation of Hunan educational Committee (19A055, 20C0142, 18B538); The International Cooperation Project of Jilin Provincial Science and Technology Department (Number 20190701047 GH).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.755507/full#supplementary-material

REFERENCES

Bai, H.-L., Lu, Z.-F., Zhao, J.-J., Ma, X., Li, X.-H., Xu, H., et al. (2019). Microarray Profiling Analysis and Validation of Novel Long Noncoding RNAs and mRNAs as Potential Biomarkers and Their Functions in Atherosclerosis. *Physiol. Genomics* 51, 644–656. doi:10.1152/physiolgenomics.00077.2019

Cao, Q., Guo, Z., Du, S., Ling, H., and Song, C. (2020). Circular RNAs in the Pathogenesis of Atherosclerosis. *Life Sci.* 255, 117837. doi:10.1016/j.lfs.2020.117837

Çekin, N., Özcan, A., Göksel, S., Arslan, S., Pınarbaşı, E., and Berkan, Ö. (2018). Decreased FENDRR and LincRNA-P21 Expression in Atherosclerotic Plaque. *Anatol J. Cardiol.* 19, 131–136. doi:10.14744/AnatolJCardiol.2017.8081

Farokhzadian, J., Mangolian Shahrbabaki, P., and Bagheri, V. (2019). S100A12-CD36 axis: A Novel Player in the Pathogenesis of Atherosclerosis. *Cytokine* 122, 154104. doi:10.1016/j.cyto.2017.07.010

Fasolo, F., Di Gregoli, K., Maegdefessel, L., and Johnson, J. L. (2019). Non-coding RNAs in Cardiovascular Cell Biology and Atherosclerosis. *Cardiovasc. Res.* 115, 1732–1756. doi:10.1093/cvr/cvz203

Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., et al. (2011). Comparative Analysis of Proteome and Transcriptome Variation in Mouse. *Plos Genet.* 7, e1001393. doi:10.1371/journal.pgen.1001393

Guo, J. U., Agarwal, V., Guo, H., and Bartel, D. P. (2014). Expanded Identification and Characterization of Mammalian Circular RNAs. *Genome Biol.* 15, 409. doi:10.1186/s13059-014-0409-z

Hung, J., Scanlon, J. P., Mahmoud, A. D., Rodor, J., Ballantyne, M., Fontaine, M. A. C., et al. (2020). Novel Plaque Enriched Long Noncoding RNA in Atherosclerotic Macrophage Regulation (PELATON). *Atvb* 40, 697–713. doi:10.1161/ATVBAHA.119.313430

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular RNAs Are Abundant, Conserved, and Associated with ALU Repeats. *RNA* 19, 141–157. doi:10.1261/rna.035667.112

Kitzenberg, D., Colgan, S. P., and Glover, L. E. (2016). Creatine Kinase in Ischemic and Inflammatory Disorders. *Clin. Translational Med.* 5, 31. doi:10.1186/s40169-016-0114-5

Lam, M. T., Li, W., Rosenfeld, M. G., and Glass, C. K. (2014). Enhancer RNAs and Regulated Transcriptional Programs. *Trends Biochem. Sci.* 39, 170–182. doi:10.1016/j.tibs.2014.02.007

Li, L., Yan, L. H., Manoj, S., Li, Y., and Lu, L. (2017). Central Role of CEMIP in Tumorigenesis and its Potential as Therapeutic Target. *J. Cancer* 8, 2238–2246. doi:10.7150/jca.19295

Ou, M., Li, X., Zhao, S., Cui, S., and Tu, J. (2020). Long Non-coding RNA CDKN2B-AS1 Contributes to Atherosclerotic Plaque Formation by Forming RNA-DNA Triplex in the CDKN2B Promoter. *EBioMedicine* 55, 102694. doi:10.1016/j.ebiom.2020.102694

Pan, R. Y., Zhao, C. H., Yuan, J. X., Zhang, Y. J., Jin, J. L., Gu, M. F., et al. (2019). Circular RNA Profile in Coronary Artery Disease. *Am. J. Transl Res.* 11, 7115–7125.

Poller, W., Dimmeler, S., Heymans, S., Zeller, T., Haas, J., Karakas, M., et al. (2018). Non-coding RNAs in Cardiovascular Diseases: Diagnostic and Therapeutic Perspectives. *Eur. Heart J.* 39, 2704–2716. doi:10.1093/eurheartj/ehx165

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

Sanjurjo, L., Aran, G., Roher, N., Valledor, A. F., and Sarrias, M. R. (2015). AIM/CD5L: a Key Protein in the Control of Immune Homeostasis and Inflammatory Disease. *J. Leukoc. Biol.* 98, 173–184. doi:10.1189/jlb.3RU0215-074R

Sanjurjo, L., Aran, G., Téllez, É., Amézaga, N., Armengol, C., López, D., et al. (2018). CD5L Promotes M2 Macrophage Polarization through Autophagy-Mediated Upregulation of ID3. *Front. Immunol.* 9, 480. doi:10.3389/fimmu.2018.00480

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

Takahashi, Y., Young, M. M., Serfass, J. M., Hori, T., and Wang, H. G. (2013). Sh3glb1/Bif-1 and Mitophagy: Acquisition of Apoptosis Resistance during Myc-Driven Lymphomagenesis. *Autophagy* 9, 1107–1109. doi:10.4161/auto.24817

Wang, L., Xiao, Y., and Feng, X. (2019). IIRWR: Internal Inclined Random Walk with Restart for LncRNA-Disease Association Prediction. *IEEE Access* 7, 54034–54041. doi:10.1109/access.2019.2912945

Wei, M. Y., Lv, R. R., and Teng, Z. (2020). Circular RNA circHIPK3 as a Novel circRNA Regulator of Autophagy and Endothelial Cell Dysfunction in Atherosclerosis. *Eur. Rev. Med. Pharmacol. Sci.* 24, 12849–12858. doi:10.26355/eurrev_202012_24187

Xiao, X., Yang, C., Qu, S. L., Shao, Y. D., Zhou, C. Y., Chao, R., et al. (2020). S100 Proteins in Atherosclerosis. *Clin. Chim. Acta* 502, 293–304. doi:10.1016/j.cca.2019.11.019

Xiao, X., Zhu, W., Liao, B., Xu, J., Gu, C., Ji, B., et al. (2018). BPLLDA: Predicting lncRNA-Disease Associations Based on Simple Paths with Limited Lengths in a Heterogeneous Network. *Front. Genet.* 9, 411. doi:10.3389/fgene.2018.00411

Xu, B. F., Liu, R., Huang, C. X., He, B. S., Li, G. Y., Sun, H. S., et al. (2020). Identification of Key Genes in Ruptured Atherosclerotic Plaques by Weighted Gene Correlation Network Analysis. *Sci. Rep.* 10, 10847. doi:10.1038/s41598-020-67114-2

Xue, Q., Wang, X., Deng, X., Huang, Y., and Tian, W. (2020). CEMIP Regulates the Proliferation and Migration of Vascular Smooth Muscle Cells in Atherosclerosis through the WNT-Beta-Catenin Signaling Pathway. *Biochem. Cel Biol.* 98, 249–257. doi:10.1139/bcb-2019-0249

Yang, J., Huang, T., Song, W. M., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016). Discover the Network Underlying the Connections between Aging and Age-Related Diseases. *Sci. Rep.* 6, 32566. doi:10.1038/srep32566

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human Geroprotector Discovery by Targeting the Converging Subnetworks of Aging and Age-Related Diseases. *Geroscience* 42, 353–372. doi:10.1007/s11357-019-00106-x

You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., et al. (2015). Neural Circular RNAs Are Derived from Synaptic Genes and Regulated by Development and Plasticity. *Nat. Neurosci.* 18, 603–610. doi:10.1038/nn.3975

Zhang, L. L. (2020). CircRNA-PTPRA Promoted the Progression of Atherosclerosis through Sponging with miR-636 and Upregulating the Transcription Factor SP1. *Eur. Rev. Med. Pharmacol. Sci.* 24, 12437–12449. doi:10.26355/eurrev_202012_24039

Zhang, S., Song, G., Yuan, J., Qiao, S., Xu, S., Si, Z., et al. (2020). Circular RNA Circ_0003204 Inhibits Proliferation, Migration and Tube Formation of Endothelial Cell in Atherosclerosis via miR-370-3p/TGFβR2/phosph-SMAD3 axis. *J. Biomed. Sci.* 27, 11. doi:10.1186/s12929-019-0595-9

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

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

frontiers in Genetics

METHODS
published: 18 November 2021
doi: 10.3389/fgene.2021.773882

PseUdeep: RNA Pseudouridine Site Identification with Deep Learning Algorithm

*Jujuan Zhuang[1], Danyang Liu[1], Meng Lin[1], Wenjing Qiu[2,3], Jinyang Liu[3] and Size Chen[4,5,6]**

[1]College of Science, Dalian Maritime University, Dalian, China, [2]Electrical and Information Engineering, Anhui University of Technology, Anhui, China, [3]Geneis (Beijing) Co., Ltd., Beijing, China, [4]Department of Oncology, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [5]Guangdong Provincial Engineering Research Center for Esophageal Cancer Precise Therapy, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [6]Central Laboratory, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China

Background: Pseudouridine (Ψ) is a common ribonucleotide modification that plays a significant role in many biological processes. The identification of Ψ modification sites is of great significance for disease mechanism and biological processes research in which machine learning algorithms are desirable as the lab exploratory techniques are expensive and time-consuming.

Results: In this work, we propose a deep learning framework, called PseUdeep, to identify Ψ sites of three species: *H. sapiens*, *S. cerevisiae*, and *M. musculus*. In this method, three encoding methods are used to extract the features of RNA sequences, that is, one-hot encoding, K-tuple nucleotide frequency pattern, and position-specific nucleotide composition. The three feature matrices are convoluted twice and fed into the capsule neural network and bidirectional gated recurrent unit network with a self-attention mechanism for classification.

Conclusion: Compared with other state-of-the-art methods, our model gets the highest accuracy of the prediction on the independent testing data set S-200; the accuracy improves 12.38%, and on the independent testing data set H-200, the accuracy improves 0.68%. Moreover, the dimensions of the features we derive from the RNA sequences are only 109,109, and 119 in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, which is much smaller than those used in the traditional algorithms. On evaluation via tenfold cross-validation and two independent testing data sets, PseUdeep outperforms the best traditional machine learning model available. PseUdeep source code and data sets are available at https://github.com/dan111262/PseUdeep.

Keywords: RNA modification, pseudouridine site prediction, feature extraction, deep learning, capsule network

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Xiangzheng Fu,
Hunan University, China
Lina Zhao,
Chinese Academy of Medical Sciences, China

****Correspondence:***
Size Chen
chensize@gdpu.edu.cn

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

Received: *10 September 2021*
Accepted: *04 October 2021*
Published: *18 November 2021*

Citation:
Zhuang J, Liu D, Lin M, Qiu W, Liu J and Chen S (2021) PseUdeep: RNA Pseudouridine Site Identification with Deep Learning Algorithm. Front. Genet. 12:773882. doi: 10.3389/fgene.2021.773882

INTRODUCTION

Pseudouridine (Ψ) is one of the most prevalent RNA modifications that occurs at the uridinebase through an isomerization reaction catalyzed by pseudouridine synthases (see **Figure 1**) (Bousquet-Antonelli et al., 1997; Chan and Huang, 2009; Ge and Yu, 2013; Kiss et al., 2010; Wolin, 2016; Yu and Meier, 2014). It is confirmed that Ψ modification occurs in several kinds of RNAs, such as small nuclear RNA, rRNA, tRNA, mRNA, and small nucleolar RNA (Ge and Yu, 2013). Ψ plays a

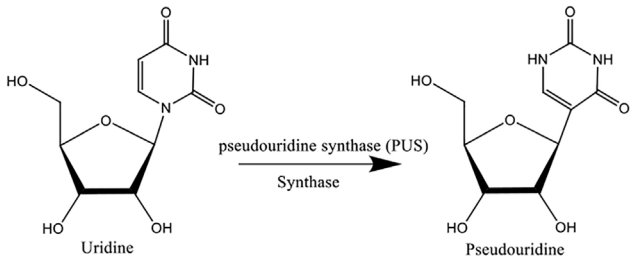


FIGURE 1 | Illustration of Ψ modification. The Ψ synthase catalyzes the uridine isomer Ψ by removing the uridine residue base from its sugar and then removing the uridine isomer, rotating it 180° along the N3–C6 axis, and finally turning the base the 5-carbon and 1′-carbons of the sugar.

significant role in many biological processes, including regulating the stability of RNA structure in tRNA and rRNA (Kierzek et al., 2014). Deficiency of Ψ might cause various diseases; the dysregulation of Ψ in mitochondrial tRNA is one of the etiologies of erythrocytic anemia and mitochondrial myopathy (Bykhovskaya et al., 2004). Moreover, the mutations of Ψ are also associated with several types of cancers, such as gastric and lung cancer (Mei et al., 2012; Carlile et al., 2014; Carlile et al., 2015; Shaheen et al., 2016; Penzo et al., 2017; Zhang et al., 2021), and Ψ is also applied in biochemical research and pharmaceuticals (C. Liu et al., 2020; Penzo et al., 2017; J. Yang et al., 2020). Undoubtedly, the identification of Ψ modification sites would be of great benefit for disease mechanism and biological processes research.

Although accurate Ψ sites can be identified by some lab exploratory techniques, they are expensive and time-consuming (Carlile et al., 2014). As an increasing number of genomic and proteomic samples are produced (J. Yang et al., 2020), it is necessary to develop some effective and robust computational models to detect Ψ sites in RNA sequences.

Many machine learning algorithms have been introduced as fast, low-cost, and efficient alternative methods to identify Ψ sites. In 2015, Li et al. established the first computational model named PPUS to identify PUS-specific Ψ sites in Saccharomyces cerevisiae and Homo sapiens. The method used the nucleotides around Ψ as features for training a support vector machine (SVM) (Y. H. Li et al., 2015). Similarly, in 2016, Chen et al. developed an SVM classifier named iRNA-PseU using the occurrence frequencies and the chemical properties of the nucleotides as well as pseudo k-tuple nucleotide composition (PseKNC) as features in *Mus musculus*, *S. cerevisiae*, and *H. sapiens* (Chen et al., 2016). He et al., in 2018, proposed PseUI, in which five types of features, nucleotide composition (NC), dinucleotide composition (DC), pseudo dinucleotide composition (PseDNC), position-specific nucleotide composition (PSNP), and position-specific dinucleotide propensity (PSDP), were combined and a sequential forward selection method was applied to select the optimal feature subset for training SVM to predict Ψ sites in *M. musculus, S. cerevisiae*, and *H. sapiens* (J. He et al., 2018). In 2019, Liu et al. proposed an ensemble model, XG-PseU, based on eXtreme gradient boosting (XGBoost) using six types of features, including NC, dinucleotide composition (DNC), trinucleotide composition (TNC), nucleotide chemical property (NCP), nucleotide density (ND), and one-hot encoding (Liu et al., 2020). In 2020, Bi et al. proposed an integrated model based on a majority voting strategy, called EnsemPseU, which contained five machine learning methods SVM, XGBoost, Naive Bays (NB), k-nearest neighbor (KNN), and random forest (RF) (Bi et al., 2020). In short, the above machine learning methods in *H. sapiens*, *S. cerevisiae*, and *M. musculus* have the highest accuracy rates of 65.44%, 68.15%, and 72.03%, respectively. Although the performance of the above machine learning methods is reasonable, there is still a lot of room for improvement. With the emergence of deep learning methods, many prediction methods based on deep learning have been applied to the field of RNA and protein modification predictions (Huang et al., 2018; Long et al., 2018; Mostavi et al., 2018; Zhang and Hamada, 2018). The above predictors do not consider deep learning methods, which can extract deeper features to improve prediction performance (B. He et al., 2020; Liang et al., 2020).

In this work, we propose a deep learning framework, PseUdeep, to identify Ψ sites of the three species *H. sapiens*, *S. cerevisiae*, and *M. musculus*. Compared with previous machine learning methods, our model applies three encoding methods, one-hot encoding, K-tuple nucleotide frequency pattern (KNFP) (Y. Yang et al., 2021), and PSNP (Dou et al., 2020) to extract RNA sequence features. Our model consists of a convolutional neural network (CNN), a capsule neural network, and a bidirectional gated recurrent unit (BiGRU) network with a self-attention mechanism (see **Figure 2**). Finally, we conduct a tenfold cross-validation test on the benchmark data set and an independent verification test on two independent data sets and compare the prediction results of our model with the results of the previous machine learning model; the accuracy of our model for *H. sapiens* increased by 1.55%, for *S. cerevisiae* by 4.58%, and for *M. musculus* by 0.42%.

METHODS

Benchmark Data Sets

Chen et al. (2016) established data sets for computationally identifying Ψ sites in *H. sapiens*, *M. musculus*, and *S. cerevisiae* based on RMBase (Sun et al., 2016). With the update of RMBase, we use three training new data sets base on RMBase2.0 (Chen et al., 2015), which include NH_990 (*H. sapiens*), NM_944 (*M. musculus*), and NS_627 (*S. cerevisiae*), and the data sets built by Liu K. et al. (2020). In *H. sapiens* and *S. cerevisiae*, we also use the independent data sets H_200 and S_200, which are built by Chen et al. (2016) to evaluate the performance of the method.

In the NH_990 and NM_944 data sets, the length of the sequence is 21 nt. However, in the NS_627 data set, the length is31 nt. In the H_200 and S_200 data sets, the RNA sequence length is 21 and 31 nt, respectively. **Table 1** shows the details of all data sets.

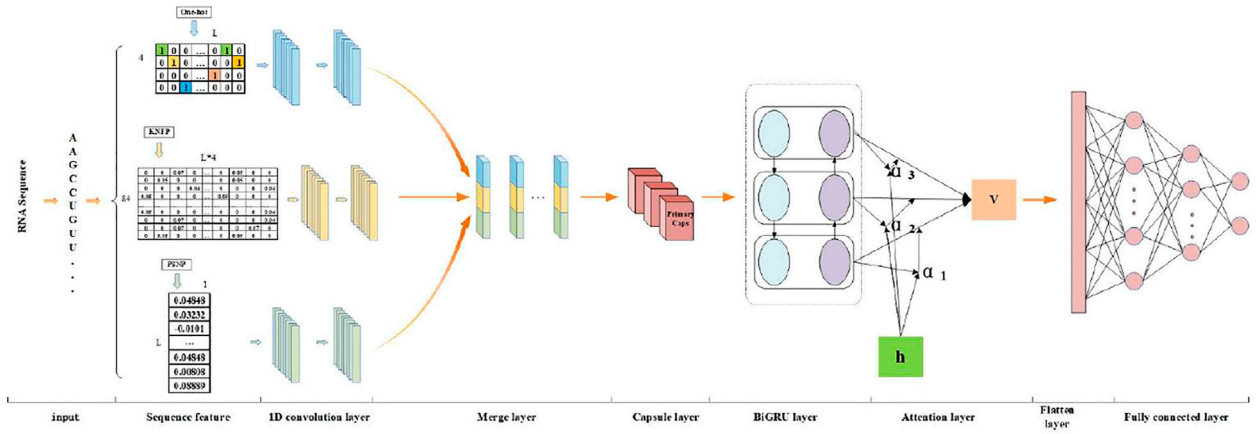


FIGURE 2 | The flowchart of PseUdeep: We use the collected RNA sequences as the input of the model and the first use three encoding methods, one-hot encoding, KNFP, and PSNP, to extract RNA sequence features. Then, the three feature matrices are convoluted twice, and the results are stitched together. Finally, it is input into the capsule neural network and the BiGRU network with a self-attention mechanism and two fully connected layers for classification.

TABLE 1 | The information on training data sets and independent testing data sets.

Species	The name of the datasets	The length of the RNA sequences (bp)	The number of positive samples	The number of negative samples
H. sapiens	NH-990 (training)	21	495	495
	H-200 (testing)	21	100	100
S. cerevisiae	NS-627 (training)	31	314	313
	S-200 (testing)	31	100	100
M. musculus	NM-944 (training)	21	472	472
	-	-	-	-

Feature Extraction

Feature extraction is the basis of the algorithm. In our work, we consider three kinds of features: one-hot encoding, KNFP (Y. Yang et al., 2021), and PSNP (Dou et al., 2020).

One-Hot Encoding

Given an RNA sequence R,

$$R_\phi = N_1 N_2 \cdots N_l, \tag{1}$$

where $N_j \in \{A, C, G, U\}\,(j = 1, 2, \cdots, l)$ represents the nucleotide at the *jth* position of the RNA segmentR. We represent each nucleotide with a four-dimensional vector, that is, nucleotide G is represented as (1, 0, 0, 0), C is (0, 1, 0, 0), Uis (0, 0, 1, 0), and A is (0, 0, 0, 1).

KNFP

The KNFP (Y. Yang et al., 2021) pattern represents the local contextual features at different levels. KNFP integrates various short-distance sequence order information and retains a large number of original sequence modes (Chen et al., 2015). We apply KNFP to extract local context features from RNA sequences. KNFP includes mononucleotide, dinucleotide, and trinucleotide composition. For an RNA sequence R_ϕ, the K-tuple nt composition can represent any RNA sequence as a 4^K dimensional vector:

$$P = \left[\varphi_1, \varphi_2, \varphi_3, \varphi_4, \ldots, \varphi_{4^K}\right]^T, \tag{2}$$

where $\phi_u\,(u = 1, 2, \cdots, 4^K)$ is the frequency of the *uth* K-tuple pattern in the RNA sequence, namely, the substring of the sequence contains K neighboring nt, and the symbol T represents the transpose operator, so it has $l - K + 1$ overlapping segments for every RNA sequence R with length l, and each segment is encoded as a one-hot vector with dimension 4^K. The frequency pattern matrix $m_K \varepsilon R^{(l-K+1)*4^K}$ is generated for each type of K-tuple nt composition. To facilitate subsequent processing, we fill the shorter part with zeros. By combining different K-tuples $M = \{m_1, m_2, m_3\}$ with $K = 1, 2, 3$, the feature of each position in the sequence is connected in one dimension of size $d = 64$.Compared with the traditional one-hot encoding, KNFP effectively compensates for the shortcomings of information insufficiency.

PSNP

PSNP (Dou et al., 2020) is an effective nucleotide encoding method, which has been successfully applied to the identification of many functional sites in biological sequences (W. He et al., 2018; W. He et al., 2018; G. Q. Li et al., 2016; Zhu et al., 2019). In this method, location-specific information can be represented by calculating the differences in nucleotide frequency

at a specific location between positive and negative RNA samples. Considering an RNA sequence $R_\phi = N_1N_2\cdots N_l$, the PSNP matrix can be written as a $4 \times l$ -dimensional vector.

First, we calculate the frequency of occurrence for four nucleotides, respectively, from bath positive and negative samples at the *j*th position. In this way, we obtain two $4 \times l$ position-specific occurrence frequency matrixes, namely, $Z^+$and Z^-, of which $Z^+$is obtained from all positive samples and $Z^-$from all negative samples. We define the location-specific nucleotide propensity matrix, represented by Z_{PSNP}, as shown below:

$$Z_{PSNP} = \left[Z_1, Z_{2,\cdots}Z_l\right] = \begin{bmatrix} Z_{1,1} & Z_{1,2} & \cdots & Z_{1,l} \\ Z_{2,1} & Z_{2,2} & \cdots & Z_{2,l} \\ Z_{3,1} & Z_{3,2} & \cdots & Z_{3,l} \\ Z_{4,1} & Z_{4,2} & \cdots & Z_{4,l} \end{bmatrix}, \quad (3)$$

where $Z_{i,j} = Z^+_{i,j} - Z^-_{i,j}$ gives the difference of frequencies of the ith nucleotide at the jth position between positive and negative samples.

Deep Learning Architecture of PseUdeep

For each input sequence, we use three feature extraction (one-hot encoding, KNFP, and PSNP) methods to form three feature matrices. For each feature matrix, a pair of 1-D CNNs are used. The first layer of each feature matrix has a filter size of 11 and a kernel size of 7. Similarly, the second 1[D CNN layer for each feature matrix has a filter size of 11 and a kernel size of 3. Two convolution layers are used to capture features from three feature matrices; all layers had a "Relu" activation function. The three convolution results are spliced together and fed into the capsule network with 14 capsules for vector convolution, and the output of the capsule network is put into the BiGRU neural network with an attention mechanism; the final feature is concatenated and fed into two dense layers to obtain the prediction results. Bayesian optimization is used to select the best performance of the hyperparameters. The adjusted parameters are the number of filters, the filter size, and epoch. To prevent the model from overfitting, the dropout algorithm with a probability of 0.5 is also used. A binary cross-entropy is used as a loss function with an early stop patience of 20. The batch size is 32, and the number of epochs is set to 200. For the stochastic gradient descent method, the Adam optimization algorithm is selected here. The total number of trainable parameters in the network is 165,365. The entire program is done in Python 3.6.

CNNs

CNNs are widely used in the fields of artificial intelligence, such as machine learning, speech recognition, document analysis, language detection, and image recognition.

Capsule Neural Networks

Capsule neural networks, first proposed by Hinton et al., provide a unique and powerful deep learning component to better simulate the various relationships represented inside the neural network. Because capsule neural networks can collect location information, they can learn a small amount of data to get good predicted results. In the data sets we collected, the amount of RNA data is small, and the length of RNA sequences is small, so to study the hierarchical relationship of local features, capsule neural networks are used in this paper.

BiGRU Networks and Attention Mechanism

BiGRU networks are used to extract the deep features of the sequences because BiGRU networks can be regarded as two unidirectional GRUs. An attention mechanism in a deep neural network is also an important part. The attention mechanism is remarkable in serialized data, such as speech recognition, machine translation, and part of speech taming, which has also been widely used in much bioinformatics research and achieved excellent performance.

Cross-Validation and Independent Testing

Because the *K*-fold (K = 5 or 10) cross-validation (Dezman et al., 2017; G. Q.; Li et al., 2016; Vučković et al., 2016) is widely used to evaluate models, we apply a tenfold cross-validation test to measure model performance in NH_990, NM_944, and NS_627, in which a data set can be divided into 10 mutually exclusive folds, one fold is reserved for testing, whereas the remaining nine folds are used for training purposes. To verify the stability of the models more objectively, the proposed models are tested on two independent data sets H_200 and S_200.

Performance Evaluations

To measure the performance of our model, we use four statistical parameters, sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC), which are used in a series of studies to evaluate the effectiveness of predictors. These parameters are defined as follows:

$$Sn = 1 - \frac{N^+_-}{N^+}, \quad (4)$$

$$Sp = 1 - \frac{N^-_+}{N^-}, \quad (5)$$

$$Acc = 1 - \frac{N^+_- + N^-_+}{N^+ + N^-}, \quad (6)$$

$$MCC = \frac{1 - \frac{N^+_- + N^-_+}{N^+ + N^-}}{\sqrt{\left(1 + \frac{N^-_+ - N^+}{N^+}\right)\left(1 + \frac{N^+_- - N^-_+}{N^-}\right)}}, \quad (7)$$

where N^+, N^- indicate the number of positive and negative sequences, respectively; N^+_-represents the number of positive RNA samples that are incorrectly predicted as negative RNA samples; and N^-_+ represents the number of negative RNA samples that are incorrectly predicted as positive RNA samples. In addition, the graph of the ROC (Fawcett, 2006) is also widely used to intuitively display the performance. Then, the AUC can be obtained to objectively evaluate performances of the proposed model.

RESULTS

Model Selection

To select a more effective model, in each data set, we first compare four models' performances based on two feature

TABLE 2 | Tenfold cross-validation performance comparison of four models based on three feature extraction methods on three benchmark data sets.

Data sets	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC
NH_990	CNN	**67.96**	68.09	67.86	**0.36**	0.737
	CNN + Capsule	66.02	63.83	67.86	0.32	0.742
	CNN + Attention	66.02	46.81	**82.14**	0.31	0.745
	PseUdeep (CNN+ +Capsule + Attention)	66.99	**74.47**	60.71	0.35	**0.746**
NS_627	CNN	69.71	70.59	68.75	0.39	0.728
	CNN + Capsule	68.18	61.76	75.00	0.37	0.735
	CNN + Attention	69.71	**76.47**	68.75	0.40	0.734
	PseUdeep (CNN +Capsule + Attention)	**72.73**	61.75	**78.13**	**0.45**	**0.737**
NM_944	CNN	70.41	57.78	**86.79**	0.41	0.741
	CNN + Capsule	69.39	73.34	66.04	0.39	0.750
	CNN + Attention	70.41	57.78	81.13	0.41	0.751
	PseUdeep (CNN +Capsule + Attention)	**72.45**	**66.70**	77.36	**0.44**	**0.756**

The bold value is the value with the best effect in the corresponding evaluation index.

TABLE 3 | Performance comparison of four models based on three feature extraction methods on independent testing data sets.

Testing data sets	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC
H_200	CNN	65.69	68.63	62.75	0.31	0.691
	CNN + Capsule	62.25	63.73	60.78	0.25	0.696
	CNN + Attention	65.19	52.94	**77.45**	0.31	0.692
	PseUdeep (CNN +Capsule + Attention)	**66.18**	**73.53**	58.82	**0.33**	**0.720**
S_200	CNN	**82.35**	**86.27**	78.43	0.65	0.899
	CNN + Capsule	80.88	77.45	84.31	0.62	0.908
	CNN + Attention	79.91	83.34	76.47	0.59	0.899
	PseUdeep (CNN +Capsule + Attention)	80.88	77.45	**84.31**	**0.65**	**0.909**

The bold value is the value with the best effect in the corresponding evaluation index.

extraction methods, one-hot encoding and KNFP (results are shown in **Supplementary Tables S1, S2**). These models are constructed by gradually adding different types of layers based on two 1-D convolution layers, a BIGRU network, and a two fully connected layers network. The four models are shown below:

1) CNN: The network consists of two layers of 1-D convolution, a BIGRU network, and a two fully connected layers network as described above. The input matrices are the one-hot encoding and KNFP features extracted from the RNA sequences.
2) CNN + Capsule: The model adds a capsule layer after the BiGRU layer on the basis of the CNN model.
3) CNN + Attention: The model adds a self-attention mechanism layer before the BiGRU layer based on the CNN model.
4) CNN + Capsule + Attention: The model adds a capsule layer based on the CNN + Attention model; on the basis of the above four models, we add PSNP features and compare the performance of the four new models (see **Tables 2**, **3**). In summary, our PseUdeep model (CNN + Capsule + Attention model on three feature extraction methods) is superior to the others.

Performance of a Single Type of Feature

We also evaluate our model (CNN + Capsule + Attention) with only one kind of feature. **Table 4** shows the comparison of performance in the tenfold cross-validation on benchmark data sets. It follows that the ACC values and AUC values of PSNP in three species, *H. sapiens*, *M. muscles*, and *S. cerevisiae*, are much higher than those of the other two characteristics. The ACC value of PSNP is increased by 11.11%, 15.6%, and 16.68%, respectively, compared with other characteristics, the AUC value increased by 0.074, 0.199, and 0.115, respectively. PSNP provides a great possibility to improve the model performance in identifying Ψ sites.

Comparison with State-of-the-Art Methods

We compare our model PseUdeep with other state-of-the-art machine learning predictors published recently to evaluate the identification ability of Ψ sites. In benchmark data sets with tenfold cross-validation and independent testing, the results obtained by PseUdeep and other predictors are listed in **Tables 5**, **6** and **Figures 3**, **4**; the ROC curves of PseUdeep are shown in **Figure 5**. The accuracy of the PseUdeep model in NH_990, NS_627, and NM_944 is increased by 1.55%,

TABLE 4 | The model performance with a single type of feature.

Benchmark data sets	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC
NH_990	one-hot	55.56	40	68.51	0.08	0.592
	PSNP	**66.67**	62.22	**70.37**	**0.32**	**0.666**
	KNFP	63.63	**80**	50	0.31	0.658
NS-627	one-hot	53.03	26.47	**81.25**	0.09	0.634
	PSNP	**69.71**	61.75	78.13	**0.40**	**0.734**
	KNFP	66.67	**64.71**	68.75	0.33	0.619
NM-944	one-hot	58.16	35.55	77.35	0.14	0.547
	PSNP	**71.42**	57.77	**83.01**	**0.42**	**0.746**
	KNFP	56.12	**62.22**	50.94	0.13	0.580

The bold value is the value with the best effect in the corresponding evaluation index.

TABLE 5 | A comparison of PseUdeep with other models on three benchmark data sets.

Training data set	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC
NH_990	iRNA-PseU	59.80	61.01	59.80	0.21	0.61
	re-Irna-PseU	61.92	65.05	58.79	0.24	0.65
	PseUI	64.24	64.85	63.64	0.28	0.68
	XG-PseU	65.44	63.64	**67.24**	0.31	0.70
	PseUdeep	**66.99**	**74.47**	60.71	**0.35**	**0.74**
NS-627	iRNA-PseU	64.49	64.65	64.33	0.29	**0.81**
	re-Irna-PseU	65.61	**66.88**	64.33	0.31	0.69
	PseUI	65.13	62.72	67.52	0.30	0.69
	XG-PseU	68.15	66.84	69.45	0.37	0.74
	PseUdeep	**72.73**	61.75	**78.13**	**0.45**	0.74
NM-944	iRNA-PseU	69.07	73.31	64.83	0.38	0.75
	re-Irna-PseU	70.34	**79.87**	60.81	0.41	0.75
	PseUI	70.44	74.58	66.31	0.41	0.77
	XG-PseU	72.03	76.48	67.57	0.45	0.77
	PseUdeep	**72.45**	66.7	**77.36**	**0.44**	**0.77**

The bold value is the value with the best effect in the corresponding evaluation index.

TABLE 6 | A comparison of PseUdeep with other models on independent data sets.

Testing dataset	Models	Accuracy (%)	Sensitivity (%)	Specificity (%)	MCC	AUC
H_200	iRNA-PseU	61.5	58	65	0.23	/
	PseUI	65.5	63	**68**	0.31	/
	PseUdeep	**66.18**	**73.53**	58.82	**0.33**	**0.720**
S_200	iRNA-PseU	60	63	57	0.2	/
	PseUI	68.5	65	72	0.37	/
	PseUdeep	**80.88**	**77.45**	**84.31**	**0.62**	**0.909**

The bold value is the value with the best effect in the corresponding evaluation index.

4.58%, and 0.32%. In addition, the performance of PseUdeep on independent data sets compared with iRNA-Pse and PseUI is shown in **Table 6** and **Figure 4**. It can be observed that the accuracy of the PseUdeep model in H_200 and S_200 is increased by 0.68% and 12.38%, respectively.

We summarize and compare our model with other state-of-the-art models in terms of feature extraction, number of features, and classifiers as shown in **Table 7**. Among them, our model PseUdeep does not further feature selection, and the feature dimension is only 109, 109, and 119 in *H. sapiens*, *M. musculus*, and *S. cerevisiae*, respectively, and our model gets the highest accuracy of the prediction.

CONCLUSION

In this study, we propose a model, PseUdeep, which can effectively identify Ψ sites in RNA sequences. To get better prediction performance, we also train a combination of three features in a simple model and then gradually add different types of layers to obtain better performance. In addition, we compare our model with other models through tenfold cross-validation and independent testing, and the results show that PseUdeep is more accurate and stable. Finally, we evaluate and compare the performance of the three features used in this study and find that PSNP shows the best effect.

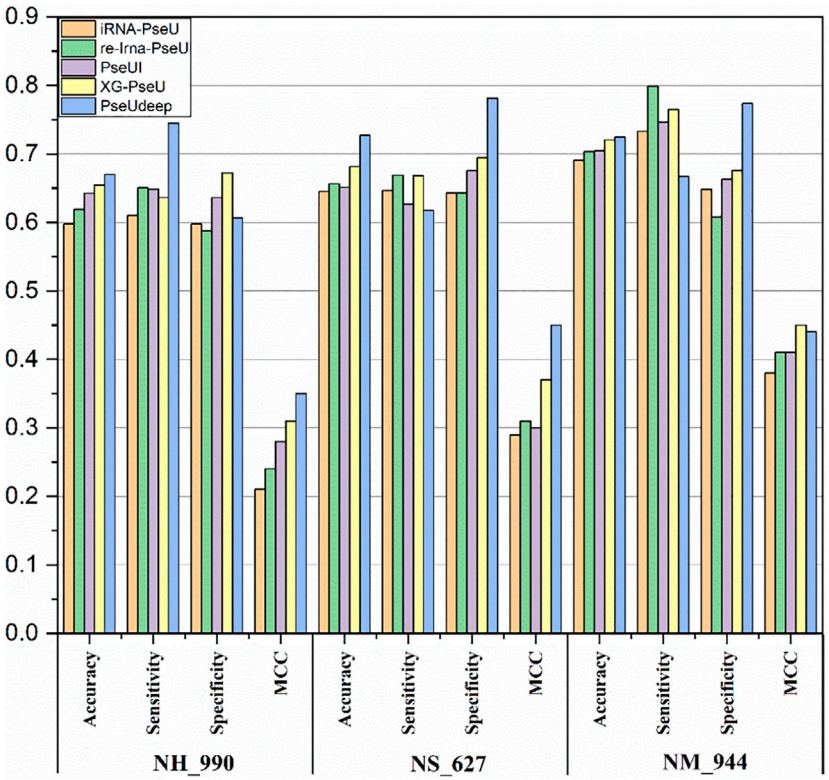


FIGURE 3 | The success rates of the PseUdeep and baseline methods on three training data sets.

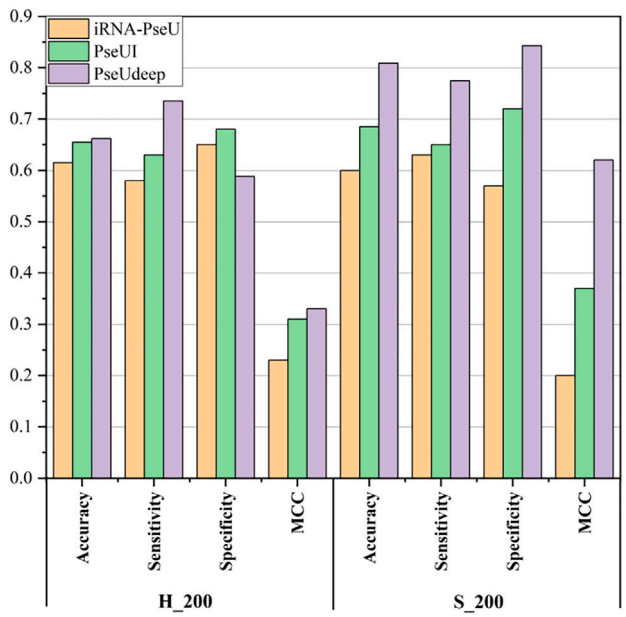


FIGURE 4 | The success rates of the PseUdeep and baseline methods on independent data sets.

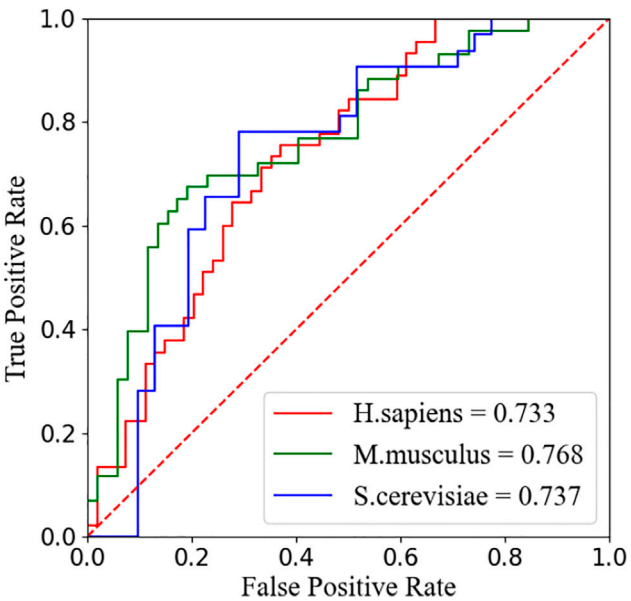


FIGURE 5 | The ROC curves of PseUdeep for *H. sapiens*, *S. cerevisiae*, and *M. musculus*, respectively.

TABLE 7 | Five methods to identify Ψ sites are summarized in all aspects.

Method	Feature extraction	Number of features		Classifiers
iRNA-PseU	PseKNC	*H. sapiens*	84	SVM
		M. musculus	84	
		S. cerevisiae	124	
PseUI	NC + DC + pseDNC + PSNP + PSDP	*H. sapiens*	1045	SVM
		M. musculus	1045	
		S. cerevisiae	1526	
XG-PseU	One-hot + TNC + NCP + ND + DNC	*H. sapiens*	1848	XGBoost
		M. musculus	1848	
		S. cerevisiae	2728	
EnsemPseU	Kmer + Binary + ENAC + NCP + ND	>1700		SVM + XGBoost + NB + KNN + RF
PseUdeep	One-hot + PSNP + KNFP	*H. sapiens*	109	Deep learning network
		M. musculus	109	
		S. cerevisiae	119	

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

JZ and DL conceived, designed, and managed the study. ML and WQ performed the experiments. ML, SC, and JL provided computational support and technical assistance. All authors approved the final manuscript.

FUNDING

This study is supported by the National Natural Science Foundation of China (Grant numbers: 61803065, 11971347, 62071079), the Fundamental Research Funds for the Central Universities of China, the Science and Technology Program of Guangzhou, China (Grant numbers: 2018059), the Science and Technology Planning Project of Guangdong Province of China (Grant numbers: 2020A0505100058), the Guangdong Educational Committee (Key Project of Regular institutions of higher learning of Guangdong Province) (Grant numbers: 2019KZDXM024).

ACKNOWLEDGMENTS

The authors thank those who contributed to this paper, as well as the reviewers for their careful reading and valuable suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.773882/full#supplementary-material

REFERENCES

Bi, Y., Jin, D., and Jia, C. (2020). *EnsemPseU: Identifying Pseudouridine Sites with an Ensemble Approach*. New Jersey: IEEE Access, 1., *PP*(99)

Bousquet-Antonelli, C., Henry, Y., G'Elugne J, P., Caizergues-Ferrer, M., and Kiss, T. (1997). A small nucleolar RNP protein is required for pseudouridylation of eukaryotic ribosomal RNAs. *Embo j* 16 (15), 4770–4776. doi:10.1093/emboj/16.15.4770

Bykhovskaya, Y., Casas, K., Mengesha, E., Inbal, A., and Fischel-Ghodsian, N. (2004). Missense mutation in pseudouridine synthase 1 (PUS1) causes mitochondrial myopathy and sideroblastic anemia (MLASA). *Am. J. Hum. Genet.* 74 (6), 1303–1308. doi:10.1086/421530

Carlile, T. M., Rojas-Duran, M. F., and Gilbert, W. V. (2015). Pseudo-Seq. *Methods Enzymol.* 560, 219–245. doi:10.1016/bs.mie.2015.03.011

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature* 515 (7525), 143–146. doi:10.1038/nature13802

Chan, C. M., and Huang, R. H. (2009). Enzymatic characterization and mutational studies of TruD - the fifth family of pseudouridine synthases. *Arch. Biochem. Biophys.* 489 (1-2), 15–19. doi:10.1016/j.abb.2009.07.023

Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K. C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. *Mol. Ther. Nucleic Acids* 5 (7), e332. doi:10.1038/mtna.2016.37

Chen, W., Lin, H., and Chou, K.-C. (2015). Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Mol. Biosyst.* 11 (10), 2620–2634. doi:10.1039/c5mb00155b

Dezman, Z. D. W., Gao, C., Yang, S., Hu, P., Yao, L., Li, H.-C., et al. (2017). Anomaly Detection Outperforms Logistic Regression in Predicting Outcomes in Trauma Patients. *Prehosp. Emerg. Care* 21 (2), 174–179. doi:10.1080/10903127.2016.1241327

Dou, L., Li, X., Ding, H., Xu, L., and Xiang, H. (2020). Prediction of m5C Modifications in RNA Sequences by Combining Multiple Sequence Features. *Mol. Ther. - Nucleic Acids* 21, 332–342. doi:10.1016/j.omtn.2020.06.004

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Lett.* 27 (8), 861–874. Retrieved from. doi:10.1016/j.patrec.2005.10.010https://www.sciencedirect.com/science/article/pii/S016786550500303X

Ge, J., and Yu, Y.-T. (2013). RNA pseudouridylation: new insights into an old modification. *Trends Biochem. Sci.* 38 (4), 210–218. doi:10.1016/j.tibs.2013.01.002

He, B., Dai, C., Lang, J., Bing, P., Tian, G., Wang, B., et al. (2020). A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1866 (11), 165916. doi:10.1016/j.bbadis.2020.165916

He, J., Fang, T., Zhang, Z., Huang, B., Zhu, X., and Xiong, Y. (2018). PseUI: Pseudouridine sites identification based on RNA sequence information. *BMC Bioinformatics* 19 (1), 306. doi:10.1186/s12859-018-2321-0

He, W., Jia, C., Duan, Y., and Zou, Q. (2018). 70ProPred: a predictor for discovering sigma70 promoters based on combining multiple features. *BMC Syst. Biol.* 12 (Suppl. 4), 44. doi:10.1186/s12918-018-0570-1

He, W., Jia, C., and Zou, Q. (2018). 4mCPred: Machine Learning Methods for DNA N4-methylcytosine sites Prediction. *Bioinformatics* 4, 4. doi:10.1093/bioinformatics/bty668

Huang, Y., He, N., Chen, Y., Chen, Z., and Li, L. (2018). BERMP: a cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach. *Int. J. Biol. Sci.* 14 (12), 1669–1677. doi:10.7150/ijbs.27819

Kierzek, E., Malgowska, M., Lisowiec, J., Turner, D. H., Gdaniec, Z., and Kierzek, R. (2014). The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Res.* 42 (5), 3492–3501. doi:10.1093/nar/gkt1330

Kiss, T., Fayet-Lebaron, E., and Jády, B. E. (2010). Box H/ACA small ribonucleoproteins. *Mol. Cel* 37 (5), 597–606. doi:10.1016/j.molcel.2010.01.032

Li, G.-Q., Liu, Z., Shen, H.-B., and Yu, D.-J. (2016). TargetM6A: Identifying N6-Methyladenosine Sites From RNA Sequences via Position-specific Nucleotide Propensities and a Support Vector Machine. *IEEE Trans.on Nanobioscience* 15 (7), 674–682. doi:10.1109/tnb.2016.2599115

Li, Y.-H., Zhang, G., and Cui, Q. (2015). PPUS: a web server to predict PUS-specific pseudouridine sites: Table 1. *Bioinformatics* 31 (20), 3362–3364. doi:10.1093/bioinformatics/btv366

Liang, Y., Wang, H., Yang, J., Li, X., Dai, C., Shao, P., et al. (2020). A Deep Learning Framework to Predict Tumor Tissue-of-Origin Based on Copy Number Alteration. *Front. Bioeng. Biotechnol.* 8, 701. doi:10.3389/fbioe.2020.00701

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, K., Chen, W., and Lin, H. (2020). XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites. *Mol. Genet. Genomics* 295 (1), 13–21. doi:10.1007/s00438-019-01600-9

Long, H., Liao, B., Xu, X., and Yang, J. (2018). A Hybrid Deep Learning Model for Predicting Protein Hydroxylation Sites. *Ijms* 19 (9), 2817. doi:10.3390/ijms19092817

Mei, Y.-P., Liao, J.-P., Shen, J., Yu, L., Liu, B.-L., Liu, L., et al. (2012). Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. *Oncogene* 31 (22), 2794–2804. doi:10.1038/onc.2011.449

Mostavi, M., Salekin, S., and Huang, Y. (20182018). Deep-2'-O-Me: Predicting 2'-O-methylation sites by Convolutional Neural Networks. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2394–2397. doi:10.1109/embc.2018.8512780

Penzo, M., Guerrieri, A., Zacchini, F., Treré, D., and Montanaro, L. (2017). RNA Pseudouridylation in Physiology and Medicine: For Better and for Worse. *Genes* 8 (11), 301. doi:10.3390/genes8110301

Shaheen, R., Han, L., Faqeih, E., Ewida, N., Alobeid, E., Phizicky, E. M., et al. (2016). A homozygous truncating mutation in PUS3 expands the role of tRNA modification in normal cognition. *Hum. Genet.* 135 (7), 707–713. doi:10.1007/s00439-016-1665-7

Sun, W.-J., Li, J.-H., Liu, S., Wu, J., Zhou, H., Qu, L.-H., et al. (2016). RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.* 44 (D1), D259–D265. doi:10.1093/nar/gkv1036

Vučković, F., Theodoratou, E., Thaçi, K., Timofeeva, M., Vojta, A., Štambuk, J., et al. (2016). IgG Glycome in Colorectal Cancer. *Clin. Cancer Res.* 22 (12), 3078–3086. doi:10.1158/1078-0432.Ccr-15-1867

Wolin, S. L. (2016). Two for the price of one: RNA modification enzymes as chaperones. *Proc. Natl. Acad. Sci. USA* 113 (50), 14176–14178. doi:10.1073/pnas.1617402113

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020). Human geroprotector discovery by targeting the converging subnetworks of aging and age-related diseases. *Geroscience* 42 (1), 353–372. doi:10.1007/s11357-019-00106-x

Yang, Y., Hou, Z., Ma, Z., Li, X., and Wong, K.-C. (2021). iCircRBP-DHN: identification of circRNA-RBP interaction sites using deep hierarchical network. *Brief Bioinform* 22 (4). doi:10.1093/bib/bbaa274

Yu, Y.-T., and Meier, U. T. (2014). RNA-guided isomerization of uridine to pseudouridine-pseudouridylation. *RNA Biol.* 11 (12), 1483–1494. doi:10.4161/15476286.2014.972855

Zhang, Y., and Hamada, M. (2018). DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. *BMC Bioinformatics* 19 (Suppl. 19), 524. doi:10.1186/s12859-018-2516-4

Zhang, Y., Xiang, J., Li, J., Lu, Q., Tian, G., and Yang, J. (2021). Identifying breast cancer-related genes based on a novel computational framework involving KEGG pathways and PPI network modularity. *Front. Genet.* 12, 876. doi:10.3389/fgene.2021.596794

Zhu, X., He, J., Zhao, S., Tao, W., Xiong, Y., and Bi, S. (2019). A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of Saccharomyces cerevisiae. *Brief. Funct. Genomics* 18 (6), 367–376. doi:10.1093/bfgp/elz018

Conflict of Interest: Authors WQ and JL were employed by the company Geneis (Beijing) Co., Ltd.

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

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

frontiers in Genetics

METHODS
published: 13 December 2021
doi: 10.3389/fgene.2021.781277

Predicting Pseudogene–miRNA Associations Based on Feature Fusion and Graph Auto-Encoder

Shijia Zhou[1], *Weicheng Sun*[1], *Ping Zhang*[1] *and Li Li*[1,2]*

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China, [2]Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan, China

Pseudogenes were originally regarded as non-functional components scattered in the genome during evolution. Recent studies have shown that pseudogenes can be transcribed into long non-coding RNA and play a key role at multiple functional levels in different physiological and pathological processes. microRNAs (miRNAs) are a type of non-coding RNA, which plays important regulatory roles in cells. Numerous studies have shown that pseudogenes and miRNAs have interactions and form a ceRNA network with mRNA to regulate biological processes and involve diseases. Exploring the associations of pseudogenes and miRNAs will facilitate the clinical diagnosis of some diseases. Here, we propose a prediction model PMGAE (Pseudogene–MiRNA association prediction based on the Graph Auto-Encoder), which incorporates feature fusion, graph auto-encoder (GAE), and eXtreme Gradient Boosting (XGBoost). First, we calculated three types of similarities including Jaccard similarity, cosine similarity, and Pearson similarity between nodes based on the biological characteristics of pseudogenes and miRNAs. Subsequently, we fused the above similarities to construct a similarity profile as the initial representation features for nodes. Then, we aggregated the similarity profiles and associations of nodes to obtain the low-dimensional representation vector of nodes through a GAE. In the last step, we fed these representation vectors into an XGBoost classifier to predict new pseudogene–miRNA associations (PMAs). The results of five-fold cross validation show that PMGAE achieves a mean AUC of 0.8634 and mean AUPR of 0.8966. Case studies further substantiated the reliability of PMGAE for mining PMAs and the study of endogenous RNA networks in relation to diseases.

Keywords: pseudogene, microRNA, ceRNA network, feature fusion, graph auto-encoder, extreme gradient boosting

OPEN ACCESS

Edited by:
Liqian Zhou,
Hunan University of Technology, China

Reviewed by:
Junfeng Xia,
Anhui University, China
Min Wu,
*Institute for Infocomm Research (A*STAR), Singapore*

***Correspondence:**
Li Li
li.li@mail.hzau.edu.cn

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

Received: *22 September 2021*
Accepted: *16 November 2021*
Published: *13 December 2021*

Citation:
Zhou S, Sun W, Zhang P and Li L (2021) Predicting Pseudogene–miRNA Associations Based on Feature Fusion and Graph Auto-Encoder. Front. Genet. 12:781277. doi: 10.3389/fgene.2021.781277

INTRODUCTION

In mammalian genomes, only about 1–2% of genes encode proteins (Carninci et al., 2005). The remaining parts involve non-coding RNAs, including pseudogenes, long non-coding RNAs (lncRNAs), and miRNAs. Pseudogenes usually refer to DNA sequences similar to genes but lack coding function in the genome. However, there is increasing evidence showing that pseudogenes can be transcribed into non-coding RNAs and become important regulators in organisms, especially in human cancer (Ma et al., 2021). Some of them may be potential therapeutic targets (Shi et al., 2015). The study of pseudogenes may help the diagnosis or clinical treatment of cancer. miRNAs are short non-coding RNAs between 19 and 25 nucleotides in length, accounting for about 3% of the genome

(Setoyama et al., 2011). miRNAs regulate gene expression by acting on mRNAs to affect many developmental processes and the occurrence of diseases (Plank, 2014; Santulli, 2015; Liu Z. et al., 2016). On the other hand, miRNAs can be used as biomarkers for the objective evaluation and diagnosis of tumors (Ruan et al., 2009; Zhang et al., 2012; Stiegelbauer et al., 2014).

Pseudogenes and miRNAs are important components of the competing endogenous RNA (ceRNA) network (Karreth et al., 2015). ceRNAs can regulate gene expression by competing with miRNAs to construct a ceRNA network (Salmena et al., 2011; Rutnam et al., 2014). The ceRNA network can be understood as a balancing mechanism regulating cell activities at the RNA level. Exploring molecular associations in the ceRNA network helps in finding more biological mechanisms at the RNA level. It is important to study various associations in the ceRNA network but this process is often time-consuming and it can be laborious to study the associations by wet experiments. Various computational methods have been developed accordingly.

Currently, non-coding RNA associations in the ceRNA network have been predicted by diverse machine learning methods, which mainly fall into three categories. The first category is based on matrix factorization (MF). MF extracts features by decomposing the input matrix into the product of two or more low-rank matrices. For instance, Zhang et al. proposed a graph-regularized generalized matrix factorization model for predicting a variety of biomolecular interactions (Zhang et al., 2020). Chen et al. and Xu et al. predicted the miRNA–disease associations based on the probability matrix decomposition and inductive matrix completion, respectively (Chen et al., 2018; Xu et al., 2019). Zheng et al. and Liu et al. respectively introduced methods based on collaborative matrix factorization and neighborhood-regularized logistic matrix factorization to predict drug–target interactions (Zheng et al., 2013; Liu Y. et al., 2016). The second category is based on graph embedding. The known associations are learned by the graph embedding method to obtain the behavior information of nodes, and then the characteristics are fused with the characteristic information of nodes, and then the classifiers use node features to predict results. Ji et al. predicted miRNA–disease associations based on the GraRep embedding model (Ji et al., 2020). Song et al. predicted lncRNA–disease associations based on the DeepWalk embedding model (Song et al., 2020). The third category is based on deep learning, among which the most representative method is the graph convolution network (GCN). The GCN is an end-to-end learning model that can deeply integrate the feature information and topological relationship of nodes in the network. Fu et al. proposed a deep learning model based on the multi-view GCN to predict multiple molecular associations (Fu et al., 2021). Xuan et al. and Long et al. proposed GCNLDA and GCNMDA based on the GCN to predict lncRNA–disease associations and microbe-drug associations, respectively (Xuan et al., 2019; Long et al., 2020).

Although pseudogenes play an important role in the ceRNA network, the computational study of associations between pseudogenes and miRNAs is under-developed. Here, we presented a method predicting pseudogene–miRNA associations (PMAs) based on feature fusion and GAE. Given there are many prediction models that can accurately predict lncRNA–miRNA associations, we proposed that the role of pseudogenes is comparable to that of lncRNAs in the ceRNA network. Thus, the expression level can be used as the node feature for pseudogenes as the methods focus on lncRNAs. We fused the node features into the pseudogene–miRNA network and predicted PMAs by a computational method. To the best of our knowledge, this is the first attempt at PMA prediction. The model achieves the mean area under the ROC curve (AUC) and mean area under the precision–recall curve (AUPR) of 0.8634 and 0.8966, respectively. The experimental results confirmed PMGAE-predicted potential PMAs. We also demonstrated the performance of PMGAE through a series of comparative experiments. Together, PMGAE is a powerful and reliable method for the prediction of PMAs as an important component of the ceRNA network.

MATERIALS AND EQUIPMENT

Datasets

We downloaded known PMAs from starBase v2.0 (Li et al., 2014), a large miRNA database that includes the association between miRNAs and lncRNAs and their associations with mRNAs, pseudogenes, and proteins. dreamBase (Zheng et al., 2018) is a database containing massive pseudogene information, including the associations between pseudogenes and the transcription factor (TF), the connection with RNA-binding protein (RBP), and the expression level of pseudogenes in various normal tissues or cancer tissues. We obtained the expression level of pseudogenes in various tissues as the characteristic information of pseudogenes. miRBase (Kozomara et al., 2019) is a comprehensive miRNA sequence database, which contains miRNA sequence information. We obtained the miRNA sequence as the characteristic information of miRNAs from it.

Data Preprocessing

After quality checking and filtering the obtained data, the dataset comprises the expression information of 444 pseudogenes, the sequence information of 173 miRNAs, and 1,884 pairs of pseudogene–miRNA associations. In addition, considering the independence of the testing set used in the case study, we firstly divided all association pairs into two parts. One is used for model training, and the other is used for the case study.

miRNA sequences are composed of four types of nucleotides: A, adenine; G, guanine; C, cytosine; U, uracil. We set k in k-mer to 3, and each miRNA sequence can be represented as a 64 ($4 \times 4 \times 4$)-dimensional vector, where each dimension can represent the frequency of each 3-mer sequence in the sequence. For example, in the miRNA sequence "AGGUUCCAGG," p ("AGG") = 2/(10−3+1). For the pseudogenes, we normalized the expression level of pseudogenes as their characteristics.

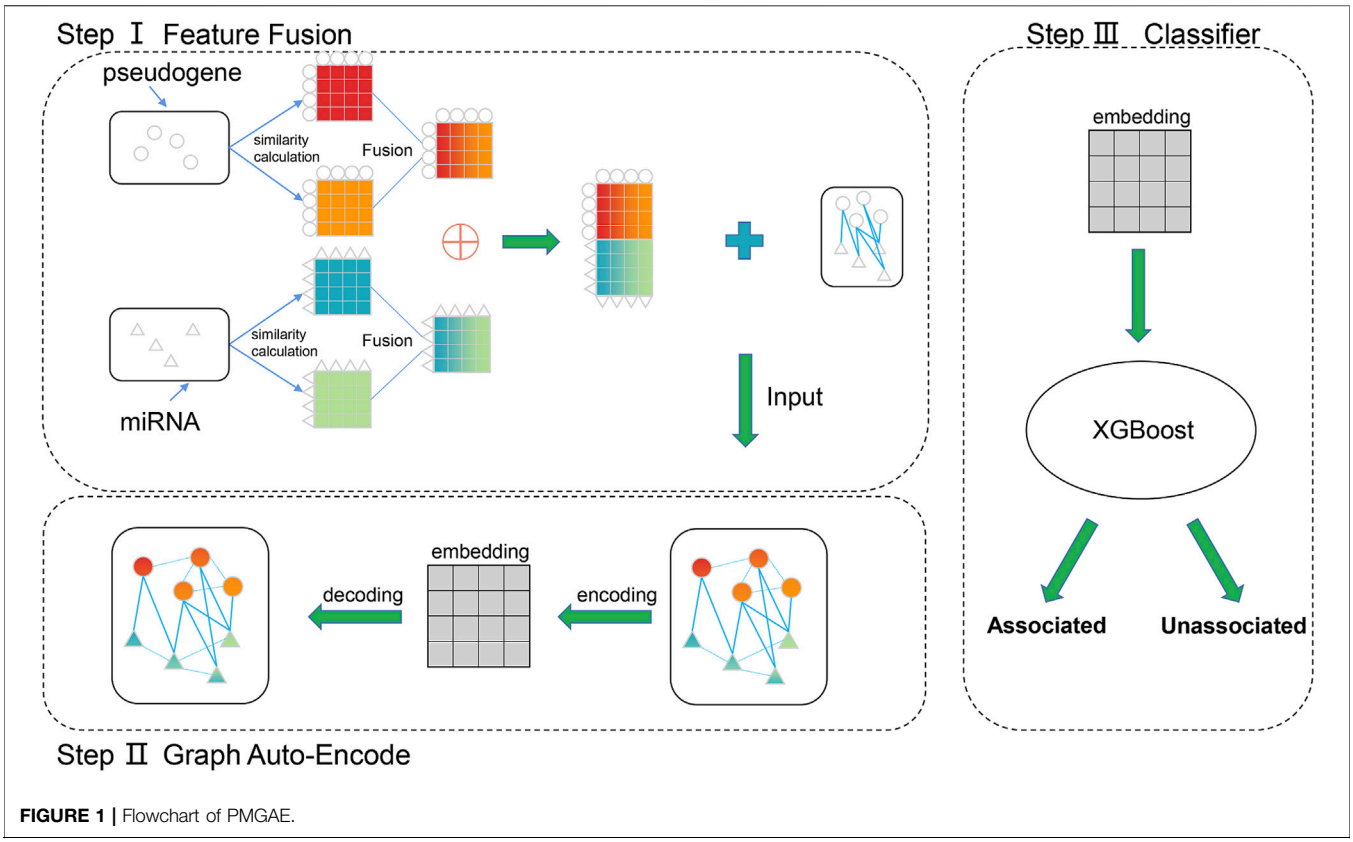


FIGURE 1 | Flowchart of PMGAE.

For the PMAs, we construct a 444 × 173 PMA matrix and put the known PMAs into the PMA matrix. If the *ith* pseudogene is associated with the *jth* miRNA, then let $\mathrm{PMA}(i, j) = 1$; otherwise, let $\mathrm{PMA}(i, j) = 0$.

METHODS

PMGAE Overview

PMGAE is composed of three steps, as shown in **Figure 1**. In step Ⅰ, we calculated and fused the biological characteristics of pseudogenes and miRNAs to obtain the similarity profiles as their features. In step Ⅱ, we obtained the low-dimensional representation vector of nodes by a GAE based on the feature information and association information of existing nodes. In step Ⅲ, we fed the low-dimensional vector into XGBoost to predict the PMAs.

Feature Fusion

We computed the Jaccard similarity coefficient, cosine similarity coefficient, and Pearson similarity coefficient based on the respective characteristics of pseudogenes and miRNAs. We calculated Gaussian kernel similarity based on PMAs to replace the zeros in the matrix (Chen, 2015). Eventually, we generated the pseudogene similarity (PS) profile of 444 × 444 in dimension and the miRNA similarity (MS) profile of 173 × 173 in dimension. Jaccard similarity, cosine similarity, and Pearson similarity can be calculated as follows:

$$\begin{aligned} Jaccard(X,Y) &= \frac{X \cap Y}{X \cup Y}, \\ Cos(x,y) &= \frac{\sum_{k=1}^{n} x_k y_k}{\sqrt{\sum_{k=1}^{n} x_k^2}\sqrt{\sum_{k=1}^{n} y_k^2}}, \\ \rho_{X,Y} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}. \end{aligned} \tag{1}$$

Individual similarity measures between pseudogenes and between miRNAs may contain noise in the data. In order to reduce the noise, we fused several similarity profiles by feature fusion. Feature fusion obtains a single output matrix by fusing all similarity profiles with non-linear methods (Wang et al., 2014). Firstly, we construct the weight matrix as

$$P(i,j) = \begin{cases} \dfrac{S(i,j)}{2\sum_{k \neq i} S(i,k)}, i \neq j \\ 1/2, i = j \end{cases}. \tag{2}$$

The local affinity matrix is defined as

$$L(i,j) = \begin{cases} \dfrac{S(i,j)}{\sum_{k \in N_i} S(i,k)}, j \in N_i \\ 0, otherwise \end{cases}, \tag{3}$$

where $S(i, j)$ represents the similarity matrix and N_i represents neighbors of the *ith* node. Then, we iteratively update the matrix as

$$P_{t+1}^{(v)} = L^{(v)} \times \left(\frac{\sum_{k \neq v} P_t^k}{n-1} \right) \times \left(L^{(v)} \right)^T, v = 1, 2, ..., n. \quad (4)$$

The final feature matrix (here, we set n to 3 in our model) is represented as

$$P_t = \frac{p_t^{(1)} + p_t^{(2)} + ... + p_t^{(n)}}{n}. \quad (5)$$

For the fusion similarity profiles *PS* and *MS*, we removed the noise by a stacked auto-encoder (SAE) and obtained the low-dimensional vector representation of pseudogenes and miRNAs. By an SAE, we obtained 128-dimensional matrix representations of $PS^{'}$ and $MS^{'}$ for pseudogenes and miRNAs, respectively. Finally, in order to improve the training speed and prediction effect of the model, we tried to standardize the obtained 128-dimensional vectors. Specifically, we carried it out using StandardScaler and RobustScaler individually. StandardScaler and RobustScaler can be expressed as

$$\begin{aligned} x' &= \frac{x - \mu}{\sigma}, \\ y' &= \frac{y - median}{IQR}, \end{aligned} \quad (6)$$

where *IQR* represents the interquartile range of the sample.

StandardScaler improves the rate of learning and prediction accuracy of the model. RobustScaler reduces the effect of outliers on results. Both of them are important, so we took the mean values of the matrix that are treated by each of them separately and obtained the final feature matrices PS'' and MS''. Finally, the node feature matrix X is constructed as

$$X = \begin{pmatrix} PS'' \\ MS'' \end{pmatrix}. \quad (7)$$

Graph Auto-Encoder

Auto-encoder is a kind of neural network, which can restore the input using output through certain training. It includes an encoder and a decoder. The encoder obtains the low-dimensional representation of the input vector (Baldi, 2012). The GAE migrates the auto-encoder to a graph (Kipf and Welling, 2016). We constructed the adjacency matrix and the feature matrix of the nodes. The goal is to obtain the low-dimensional representation of the nodes by deeply integrating the association information between nodes and the feature information of nodes themselves through the GAE. The GAE uses a two-layer graph convolution network as an encoder, which can be described as follows:

$$GCN(X, A) = \tilde{A}\mathrm{ReLu}\left(\tilde{A}XW_0 \right)W_1, \quad (8)$$

where $\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, $\mathrm{ReLu}(X) = \max(X, 0)$ represents the activation function, and W_0 and W_1 are parameters to be learned.

We built the adjacency matrix based on the PMA network as follows:

$$A = \begin{pmatrix} 0 & PMA \\ PMA^T & 0 \end{pmatrix}, \quad (9)$$

where PMA^T represents the transpose of the matrix PMA.

We used the adjacency matrix A and feature matrix X to obtain the low-dimensional representation vector of nodes by an encoder, which can be defined as

$$Z = GCN(X, A). \quad (10)$$

The decoder also obtains the low-dimensional vector recomposition map based on the neural network. The decoder generates a graph according to the probability of edges between nodes. It can be defined as

$$\hat{A} = sigmoid\left(ZZ^T\right), \quad (11)$$

where $sigmoid(x) = \frac{1}{1+e^{-x}}$ represents the activation function. $\hat{A}$ is the reconstructed network matrix. In this study, in order to make the model more explanatory, we do not use the decoder layer but put the low-dimensional representation vector of nodes into the best classifier we trained to predict the PMAs.

To measure the error between the predicted and the real association, the loss function is defined as

$$L = -\frac{1}{N} \sum y \log\hat{y} + (1 - y)\log(1 - \hat{y}), \quad (12)$$

where y represents the value of an element in the adjacency matrix A (0 or 1) and $\hat{y}$ represents the value of the same element in the reconstructed adjacency matrix $\hat{A}$ (0–1). We took multiple epochs to minimize the loss function to make the reconstituted data as similar to the original data as possible.

Subsequently, we predicted potential PMAs by XGBoost. XGBoost is a machine learning algorithm whose core idea is to integrate multiple decision trees and continuously add trees to them. Each addition of trees is a process of iteratively adding new functions. Its purpose is to make the final predicted value as close as possible to the real value. Its implementation process can be expressed as

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i). \quad (13)$$

The objective function of XGBoost is defined as follows:

$$L(\varphi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \quad (14)$$

where $l(y_i, \hat{y}_i)$ is the training error and $\Omega(f_k)$ is the regularization term to suppress over-fitting.

Graph Embedding

In contrast to the traditional machine learning algorithm which may only consider the mapping from input to output without considering the associations in the network, the graph-based algorithm can obtain the associations between nodes together with their own characteristics to improve the accuracy of prediction. The graph data we obtain from real life are often

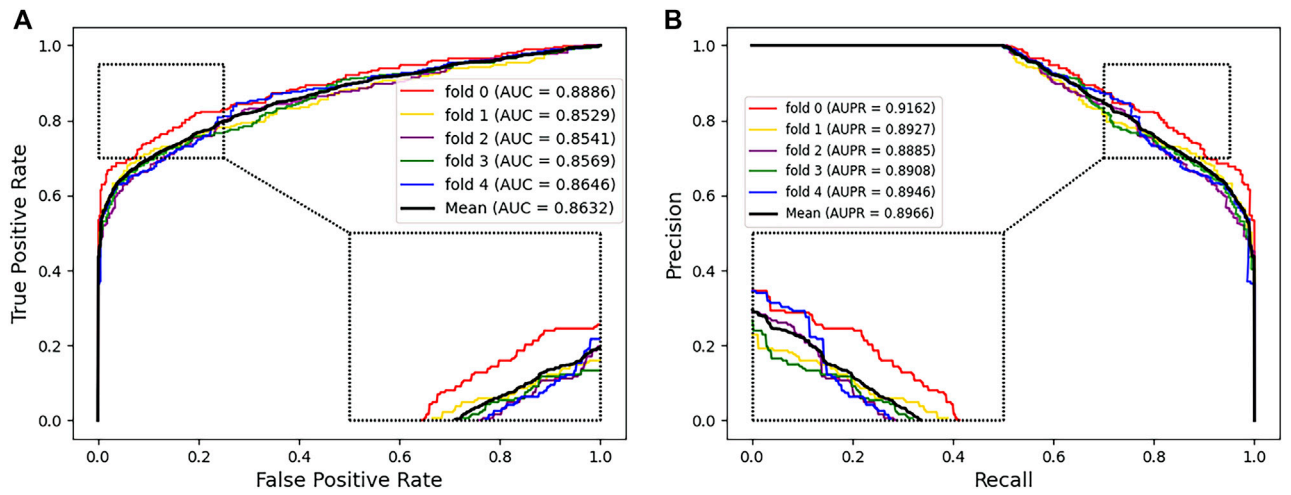


FIGURE 2 | AUC **(A)** and AUPR **(B)** of PMGAE using five-fold cross validation. Insets represent the zoom-in view of local regions.

high-dimensional and sparse. Graph embedding is the process of mapping the input graph data to low-dimensional dense vectors, which can reinforce the efficiency of machine learning and improve the accuracy of prediction.

We selected several representative graph embedding methods including Line (Tang et al., 2015), GraRep (Cao et al., 2015), Node2vec (Grover and Leskovec, 2016), and DeepWalk (Perozzi et al., 2014) to predict the PMAs and compared the results of PMGAE in *Results*.

RESULTS

Experimental Setup and Performance Evaluation

For the experiment parameters in the GAE, we set a learning rate of 0.001 and trained the model for 8,000 epochs. We obtained a 32-dimensional representation for each node. Then, they were put into XGBoost for prediction. In addition, we used five-fold cross validation to evaluate the performance of the model. We take the known PMAs as a positive sample. The remaining unknown PMAs can be considered potential negatives from which we randomly selected PMAs with equal size to the positive samples as negative samples. Subsequently, we randomly divided the positive and negative samples into five parts. One in the five parts was taken out in turn as a test set, and the remaining were used as the training sets.

We used several evaluation metrics including accuracy, sensitivity, specificity, and precision. In addition, we also adopted the AUC and AUPR to evaluate the prediction performance. We took multiple independent experiments of five-fold cross validation to reduce the error. The mean AUC and AUPR were shown under the corresponding curve (**Figure 2**). The AUC and AUPR of our prediction model reached 0.8634 and 0.8966, respectively, which showed that PMGAE has satisfactory performance in PMA prediction.

Comparison of the Performance of PMGAE and MF-Based Methods

MF-based methods have shown excellent performance in predicting the correlation of various biomolecules. To evaluate the performance of PMGAE, we compared it with MF-based methods including multiple similarities collaborative matrix factorization (MSCMF), inductive matrix completion for miRNA–disease association (IMCMDA), and neighborhood-regularized logistic matrix factorization (NRLMF). MSCMF is a collaborative filtering model integrating multiple similarities for predicting drug–target interactions (Zheng et al., 2013). IMCMDA is a matrix completion–based model, integrating miRNA–disease associations, individual miRNA and disease characteristics, and Gaussian interaction profile kernel similarity between them to predict miRNA–disease associations (Chen et al., 2018). NRLMF combined logical matrix factorization and neighborhood regularization to predict drug–target interactions (Liu Y. et al., 2016).

As shown in **Figure 3**, PMGAE showed the best performance in terms of AUC and AUPR. Relative to the MF-based methods, the GAE can effectively extract node features, with the best prediction achieved through XGBoost.

Visualization of Embedding Effect

Because the features are high-dimensional, it is difficult to visualize the clustering results directly. In order to make the model more interpretable and validate the embedded effects, we mapped the features of the nodes before and after embedding them into the three-dimensional space through t-SNE (Maaten and Hinton, 2008). t-SNE can reduce the high-dimensional data to two or three dimensions. Through t-SNE, we can do an intuitive observation on the embedding method for the node clustering effect.

As shown in **Figure 4**, nodes are randomly distributed before embedding, and our embedding method leads to clustering of the nodes based on their characteristics. Since similar molecules may

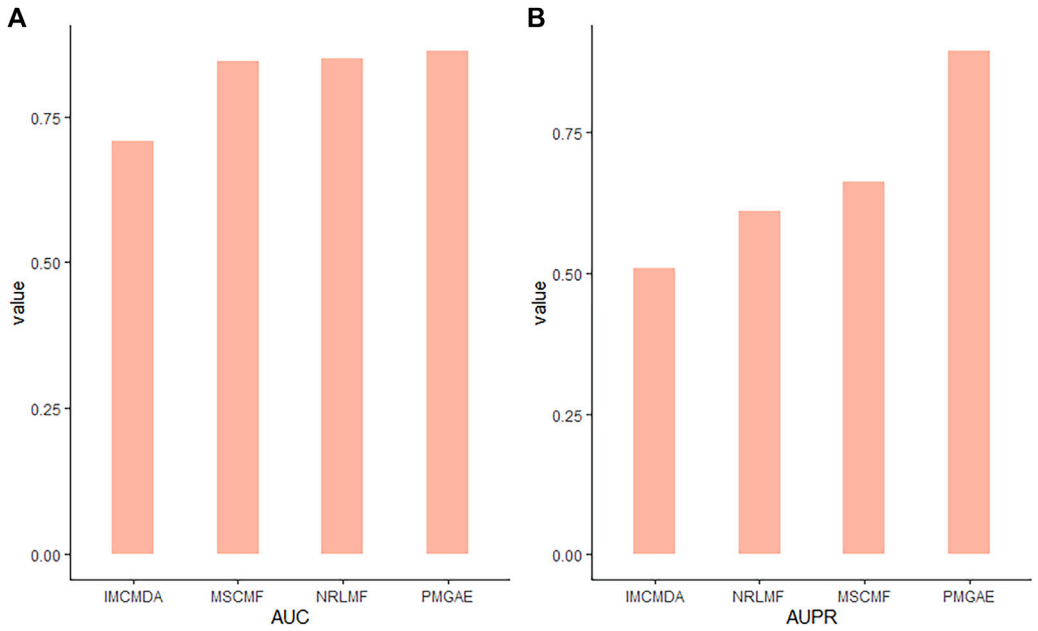


FIGURE 3 | Comparison of AUC **(A)** and AUPR **(B)** of PMGAE and MF-based models.

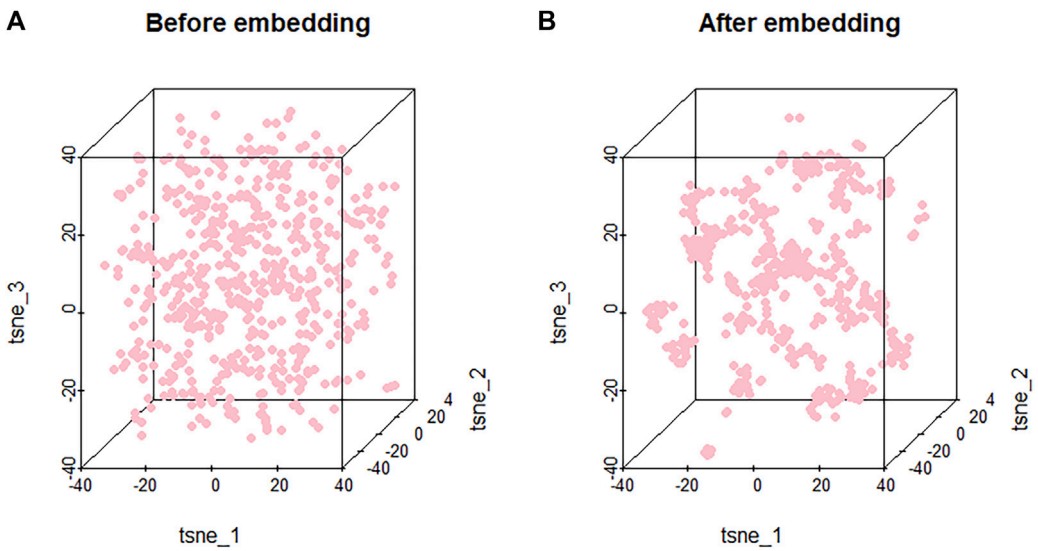


FIGURE 4 | Clustering results of nodes before **(A)** and after **(B)** embedding.

have similar or related biological functions, effective clustering can facilitate potential association prediction and improve the performance of the model. The effective clustering through embedding validates it as an important component of PMGAE.

Feature Fusion With Various Similarity Measures

Using the expression information of pseudogenes and the k-mer sequence information of miRNAs, we calculated the Jaccard similarity coefficient, cosine similarity coefficient, and Pearson similarity coefficient of pseudogenes and miRNAs, respectively. Then, pairwise fusion and full fusion were performed and compared. **Table 1** shows the performance of specific fusions and no fusion.

Individual similarity has its own limitations. For example, the cosine similarity coefficient tends to distinguish differences from directions; thus, it has a good effect on the calculation of different directions but is not sensitive to the change of values. The Jaccard similarity coefficient has a good effect on the binary data, but it

TABLE 1 | Model performance comparison using similarity profile fusions and using individual similarity profiles.

Methods	Evaluation metrics					
	Acc.	Sen.	Spec.	Prec.	AUC	AUPR
Jaccard	0.7641	0.6443	0.8838	0.8475	0.8416	0.8676
Pearson	0.7633	0.6555	0.8710	0.8356	0.8381	0.8637
Cosine	0.7901	0.6491	0.9310	0.9040	0.8562	0.8872
Cosine + Jaccard	0.7927	0.6433	0.9421	0.9176	0.8607	0.8912
Cosine + Pearson	0.7964	0.6396	0.9533	0.9320	0.8591	0.8935
Jaccard + Pearson	0.7954	0.6460	0.9448	0.9214	0.8565	0.8913
Full fusion	0.8015	0.6592	0.9437	0.9216	0.8632	0.8966

cannot measure the specific value of the difference. The Pearson similarity coefficient tends to give better results when the data do not conform to a certain rule, but the effect on overlapping data is compromised. Considering these shortcomings, we tried to fuse these similarity measures in a non-linear way for a better similarity representation by integrating the advantages. The experimental results in **Table 1** show that our full similarity fusion method can effectively improve the performance of the model.

Comparison of the Performance of Various Embedding Methods

For each method, the mean of individual runs is used to measure its performance. As shown in **Figure 5**, the PMGAE model shows the best prediction. The performance of GAE is superior to that of other graph embedding methods. The GAE more effectively mines the topology structure in the scenario of node information in the network than other embeddings.

Although the graph embedding models mentioned above have many advantages, according to our experimental study, we found that these models still have some drawbacks. Specifically, the Line model only considers the first-order relationship and second-order relationship of nodes. It cannot construct the global structure of the network well, and the embedding of Line for low-level nodes is not accurate enough. Thus, the prediction outcome of Line is the least accurate in our data. DeepWalk takes into account each first-order relationship of the node with all relationships stored in a subspace. But it cannot distinguish the order of the node's neighbors during training. At the same time, DeepWalk is only applicable to unweighted graphs and has obvious limitations. The Node2vec model combines some advantages of Line and DeepWalk and also can control the preference of random walk by adjusting the hyperparameters. However, when the number of samples is limited as in the case of PMGAE, the length of random walk is also limited. So, the learning effect for remote neighbors in the network is far from optimum. The GraRep model can put each first-order relationship between nodes in different subspaces, which well constructs the global structure of the network. However, the calculation of each first-order relationship A^k and the optimization loss function is large, so it cannot be used for large-scale graph data. Besides, the above-mentioned graph embedding models often only take into account the topological information of nodes but do not well incorporate the characteristic information of nodes themselves. The GAE can achieve the best predictions, mainly because it uses the graph convolution neural network to learn the characteristics of nodes in an end-to-end way. At the same time, the GAE has better robustness and stability, together with good learning effect for poor datasets.

Comparison of the Performance of Various Classifiers

Classifiers play a key role in the model. To compare the prediction performance of our model under different classifiers and select the best classifier, we seek to check its predictive performances with five representative classifiers: eXtreme Gradient Boosting

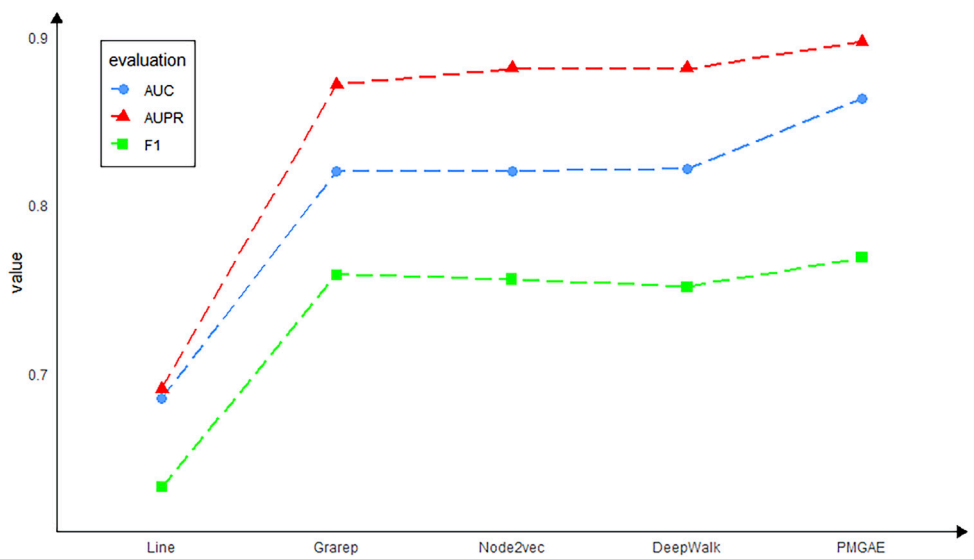


FIGURE 5 | Model performance using various embedding methods.

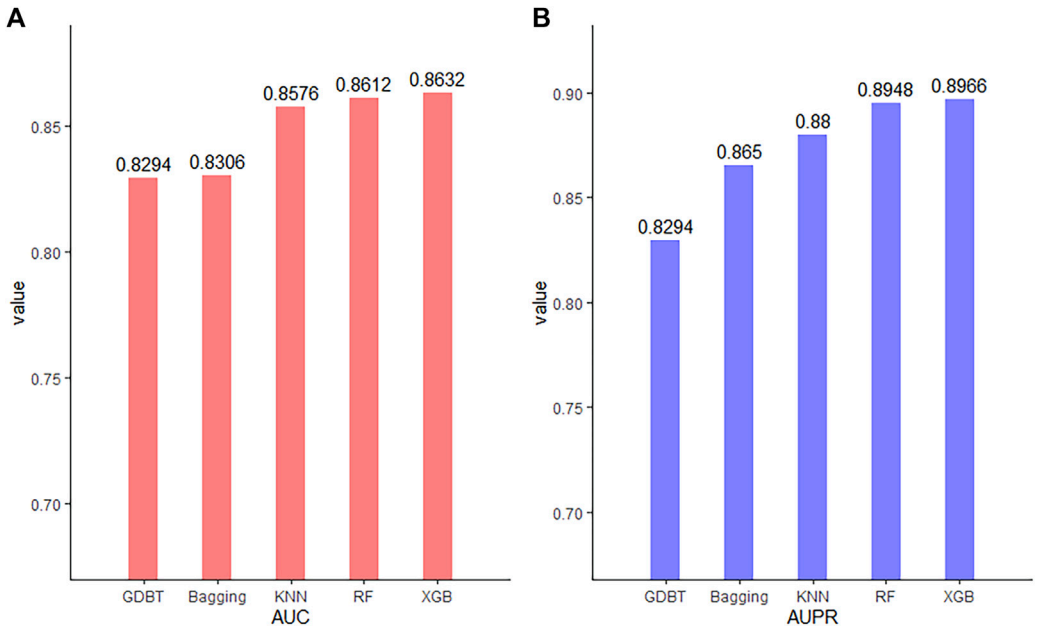


FIGURE 6 | AUC **(A)** and AUPR **(B)** using various classifiers.

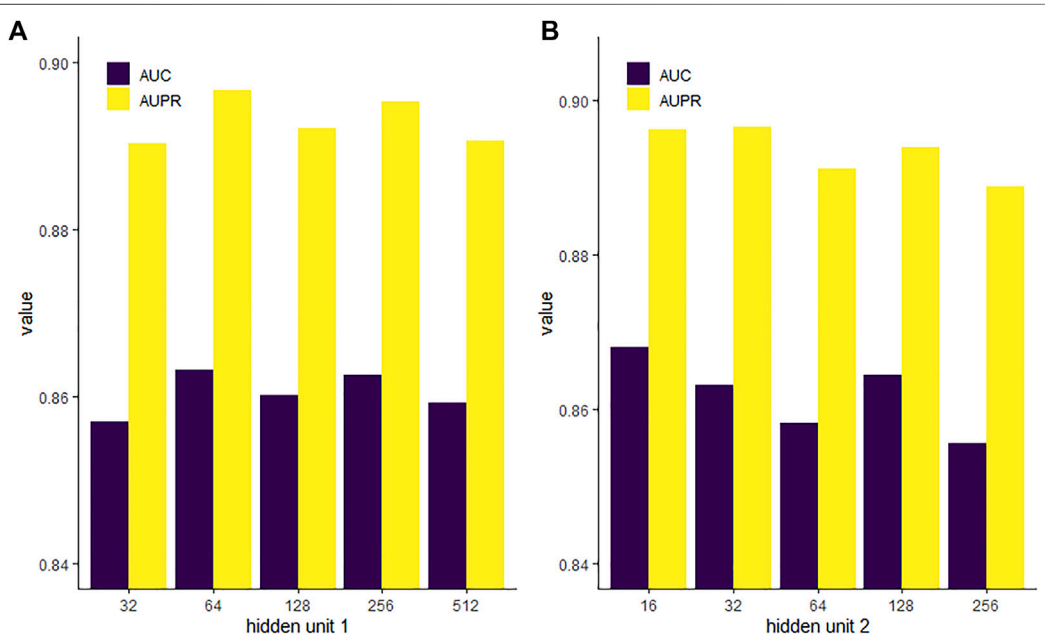


FIGURE 7 | AUC and AUPR of various hidden unit setups in the first **(A)** and second **(B)** layers of GAE.

(XGBoost), random forest (RF), K-nearest neighbor (KNN), bagging, and gradient boosting decision tree (GBDT). The AUC and AUPR were used to evaluate their performance. As shown in **Figure 6**, while all the classifiers have an AUC and AUPR above 0.8, XGBoost yields the best performance. Thus, XGBoost is most suitable for our model.

Comparison of GAE With Various Setups of Hidden Units

The GAE contains two layers of hidden units in the neural network. We evaluated the impact of different dimensions of each layer on the performance of the model. We fixed the second hidden layer with 32 units and then set the first hidden layer with

TABLE 2 | Model performance under various setups of positive: negative sample ratios.

Evaluation metrics	Positive: negative sample ratio				
	1:1	1:2	1:5	1:10	1:20
AUC	0.8632	0.8548	0.8557	0.8596	0.8626
AUPR	0.8966	0.8388	0.7653	0.7193	0.6693
Acc.	0.8015	0.8523	0.9218	0.9554	0.9753
Sen.	0.6592	0.6008	0.5594	0.5419	0.5196
Spec.	0.9437	0.9782	0.9943	0.9968	0.9981
Prec.	0.9216	0.9323	0.9513	0.9447	0.9323
MCC	0.6292	0.6646	0.6938	0.6965	0.6858

units of 32, 64, 128, 256, and 512, respectively. **Figure 7** shows that when the first hidden unit is 64, the GAE has the best performance. Then, we set the first hidden layer with units of 64 and set the second hidden layer with units of 16, 32, 64, 128, and 256, respectively. We found that model performance was slightly improved with the decrease of the unit number. The AUPR is highest when the unit number is reduced to 32, and the AUC is highest when the unit number is reduced to 16. High-dimensional representation may lead to data sparsity, which is not conducive to classification. While reducing dimension can improve the training speed of the model, dimensions too low may cause loss of key information. For the task of PMA prediction, we chose the first hidden unit to be 64 and the second hidden unit to be 32.

Effect of Ratio of Positive to Negative Samples

Unbalanced test sets containing too many negative samples may affect the performance of the model. To explore the impact of this data imbalance on PMGAE, we used various setups of positive: negative sample ratios. In the five-fold cross validation, we constructed 1:1, 1:2, 1:5, 1:10, and 1:20 test sets by changing sizes of potentially negative samples. **Table 2** shows the experimental results. The test set with different proportions has a moderate effect on the results. It suggests that, for the evaluation of model performance in predicting PMAs, the influence of different positive: negative sample ratios cannot be omitted.

Case Studies

Exploring cases of PMAs is of great significance to provide insights for research of diseases. Seeking support of our predictions from independent sources can evaluate the effectivity and robustness of PMGAE. For the case study, we used all other associations that did not contain three pseudogenes RPLP0P2, HLA-H, and HLA-J to train the model and then predicted the probability of all miRNAs associated with each of these three pseudogenes. The top 15 predicted associations were used to verify the predictions through starBase.

Three pseudogenes, RPLP0P2, HLA-H, and HLA-J, were used for case studies. RPLP0P2 is a pseudogene associated with a variety of cancers including lung adenocarcinoma and colorectal cancer. Several studies have shown that low expression of RPLP0P2 can lead to decreased proliferation and adhesion of tumor cells (Chen et al., 2016; Yuan et al., 2021). **Table 3** shows the top 15 candidate miRNAs associated with RPLP0P2, 11 of which are supported by starBase.

HLA-H is a kind of transmembrane molecule, and it can mobilize HLA-E at the cell surface of multiple immune cells (Jordier et al., 2019). At the same time, HLA-H gene mutations cause many cases of hereditary hemochromatosis. **Table 3** shows the top 15 candidate miRNAs associated with HLA-H, 12 of which are proved by starBase.

HLA-J is also a class of HLA gene. HLA-J has an immunosuppressive effect and is potentially a predictor of breast cancer (Würfel et al., 2020). Besides, HLA-A has been shown to be associated with schizophrenia. The presence of HLA-AM80468 significantly reduces the incidence of schizophrenia, whereas the presence of HLA-JM80469 increases the incidence of schizophrenia (Gu et al., 2013). As shown in **Table 3**, 11 of the top 15 candidate miRNAs associated with HLA-J are proved by starBase.

DISCUSSION

Genome-wide prediction of PMAs has great significance in both biology and medicine. It can not only help us understand the cellular role of pseudogenes but also provide clues and directions for the clinical treatment of various diseases. In this work, full potential PMAs are predicted for the first time. Feature fusion and GAE were used to construct the model, PMGAE. The performance of PMGAE was evaluated by five-fold cross validation, with an AUC of 0.8634 and AUPR of 0.8966 obtained. Extensive experiments on feature fusion, model framework, and setup were conducted.

The good performance of PMGAE may be attributed to the optimization of each step and flexibility together with the good interpretability of the model. First, we integrated the attribute information from different perspectives of nodes by feature fusion. Subsequently, the GAE was used to integrate the correlation information and attribute information to obtain the low-dimensional representation of nodes. Finally, we selected the most suitable classifier for the model as an association prediction task. By comparative experiments on the feature construction, embedding method, and classifiers, the best integrated model can be selected. The resultant PMGAE model has the optimal effect in predicting the PMAs.

In the ceRNA network, pseudogene–miRNA is the only pair of relationships that have not been studied computationally. By predicting PMAs for the first time, using PMGAE, our work fills the gap in the ceRNA network, so that all known relational pairs in the ceRNA network can be predicted by computational methods. The completed map will facilitate the studies of ceRNA network architecture and its biological implications.

Based on the successful application of PMGAE, there is space for further improvement. First, only one type of feature for each node was used when constructing a similarity feature profile. Fusing more types of node features may provide more information for model training. Second, one can also

TABLE 3 | The top 15 candidate miRNAs associated with pseudogenes RPLP0P2, HLA-H, and HLA-J and the evidence from starBase.

Rank	RPLP0P2		HLA-H		HLA-J	
	miRNA	starBase	miRNA	starBase	miRNA	starBase
1	hsa-miR-15a-5p	Yes	hsa-miR-15a-5p	Yes	hsa-miR-497-5p	Yes
2	hsa-miR-424-5p	Yes	hsa-miR-15b-5p	Yes	hsa-miR-424-5p	Yes
3	hsa-miR-15b-5p	Yes	hsa-miR-16-5p	Yes	hsa-miR-195-5p	Yes
4	hsa-miR-195-5p	Yes	hsa-miR-195-5p	Yes	hsa-miR-16-5p	Yes
5	hsa-miR-497-5p	Yes	hsa-miR-497-5p	Yes	hsa-miR-15b-5p	Yes
6	hsa-miR-16-5p	Yes	hsa-miR-424-5p	Yes	hsa-miR-15a-5p	Yes
7	hsa-miR-34c-5p	No	hsa-miR-199b-5p	No	hsa-miR-23c	Yes
8	hsa-miR-449a	No	hsa-miR-3619-5p	Yes	hsa-miR-103a-3p	No
9	hsa-miR-378b	No	hsa-miR-761	Yes	hsa-miR-204-5p	No
10	hsa-miR-320c	Yes	hsa-miR-106b-5p	No	hsa-miR-3619-5p	Yes
11	hsa-miR-761	Yes	hsa-miR-125a-5p	Yes	hsa-miR-134-5p	Yes
12	hsa-miR-99a-5p	No	hsa-miR-4319	Yes	hsa-miR-613	No
13	hsa-miR-320d	Yes	hsa-miR-146a-5p	No	hsa-miR-29b-3p	Yes
14	hsa-let-7d-5p	Yes	hsa-miR-875-5p	Yes	hsa-miR-125b-3p	No
15	hsa-let-7b-5p	Yes	hsa-miR-503-5p	Yes	hsa-miR-761	Yes

introduce intermediate layers to incorporate pseudogene–lncRNA associations and lncRNA–miRNA associations. Whether adding intermediate layers will improve the prediction effect of the model is a problem worth further exploration. Third, when constructing negative samples, we simply used non-positive samples as potential negative samples and then randomly extracted them. How to build negative samples more accurately is also a question worth exploring. Fourth and more importantly, in PMGAE, embedding and classifier are sequentially, also separately trained. For the task of PMA prediction, end-to-end modeling seeking a global optimal solution is worth further exploration. Toward a full description and understanding, we will incorporate all relation pairs to build a complete graph of the ceRNA network, together with diverse information of all types of nodes.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

LL and PZ designed the methods and arranged the datasets. SZ implemented the methods and performed the analyses. SZ and WS tested the methods. SZ and LL wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

We acknowledge the financial support from the National Natural Science Foundation of China (no. 31771430 to LL), Huazhong Agricultural University Scientific and Technological Self-innovation Foundation (to LL), and Hubei Hongshan Laboratory (to LL).

ACKNOWLEDGMENTS

The authors thank the lab members for their assistance.

REFERENCES

Baldi, P. (2012). *Autoencoders, Unsupervised Learning, and Deep Architectures*. Bellevue, WA: ICML Unsupervised and Transfer Learning, 37–49.

Cao, S., Lu, W., and Xu, Q. (2015). "GraRep: Learning Graph Representations with Global Structural Information," in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, 891–900.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The Transcriptional Landscape of the Mammalian Genome. *Science* 309, 1559–1563. doi:10.1126/science.1112014

Chen, J., Hu, L., Chen, J., Wu, F., Hu, D., Xu, G., et al. (2016). Low Expression lncRNA RPLP0P2 Is Associated with Poor Prognosis and Decreased Cell Proliferation and Adhesion Ability in Lung Adenocarcinoma. *Oncol. Rep.* 36 (3), 1665–1671. doi:10.3892/or.2016.4965

Chen, X. (2015). KATZLDA: KATZ Measure for the lncRNA-Disease Association Prediction. *Sci. Rep.* 5 (1), 16840. doi:10.1038/srep16840

Chen, X., Wang, L., Qu, J., Guan, N. N., and Li, J. Q. (2018). Predicting miRNA-Disease Association Based on Inductive Matrix Completion. *Bioinformatics* 34 (24), 4256–4265. doi:10.1093/bioinformatics/bty503

Fu, H., Huang, F., Liu, X., Qiu, Y., and Zhang, W. (2021). MVGCN: Data Integration through Multi-View Graph Convolutional Network for Predicting Links in Biomedical Bipartite Networks. *Bioinformatics*. doi:10.1093/bioinformatics/btab651

Grover, A., and Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. *KDD* 2016, 855–864. doi:10.1145/2939672.2939754

Gu, S., Fellerhoff, B., Müller, N., Laumbacher, B., and Wank, R. (2013). Paradoxical Downregulation of HLA-A Expression by IFNγ Associated with Schizophrenia and Noncoding Genes. *Immunobiology* 218 (5), 738–744. doi:10.1016/j.imbio.2012.08.275

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. doi:10.1038/s41598-020-63735-9

Jordier, F., Gras, D., De Grandis, M., D'Journo, X. B., Thomas, P. A., Chanez, P., et al. (2019). HLA-H: Transcriptional Activity and HLA-E Mobilization. *Front. Immunol.* 10, 2986. doi:10.3389/fimmu.2019.02986

Karreth, F. A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., et al. (2015). The BRAF Pseudogene Functions as a Competitive Endogenous RNA and Induces Lymphoma *In Vivo*. *Cell* 161 (2), 319–332. doi:10.1016/j.cell.2015.02.043

Kipf, T., and Welling, M. (2016). Variational Graph Auto-Encoders. ArXiv abs/1611.07308.

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA Sequences to Function. *Nucleic Acids Res.* 47, D155–D162. doi:10.1093/nar/gky1141

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucl. Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Liu, Y., Wu, M., Miao, C., Zhao, P., and Li, X.-L. (2016a). Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *Plos Comput. Biol.* 12 (2), e1004760. doi:10.1371/journal.pcbi.1004760

Liu, Z., Zhang, X.-H., Callejas-Díaz, B., and Mullol, J. (2016b). MicroRNA in United Airway Diseases. *Ijms* 17 (5), 716. doi:10.3390/ijms17050716

Long, Y., Wu, M., Kwoh, C. K., Luo, J., and Li, X. (2020). Predicting Human Microbe-Drug Associations *via* Graph Convolutional Network with Conditional Random Field. *Bioinformatics* 36 (19), 4918–4927. doi:10.1093/bioinformatics/btaa598

Ma, Y., Liu, S., Gao, J., Chen, C., Zhang, X., Yuan, H., et al. (2021). Genome-wide Analysis of Pseudogenes Reveals HBBP1's Human-specific Essentiality in Erythropoiesis and Implication in β-thalassemia. *Dev. Cel.* 56 (4), 478–493. doi:10.1016/j.devcel.2020.12.019

Maaten, L. V. D., and Hinton, G. E. (2008). Visualizing Data Using T-SNE. *J. Machine Learn. Res.* 9, 2579–2605.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "DeepWalk: Online Learning of Social Representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 701–710.

Plank, M. (2014). *The Role of microRNAs in Allergic Airways Disease and T Cell Biology*.

Ruan, K., Fang, X., and Ouyang, G. (2009). MicroRNAs: Novel Regulators in the Hallmarks of Human Cancer. *Cancer Lett.* 285 (2), 116–126. doi:10.1016/j.canlet.2009.04.031

Rutnam, Z. J., Du, W. W., Yang, W., Yang, X., and Yang, B. B. (2014). The Pseudogene TUSC2P Promotes TUSC2 Function by Binding Multiple microRNAs. *Nat. Commun.* 5, 2914. doi:10.1038/ncomms3914

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

Santulli, G. (2015). *MicroRNA : Medical Evidence : From Molecular Biology to Clinical Practice*.

Setoyama, T., Ling, H., Natsugoe, S., and Calin, G. A. (2011). Non-coding RNAs for Medical Practice in Oncology. *Keio J. Med.* 60 (4), 106–113. doi:10.2302/kjm.60.106

Shi, X., Nie, F., Wang, Z., and Sun, M. (2015). Pseudogene-expressed RNAs: a New Frontier in Cancers. *Tumor Biol.* 37, 1471–1478. doi:10.1007/s13277-015-4482-z

Song, X.-Y., Liu, T., Qiu, Z.-Y., You, Z.-H., Sun, Y., Jin, L.-T., et al. (2020). Prediction of lncRNA-Disease Associations from Heterogeneous Information Network Based on DeepWalk Embedding Model. ICIC 12465.

Stiegelbauer, V., Perakis, S. O., Deutsch, A., Ling, H., Gerger, A., and Pichler, M. (2014). MicroRNAs as Novel Predictive Biomarkers and Therapeutic Targets in Colorectal Cancer. *Wjg* 20 (33), 11727–11735. doi:10.3748/wjg.v20.i33.11727

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). "LINE: Large-Scale Information Network Embedding," in Proceedings of the 24th International Conference on World Wide Web, 1067–1077.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity Network Fusion for Aggregating Data Types on a Genomic Scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810

Würfel, F. M., Wirtz, R. M., Winterhalter, C., Taffurelli, M., Santini, D., Mandrioli, A., et al. (2020). HLA-J, a Non-pseudogene as a New Prognostic Marker for Therapy Response and Survival in Breast Cancer. *Geburtshilfe Frauenheilkd* 80 (11), 1123–1133. doi:10.1055/a-1128-6664

Xu, J., Cai, L., Liao, B., Zhu, W., Wang, P., Meng, Y., et al. (2019). Identifying Potential miRNAs-Disease Associations with Probability Matrix Factorization. *Front. Genet.* 10, 1234. doi:10.3389/fgene.2019.01234

Xuan, P., Pan, S., Zhang, T., Liu, Y., and Sun, H. (2019). Graph Convolutional Network and Convolutional Neural Network Based Method for Predicting lncRNA-Disease Associations. *Cells* 8 (9), 1012. doi:10.3390/cells8091012

Yuan, H., Tu, S., Ma, Y., and Sun, Y. (2021). Downregulation of lncRNA RPLP0P2 Inhibits Cell Proliferation, Invasion and Migration, and Promotes Apoptosis in Colorectal Cancer. *Mol. Med. Rep.* 23 (5), 309. doi:10.3892/mmr.2021.11948

Zhang, Z.-C., Zhang, X.-F., Wu, M., Ou-Yang, L., Zhao, X.-M., and Li, X.-L. (2020). A Graph Regularized Generalized Matrix Factorization Model for Predicting Links in Biomedical Bipartite Networks. *Bioinformatics* 36 (11), 3474–3481. doi:10.1093/bioinformatics/btaa157

Zhang, Z., Liu, Z.-B., Ren, W.-M., Ye, X.-G., and Zhang, Y.-Y. (2012). The miR-200 Family Regulates the Epithelial-Mesenchymal Transition Induced by EGF/EGFR in Anaplastic Thyroid Cancer Cells. *Int. J. Mol. Med.* 30 (4), 856–862. doi:10.3892/ijmm.2012.1059

Zheng, L.-L., Zhou, K.-R., Liu, S., Zhang, D.-Y., Wang, Z.-L., Chen, Z.-R., et al. (2018). dreamBase: DNA Modification, RNA Regulation and Protein Binding of Expressed Pseudogenes in Human Health and Disease. *Nucleic Acids Res.* 46 (D1), D85–D91. doi:10.1093/nar/gkx972

Zheng, X., Ding, H., Mamitsuka, H., and Zhu, S. (2013). "Collaborative Matrix Factorization with Multiple Similarities for Predicting Drug-Target Interactions," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 1025–1033. doi:10.1145/2487575.2487670

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 January 2022
doi: 10.3389/fgene.2021.832627

A Computational Framework to Identify Biomarkers for Glioma Recurrence and Potential Drugs Targeting Them

Shuzhi Ma[1,2†], *Zhen Guo*[3,4†], *Bo Wang*[5], *Min Yang*[5], *Xuelian Yuan*[5], *Binbin Ji*[5], *Yan Wu*[5] *and Size Chen*[1,6,7*]

[1]Department of Oncology, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [2]Department of Pathology, Zhujiang Hospital, Southern Medical University, Guangzhou, China, [3]Academician Workstation, Changsha Medical University, Changsha, China, [4]Hunan Key Laboratory of the Research and Development of Novel Pharmaceutical Preparations, Changsha Medical University, Changsha, China, [5]Geneis (Beijing) Co., Ltd., Beijing, China, [6]Guangdong Provincial Engineering Research Center for Esophageal Cancer Precise Therapy, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [7]Central Laboratory, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Xiangzheng Fu,
Hunan University, China
Taigang Liu,
Shanghai Ocean University, China

***Correspondence:**
Size Chen
chensize@gdpu.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: *10 December 2021*
Accepted: *29 December 2021*
Published: *17 January 2022*

Citation:
Ma S, Guo Z, Wang B, Yang M, Yuan X, Ji B, Wu Y and Chen S (2022) A Computational Framework to Identify Biomarkers for Glioma Recurrence and Potential Drugs Targeting Them. Front. Genet. 12:832627. doi: 10.3389/fgene.2021.832627

Background: Recurrence is still a major obstacle to the successful treatment of gliomas. Understanding the underlying mechanisms of recurrence may help for developing new drugs to combat gliomas recurrence. This study provides a strategy to discover new drugs for recurrent gliomas based on drug perturbation induced gene expression changes.

Methods: The RNA-seq data of 511 low grade gliomas primary tumor samples (LGG-P), 18 low grade gliomas recurrent tumor samples (LGG-R), 155 glioblastoma multiforme primary tumor samples (GBM-P), and 13 glioblastoma multiforme recurrent tumor samples (GBM-R) were downloaded from TCGA database. DESeq2, key driver analysis and weighted gene correlation network analysis (WGCNA) were conducted to identify differentially expressed genes (DEGs), key driver genes and coexpression networks between LGG-P *vs* LGG-R, GBM-P *vs* GBM-R pairs. Then, the CREEDS database was used to find potential drugs that could reverse the DEGs and key drivers.

Results: We identified 75 upregulated and 130 downregulated genes between LGG-P and LGG-R samples, which were mainly enriched in human papillomavirus (HPV) infection, PI3K-Akt signaling pathway, Wnt signaling pathway, and ECM-receptor interaction. A total of 262 key driver genes were obtained with frizzled class receptor 8 (*FZD8*), guanine nucleotide-binding protein subunit gamma-12 (*GNG12*), and G protein subunit β2 (*GNB2*) as the top hub genes. By screening the CREEDS database, we got 4 drugs (Paclitaxel, 6-benzyladenine, Erlotinib, Cidofovir) that could downregulate the expression of up-regulated genes and 5 drugs (Fenofibrate, Oxaliplatin, Bilirubin, Nutlins, Valproic acid) that could upregulate the expression of down-regulated genes. These drugs may have a potential in combating recurrence of gliomas.

Conclusion: We proposed a time-saving strategy based on drug perturbation induced gene expression changes to find new drugs that may have a potential to treat recurrent gliomas.

Keywords: low grade gliomas, RNA-seq, differentially expressed genes, WGCNA, key driver genes, drug discovery

INTRODUCTION

Gliomas are the most common type of central nervous system (CNS) tumors, which are composed of various distinct subtype tumors (Galbraith and Snuderl, 2021). Differed from non-CNS neoplasms, the gliomas' grading system is quite complicate. The latest 2021 WHO Classification of Tumors of the Central Nervous System (WHO CNS5) has integrated certain molecular markers and histological features for more accurate staging of gliomas (Louis et al., 2021). Briefly, the gliomas can be classified as two categories, low-grade gliomas (grade 1–2) with a relatively benign slow-growing feature and favorable prognosis, and high-grade gliomas (grade 3–4) with highly infiltrative ability and malignant form. Glioblastoma multiforme (GBM) accounts for approximately 55% of gliomas and is considered as the most aggressive type of gliomas with a 5-years survival rate less than 5% and a median survival time of 12–15 months (Ostrom et al., 2015a; Ostrom et al., 2015b). Currently, the conventional therapeutic regimen for gliomas is surgical resection followed by radiotherapy and chemotherapy. However, the curative effect is far from satisfaction and recurrence is still the major obstacle to the success of chemoradiotherapy since the majority of GBM would experience recurrence within 6.2 months after diagnosis (Bähr et al., 2009; King and Benhabbour, 2021). Therefore, it is imperative to find new therapeutic target and novel therapeutic strategies for patient with gliomas.

Over the past decades, molecular biomarkers have gained important value in providing diagnostic information and therapeutic target for gliomas. Bevacizumab, an inhibitor of vascular endothelial growth factor (*VEGF*), was approved to treat recurrent GBM by the Food and Drug Administration (FDA) in March 2009 (Friedman et al., 2009; Kreisl et al., 2009). By targeting *VEGF*, Bevacizumab inhibits angiogenesis and blocks the nutrient supply, which ultimately impedes the growth and metastasis of GBM. In addition, methylation of the O^6 -methylguanine-DNA methyltransferase (*MGMT*) promoter might serve as a predictive marker for temozolomide (TMZ) treatment response of GBM (Hegi et al., 2008). Poly (ADP ribose) polymerase (*PARP*) inhibitors can increase tumor sensitivity to TMZ chemotherapy and synergize with radiation therapy (Lesueur et al., 2018). A novel nano-compounds encapsulating wild-type p53 (SGT-53) could enhance the inhibitory effects of TMZ on TMZ-resistant GBM cells (Kim et al., 2015). Moreover, mutations/deregulation in the platelet-derived growth factor receptor alpha (*PDGFRα*), telomerase reverse transcriptase (*TERT*), epidermal growth factor receptor (*EGFR*), c-Myc, phosphatase and tensin homolog (*PTEN*), serine/threonine-protein kinase (*BRAF*) are frequently observed in glioma, which have become attractive markers for targeted therapy [Mukasa et al., 2010; Sampson et al., 2010; Killela et al., 2013; Johnson et al., 2014; Liu et al., 2020a; Liu et al., 2020b)].

Though a handful of molecular biomarkers have been discovered, the targeted therapies in clinical trials displayed limited curative effect for gliomas (Wu et al., 2021). This may be attributed to the inter- and intra-heterogeneity in driver mutations and plasticity of gliomas. The recurrent tumor might have a totally distinct gene expression signature in comparison with the primary tumor. Notably, 90% of druggable targets identified at initial diagnosis of gliomas are differentially expressed in a recurrent tumor (Schäfer et al., 2019). Ideally, a patient would select the specific drug according to his/her own molecular genetic feature and change the drugs over the course as the tumor evolves. With the advent of multiomics era, it is becoming possible.

In the era of big data, genome-wide molecular profiling at genome, transcriptome, proteome, and metabolome level have revealed comprehensive landscapes for all major types of gliomas. This has not only enriched our understanding of the molecular mechanism of gliomas pathogenesis and progression, but also broadened our ideas in discovering new therapeutic drugs. As we all know, new drug developing is a time-consuming course with high capital input and low yield, and drug repositioning based on computational tools could largely shorten the process (Liu et al., 2016; Xu et al., 2019; Zhou et al., 2019; Yang et al., 2020a; Liu et al., 2020; Tang et al., 2020; Zhou et al., 2020; Peng et al., 2021). In the present study, we proposed a fast, economical, and comprehensive strategy to find old drugs with new function for combating recurrence of gliomas. Our results on TCGA data suggested that this strategy provides a new direction in discovering drugs and brings hope for people in treating gliomas.

MATERIALS AND METHODS

Data Collecting and Grouping

The RNA sequencing data of gliomas samples were downloaded from TCGA database. The samples were divided into 4 groups: low grade gliomas primary tumor samples (LGG-P, $n = 511$), low grade gliomas recurrent tumor samples (LGG-R, $n = 18$), GBM primary tumor samples (GBM-P, $n = 155$), and GBM recurrent tumor samples (GBM-R, $n = 13$).

Screening of DEGs in Gliomas

The R package DESeq2 was applied to identify DEGs in the following data pairs: LGG-P *vs* LGG-R, GBM-P *vs* GBM-R. $|\log2\text{fold change (FC)}| \geq 2$, false discovery rate (FDR) < 0.5 and adjusted p value < 0.001 were set as threshold. R package clusterProfiler was used for Gene Ontology (GO) enrichment analysis and calculations. The enriched pathways in the up- or down-regulated gene set were generated using R package ggplot2.

Weighted Gene Correlation Network Analysis and Key Driver Analysis

Network-based methods have been widely used to analyze the associations between various biological entities (Chen et al., 2018; Peng et al., 2018; Peng et al., 2020; Zhang et al., 2021). The R package WGCNA was used to construct a weighted gene co-expression network. The key driver analysis was performed using a software package described by Yang et al.(Yang et al., 2016). The first step was to generate a subnetwork NG, which is located within 2 steps of nodes in a given gene set. Next, the dynamic neighborhood search (DNS) was used to find the gene within 2 steps of each gene in NG. Lastly, by taking the gene set in the first step as the background, the hypergeometric test is carried out to calculate the enrichment value between the gene set in the second step and the input gene set. p value < 0.05 for DEG and p value < 0.01 for subnet were set as threshold in the key driver analysis.

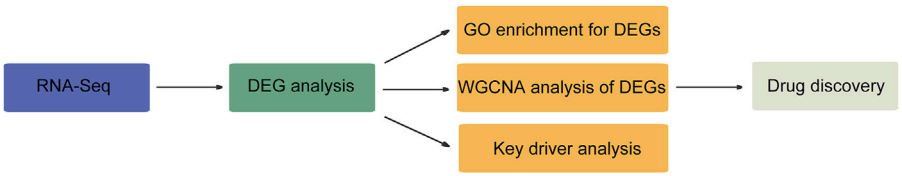


FIGURE 1 | The framework of this study. The RNA-seq data was used to find DEGs, which was followed by GO enrichment, WGCNA, and key driver analysis. Potential drugs that could reverse the DEGs and key drivers were screened through the CREEDS database.

Drug Discovery

Based on the DEGs and key drivers, the CREEDS database was used to find potential drugs. The CREEDS database contains 906 drug perturbation gene expression signatures collected from GEO database (Wang et al., 2016). We screened the drug-gene pairs to identify potential drugs that could reverse the expression of DEGs and key drivers in gliomas. p value < 10^{-10} was set as the threshold.

RESULTS

A Computational Framework to Identify Biomarkers for Glioma Recurrence and Potential Drugs Targeting Them

We proposed a computational framework biomarker identification and drug discovering for glioma recurrence **(Figure 1)**. Firstly, the differentially expressed genes (DEGs) of primary gliomas samples and recurrent gliomas samples were identified from RNA sequencing data downloaded from The Cancer Genome Atlas (TCGA) database. Secondly, weighted gene correlation network (WGCNA) analysis and key driver analysis were conducted to find co-expression modules and key driver genes. Thirdly, the CREEDS database was applied to find potential drugs that could reverse the DEGs and key drivers. We then applied this framework to the downloaded TCGA data and identified important genes involving in glioma recurrence and potential drugs targeting them.

Many DEGs Were Identified Between LGG-P and LGG-R, and Between GBM-P and GBM-R

We conducted a comprehensive analysis of the DEGs between LGG-P and LGG-R, and between GBM-P and GBM-R. The difference between GBM-P and GBM-R samples was not significant as we only obtained 2 upregulated and 29 downregulated genes. A total of 205 DEGs with 75 upregulated and 130 downregulated genes were identified between LGG-P and LGG-R samples. The specific details of each DEGs were shown in **Supplementary Table S1**. Since the number of DEGs between GBM-P and GBM-R was not large enough, we chose LGG-P and LGG-R pairs for further study. We randomly selected 25 samples to draw the heat map and the top 10 differentially expressed genes were shown in **Figure 2**.

GO analysis was utilized to annotate the function of DEGs between LGG-P and LGG-R. The upregulated genes could not be enriched owing to the relatively large p value. The downregulated

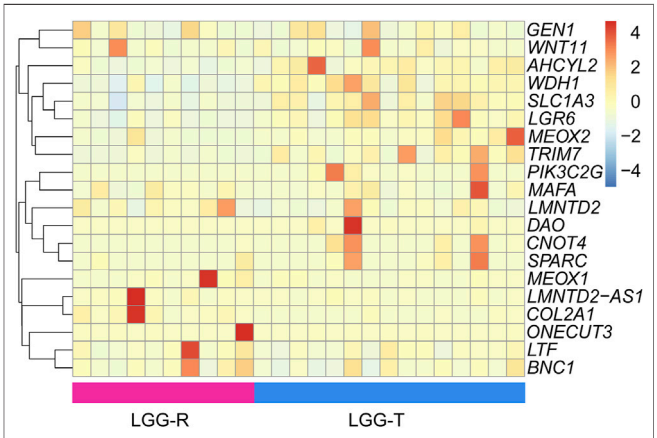


FIGURE 2 | The heat map of the DEGs between LGG-P and LGG-R samples. Dao, mdh1, slc1a3, CNOT4, meox2, pik3c2g, LGR6, ahcyl2, SPARC and LTF were the top 10 differentially upregulated genes; LMNTD2-AS1, MAFA, COL2A1, ONECUT3, MEOX1, BNC1, LMNTD2, GEN1, Wnt11 and trim7 were the top 10 differentially down regulated genes.

genes were predominantly enriched in DNA-binding transcription activator activity, extracellular matrix structural constituent, growth factor binding, and platelet-derived growth factor binding for the molecular function (MF) category (**Figure 3A**). For the cellular component (CC) category, the downregulated genes were correlated with extracellular matrix, collagen-containing extracellular matrix, endoplasmic reticulum lumen, and collagen trimer (**Figure 3B**). For the biological process (BP) category, the downregulated genes were mainly involved in skeletal system development, extracellular matrix organization, extracellular structure organization, and connective tissue development (**Figure 3C**). KEGG enrichment analysis was further performed to explore the underling pathological pathways for LGG. As shown in **Figure 4**, the enrichment pathways include human papillomavirus infection, PI3K-Akt signaling pathway, Wnt signaling pathway, ECM-receptor interaction, and proteoglycans in cancer.

Coexpression Analysis Revealed Chemical Synaptic Transmission Pathway and T Cell Activation Pathway as Key Modules for Glioma Recurrence

To better understand the function of differentially expressed genes, WGCNA was carried out to identify highly correlated

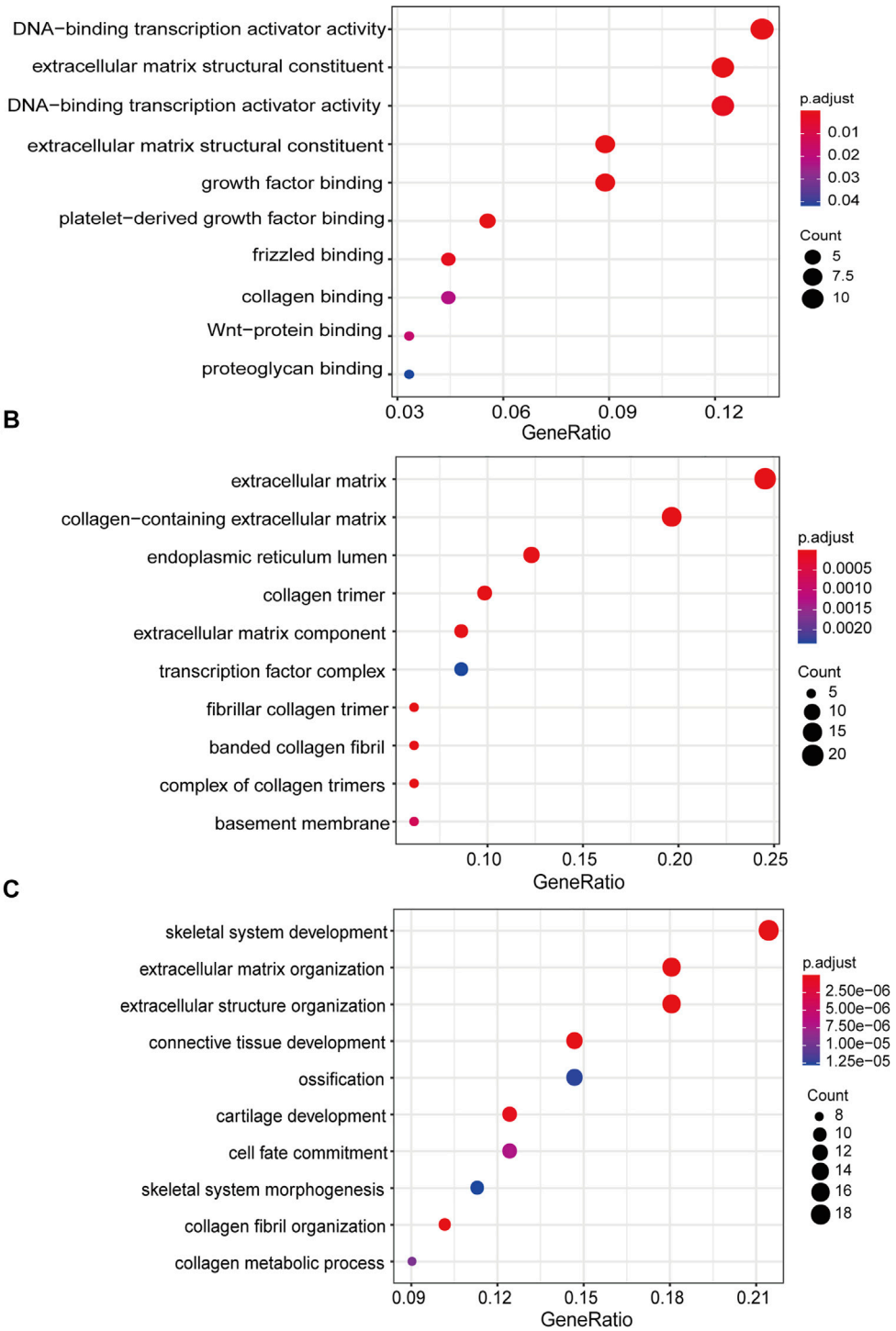


FIGURE 3 | GO analysis of the downregulated genes between LGG-P and LGG-R. **(A)** molecular function category; **(B)** cellular component category; **(C)** biological process category. The *X*-axis is the ratio of differentially expressed genes enriched in the corresponding pathway, and the *Y*-axis is the name of the pathway.

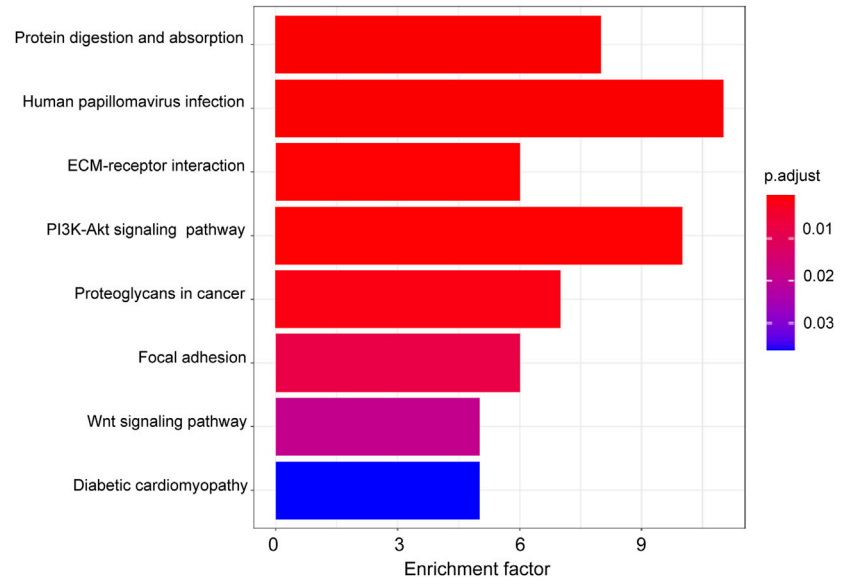


FIGURE 4 | KEGG analysis of the downregulated genes between LGG-P and LGG-R. The DEGs were mainly enriched in protein digestion and absorption, human papillomavirus infection, ECM-receptor interaction, PI3K-Akt signaling pathway, and proteoglycans in cancer pathways.

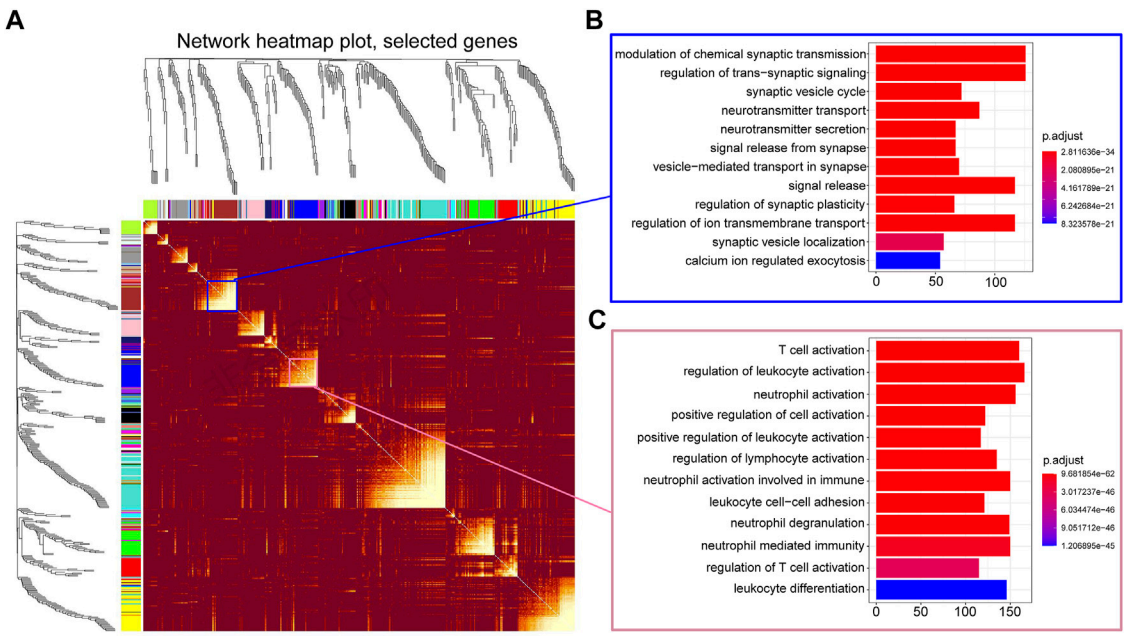


FIGURE 5 | Highly correlated gene clusters were identified by WGCNA. **(A)** Topological overlapping heat map of 400 genes. **(B)** GO enrichment analysis of module 5. **(C)** GO enrichment analysis of module 8.

gene clusters. Genes with zero expression were deleted from all samples, and the samples of some separated groups were removed from the hierarchical clustering results. WGCNA finally yielded 150 significant gene modules in LGG-P group and 65 gene modules in LGG-R group. Since there are too many genes to visualize, we randomly selected 400 genes to construct a topological overlapping heat map (**Figure 5A**) and performed functional enrichment analysis. As shown in **Figure 5B** and **Figure 5C**, the highly coexpression genes were mainly enriched in chemical synaptic transmission pathway and T cell activation pathway.

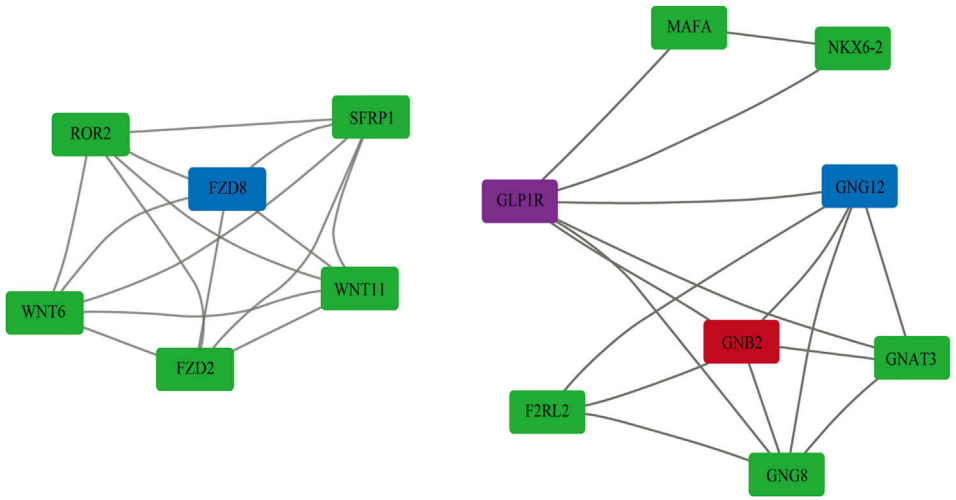


FIGURE 6 | A subnetwork of the key drivers that connected the up and down regulated genes. Purple indicates upregulated genes, red indicates key drivers of upregulated genes, green indicates downregulated genes, and blue indicates key drivers of downregulated genes.

Many Genes Including *FZD8* and *GNG12* Were Identified as Key Driver Genes for Glioma Recurrence

Key driver gene was considered as the hub gene that connected the up- or downregulated genes. For LGG-P *vs* LGG-R pair, we obtained 2 key drivers in the upregulated gene set and 260 key drivers in the downregulated gene set. The detailed information of the key drivers can be found in **Supplementary Table S2**. The most siginificant key drivers and their corresponding subnetwork were shown in **Figure 6**. We demonstrated that frizzled class receptor 8 (*FZD8*) and guanine nucleotide-binding protein subunit gamma-12 (*GNG12*) were two of the hub gene of the downregulated genes. *FZD8* is a G protein-coupled receptor protein that plays an important role in β-catenin signaling pathway and regulates cancer invasion and metastasis (Li et al., 2017). *GNG12* is a member of the G protein family and participate in a handful of trans-membrane signal transducer pathways (Yuan et al., 2021). G protein subunit β2 (*GNB2*) is a hub gene of up regulated genes and belongs to the guanine nucleotide-binding proteins family. *GNB2* may activate the canonical G protein signaling and involved in cancer initiation and progression (O'hayre et al., 2014). The functions and implications of these hub genes in cancers will be discussed further.

A Few Drugs Including Paclitaxel and Fenofibrate Were Identified as Potential Drugs for Preventing Glioma Recurrence

Drugs that have a potential to reverse the expression of DEGs may be valuable for further treatment of gliomas. We obtained 26 drugs that could perturb the expression of up regulated gene sets and 50 drugs that could perturb the expression of down regulated gene sets by CREEDS database. In consideration of drug profile and previous studies, we focused on 4 perturbation drugs of up regulated genes (Paclitaxel, 6-benzyladenine, Erlotinib, Cidofovir) and 5 perturbation drugs of down regulated genes (Fenofibrate, Oxaliplatin, Bilirubin, Nutlins, Valproic acid) (**Table 1**). These drugs may provide new insight into preventing recurrence of LGG.

DISCUSSION

Gliomas are highly malignant tumors and recurrence is still the main obstacle to treatment. Though researchers have identified a handful of biomarkers for glioma, the therapeutic effect of targeted drugs is far from satisfaction. Currently, our study provides a fast, economical, and comprehensive method for finding potential drugs to treat gliomas.

Based on RNA sequencing data of LGG-P and LGG-R samples from TCGA, we yielded 75 upregulated and 130 downregulated genes, which were predominantly correlated with human papillomavirus (HPV) infection, PI3K-Akt signaling pathway, Wnt signaling pathway, ECM-receptor interaction. HPV infection is a major cause of cervical cancer, and associated with several epithelial malignancies, including oral cavity, anal, oropharyngeal, penile, vulvar, vaginal, and laryngeal cancers (Lu et al., 2020). Until now, there is no direct evidence indicates that HPV infection is involved in gliomas. We speculate HPV infection may interfere host immune system and participate in LGG recurrence. PI3K-Akt signaling pathway and Wnt signaling pathway are two of the canonical signaling transduction pathways in various cancers. PI3K/Akt pathway controls cell fate by regulating cell growth, apoptosis, angiogenesis, metabolism, autophagy, and chemotherapy resistance of gliomas (Shahcheraghi et al., 2020). Activation of PI3K-Akt pathway is associated with migration and invasion of glioblastoma cells (Huang et al., 2018). Wnt/beta-catenin signaling pathway plays

TABLE 1 | Potential drugs for the treatment of LGG recurrence.

Class	Drug	Antitumor mechanism	Evidence (DIO)
up	Paclitaxel	prevents mitosis, blocks cell cycle progression, and inhibits cell growth.	10.1186/s11658-019-0164-y
up	6-benzyladenine	stimulates cell division, and inhibits respiratory kinase, leading to plant growth and development.	10.1111/plb.13154
up	Erlotinib	inhibits tyrosine kinase activity, blocks EGFR signaling pathway	10.1016/bs.podrm.2019.10.004
up	Cidofovir	inhibits viral DNA polymerase	10.1542/peds.2019-1632
down	Fenofibrate	activates PPARα-RXR signaling	10.7150/jca.24488
down	Oxaliplatin	blocks DNA replication	10.1016/j.ctrv.2020.102112
down	Bilirubin	an endogenous metabolite from haem	10.1021/np4005807
down	Nutlins	binds to p53/MDM2 complex and displace p53 protein	10.2174/138161210791033932
down	Valproic acid	inhibits histone deacetylase (HDAC)	10.2174/1574892810666150317144511

a vital role in ionizing radiation-induced invasion of glioblastoma cells (Dong et al., 2015). It has been reported that high level of beta-catenin was associated with a poor prognosis in glioblastoma patients (Gao et al., 2017). Currently, a number of PI3K inhibitors and wnt inhibitors have entered clinical trials for gliomas treatment, such as BKM120, XL147 and XL765 (Lee et al., 2016; Zhao et al., 2017). ECM-receptor interaction pathway mediates cell migration by regulating neovascularization and diffuse infiltration of tumor cells (Cui et al., 2018). Previous study also indicated ECM-receptor interaction pathway was abnormal in development and survival of glioblastoma (Bo et al., 2015; Yang et al., 2020b).

We also identified several key driver genes that contributed to LGG recurrence, including *FZD8, GNG12, GNB2*. *FZD8* could activate the β-catenin pathway and play a vital role in cancer invasion and metastasis (Chen et al., 2020). Aberrant expression of *FZD8* has been reported in gastric cancer, prostate cancer, renal cell carcinoma, lung cancer, pancreatic adenocarcinoma, and overexpression of *FZD8* was considered to promote tumor metastasis (Li et al., 2017; Yang et al., 2017; Liu et al., 2019; Chen et al., 2020; Li et al., 2021). In addition, overexpression of *FZD8* leaded to chemotherapy resistance in breast cancer patients (Yin et al., 2013). *GNG12* acted as an important modulator or transducer in various transmembrane signaling systems. Researchers have demonstrated that *GNG12* could regulate cancer cell proliferation, inflammatory response, and immune response *via* activating the mTORC1 pathway and NF-κB signaling pathway (Larson et al., 2010; Luo et al., 2018; Li et al., 2020). *GNB2* was involved in cancer initiation and progression by activating AKT/mTOR pathway, MAPK pathway, and Hippo signaling pathway. Mutations of *GNB2* may result in targeted kinase inhibitors resistance to numerous types of cancer (Yoda et al., 2015). The roles of *FZD8, GNG12, GNB2* have not been fully illustrated in gliomas and needs further investigation. These key driver genes may help for understanding the pathogenesis of for LGG recurrence and shed new insight for developing new drugs.

However, developing new drugs from a molecular biomarker is a great project and still has a long way to go. In the present study, a total of 9 drugs with potential therapeutic effect against LGG recurrence were selected through a drug-gene perturbation method. Paclitaxel is a natural anticancer drug that has been widely used in the therapy of breast cancer, ovarian cancer, lung cancer, and several head and neck cancers. Paclitaxel binds to tubulin, promotes its assembly with microtubules and inhibits dissociation, which finally prevents mitosis and hinders cell cycle progression (Zhu and Chen, 2019). Erlotinib is used to treat some types of lung cancer and advanced or metastatic pancreatic cancer in clinical. Erlotinib blocks EGFR pathway by inhibiting tyrosine kinase activity and impeding cell proliferation, apoptosis, angiogenesis, invasions, and metastasis. Benzylaminopurine is a first generation cytokinin that stimulates cell division and inhibits respiratory kinase, leading to plant growth and development. Cidofovir is used to treat cytomegalovirus (CMV) infection through inhibiting viral DNA polymerase (Alcamo et al., 2020). Cidofovir also has a strong activity against herpes simplex virus (HSV), varicella zoster virus (VZV), adenovirus (AV), and human papillomavirus (HPV). Fenofibrate is widely used as a lipid-lowering drug through activating PPARα-RXR signal and transcription of lipid metabolism related genes. Numerous evidence has indicated that fenofibrate might exert anticancer effects through regulating cell apoptosis, cell-cycle arrest, invasion, and migration (Lian et al., 2018). Oxaliplatin is a third-generation platinum analog that has been widely used as the first-line drugs for metastatic colorectal cancer. It blocks DNA replication by binding to DNA and forming cross-linked DNA adducts, which consequently leading to cancer cell death (Mauri et al., 2020). Bilirubin is an endogenous metabolite from haem. Recent studies have indicated bilirubin levels may serve as biomarker for several cancers and vascular disease (Horsfall et al., 2020; Seyed Khoei et al., 2020). Nutlins is a small molecule that could displace p53 protein from p53/MDM2 complex, thereby preventing the degradation of p53. It has been revealed that Nutlins could induce p53 dependent cell cycle arrest and apoptosis in a number of tumors (Impicciatore et al., 2010). Valproic acid (VPA), a histone deacetylase (HDAC) inhibitor, is widely used to treat epilepsia, bipolar disorders, migraine, and schizophrenia. In addition, VPA may exert anti-tumor activity by regulating cell proliferation, apoptosis, differentiation, adhesion, invasion, migration, angiogenesis, and inflammation (Michaelis et al., 2007). As shown above, some candidate drugs were already commonly used in clinical, some candidates only showed preclinical antitumor activity. Weather these drugs/agents

could prevent LGG recurrence still needs further preclinical and clinical trial validation.

In conclusion, we conducted a comprehensive analysis of LGG-P and LGG-R samples to find DEGs and key driver genes. By using a drug-gene perturbation method, a serious potential drugs/agents were screened to treat LGG recurrence. However, the exact effect of these drugs on glioma recurrence needs further experimental data for verification. Besides, independent dataset with paired primary tumor and recurrent tumor samples is needed to validate the findings in the future. This study may broaden our understanding of the molecular mechanism of LGG recurrence and provide new sights for drug discovery.

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

SC contributed to conception and design of the study. BW, MY, and XY organized the database. SM, ZG, and BW performed the statistical analysis. SM and ZG wrote the first draft of the manuscript. MY, XY, BJ, and YW wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of China (No. 82104309), Natural Science Foundation of Hunan Province (No.2021JJ40639), Outstanding Youth Project of Hunan Provincial Education Department (No.20B075). Science and Technology Program of Guangzhou, China (Grant numbers: 201904010047); Science and Technology Planning Project of Guangdong Province of China (Grant numbers: 2020A0505100058); Guangdong Educational Committee (Key Project of Regular institutions of higher learning of Guangdong Province (Grant numbers: 2019KZDXM024).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.832627/full#supplementary-material

Supplementary Table S1 | DEGs discovered in LGG-P and LGG-R samples.

Supplementary Table S2 | Key drivers discovered in LGG-P and LGG-R samples.

REFERENCES

Alcamo, A. M., Wolf, M. S., Alessi, L. J., Chong, H. J., Green, M., Williams, J. V., et al. (2020). Successful Use of Cidofovir in an Immunocompetent Child with Severe Adenoviral Sepsis. *Pediatrics* 145. doi:10.1542/peds.2019-1632

Bähr, O., Herrlinger, U., Weller, M., and Steinbach, J. P. (2009). Very Late Relapses in Glioblastoma Long-Term Survivors. *J. Neurol.* 256, 1756–1758. doi:10.1007/s00415-009-5167-6

Bo, L. J., Wei, B., Li, Z. H., Wang, Z. F., Gao, Z., and Miao, Z. (2015). Bioinformatics Analysis of miRNA Expression Profile between Primary and Recurrent Glioblastoma. *Eur. Rev. Med. Pharmacol. Sci.* 19, 3579–3586.

Chen, W., Liu, Z., Mai, W., Xiao, Y., You, X., and Qin, L. (2020). FZD8 Indicates a Poor Prognosis and Promotes Gastric Cancer Invasion and Metastasis via B-Catenin Signaling Pathway. *Ann. Clin. Lab. Sci.* 50, 13–23.

Chen, X., Yin, J., Qu, J., and Huang, L. (2018). MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-Disease Association Prediction. *Plos Comput. Biol.* 14 (8), e1006418. doi:10.1371/journal.pcbi.1006418

Cui, X., Morales, R.-T. T., Qian, W., Wang, H., Gagner, J.-P., Dolgalev, I., et al. (2018). Hacking Macrophage-Associated Immunosuppression for Regulating Glioblastoma Angiogenesis. *Biomaterials* 161, 164–178. doi:10.1016/j.biomaterials.2018.01.053

Dong, Z., Zhou, L., Han, N., Zhang, M., and Lyu, X. (2015). Wnt/β-catenin Pathway Involvement in Ionizing Radiation-Induced Invasion of U87 Glioblastoma Cells. *Strahlenther Onkol* 191, 672–680. doi:10.1007/s00066-015-0858-7

Friedman, H. S., Prados, M. D., Wen, P. Y., Mikkelsen, T., Schiff, D., Abrey, L. E., et al. (2009). Bevacizumab Alone and in Combination with Irinotecan in Recurrent Glioblastoma. *Jco* 27, 4733–4740. doi:10.1200/jco.2008.19.8721

Galbraith, K., and Snuderl, M. (2021). Molecular Pathology of Gliomas. *Surg. Pathol. Clin.* 14, 379–386. doi:10.1016/j.path.2021.05.003

Gao, L., Chen, B., Li, J., Yang, F., Cen, X., Liao, Z., et al. (2017). Wnt/β-catenin Signaling Pathway Inhibits the Proliferation and Apoptosis of U87 Glioma Cells via Different Mechanisms. *PLoS One* 12, e0181346. doi:10.1371/journal.pone.0181346

Hegi, M. E., Liu, L., Herman, J. G., Stupp, R., Wick, W., Weller, M., et al. (2008). Correlation of O6-Methylguanine Methyltransferase (MGMT) Promoter Methylation with Clinical Outcomes in Glioblastoma and Clinical Strategies to Modulate MGMT Activity. *Jco* 26, 4189–4199. doi:10.1200/jco.2007.11.5964

Horsfall, L. J., Burgess, S., Hall, I., and Nazareth, I. (2020). Genetically Raised Serum Bilirubin Levels and Lung Cancer: a Cohort Study and Mendelian Randomisation Using UK Biobank. *Thorax* 75, 955–964. doi:10.1136/thoraxjnl-2020-214756

Huang, W., Ding, X., Ye, H., Wang, J., Shao, J., and Huang, T. (2018). Hypoxia Enhances the Migration and Invasion of Human Glioblastoma U87 Cells through PI3K/Akt/mTOR/HIF-1α Pathway. *Neuroreport* 29, 1578–1585. doi:10.1097/wnr.0000000000001156

Impicciatore, G., Sancilio, S., Miscia, S., and Di Pietro, R. (2010). Nutlins and Ionizing Radiation in Cancer Therapy. *Cpd* 16, 1427–1442. doi:10.2174/138161210791033932

Johnson, B. E., Mazor, T., Hong, C., Barnes, M., Aihara, K., Mclean, C. Y., et al. (2014). Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science* 343, 189–193. doi:10.1126/science.1239947

Killela, P. J., Reitman, Z. J., Jiao, Y., Bettegowda, C., Agrawal, N., Diaz, L. A., Jr., et al. (2013). TERT Promoter Mutations Occur Frequently in Gliomas and a Subset of Tumors Derived from Cells with Low Rates of Self-Renewal. *Proc. Natl. Acad. Sci.* 110, 6021–6026. doi:10.1073/pnas.1303607110

Kim, S.-S., Rait, A., Kim, E., Pirollo, K. F., and Chang, E. H. (2015). A Tumor-Targeting P53 Nanodelivery System Limits Chemoresistance to Temozolomide Prolonging Survival in a Mouse Model of Glioblastoma Multiforme. *Nanomedicine: Nanotechnology, Biol. Med.* 11, 301–311. doi:10.1016/j.nano.2014.09.005

King, J. L., and Benhabbour, S. R. (2021). Glioblastoma Multiforme-A Look at the Past and a Glance at the Future. *Pharmaceutics* 13. doi:10.3390/pharmaceutics13071053

Kreisl, T. N., Kim, L., Moore, K., Duic, P., Royce, C., Stroud, I., et al. (2009). Phase II Trial of Single-Agent Bevacizumab Followed by Bevacizumab Plus Irinotecan at

Tumor Progression in Recurrent Glioblastoma. *Jco* 27, 740–745. doi:10.1200/jco.2008.16.3055

Larson, K. C., Draper, M. P., Lipko, M., and Dabrowski, M. (2010). Gng12 Is a Novel Negative Regulator of LPS-Induced Inflammation in the Microglial Cell Line BV-2. *Inflamm. Res.* 59, 15–22. doi:10.1007/s00011-009-0062-2

Lee, Y., Lee, J.-K., Ahn, S. H., Lee, J., and Nam, D.-H. (2016). WNT Signaling in Glioblastoma and Therapeutic Opportunities. *Lab. Invest.* 96, 137–150. doi:10.1038/labinvest.2015.140

Lesueur, P., Chevalier, F., El-Habr, E. A., Junier, M.-P., Chneiweiss, H., Castera, L., et al. (2018). Radiosensitization Effect of Talazoparib, a Parp Inhibitor, on Glioblastoma Stem Cells Exposed to Low and High Linear Energy Transfer Radiation. *Sci. Rep.* 8, 3664. doi:10.1038/s41598-018-22022-4

Li, J., Jin, C., Zou, C., Qiao, X., Ma, P., Hu, D., et al. (2020). GNG12 Regulates PD-L1 Expression by Activating NF-κB Signaling in Pancreatic Ductal Adenocarcinoma. *FEBS Open Bio* 10, 278–287. doi:10.1002/2211-5463.12784

Li, Q., Ye, L., Zhang, X., Wang, M., Lin, C., Huang, S., et al. (2017). FZD8, a Target of P53, Promotes Bone Metastasis in Prostate Cancer by Activating Canonical Wnt/β-Catenin Signaling. *Cancer Lett.* 402, 166–176. doi:10.1016/j.canlet.2017.05.029

Li, Y., Liu, Z., and Zhang, Y. (2021). Expression and Prognostic Impact of FZDs in Pancreatic Adenocarcinoma. *BMC Gastroenterol.* 21, 79. doi:10.1186/s12876-021-01643-6

Lian, X., Wang, G., Zhou, H., Zheng, Z., Fu, Y., and Cai, L. (2018). Anticancer Properties of Fenofibrate: A Repurposing Use. *J. Cancer* 9, 1527–1537. doi:10.7150/jca.24488

Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., et al. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Mol. Ther. - Nucleic Acids* 21, 676–686. doi:10.1016/j.omtn.2020.07.003

Liu, F., Peng, L., Tian, G., Yang, J., Chen, H., Hu, Q., et al. (2020a). 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, R., Chen, Y., Shou, T., Hu, J., and Qing, C. (2019). miRNA-99b-5p Targets FZD8 to Inhibit Non-small Cell Lung Cancer Proliferation, Migration and Invasion. *Ott* 12, 2615–2621. doi:10.2147/ott.s199196

Liu, X., Lang, J., Li, S., Wang, Y., Peng, L., Wang, W., et al. (2020b). Fragment Enrichment of Circulating Tumor DNA with Low-Frequency Mutations. *Front. Genet.* 11, 147. doi:10.3389/fgene.2020.00147

Liu, X., Yang, J., Zhang, Y., Fang, Y., Wang, F., Wang, J., et al. (2016). A Systematic Study on Drug-Response Associated Genes Using Baseline Gene Expressions of the Cancer Cell Line Encyclopedia. *Sci. Rep.* 6, 22811. doi:10.1038/srep22811

Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a Summary. *Neuro Oncol.* 23, 1231–1251. doi:10.1093/neuonc/noab106

Lu, Y., Li, P., Luo, G., Liu, D., and Zou, H. (2020). Cancer Attributable to Human Papillomavirus Infection in China: Burden and Trends. *Cancer* 126, 3719–3732. doi:10.1002/cncr.32986

Luo, C., Zhao, S., Dai, W., Zheng, N., and Wang, J. (2018). Proteomic Analyses Reveal GNG12 Regulates Cell Growth and Casein Synthesis by Activating the Leu-Mediated mTORC1 Signaling Pathway. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1866, 1092–1101. doi:10.1016/j.bbapap.2018.08.013

Mauri, G., Gori, V., Bonazzina, E., Amatu, A., Tosi, F., Bencardino, K., et al. (2020). Oxaliplatin Retreatment in Metastatic Colorectal Cancer: Systematic Review and Future Research Opportunities. *Cancer Treat. Rev.* 91, 102112. doi:10.1016/j.ctrv.2020.102112

Michaelis, M., Doerr, H., and Cinatl Jr., J., Jr. (2007). Valproic Acid as Anti-cancer Drug. *Cpd* 13, 3378–3393. doi:10.2174/138161207782360528

Mukasa, A., Wykosky, J., Ligon, K. L., Chin, L., Cavenee, W. K., and Furnari, F. (2010). Mutant EGFR Is Required for Maintenance of Glioma Growth *In Vivo*, and its Ablation Leads to Escape from Receptor Dependence. *Proc. Natl. Acad. Sci.* 107, 2616–2621. doi:10.1073/pnas.0914356107

O'hayre, M., Degese, M. S., and Gutkind, J. S. (2014). Novel Insights into G Protein and G Protein-Coupled Receptor Signaling in Cancer. *Curr. Opin. Cel Biol* 27, 126–135. doi:10.1016/j.ceb.2014.01.005

Ostrom, Q. T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., et al. (2015a). CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012. *Neuro Oncol.* 17 Suppl 4 (Suppl. 4), iv1. iv1-iv62. doi:10.1093/neuonc/nov189

Ostrom, Q. T., Gittleman, H., Stetson, L., Virk, S. M., and Barnholtz-Sloan, J. S. (2015b). Epidemiology of Gliomas. *Cancer Treat. Res.* 163, 1–14. doi:10.1007/978-3-319-12048-5_1

Peng, L.-H., Sun, C.-N., Guan, N.-N., Li, J.-Q., and Chen, X. (20182018). HNMDA: Heterogeneous Network-Based miRNA-Disease Association Prediction. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L. H., Shen, L., Tian, X. F., Liu, F. X., Wang, J. J., Tian, G., et al. (20212010). Prioritizing Antiviral Drugs against SARS-CoV-2 by Integrating Viral Complete Genome Sequences and Drug Chemical Structures. Scientific reportsImmunologic Escape after Prolonged Progression-free Survival with Epidermal Growth Factor Receptor Variant III Peptide Vaccination in Patients with Newly Diagnosed Glioblastoma. *J. Clin. Oncol.* 1128 (1), 14722–114729.

Peng, L. H., Tian, X. F., Shen, L., Kuang, M., Tian, G., Yang, J. L., et al. (2020). Identifying Effective Antiviral Drugs against SARS-CoV-2 by Drug Repositioning through Virus-Drug Association Prediction[J]. *Front. Genet.* 11, 1072. doi:10.3389/fgene.2020.577387

Schäfer, N., Gielen, G. H., Rauschenbach, L., Kebir, S., Till, A., Reinartz, R., et al. (2019). Longitudinal Heterogeneity in Glioblastoma: Moving Targets in Recurrent versus Primary Tumors. *J. Transl Med.* 17, 96. doi:10.1186/s12967-019-1846-y

Seyed Khoei, N., Jenab, M., Murphy, N., Banbury, B. L., Carreras-Torres, R., Viallon, V., et al. (2020). Circulating Bilirubin Levels and Risk of Colorectal Cancer: Serological and Mendelian Randomization Analyses. *BMC Med.* 18, 229. doi:10.1186/s12916-020-01703-w

Shahcheraghi, S. H., Tchokonte-Nana, V., Lotfi, M., Lotfi, M., Ghorbani, A., and Sadeghnia, H. R. (2020). Wnt/beta-catenin and PI3K/Akt/mTOR Signaling Pathways in Glioblastoma: Two Main Targets for Drug Design: A Review. *Cpd* 26, 1729–1741. doi:10.2174/1381612826666200131100630

Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2020). Indicator Regularized Non-negative Matrix Factorization Method-Based Drug Repurposing for COVID-19. *Front. Immunol.* 11, 603615. doi:10.3389/fimmu.2020.603615

Wang, Z., Monteiro, C. D., Jagodnik, K. M., Fernandez, N. F., Gundersen, G. W., Rouillard, A. D., et al. (2016). Extraction and Analysis of Signatures from the Gene Expression Omnibus by the Crowd. *Nat. Commun.* 7, 12846. doi:10.1038/ncomms12846

Wu, W., Klockow, J. L., Zhang, M., Lafortune, F., Chang, E., Jin, L., et al. (2021). Glioblastoma Multiforme (GBM): An Overview of Current Therapies and Mechanisms of Resistance. *Pharmacol. Res.* 171, 105780. doi:10.1016/j.phrs.2021.105780

Xu, X., Long, H., Xi, B., Ji, B., Li, Z., Dang, Y., et al. (2019). Molecular Network-Based Drug Prediction in Thyroid Cancer. *Int. J. Mol. Sci.* 20. doi:10.3390/ijms20020263

Yang, J. a., Wang, L., Xu, Z., Wu, L., Liu, B., Wang, J., et al. (2020b). Integrated Analysis to Evaluate the Prognostic Value of Signature mRNAs in Glioblastoma Multiforme. *Front. Genet.* 11, 253. doi:10.3389/fgene.2020.00253

Yang, J., Huang, T., Song, W.-m., Petralia, F., Mobbs, C. V., Zhang, B., et al. (2016). Discover the Network Mechanisms Underlying the Connections between Aging and Age-Related Diseases. *Sci. Rep.* 6, 32566. doi:10.1038/srep32566

Yang, J., Peng, S., Zhang, B., Houten, S., Schadt, E., Zhu, J., et al. (2020a). Human Geroprotector Discovery by Targeting the Converging Subnetworks of Aging and Age-Related Diseases. *Geroscience* 42, 353–372. doi:10.1007/s11357-019-00106-x

Yang, Q., Wang, Y., Pan, X., Ye, J., Gan, S., Qu, F., et al. (2017). Frizzled 8 Promotes the Cell Proliferation and Metastasis of Renal Cell Carcinoma. *Oncotarget* 8, 78989–79002. doi:10.18632/oncotarget.20742

Yin, S., Xu, L., Bonfil, R. D., Banerjee, S., Sarkar, F. H., Sethi, S., et al. (2013). Tumor-initiating Cells and FZD8 Play a Major Role in Drug Resistance in Triple-Negative Breast Cancer. *Mol. Cancer Ther.* 12, 491–498. doi:10.1158/1535-7163.mct-12-1090

Yoda, A., Adelmant, G., Tamburini, J., Chapuy, B., Shindoh, N., Yoda, Y., et al. (2015). Mutations in G Protein β Subunits Promote Transformation and Kinase Inhibitor Resistance. *Nat. Med.* 21, 71–75. doi:10.1038/nm.3751

Yuan, J., Yuan, Z., Ye, A., Wu, T., Jia, J., Guo, J., et al. (2021). Low GNG12 Expression Predicts Adverse Outcomes: A Potential Therapeutic Target for

Osteosarcoma. *Front. Immunol.* 12, 758845. doi:10.3389/fimmu.2021.758845

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (20212021). Using Network Distance Analysis to Predict lncRNA-miRNA Interactions. *Interdiscip. Sci. Comput. Life Sci.* 13 (3), 535–545. doi:10.1007/s12539-021-00458-z

Zhao, H.-f., Wang, J., Shao, W., Wu, C.-p., Chen, Z.-p., To, S.-s. T., et al. (2017). Recent Advances in the Use of PI3K Inhibitors for Glioblastoma Multiforme: Current Preclinical and Clinical Development. *Mol. Cancer* 16, 100. doi:10.1186/s12943-017-0670-3

Zhou, L., Li, Z., Yang, J., Tian, G., Liu, F., Wen, H., et al. (2019). Revealing Drug-Target Interactions with Computational Models and Algorithms. *Molecules* 24 (9), 1714. doi:10.3390/molecules24091714

Zhou, L., Wang, J., Liu, G., Lu, Q., Dong, R., Tian, G., et al. (2020). Probing Antiviral Drugs against SARS-CoV-2 through Virus-Drug Association Prediction Based on the KATZ Method. *Genomics* 112 (6), 4427–4434. doi:10.1016/j.ygeno.2020.07.044

Zhu, L., and Chen, L. (2019). Progress in Research on Paclitaxel and Tumor Immunotherapy. *Cell Mol Biol Lett* 24, 40. doi:10.1186/s11658-019-0164-y

Conflict of Interest: BW, MY, XY, BJ, and YW were employed by the company Geneis (Beijing) Co., Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 February 2022
doi: 10.3389/fgene.2022.829937

Using Graph Attention Network and Graph Convolutional Network to Explore Human CircRNA–Disease Associations Based on Multi-Source Data

Guanghui Li[1]*, *Diancheng Wang*[1], *Yuejin Zhang*[1], *Cheng Liang*[2], *Qiu Xiao*[3] and *Jiawei Luo*[4]*

[1]School of Information Engineering, East China Jiaotong University, Nanchang, China, [2]School of Information Science and Engineering, Shandong Normal University, Jinan, China, [3]College of Information Science and Engineering, Hunan Normal University, Changsha, China, [4]College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

OPEN ACCESS

Edited by:
Lihong Peng,
Hunan University of Technology, China

Reviewed by:
Min Chen,
Hunan Institute of Technology, China
Aiping Yao,
Lanzhou University, China
Fei Guo,
Tianjin University, China

***Correspondence:**
Guanghui Li
ghli16@hnu.edu.cn
Jiawei Luo
luojiawei@hnu.edu.cn

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

Received: 06 December 2021
Accepted: 10 January 2022
Published: 07 February 2022

Citation:
Li G, Wang D, Zhang Y, Liang C, Xiao Q and Luo J (2022) Using Graph Attention Network and Graph Convolutional Network to Explore Human CircRNA–Disease Associations Based on Multi-Source Data. Front. Genet. 13:829937. doi: 10.3389/fgene.2022.829937

Cumulative research studies have verified that multiple circRNAs are closely associated with the pathogenic mechanism and cellular level. Exploring human circRNA–disease relationships is significant to decipher pathogenic mechanisms and provide treatment plans. At present, several computational models are designed to infer potential relationships between diseases and circRNAs. However, the majority of existing approaches could not effectively utilize the multisource data and achieve poor performance in sparse networks. In this study, we develop an advanced method, GATGCN, using graph attention network (GAT) and graph convolutional network (GCN) to detect potential circRNA–disease relationships. First, several sources of biomedical information are fused *via* the centered kernel alignment model (CKA), which calculates the corresponding weight of different kernels. Second, we adopt the graph attention network to learn latent representation of diseases and circRNAs. Third, the graph convolutional network is deployed to effectively extract features of associations by aggregating feature vectors of neighbors. Meanwhile, GATGCN achieves the prominent AUC of 0.951 under leave-one-out cross-validation and AUC of 0.932 under 5-fold cross-validation. Furthermore, case studies on lung cancer, diabetes retinopathy, and prostate cancer verify the reliability of GATGCN for detecting latent circRNA–disease pairs.

Keywords: circRNA-disease associations, deep learning, graph attention network, graph convolutional network, centered kernel alignment

INTRODUCTION

Circular RNA (circRNA) is a novel endogenous non-coding RNA forming a covalently closed loop structure, which lacks a 50-end cap and a 30-end ployA tail (Memczak et al., 2013; Meng et al., 2017). This structure is beneficial for circRNA to develop resistance to RNA exonuclease degradation and provides a more stable biological expression (Li et al., 2015). As a result, in most species, the average half-life of circRNAs is substantially increased than their linear equivalent. When circRNAs were first found as early as 1970s, they had been regarded as the abnormal shear or product of "shear noise,"

limited to the level of technology and knowledge at that time. In previous studies, multiple circRNAs were verified to be widespread in eukaryotes and play an essential role in biological functions with the advancement of biology and sequencing technologies. Currently, the biological functions of circRNA are reflected as follows (Rong et al., 2017): regulation of alternative splicing or transcription, miRNA sponges, regulation of protein binding, and generation of pseudogenes.

CircRNA has become a new biomarker due to its abundance, structural stability, developmental stage specificity, and tissue specificity (Zhang Z. et al., 2018), which can be discovered in saliva, blood, and exosomes. Cumulative research studies have confirmed that multiple circRNAs are significant to the expression of various pathological conditions (Han et al., 2018; Zhu et al., 2017; Zhang S. et al., 2018), especially cancer (Vo et al., 2019), cardiovascular, cerebrovascular, and nervous system diseases. For instance, circRNA hsa_circ_0027599 is overexpressed in gastric cancer (Wang L. et al., 2018), thereby regulating the expression of the gene PHLDA1 and promoting tumorigenesis. In cardiovascular and cerebrovascular diseases, circRNA circWDR77Z targets and regulates miRNA miR-124/FGF-2 through the "sponge" function (Chen et al., 2017), which affects the migration and proliferation for vascular smooth muscle cells, thereby promoting atherosclerosis development. For myocardial infarction, overexpression of circRNA CDR1 leads to the upregulation of downstream corresponding enzymes and proteins (Zhang et al., 2016), thereby aggravating myocardial infarction. In neurological diseases, the expression of circRNA in brain tissue is different, and its distribution in the brain is uneven (Zhang et al., 2021b).

Although circRNA is commonly expressed in various cell lines and tissues with strong tissue specificity and development stage specificity, the pathogenic mechanism of circular RNA and how it interacts with other biological molecules remain unknown. In recent years, researchers have established many experimentally verified or reported databases on relationships between circRNAs and diseases, such as circBase (Glažar et al., 2014), circRNADb (Chen et al., 2016), circR2Disease (Fan et al., 2018b), circRNADisease (Zhao et al., 2018), circ2Disease (Yao et al., 2018) and circ2Traits databases (Ghosal et al., 2013). Considering that conventional biological studies are cost-ineffective and time-consuming, several computational approaches have been designed to detect relationships between diseases and circRNAs efficiently (Xiao et al., 2022; Lei et al., 2021). At present, the proposed computational models for discovering relationships between diseases and circRNAs are mainly classified into the following groups:

Network propagating methods have been widely applied to detect correlations between diseases and various biological entities, including circRNAs, due to the efficient use of network structure information (Peng et al., 2018). Zhang et al. designed a linear neighbor marker propagation approach named CD-LNLP *via* neighbor similarity to reveal relationships between diseases and circRNAs (Zhang et al., 2019). Li et al. presented the DWNCPCDA using DeepWalk and network consistency projection (Chen et al., 2018) to detect unobserved associations between diseases and circRNAs (Li G. et al., 2020). Lei et al. constructed a prediction model named RWRKNN, which combined the k-nearest neighbor and RWR to calculate weighted features for diseases and circRNAs (Lei and Bian, 2020).

Path-based methods are widely adopted to calculate potential interactions between diseases and circRNAs by measuring the weight of paths in different networks. Lei et al. presented a path-weighted method named PWCDA, which predicted the circRNA–disease relationships by calculating the probability value for each circRNA–disease pair *via* path information (Lei et al., 2018). Fan et al. presented the model named KATZHCDA *via* the circRNA expression profile, the similarity of the disease phenotype, and the nuclear similarity of the Gaussian interaction profile using the KATZ method to detect potential interactions between diseases and circRNAs through the heterogenous network (Fan et al., 2018a). Zhao et al. revealed a computed method named IBNPKATZ using the bipartite network projection model and the KATZ (Zhang et al., 2021a) model to discover circRNA–disease interactions (Zhao et al., 2019).

Matrix factorization–based methods have been carried out for detecting circRNA–disease relationships by constructing a low-dimensional matrix to represent the initial input features (Wang P. et al., 2018; Peng et al., 2020a). Wei et al. used weight-based nearest neighbor nodes to reconstruct the association matrix and designed a graph regularized non-negative matrix factorization algorithm iCircDA-MF to detect relationships between diseases and circRNAs (Wei and Liu, 2020). Lu et al. constructed a model named DMFCDA with deep matrix factorization, which infers potential circRNA–disease interactions by mapping features of diseases and circRNAs into low-dimensional spaces (Lu et al., 2021). Yan et al. used the Kronecker product kernel to design a regularized least squares algorithm called DWNN-RLS to detect relationships (Yan et al., 2018). Li et al. presented an advanced approach named SIMCCDA by regarding predicting associations as a recommendation system task, which achieves outstanding performance for discovering circRNA–disease associations (Li M. et al., 2020).

Deep learning integrates low-level features to construct high-level representations of features or attribute categories through the deep non-linear network structure (Peng et al., 2021; Zhou et al., 2021). Wang et al. designed a model to reveal interactions between diseases and circRNAs using deep convolutional neural networks and deep generative adversarial networks (Wang et al., 2020a). Wang et al. designed an approach named GCNCDA to identify disease-related circRNAs, which extracts high-level features contained in the circRNA–disease heterogenous network through graph convolutional networks to calculate association scores (Wang et al., 2020b). GATCDA is a novel model for discovering the correlation between diseases and circRNAs, which learns the latent representation of nodes by assigning corresponding weights to each neighbor node (Bian et al., 2021). Xiao et al. designed a computational model named NSL2CD that adopts network embedding by adaptive subspace learning (Xiao et al., 2021).

Although the abovementioned approaches have achieved excellent predictive performance, there are still several

limitations given as follows: First, network-based methods achieve poor performance in sparse networks due to a small amount of network structure information. Second, path-based methods fail to dynamically calculate weights based on known associations, which makes it unable to efficiently detect relationships between diseases and circRNAs with new diseases or circRNAs. Third, matrix factorization–based methods could not discover a non-linear interaction between diseases and circRNAs. Last, current deep learning–based methods could not effectively utilize the multisource data and only pay more attention to features of the neighbor nodes or the node itself, respectively.

To solve the abovementioned challenges, we develop an advanced method GATGCN *via* graph attention network (GAT) and graph convolutional network (GCN) to detect potential circRNA–disease relationships. The complete process could be summarized as four steps: First, multisource similarity data for circRNAs and diseases are fused by the centered kernel alignment model (CKA) (Cristianini et al., 2006). Second, we adopt the graph attention network to learn the dense representation of nodes on fused disease similarity network and fused circRNA similarity network. Third, we construct the heterogenous network by connecting circRNA–disease interaction network, feature matrix of diseases, and feature matrix of circRNAs. Finally, the graph convolutional network is adopted to get prediction scores based on the heterogenous network. According to reliable computer experiments, GATGCN outperforms several state-of-the-art methods with a prominent AUC of 0.932.

MATERIALS

Human CircRNA–Disease Associations

The circR2Disease provides verified relationships between diseases and circRNAs, which is a manually curated database including 739 known relationships between 100 diseases and 676 circRNAs. We eventually extract 661 associations between 88 diseases and 585 circRNAs for humans after removing the associations unrelated to human species and duplicate associations.

Human Disease–MiRNA Associations

MiRNAs are significant to pathogenesis and treatment of diseases as the important regulatory molecule for genes. On dataset, we collect 1,883 experimentally verified disease–miRNA relationships between 462 miRNAs and 88 diseases from the HMDD (Li et al., 2014), which provides disease-associated miRNAs and their target genes, including 8,802 known relationships between 350 diseases and 32281 miRNAs.

Human Disease–Gene Associations

Due to gene mutation and expression affecting diseases, diseases are closely related to genes. On the dataset, 74 experimentally verified disease–gene associations between 61 genes and 88 diseases are filtered out, downloaded from http://cssb2.biology.gatech.edu/knowgene/.

Human CircRNA–MiRNA Associations

With plenty miRNA binding sites (Hansen et al., 2013; Peng et al., 2020b), circRNAs actively affect the expression of miRNA's downstream genes as miRNA sponges (Peng et al., 2017; Zeng et al., 2020). We obtain 17844 known circRNA–miRNA associations between 640 miRNAs and 585 circRNAs from ENCORI (available at http://starbase.sysu.edu.cn/agoClipRNA.php? source=circRNA).

Human CircRNA–Gene Associations

According to the previous research, circRNAs are verified to be significant in regulating the expression of genes. On the dataset, 487 known circRNA–gene associations between 418 genes and 585 circRNAs are extracted from http://cssb2.biology.gatech.edu/knowgene/search.html.

Disease Semantic Similarity

The semantic information of the diseases has been wildly adopted to measure the similarity of diseases because of its effectiveness and stability. In this study, we obtain the related annotation terms for each disease from MeSH.

In MeSH, the directed acyclic graph (DAG) is applied to represent the semantic relationship among diseases, in which nodes denote corresponding disease information and directed edges denote the relationship among diseases. Specifically, disease d_i can be described as three items $DAG_i = [d_i, T(d_i, E(d_i))]$, where $T(d_i)$ represents d_i itself and its ancestor nodes and $E(di)$ is relationships between d_i and all diseases. The contribution of disease d_i in DAG_i is formulated as follows:

$$\begin{cases} D_{d_i}(n) = 1 & \text{if } n = d \\ D_{d_i}(n) = \max\{\sigma \cdot D_{d_i}(n') | n' \in children\ of\ n\} & \text{if } n \neq d \end{cases}, \quad (1)$$

where σ denotes the attenuation factor for semantic contribution, which is defined as the optimal value of 0.5 according to Wang's experience Wang et al. (2010); n' represents the child node of the node n. Therefore, the overall semantic score of the disease d_i is measured by accumulating the contribution scores from its ancestor diseases and itself as follows:

$$D(d_i) = \sum_{n \in T(d_i)} D_{d_i}(n). \quad (2)$$

In general, diseases with more common parts shared in the DAG achieve higher semantic similarities. Based on this hypothesis, the value of disease semantic similarity between disease d_i and disease d_j is formulated *via* **Eq.3**:

$$DS(d_i, d_j) = \frac{\sum_{n \in T_{d_i} \cap T_{d_j}} \left(D_{d_i}(n) + D_{d_j}(n)\right)}{D(d_i) + D(d_j)}. \quad (3)$$

CircRNA Functional Similarity

According to previous studies, circRNAs that are relevant to more similar diseases are prone to be more similar in functions (Li et al., 2019). Then, the BMA method is deployed to measure the functional similarity score among different circRNAs according to relevant disease sets. Given a specific disease d_i

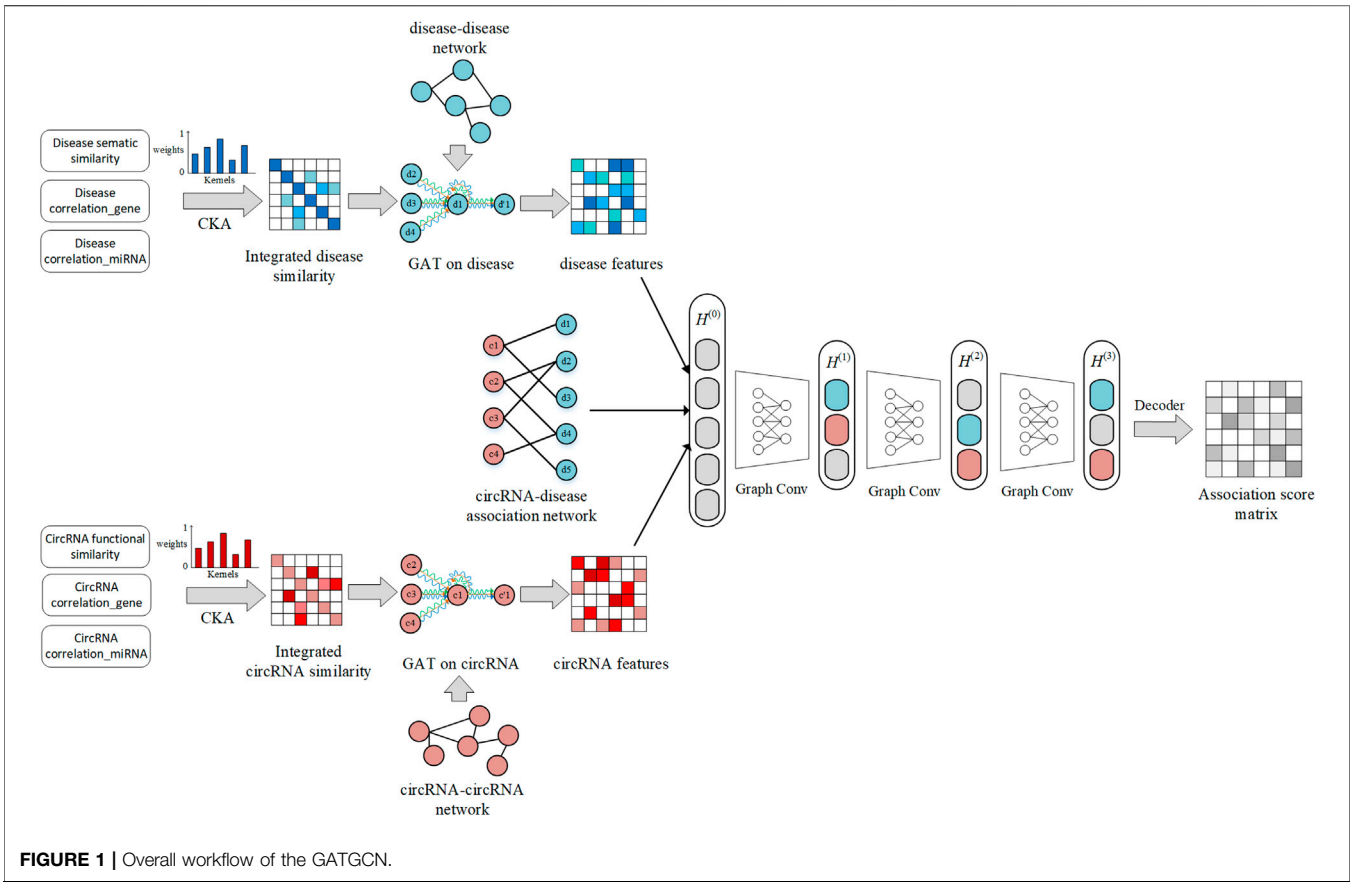


FIGURE 1 | Overall workflow of the GATGCN.

and $D = (d_1, d_2, \ldots, d_t)$, the score of functional similarity between circRNA c_i and circRNA c_j is measured *via* **Eqs 4**, **5**:

$$FS(c_i, c_j) = \frac{\sum_{m=1}^{|D_i|} S(d_m, D_j) + \sum_{n=1}^{|D_j|} S(d_n, D_i)}{|D_i| + |D_j|}, \quad (4)$$

$$S(d_m, D_j) = \max_{1 \le t \le |D_j|} (S(d_m, d_t)), \quad (5)$$

where D_j represents the collection of diseases associated with circRNA c_j. $S(d_m, D_j)$ represents the similarity between disease d_m associated with circRNA c_i and disease collection D_j associated with circRNA c_j.

Pearson's Correlation Coefficient Similarity

Since the circRNA functional similarity network and the disease semantic similarity network are prone to be sparse, we adopt Pearson's correlation coefficient approach to enrich multisource similarity data by calculating the linear correlation among different variables. To be specific, the value of Pearson's correlation between variable M and variable N is measured as follows:

$$Cor(M, N) = \frac{cov(M, N)}{\sqrt{var(M) var(N)}}, \quad (6)$$

where $var(M)$ measures the variance of M; $cov(M, N)$ calculates the covariance between M and N; the value of $Cor(M, N)$ ranges from −1 to 1, which reflects the strength of the linear correlation between M and N.

Four binary networks have been built including the disease–gene network, circRNA–miRNA network, circRNA–gene network, and disease–miRNA network. Then, Pearson's correlation coefficient approach is adopted to compute disease similarity and circRNA similarity *via* corresponding bipartite networks. The equation is computed as follows:

$$Cor(n_i, n_j) = \frac{cov(IP(n_i), IP(n_j))}{\sqrt{var(IP(n_i)) var(IP(n_j))}}, \quad (7)$$

where $IP(n_i)$ denotes the ith row of the corresponding association network. $Cor(n_i, n_j)$ denotes the value of Pearson's correlation similarity between node n_i and node n_j based on the corresponding association network.

METHODS

In this work, we develop an advanced method GATGCN *via* the graph attention network and graph convolutional network to detect potential circRNA–disease relationships. As shown in **Figure 1**, the complete process could be summarized in four

steps: First, the CKA-based model is adopted to fuse multisource similarity data for circRNAs and diseases. Second, we adopt the graph attention network to calculate the dense representation of nodes on the fused disease similarity network and fused circRNA similarity network. Third, we construct the heterogenous network, including circRNA–disease interactions network, feature matrix of diseases, and feature matrix of circRNAs. Eventually, the graph convolutional network is adopted to get prediction scores based on the constructed heterogenous network.

Centered Kernel Alignment

In previous studies, multisource data are usually fused by calculating the average value, which ignores the importance among different kernels. Thus, the centered kernel alignment (CKA) model (Wang et al., 2021) is adopted to fuse several kinds of similarities for diseases and circRNAs based on different weights. We consider $K_d = \{K_d^1, \ldots, K_d^v\}$ and $K_c = \{K_c^1, \ldots, K_c^u\}$ as different kernels for disease space and circRNA space. The v and u denote the number of kernels from disease space and circRNA space, respectively. Meanwhile, the basic CKA model (Cristianini et al., 2006) is used as the objective of MKL (Ding et al., 2019) to measure the corresponding weight of each kernel.

To be specific, the kernels K_c^* and $K^*{}_d$ based on optimal weight are calculated as follows:

$$K_c^* = \sum_{p=1}^{u} a_c^p K_c^p, \quad K_c^p \in R^{m\times m}, \tag{8}$$

$$K_d^* = \sum_{p=1}^{v} a_d^q K_d^q, \quad K_d^q \in R^{n\times n}, \tag{9}$$

where $\alpha_c = \{\alpha_c^1, \ldots, \alpha_c^u\}$and $\alpha_d = \{\alpha_d^1, \ldots, \alpha_d^v\}$.

Basic CKA (Cristianini et al., 2006) is adopted to calculate the weights of each kernel on the training set. The kernel alignment score between the two kernels is formulated as follows:

$$U(E, I) = \frac{\langle E, I\rangle_F}{\|E\|_F \|I\|_F}, \tag{10}$$

where E, I denotes the corresponding similarity matrix, $||E||_F$ denotes the Frobenius norm, and $<E, I> = Trace(E^T I)$ denotes the Frobenius inner product. The kernel alignment score represents the similarity among different kernels. Specifically, the kernel alignment score between the similarity kernel (disease kernel or circRNA kernel) and the ideal kernel matrix is measured as follows:

$$\max_{\beta\geq 0} CU(K^*, K_{ideal}) = \max_{\beta\geq 0} \frac{\langle Z_N K^* Z_N, K_{ideal}\rangle_F}{\|Z_N K^* Z_N\|_F \|K_{ideal}\|_F}, \tag{11}$$

$$subject\sim to\sim K^* = \sum_{p=1}^{N} \beta^p K^p \quad \beta \geq 0, p = 1, 2..., N, \tag{12}$$

$$\sum_{p=1}^{N} \beta^p = 1, \tag{13}$$

where K_{ideal} denotes a label kernel constructed by known associations; $K_{ideal,\ d} = Y_{train}^T Y_{train} \in R^{n\times n}$ and $K_{ideal,\ c} = Y_{train} Y_{train}^T \in R^{m\times m}$ denote the ideal kernel of diseases and circRNAs, respectively.

Attention Mechanism on Similarity

Considering that current methods did not capture potential features on the similarity network, we adopt the graph attention method to learn latent representation of diseases and circRNAs, which assigns corresponding weights to different node features based on the local graph structure to ignore noise and redundancy. The advantage of the attention mechanism is to directly evaluate which features are preferred embedding for specific downstream tasks by calculating the weights. First, we obtain the corresponding association matrix by setting a threshold on the similarity network for diseases and circRNAs. Then, the GAT (Veličković et al., 2017) is applied to learn dense representation for diseases and circRNAs as follows:

The input layer of the graph attention network is formulated as follows:

$$f = \{f_1, f_2, ..., f_N\}, f_i \in R^F, \tag{14}$$

where F denotes the dimension of features, and N represents the number of nodes in the corresponding similarity network. $f \in R^{N\times F}$ is constructed by the features of nodes in the corresponding similarity network. The output layer of the graph attention network is defined as follows:

$$f' = \{f_1', f_2', ..., f_i'\}, f_i' \in R^{F'}, \tag{15}$$

where F' denotes the length of learned features, and $f' \in R^{N\times F'}$ represents the learned latent representations of nodes in the network. The first step is to calculate the weight of the corresponding neighbor node. The importance of the given nodes is computed by the self-attention mechanism. For each association pair between node n_i and node n_j, the attention coefficient e_{ij} is calculated as follows:

$$e_{ij}(n_i, n_j) = att(Wf_i, Wf_j), \tag{16}$$

where att represents a mapping function transforming high-level features to a real number for association pair between node n_i and node n_j based on input features, and $W \in R^{F'\times F}$ denotes a trainable weight matrix. To avoid the influence of dimension between different attention coefficients, e_{ij} is further normalized as follows:

$$\theta_{ij} = soft\,max(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{t\in N_i}\exp(eit)}, \tag{17}$$

where N_i represents the collection of neighbor nodes of node n_i. θ_{ij} denotes the normalized weight representing the importance between node n_i and node n_j in the network.

From the abovementioned formula, we obtain the combined attention mechanism as follows:

$$\theta_{ij} = \frac{\exp(leakyRelu(a^T[Wf_i||Wf_t]))}{\sum_{t\in N_i}\exp(leakyRelu(a^T[Wf_i||Wf_t]))}, \tag{18}$$

where $leakyRelu$ denotes a non-saturated activation function, which can solve the vanishing gradients and accelerate convergence. $a \in R^{2F'}$ denotes the weight matrix, which maps features to a real number. The second step is to aggregate the features of all neighbors for a given node by integrating the

corresponding weight. The aggregation between the given node and neighbors is formulated as follows:

$$f_i^{'} = \sigma\Big(\sum\nolimits_{t\in N_i} \theta_{it} W f_t\Big) \tag{19}$$

where σ denotes a non-saturated activation function. Multi-head attention mechanism is applied in GAT to integrate features and prevent overfitting. The output with the multi-head attention mechanism contains the features in different representation subspaces, which enhances the expressive capacity of the model. To be specific, the multi-head attention model based on the combination of K-independent attention mechanisms learns latent features as follows:

$$f_i^{'} = \sigma\Big(\frac{1}{K}\sum\nolimits_{K=1}^{K}\sum_{t\in N_i}\theta_{it}^{k}\cdot W^{K} f_t\Big), \tag{20}$$

where K represents the number of self-attention models. W^k denotes the trained weight matrix of the kth attention model.

Heterogenous Network

The heterogenous network is constructed as initial features of GCN, including circRNA–disease associations, learned feature matrix of circRNAs, and learned feature matrix of diseases. The binary matrix A is constructed, and $A_{ij} = 1$ if the interaction between circRNA c_i and disease d_j has been verified; otherwise $A_{ij} = 0$. The learned feature matrix of circRNAs and learned feature matrix of diseases based on GAT are denoted as matrix S_c and matrix S_d, respectively. The heterogenous network A_H is defined as follows:

$$A_H = \begin{bmatrix} S^c & A \\ A^T & S^d \end{bmatrix} \in R^{(M+N)\times(M+N)}. \tag{21}$$

Graph Convolutional Network on Heterogenous Network

In recent years, GCN has achieved superior performance in node prediction, node classification, and edge prediction tasks (Kipf and Welling, 2016). In order to discover potential relationships between diseases and circRNAs, GCN models (Wang et al., 2020b) are designed to effectively extract features of circRNA–disease relationships based on the global graph structure by aggregating feature vectors of neighbors. To be specific, given a network G, each layer of the GCN model embedding is formulated as follows:

$$H^{(l+1)} = \sigma\Big(D^{-\frac{1}{2}} G D^{-\frac{1}{2}} H^{(l)} W^{(l)}\Big), \tag{22}$$

where $H^{(l)}$ denotes the propagation of features at the lth layer, $\sigma(\cdot)$ represents a nonlinear activation function, $D = \mathrm{diag}(\sum G_{ij})$ denotes the degree matrix of G, and $W^{(l)}$ is the trained weight matrix at the lth layer. GCN integrates low-level features to construct high-level representations of nodes on the constructed heterogenous network A_H. In addition, we adjust the number of graph convolutional network layers and set node dropout to avoid overfitting, which can reduce excessive parameters and improve the generalization ability of the GATGCN. The penalty factor μ is set to regulate the contribution of learned similarity features in the embedding of graph convolutional layers. Specifically, the input heterogenous network G is defined as follows:

$$G = \begin{bmatrix} \mu\cdot S^c & A \\ A^T & \mu\cdot S^d \end{bmatrix}. \tag{23}$$

Then, the initial embedding is defined as follows:

$$H^{(0)} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \tag{24}$$

The first layer of the GCN model embedding is calculated as follows:

$$H^{(1)} = \sigma\Big(D^{-\frac{1}{2}} G D^{-\frac{1}{2}} H^{(0)} W^{(0)}\Big), \tag{25}$$

where $W^{(0)} \in R^{(M+N)\times k}$ represents an input-to-hidden trained weight matrix, $H^{(1)} \in R^{(M+N)\times k}$ represents the first-layer propagation of features, including circRNAs and diseases. K denotes the embedding dimension in graph conventional layers. We adopt the exponential linear unit (Clevert et al., 2016) as the nonlinear activation function to enhance noise robustness and expressive capacity of the model in all graph convolutional layers. Eventually, the bilinear decoder A' proposed by Huang et al., (2020) is deployed to reconstruct the circRNA–disease association matrix as follows:

$$A' = sigmoid\Big(H_C W' H_D^T\Big), \tag{26}$$

where $W' \in R^{k\times k}$ denotes a trained weight matrix. $H_D \in R^{N\times k}$ and $H_C \in R^{M\times k}$ represent the last embedding for diseases and circRNAs, respectively. The final predicted relationship score a'_{ij} between circRNA c_i and disease d_j is obtained according to the corresponding (i, j)th entry of A'.

RESULTS

In this section, several verification experiments are deployed to assess the predictive capacity of GATGCN. First, we assess the influence of different parameters setting on GATGCN. Second, we introduce the evaluation metrics under leave-one-out cross-validation and 5-fold cross-validation to analyze the predictive capacity of GATGCN. Third, we design the ablation study to assess the impact of each part on GATGCN. Fourth, we discuss and compare GATGCN with state-of-the-art models on the same dataset. Last, case studies are deployed to further assess the performance in detecting potential relationships on GATGCN.

Parameter Setting

The performance of the model is frequently impacted by hyperparameter settings. Analysis of the parameters can quantitatively evaluate the stability of the model and provide a reference for parameter selection. The learning rate is significant to the convergence of the gradient descent algorithm in the model. **Figure 2** indicates that the model will converge slowly

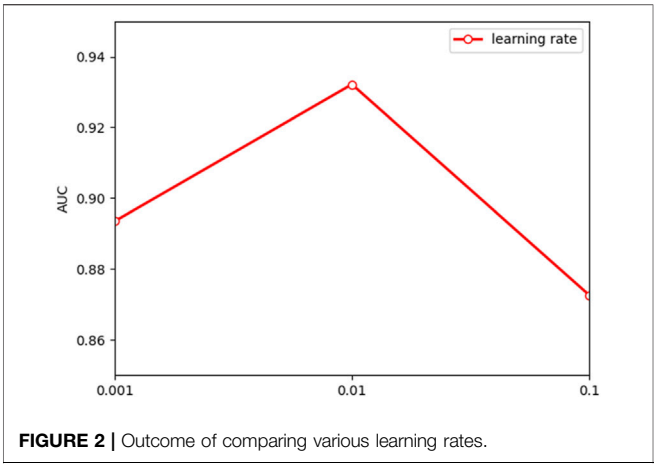


FIGURE 2 | Outcome of comparing various learning rates.

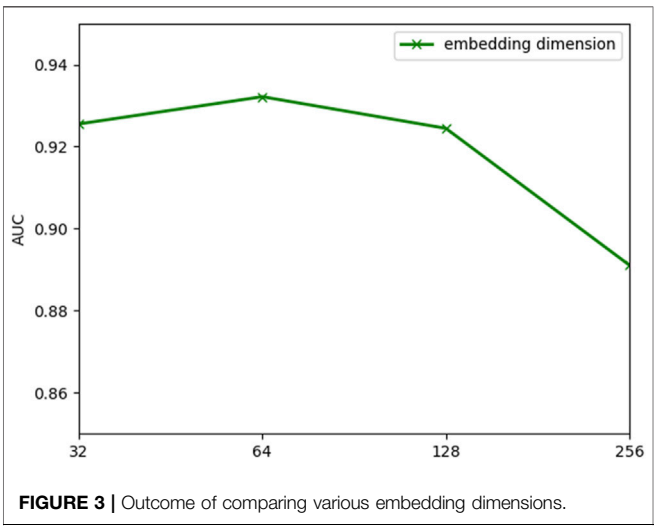


FIGURE 3 | Outcome of comparing various embedding dimensions.

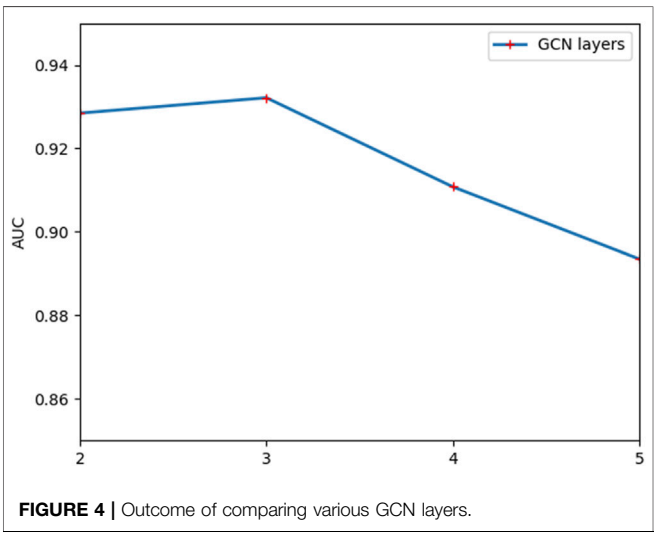


FIGURE 4 | Outcome of comparing various GCN layers.

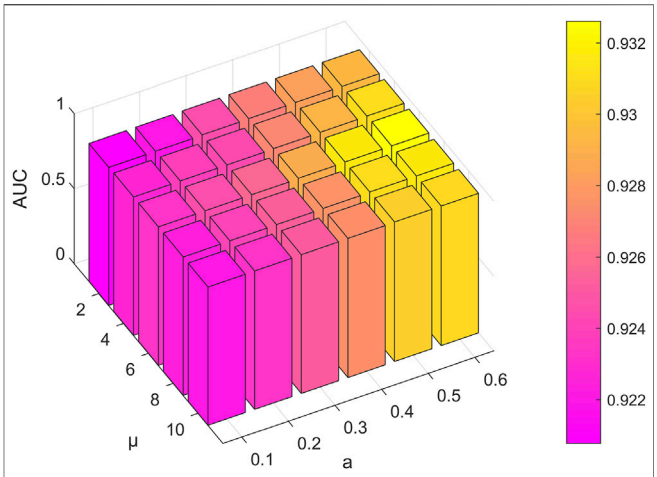


FIGURE 5 | Outcome of comparing various dropout rates and penalty factors.

with too small a learning rate, while too large a learning rate makes it hard to converge. According to the results in **Figure 3**, the embedding dimension within a certain size range has less impact on the convergence of our model. However, when the embedding dimension is too large, the model is prone to overfitting due to plenty of parameters. As shown in **Figure 4**, the model performs better with small layers of the graph convolutional network, and the performance drops significantly when the number of layers of GCN is $l > 4$. The reason is that the GCN with more layers not only captures more global prior information but also captures a lot of noise at the same time. Meanwhile, the penalty factor μ is set to regulate the contribution of learned similarity features in the propagation of convolutional layers, and the dropout rate a is adopted to avoid overfitting. As shown in **Figure 5**, the model achieves best performance at $\mu = 6$ and $a = 0.6$.

Evaluation Metrics

Cross-validation is a self-consistent testing approach widely adopted to demonstrate the predictive capacity of a method. The basic idea is to carry out the resampling method to select a portion of the benchmark data set as the training set to train the model, and the remaining samples to verify the model. Five-fold cross-validation and leave-one-out cross-validation are deployed to assess the predictive capacity of GATGCN. For five-fold cross-validation, the whole samples in the dataset are randomly separated into five roughly identical sections, four of which are adopted to train the GATGCN and the other is used to test the GATGCN. In order to decrease the bias produced by sample segmentation, the five-fold cross-validation is repeated 30 times to calculate the average result as the ultimate output. For leave-one-out cross-validation, each time only one sample in the dataset is selected among all recorded circRNA–disease relationships to test the model, and the remaining known relationships are utilized as training samples. In this study, since circRNA functional similarity relies on known associations; we recalculate the circRNA functional similarity in each repetition of the experiment.

TABLE 1 | Results generated by the GATGCN under five-fold CV and LOOCV.

Test set	Accu	Rec	Spe	F1	AUC
5-fold CV_1	0.988	0.682	0.989	0.437	0.956
5-fold CV_2	0.987	0.568	0.991	0.361	0.918
5-fold CV_3	0.987	0.644	0.988	0.373	0.922
5-fold CV_4	0.990	0.627	0.991	0.414	0.931
5-fold CV_5	0.991	0.647	0.990	0.402	0.934
Average	0.9886 ± 0.0024	0.6336 ± 0.0656	0.9898 ± 0.0012	0.3974 ± 0.0396	0.9322 ± 0.0238
LOOCV	0.987	0.782	0.992	0.542	0.951

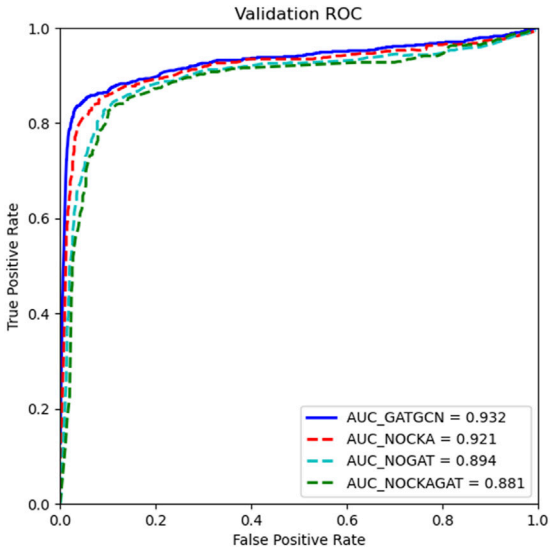


FIGURE 6 | Performance of the GATGCN based on various model combinations.

In this study, the area under the curve (AUC) is applied as the primary metric to assess our model, which can visually show the predictive ability of GATGCN under each decision threshold. The basic principle is to treat the false-positive rate (FPR) and the true rate (TPR) as a two-dimensional coordinate point in a Cartesian coordinate system with FPR as the abscissa and TPR as the ordinate under different discrimination thresholds. Besides, several threshold-based metrics are adopted to further evaluate the predictive performance of the GATGCN including recall, specificity, accuracy, and F1. The detailed results of five-fold cross-validation and leave-one-out cross-validation are summarized in **Table 1**.

Ablation Study

The model GATGCN is used to detect potential relationships between diseases and circRNAs based on the centered kernel alignment model (CKA), graph attention network (GAT), and graph convolutional network (GCN). In order to verify the importance of CKA, GAT, and GCN in our model, we apply the ablation study to our model. In this part, we replace the CKA model with calculated average to fuse multisource similarity as NOCKA. Meanwhile, we only combine the CKA model and GCN model as NOGAT to calculate association scores. In addition, we only adopt the GCN to predict associations between diseases and circRNAs as NOCKAGAT. According to the results in **Figure 6**, the complete model GATGCN is compared with NOCKA, NOGAT, and NOCKAGAT with five-fold cross-validation, which achieves the best AUC of 0.932. In general, using the the graph attention network on the similarity network is beneficial to learn the latent representation of nodes. The AUC of GATGCN and NOCKA is significantly higher than that of the other two models, which indicates that GAT is significant to detect relationships between diseases and circRNAs. Moreover, the comparison between GATGCN and NOCKA suggests that the fusion of multisource similarity based on weights can improve performance in circRNA–disease relationship prediction.

Comparison With Other Methods

To confirm the advantage of GATGCN, we compare it with several classic prediction methods with five-fold cross-validation. Since these methods adopt various datasets and evaluation metrics, we apply the same dataset and AUC as the metrics to compare the predictive capacity of models fairly and reasonably. In this part, the GATGCN is compared with several state-of-the-art methods, including KATZHCDA (Fan et al., 2018a), DWNN-RLS (Yan et al., 2018), PWCDA (Lei et al., 2018), GCNCDA (Wang et al., 2020b), and GATCDA (Bian et al., 2021). KATZHCDA is a graph-based method that uses the walking lengths and number of walks among nodes to measure the similarity among nodes in the heterogenous network. The DWNN-RLS measures initial relational values of new diseases and circRNAs *via* the decreasing weight k-nearest neighbor model and adopts the Kronecker product kernel to predict associations between diseases and circRNAs. The PWCDA predicts the circRNA–disease relationships by searching the paths without repeating for all circRNA–disease pairs based on the constructed heterogenous network. The GCNCDA extracts high-level features in the heterogenous network through graph convolutional neural networks and predicts the correlation between circRNAs and diseases *via* Forest by Penalizing Attributes. GATCDA learns the latent representation of nodes by assigning corresponding weights to each neighbor node, which efficiently aggregates the information of neighbor nodes and utilizes the local features of the graph. The results in **Figure 7** indicate that

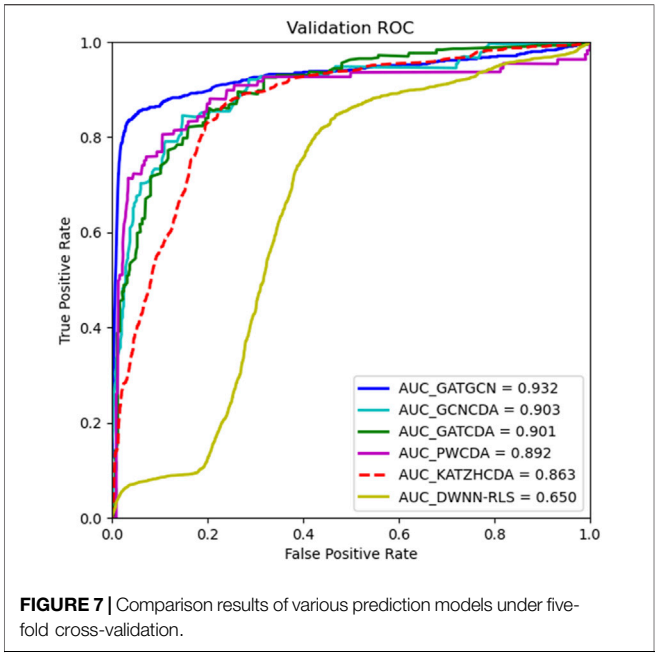


FIGURE 7 | Comparison results of various prediction models under five-fold cross-validation.

GATGCN achieves the best AUC of 0.932, which is substantially greater than that of other models, and produces 7.9%, 43.3%, 4.5%, 3.2%, and 3.4% improvement in the AUC compared with KATZHCDA, DWNN-RLS, PWCDA, GCNCDA, and GATCDA respectively.

Furthermore, the number of known interactions between diseases and circRNAs in the dataset can greatly affect the performance of the method, which also indicates the robustness of the method. Thus, we randomly remove a part of known associations between diseases and circRNAs at a ratio $r \in \{80\%, 85\%, 90\%, 95\%, \text{and } 100\%\}$ with five-fold cross-validation. As shown in **Figure 8**, the performance of GATGCN improves with increasingly known associations. Meanwhile, the GATGCN achieves the best result across different data richness among KATZ, DWNN-RLS, PWCDA, GCNCDA, and GATCDA.

TABLE 2 | Top 10 candidate circRNAs related to lung cancer.

Rank	circRNA	Evidence (PMID)
1	hsa_circ_0007385	29372377
2	hsa_circ_0014130	29440731
3	hsa_circ_0016760	33416186
4	hsa_circ_0043256	28958934
5	hsa_circ_0012673	32141533
6	hsa_circRNA_404833	unconfirmed
7	hsa_circRNA_006411	unconfirmed
8	hsa_circRNA_401977	unconfirmed
9	hsa_circ_0013958	28685964
10	hsa_circ_0006404	unconfirmed

Case Studies

In this part, two kinds of case studies are utilized to further assess the reliability of the GATGCN for detecting potential circRNA–disease associations, which calculated the predicted probability matrix *via* a candidate set comprising unproven circRNAs. For the first kind of case study, all known circRNA–disease relationships are selected as training samples, and all unknown circRNA–disease relationships are prioritized

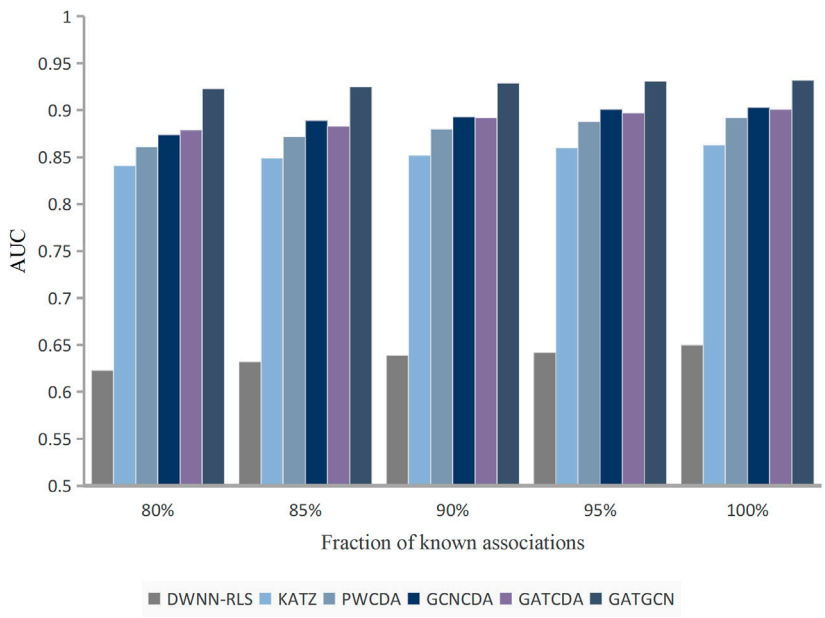


FIGURE 8 | Performance of methods based on various percentages of known relationships.

TABLE 3 | Top 10 candidate circRNAs related to diabetes retinopathy.

Rank	circRNA	Evidence (PMID)
1	hsa_circRNA_063981	28817829
2	hsa_circRNA_404457	28817829
3	hsa_circRNA_100750	28817829
4	hsa_circRNA_406918	28817829
5	hsa_circRNA_104387	28817829
6	hsa_circRNA_103410	28817829
7	hsa_circRNA_100192	28817829
8	hsa_circ_0013509	unconfirmed
9	circSLC8A1-1	unconfirmed
10	hsa_circ_101396	unconfirmed

TABLE 4 | Top 10 candidate circRNAs related to prostate cancer.

Rank	circRNA	Evidence (PMID)
1	circHIPK3	32547085
2	hsa_circ_0004383	unconfirmed
3	circ-Foxo3	31733095
4	hsa-circRNA 2149	unconfirmed
5	circR-284	unconfirmed
6	circDLGAP4	unconfirmed
7	hsa_circ_0008887	unconfirmed
8	hsa_circ_0044516	31625175
9	CDR1as	23900077
10	Cir-ITCH	32904490

TABLE 5 | Top 10 candidate circRNAs related to cholangiocarcinoma.

Rank	circRNA	Evidence (PMID)
1	hsa_circ_000438	unconfirmed
2	circHIPK3	31654054
3	ciRS-7	33390857
4	circR-284	unconfirmed
5	circDLGAP4	unconfirmed
6	circSMARCA5	31880360
7	hsa_circ_0008887	unconfirmed
8	hsa_circ_0006404	unconfirmed
9	hsa_circRNA_000585	34182814
10	hsa_circ_0000673	33221765

TABLE 6 | Top 10 candidate circRNAs related to clear cell renal cell carcinoma.

Rank	circRNA	Evidence (PMID)
1	circHIPK3	32409849
2	circR-284	unconfirmed
3	circDLGAP4	unconfirmed
4	hsa_circ_0004383	unconfirmed
5	Cir-ITCH	unconfirmed
6	hsa_circRNA_003251	unconfirmed
7	circPVT1	33453148
8	hsa_circ_0001451	30271486
9	ciRS-7	32496306
10	circZFR	31571906

according to the corresponding prediction scores. We select the top 10 scores by sorting the scores of the probability matrix in descending order and verified those predicted candidates through validated databases and literature, such as CircR2Disease, CircBase, and PubMed. Eventually, we adopt case studies on lung cancer, diabetes retinopathy, and prostate cancer.

Lung cancer occurs in the bronchial mucosa or glands with the highest incidence and the highest number of deaths in the world. The results in **Table 2** show that six associations are verified by experiments among top 10 predicted candidate circRNAs for lung cancer. For example, the hsa_circ_0007385 (top 1) knockdown resulted in considerable inhibition of the proliferation, invasion, and migration of lung cancer cells (Jiang et al., 2018). Zhang et al. discovered that hsa_circ_0014130 (top 2) exhibited substantially overexpression in NSCLC tissues (Zhang S. et al., 2018). Zhu et al. indicated that hsa_circ_0016760 (top 3) accelerated the malignant growth of NSCLC by sponging miR-145-5p/FGF5 (Zhu et al., 2021).

Diabetes retinopathy is a microvascular complication caused by diabetes, which can be divided into proliferative diabetic retinopathy and non-proliferative diabetic retinopathy. As shown in **Table 3**, the predictive results contain seven experimentally verified associations among the top 10 ranked candidate circRNAs. For instance, hsa_circRNA_063981 (top 1), hsa_circRNA_404457 (top 2), and hsa_circRNA_100750 (top 3) are considerably elevated in the serum of T2DR patients compared to T2DM and control patients (Gu et al., 2017).

Prostate cancer refers to malignant tumors produced by the epithelial cells of the prostate under the action of a variety of carcinogenic factors, which causes bone pain, pathological fractures, and paraplegia. Using the GATGCN, we successfully predict five of 10 top candidate circRNAs for prostate cancer (**Table 4**). The results in the literature indicate that circHIPK3 (top 1) expression is upregulated in prostate cancer cells and prostate cancer tissues (Liu et al., 2020). Kong et al. found that circFOXO3 (top 3) acted as a sponge for miR-29a-3p, exhibiting oncogenic activity in prostate cancer (Kong et al., 2020). Li et al. revealed that hsa_circ_0044516 (top 8) downregulation suppressed prostate cancer cell metastasis and growth (Li T. et al., 2020).

In order to further assess the capacity of GATGCN for detecting new diseases, two common diseases, that is, clear cell renal cell carcinoma and cholangiocarcinoma are chosen for case studies. Specifically, all known associations about clear cell renal cell carcinoma and cholangiocarcinoma are reset to unknown and all candidate circRNAs are prioritized according to corresponding prediction scores. Eventually, we select the top 10 scores to assess the performance of GATGCN for detecting new circRNAs and diseases.

Cholangiocarcinoma is a malignant tumor that originates from the extrahepatic bile duct. The result in **Table 5** shows that five associations are verified among the top 10 ranked candidate circRNAs. For example, Louis et al. demonstrated that the expression of circHIPK3 (top 2) was specifically elevated in cholangiocarcinoma cell lines (Louis et al., 2019). Chen et al. discovered that in cholangiocarcinoma, ciRS-7 (top 3) acts as an oncogene and promotes tumor development by competitively inhibiting miR-7. (Chen et al., 2021). Lu et al.

indicated that circSMARCA5 (top 6) expression was lower in ICC tumor tissues than surrounding tissues (Lu and Fang, 2020).

Clear cell renal cell carcinoma is derived from adenocarcinoma of renal tubular epithelial cells, which forms hemangioma thrombus or metastasizes to lymph nodes and other organs. As shown in **Table 6**, the predicted results contain five experimental verified associations among the top 10 ranked candidate circRNAs. For example, Li et al. discovered that overexpression of circHIPK3 (top 1) substantially reduced CCRCC cell invasion and migration *in vitro* (Li H. et al., 2020). Zheng et al. discovered that circPVT1 (top 7) promotes progression in CCRCC cells by regulating TBX15 expression and sponging miR-145-5p (Zheng et al., 2021). Wang et al. indicated that hsa_circ_0001451 (top 8) upregulation could promote CCRCC cell invasion and proliferation (Wang G. et al., 2018).

The results of the case studies show that GATGCN can efficiently detect the potential circRNA–disease relationships and provide clues for exploring the mechanism between human complex diseases and circRNAs.

CONCLUSION

Cumulative evidence has proved that the development of powerful calculation methods is significant to infer the interactions between diseases and circRNAs. These calculation models address challenges of high cost and high time consumption in conventional biological experiments. In this study, an advanced calculation method called GATGCN is designed to discover potential circRNA–disease relationships *via* graph attention mechanism and graph convolutional network. First, multisource similarity data for circRNAs and diseases are fused by the centered kernel alignment model. Second, the graph attention network is deployed to learn the dense representation of nodes on the disease–disease similarity network and circRNA–circRNA similarity network. Third, the heterogenous network is constructed by connecting known circRNA–disease associations, feature matrix of diseases, and feature matrix of circRNAs. Finally, the graph convolutional network is applied to get prediction scores based on the constructed heterogenous network. To further confirm the advantage of GATGCN for detecting circRNA–disease interactions, we compare it with several state-of-the-art prediction models under five-fold cross-validation. The results indicate that GATGCN achieves significant performance among compared methods. Meanwhile, the case study substantiates the excellent capability of the GATGCN for detecting potential circRNA–disease relationships. In conclusion, GATGCN is a powerful and promising approach for detecting circRNA–disease relationships.

Although we have integrated multisource biological information and utilized graph attention network and graph convolutional network to learn latent representation for diseases and circRNAs, there is still room to strengthen the predictive capability of the model. On the one hand, a large number of nonlinear features are extracted to detect circRNA–disease associations, which ignore the importance of linear features. We could further solve this problem by fusing nonlinear features and linear features to enhance the stability of our model. On the other hand, feature aggregation in excessive network layers could affect the expression of initial feature information. In the future, we can splice the representations of nodes in different layers as node features.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding authors. The GATGCN dataset and code can be downloaded from https://github.com/ghli16/GATGCN.

AUTHOR CONTRIBUTIONS

GL and JL conceived, designed, and managed the study. DW developed the GATGCN model and wrote the original manuscript. YZ revised the original draft. CL and QX discussed the GATGCN model and provided further research. All authors read and approved the final manuscript.

FUNDING

This work has been supported by the National Natural Science Foundation of China (Grant Nos. 61862025, 61873089, 62002116, 11862006, and 92159102), Natural Science Foundation of Jiangxi Province of China (Grant Nos. 20212BAB202009, 20181BAB211016, and 20202BAB205011).

ACKNOWLEDGMENTS

We would like to thank all authors of the cited references.

REFERENCES

Bian, C., Lei, X.-J., and Wu, F.-X. (2021). GATCDA: Predicting circRNA-Disease Associations Based on Graph Attention Network. *Cancers* 13 (11), 2595. doi:10.3390/cancers13112595

Chen, X., Han, P., Zhou, T., Guo, X., Song, X., and Li, Y. (2016). circRNADb: a Comprehensive Database for Human Circular RNAs with Protein-Coding Annotations. *Sci. Rep.* 6 (1), 1–6. doi:10.1038/srep34985

Chen, J., Cui, L., Yuan, J., Zhang, Y., and Sang, H. (2017). Circular RNA WDR77 Target FGF-2 to Regulate Vascular Smooth Muscle Cells Proliferation and Migration by Sponging miR-124. *Biochem. Biophys. Res. Commun.* 494 (1-2), 126–132. doi:10.1016/j.bbrc.2017.10.068

Chen, M., Peng, Y., Li, A., Li, Z., Deng, Y., Liu, W., et al. (2018). A Novel Information Diffusion Method Based on Network Consistency for Identifying Disease Related Micrornas. *RSC Adv.* 8 (64), 36675–36690. doi:10.1039/C8RA07519K

Chen, J., Yang, J., Fei, X., Wang, X., and Wang, K. (2021). CircRNA ciRS-7: a Novel Oncogene in Multiple Cancers. *Int. J. Biol. Sci.* 17 (1), 379–389. doi:10.7150/ijbs.54292

Clevert, D-A., Unterthiner, T., and Hochreiter, S. (2016). "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," in International

Conference on Learning Representations (ICLR), San Juan, Puerto Rico, May 2–4, 2016. Available at: https://arxiv.org/abs/1511.07289.

Cristianini, N., Kandola, J., Elisseeff, A., and Shawe-Taylor, J. (2006). "On Kernel Target Alignment," in *Innovations in Machine Learning* (Berlin, Heidelberg: Springer), 205–256. doi:10.1007/3-540-33486-6_8

Ding, Y., Tang, J., and Guo, F. (2019). Identification of Drug-Side Effect Association *via* Multiple Information Integration with Centered Kernel Alignment. *Neurocomputing* 325, 211–224. doi:10.1016/j.neucom.2018.10.028

Fan, C., Lei, X., and Wu, F.-X. (2018a). Prediction of CircRNA-Disease Associations Using KATZ Model Based on Heterogeneous Networks. *Int. J. Biol. Sci.* 14 (14), 1950–1959. doi:10.7150/ijbs.28260

Fan, C., Lei, X., Fang, Z., Jiang, Q., and Wu, F.-X. (2018b). CircR2Disease: a Manually Curated Database for Experimentally Supported Circular RNAs Associated with Various Diseases. *Database* 2018, bay044. doi:10.1093/database/bay044

Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013). Circ2Traits: a Comprehensive Database for Circular RNA Potentially Associated with Disease and Traits. *Front. Genet.* 4, 283. doi:10.3389/fgene.2013.00283

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

Gu, Y., Ke, G., Wang, L., Zhou, E., Zhu, K., and Wei, Y. (2017). Altered Expression Profile of Circular RNAs in the Serum of Patients with Diabetic Retinopathy Revealed by Microarray. *Ophthalmic Res.* 58 (3), 176–184. doi:10.1159/000479156

Han, B., Chao, J., and Yao, H. (2018). Circular RNA and its Mechanisms in Disease: from the Bench to the Clinic. *Pharmacol. Ther.* 187, 31–44. doi:10.1016/j.pharmthera.2018.01.010

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 (7441), 384–388. doi:10.1038/nature11993

Huang, Y.-a., Hu, P., Chan, K. C. C., and You, Z.-H. (2020). Graph Convolution for Predicting Associations between miRNA and Drug Resistance. *Bioinformatics* 36 (3), 851–858. doi:10.1093/bioinformatics/btz621

Jiang, M.-M., Mai, Z.-T., Wan, S.-Z., Chi, Y.-M., Zhang, X., Sun, B.-H., et al. (2018). Microarray Profiles Reveal that Circular RNA Hsa_circ_0007385 Functions as an Oncogene in Non-small Cell Lung Cancer Tumorigenesis. *J. Cancer Res. Clin. Oncol.* 144 (4), 667–674. doi:10.1007/s00432-017-2576-2

Kipf, T. N., and Welling, M. (2016). *Semi-supervised Classification with Graph Convolutional Networks*. arXiv preprint arXiv:1609.02907. Available at: https://arxiv.org/abs/1609.02907.

Kong, Z., Wan, X., Lu, Y., Zhang, Y., Huang, Y., Xu, Y., et al. (2020). Circular RNA circFOXO3 Promotes Prostate Cancer Progression through Sponging miR-29a-3p. *J. Cel Mol Med* 24 (1), 799–813. doi:10.1111/jcmm.14791

Lei, X., and Bian, C. (2020). Integrating Random Walk with Restart and K-Nearest Neighbor to Identify Novel circRNA-Disease Association. *Sci. Rep.* 10 (1), 1–9. doi:10.1038/s41598-020-59040-0

Lei, X., Fang, Z., Chen, L., and Wu, F.-X. (2018). PWCDA: Path Weighted Method for Predicting circRNA-Disease Associations. *Ijms* 19 (11), 3410. doi:10.3390/ijms19113410

Lei, X., Mudiyanselage, T. B., Zhang, Y., Bian, C., Lan, W., Yu, N., et al. (2021). A Comprehensive Survey on Computational Methods of Non-coding RNA and Disease Association Prediction. *Brief. Bioinform.* 22 (4), bbaa350. doi:10.1093/bib/bbaa350

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. *Nucl. Acids Res.* 42 (D1), D1070–D1074. doi:10.1093/nar/gkt1023

Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA Is Enriched and Stable in Exosomes: a Promising Biomarker for Cancer Diagnosis. *Cel. Res.* 25, 981–984. doi:10.1038/cr.2015.82

Li, G., Yue, Y., Liang, C., Xiao, Q., Ding, P., and Luo, J. (2019). NCPCDA: Network Consistency Projection for circRNA-Disease Association Prediction. *RSC Adv.* 9 (57), 33222–33228. doi:10.1039/C9RA06133A

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. Inform.* 112, 103624. doi:10.1016/j.jbi.2020.103624

Li, H., Heng, B., Ouyang, P., Xie, X., Zhang, T., Chen, G., et al. (2020). Comprehensive Profiling of circRNAs and the Tumor Suppressor Function of circHIPK3 in clear Cell Renal Carcinoma. *J. Mol. Hist.* 51 (3), 317–327. doi:10.1007/s10735-020-09882-9

Li, M., Liu, M., Bin, Y., and Xia, J. (2020). Prediction of circRNA-Disease Associations Based on Inductive Matrix Completion. *BMC Med. Genomics* 13 (5), 1–13. doi:10.1186/s12920-020-0679-0

Li, T., Sun, X., and Chen, L. (2020). Exosome Circ_0044516 Promotes Prostate Cancer Cell Proliferation and Metastasis as a Potential Biomarker. *J. Cel. Biochem.* 121 (3), 2118–2126. doi:10.1002/jcb.28239

Liu, F., Fan, Y., Ou, L., Li, T., Fan, J., Duan, L., et al. (2020). CircHIPK3 Facilitates the G2/M Transition in Prostate Cancer Cells by Sponging miR-338-3p. *Ott* 13, 4545–4558. doi:10.2147/OTT.S242482

Lihong, P., Wang, C., Tian, X., Zhou, L., and Li, K. (2021). Finding lncRNA-Protein Interactions Based on Deep Learning with Dual-Net Neural Architecture. *Ieee/acm Trans. Comput. Biol. Bioinf.* 14 (8), 1. doi:10.1109/TCBB.2021.3116232

Louis, C., Desoteux, M., and Coulouarn, C. (2019). Exosomal circRNAs: New Players in the Field of Cholangiocarcinoma. *Clin. Sci. (Lond).* 133 (21), 2239–2244. doi:10.1042/CS20190940

Lu, Q., and Fang, T. (2020). Circular RNA SMARCA5 Correlates with Favorable Clinical Tumor Features and Prognosis, and Increases Chemotherapy Sensitivity in Intrahepatic Cholangiocarcinoma. *J. Clin. Lab. Anal.* 34 (4), e23138. doi:10.1002/jcla.23138

Lu, C., Zeng, M., Zhang, F., Wu, F.-X., Li, M., and Wang, J. (2021). Deep Matrix Factorization Improves Prediction of Human circRNA-Disease Associations. *IEEE J. Biomed. Health Inform.* 25 (3), 891–899. doi:10.1109/JBHI.2020.2999638

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

Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., et al. (2017). CircRNA: Functions and Properties of a Novel Potential Biomarker for Cancer. *Mol. Cancer* 16 (1), 1–8. doi:10.1186/s12943-017-0663-2

Peng, L., Chen, Y., Ma, N., and Chen, X. (2017). NARRMDA: Negative-Aware and Rating-Based Recommendation Algorithm for miRNA-Disease Association Prediction. *Mol. Biosyst.* 13 (12), 2650–2659. doi:10.1039/c7mb00499k

Peng, L.-H., Sun, C.-N., Guan, N.-N., Li, J.-Q., and Chen, X. (2018). HNMDA: Heterogeneous Network-Based miRNA-Disease Association Prediction. *Mol. Genet. Genomics* 293 (4), 983–995. doi:10.1007/s00438-018-1438-1

Peng, L., Shen, L., Liao, L., Liu, G., and Zhou, L. (2020a). RNMFMDA: A Microbe-Disease Association Identification Method Based on Reliable Negative Sample Selection and Logistic Matrix Factorization with Neighborhood Regularization. *Front. Microbiol.* 11, 592430. doi:10.3389/fmicb.2020.592430

Peng, L.-H., Zhou, L.-Q., Chen, X., and Piao, X. (2020b). A Computational Study of Potential miRNA-Disease Association Inference Based on Ensemble Learning and Kernel ridge Regression. *Front. Bioeng. Biotechnol.* 8, 40. doi:10.3389/fbioe.2020.00040

Rong, D., Sun, H., Li, Z., Liu, S., Dong, C., Fu, K., et al. (2017). An Emerging Function of circRNA-miRNAs-mRNA axis in Human Diseases. *Oncotarget* 8 (42), 73271–73281. doi:10.18632/oncotarget.19154

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). *Graph Attention Networks*. arXiv preprint arXiv:1710.10903. Available at: https://arxiv.org/abs/1710.10903.

Vo, J. N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., et al. (2019). The Landscape of Circular RNA in Cancer. *Cell* 176 (4), 869–881. doi:10.1016/j.cell.2018.12.021

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, L., You, Z.-H., Huang, Y.-A., Huang, D.-S., and Chan, K. C. C. (2020a). An Efficient Approach Based on Multi-Sources Information to Predict circRNA-Disease Associations Using Deep Convolutional Neural Network. *Bioinformatics* 36 (13), 4038–4046. doi:10.1093/bioinformatics/btz825

Wang, L., You, Z.-H., Li, Y.-M., Zheng, K., and Huang, Y.-A. (2020b). GCNCDA: A New Method for Predicting circRNA-Disease Associations Based on Graph Convolutional Network Algorithm. *Plos Comput. Biol.* 16 (5), e1007568. doi:10.1371/journal.pcbi.1007568

Wang, H., Tang, J., Ding, Y., and Guo, F. (2021). Exploring Associations of Non-coding Rnas in Human Diseases *via* Three-Matrix Factorization with

Hypergraph-Regular Terms on center Kernel Alignment. *Brief. Bioinform.* 22, bbaa409. doi:10.1093/bib/bbaa409

Wang, G., Xue, W., Jian, W., Liu, P., Wang, Z., Wang, C., et al. (2018). The Effect of Hsa_circ_0001451 in clear Cell Renal Cell Carcinoma Cells and its Relationship with Clinicopathological Features. *J. Cancer* 9 (18), 3269–3277. doi:10.7150/jca.25902

Wang, L., Shen, J., and Jiang, Y. (2018). Circ_0027599/PHDLA1 Suppresses Gastric Cancer Progression by Sponging miR-101-3p.1. *Cel. Biosci.* 8, 58. doi:10.1186/s13578-018-0252-0

Wang, P., Zhu, W., Liao, B., Cai, L., Peng, L., Yang, J., et al. (2018). Predicting Influenza Antigenicity by Matrix Completion with Antigen and Antiserum Similarity. *Front. Microbiol.* 9, 2500. doi:10.3389/fmicb.2018.02500

Wei, H., and Liu, B. (2020). iCircDA-MF: Identification of circRNA-Disease Associations Based on Matrix Factorization. *Brief. Bioinformatics* 21 (4), 1356–1367. doi:10.1093/bib/bbz057

Xiao, Q., Dai, J., and Luo, J. (2022). A Survey of Circular RNAs in Complex Diseases: Databases, Tools and Computational Methods. *Brief. Bioinformatics* 23(1), bbab444. doi:10.1093/bib/bbab444

Xiao, Q., Fu, Y., Yang, Y., Dai, J., and Luo, J. (2021). NSL2CD: Identifying Potential circRNA-Disease Associations Based on Network Embedding and Subspace Learning. *Brief. Bioinform.* 22, bbab177. doi:10.1093/bib/bbab177

Yan, C., Wang, J., and Wu, F.-X. (2018). DWNN-RLS: Regularized Least Squares Method for Predicting circRNA-Disease Associations. *BMC Bioinformatics* 19 (19), 73–81. doi:10.1186/s12859-018-2522-6

Yao, D., Zhang, L., Zheng, M., Sun, X., Lu, Y., and Liu, P. (2018). Circ2Disease: a Manually Curated Database of Experimentally Validated circRNAs in Human Disease. *Sci. Rep.* 8 (1), 1–6. doi:10.1038/s41598-018-29360-3

Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting Disease-Associated Circular RNAs Using Deep Forests Combined with Positive-Unlabeled Learning Methods. *Brief. Bioinformatics* 21 (4), 1425–1436. doi:10.1093/bib/bbz080

Zhang, Y., Sun, L., Xuan, L., Pan, Z., Li, K., Liu, S., et al. (2016). Reciprocal Changes of Circulating Long Non-Coding RNAs ZFAS1 and CDR1AS Predict Acute Myocardial Infarction. *Sci. Rep.* 6, 22384. doi:10.1038/srep22384

Zhang, W., Yu, C., Wang, X., and Liu, F. (2019). Predicting CircRNA-Disease Associations through Linear Neighborhood Label Propagation Method. *IEEE Access* 7, 83474–83483. doi:10.1109/ACCESS.2019.2920942

Zhang, Y., Chen, M., Huang, L., Xie, X., Li, X., Jin, H., et al. (2021a). Fusion of KATZ Measure and Space Projection to Fast Probe Potential lncRNA-Disease Associations in Bipartite Graphs. *PLoS ONE* 16 (11), e0260329. doi:10.1371/journal.pone.0260329

Zhang, S., Zeng, X., Ding, T., Guo, L., Li, Y., Ou, S., et al. (2018). Microarray Profile of Circular RNAs Identifies Hsa_circ_0014130 as a New Circular RNA Biomarker in Non-small Cell Lung Cancer. *Sci. Rep.* 8 (1), 2878. doi:10.1038/s41598-018-21300-5

Zhang, Y., Gao, T., Li, X., Wen, C.-C., Yan, X.-T., Peng, C., et al. (2021b). Circ_0005075 Targeting miR-151a-3p Promotes Neuropathic Pain in CCI Rats *via* Inducing NOTCH2 Expression. *Gene* 767, 145079. doi:10.1016/j.gene.2020.145079

Zhang, Z., Yang, T., and Xiao, J. (2018). Circular RNAs: Promising Biomarkers for Human Diseases. *EBioMedicine* 34, 267–274. doi:10.1016/j.ebiom.2018.07.036

Zhao, Z., Wang, K., Wu, F., Wang, W., Zhang, K., Hu, H., et al. (2018). circRNA Disease: a Manually Curated Database of Experimentally Supported circRNA-Disease Associations. *Cel. Death Dis.* 9 (5), 1–2. doi:10.1038/s41419-018-0503-3

Zhao, Q., Yang, Y., Ren, G., Ge, E., and Fan, C. (2019). Integrating Bipartite Network Projection and KATZ Measure to Identify Novel CircRNA-Disease Associations. *IEEE Trans.on Nanobiosci.* 18 (4), 578–584. doi:10.1109/TNB.2019.2922214

Zheng, Z., Chen, Z., Zhong, Q., Zhu, D., Xie, Y., Shangguan, W., et al. (2021). CircPVT1 Promotes Progression in clear Cell Renal Cell Carcinoma by Sponging miR-145-5p and Regulating TBX15 Expression. *Cancer Sci.* 112 (4), 1443–1456. doi:10.1111/cas.14814

Zhou, L., Wang, Z., Tian, X., and Peng, L. (2021). LPI-DeepGBDT: A Multiple-Layer Deep Framework Based on Gradient Boosting Decision Trees for lncRNA-Protein Interaction Identification. *BMC Bioinformatics* 22 (479), 1–24. doi:10.1186/s12859-021-04399-8

Zhu, L.-P., He, Y.-J., Hou, J.-C., Chen, X., Zhou, S.-Y., Yang, S.-J., et al. (2017). The Role of circRNAs in Cancers. *Biosci. Rep.* 37 (5), BSR20170750. doi:10.1042/BSR20170750

Zhu, Z., Wu, Q., Zhang, M., Tong, J., Zhong, B., and Yuan, K. (2021). Hsa_circ_0016760 Exacerbates the Malignant Development of Non-Small Cell Lung Cancer by Sponging miR-145-5p/FGF5. *Oncol. Rep.* 45 (2), 501–512. doi:10.3892/or.2020.7899

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 February 2022
doi: 10.3389/fgene.2022.816035

Genome-Wide Identification of Immune-Related Alternative Splicing and Splicing Regulators Involved in Abdominal Aortic Aneurysm

Shiyong Wu[1], *Shibiao Liu*[1], *Ningheng Chen*[1], *Chuang Zhang*[1], *Hairong Zhang*[2]* and *Xueli Guo*[1]*

[1]Department of Vascular Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China, [2]Department of Colorectal and Anal Surgery, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

The molecular mechanism of AAA formation is still poorly understood and has not been fully elucidated. The study was designed to identify the immune-related genes, immune-RAS in AAA using bioinformatics methods. The GSE175683 datasets were downloaded from the GEO database. The DEseq2 software was used to identify differentially expressed genes (DEGs). SUVA pipeline was used to quantify AS events and RAS events. KOBAS 2.0 server was used to identify GO terms and KEGG pathways to sort out functional categories of DEGs. The CIBERSORT algorithm was used with the default parameter for estimating immune cell fractions. Nine samples from GSE175683 were used to construct the co-disturbed network between expression of SFs and splicing ratio of RAS events. PCA analysis was performed by R package factoextra to show the clustering of samples, and the pheatmap package in R was used to perform the clustering based on Euclidean distance. The results showed that there were 3,541 genes significantly differentially expressed, of which 177 immune-related genes were upregulated and 48 immune-related genes were downregulated between the WT and WTA group. Immune-RAS events were mainly alt5P and IR events, and about 60% of it was complex splicing events in AAA. The WT group and the WTA group can be clearly distinguished in the first principal component by using the splicing ratio of immune-RAS events. Two downregulated genes, Nr4a1 and Nr4a2, and eight upregulated genes, Adipor2, Akt2, Bcl3, Dhx58, Pparg, Ptgds, Sytl1, and Vegfa were identified among the immune-related genes with RAS and DEGs. Eighteen differentially expressed SFs were identified and displayed by heatmap. The proportion of different types of cells and ratio of the average ratio of different cells were quite different. Both M1 and M2 types of macrophages and plasma cells were upregulated, while M0 type was downregulated in AAA. The proportion of plasma cells in the WTA group had sharply increased. There is a correlation between SF expression and immune cells/immune-RAS. Sf3b1, a splicing factor with significantly different expression, was selected to bind on a mass of immune-related genes. In conclusion, our results showed that immune-related genes, immune-RAS, and SFs by genome-wide identification were involved in AAA.

Keywords: abdominal aortic aneurysm, alternative splicing, RNA-Seq, immune-related genes, splicing factor, genome-wide identification

OPEN ACCESS

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

Reviewed by:
Manali Rupji,
Emory University, United States
Basavaraj Mallikarjunayya Vastrad,
KLE Society's College Of Pharmacy, India

***Correspondence:**
Hairong Zhang
18838221713@163.com
Xueli Guo
guoxueli2000@163.com

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

Received: *16 November 2021*
Accepted: *06 January 2022*
Published: *17 February 2022*

Citation:
Wu S, Liu S, Chen N, Zhang C, Zhang H and Guo X (2022) Genome-Wide Identification of Immune-Related Alternative Splicing and Splicing Regulators Involved in Abdominal Aortic Aneurysm. Front. Genet. 13:816035. doi: 10.3389/fgene.2022.816035

INTRODUCTION

Abdominal aortic aneurysm (AAA) refers to the permanent and localized expansion of the abdominal aortic wall exceeding 50% of the normal vascular diameter, and is usually diagnosed when the abdominal aorta is more than 3 cm in diameter (Chaikof et al., 2018; Cai et al., 2021). AAA is a disease of the cardiovascular system with severe complications, mainly manifested in the lower renal aorta. With the progression of the disease and the increase in the inner diameter of the aorta, the risk of AAA rupture increases. AAA rupture represents a life-threatening complication of aneurysms with an overall mortality rate of up to 90% in western countries. AAA accounted for 1.3% of deaths among men aged 65 to 85 in developed countries (Sakalihasan et al., 2005). The incidence of AAA has steadily increased in most developed countries, rising from 1.6% to 7.2% of the general population 60–65 years or older (Guirguis-Blake et al., 2019). AAA is now the 10th leading cause of death in western countries, and its incidence is rising (Brangsch et al., 2017). AAA is related to advanced age, men, smoking, atherosclerosis, high blood pressure, and genetic predisposition (Annambhotla et al., 2008). Although important evidence has emerged in the past decade, the molecular mechanism of AAA formation is still poorly understood, and the exact reasons for the occurrence and its development have not been fully elucidated (Brangsch et al., 2017; Li H. et al., 2021). At present, the treatment of AAA is still mainly surgery, only the innovation of endovascular treatment (Wanhainen et al., 2019). With the advancement of the human genome, understanding exactly which molecules and genes mediate the development of AAA and blocking their activity at the molecular level may lead to important new discoveries and treatments.

AAA is a fatal vascular disease in human, which is a chronic degenerative disease of abdominal aorta. In this process, the inflammatory responses and immune system work effectively through the attraction of inflammatory cells, the secretion of proinflammatory factors, and the subsequent upregulation of MMP (Li et al., 2018). Inflammation is an important part of the immune system. A large number of exogenous immune cells, including macrophages, lymphocytes, neutrophils, mast cells, and natural killer cells, gradually infiltrate into the tissue from adventitia to intima, triggering a series of inflammatory reactions (Rateri et al., 2011; Wang et al., 2014; Yan et al., 2016). The adaptive and innate immune system plays an important role in the initiation and propagation of the inflammatory response in aortic tissue. Recently increased knowledge indicates that the immune process is involved in the pathogenesis of AAA (Jagadesham et al., 2008; Liu et al., 2015). Some immune cells such as macrophages, CD^{4+} T cells, and B cells play an important role in the diseased aortic wall through phenotypic regulation (Maiellaro and Taylor, 2007; Schaheen et al., 2016). Additionally, immunoglobulin also has a great influence on the function and differentiation of immune cells in AAA. Recent evidence suggests that innate immune system, especially Toll-like receptors, chemokine receptors, and complements are involved in the progression of AAA (Li et al., 2018). The current understanding may provide new insights into the role of inflammation and immune response in AAA. Based on tissue gene expression profiles and specific gene expression profiles of various immune cells, some methods have been developed to allow the quantification of immune cell composition through traditional gene profiling methods, including a large number of RNA-seq, such as EPIC, TIMER. and CIBERSORT (Finotello and Trajanoski, 2018). However, the composition of immune cells in the process of AAA is dynamic, and the factors affecting immune infiltration are not fully understood. Therefore, regulation of immune inflammatory response is an emerging molecular target for AAA (Li et al., 2018).

Alternative splicing (AS) plays an immunomodulatory role in many diseases. The regulated alternative splicing events located in immune-related genes (immune-RAS) is a new kind of drug target and an important biomarker in clinical diagnosis. Abnormal immune-RAS is an important factor in the occurrence and development of many diseases including tumors (Li et al., 2019; Bonnal et al., 2020). The research of Sanela et al. showed that AS was a common feature of thoracic aortic aneurysm (TAA) formation, and AS in the TGF-β pathway could be used to characterize patients with bicuspid aortic valve and tricuspid aortic valve TAA (Kurtovic et al., 2011). The study of Zhao et al. revealed the pivotal role of the AS change of XBP1 in maintaining the VSMC contractile phenotype and providing protection from aortic aneurysm formation (Zhao et al., 2017). At present, there are a small number of reports on the role of AS in the development of AAA, but there is a rare report on the role of AS in the immune regulation of AAA. In this view, we will discuss immune-related genes and its regulation of AS, and provide new mechanism insights for the development of immune-targeted therapy in AAA.

MATERIALS AND METHODS

Retrieval and Process of Public Data

Public sequence data files GSE175683 (Li H. et al., 2021) were downloaded from the Sequence Read Archive (SRA). SRA run files were converted to fastq format with NCBI SRA Tool fastq-dump. The raw reads were trimmed of low-quality bases by using a FASTX-Toolkit (v.0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). Then the clean reads were evaluated using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
SF3B1-bound peaks were downloaded from Encodeproject (https://www.encodeproject.org/) (ENCSR133QEA).

Read Alignment and Differentially Expressed Gene Analysis

Clean reads were aligned to the mouse GRCm39 genome by HISAT2 (Kim et al., 2019). Uniquely mapped reads were ultimately used to calculate read number and reads per kilobase of exon per million fragments mapped (FPKM) for each gene. The expression levels of genes were evaluated using FPKM. When we do gene differential expression analysis, we choose the software DEseq2 (Love et al., 2014). DEseq2 will model the original reads and use the scale factor to explain the

difference of Library depth. Then DEseq2 estimates the gene dispersion, and reduces these estimates to produce more accurate dispersion estimates, so as to model the reads count. Finally, the model of negative binomial distribution is fitted by DEseq2, and the hypothesis is tested by Wald test or likelihood ratio test. DEseq2 can be used to analyze the differential expression between two or more samples, and the analysis results can be used to determine whether a gene is differentially expressed by fold change (FC) and false discovery rate (FDR).

There are two important parameters

1) FC: fold change, the absolute ratio of expression change.
2) FDR: false discovery rate.

The criteria of significant difference expression were as follows

FC ≥ 2(up) or ≤0.5(down), FDR ≤ 0.05.

Alternative Splicing Analysis

The AS events and regulated alternative splicing events (RAS) among different groups were defined and quantified by using the SUVA pipeline as described previously (Cheng et al., 2021). Reads proportion of SUVA AS event (pSAR) of each AS events were calculated. Immune-related genes (1,793) (https://www.immport.org/shared/genelists/) were regained from the ImmPort database. The regulated alternative splicing events located in immune-related genes (immune-RAS) were screened and analyzed.

Co-Expression Analysis

The co-disturbed network between expression of splicing factors and splicing ratio of RAS events (pSAR ≥90%) was constructed using nine samples from GSE175683. We calculated the Pearson's correlation coefficients (PCCs) between them and classified their relation into three classes: positive correlated, negative correlated, and non-correlated based on the PCCs value. |Pearson's correlation| ≥0.8 and p-value ≤0.01 were retained.

Functional Enrichment Analysis

To sort out functional categories of DEGs, Gene Ontology (GO) terms and KEGG pathways were identified using the KOBAS 2.0 server (Xie et al., 2011). Hypergeometric test and Benjamini–Hochberg FDR controlling procedure were used to define the enrichment of each term.

Cell-type Quantification

The CIBERSORT algorithm (Newman et al., 2015) (v1.03) was used with the default parameter for estimating immune cell fractions using FPKM values of each expressed gene. A total of 22 immune cell phenotypes were analyzed in the study, including seven T-cell types [CD8 T cells, naive CD4 T cells, memory CD4 resting T cells, memory CD4 activated T cells, T follicular helper cells, and regulatory T cells (Tregs)]; naive and memory B cells; plasma cells; resting and activated NK cells; monocytes; macrophages M0, M1, and M2; resting and activated dendritic cells; resting and activated mast cells; eosinophils; and neutrophils.

Other Statistical Analysis

Principal component analysis (PCA) was performed by R package factoextra (https://cloud.r-project.org/package=factoextra/) to show the clustering of samples with the first two components. After normalizing the reads by TPM (tags per million) of each gene in samples, in house-script (Sogen) was used for visualization of next-generation sequence data and genomic annotations. The pheatmap package (https://cran.r-project.org/web/packages/pheatmap/index.html/) in R was used to perform the clustering based on Euclidean distance. Student's t-test was used for comparisons between two groups.

RESULT

Transcriptome Analysis of DEGs in WT-AngII Group and WT-Saline Group Samples

In the study, the RNA-seq data of 10 mice AAA model samples of GSE175683 were downloaded from GEO database. Five mice were the control group (WT) perfused with saline and five mice were the experimental group (WTA) perfused with angiotensin II. In our basic analysis, it was found that the sample WTA4 was seriously outlier, which may affect the subsequent analysis results, so this sample was eliminated. Compared with the WT group, a large number of gene transcription levels have changed in the WTA group. There were 3,541 genes significantly differentially expressed, of which 2,627 genes were upregulated and 914 genes were downregulated (**Figure 1A**; **Supplementary Figure S1**). Compared with the WT group, functional enrichment analysis of DEGs was conducted in the WTA group, and it was found that the upregulated genes were mainly enriched in the signaling pathways of mitochondrial respiratory chain complex I assembly, fatty acid metabolism process, oxidation–reduction process, fatty acid beta-oxidation, mitochondrial electron transport, NADH to ubiquitin, electron transport chain, lipid metabolism process, brown fat cell differentiation, respiratory electron transport chain, and fatty acid biosynthetic process (**Figure 1B**). Compared with the WT group, the WTA group has downregulated genes mainly enriched in signaling pathways in cell adhesion, extracellular matrix organization, positive regulation of pri-miRNA transcription by RNA polymerase II, axon guidance, response to mechanical stimulus, transforming growth factor beta receptor signaling pathway, positive regulation of apoptotic process, nervous system development, angiogenesis, and cell-substrate adhesion (**Figure 1C**). As shown in **Figure 1D**, 1793 immune-related genes were downloaded from the ImmunePort database, and Venn diagram was used to show the number of immune genes with significantly different expression levels in the WTA group compared with the WT group. Among them, 177 genes were upregulated, and 48 genes were downregulated, indicating that the expression levels of a large number of immune-related genes were also regulated. In this study, we focused on the AS of

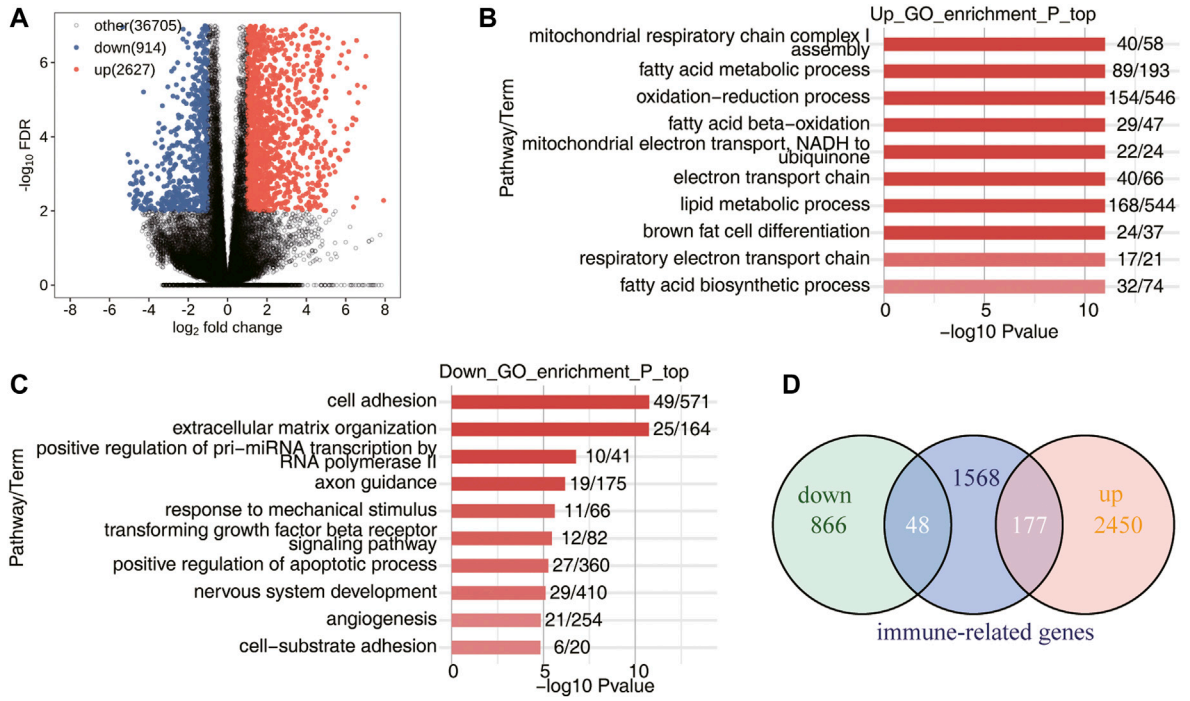


FIGURE 1 | Transcriptome analysis of DEGs in WT-AngII group (WTA) and WT-Saline group (WT) samples. **(A)** Volcano plot shows all DEGs between WTA and WT groups. False discovery rate (FDR) ≤0.05 and FC (fold change) ≥2 (up) or ≤0.5 (down). **(B)** Bar plot exhibited the most enriched Gene Ontology (GO) biological process results of the upregulated genes between the WTA and WT groups. **(C)** Bar plot exhibits the most enriched GO biological process results of the downregulated genes between the WTA and WT groups. **(D)** Venn diagram shows the immune-related genes involved in DEGs between WTA and WT groups.

immune genes, but it also shows that the regulation of immune genes of the body is multilayered.

Identification of WTA-Associated Alternative Splicing vents Located in Immune-Related Genes

We used the recently published AS analysis software SUVA to analyze and identify AS events that are significantly different between the WT and WTA group. Immune-RAS events were specifically showed, and the main splicing events in the transcript (pSAR ≥90%) were displayed. As shown in **Figure 2A** and **Supplementary Figure S2A**, immune-RAS events identified by SUVA were mainly alt5P, alt3p and IR. The splicing events were corresponding to classical splicing events, in which A5SS events accounted for a large proportion, which may be one of the characteristics of immune-RAS and RAS (**Figure 2B** and **Supplementary Figure S2B**). As shown in **Figure 2C**, about 60% of immune-RAS events were complex splicing events, indicating the complexity of immune-RAS regulation of AAA. As shown in **Figure 2D**, the WTA group and the WT group can be clearly distinguished in the first principal component by using the splicing ratio of immune-RAS events for PCA analysis. A splicing event involves two transcripts, and these two transcripts may only account for a very small part of the expression of the whole gene. We hope to find a more dominant transcript undergo AS which was quantified as "pSAR" value by SUVA. The number of splicing events accounting for different proportions of all reads in the region is shown in **Figure 2E**; **Supplementary Figure S2C**. Some splicing events accounted for only a small proportion, so the immune-RAS events with pSAR ≥90% were selected for follow-up analysis. As shown in **Figure 2F** the heatmap was used to show the splicing events of the dominant transcripts in immune-RAS events. As shown in **Figure 2G**, the Venn diagram showed the common and unique genes among the immune-related genes with RAS (RASGs) and DEGs. These splicing events were significantly regulated in AAA, and the transcripts produced by its splicing were the main transcription products of genes, which were worthy of in-depth study and also potential therapeutic targets. A splicing event on Vegfa gene is demonstrated in **Figures 2H, I**, which is an exon skipping event. The intermediate exon reservation transcripts are mainly selected in WTA samples, while exon jump transcripts are more selected in WT samples. A splicing event on another immune gene Pparg is shown in **Figures 2J, K**, which was a altered first exon event. The shorter transcripts were mainly selected in the WTA group, and longer transcripts were more selected in the WT group.

Construction of Co-disturbed Network Between SFs and immune-Regulated Alternative Splicing

This part mainly displayed the differentially expressed SFs in WTA and WT samples, as well as immune-RAS that may be

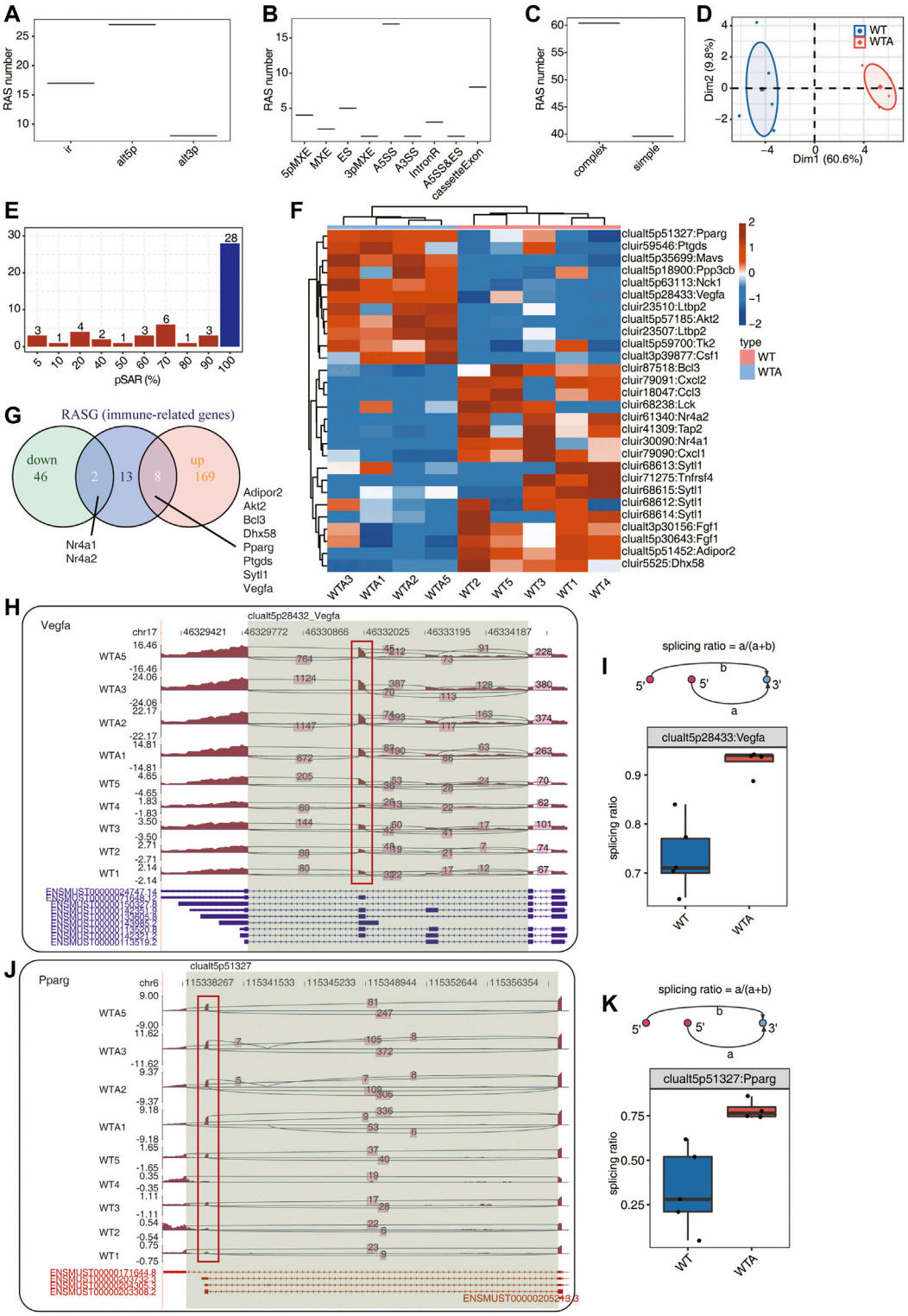


FIGURE 2 | Identification of WTA-associated alternative splicing (AS) events located in immune-related genes (immune-RAS). **(A)** Boxplot showing the number of immune-RAS detected by SUVA, which were altered spliced between the WTA and WT groups. **(B)** Splice junction constituting immune-RAS events detected by SUVA was annotated to classical AS event types, and the number of each classical AS event types is shown with boxplot. **(C)** Boxplot showing number of SUVA immune-RAS events, which contains SJs involved in two or more different classical splicing events (complex) or in the same classical splicing event (simple). **(D)** PCA of splicing ratio of immune-RAS in which frequency ≥40% and pSAR (read proportion of SUVA AS event) ≥50%. The samples were grouped by tumor or normal, and the ellipse for

(*Continued*)

FIGURE 2 | each group is the confidence ellipse. **(E)** Bar plot showing immune-RAS number with different abundance (pSAR) of all detected regulated alternative splicing events (RAS). **(F)** Heatmap of splicing ratio across all samples for immune-RAS which pSASR ≥90% and corresponding genes. **(G)** Venn diagram showing the common and unique genes among the immune-related RASGs and DEGs. **(H)** Visualization of junction reads distribution of vascular endothelial growth factor A (Vegfa) in AS events clualt5p28432 from different groups. Splice junctions were labeled with SJ reads number, and altered exon was marked out with a red box. **(I)** Splicing events model is shown in the top panel. Boxplot in the bottom panel showing splicing ratio profile of the splicing event from Vegfa shown in **(H)**. **(J)** Visualization of junction read distribution of Pparg in AS events clualt5p51327 from different groups. Splice junctions were labeled with SJ reads number, and altered exon was marked out with red box. **(K)** Splicing events model is shown in the top panel. Boxplot in the bottom panel showing splicing ratio profile of the splicing event from Pparg shown in **(J)**.

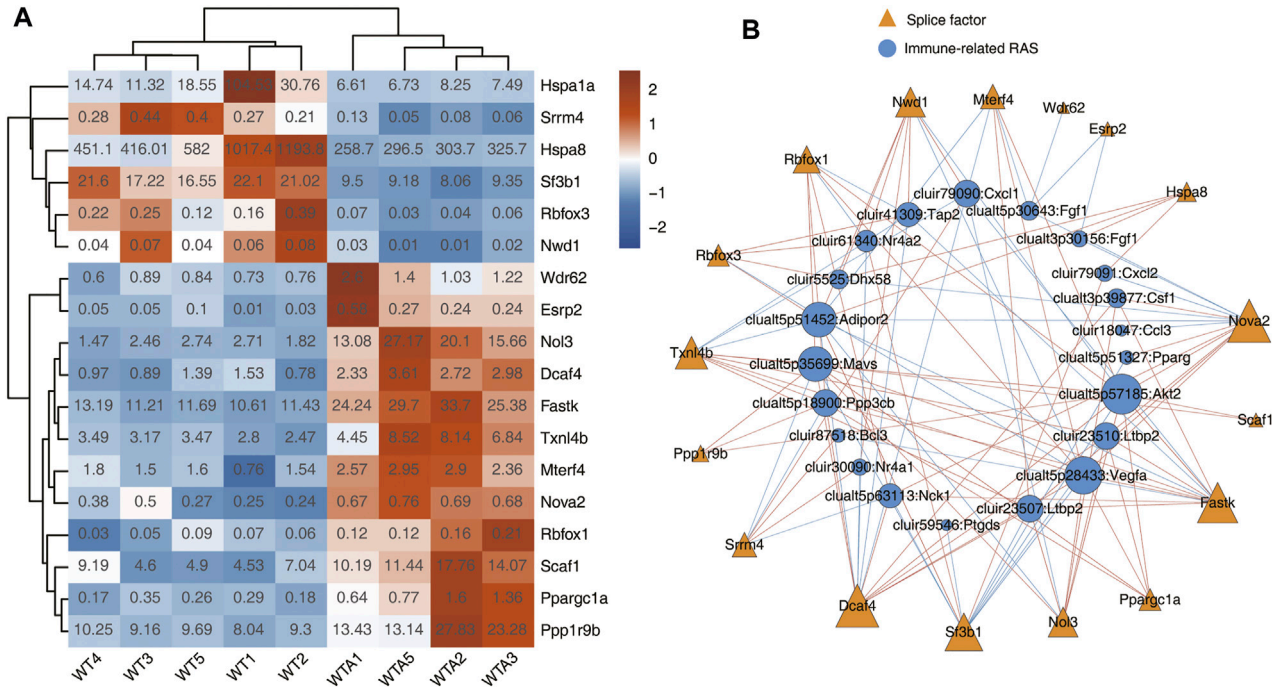


FIGURE 3 | Construction of co-disturbed network between SFs and immune-RAS. **(A)** Expression heatmap of all 18 significant differentially expressed SFs between WTA and WT sample. **(B)** The co-disturbed network between expression of WTA-associated SFs and splicing ratio of immune-RAS events (pSAR ≥90%) was constructed. |Pearson's correlation| ≥0.8 and *p*-value ≤0.01 were retained. Triangle represents SF genes. Circles indicate immune-RAS.

regulated by SFs, and constructed a SF-RAS interaction network structure diagram. As shown in **Figure 3A**; **Supplementary Figure S3A**, the expression levels of all these different SFs were displayed by heatmap, and a total of 18 differentially expressed SFs were identified. In subsequent verification and experiments, SFs with higher expression levels and significant differences, such as Sf3b1, Nol3, Fastk, Scaf1, etc., should be selected and based on existing literature reports. As shown in **Figure 3B**, the expression of differentially expressed SFs and the splicing ratio of immune-RAS events (pSAR ≥90%) were used for Pearson's correlation analysis (correlation coefficient ≥0.8, *p*-value ≤0.01). It mainly showed the co-variation relationship between differentially expressed SFs and immune-RAS, which meant that these SFs might potentially regulate immune-RAS. The size of nodes in the figure represents the number of gene/splicing events associated with them. We can focus on the larger SFs and splicing events.

Immune Infiltration Altered and is Associated With Candidate splicing factors

As shown in **Figure 4A,B** and **Supplementary Figure S4**, the proportion of different types of cells and the ratio of the average ratio of different cells were quite different between the WTA and WT group. In particular, both M1 and M2 types of macrophages were upregulated, while M0 type was downregulated in the WTA group. It implied that the polarization of macrophages in AAA had changed. It was also worth noting that the proportion of plasma cells had sharply increased in the WTA group. As shown

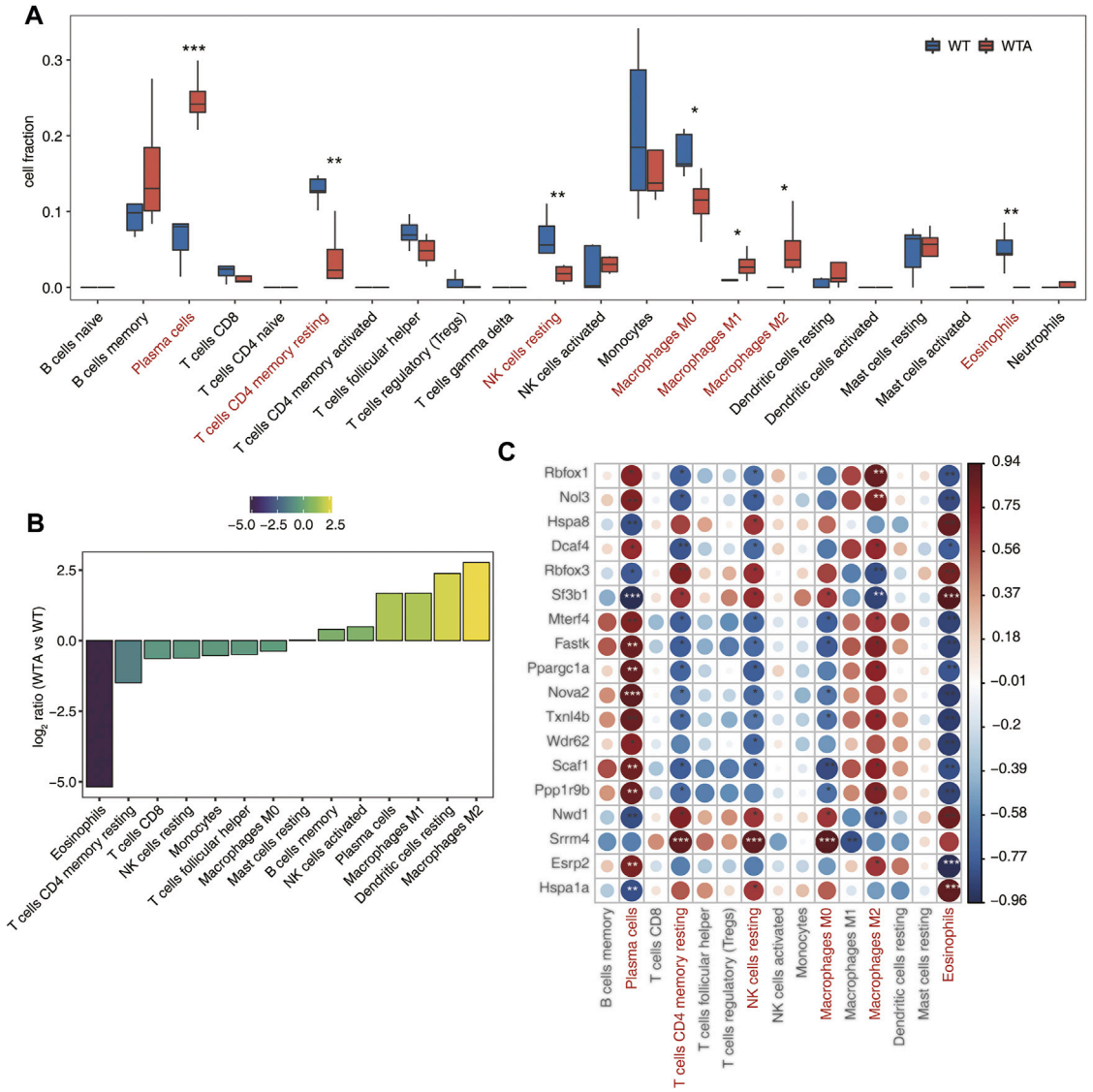


FIGURE 4 | Immune infiltration altered, and is associated with, candidate SFs. **(A)** Boxplot showing the fraction of each immune cell type in WTA or WT samples; the significant difference in the immune cell fractions between WTA and WT samples was calculated using the Student's t test. *$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$ **(B)** The WTA group relative to the WT group rank ordered based on decreasing values of the relative frequency ratio of cell populations. **(C)**. The dot-plot demonstrated the correlations between each immune microenvironment infiltration cell type and each dysregulated SF regulator. Different colors indicate correlation of immunocyte-RBP regulator, and significant ones were labeled with a star. *$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$.

in **Figure 4C**, correlation analysis between SF expression and the proportion of different cell types indicated that SF and its regulated immune genes played an important role in immune infiltration.

SF3B1 Bound on a Mass of Immune-Related Genes

We selected a splicing factor Sf3b1 with significantly different expression and studied the binding characteristics of Sf3b1 homologous gene in human K562 cells using ECLIP data, and speculated its regulation effect on immune-related genes. As shown **Figure 5A** and **Supplementary Figure S5A**, Sf3b1 interacts with AS of many immune genes and is associated with many processes. Then we used the ECLIP data of human K562 cells to analyze the distribution of the homologous gene Sf3b1 binding peak in different regions of the genome, mainly the intron region, followed by the CDS region in **Figure 5B**. As shown in **Figure 5C**, 142 peak genes (common to IP1 and IP2) bound to Sf3b1 were immune-related genes, including AKT2, MAVS, and VEGFA, which are shown in **Figure 5A**. As shown in **Figure 5D**; **Supplementary Figure S5B**, the motif enrichment analysis of peak on Sf3b1-bound immune genes showed the top five genetic sequence.

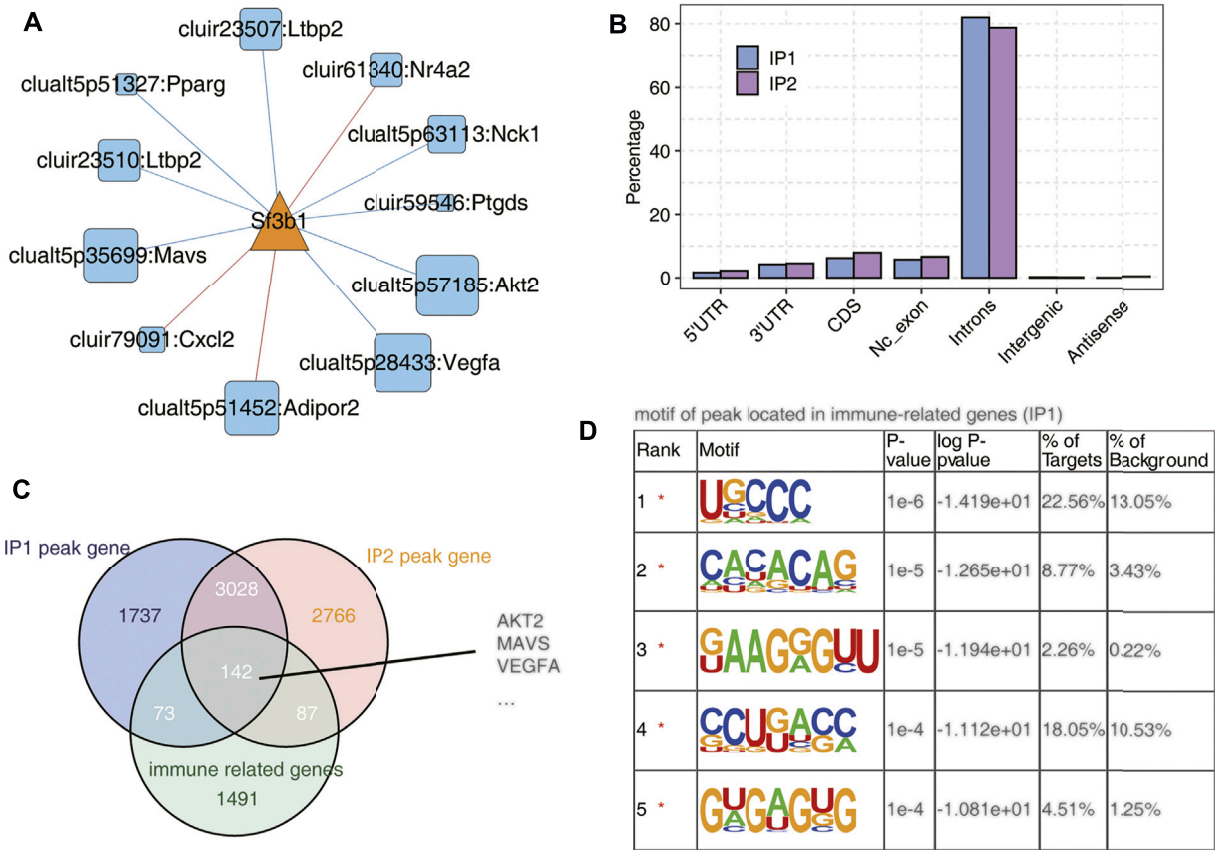


FIGURE 5 | SF3B1 bound on a mass of immune-related genes. **(A)** The co-disturbed network between expression of Sf3b1 and splicing ratio of immune-RAS events (pSAR ≥90%). **(B)** Bar plot showing the distribution of Sf3b1-bound peaks across different genomic regions. **(C)** Venn diagram of Sf3b1-bound peak genes and immune-related genes. **(D)** Motif analysis showing the top five peaks preferred bound motifs of Sf3b1 on immune-related genes by the HOMER software.

DISCUSSION

AAA has always been the research focus of vascular surgery. With the development of second-generation sequencing technology, more and more researchers have begun to use bioinformatics technology to study AAA. Based on the RNA-seq data of GSE175683, we explored the genome-wide identification of immune-RAS and splicing regulators involved in AAA. We discovered that there were 3,541 genes significantly differentially expressed, of which 2,627 genes were upregulated and 914 genes were downregulated, and 177 upregulated genes and 48 downregulated genes were immune-related regulatory genes (**Figure 1**; **Supplementary Figure S1**). We further focused on the AS of immune genes and used software SUVA to analyze and identify immune-RAS events. We found that immune-RAS events were mainly alt5p and IR events, about 60% of it was complex splicing events, and some immune-related genes could regulate AS events in AAA (**Figure 2**; **Supplementary Figure S2**). Next, we explored the differentially expressed SFs in WTA and WT samples and constructed an interaction network structure diagram between SFs and immune-RAS. We found that a total of 18 differentially expressed SFs were identified and constructed a co-variation relationship between differentially expressed SFs and immune-RAS (**Figure 3**; **Supplementary Figure S3**). Interestingly, we found that the proportion of different types of immune cells in WTA and WT group changed, individual cell types also changed significantly, and the expression levels of SFs were correlated with the ratio of different cell types (**Figure 4**; **Supplementary Figure S4**). Finally, we selected Sf3b1, an SF with a significant difference in expression, and found that many of the genes of the distribution of Sf3b1-bound peaks were related to the immune-RAS (**Figure 5**; **Supplementary Figure S5**). In summary, we found that immune-related genes and its regulation of AS events played an important role in AAA, and immune-RAS events were regulated by SFs.

With the change in lifestyle, the incidence of cardiovascular disease is increasing year by year, and it has become a serious public health problem. AAA is a serious aortic disease that has become an important cause of death in elderly people over 65. According to reports, about 13,000 people die from AAA every year (Nie et al., 2020). Most AAA patients are asymptomatic and cannot be treated before aneurysm ruptures, or the patient dies (Summers et al., 2021). No drugs can slow the development of AAA. The occurrence and development of AAA is a complicated

process involving many factors. It is usually believed that AAA is directly associated with atherosclerosis, hypertension, chronic obstructive pulmonary disease, and a variety of proteases, but there is no clear evidence that these factors play a role in the pathogenesis of AAA (Nie et al., 2020). The pathophysiological progresses of AAA include infiltration of inflammatory cells, degradation of collagen and elastic fibers, death of smooth muscle cells, increase of oxidative stress and defects of the arterial wall (Li H. et al., 2021; Thanigaimani et al., 2021). Although AAA has several established biological characteristics, convincing evidence shows that immune-mediated processes play a clear and prominent role in the pathogenesis of AAA (Li et al., 2018). The immune-inflammatory response is mediated by some special immune cell types, which interact in a highly coordinated manner and are functionally vital to the initiation and progression of AAA (Dale et al., 2015; Liu et al., 2015). In our study, compared with the WT group, a large number of gene transcription levels have changed, functional enrichment analysis of DEGs was conducted, and expression levels of immune genes changed in the WTA group. These suggested that there was regulation of immune-related genes in AAA, which was consistent with some previous reports (Nie et al., 2020; Li T. et al., 2021).

According to previous human and AAA experimental studies, several exogenous immune cells, including lymphocytes, macrophages, natural killer cells (NK), neutrophils, and dendritic cells, have been found to penetrate into aneurysm tissue and release extensively proinflammatory cytokines to trigger a series of inflammatory responses that lead to the direct structural protein degradation of the abdominal aorta (Hendel et al., 2015; Li et al., 2018; Blassova et al., 2019). Lei et al. found that several kinds of immune cells including naive B cells, resting and activated $CD4^+$ T cells were identified to be pointedly higher in ruptured AAA, while regulatory T cells, together with activated mast cells, were more in stable AAA conversely (Lei et al., 2020). Research showed that there was a significant difference in immune cell infiltration between normal vascular and AAA specimens, and high proportions of $CD4^+$ T cells, resting natural killer cells, activated mast cells, and 12 other types of immune cells were found in normal vascular tissues, whereas high proportions of macrophages, resting mast cells, $CD8^+$ T cells, and six other types of immune cells were found in AAA tissues (Nie et al., 2020). In the same situation, our research also found that both M1 and M2 types of macrophages and plasma cells were upregulated in the WTA group, while M0 type of macrophages, NK cells, $CD4^+$ T cells, and eosinophils were upregulated in the WT group. However, most reports only focus on a perspective of the immune response, and generally only discuss the types of immune cells and their roles. We know little about the AS events of immune-related genes and the regulation of SFs in AAA.

In order to better understand the immune-RAS events and SFs that play a role in AAA, further research is necessary to determine their special role in the pathophysiology of AAA. In our study, we used the software SUVA to analyze and identify immune-RAS events. The results showed that immune-RAS events were mainly alt5p and IR events, and about 60% of it was complex in AAA. In order to find a more dominant transcript for splicing, the events with pSAR ≥90% were selected for follow-up analysis. We newly discovered the AS of immune-related genes, such as Vegfa, Pparg, Adipor2, Ltbp2, Nr4a1, etc., by using heatmap and Venn diagram analysis. We choose Vegfa gene as a representative to discuss its effect of AS. Vegfa (vascular endothelial growth factor A) gene expresses multiple protein isoforms due to its AS exons. Dou et al. found that AS of Vegfa may regulate the development of colorectal cancer and represent new targets for its diagnosis, prognosis, and treatment (Zhao et al., 2015). Dou et al. found that Vegfa gene had AS in endometrial cancer, which may also provide new biomarkers for the diagnosis of endometrial cancer (Dou et al., 2020). The use of AS to produce VEGFA protein isoforms with different bioavailabilities is a key mechanism to control the development and function of blood vessel (Bridgett et al., 2017). Chesnokov et al. discovered that Vegfa isoform ratio produced by AS may be a promising factor for prediction of anti-angiogenic therapy efficiency in human hepatocellular carcinoma (Chesnokov et al., 2018). These fully indicate that the AS of genes has important effects, and it is necessary to further explore the role of these immune-RAS in AAA.

SFs are involved in removing introns from mRNA so that exons can be joined together. AS of precursor mRNA is an important mechanism to increase the complexity of gene expression and plays an important role in cell differentiation and organism development (Bonnal et al., 2020; Zhang et al., 2021). The regulation of AS is a complex process in which many interacting components are at work. Any error in this process may lead to the destruction of normal cell functions and the occurrence of diseases. In particular, immune-related genes are also regulated by SFs, which can cause changes in immune response or immune cell composition (Blake and Lynch, 2021). SFs may be the basis for identifying new diagnostic and prognostic biomarkers and new treatment strategies. In our study, a total of 18 differentially expressed SFs, such as Sf3b1, Nol3, Fastk, Scaf1, etc., were identified between the WTA and WT group, and SFs show interaction with immune-RAS, which meant that these SFs might potentially regulate immune-RAS. Recently, with the development of second-generation sequencing technology, many mutations related to RNA splicing have been gradually identified and reported one after another. Among these different SFs, human Sf3b1 (splicing factor 3b subunit 1) is the gene with the higher mutation frequency. So, we selected Sf3b1 for analysis based on higher expression levels and significant differences, and combined it with existing literature reports. Furney et al. found that Sf3b1 was repeatedly mutated in uveal melanoma, and the mutation was associated with abnormal AS (Furney et al., 2013). Maguire et al. discovered that Sf3b1 mutations resulted in AS events, and might constitute drivers and a novel therapeutic target in a subset of breast cancers (Maguire et al., 2015). Chang et al. found that Sf3b1 may not only induce direct cancer cell cytotoxicity but also initiate an innate immune response *via* activation of RNA-sensing pathways (Chang et al., 2021). All these suggest that Sf3b1 can regulate AS events, which is consistent with our research.

In conclusion, our research discovered that immune-related genes and immune cells played an important role in the occurrence and development of AAA, and the immune-RAS

events affected the formation of AAA. The regulation of SFs on AS events may be a new target for diagnostic and therapeutic intervention. However, there are several limitations to the study. First, because of the secondary analysis of the original data, it is difficult to evaluate the reliability of the original samples. Second, a small sample size may cause certain deviations in the comparison results, including DEGs, immune-RAS, and SFs. Since the samples we analyzed are from mouse AAA models and are relatively consistent, the pathophysiological process of patients may be different in clinical practice. Although we have removed outliers and used powerful tools, such as the latest algorithms for evaluation, it may still cause a certain error in the clinically actual situation. Therefore, further research is needed to provide more direct evidence for the immune-RAS and SF regulation of AAA. Altogether, we take the lead in discussing the vital role of immune-RAS and SFs, and provide new mechanism insights for the development of immune-targeted therapy in AAA.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SW and XG proposed and designed this research. SW wrote this manuscript. SW, HZ, and SL participated in the data analysis. NC and CZ participated in the design of the study. SW, HZ, and XG reviewed and edited the manuscript.

FUNDING

This work was financially supported by the Joint Co-construction Project of Henan Medical Science and Technology Research Plan (No. LHGJ20200342), the Key Research and Development and Promotion of Special Project (Scientific Problem Tackling) of Henan Province (No. 202102310122), and the Medical Science and Technology Research Project (co-constructed by province and ministry) of Henan Province (No. SB201901009).

ACKNOWLEDGMENTS

We thank the government for its fund support and the efforts of the co-authors, as well as Chao Cheng of the Center for Genome Analysis of Wuhan Ruixing Biotechnology Co. Ltd. for his guidance and help in the analysis process.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.816035/full#supplementary-material

Supplementary Figure S1 | Transcriptome analysis of DEGs in WTA and WT. **(A)**. The sample correlation heatmap exhibited the clustering of samples from the WT and WTA groups. **(B)**. Expression heatmap of all significant DEGs between WTA and WT samples.

Supplementary Figure S2 | Identification of WTA-associated AS events located in immune-related genes (immune-RAS). **(A)**. Boxplot showing all regulated AS events (RAS) by SUVA between WTA and WT. **(B)**. Splice junction constituting RAS detected by SUVA was annotated to classical AS event types. And the number of each classical AS event types were showed with boxplot. **(C)**. Bar plot showing number with different abundance (pSAR) of all detected RAS. **(D)**. Bar plot exhibited the most enriched GO biological process results of the RASG between WTA and WT groups. **(E)**. Bar plot exhibited the most enriched KEGG pathways results of RASG between WTA and WT groups. **(F)**. Boxplot showing splicing ratio profile across 5 WT tumor and 4 WTA samples of 3 immune-related splicing events (pSASR ≥ 90%).

Supplementary Figure S3 | Construction of co-disturbed network between SFs and immune-RAS. A. Boxplot showing expression profile of co-disturbed SFs in WTA and WT samples.

Supplementary Figure S4 | Immune infiltration altered and is associated with candidate SFs. **(A)**. Box plots showing proportion of 7 significantly altered cell type ($P \leq 0.05$) in WTA or WT samples. **(B)**. Scatter plot comparing the mean proportions of cell populations of each cell type in two groups.

Supplementary Figure S5 | SF3B1 bound on a mass of immune-related genes. **(A)**. Bar plot exhibited the most enriched GO biological process results of the genes bound by Sf3b1 in both replicates. **(B)**. Motif analysis showing the top 5 peaks preferred bound motifs of Sf3b1 on immune-related genes by HOMER software.

REFERENCES

Annambhotla, S., Bourgeois, S., Wang, X., Lin, P. H., Yao, Q., and Chen, C. (2008). Recent Advances in Molecular Mechanisms of Abdominal Aortic Aneurysm Formation. *World J. Surg.* 32 (6), 976–986. doi:10.1007/s00268-007-9456-x

Blake, D., and Lynch, K. (2021). The Three as: Alternative Splicing, Alternative Polyadenylation and Their Impact on Apoptosis in Immune Function. *Immunol. Rev.* 304 (1), 30–50. doi:10.1111/imr.13018

Blassova, T., Tonar, Z., Tomasek, P., Hosek, P., Hollan, I., Treska, V., et al. (2019). Inflammatory Cell Infiltrates, Hypoxia, Vascularization, Pentraxin 3 and Osteoprotegerin in Abdominal Aortic Aneurysms - A Quantitative Histological Study. *PLoS One* 14 (11), e0224818. doi:10.1371/journal.pone.0224818

Bonnal, S. C., López-Oreja, I., and Valcárcel, J. (2020). Roles and Mechanisms of Alternative Splicing in Cancer - Implications for Care. *Nat. Rev. Clin. Oncol.* 17 (8), 457–474. doi:10.1038/s41571-020-0350-x

Brangsch, J., Reimann, C., Collettini, F., Buchert, R., Botnar, R. M., and Makowski, M. R. (2017). Molecular Imaging of Abdominal Aortic Aneurysms. *Trends Mol Medicine* 23 (2), 150–164. doi:10.1016/j.molmed.2016.12.002

Bridgett, S., Dellett, M., and Simpson, D. A. (2017). RNA-sequencing Data Supports the Existence of Novel VEGFA Splicing Events but Not of VEGFAxxxb Isoforms. *Sci. Rep.* 7 (1), 58. doi:10.1038/s41598-017-00100-3

Cai, D., Sun, C., Zhang, G., Que, X., Fujise, K., Weintraub, N. L., et al. (2021). A Novel Mechanism Underlying Inflammatory Smooth Muscle Phenotype in Abdominal Aortic Aneurysm. *Circ. Res.* 129 (10), e202–e214. doi:10.1161/CIRCRESAHA.121.319374

Chaikof, E. L., Dalman, R. L., Eskandari, M. K., Jackson, B. M., Lee, W. A., Mansour, M. A., et al. (2018). The Society for Vascular Surgery Practice Guidelines on the Care of Patients with an Abdominal Aortic Aneurysm. *J. Vasc. Surg.* 67 (1), 2–77.e2. doi:10.1016/j.jvs.2017.10.044

Chang, A. Y., Zhou, Y. J., Iyengar, S., Pobiarzyn, P. W., Tishchenko, P., Shah, K. M., et al. (2021). Modulation of SF3B1 in the Pre-mRNA Spliceosome Induces a

RIG-I-dependent Type I IFN Response. *J. Biol. Chem.* 297, 101277. doi:10.1016/j.jbc.2021.101277

Cheng, C., Liu, L., Bao, Y., Yi, J., Quan, W., Xue, Y., et al. (2021). SUVA: Splicing Site Usage Variation Analysis from RNA-Seq Data Reveals Highly Conserved Complex Splicing Biomarkers in Liver Cancer. *RNA Biol.* 18 (Suppl. 1), 157–171. doi:10.1080/15476286.2021.1940037

Chesnokov, M. S., Khesina, P. A., Shavochkina, D. A., Kustova, I. F., Dyakov, L. M., Morozova, O. V., et al. (2018). Shift in VEGFA Isoform Balance towards More Angiogenic Variants Is Associated with Tumor Stage and Differentiation of Human Hepatocellular Carcinoma. *PeerJ* 6, e4915. doi:10.7717/peerj.4915

Dale, M. A., Ruhlman, M. K., and Baxter, B. T. (2015). Inflammatory Cell Phenotypes in AAAs. *Arterioscler Thromb. Vasc. Biol.* 35 (8), 1746–1755. doi:10.1161/atvbaha.115.305269

Dou, X. Q., Chen, X. J., Wen, M. X., Zhang, S. Z., Zhou, Q., and Zhang, S. Q. (2020). Alternative Splicing of VEGFA Is Regulated by RBM10 in Endometrial Cancer. *Kaohsiung J. Med. Sci.* 36 (1), 13–19. doi:10.1002/kjm2.12127

Finotello, F., and Trajanoski, Z. (2018). Quantifying Tumor-Infiltrating Immune Cells from Transcriptomics Data. *Cancer Immunol. Immunother.* 67 (7), 1031–1040. doi:10.1007/s00262-018-2150-z

Furney, S. J., Pedersen, M., Gentien, D., Dumont, A. G., Rapinat, A., Desjardins, L., et al. (2013). SF3B1 Mutations Are Associated with Alternative Splicing in Uveal Melanoma. *Cancer Discov.* 3 (10), 1122–1129. doi:10.1158/2159-8290.cd-13-0330

Guirguis-Blake, J. M., Beil, T. L., Senger, C. A., and Coppola, E. L. (2019). Primary Care Screening for Abdominal Aortic Aneurysm. *JAMA* 322 (22), 2219–2238. doi:10.1001/jama.2019.17021

Hendel, A., Ang, L., and Granville, D. (2015). Inflammaging and Proteases in Abdominal Aortic Aneurysm. *Curr. Vasc. Pharmacol.* 13 (1), 95–110. doi:10.2174/15701611130115030313234 8

Jagadesham, V. P., Scott, D. J. A., and Carding, S. R. (2008). Abdominal Aortic Aneurysms: an Autoimmune Disease? *Trends Mol. Med.* 14 (12), 522–529. doi:10.1016/j.molmed.2008.09.008

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi:10.1038/s41587-019-0201-4

Kurtovic, S., Paloschi, V., Folkersen, L., Gottfries, J., Franco-Cereceda, A., and Eriksson, P. (2011). Diverging Alternative Splicing Fingerprints in the Transforming Growth Factor-β Signaling Pathway Identified in Thoracic Aortic Aneurysms. *Mol. Med.* 17, 665–675. doi:10.2119/molmed.2011.00018

Lei, C., Yang, D., Chen, S., Chen, W., Sun, X., Wu, X., et al. (2020). Patterns of Immune Infiltration in Stable and Raptured Abdominal Aortic Aneurysms: A Gene-Expression-Based Retrospective Study. *Gene* 762, 145056. doi:10.1016/j.gene.2020.145056

Li, H., Bai, S., Ao, Q., Wang, X., Tian, X., Li, X., et al. (2018). Modulation of Immune-Inflammatory Responses in Abdominal Aortic Aneurysm: Emerging Molecular Targets. *J. Immunol. Res.* 2018, 7213760. doi:10.1155/2018/7213760

Li, H., Xu, H., Wen, H., Wang, H., Zhao, R., Sun, Y., et al. (2021). Lysyl Hydroxylase 1 (LH1) Deficiency Promotes Angiotensin II (Ang II)-induced Dissecting Abdominal Aortic Aneurysm. *Theranostics* 11 (19), 9587–9604. doi:10.7150/thno.65277

Li, T., Wang, T., and Zhao, X. (2021). Profiles of Immune Infiltration in Abdominal Aortic Aneurysm and Their Associated Marker Genes: a Gene Expression-Based Study. *Braz. J. Med. Biol. Res.* 54 (11), e11372. doi:10.1590/1414-431X2021e11372

Li, Z.-X., Zheng, Z.-Q., Wei, Z.-H., Zhang, L.-L., Li, F., Lin, L., et al. (2019). Comprehensive Characterization of the Alternative Splicing Landscape in Head and Neck Squamous Cell Carcinoma Reveals Novel Events Associated with Tumorigenesis and the Immune Microenvironment. *Theranostics* 9 (25), 7648–7665. doi:10.7150/thno.36585

Liu, Y., Liao, J., Zhao, M., Wu, H., Yung, S., Chan, T. M., et al. (2015). Increased Expression of TLR2 in CD4+T Cells from SLE Patients Enhances Immune Reactivity and Promotes IL-17 Expression through Histone Modifications. *Eur. J. Immunol.* 45 (9), 2683–2693. doi:10.1002/eji.201445219

Love, M. I., Huber, W., and Anders, S. (2014). Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Maguire, S. L., Leonidou, A., Wai, P., Marchiò, C., Ng, C. K., Sapino, A., et al. (2015). SF3B1 Mutations Constitute a Novel Therapeutic Target in Breast Cancer. *J. Pathol.* 235 (4), 571–580. doi:10.1002/path.4483

Maiellaro, K., and Taylor, W. (2007). The Role of the Adventitia in Vascular Inflammation. *Cardiovasc. Res.* 75 (4), 640–648. doi:10.1016/j.cardiores.2007.06.023

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 (5), 453–457. doi:10.1038/nmeth.3337

Nie, H., Qiu, J., Wen, S., and Zhou, W. (2020). Combining Bioinformatics Techniques to Study the Key Immune-Related Genes in Abdominal Aortic Aneurysm. *Front. Genet.* 11, 579215. doi:10.3389/fgene.2020.579215

Rateri, D. L., Howatt, D. A., Moorleghen, J. J., Charnigo, R., Cassis, L. A., and Daugherty, A. (2011). Prolonged Infusion of Angiotensin II in apoE−/− Mice Promotes Macrophage Recruitment with Continued Expansion of Abdominal Aortic Aneurysm. *Am. J. Pathol.* 179 (3), 1542–1548. doi:10.1016/j.ajpath.2011.05.049

Sakalihasan, N., Limet, R., and Defawe, O. (2005). Abdominal Aortic Aneurysm. *The Lancet* 365 (9470), 1577–1589. doi:10.1016/s0140-6736(05)66459-8

Schaheen, B., Downs, E. A., Serbulea, V., Almenara, C. C. P., Spinosa, M., Su, G., et al. (2016). B-cell Depletion Promotes Aortic Infiltration of Immunosuppressive Cells and Is Protective of Experimental Aortic Aneurysm. *Arterioscler. Thromb. Vasc. Biol.* 36 (11), 2191–2202. doi:10.1161/atvbaha.116.307559

Summers, K. L., Kerut, E. K., Sheahan, C. M., and Sheahan, M. G. (2021). Evaluating the Prevalence of Abdominal Aortic Aneurysms in the United States through a National Screening Database. *J. Vasc. Surg.* 73 (1), 61–68. doi:10.1016/j.jvs.2020.03.046

Thanigaimani, S., Singh, T., Unosson, J., Phie, J., Moxon, J., Wanhainen, A., et al. (2021). Association between Metformin Prescription and Abdominal Aortic Aneurysm Growth and Clinical Events: a Systematic Review and Meta-Analysis. *Eur. J. Vasc. endovascular Surg.* 62 (5), 747–756. doi:10.1016/j.ejvs.2021.06.013

Wang, J., Lindholt, J. S., Sukhova, G. K., Shi, M. A., Xia, M., Chen, H., et al. (2014). IgE Actions on CD 4 + T Cells, Mast Cells, and Macrophages Participate in the Pathogenesis of Experimental Abdominal Aortic Aneurysms. *EMBO Mol. Med.* 6 (7), 952–969. doi:10.15252/emmm.201303811

Wanhainen, A., Verzini, F., Van Herzeele, I., Allaire, E., Bown, M., Cohnert, T., et al. (2019). Editor's Choice - European Society for Vascular Surgery (ESVS) 2019 Clinical Practice Guidelines on the Management of Abdominal Aorto-Iliac Artery Aneurysms. *Eur. J. Vasc. Endovascular Surg.* 57 (1), 8–93. doi:10.1016/j.ejvs.2018.09.020

Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., et al. (2011). KOBAS 2.0: a Web Server for Annotation and Identification of Enriched Pathways and Diseases. *Nucleic Acids Res.* 39 (Suppl. l), W316–W322. doi:10.1093/nar/gkr483

Yan, Y.-W., Fan, J., Bai, S.-L., Hou, W.-J., Li, X., and Tong, H. (2016). Zinc Prevents Abdominal Aortic Aneurysm Formation by Induction of A20-Mediated Suppression of NF-Kb Pathway. *PLoS One* 11 (2), e0148536. doi:10.1371/journal.pone.0148536

Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative Splicing and Cancer: a Systematic Review. *Sig Transduct Target. Ther.* 6 (1), 78. doi:10.1038/s41392-021-00486-7

Zhao, G., Fu, Y., Cai, Z., Yu, F., Gong, Z., Dai, R., et al. (2017). Unspliced XBP1 Confers VSMC Homeostasis and Prevents Aortic Aneurysm Formation via FoxO4 Interaction. *Circ. Res.* 121 (12), 1331–1345. doi:10.1161/circresaha.117.311450

Zhao, Y.-J., Han, H., Liang, Y., Shi, C.-Z., Zhu, Q.-C., Yang, J., et al. (2015). Alternative Splicing ofVEGFA,APPandNUMBgenes in Colorectal Cancer. *World J. Gastroenterol.* 21 (21), 6550–6560. doi:10.3748/wjg.v21.i21.6550

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 nor endorsed by the publisher.

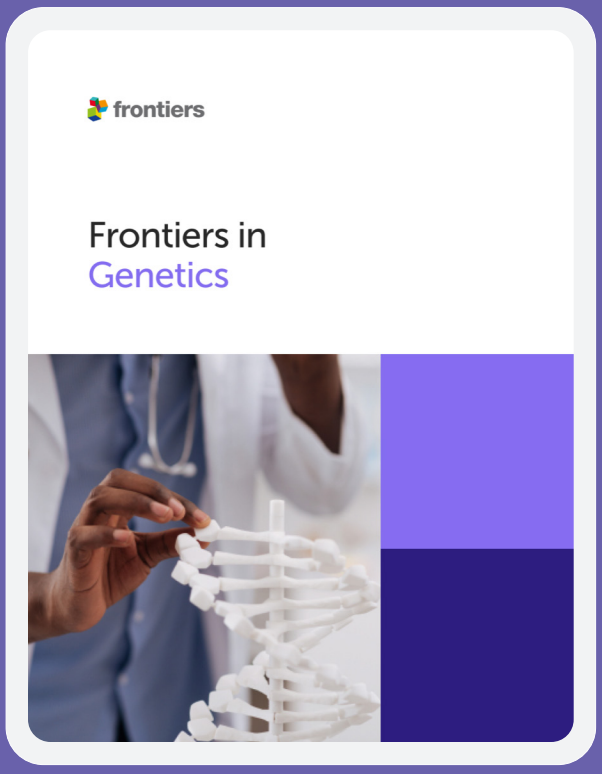
Frontiers in Genetics

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