# MACHINE LEARNING-BASED METHODS FOR RNA DATA ANALYSIS

EDITED BY: Lihong Peng, Jialiang Yang, Minxian Wallace Wang and Liqian Zhou

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-384-9
DOI 10.3389/978-2-88976-384-9

## About Frontiers

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

## Frontiers Journal Series

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

## Dedication to Quality

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

## What are Frontiers Research Topics?

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

# MACHINE LEARNING-BASED METHODS FOR RNA DATA ANALYSIS

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

**Citation:** Peng, L., Yang, J., Wang, M. W., Zhou, L., eds. (2022). Machine Learning-Based Methods for RNA Data Analysis. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-384-9

# Table of Contents


*04* ***Editorial: Machine Learning-Based Methods for RNA Data Analysis***
Lihong Peng, Jialiang Yang, Minxian Wang and Liqian Zhou

*08* ***A Novel Framework to Predict Breast Cancer Prognosis Using Immune-Associated LncRNAs***
Zhijian Huang, Chen Xiao, Fushou Zhang, Zhifeng Zhou, Liang Yu, Changsheng Ye, Weiwei Huang and Nani Li

*22* ***Comprehensive circRNA Expression Profile and Construction of circRNAs-Related ceRNA Network in a Mouse Model of Autism***
Ji Wang, Zhongxiu Yang, Canming Chen, Yang Xu, Hongguang Wang, Bing Liu, Wei Zhang and Yanan Jiang

*33* ***Identifying the Signatures and Rules of Circulating Extracellular MicroRNA for Distinguishing Cancer Subtypes***
Fei Yuan, Zhandong Li, Lei Chen, Tao Zeng, Yu-Hang Zhang, Shijian Ding, Tao Huang and Yu-Dong Cai

*43* ***Identification of miRNA-Mediated Subpathways as Prostate Cancer Biomarkers Based on Topological Inference in a Machine Learning Process Using Integrated Gene and miRNA Expression Data***
Ziyu Ning, Shuang Yu, Yanqiao Zhao, Xiaoming Sun, Haibin Wu and Xiaoyang Yu

*56* ***Deep Learning Enables Fast and Accurate Imputation of Gene Expression***
Ramon Viñas, Tiago Azevedo, Eric R. Gamazon and Pietro Liò

*68* ***Bioinformatic Analysis of Crosstalk Between circRNA, miRNA, and Target Gene Network in NAFLD***
Cen Du, Lan Shen, Zhuoqi Ma, Jian Du and Shi Jin

*76* ***Transcriptome Analysis of Choroid and Retina From Tree Shrew With Choroidal Neovascularization Reveals Key Signaling Moieties***
Jie Jia, Dandan Qiu, Caixia Lu, Wenguang Wang, Na Li, Yuanyuan Han, Pinfen Tong, Xiaomei Sun, Min Wu and Jiejie Dai

*90* ***Correlations Between the Characteristics of Alternative Splicing Events, Prognosis, and the Immune Microenvironment in Breast Cancer***
Youyuan Deng, Hongjun Zhao, Lifen Ye, Zhiya Hu, Kun Fang and Jianguo Wang

*101* ***Development of a Four-mRNA Expression-Based Prognostic Signature for Cutaneous Melanoma***
Haiya Bai, Youliang Wang, Huimin Liu and Junyang Lu

*111* ***Establishment of A Nomogram for Predicting the Prognosis of Soft Tissue Sarcoma Based on Seven Glycolysis-Related Gene Risk Score***
Yuhang Liu, Changjiang Liu, Hao Zhang, Xinzeyu Yi and Aixi Yu

frontiers | Frontiers in Genetics

EDITORIAL
published: 25 May 2022
doi: 10.3389/fgene.2022.828575

# Editorial: Machine Learning-Based Methods for RNA Data Analysis

*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, circRNA, microRNA, mRNA, gene expression**

## Editorial on the Research Topic

## Machine Learning-Based Methods for RNA Data Analysis

RNA is a type of extremely important biological macromolecules, which play key roles in all aspects of life activities and biological processes through its interactions with other biological entities Wang et al. (2021); Zhang et al. (2021). Thus, it is critical to identify complex biological associations between RNA and other biological entities Mu et al. (2020); Deng et al. (2018). Although experimental methods have been applied to analyze RNA data, especially identify various associations between RNA molecules and complex diseases, they are usually time-consuming and resource demanding. Machine learning aims to simulate human learning ways in real time and divide the existing content into knowledge structures to advance learning efficiency. It can effectively use available electronic data to boost learning performance or implement accurate prediction Mohri et al. (2018). Furthermore, it still improves more evidence-based decision-making in the area of life science Jordan and Mitchell (2015). With the advancement of next generation sequencing techniques, machine learning-based methods discovered a large number of useful information from abundant RNA data and thus provide an effective way for the analysis of RNA data. Consequently, through machine learning techniques, we can design powerful models and algorithms to discovery diverse associations between RNA molecules themselves (such as microRNAs, mRNA, circular RNAs, and long noncoding RNAs) and between RNA molecules and complex diseases. We can further infer novel molecular markers for diagnosis and prognosis of corresponding diseases based on the identified associations.

Based on the assumption of "guilt-by-association" and machine learning technologies, accumulated computational methods have been developed to analyze RNA data Liu et al. (2020); Chu et al. (2021). However, the performance of most methods remains unsatisfying due to data complexity and heterogeneity. Therefore, this research topic serves as a forum to develop new machine learning algorithms to improve RNA data analyses.

MicroRNAs (miRNAs) are a class of short and endogenous noncoding RNAs Wang et al. (2020); Chen et al. (2019a). miRNAs can control gene expression based on translational repression or messenger RNA (mRNA) degradation and exhibit strong associations with a variety of disease including neurodegenerative diseases and cancers Saliminejad et al. (2019). Chen et al. designed a few representative machine learning-based algorithms to identify potential microRNA-disease associations Chen et al. (2018, 2019b).

To find robust biomarkers associated with prostate cancer, Ning et al. designed a multi-omics data fusion method by integrating directed random walk and Support Vector Machine (SVM). They compared their proposed pathway-based method with five other methods including the Median method, Mean method, component analysis method, pathway activity inference method based on



***Edited by:***
*William C. Cho,*
*QEH, Hong Kong SAR, China*

***Reviewed by:***
*Maarten M. G. van den Hoogenhof,*
*Heidelberg University Hospital, Germany*

****Correspondence:***
*Liqian Zhou*
*zhoulq11@163.com*

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

***Received:*** *03 December 2021*
***Accepted:*** *12 April 2022*
***Published:*** *25 May 2022*

***Citation:***
*Peng L, Yang J, Wang M and Zhou L (2022) Editorial: Machine Learning-Based Methods for RNA Data Analysis. Front. Genet. 13:828575. doi: 10.3389/fgene.2022.828575*

condition-responsive gene analysis, and directed random walk method. The results from cross validation showed that their proposed method computed the best average AUC and accuracy in three within-datasets and other 10 cancer datasets. They inferred that hsa-miR-106b and hsa-miR-20b may be the shared miRNA-mediated subpathway biomarkers in GSE21036, GSE14794, and "PRAD-TCGA" datasets.

The inference of cancer-related circulating biomarkers has become one of the most important research directions on clinical cancer diagnosis. To identify extracellular microRNAs for pan-cancers, Yuan et al. integrated Boruta feature filtering, max-relevance and min-redundancy feature selection, incremental feature selection, synthetic minority oversampling method, and four classification models including random forest, SVM, *k*-nearest neighbors, and decision trees. They conducted 10-fold cross validation for 20 times. The results showed that SVM obtained better accuracies and Matthew correlation coefficients compared to other three black-box classifiers. They predicted that hsa-miR-5100 and hsa-miR-6088 have strong associations with pan-cancer.

Most of chronic liver diseases are caused by non-alcoholic fatty liver diseases (NAFLD). To capture miRNAs, circRNAs, and genes associated with NAFLD, Du et al. built a circRNA-miRNA-mRNA network and provided a novel perspective for the inhibition and therapy of NAFLD using functional enrichment analysis and protein interaction network analysis. They found that the crosstalk between hsa_circ_000031, miR-6512-3p, and PEG10 may participate in NAFLD's pathogenesis and the crosstalk could be the underlying biomarkers of NAFLD.

mRNA stability affects gene expression in almost all organisms from bacteria to human. Clinical trials have revealed that mRNA vaccines provide a safe and effective immune response in human Zhang et al. (2019). Therefore, mRNA vaccines exhibit a powerful alternative to traditional vaccine approaches Pardi et al. (2018). To discover potential biomarkers associated with cutaneous melanoma, Bai et al. combined univariable Cox proportional hazards regression and random survival forest algorithm and designed a four-mRNA signature approach. The proposed four-mRNA signature method was compared with two clinical prognostic markers (melanoma clark level and tumor stage) and obtained the best AUC and sensitivity. They found the four-mRNA signature (CD276, UQCRFS1, HAPLN3, and PIP4P1) could be a prognostic signature for cutaneous melanoma patients.

Circular RNAs (circRNAs) are a class of RNAs with covalently closed structure and high stability Vo et al. (2019). circRNAs can serve as a new strategy for diagnosis and treatment of diseases Zeng et al. (2020). Autism is a multifactorial neurodevelopmental disease and usually involves in mental disorder, attention deficit, and intellectual disability. To analyze circRNA expression in autism in the mouse brain, Wang et al. built a circRNA-based competing endogenous RNA network. They successfully established a mouse autism model and measured repetitive self-grooming behaviors. Furthermore, they constructed a circRNA-miRNA-mRNA network composed of 1,059 circRNAs, 1,926 miRNAs, and 6,730 mRNA. Third, they performed gene ontology and pathway enrichment analysis and statistical analysis. Finally, they identified 1,059 differentially expression circRNAs associated to autism.

Long noncoding RNAs (lncRNAs) are a class of noncoding RNAs involved in diverse biological processes Peng et al. (2022); Jia and Luan (2022). lncRNAs are aberrantly expressed in numerous cancers and demonstrate crucial roles in oncogenic and tumor suppressive activities Zhou et al. (2021); Liu et al. (2022). lncRNAs implement their biological functions by linking to RNA-binding proteins. Deep learning-based methods, such as LPIDF Tian et al. (2021), deep forest Wang et al. (2021), Capsule-LPI Li et al. (2021), and LPI-DLDN Peng et al. (2021), were widely applied to detect lncRNA-protein interactions. Breast cancer is one of the most common malignant tumors and causes the leading mortality in women. To identify immune-associated lncRNAs for breast cancer prognosis, Huang et al. designed a novel framework to identify immune lncRNA signatures as prognostic marker for breast cancer combining Cox regression analysis and iterative Lasso Cox regression analysis. The proposed model was validated the performance in two independent cohorts by comparing with known prognostic biomarkers and obtained an AUC of 0.86. The results showed that the proposed model can effectively analyze immune-associated lncRNAs for breast cancers based on ROC analysis, Kaplan-Meier analysis, univariate and multivariate Cox regression analysis, gene set enrichment analysis, and gene set variation analysis. Furthermore, they confirmed that lncRNA signatures could independently assess breast cancer survival.

Gene expression analyses contribute to prioritizing potential disease genes and identifying transcriptional regulatory programmes Toro-Domínguez et al. (2021). With the development of single-cell RNA sequencing technologies, a large number of machine learning-related models and algorithms are increasingly exploited to analyze single-cell RNA sequencing data Zhu et al. (2022); Xu et al. (2020). Moni et al. Mohri et al. (2018) conducted a large number of comparative genomic and transcriptomic analyses to capture key gene expression pathways associated with SARS-CoV-2.

To accurately impute gene expression information for multiple tissue types with minimal reconstruction error, Vinas et al. developed two deep learning models, pseudo-mask imputer and generative adversarial imputation network-based method. They compared their proposed methods with several state-of-the-art imputation methods on RNA-seq data from the GTEx project. The results showed that pseudo-mask imputer outperformed all other methods in inductive imputation and generative adversarial imputation network-based method obtained the highest performance in in-place imputation in terms of the coefficient of determination and runtime. They observed that several genes (such as PSMB6, COX6C, PSMD7 and PSMA2) exhibited different distributions in the Alzheimer's disease pathway.

Soft tissue sarcoma is a type of tumors accounting for 1% in adult cancers. To precisely observe new biomarkers and therapeutic targets for the disease, Liu et al. used risk characteristics and transcriptome data and built a risk signature and nomograms for patients with soft tissue sarcoma based on glycolysis-related genes. The results

demonstrated that the proposed model computed the best AUCs. They screened seven glycolysis-related genes associated with soft tissue sarcoma.

Dysregulation of alternative splicing is very important to tumorigenesis and microenvironment formation. To predict splicing factors related to alternative splicing in breast cancer, Deng et al. performed genome-wide analysis of the alternative splicing events in breast cancers based on differential and prognostic analyses. The proposed method computed relatively lower false discovery rate. They detected a few differentially expressed alternative splicing events and independent prognostic factors associated with breast cancer.

Pathological neovascularization in choroid is a major cause of blindness. To capture key signaling pathways in choroidal neovascularization, Jia et al. first performed three bioinformatics analyses, which include hierarchical cluster analysis, weighted gene co-expression network analysis, and protein-protein interaction network analysis. They then implemented hematoxylin and eosin staining, CD31 immunohistochemistry, and reverse transcription quantitative PCR. The results showed that differentially expressed genes in chroid were mainly linked to membrane transport.

RNA molecules have close linages with various diseases. The inference of diverse associations between RNAs and diseases contributes to revealing the pathogenic mechanism of complex diseases and investigating corresponding biomarkers, and further designing appropriate therapeutic strategies. On the research topic, researchers designed various machine learning-based methods and used multiple bioinformatics tools to analyze diverse RAN molecules. They obtained relatively better results and found a few biomarkers. We hope that the topic could improve RNA analyses and promote the diagnosis and treatment of related diseases.

## AUTHOR CONTRIBUTIONS

LP, JY, MW, and LZ wrote the Editorial.

## FUNDING

This work is supported in part by the National Natural Science Foundation of China (Grant 62172158).

## REFERENCES

Chen, X., Xie, D., Wang, L., Zhao, Q., You, Z.-H., and Liu, H. (2018). 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. Bioinformatics* 20, 515–539. doi:10.1093/bib/bbx130

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

Chu, Y., Wang, X., Dai, Q., Wang, Y., Wang, Q., Peng, S., et al. (2021). Mda-gcnftg: Identifying Mirna-Disease Associations Based on Graph Convolutional Networks via Graph Sampling through the Feature and Topology Graph. *Brief Bioinform* 22, bbab165. doi:10.1093/bib/bbab165

Deng, L., Wang, J., Xiao, Y., Wang, Z., and Liu, H. (2018). Accurate Prediction of Protein-Lncrna Interactions by Diffusion and Hetesim Features across Heterogeneous Network. *BMC bioinformatics* 19, 370. doi:10.1186/s12859-018-2390-0

Jia, L., and Luan, Y. (2022). Multi-feature Fusion Method Based on Linear Neighborhood Propagation Predict Plant Lncrna–Protein Interactions. *Interdiscip. Sci. Comput. Life Sci.* 2022, 1–10. doi:10.1007/s12539-022-00501-7

Jordan, M. I., and Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science* 349, 255–260. doi:10.1126/science.aaa8415

Li, Y., Sun, H., Feng, S., Zhang, Q., Han, S., and Du, W. (2021). Capsule-lpi: a Lncrna-Protein Interaction Predicting Tool Based on a Capsule Network. *BMC bioinformatics* 22, 1–19. doi:10.1186/s12859-021-04171-y

Liu, H., Ren, G., Chen, H., Liu, Q., Yang, Y., and Zhao, Q. (2020). Predicting lncRNA-miRNA Interactions Based on Logistic Matrix Factorization with Neighborhood Regularized. *Knowledge-Based Syst.* 191, 105261. doi:10.1016/j.knosys.2019.105261

Liu, Y., Yu, Y., and Zhao, S. (2022). Dual Attention Mechanisms and Feature Fusion Networks Based Method for Predicting Lncrna-Disease Associations. *Interdiscip. Sci. Comput. Life Sci.* 2022, 1–14. doi:10.1007/s12539-021-00492-x

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. Cambridge, MA, USA: MIT press.

Mu, Y., Zhang, R., Wang, L., and Liu, X. (2020). Ipseu-Layer: Identifying Rna Pseudouridine Sites Using Layered Ensemble Model. *Interdiscip. Sci.* 12, 193–203. doi:10.1007/s12539-020-00362-y

Pardi, N., Hogan, M. J., Porter, F. W., and Weissman, D. (2018). mRNA Vaccines - a new era in Vaccinology. *Nat. Rev. Drug Discov.* 17, 261–279. doi:10.1038/nrd.2017.243

Peng, L., Tan, J., Tian, X., and Zhou, L. (2022). Enanndeep: An Ensemble-Based Lncrna–Protein Interaction Prediction Framework with Adaptive K-Nearest Neighbor Classifier and Deep Models. *Interdiscip. Sci. Comput. Life 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

Saliminejad, K., Khorram Khorshid, H. R., Soleymani Fard, S., and Ghaffari, S. H. (2019). An Overview of Micrornas: Biology, Functions, Therapeutics, and Analysis Methods. *J. Cell Physiol.* 234, 5451–5465. doi:10.1002/jcp.27486

Tian, X., Shen, L., Wang, Z., Zhou, L., and Peng, L. (2021). A Novel Lncrna–Protein Interaction Prediction Method Based on Deep forest with cascade forest Structure. *Scientific Rep.* 11, 1–15. doi:10.1038/s41598-021-98277-1

Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., and Carmona-Sáez, P. (2021). A Survey of Gene Expression Meta-Analysis: Methods and Applications. *Brief. Bioinformatics* 22, 1694–1705. doi:10.1093/bib/bbaa019

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, 869–881. doi:10.1016/j.cell.2018.12.021

Wang, W., Dai, Q., Li, F., Xiong, Y., and Wei, D. Q. (2021). 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

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

Zeng, B., Chen, T., Luo, J., Xie, M., Wei, L., Xi, Q., et al. (2020). Exploration of Long Non-coding Rnas and Circular Rnas in Porcine Milk Exosomes. *Front. Genet.* 11, 652. doi:10.3389/fgene.2020.00652

Zhang, C., Maruggi, G., Shan, H., and Li, J. (2019). Advances in Mrna Vaccines for Infectious Diseases. *Front. Immunol.* 10, 594. doi:10.3389/fimmu.2019.00594

Zhang, L., Yang, P., Feng, H., Zhao, Q., and Liu, H. (2021). Using Network Distance Analysis to Predict lncRNA-miRNA Interactions. *Interdiscip. Sci. Comput. Life Sci.* 13, 535–545. doi:10.1007/s12539-021-00458-z

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, 1–24. doi:10.1186/s12859-021-04399-8

Zhu, Y.-L., Yuan, S.-S., and Liu, J.-X. (2022). Similarity and Dissimilarity Regularized Nonnegative Matrix Factorization for Single-Cell Rna-Seq Analysis. *Interdiscip. Sci. Comput. Life Sci.* 14, 45–54. doi:10.1007/s12539-021-00457-0

**Conflict of Interest:** Author JY was 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

ORIGINAL RESEARCH
published: 21 January 2021
doi: 10.3389/fgene.2020.634195

# A Novel Framework to Predict Breast Cancer Prognosis Using Immune-Associated LncRNAs

*Zhijian Huang[1,2†], Chen Xiao[3†], Fushou Zhang[4], Zhifeng Zhou[5], Liang Yu[6], Changsheng Ye[2*], Weiwei Huang[7*] and Nani Li[7*]*

[1] Department of Breast Surgical Oncology, Fujian Medical University Cancer Hospital, Fujian Cancer Hospital, Fuzhou, China, [2] Breast Center, Nanfang Hospital, Southern Medical University, Guangzhou, China, [3] Department of Gastroenterology, Fuzhou Second Hospital Affiliated to Xiamen University, Fuzhou, China, [4] Department of General Surgery, The Hospital of Changle District, Fuzhou, China, [5] Laboratory of Immuno-Oncology, Fujian Medical University Cancer Hospital, Fujian Cancer Hospital, Fuzhou, China, [6] Department of Thyroid and Breast Surgery, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [7] Department of Medical Oncology, Fujian Medical University Cancer Hospital, Fujian Cancer Hospital, Fuzhou, China

OPEN ACCESS

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

**Reviewed by:**
Qiqi Xie,
Lanzhou University, China
Jialiang Yang,
Genesis (Beijing) Co., Ltd., China

***Correspondence:**
Changsheng Ye
yechsh2006@126.com
Weiwei Huang
huangstudenth@163.com
Nani Li
linanifujian@126.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:** 27 November 2020
**Accepted:** 31 December 2020
**Published:** 21 January 2021

**Citation:**
Huang Z, Xiao C, Zhang F, Zhou Z, Yu L, Ye C, Huang W and Li N (2021) A Novel Framework to Predict Breast Cancer Prognosis Using Immune-Associated LncRNAs. Front. Genet. 11:634195. doi: 10.3389/fgene.2020.634195

**Background:** Breast cancer (BC) is one of the most frequently diagnosed malignancies among females. As a huge heterogeneity of malignant tumor, it is important to seek reliable molecular biomarkers to carry out the stratification for patients with BC. We surveyed immune- associated lncRNAs that may be used as potential therapeutic targets in BC.

**Methods:** LncRNA expression data and clinical information of BC patients were downloaded from the TCGA database for a comprehensive analysis of candidate genes. A model consisting of immune-related lncRNAs enriched in BC cancerous tissues was established using the univariate Cox regression analysis and the iterative Lasso Cox regression analysis. The prognostic performance of this model was validated in two independent cohorts (GSE21653 and BC-KR), and compared with known prognostic biomarkers. A nomogram that integrated the immune-related lncRNA signature and clinicopathological factors was constructed to accurately assess the prognostic value of this signature. The correlation between the signature and immune cell infiltration in BC was also analyzed.

**Results:** The Kaplan-Meier analysis showed that the OS of Patients in the low-risk group had significantly better survival than those in the high-risk group, Clinical subgroup analysis showed that the predictive ability was independent of clinicopathological factors. Univariate/multivariate Cox regression analysis showed immune lncRNA signature is an important prognostic factor and an independent prognostic marker. In addition, GSEA and GSVA analysis as well as comprehensive analysis of immune cells showed that the signature was significantly correlated with the infiltration of immune cells.

**Conclusion:** We successfully constructed an immune-associated lncRNA signature that can accurately predict BC prognosis.

**Keywords: TCGA, GEO, immune, breast cancer, risk score, framework, prognosis**

## INTRODUCTION

Globally, Cancer Statistics 2018 estimates that more than 2.1 million females are diagnosed with breast cancer (BC) every year and about 62,000 deaths occur, making it the most common female malignancy and the leading cause of cancer mortality in women. Developing countries account for nearly 60% of BC mortality (Bray et al., 2018). Although the progress of early screening and recent advances in anti-cancer strategies offer improvements in the main outcomes of BC patients (Kim et al., 2018; Siegel et al., 2020), the recurrence rate of BC remains high (Li et al., 2019). Several studies have identified a large amount of poor-prognosis related biomarkers for BC, including age, tumor size, histological grade, lymphatic vessel invasion (LVI), number of metastatic lymph nodes, hormone receptor status, c-erbB2 status, and positive margin status. However, because of the complexity of BC incidence and the heterogeneity of tumors, pre-existing prognostic markers are underpowered to increase the prediction efficiency. The emerging technique, an optimal model consisting of relevant lncRNAs, will renew hope for pronounced improvements in predicting the BC prognosis.

LncRNAs, located in the nucleus or cytoplasm, are a group of RNA molecules of greater than 200 nt in length. The majority of them have no protein-coding function, and a few can encode a limited number of polypeptides. LncRNAs are engaged in precise regulations at pre-transcription and transcription levels and a broad spectrum of biological processes, such as tumor invasion, metastasis, apoptosis, and drug resistance. Therefore, lncRNA abnormalities in the peripheral blood of BC patients may promote BC formation and progression, promoting early diagnosis and timely treatment for patients. Many studies have confirmed the regulatory effects of lncRNAs on malignancy behaviors in tumorigenesis (Mercer et al., 2009; Fang and Fullwood, 2016; Bin et al., 2018), including proliferation, adhesion, migration, and apoptosis of tumor cells (Wang et al., 2020). HOXA11-AS and MALAT1 are identified as the lncRNAs related to cell cycle progression, which can be used as prognostic biomarkers for the survival of glioma patients (Wang et al., 2016).

Immune-related lncRNAs refer to those with various functions in regulating gene expressions and critical roles (e.g., the control of the differentiation and function of immune cell types, dendritic cell activity, T cell ratio, and metabolism) in innate or adaptive immune responses (Denaro et al., 2019). Hu et al. (2013) demonstrated that a lincRNA Ccr2-5′AS was an important part of the regulatory circuit in gene expression specific to the TH2 subset of helper T cells. Yang et al. (2020) found that LNC TANCR play important roles in γδ T cell activation. Huang et al. (2018) noticed that lncRNA NKILA, regulated T cell sensitivity to Activation-induced cell death (AICD) by inhibiting NF-κB activity. However, details of immune-related lncRNAs related to BC are rarely reported. In this study, an immune-related lncRNA signature was constructed and verified in an independent set and its association with immune cell infiltration was analyzed to evaluate the clinical predictive value of the signature.

## MATERIALS AND METHODS

### Data Download, Pre-processing and Screening of Prognostic IRGs

To identify BC-related lncRNAs, lncRNA profile, and clinical data from the TCGA training set were downloaded from the UCSC Xena database[1]; To verify the multi-lncRNA signature, gene profile and clinical information of the Breast Cancer-KR (BC-KR) data set were downloaded from ICGC[2]; GEOquery package with R language (Davis and Meltzer, 2007) was used to download and analyze the reliable BC expression data set GSE21653 from the GEO database, the source of species was human, and the platform was based on the GPL570 ([HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array), including 266 BC samples included in this study. The BC-KR data set and the GSE21653 data set were used as verification sets. Firstly, the gene expression data and the corresponding clinical data of BC patients were pre-processed; the immune response genes (IRGs) were downloaded and sorted out from ImmPort database[3]. The prognostic immune-related lncRNAs were screened out from the TCGA data set using the univariate Cox regression method, and the statistical standard of $p < 0.05$ was used. We performed GO and KEGG enrichment analysis to extract prognostic IRGs (**Supplementary Figure 1**).

### Screening, Co-expression Analysis and GO Enrichment Analysis of Immune-Related LncRNAs

The gene list of expression matrix was annotated and classified by mRNA and lncRNA, the mRNA expression matrix and lncRNA expression matrix were distinguished, and Pearson coefficient between lncRNA gene expression and prognostic IRGs was calculated. Pearson coefficient $<-0.4$ and $P < 0.01$ were set as the identification criteria of lncRNAs related to immune. In order to explore the possible function of lncRNAs related to immune in the occurrence and development of BC, co-expression analysis was performed on mRNA expression matrix and immune-related lncRNAs expression matrix. GO enrichment analysis was performed to annotate the immune functions of target immune-related lncRNAs using the ClusterProfiler package (Yu et al., 2012). Adj $p$ value $< 0.05$ was considered statistically significant.

### Construction and Evaluation of a Prognosis Model

A least absolute shrinkage and selection operator (Lasso) Cox regression model was used to identify optimal lncRNA candidates and construct the immune-related lncRNA signature. The process of cross-verification is to select samples randomly. The optimal value of λ was identified to train the model. Optimal lncRNA candidates were included during the training and stopped when the area under the curve (AUC) of

[1]https://xenabrowser.net/datapages/

[2]https://dcc.icgc.org/

[3]https://ImmPort.niaid.nih.gov

ROC reached the peak. Therefore, the optimal multi-lncRNA signature containing the least number of prognostic immune-related lncRNAs was refined (Tibshirani, 1997). Subsequently, a univariate Cox regression analysis was implemented to identify and confirm prognostic immune-related lncRNAs in the signature, with $p < 0.05$ as the standard. An iterative Lasso Cox regression analysis was performed to construct the prognostic immune-related lncRNA signature using the glmnet package (Friedman et al., 2010). The specific method was to count the consensus genes whose gene frequency is more than 100 after 1000 Lasso Cox regression. Then, a multivariate Cox regression analysis was conducted to assess the prognostic capacity of the refined prognostic gene signature, using the expression values of genes as covariates. Subsequently, a prognostic risk signature was conducted for each patient from the TCGA training cohort according to the refined prognostic gene-expression signature. Using the median risk score as the threshold, these patients were divided into the high- and low-risk groups and the survival differences between the high- and low-risk groups were analyzed using the R software.

## Comparison of the Prognostic Multi-LncRNA Signature With Known Biomarkers and Validation of the Signature

To explore the prognostic capability and prediction efficiency of our signature for BC, a comparative survival analysis was used to compare the prognostic capability between the signature and known prognostic biomarkers, and

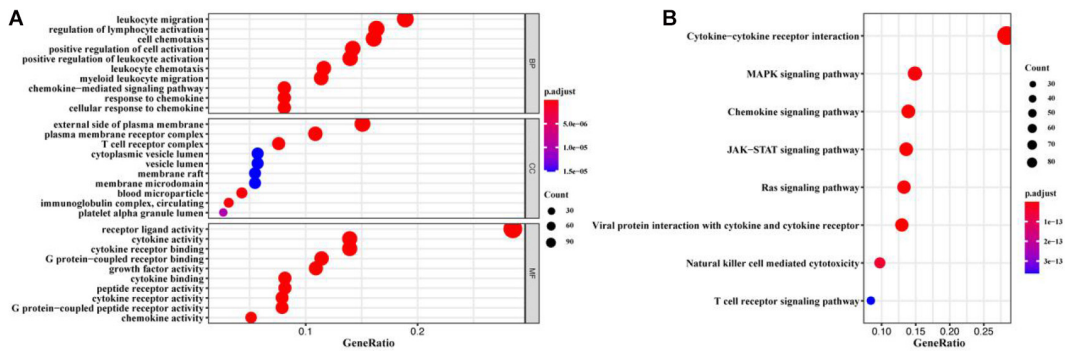


**FIGURE 1 |** GO and KEGG enrichment analysis of prognostic IRGs. **(A)** GO enrichment result. **(B)** KEGG enrichment result. The *x*-axis represented the ratio of the number of genes enriched in a single GO (or KEGG) term to the total number of genes changed in all GO (or KEGG) terms, and the *y*-axis represents GO (or KEGG) terms. The color of the circle indicated the adjusted *p* value, and the size represented the number of genes enriched in the term or the pathway.

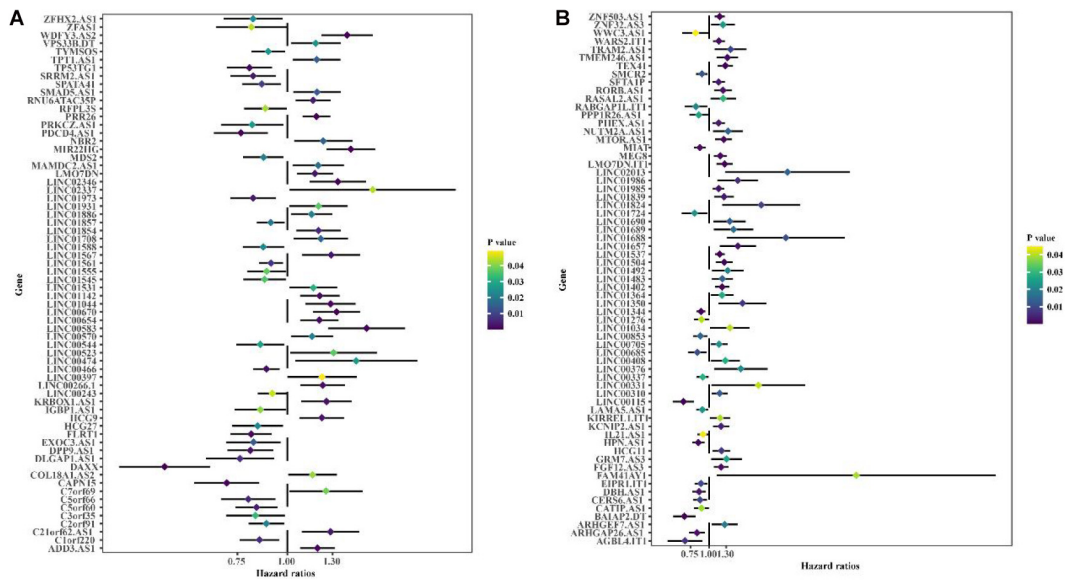


**FIGURE 2 |** Univariate Cox regression is used to identify the forest graph of 166 immune-associated lncRNAs related to prognosis. **(A)** Part A of 166 immune-associated lncRNAs. **(B)** Part B of 166 immune-associated lncRNAs. The yellower the color is, the higher the *p* value; the bluer the color is, the smaller the *p* value.

the ROC curve was used to analyze the sensitivity and specificity of the prognostic risk signature. Based on the prognostic multi-lncRNA signature in the training set, its prognostic performances were verified in two independent validation sets (BC-KR and GSE21653 validation sets).

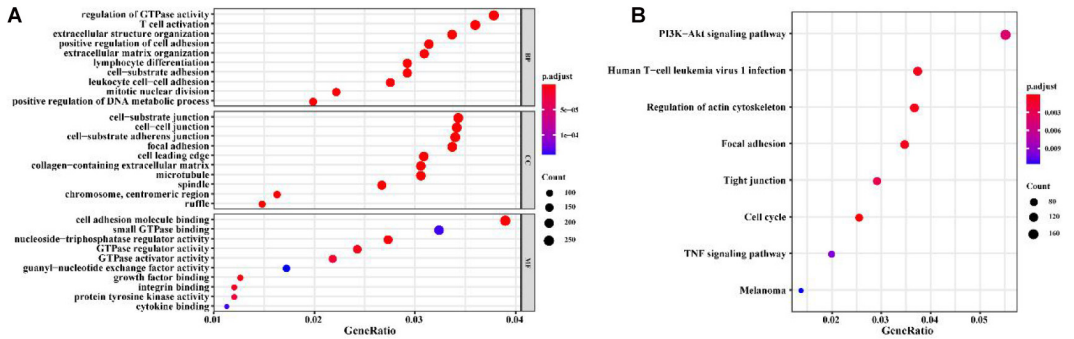


**FIGURE 3 |** GO and KEGG enrichment analysis for the immune-related lncRNAs co-expressed with prognostic mRNAs in BC. **(A)** GO enrichment result. **(B)** KEGG enrichment result. The *x*-axis represents the ratio of the number of genes enriched in a single GO (or KEGG) term to the total number of genes changed in all GO (or KEGG) terms, and the *y*-axis represents the GO (or KEGG) terms. The color of the circle indicates the adjusted *p* value, and the size represents the number of genes enriched in the term or the pathway.

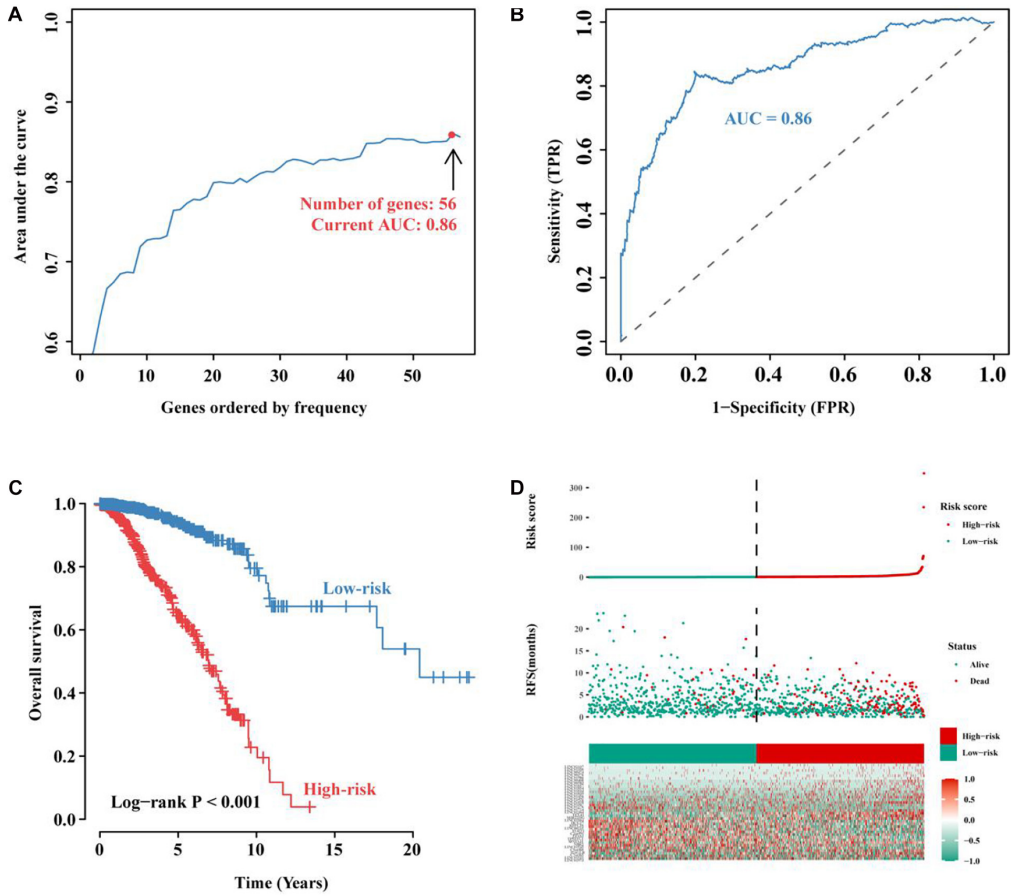


**FIGURE 4 |** Construction and evaluation of the immune-related 56-lncRNA signature. **(A)** The immune-related lncRNA signature constructed by the iterative Lasso Cox regression analysis. **(B)** ROC curve used to evaluate the predictive efficiency of the signature. **(C)** Kaplan-Meier curve for the OS of BC patients in the high- and low-risk groups. **(D)** Risk-score distribution. In the *x*-axis, red represents higher expression levels while green represents lower expression levels. Risk scores for BC patients in the high- (red, the *y*-axis) and low-risk groups (green, the *y*-axis) were listed in ascending order.

## Subgroup Survival Analysis, the Prognostic Value of the Signature, and Construction of Nomograms

The prognostic value of any biomarker can be evaluated by whether they are independent of the existence of clinicopathological prognostic factors. In this study, to evaluate the independence and applicability of the validated signature, the BC patients in the TCGA cohort were reassigned according to different clinicopathological characteristics. Kaplan-Meier survival analysis was performed to compare the prognostic capability between subgroups. Decision curve analysis was used to evaluate the net benefit and clinical value of the signature. Univariate Cox regression and multivariate Cox regression were used to evaluate the associations between clinical features and the risk score values, which were illustrated as nomograms.

## Phenotype Differences Between the High- and Low-Risk Groups Using GSEA and GSVA Analysis

Differences in functional phenotypes between the high- and low-risk groups were analyzed using the gene set enrichment analysis (GSEA) and GSVA analysis. We selected "c2.all.v7.1.symbols.gmt," "c5.all.v7.1.symbols.gmt," and "h.all.v7.1 symbols.gmt" as the reference gene sets. The differences in activated pathways between the high- and low-risk groups were identified using GSEA v4.0.3 software, and permutations of the gene sets were performed 1000 times for each analysis. A nominal $p < 0.05$ and a false discovery rate (FDR) $< 0.05$ were considered as statistical significance. We selected "h.all.v7.1 symbols.gmt" as the reference gene set. The differences in activated pathways were also analyzed using the ClusterProfiler package and GSVA package (Hänzelmann et al., 2013). A $p$ value $< 0.05$ was considered statistically different.

## Correlation Between the Prognostic Signature and Immune Cell Infiltration

To explore the association between the lncRNA signature and immune cell infiltration in BC, gene expression matrix data was uploaded to CIBERSORT (Newman et al., 2015) for cell type-specific gene expression purification. Infiltrating immune cells were filtered for analysis, and the abundances of infiltrating immune cells between the high- and low-risk groups were compared with CIBERSORT $p < 0.05$ for the eligible samples. Survival-related infiltrating immune cells were identified using Kaplan-Meier survival analysis. Finally, the correlation analysis was performed for the obtained prognostic biomarkers and immune cell infiltration. The results were visualized by using ggplot2 package (Ginestet, 2011).

# RESULTS

## Screening of Immune-Related LncRNAs for the Prognosis of BC Patients

A total of 1900 IRGs were retrieved from the ImmPort database. Univariate Cox regression analysis showed that 516 IRGs (**Supplementary Table 1**) were associated with the prognosis of BC patients. We further identified immune functions of the prognostic IRGs via the gene GO and KEGG enrichment. We found that the 516 prognostic IRGs

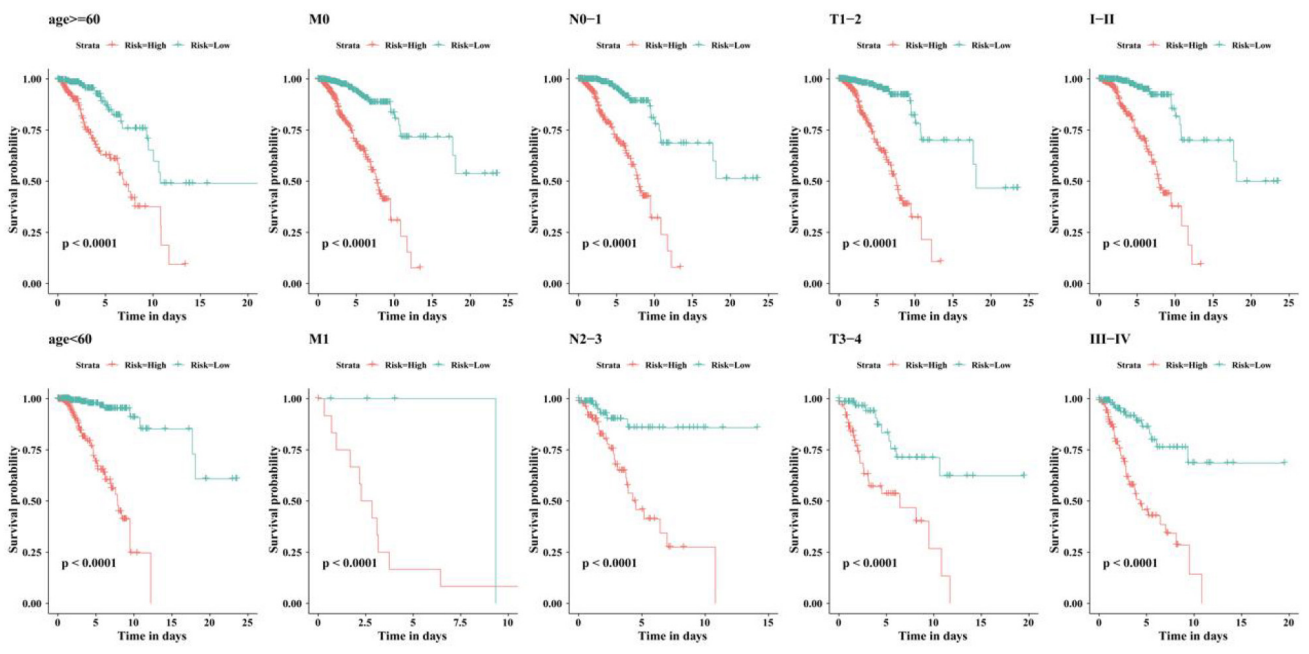


**FIGURE 5 |** Kaplan-Meier survival curves of BC patients in different clinical subgroups.

were significantly enriched in GO terms related to leukocyte migration, regulation of lymphocyte activation, cell chemotaxis, external side of plasma membrane, plasma membrane receptor complex, T cell receptor complex, and receptor-ligand activity, and in KEGG pathways related to cytokine-cytokine receptor interaction, MAPK signaling pathway, chemokine signaling pathway, and JAK-STAT signaling pathway (**Figure 1** and **Supplementary Table 2**).

## Co-expression Analysis and GO Enrichment Analysis of Immune-Related LncRNAs

Co-expression analysis showed 948 immune-related lncRNAs linked to immune responses in BC (**Supplementary Table 3**). To explore the prognostic capacity of the immune-related lncRNAs, univariate Cox regression analysis was implemented and identified that 166 immune-related lncRNAs (**Supplementary Table 4**) were related to the maximum prognostic value ($P < 0.05$, **Figure 2**; $P < 0.05$). To understand the immune function of the 166 genes, we performed functional enrichment analysis for these genes. GO enrichment analysis revealed that these genes were mainly enriched in GO terms related to the regulation of GTPase activity, T cell activation, extractable structure organization, cell-substitute junction, cell adhesion, and molecular binding (**Figure 3** and **Supplementary Table 5**). These results suggest that immune-related lncRNAs may participate in the occurrence and development of BC by regulating immune responses.

## Construction and Evaluation of the Prognostic Model

To further improve the prediction efficiency of the signature, we performed iterative Lasso Cox regression analysis and built an optimal signature containing 56 immune-related lncRNAs (**Figure 4A**). To verify its predictive efficiency, we

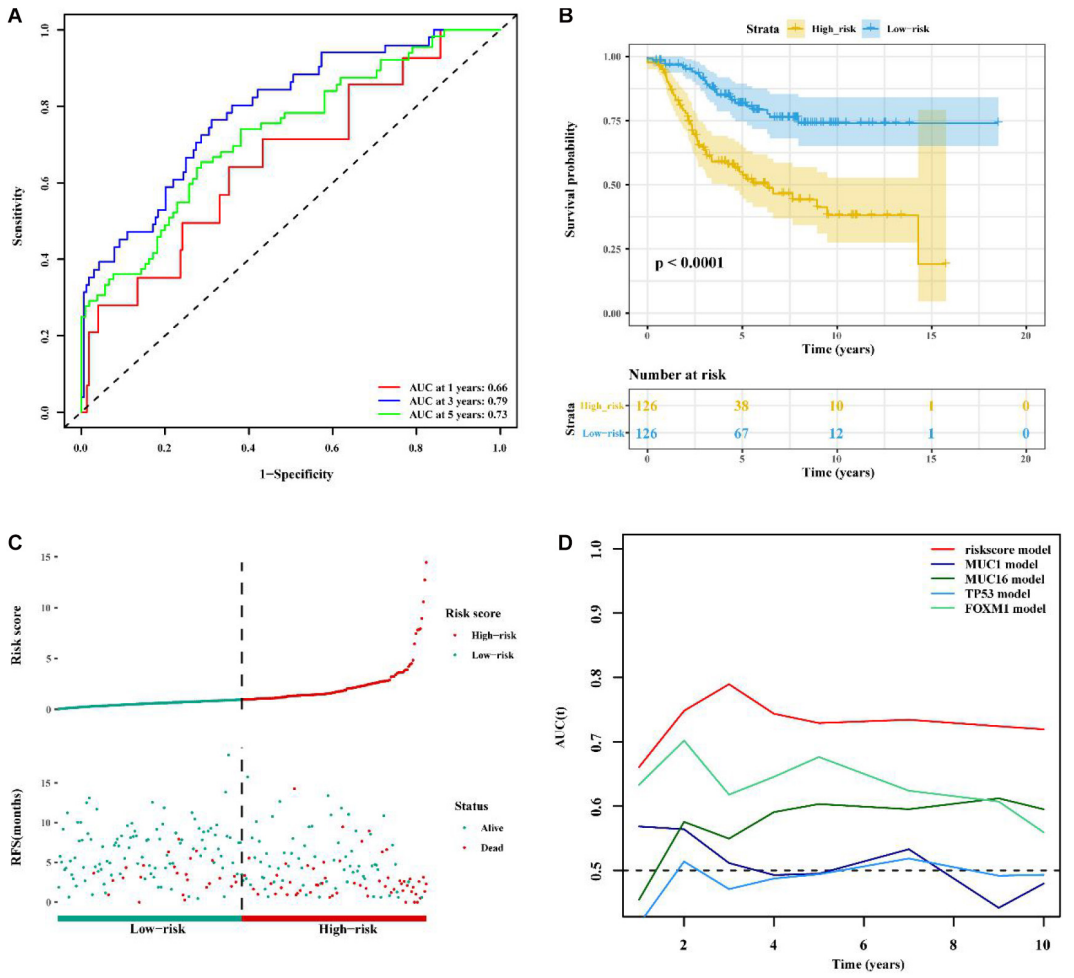


**FIGURE 6 |** Comparison of prognostic performances between the lncRNA signature and known prognostic biomarkers and validation using the GSE21653 cohort. **(A)** The ROC curve to verify the prediction efficiency of the lncRNA signature. **(B)** Kaplan-Meier curve for the OS of BC patients in the high- and low-risk groups. **(C)** The risk-score distribution and duration of RFS of the lncRNA signature in the GSE21653 validation cohort. **(D)** Time-dependent ROC curve to compare prediction efficiency between the lncRNA signature and the known prognostic biomarkers.

performed ROC analysis and showed that the immune-related 56-lncRNA signature had a good prognostic performance in BC (AUC = 0.86, **Figure 4B**). Kaplan-Meier survival analysis showed that the OS of patients in the high-risk group was significantly worse than that in the low-risk group ($P < 0.001$, **Figure 4C**). The distribution risk scores, duration of recurrence-free survival (RFS), and lncRNA expression levels were illustrated in **Figure 4D**. The number of deaths in the high-risk group was significantly higher than that in the low-risk group. These results suggest that the immune-related 56-lncRNA signature can distinguish high-risk patients from low-risk individuals, and it is of great significance in predicting the prognosis of BC.

## The LncRNA Signature as an Independent Prognostic Predictor

We examined the independent prognostic ability of the immune-related lncRNA signature. The BC patients from the TCGA cohort were reassigned according to different prognostic and clinicopathological factors consisting of age, metastasis, and tumor staging (TNM and clinical staging). Kaplan-Meier survival analysis was performed for each subgroup and showed that the survival time of BC patients in the low-risk group was significantly prolonged regardless of age, TNM staging, and clinical staging ($p < 0.0001$, **Figure 5**). This indicates that the immune-related 56-lncRNA signature is independent of gender, age, and metastasis in survival prediction, which can be considered to be an independent predictor of the prognosis of BC patients.

## Validation of the LncRNA Signature and the Comparison Between the Signature and Known Prognostic Biomarkers

To determine whether the immune-related 56-lncRNA signature was superior to known prognostic biomarkers in terms of prognostic capability and prediction efficiency, the validation was performed in the other two independent verification set

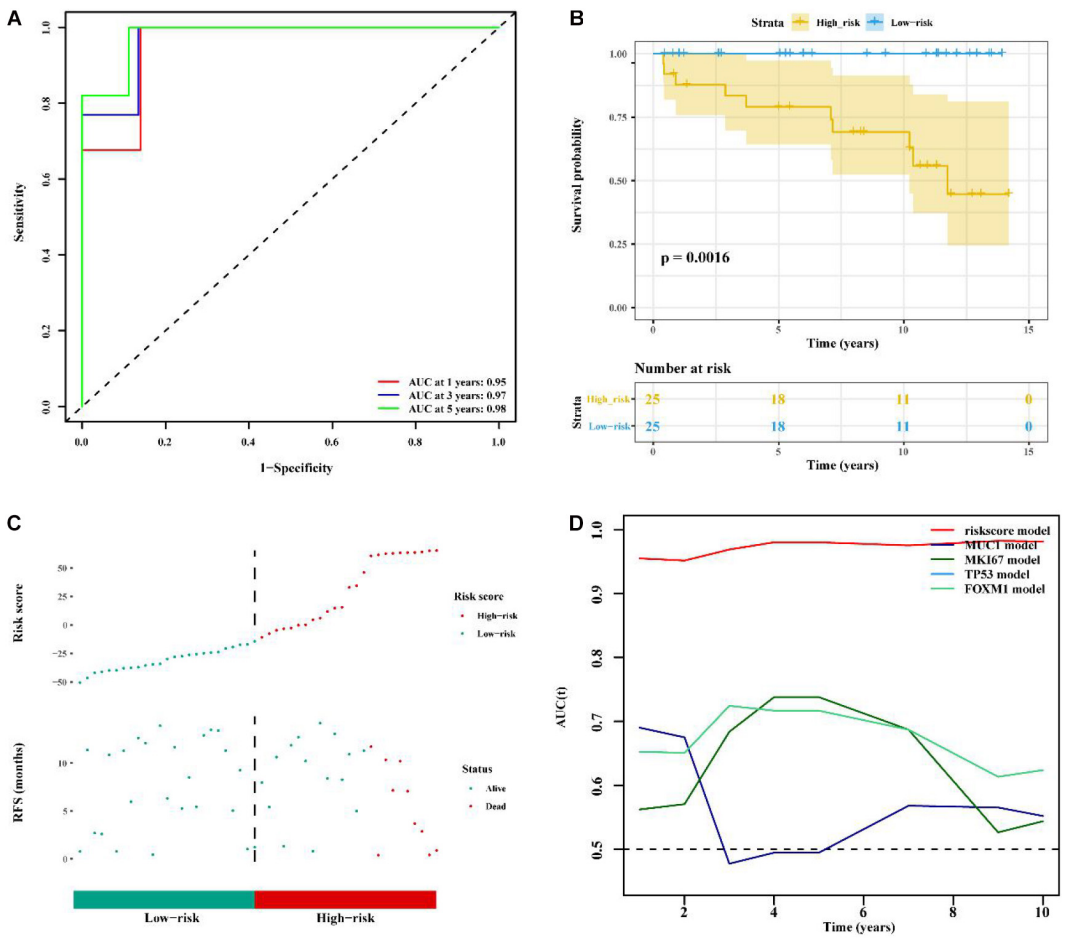


**FIGURE 7 |** Comparison of prognostic performances between the lncRNA signature and known prognostic biomarkers and validation using the BC-KR cohort. **(A)** The ROC curve to verify the prediction efficiency of the lncRNA signature. **(B)** Kaplan-Meier curve for the OS of BC patients in the high- and low-risk groups. **(C)** The risk-score distribution and duration of RFS of the lncRNA signature in the BC-KR validation cohort. **(D)** Time-dependent ROC curve to compare prediction efficiency between the lncRNA signatures and known prognostic biomarkers.

GSE21653 data set and the BC-KR data set cohort. The ROC analysis showed a similar prediction efficiency between the training cohort and the GSE21653 validation cohort (**Figure 6A**). Kaplan-Meier survival analysis showed that the OS of BC patients in the high-risk group was significantly worse than that in the low-risk group ($P < 0.001$, **Figure 6B**). The distribution of risk scores and duration of RFS were illustrated in **Figure 6C**. The number of deaths in the high-risk group was significantly higher than that in the low-risk group. ROC analysis showed that the AUC of the lncRNA signature was higher than that of known prognostic biomarkers (**Figure 6D**). The lncRNA signature had better stability and reliability in predicting the survival of BC patients, which should be considered as a good prognostic biomarker. The validation using the BC-KR cohort showed similar results in prognostic capability and prediction efficiency of the signature, more deaths in the high-risk group, and the superiority over known biomarkers (**Figure 7**). These results suggest that the immune-related 56-lncRNA signature can predict the survival of BC patients in other independent cohorts.

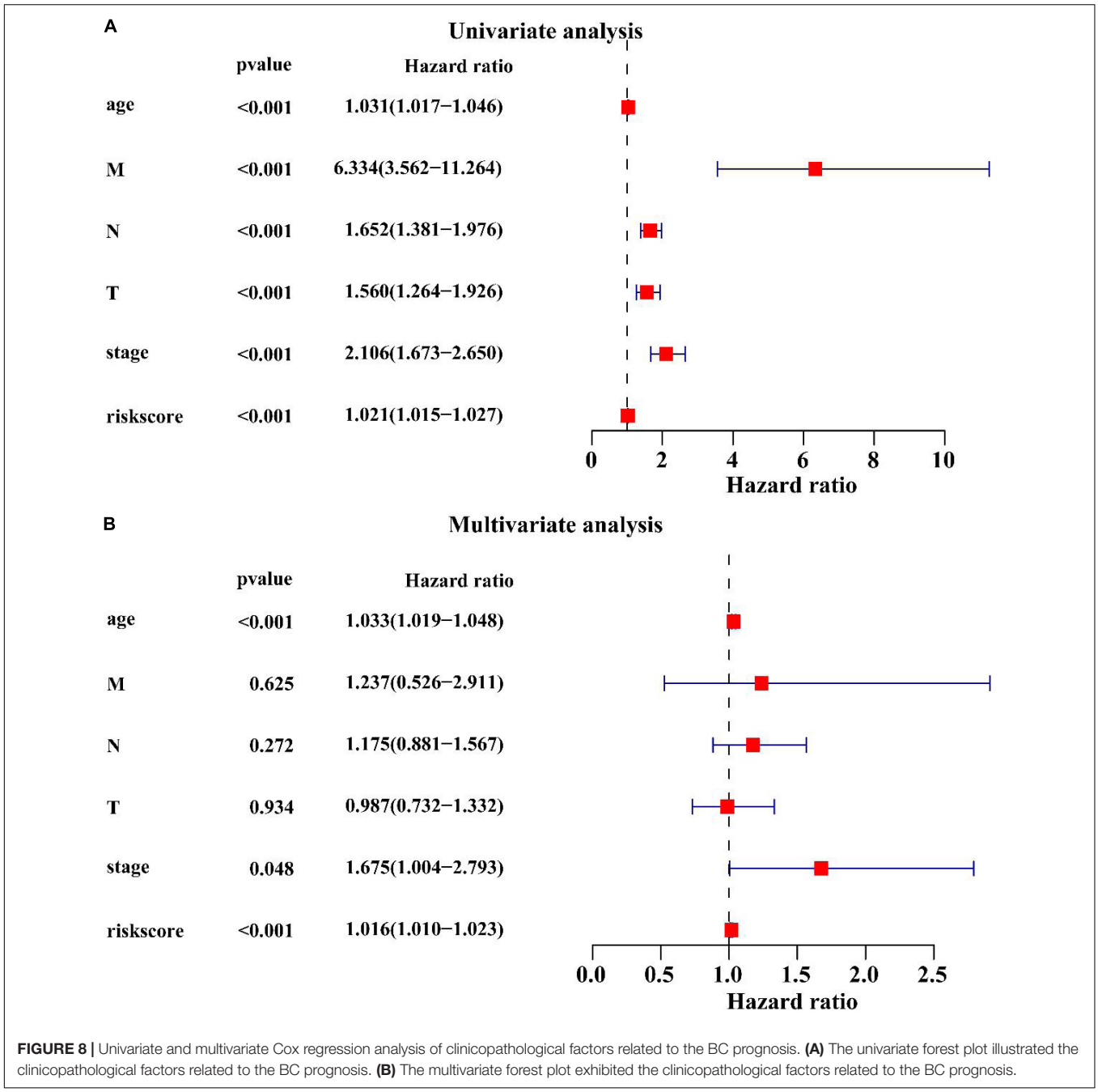


**FIGURE 8 |** Univariate and multivariate Cox regression analysis of clinicopathological factors related to the BC prognosis. **(A)** The univariate forest plot illustrated the clinicopathological factors related to the BC prognosis. **(B)** The multivariate forest plot exhibited the clinicopathological factors related to the BC prognosis.

## Univariate and Multivariate Cox Regression Analysis of Clinicopathological Factors Related to the BC Prognosis

Univariate Cox regression analysis showed that age (HR = 1.031, $P < 0.001$), M stage (HR = 6.334, $P < 0.001$), N stage (HR = 1.652, $P < 0.001$), T stage (HR = 1.560, $P < 0.001$), clinical stage (HR = 2.106, $P < 0.001$), the lncRNA signature (HR = 1.021, $P < 0.001$) were associated with the BC prognosis (**Figure 8A**). Multivariate Cox regression analysis revealed that age (HR = 1.033, $P < 0.001$), clinical stage (HR = 1.675, $P < 0.001$) and the lncRNA signature (HR = 1.016, $P < 0.001$) were significantly associated with the BC prognosis (**Figure 8B**). These results suggest that the lncRNA signature is an important prognostic factor for BC.

## Construction of a Nomogram to Predict the OS of Patients With BC

Combined with clinicopathological characteristics, we constructed a nomogram containing the immune-related 56-lncRNA signature (**Figure 9**) to predict the 3-, 5-, and 8-year survival rate of BC patients. The results suggest that the nomogram can independently evaluate the survival of BC patients, which may help physicians make better medical decisions and follow-up plans.

## GSEA and GSVA Analysis for Phenotype Differences Between the High- and Low-Risk Groups

We conducted GSEA and GSVA analysis to identify differences in main functional phenotypes between the high- and low-risk groups. As shown in **Figures 10A,B** and **Supplementary Figure 2A**, GSEA analysis revealed that the pathways involved in BC patients from the high-risk group consisted of CELL_CELL_JUNCTION, DUTERTRE_ESTRADIOL_RESPONSE_24HR_DN, EPITHELIAL _MESENCHYMAL_TRANSITION, KRAS_SIGNALING _Up, Zhang_INTERFERON_ RESPONSE, and POSITIVE _T_CELL_SELECTION. GSVA analysis showed that HEME _ METABOLISM, TGF_BETA_SIGNALING, KRAS_SIGNALING_UP, IL6_JAK_STAT3_ SIGNALING, INFLAMMATORY_RESPONSE were activated in BC patients from the high-risk group, and ESTROGEN_RESPONSE_LATE, MYC_TARGETS_V1, DNA_REPAIR, and INTERFERON_GAMMA_RESPONSE were inhibited (**Supplementary Figure 2B** and **Supplementary Table 6**).

## Correlation Between Immune Cell Infiltration and the LncRNA Signature

Based on the mRNA-lncRNA co-expression network, GO enrichment analysis had confirmed that the immune-related lncRNA signature was involved in the occurrence and development of BC by regulating immune responses. To explore the association between the lncRNA signature and immune cell infiltration, a comprehensive analysis was carried out to figure out the composition of infiltrating immune cells in the tumor microenvironment. Correlation analysis showed that cytotoxic cells were positively correlated with T cells and B cells, while Th2 cells were negatively correlated with other immune cells (**Figure 11A**). We predicted the interaction network between immune cells, and the results showed that $CD8^+$ T cells, Th1 cells, and B cells had a strong interaction with other immune cells, while Tgd, T helper cells, and Th17 cells had a weak interaction with other immune cells (**Figure 11B**). The analysis of immune cell composition showed that BC patients had higher numbers of T helper cells, $CD8^+$ T cells, and T helper cells but lower numbers of B cells and regulatory T (TReg) cells in the tumor microenvironment (**Figure 11C**). We compared the difference in the expressions of infiltrating immune cells in BC patients between the high-risk group and the low-risk group. As shown in **Figure 11D**, the levels of aDC, Macrophages, Tgd, Macrophages, and TReg were significantly different between the two groups ($P < 0.01$). The prognosis analysis for infiltrating immune cells showed that aDC, Neutrophils, Tem, DC, NK CD56dim cells, Tgd, Mast cells, TCM, and TReg were correlated with the prognosis of BC patients (**Figure 12**). The correlation analysis confirmed a significant correlation between the infiltrating

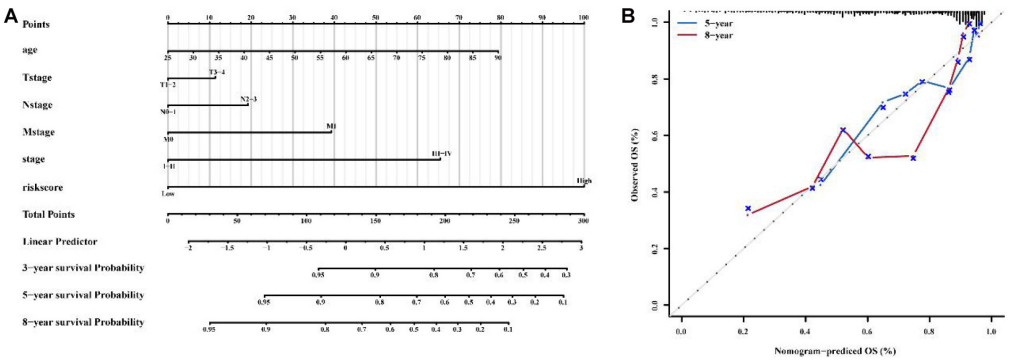


**FIGURE 9 |** Construction of a nomogram to predict the OS of patients with BC. **(A)** nomogram to predict survival probability at 1, 3, 5 and 8 years for BC. **(B)** Calibration curve with Hosmer-Lemeshow test of the nomogram.

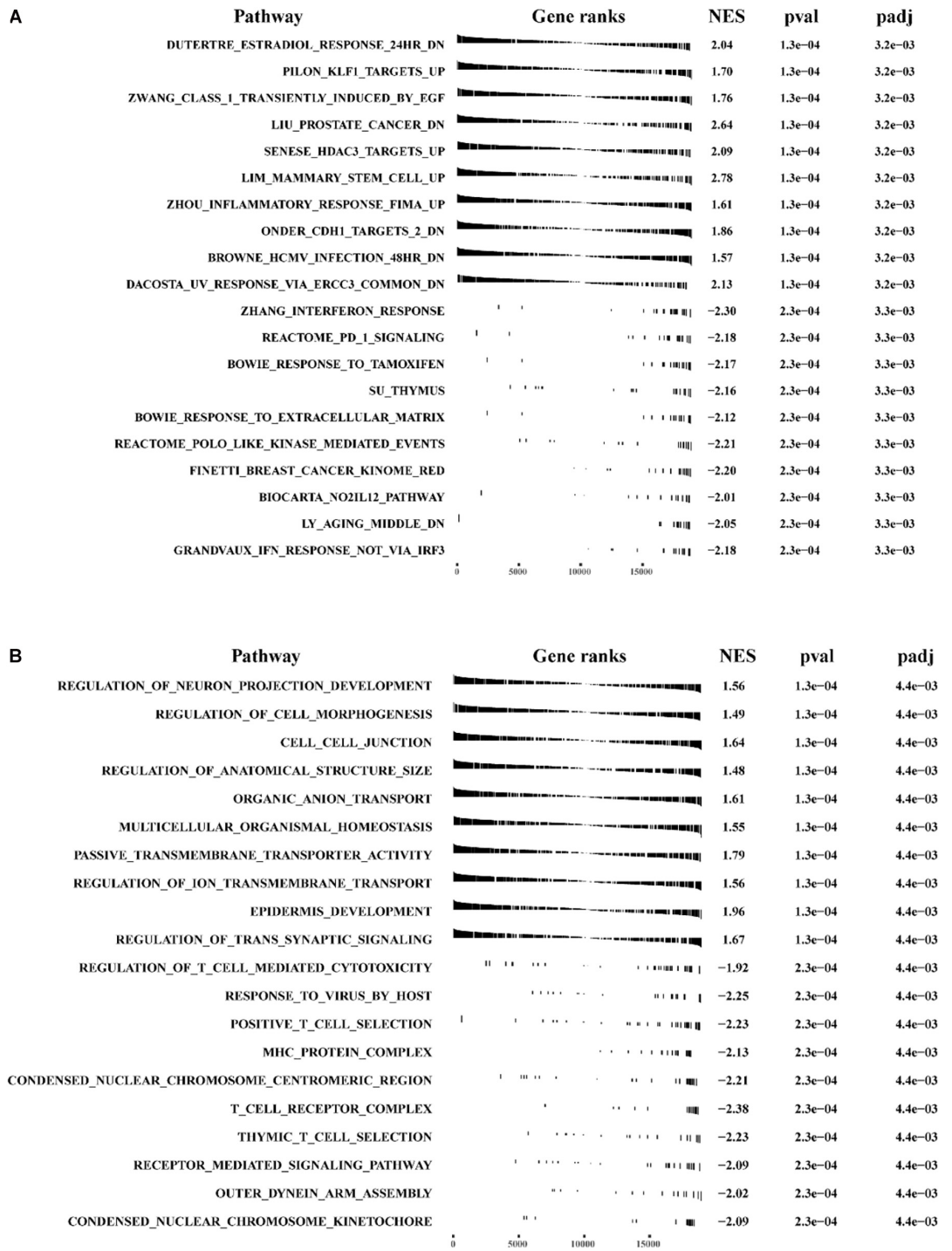


**FIGURE 10 |** The GSEA analysis for pathway enrichment between the high-risk group and the low-risk group. GSEA enrichment analysis was performed using **(A)** "c2.all.v7.0 symbols.gmt" and **(B)** "c5.all.v7.0 symbols.gmt" as the reference gene sets.

immune cells and the lncRNAs in the signature (**Supplementary Figure 3**). These results suggest that the lncRNAs in the signature have a complex regulatory relationship with immune cell infiltration, which may be an important way to participate in the development of BC and have high research value.

## DISCUSSION

Breast cancer remains one of the most lethal malignancies in the world. Although currently available cancer treatments (such as surgery, chemotherapy, radiotherapy, endocrine therapy,

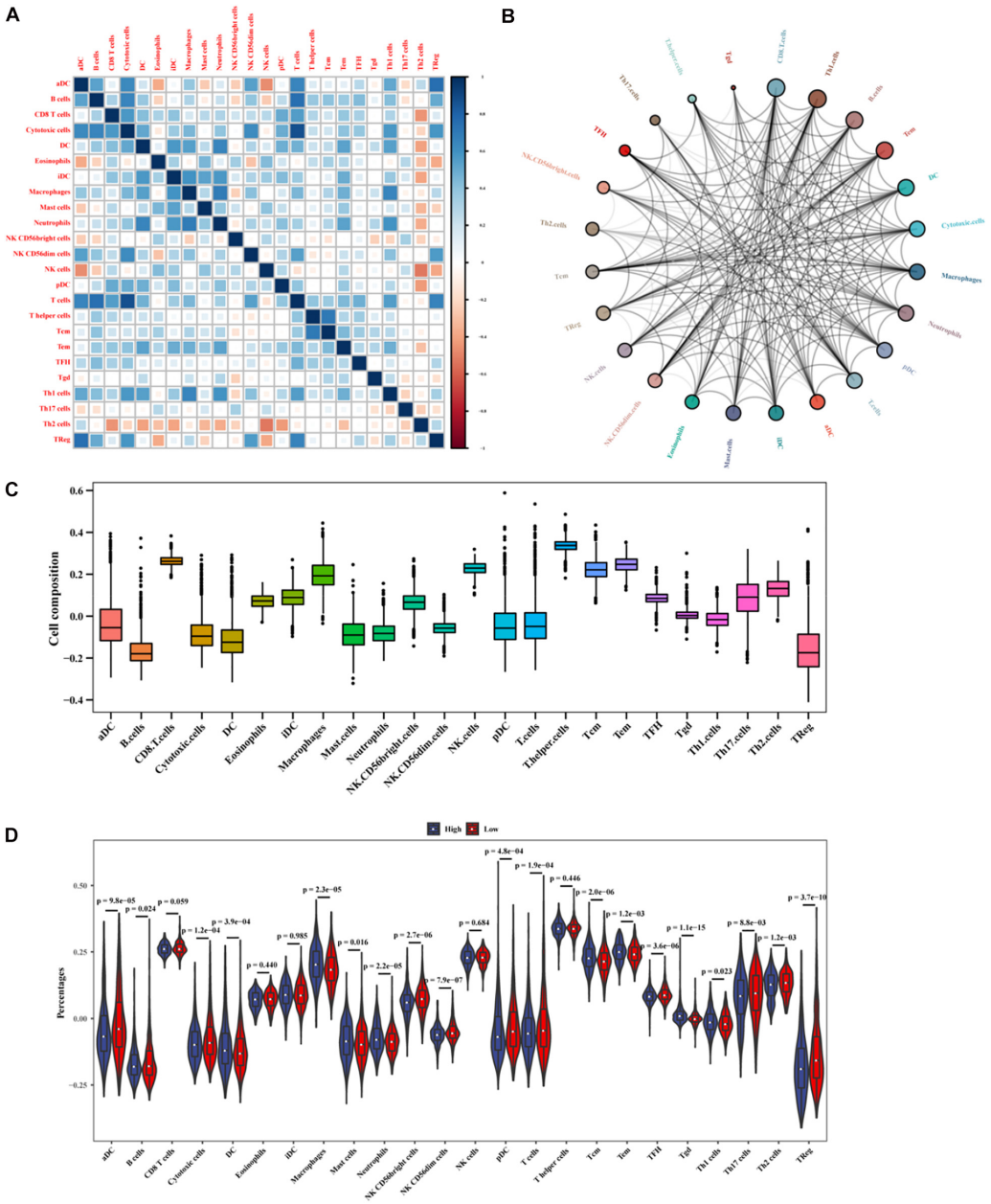


**FIGURE 11 |** Analysis of immune cell infiltration in BC patients. **(A)** The correlation between infiltrating immune cells. **(B)** The cycle indicates the immune cell-cell interactions in the tumor microenvironment in BC. A larger circle corresponds to more interactions between immune cells and vice versa. **(C)** The *x*-axis represents immune cell types, and the *y*-axis represents the percentage of immune cells. **(D)** The differences in the expressions of infiltrating immune cells between the high- and low-risk groups.

targeted therapy, and immunotherapy) have a positive impact on BC survival, increasing life expectancy in a short period seems unrealistic. The risk of BC recurrence is particularly high, suggesting that BC recurrence is one of the most important causes of mortality. The difficulty of predicting the BC prognosis among a complex tumor microenvironment comprised of molecular and cellular heterogeneity between some commonly used markers in several cancers, such as CA 15-3, CA 27.29, CEA, MUC1, MUC16, TP53, FOXM1, ER, PR, HER2, KI 67, and others (Børresen-Dale, 2003; Lakshmanan et al., 2012; Graham et al., 2014; Kabel, 2017; Song et al., 2017; Stergiou et al., 2019), makes a prediction inaccurate. Therefore, the proposal of an effective and reliable biomarker using a risk prediction model and cohort-specific imputation of gene expression to encourage the implementation of relevant clinical trials to validate this biomarker is currently possible. Advantages of a reliable biomarker are that high-risk populations can be closely monitored, management for the prevention and early detection of BC recurrence can be implemented, and once the recurrence is timely detected,

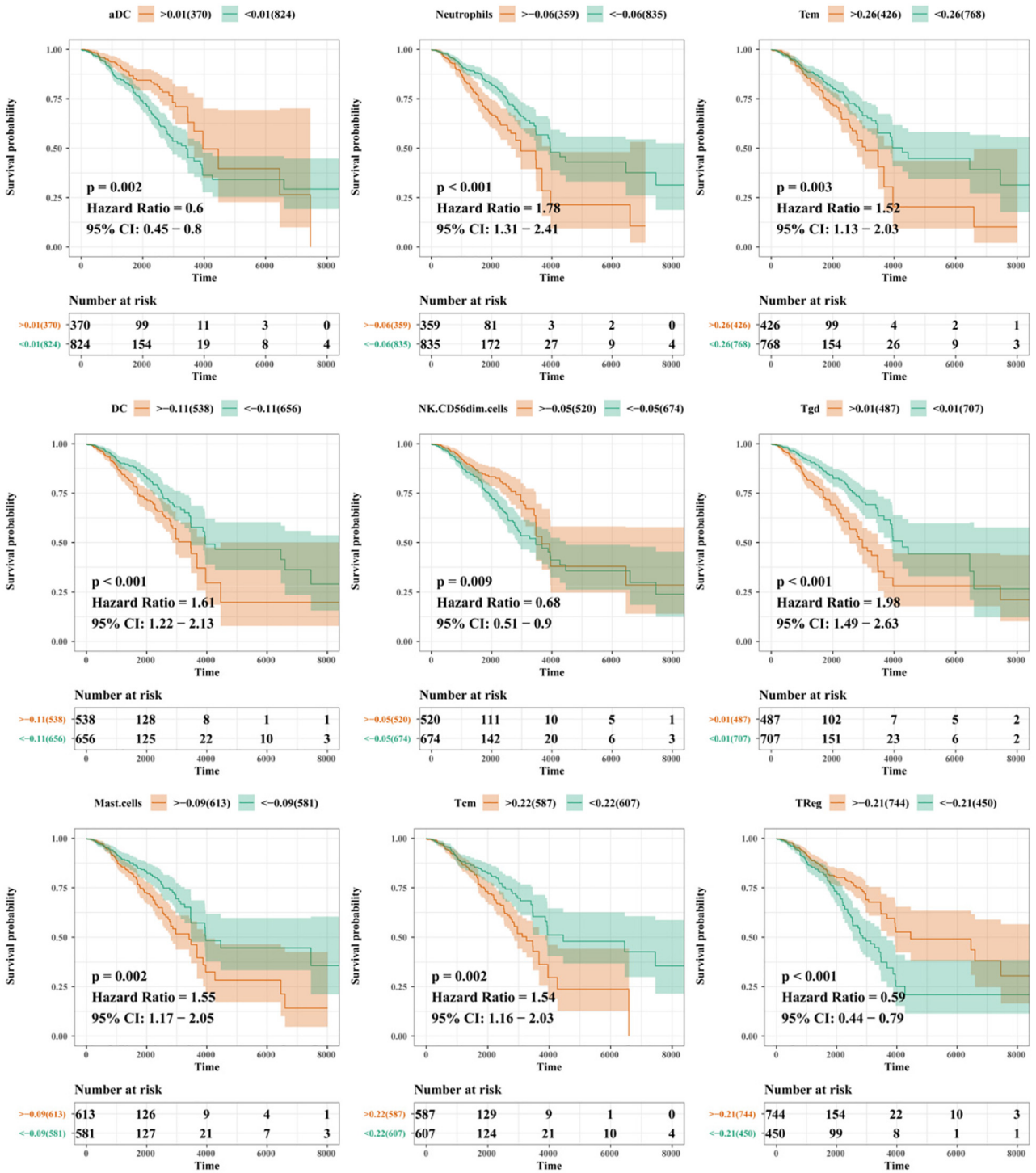


**FIGURE 12 |** Prognosis analysis for immune cells.

individualized treatment plans for prolonged survival can be prescribed.

Previous studies of prognostic biomarkers of BC have focused on coding genes and microRNAs. LncRNAs, especially immune-related lncRNAs, recently emerging as novel biomarkers have gained primary results. In-depth sequencing studies have found that lncRNAs directly interact with a variety of signaling pathways and regulators of oncogenes or tumor suppressor genes, thereby affecting tumor occurrence and development (Lin and Yang, 2018).

In this study, we initially identified 948 immune-related lncRNA candidates. Using a signature risk prediction model and the iterative Lasso Cox regression analysis, 56 lncRNAs with the maximum prognostic values were selected from the candidates to form a signature. ROC analysis showed good prognostic efficiency, with an AUC of 0.86. Finally, a 56-lncRNA signature was constructed. In the test set, BC patients were divided into the high- and low-risk groups according to their risk scores. Kaplan-Meier curve showed that patients in the low-risk group had a significantly longer OS than those in the high-risk group. Moreover, multivariate Cox regression analysis showed that the lncRNA signature was an independent predictor for BC prognosis in the two other validation cohorts from GEO and ICGC. Besides, the lncRNA signature offers a more effective prediction than known biomarkers. Clinical subgroup analyses showed the signature was independent of the presence of clinicopathological factors (gender, age, metastatic lymph nodes, and tumor metastasis) related to BC prognosis, indicating that the signature is an independent predictor of BC survival and the perspective of its clinical applicability is expected. The construction of a nomogram confirmed that the lncRNA signature can independently evaluate BC survival.

Cancer immunotherapy is a rapidly developing and exciting field of oncology. Eliciting adaptive immunity, particularly the adaptive infiltrating immune system which is an important component in the tumor microenvironment, is the goal of immunotherapy for effective and sustained anti-tumor responses to limit tumor growth and metastasis. Studies have shown that a combination of immunotherapy and other therapies can bring better benefits to patients with BC, especially advanced-stage, triple-negative BC (Schmid et al., 2018). In this study, GEVA analysis of mRNAs that are co-expressed with immune-related lncRNAs showed that TGF_BETA_SIGNALING and IL-6/JAK/STAT3TGF were the main enriched pathways. Hao et al. (2017) reported that TGF - β signaling pathway plays an important role in regulating the migration and invasion of lung cancer and BC cells. The IL-6/JAK/STAT3 signaling pathway is critical in regulating cell growth, survival, and differentiation in tumor cells, and also in mediating the differentiation and activation of T lymphocytes (Eskiler et al., 2019). Immune-related lncRNAs may participate in the occurrence and development of BC via directly interacting with a variety of signaling pathways and regulators of oncogenes or tumor suppressor genes to regulate adaptive immune responses. We further studied the relationship between lncRNAs in the signature and immune cell infiltration, and found that the levels of aDC, B cells, NK, CD56dim cells, T cells, TFH, and TReg in the low-risk group were higher than those in the high-risk group, ADC, NK, CD56DIM cells and TReg were correlated with BC prognosis, and ADC, NK, CD56DIM cells were positively correlated with T cells. Tumor-infiltrating lymphocytes are an important component of the tumor microenvironment in BC. Pantelidou et al. (2019) reported that Olaparib, a PARP inhibitor, inhibited tumor via activating the cGAs/STING pathway in tumor cells to induce T cell recruitment.

Pei et al. (2018) reported that LNC SNHG1 functioning as a competing endogenous RNA (CeRNA) inhibited the differentiation of TReg cells, thereby preventing BC immune escape. Wang et al. (2019) reported that TReg cells play a major role in the development of immunosuppressive tumor microenvironment mainly through cytokine signal transductions–a key determinant of immunosuppressive potential and unfavorable clinical outcomes. Inconsistent with these findings, we found that BC patients with a high TReg level had a better prognosis than those with a low TReg level (Takasato et al., 2020). The origin of tumor-infiltrating TReg cells and their association with circulating TReg cells are rarely reported, which will be explained in the follow-up study in our following publications.

Besides, there are some limitations of our study. Further cell and animal experiments are needed to verify the correlation between lncRNA expressions and immunophenotype in BC. To explore the underlying immune mechanisms, a larger data set is needed to verify the accuracy of the prediction model.

## CONCLUSION

In conclusion, we have constructed an immune-related 56-lncRNA signature which can be used as an independent prognostic marker for the survival risk subgroup of BC patients. This lncRNA signature is superior to known predictors in terms of prognostic performance in risk stratification, which will renew hope for experimental and clinical validations as well as clinical application in the future.

## 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

CY, WH, and NL conceived and designed the study. ZH, CX, and LY performed the experiments. FZ, WH, and ZZ analyzed the data. ZH wrote the manuscript. All authors have read and approved this manuscript.

## FUNDING

This work was supported by Science and Technology Program of Fujian Province, China (No. 2018Y2003); the Natural Science Foundation of Fujian Province (No. 2020J011112); and the Natural Science Foundation of Fujian Province (No. 2018J01265).

## ACKNOWLEDGMENTS

The authors thank reviewers for helpful comments on the manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.634195/full#supplementary-material

## REFERENCES


Bin, X., Hongjian, Y., Xiping, Z., Bo, C., Shifeng, Y., and Binbin, T. (2018). Research progresses in roles of LncRNA and its relationships with breast cancer. *Cancer Cell Int.* 18:179. doi: 10.1186/s12935-018-0674-0

Børresen-Dale, A.-L. (2003). TP53 and breast cancer. *Hum. Mutat.* 21, 292–300. doi: 10.1002/humu.10174

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

Davis, S. R., and Meltzer, P. S. (2007). GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Cancer Res.* 23, 1846–1847. doi: 10.1093/bioinformatics/btm254

Denaro, N., Merlano, M. C., and Lo Nigro, C. (2019). Long noncoding RNA s as regulators of cancer immunity. *Mol. Oncol.* 13, 61–73.

Eskiler, G. G., Bezdegumeli, E., Ozman, Z., Ozkan, A. Deveci, Bilir, C., Kucukakca, B. N., et al. (2019). IL-6 mediated JAK/STAT3 signaling pathway in cancer patients with cachexia[J]. *Bratislava Med. J.* 120, 819–826.

Fang, Y., and Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteom. Bioinform.* 14, 42–54. doi: 10.1016/j.gpb.2015.09.006

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.

Ginestet, C. (2011). ggplot2: elegant graphics for data analysis. *R. Stat. Soc.* 174, 245–246.

Graham, L. J., Shupe, M. P., Schneble, E. J., Flynt, F. L., Clemenshaw, M. N., Kirkpatrick, A. D., et al. (2014). Current approaches and challenges in monitoring treatment responses in breast cancer. *J. Cancer* 5, 58–68. doi: 10.7150/jca.7047

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

Hao, Y., Yang, X., Zhang, D., Luo, J., and Chen, R. (2017). Long noncoding RNA LINC01186, regulated by TGF-β/SMAD3, inhibits migration and invasion through Epithelial-Mesenchymal-Transition in lung cancer[J]. *Gene* 608, 1–12.

Hu, G., Tang, Q., Sharma, S., Yu, F., Escobar, T. M., Muljo, S. A., et al. (2013). Expression and regulation of intergenic long noncoding RNAs during T cell development and differentiation. *Nat. Immunol.* 14, 1190–1198. doi: 10.1038/ni.2712

Huang, D., Chen, J., Yang, L., Ouyang, Q., Li, J., Lao, L., et al. (2018). NKILA lncRNA promotes tumor immune evasion by sensitizing T cells to activation-induced cell death[J]. *Nat. Immunol.* 19, 1112–1125.

Kabel, A. M. (2017). Tumor markers of breast cancer: new prospectives. *J. Oncol. Sci.* 3, 5–11. doi: 10.1016/j.jons.2017.01.001

Kim, Y. A., Cho, H., Lee, N., Jung, S. Y., Sim, S. H., Park, I. H., et al. (2018). Doxorubicin-induced heart failure in cancer patients: a cohort study based on the Korean National Health Insurance Database. *Cancer Med.* 7, 6084–6092. doi: 10.1002/cam4.1886

Lakshmanan, I., Ponnusamy, M. P., Das, S., Chakraborty, S., Haridas, D., Mukhopadhyay, P., et al. (2012). MUC16 induced rapid G2/M transition via interactions with JAK2 for increased proliferation and anti-apoptosis in breast cancer cells. *Oncogene* 31, 805–817. doi: 10.1038/onc.2011.297

Li, N., Deng, Y., Zhou, L., Tian, T., Yang, S., Wu, Y., et al. (2019). Global burden of breast cancer and attributable risk factors in 195 countries and territories, from 1990 to 2017: results from the Global Burden of Disease Study 2017. *J. Hematol. Oncol.* 12:140. doi: 10.1186/s13045-019-0828-0

Lin, C., and Yang, L. (2018). Long noncoding RNA in cancer: wiring signaling circuitry. *Trends Cell. Biol.* 28, 287–301. doi: 10.1016/j.tcb.2017.11.008

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159. doi: 10.1038/nrg2521

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

Pantelidou, C., Sonzogni, O., De Oliveria Taveira, M., Mehta, A. K., Kothari, A., Wang, D., et al. (2019). PARP inhibitor efficacy depends on CD8 T-cell recruitment via intratumoral STING pathway activation in BRCA-deficient models of triple-negative breast cancer. *Cancer Discov.* 9, 722–737. doi: 10.1158/2159-8290.Cd-18-1218

Pei, X., Wang, X., and Li, H. (2018). LncRNA SNHG1 regulates the differentiation of Treg cells and affects the immune escape of breast cancer via regulating miR-448/IDO. *Int. J. Biol. Macromol.* 118, 24–30. doi: 10.1016/j.ijbiomac.2018.06.033

Schmid, P., Adams, S., and Rugo, H. S. (2018). Atezolizumab and NabPaclitaxel in advanced triple-negative breast cancer. *N. Engl. J. Med.* 379, 2108–2121.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590

Song, X., Fiati Kenston, S. S., Zhao, J., Yang, D., and Gu, Y. (2017). Roles of FoxM1 in cell regulation and breast cancer targeting therapy. *Med. Oncol.* 34:41. doi: 10.1007/s12032-017-0888-3

Stergiou, N., Gaidzik, N., Heimes, A.-S., Dietzen, S., Besenius, P., Jäkel, J., et al. (2019). Reduced breast tumor growth after immunization with a tumor-restricted MUC1 glycopeptide conjugated to tetanus toxoid. *Cancer Immunol. Res.* 7, 113–122. doi: 10.1158/2326-6066.Cir-18-0256

Takasato, Y., Kurashima, Y., Kiuchi, M., Hirahara, K., Murasaki, S., Arai, F., et al. (2020). Orally desensitized mast cells form a regulatory network with Treg cells for the control of food allergy. *Mucosal Immunol.* [Epub ahead of print]. doi: 10.1038/s41385-020-00358-3

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* 16, 385–395. doi: 10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380<3.0.co;2-3

Wang, J., Yang, S., Ji, Q., Li, Q., Zhou, F., Li, Y., et al. (2020). Long non-coding RNA EPIC1 promotes cell proliferation and motility and drug resistance in glioma. *Mol. Ther. Oncolytics* 17, 130–137. doi: 10.1016/j.omto.2020.03.011

Wang, L., Simons, D. L., Lu, X., Tu, T. Y., Solomon, S., Wang, R., et al. (2019). Connecting blood and intratumoral T cell activity in predicting future relapse in breast cancer. *Nat. Immunol.* 20, 1220–1230. doi: 10.1038/s41590-019-0429-7

Wang, Q., Zhang, J., Liu, Y., Zhang, W., Zhou, J., Duan, R., et al. (2016). A novel cell cycle-associated lncRNA, HOXA11-AS, is transcribed from the 5-prime end of the HOXA transcript and is a biomarker of progression in glioma. *Cancer Lett.* 373, 251–259. doi: 10.1016/j.canlet.2016.01.039

Yang, C., Feng, T., Lin, F., Gong, T., Yang, S., Tao, Y., et al. (2020). Long noncoding RNA TANCR promotes γδ T cells activation by regulating TRAIL expression in cis. *Cell Biosci.* 10:15. doi: 10.1186/s13578-020-00383-6

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


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

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 February 2021
doi: 10.3389/fgene.2020.623584

# Comprehensive circRNA Expression Profile and Construction of circRNAs-Related ceRNA Network in a Mouse Model of Autism

*Ji Wang[1,2], Zhongxiu Yang[3], Canming Chen[1], Yang Xu[1], Hongguang Wang[4], Bing Liu[5], Wei Zhang[5]* and Yanan Jiang[5,6]**

[1] Yangzhou Maternal and Child Health Hospital, Yangzhou, China, [2] Harbin Children's Hospital, Harbin, China, [3] Xuzhou Children's Hospital, Xuzhou Medical University, Xuzhou, China, [4] School of Civil Engineering, Northeast Forestry University, Harbin, China, [5] Translational Medicine Research and Cooperation Center of Northern China, Heilongjiang Academy of Medical Sciences, Harbin, China, [6] Department of Pharmacology (State-Province Key Laboratories of Biomedicine-Pharmaceutics of China, Key Laboratory of Cardiovascular Research, Ministry of Education), College of Pharmacy, Harbin Medical University, Harbin, China

Autism is a common disease that seriously affects the quality of life. The role of circular RNAs (circRNAs) in autism remains largely unexplored. We aimed to detect the circRNA expression profile and construct a circRNA-based competing endogenous RNA (ceRNA) network in autism. Valproate acid was used to establish an *in vivo* model of autism in mice. A total of 1,059 differentially expressed circRNAs (477 upregulated and 582 downregulated) in autism group was identified by RNA sequencing. The expression of novel_circ_015779 and novel_circ_035247 were detected by real-time PCR. A ceRNA network based on altered circRNAs was established, with 9,715 nodes and 150,408 edges. Module analysis was conducted followed by GO and KEGG pathway enrichment analysis. The top three modules were all correlated with autism-related pathways involving "TGF-beta signaling pathway," "Notch signaling pathway," "MAPK signaling pathway," "long term depression," "thyroid hormone signaling pathway," etc. The present study reveals a novel circRNA involved mechanisms in the pathogenesis of autism.

**Keywords: autism, circular RNA (circRNA), RNA sequencing (RNA-Seq), ceRNA network, *in silico* analysis**

OPEN ACCESS

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

***Reviewed by:***
*Feng Yang,*
*Shenzhen Second People's Hospital, China*
*Yanjuan Wang,*
*Lianyungang Maternal and Children's Hospital, China*

***Correspondence:**
*Yanan Jiang*
*jiangyanan@hrbmu.edu.cn*
*Wei Zhang*
*zhangwei830530@163.com*

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

***Received:*** *30 October 2020*
***Accepted:*** *23 December 2020*
***Published:*** *16 February 2021*

***Citation:***
*Wang J, Yang Z, Chen C, Xu Y, Wang H, Liu B, Zhang W and Jiang Y (2021) Comprehensive circRNA Expression Profile and Construction of circRNAs-Related ceRNA Network in a Mouse Model of Autism. Front. Genet. 11:623584. doi: 10.3389/fgene.2020.623584*

## INTRODUCTION

Autism spectrum disorder (ASD) is a multifactorial neurodevelopmental disorder diagnosed mainly during early life onset, which is often combined with attention deficit/hyperactivity disorder, mental disorder, and intellectual disability (Vahabzadeh et al., 2018; Valiente-Palleja et al., 2018; Miryounesi et al., 2019). According to the Diagnostic and Statistical Manual of Mental Disorders, 5th ed (DSM-5), it is characterized by impaired social interactions and elevated stereotyped activities, and the global prevalence is about 1% (Lai et al., 2014). Autism prevalence is increasing globally. According to the latest data of Autism and Developmental Disabilities Monitoring (ADDM) network in 2018, ASD prevalence was 1 in 54 (https://www.cdc.gov/ncbddd/autism/data.html). Early screening and diagnosis are very important to improve the outcome of autism patients. Unfortunately, primary health care professionals are usually unaware of the early manifestations of autism, and the gold standard diagnostic tools, autism Diagnostic Interview Revised (ADI-R) and Autism Diagnostic Observation Schedule (ADOS), are time-consuming and expensive. The

pathogenesis of autism is associated with genetic and epigenetic alterations (Olde Loohuis et al., 2015; Turner et al., 2017). However, there is no specific autism therapeutic drug because the pathogenesis mechanism of autism is still not fully clarified.

Circular RNAs (circRNAs) are a kind of non-coding RNAs with circular structure, which are rapidly becoming considered as critical regulators of gene expression networks (Salzman et al., 2013). Furthermore, studies have proved that circRNAs are widely expressed in tissue-and-developmental-stage-specific patterns, and a fraction of them displays conservation across species (Rybak-Wolf et al., 2015). Importantly, a general phenomenon was discovered that circRNAs bind to miRNAs, acting as miRNA sponges, and thereby affect the expression of target genes (Hansen et al., 2013). Several studies have shown that circRNAs are more abundantly expressed in the brain than other tissues in mammals (You et al., 2015; Li et al., 2017). CircRNAs also show dynamic expression during neurogenesis, synaptogenesis, and neuronal differentiation (Izuogu et al., 2018). These findings indicated that circRNAs are likely to play functional roles in neuron development and diseases.

Even though studies revealed that circRNAs are highly abundant in the brain, the expression profile, function, and mechanism of most circRNAs in autism are still largely unelucidated. Chen et al. (2020) found a series of autism-associated circRNAs in autism cortex samples. Our study aimed to determine the circRNA profile in brain tissues from valproic acid-induced mouse autism model and analyze their function and potential mechanism.

# MATERIALS AND METHODS

## Establishment of A Mouse Model of Autism

C57BL/6 mice were bought from Liaoning Changsheng Biotechnology Co., Ltd. (Liaoning, China). Exposure to some neurotoxic drugs may induce fetal nervous system development disorders. Valproate acid (VPA) is a widely used drug to induce autism (Baronio et al., 2018; Eissa et al., 2018). The animal model of autism was established as previously described (Zheng et al., 2019). Briefly, female adult mice were mated to males overnight. Pregnant mice were injected with valproic acid (Sigma, Ronkonkoma, NY, USA) 500 mg/kg at embryonic day 12.5 (E12.5). Mice in control group received a considerable amount of normal saline. All pregnant mice were allowed to give birth naturally, and the 1st day of birth was recorded as postnatal day 1 (P1). Pups were weaned at the 23rd day (p23) after birth. The animal protocol was approved by the Harbin Children's Hospital Animal Care and Use Committee (JJ2017ZR0484).

## Repetitive Self-Grooming Behavior Measurement

On the 32nd day after birth, mice were placed in open field box for 10 min to adapt to the environment, then start timing for 10 min to observe mice behavior (Onaolapo et al., 2017). The number of body cleaning with paws and face-washing actions was calculated.

## Social Interaction Test

Social interaction test was used to assess active interaction time in a test mouse with a novel mouse. (1) Adaptation stage: Gently put the tested mice into the device (material: acrylic glass, 40 × 40 × 30 cm, covered with 2–3 cm thick bedding), and let them move freely inside 10 min. (2) Test stage: Take out the test mice that have just adapted, and put them in the test mice and interactive mice (the same strains of mice that have the same sex and age as the test mice that have not been raised in the same cage). The number of social behaviors (physical contact and following peers) and non-social behaviors (investigation) of the mice were recorded within 10 min.

## Real-Time PCR Analysis

TRIzol reagent (Invitrogen, CA, USA) was used to extract total RNA. A reverse transcriptase kit (Roche, Mannheim, Germany) was used to synthesize first-strand cDNA. Reverse transcription was performed using HiScript® II Q RT SuperMixfor qPCR (Vazyme Biotech Co., Ltd., Nanjing, China). Real-time PCR analysis was performed on an ABI step one plus system (Applied Biosystems, CA, USA). Primer sequences were as follows: novel_circ_000430, 5′-ACCCGTCTTCAGTCTCCGT-3′ (forward), 5′-AATATCACCCACACCCTCAGC-3′ (reverse); novel_circ_015779, 5′-CTCTGCCTGGTGCTGGTATTG-3′ (forward), 5′-ATGTAACTCTCTCCCTCCCCTG-3′ (reverse); novel_circ_063340, 5′-GCTTACCGTGGAGATGTTTGAC-3′ (forward), 5′-CGCCTTCTCCAACACCTCA-3′ (reverse); β-actin, 5′-ACCCATTCTCTGTCTCGCAC-3′ (forward), 5′-ATCGTCACCCCCAAAACCTG-3′ (reverse). β-actin was used as an internal control. Target gene expression was analyzed using the $2^{-\Delta\Delta CT}$ method.

## RNA Sequencing

RNA sequencing was conducted by Gene Denovo Biotechnology Co. (Guangzhou, China). The circRNA were identified using find_circ (Memczak et al., 2013). The edge R package (http://www.rproject.org/) was used to identify differentially expressed circRNAs. CircRNAs with $P < 0.05$ and $|\log2FoldChange| > 1$ in a comparison between control and autism groups was identified as differentially expressed.

## CircRNA-miRNA-mRNA Network Construction

The miRNA binding sites of annotated circRNAs in circBase was predicted by StarBase (v2.0) (Li et al., 2014). The miRNA binding sites of novel circRNAs were predicted by Mireap, Miranda (v3.3a) (Betel et al., 2010), and TargetScan (v7.0) (Agarwal et al., 2015). Subsequently, miRTarBase (v6.1) (Hsu et al., 2011) was used to predict mRNAs that interact with circRNAs through miRNAs. The established competing endogenous RNA (ceRNA) network was visualized by Cytoscape (Version 3.7.2) (Kohl et al., 2011).

## Gene Ontology and Pathway Enrichment Analysis

Significantly enriched Gene Ontology (GO) terms in source genes comparing to the genome background were

defined by hypergeometric test (Ashburner et al., 2000). Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used to perform enrichment analysis (Kanehisa et al., 2008).

## Statistical Analysis

Data are expressed as mean ± SEM and compared using Student's *t*-test. A two-tailed $P < 0.05$ was required for significance.

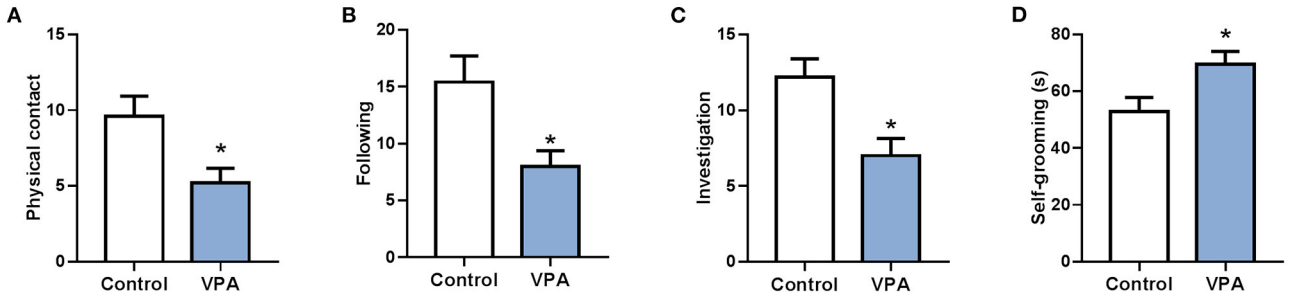


**FIGURE 1 |** A mouse model of autism was successfully established. **(A)** The social interaction test (physical contact). **(B)** The social interaction test (following). **(C)** The social interaction test (investigation). **(D)** Repetitive self-grooming behavior. *$P < 0.05$ vs. Control; $n = 4$–5.

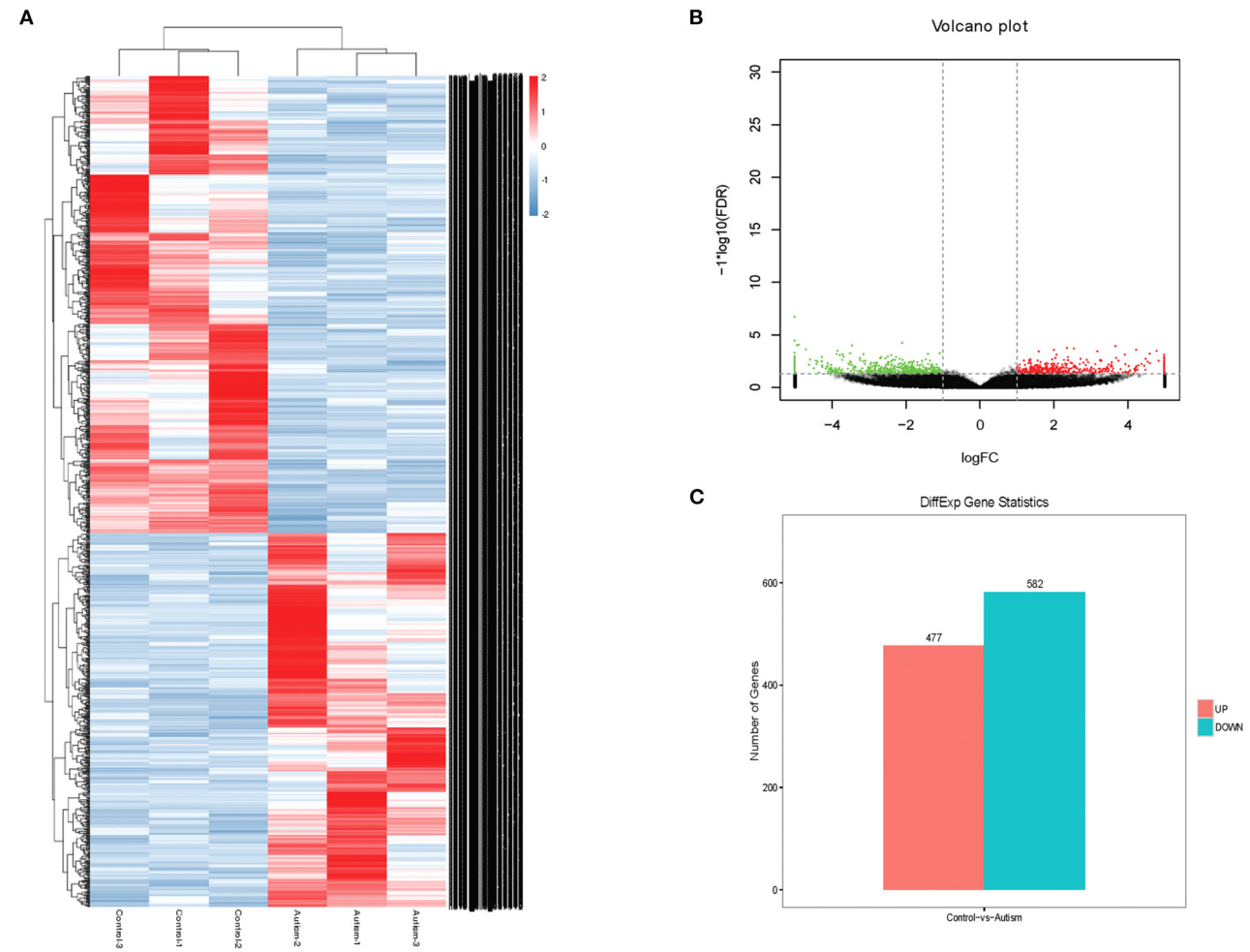


**FIGURE 2 |** Differentially expressed circRNAs in brain from autism mice. **(A)** Heat map. **(B)** Volcano plots. **(C)** The number of differentially expressed circRNAs in a mouse model of autism.

# RESULTS

## A Mouse Model of Autism Was Successfully Established

To analyze different expression of circRNA in the brain of autistic mice, we established a mouse model of autism and verified the model through behavioral testing. The animal model of autism was induced as previously described (Zheng et al., 2019). On the 28th day after birth, their behavioral ontogeny was evaluated using the social interaction test, repetitive self-grooming behavior. Results showed that the offspring mice in the model group exhibited autism-like behavioral abnormalities (**Figure 1**).

**TABLE 1 |** Biological information regarding the top 10 upregulated and downregulated circRNAs in autism mouse model.

| | circRNA | log2(Fold change) | *p*-value |
|---|---|---|---|
| Upregulated | novel_circ_015779 | 14.07865 | 1.88E-03 |
| | novel_circ_063340 | 13.83892 | 1.82E-03 |
| | novel_circ_000430 | 13.75961 | 1.23E-03 |
| | novel_circ_052619 | 13.754 | 3.88E-03 |
| | novel_circ_011355 | 13.72869 | 6.55E-03 |
| | novel_circ_002576 | 13.67244 | 7.79E-04 |
| | novel_circ_013547 | 13.65162 | 9.93E-03 |
| | novel_circ_058403 | 13.6008 | 1.33E-03 |
| | novel_circ_028696 | 13.58711 | 8.98E-03 |
| | novel_circ_018711 | 13.57057 | 1.20E-02 |
| Downregulated | novel_circ_014536 | −14.9586 | 1.90E-07 |
| | novel_circ_066322 | −14.168 | 3.44E-05 |
| | novel_circ_022010 | −13.7173 | 5.14E-04 |
| | novel_circ_056205 | −13.5461 | 1.36E-03 |
| | novel_circ_001586 | −13.5253 | 8.62E-03 |
| | novel_circ_065910 | −13.5045 | 1.93E-03 |
| | novel_circ_000370 | −13.4837 | 1.10E-03 |
| | novel_circ_018104 | −13.479 | 1.09E-03 |
| | novel_circ_012200 | −13.436 | 2.96E-03 |
| | novel_circ_061645 | −13.392 | 1.61E-02 |

## Expression Profile of circRNAs in A Mouse Model of Autism

The expression profile of circRNAs in brain tissues from autism mouse model and corresponding controls was evaluated using RNA sequencing. The hierarchical cluster analysis of circRNA is shown in **Figure 2A**. The volcano plots of circRNAs are shown in **Figure 2B**. A total of 1,059 altered circRNAs were identified in the autism group, with 477 upregulated and 582 downregulated (GSE163904, **Figure 2C**, **Table 1**, and **Supplementary Material**).

## CircRNA Expression Verified by Real-Time PCR

The expressions of novel_circ_015779 and novel_circ_035247 were detected by real-time PCR assay. The expression levels of novel_circ_015779 and novel_circ_035247 were upregulated in the autism group (**Figure 3**). These results were in accordance with that in RNA sequencing.

## The ceRNA Network Construction

The interactions between circRNAs/mRNA and miRNAs were predicted using StarBase (v2.0), Miranda (v3.3a), TargetScan (v7.0), and miRTarBase (v6.1). Then, a ceRNA network was constructed using differentially expressed circRNAs and bioinformatic predication results. The established circRNAs-miRNA-mRNA ceRNA network contains 9,715 nodes (including 1,059 circRNAs, 6,730 mRNA, and 1,926 miRNA) and 150,408 edges. A top 5 miRNA-based circRNA-miRNA-mRNA network is shown in **Figure 4**. The top 10 hub nodes are shown in **Table 2**. GO analysis showed that the ceRNA network was mainly associated with several biological processes, including "cellular process," "biological regulation," "regulation of biological process," and "metabiological process." Meanwhile, KEGG pathway analysis showed the ceRNA network was mainly associated with "axon guidance," "MAPK signaling," "Hippo signaling pathway," and "ErbB signaling pathway" (**Figure 5**).

## Module Analysis of the ceRNA Network

Modules were screened using plug-in MCODE. GO and KEGG pathway enrichments were performed in the top

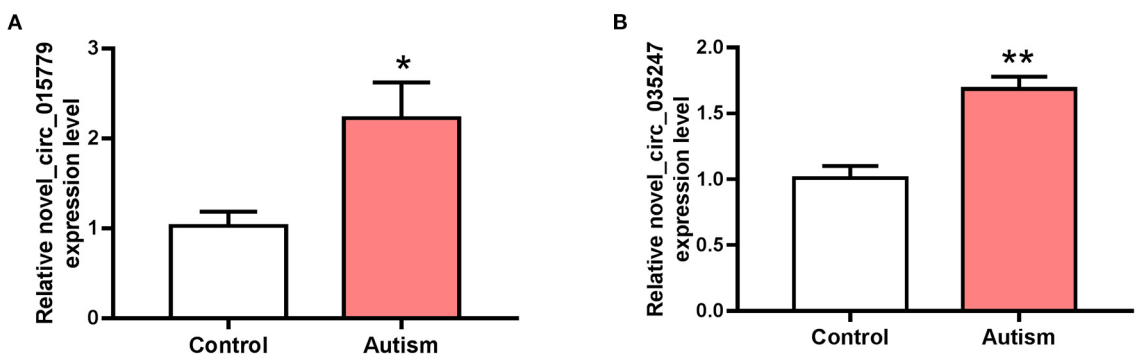


**FIGURE 3 |** The expression of novel_circ_015779 and novel_circ_035247 in control and autism groups. **(A)** The expression of novel_circ_015779. **(B)** The expression of novel_circ_035247. $^{*}P < 0.05$, $^{**}P < 0.01$ vs. Control, $n = 3$.

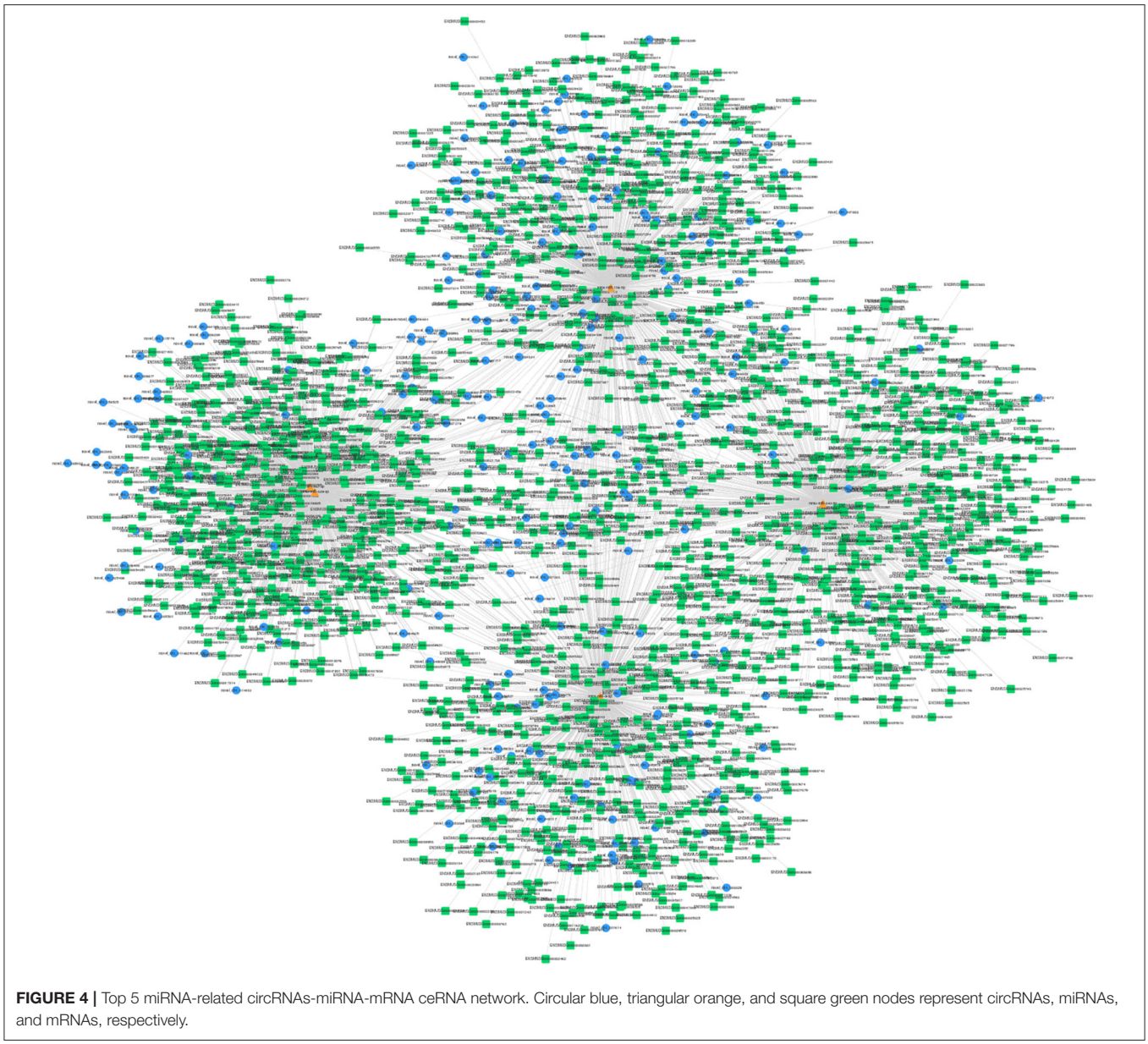

**FIGURE 4 |** Top 5 miRNA-related circRNAs-miRNA-mRNA ceRNA network. Circular blue, triangular orange, and square green nodes represent circRNAs, miRNAs, and mRNAs, respectively.

three modules, respectively. Module 1 consists of 393 nodes (68 circRNAs, 239 miRNAs, and 86 mRNAs) and 1,508 edges, which are mainly associated with "HIF-1 signaling pathway," "arginine and proline metabolism," "TGF-beta signaling pathway," "IL-17 signaling pathway," etc. (**Figure 6**). Module 2 consists of 242 nodes (67circRNAs, 73 miRNAs, and 102 mRNAs) and 891 edges, which are mainly related to "renin secretion," "cGMP-PKG signaling pathway," "Notch signaling pathway," "long term depression," etc. (**Figure 7**). Module 3 consists of 289 nodes (75 circRNAs, 95 miRNAs, and 119 mRNAs) and 1,011 edges, which are mainly associated with "IL-17 signaling pathway," "MAPK signaling pathway," "C-type lectin receptor signaling pathway," "thyroid hormone signaling pathway," etc. (**Figure 8**).

## DISCUSSION

The incidence of autism is still increasing, which seriously affects the quality of human life. Despite the research progress that has been made in recent years, the pathogenesis mechanisms of autism are still not fully clarified. The development of bioinformatics facilitated the investigation of disease mechanisms and therapeutic strategies (Peng et al., 2017, 2018; Zhou et al., 2019). Our study comprehensively elucidated circRNA expression profile in a mouse model of autism and constructed a circRNA-associated ceRNA network.

CircRNAs are a special kind of non-coding RNAs with circular structure. CircRNAs could exert biological function through a ceRNA mechanism (circRNA-miRNA-mRNA), circRNA–protein interaction, or regulate translation (Du et al., 2017; Sun et al.,

2019; Yi et al., 2019). This kind of non-coding RNAs is involved in various diseases, including cancers, cardiovascular diseases, and neuronal diseases (Chen et al., 2017; Sekar et al., 2018; Yang et al., 2019). CircRNAs are shown to be associated with schizophrenia and cognitive dysfunction (Zhang et al., 2017; Mahmoudi et al., 2019). However, the expression and role of circRNAs in autism is still poorly understood. Therefore, we detected the circRNA expression profile using RNA sequencing in a mouse model of autism.

RNA sequencing identified a total of 1,059 differentially expressed circRNAs in the autism group. The expression level of novel_circ_015779 and novel_circ_035247 was detected by real-time PCR assay. The RNA sequencing and real-time PCR assay achieved similar results.

CircRNAs could act as a ceRNA and thus exert biological functions. Using bioinformatic methods, we predicted the interaction between circRNAs and miRNA and constructed an autism-related ceRNA network. The ceRNA network contains 150,408 edges and 9,715 nodes, including 1,059 circRNAs, 1,926 miRNA, and 6,730 mRNA.

We then performed GO and KEGG analysis. The autism-related ceRNA network was signaling pathways including "axon guidance," "MAPK signaling," "Hippo signaling pathway," and "ErbB signaling pathway." Among these signaling pathways, the role of "axon guidance" and "MAPK signaling" in autism has been proved (Suda et al., 2011; Rosina et al., 2019). It has been reported that aberrant axon-guidance protein expression was observed in the brains of people with autism, including the decreased expression of PLXNA4 and ROBO2 (Suda et al., 2011). And the disruption of the MAPK pathways correlates with severity in idiopathic autism (Rosina et al., 2019).

The hub nodes, take miR-142-3p, miR-142a-5p, and miR-142b for example, may play an important role in autism. Mor et al. (2015) found that miR-142a-5p and miR-142a-3p were upregulated in brain tissues of autism patients compared with that in controls. It has been reported that the upregulation of these two miRNAs could induce apoptosis (Lu et al., 2015; Kim et al., 2018). Apoptosis was also related to autism (Dong et al., 2018). Therefore, miR-142 cluster and its related circRNAs may be involved in the pathogenesis of autism, which needs further investigation. miR-9-5p was also a hub nod, which targets Dlg4 (encode PSD-95). The deletion or variation of Dlg4 was associated with autism (Feyder et al., 2010). In the ceRNA network, novel_circ_000370, novel_circ_050499, novel_circ_021126, etc. may regulate Dlg4 through miR-9-5p.

Subsequently, network modules analysis was conducted. The top three modules were all correlated with autism-related pathways including "TGF-beta signaling pathway," "Notch signaling pathway," "MAPK signaling pathway," "long-term depression," etc. Take "TGF-beta signaling pathway" as an example; TGF-β1 was upregulated in the brain of autistic patients (Vargas et al., 2005). Early and adult hippocampal TGF-β1 overexpression led to a series of autism-related behaviors (Depino et al., 2011). The circRNAs involved in these signaling pathways may play a regulatory role in autism.

Circulating miRNAs were aberrantly expressed in autism individuals compared with that in controls (Mundalil Vasu et al., 2014; Hicks et al., 2016; Jyonouchi et al., 2019). In the serum of individuals with autism, 13 miRNAs were

**TABLE 2 |** Hub nodes with top degrees in the ceRNA network.

| Rank | circRNA | miRNA | mRNA |
|---|---|---|---|
| 1 | novel_circ_050499 | mmu-miR-15a-5p | Tbc1d24 |
| 2 | novel_circ_021126 | mmu-miR-340-5p | Gabpb2 |
| 3 | novel_circ_002396 | mmu-miR-362-3p | Ybey |
| 4 | novel_circ_054171 | mmu-miR-9-5p | Chic1 |
| 5 | novel_circ_002378 | mmu-miR-329-3p | Hecw1 |
| 6 | novel_circ_025446 | mmu-miR-7b-5p | Cd28 |
| 7 | novel_circ_028364 | mmu-miR-181a-5p | Asxl2 |
| 8 | novel_circ_036195 | mmu-miR-466l-5p | Ppm1k |
| 9 | novel_circ_041509 | mmu-miR-466i-5p | Kcnk6 |
| 10 | novel_circ_000669 | mmu-miR-466k | Car5b |

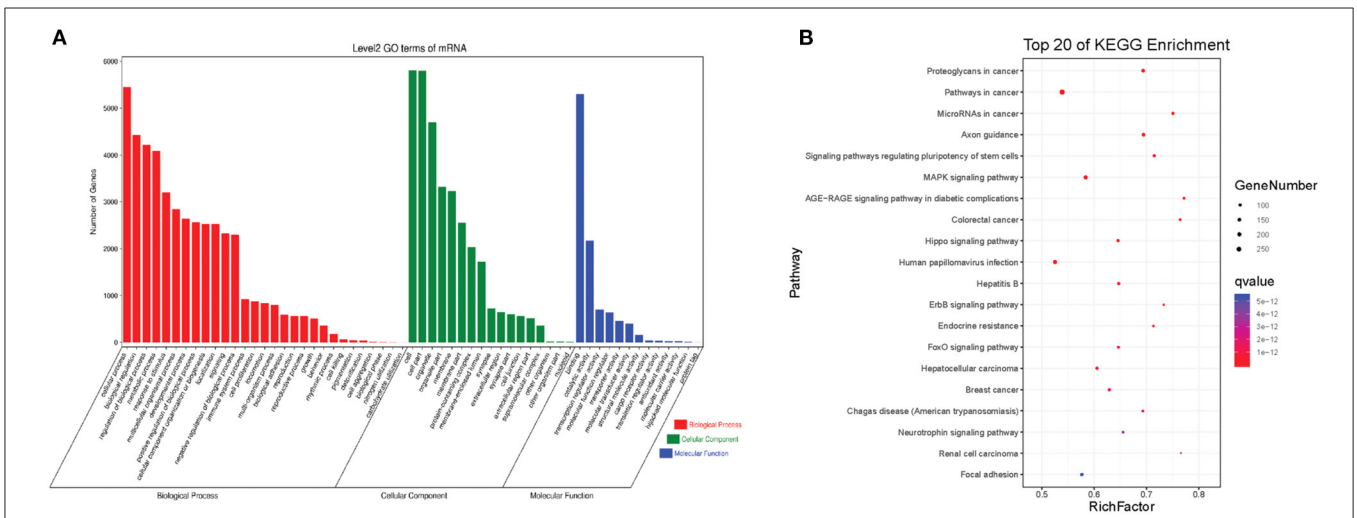


**FIGURE 5 |** The GO and KEGG pathway enrichment of autism-related circRNAs-miRNA-mRNA ceRNA network. **(A)** GO enrichment. **(B)** KEGG pathway enrichment. The color bar represents the *q*-value.

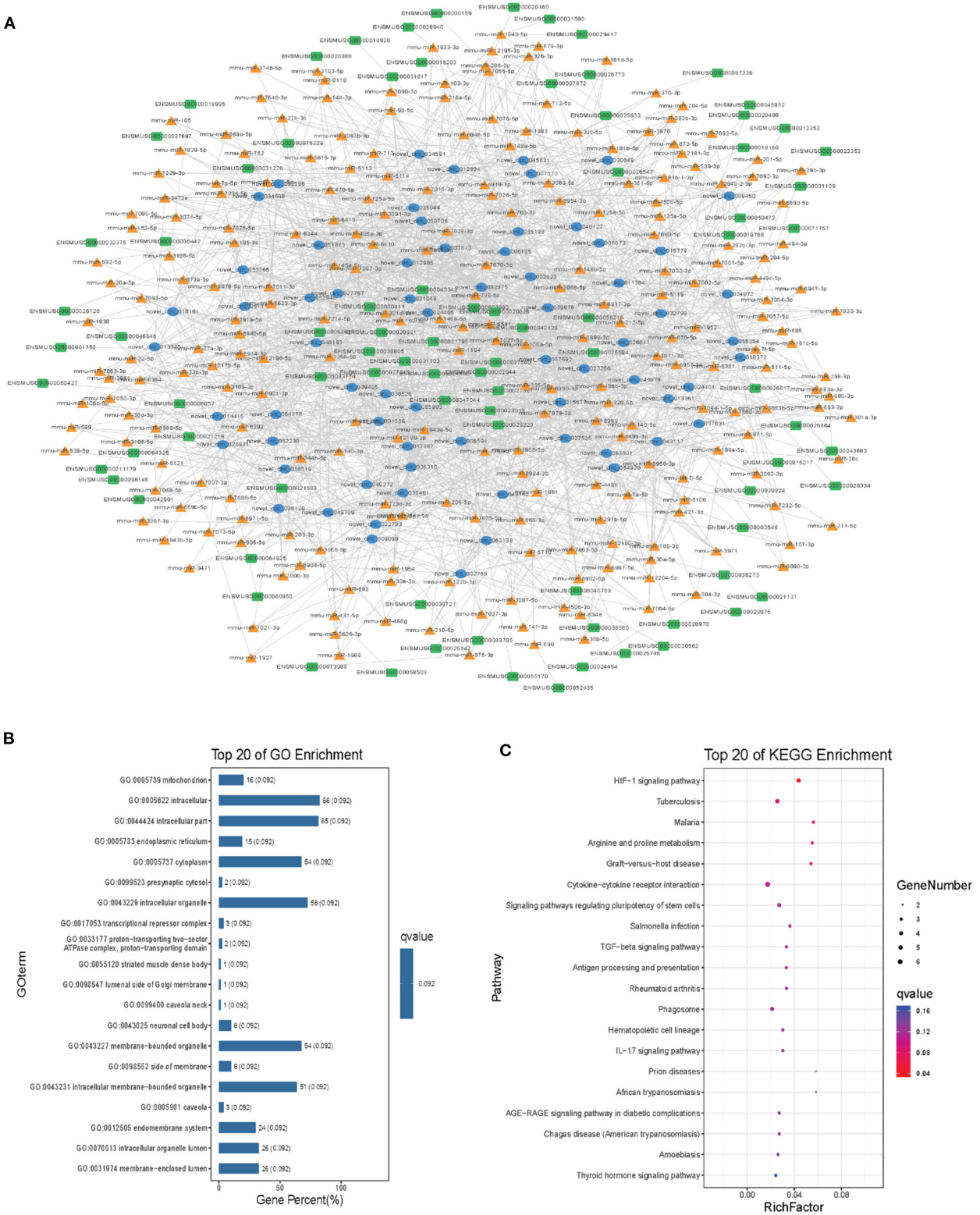


**FIGURE 6 |** Module 1 from the autism-related ceRNA network. **(A)** Module 1 network. **(B)** GO enrichment. **(C)** KEGG pathway enrichment. Circular blue, triangular orange, and square green nodes represent circRNAs, miRNAs, and mRNAs, respectively.

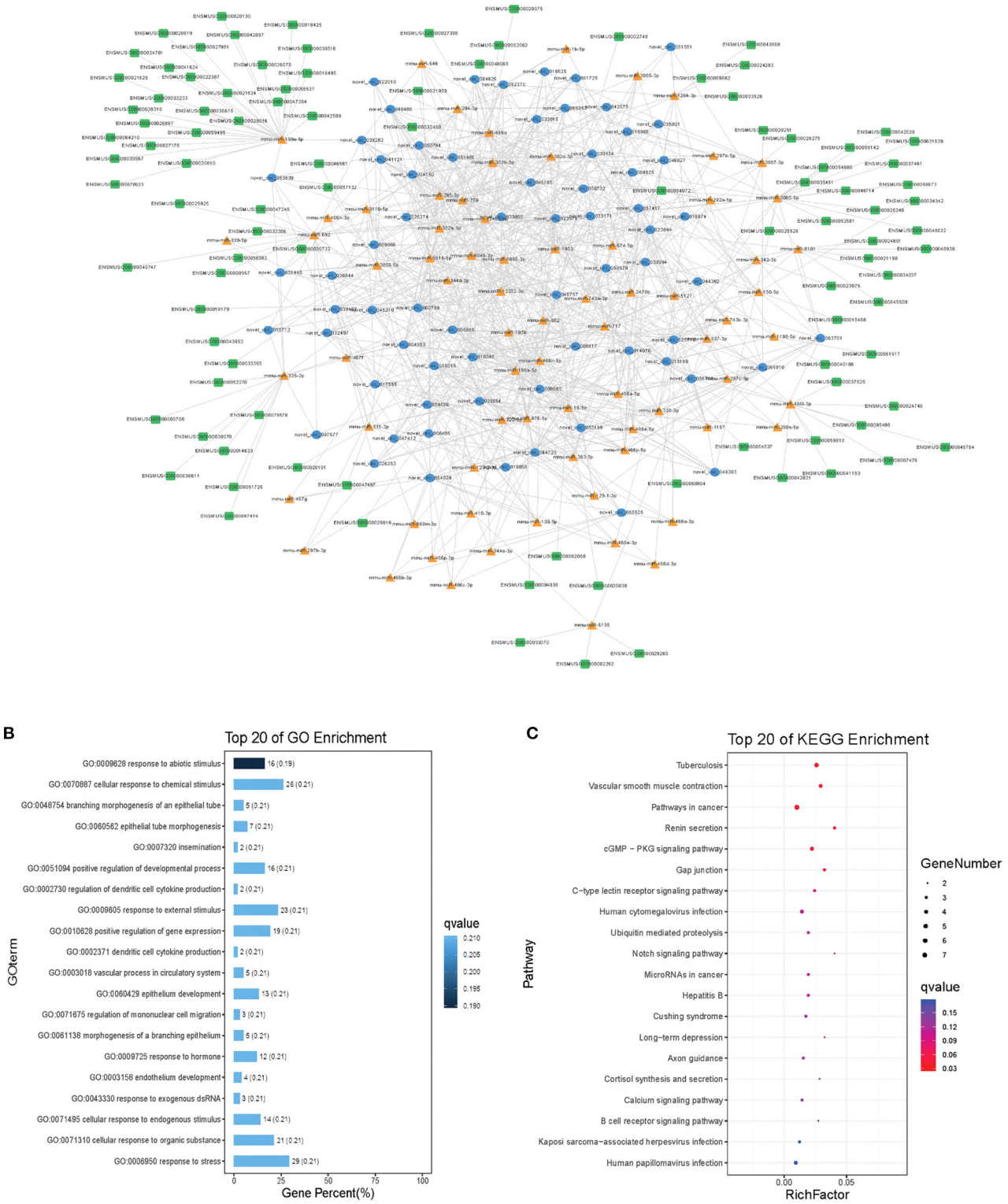


**FIGURE 7 |** Module 2 from the autism-related ceRNA network. **(A)** Module 2 network. **(B)** GO enrichment. **(C)** KEGG pathway enrichment. Circular blue, triangular orange, and square green nodes represent circRNAs, miRNAs, and mRNAs, respectively.

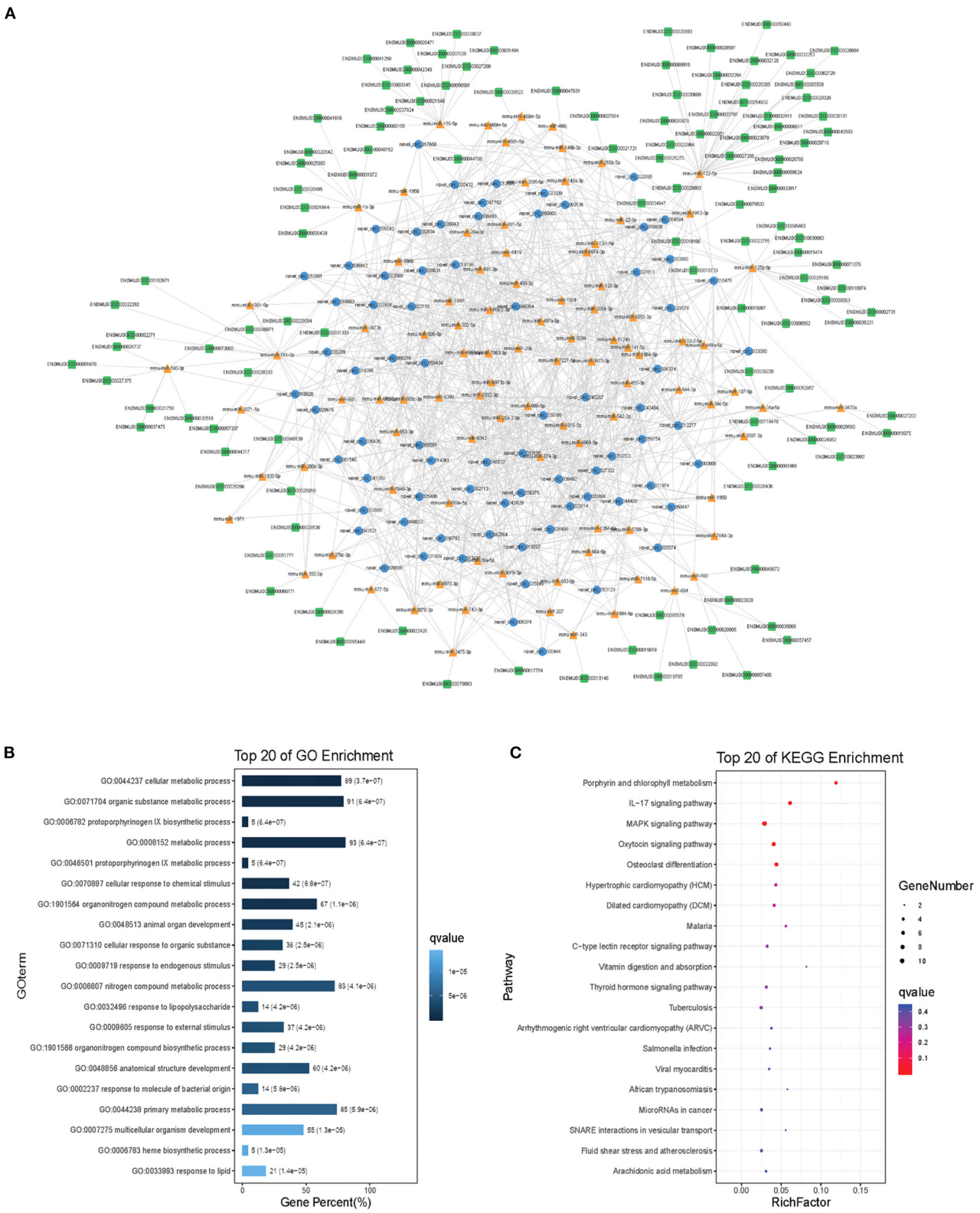


**FIGURE 8 |** Module 3 from the autism-related ceRNA network. **(A)** Module 3 network. **(B)** GO enrichment. **(C)** KEGG pathway enrichment. Circular blue, triangular orange, and square green nodes represent circRNAs, miRNAs, and mRNAs, respectively.

differentially expressed, some of which showed diagnostic potential for patients with autism (Mundalil Vasu et al., 2014). Subsequent research achieved similar results, that serum miRNAs may be promising biomarkers for autism (Kichukova et al., 2017; Jyonouchi et al., 2019). Besides, 14 miRNAs were identified differentially expressed in salivary samples from autism patients, and most of them significantly correlate with Vineland neurodevelopmental scores, including miR-27a, miR-23a, and miR-628-5p (Hicks et al., 2016). However, the function of these miRNAs in autism remains largely uninvestigated. In our ceRNA network, we identified a series of potential signaling pathways involved in the pathogenesis of autism, which may be helpful to clarify the role of miRNAs in autism. Besides, circRNAs also exist in the circulation system, which may also serve as biomarkers for autism. The expression of circRNAs in serum from autism patients and animal models still needs further investigation.

In conclusion, our study provides the expression profile of circRNAs in a mouse model of autism, which promotes our understanding in the pathogenesis mechanism of autism.

## DATA AVAILABILITY STATEMENT

The RNA sequencing results for this study could be found in the GEO repository with the accession number of GSE163904.

## ETHICS STATEMENT

The animal study was reviewed and approved by Harbin children's hospital Animal Care and Use Committee [JJ2017ZR0484].

## AUTHOR CONTRIBUTIONS

JW, YJ, and WZ designed the study. HW and BL performed the animal experiment. ZY, CC, and YX analyzed the data. JW and YJ wrote the manuscript. All authors reviewed the manuscript.

## FUNDING

This project was supported by the National Natural Science Foundation of China (81803524 and 51708092); the China Postdoctoral Science Foundation (2018M641878); the Heilongjiang Postdoctoral Foundation (LBH-Z18168); the Key Laboratory of Myocardial ischemia, Harbin Medical University, Ministry of Education, Heilongjiang Province, China (KF201611); Natural Science Foundation of Heilongjiang Province of China (H2017004); research project of maternal and child health in Jiangsu Province (F201920); and research project of Jiangsu Province Health Commission (H2019021).

## ACKNOWLEDGMENTS

The authors want to thank all the participants in the research.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.623584/full#supplementary-material

## REFERENCES

Agarwal, V., Bell, G. W., Nam, J. W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 4:e05005. doi: 10.7554/eLife.05005.028

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

Baronio, D., Bauer-Negrini, G., Castro, K., Della-Flora Nunes, G., Riesgo, R., Mendes-da-Cruz, D. A., et al. (2018). Reduced CD4 T lymphocytes in lymph nodes of the mouse model of autism induced by valproic acid. *Neuroimmunomodulation* 25, 280–284. doi: 10.1159/000491395

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* 11:R90. doi: 10.1186/gb-2010-11-8-r90

Chen, J., Li, Y., Zheng, Q., Bao, C., He, J., Chen, B., et al. (2017). Circular RNA profile identifies circPVT1 as a proliferative factor and prognostic marker in gastric cancer. *Cancer Lett.* 388, 208–219. doi: 10.1016/j.canlet.2016.12.006

Chen, Y. J., Chen, C. Y., Mai, T. L., Chuang, C. F., Chen, Y. C., Gupta, S. K., et al. (2020). Genome-wide, integrative analysis of circular RNA dysregulation and the corresponding circular RNA-microRNA-mRNA regulatory axes in autism. *Genome Res.* 30, 375–391. doi: 10.1101/gr.255463.119

Depino, A. M., Lucchina, L., and Pitossi, F. (2011). Early and adult hippocampal TGF-beta1 overexpression have opposite effects on behavior. *Brain Behav. Immun.* 25, 1582–1591. doi: 10.1016/j.bbi.2011.05.007

Dong, D., Zielke, H. R., Yeh, D., and Yang, P. (2018). Cellular stress and apoptosis contribute to the pathogenesis of autism spectrum disorder. *Autism Res.* 11, 1076–1090. doi: 10.1002/aur.1966

Du, W. W., Zhang, C., Yang, W., Yong, T., Awan, F. M., and Yang, B. B. (2017). Identifying and characterizing circRNA-protein interaction. *Theranostics* 7, 4183–4191. doi: 10.7150/thno.21299

Eissa, N., Jayaprakash, P., Azimullah, S., Ojha, S. K., Al-Houqani, M., Jalal, F. Y., et al. (2018). The histamine H3R antagonist DL77 attenuates autistic behaviors in a prenatal valproic acid-induced mouse model of autism. *Sci. Rep.* 8:13077. doi: 10.1038/s41598-018-31385-7

Feyder, M., Karlsson, R. M., Mathur, P., Lyman, M., Bock, R., Momenan, R., et al. (2010). Association of mouse Dlg4 (PSD-95) gene deletion and human DLG4 gene variation with phenotypes relevant to autism spectrum disorders and Williams' syndrome. *Am. J. Psychiatry* 167, 1508–1517. doi: 10.1176/appi.ajp.2010.10040484

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

Hicks, S. D., Ignacio, C., Gentile, K., and Middleton, F. A. (2016). Salivary miRNA profiles identify children with autism spectrum disorder, correlate with adaptive behavior, and implicate ASD candidate genes involved in neurodevelopment. *BMC Pediatr.* 16:52. doi: 10.1186/s12887-016-0586-x

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., et al. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 39, D163–D169. doi: 10.1093/nar/gkq1107

Izuogu, O. G., Alhasan, A. A., Mellough, C., Collin, J., Gallon, R., Hyslop, J., et al. (2018). Analysis of human ES cell differentiation establishes that the dominant isoforms of the lncRNAs RMST and FIRRE are circular. *BMC Genomics* 19:276. doi: 10.1186/s12864-018-4660-7

Jyonouchi, H., Geng, L., Toruner, G. A., Rose, S., Bennuri, S. C., and Frye, R. E. (2019). Serum microRNAs in ASD: association with monocyte

cytokine profiles and mitochondrial respiration. *Front. Psychiatry* 10:614. doi: 10.3389/fpsyt.2019.00614

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–D484. doi: 10.1093/nar/gkm882

Kichukova, T. M., Popov, N. T., Ivanov, I. S., and Vachev, T. I. (2017). Profiling of circulating serum MicroRNAs in children with autism spectrum disorder using stem-loop qRT-PCR assay. *Folia Med.* 59, 43–52. doi: 10.1515/folmed-2017-0009

Kim, J. O., Park, J. H., Kim, T., Hong, S. E., Lee, J. Y., Nho, K. J., et al. (2018). A novel system-level approach using RNA-sequencing data identifies miR-30-5p and miR-142a-5p as key regulators of apoptosis in myocardial infarction. *Sci. Rep.* 8:14638. doi: 10.1038/s41598-018-33020-x

Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. *Methods Mol. Biol.* 696, 291–303. doi: 10.1007/978-1-60761-987-1_18

Lai, M. C., Lombardo, M. V., and Baron-Cohen, S. (2014). Autism. *Lancet* 383, 896–910. doi: 10.1016/S0140-6736(13)61539-1

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Li, L., Zheng, Y. C., Kayani, M. U. R., Xu, W., Wang, G. Q., Sun, P., et al. (2017). Comprehensive analysis of circRNA expression profiles in humans by RAISE. *Int. J. Oncol.* 51, 1625–1638. doi: 10.3892/ijo.2017.4162

Lu, X., Wei, Y., and Liu, F. (2015). Direct regulation of p53 by miR-142a-3p mediates the survival of hematopoietic stem and progenitor cells in zebrafish. *Cell Discov.* 1:15027. doi: 10.1038/celldisc.2015.27

Mahmoudi, E., Fitzsimmons, C., Geaghan, M. P., Shannon Weickert, C., Atkins, J. R., Wang, X., et al. (2019). Circular RNA biogenesis is decreased in postmortem cortical gray matter in schizophrenia and may alter the bioavailability of associated miRNA. *Neuropsychopharmacology* 44, 1043–1054. doi: 10.1038/s41386-019-0348-1

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

Miryounesi, M., Bahari, S., Salehpour, S., Alipour, N., and Ghafouri-Fard, S. (2019). ELMO domain containing 1 (ELMOD1) gene mutation is associated with mental retardation and autism spectrum disorder. *J. Mol. Neurosci.* 69, 312–315. doi: 10.1007/s12031-019-01359-z

Mor, M., Nardone, S., Sams, D. S., and Elliott, E. (2015). Hypomethylation of miR-142 promoter and upregulation of microRNAs that target the oxytocin receptor gene in the autism prefrontal cortex. *Mol. Autism* 6:46. doi: 10.1186/s13229-015-0040-1

Mundalil Vasu, M., Anitha, A., Thanseem, I., Suzuki, K., Yamada, K., Takahashi, T., et al. (2014). Serum microRNA profiles in children with autism. *Mol. Autism* 5:40. doi: 10.1186/2040-2392-5-40

Olde Loohuis, N. F., Kole, K., Glennon, J. C., Karel, P., G., Van der Borg, Van Gemert, Y., et al. (2015). Elevated microRNA-181c and microRNA-30d levels in the enlarged amygdala of the valproic acid rat model of autism. *Neurobiol. Dis.* 80, 42–53. doi: 10.1016/j.nbd.2015.05.006

Onaolapo, O. J., Paul, T. B., and Onaolapo, A. Y. (2017). Comparative effects of sertraline, haloperidol or olanzapine treatments on ketamine-induced changes in mouse behaviours. *Metab. Brain Dis.* 32, 1475–1489. doi: 10.1007/s11011-017-0031-3

Peng, L., Zhu, W., Liao, B., Duan, Y., Chen, M., Chen, Y., et al. (2017). Screening drug-target interactions with positive-unlabeled learning. *Sci. Rep.* 7:8087. doi: 10.1038/s41598-017-08079-7

Peng, L. H., Yin, J., Zhou, L., Liu, M. X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Rosina, E., Battan, B., Siracusano, M., Di Criscio, L., Hollis, F., Pacini, L., et al. (2019). Disruption of mTOR and MAPK pathways correlates with severity in idiopathic autism. *Transl. Psychiatry* 9:50. doi: 10.1038/s41398-018-0335-z

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

Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L., and Brown, P. O. (2013). Cell-type specific features of circular RNA expression. *PLoS Genet.* 9:e1003777. doi: 10.1371/journal.pgen.1003777

Sekar, S., Cuyugan, L., Adkins, J., Geiger, P., and Liang, W. S. (2018). Circular RNA expression and regulatory network prediction in posterior cingulate astrocytes in elderly subjects. *BMC Genomics* 19:340. doi: 10.1186/s12864-018-4670-5

Suda, S., Iwata, K., Shimmura, C., Kameno, Y., Anitha, A., Thanseem, I., et al. (2011). Decreased expression of axon-guidance receptors in the anterior cingulate cortex in autism. *Mol. Autism* 2:14. doi: 10.1186/2040-2392-2-14

Sun, Y. M., Wang, W. T., Zeng, Z. C., Chen, T. Q., Han, C., Pan, Q., et al. (2019). circMYBL2, a circRNA from MYBL2, regulates FLT3 translation by recruiting PTBP1 to promote FLT3-ITD AML progression. *Blood* 134, 1533–1546. doi: 10.1182/blood.2019000802

Turner, T. N., Coe, B. P., Dickel, D. E., Hoekzema, K., Nelson, B. J., Zody, M. C., et al. (2017). Genomic patterns of *de novo* mutation in simplex autism. *Cell* 171, 710–722 e12. doi: 10.1016/j.cell.2017.08.047

Vahabzadeh, A., Keshav, N. U., Salisbury, J. P., and Sahin, N. T. (2018). Improvement of attention-deficit/hyperactivity disorder symptoms in school-aged children, adolescents, and young adults with autism via a digital smartglasses-based socioemotional coaching aid: short-term, uncontrolled pilot study. *JMIR Ment. Health* 5:e25. doi: 10.2196/mental.9631

Valiente-Palleja, A., Torrell, H., Muntane, G., Cortes, M. J., Martinez-Leal, R., Abasolo, N., et al. (2018). Genetic and clinical evidence of mitochondrial dysfunction in autism spectrum disorder and intellectual disability. *Hum. Mol. Genet* 27, 891–900. doi: 10.1093/hmg/ddy009

Vargas, D. L., Nascimbene, C., Krishnan, C., Zimmerman, A. W., and Pardo, C. A. (2005). Neuroglial activation and neuroinflammation in the brain of patients with autism. *Ann. Neurol.* 57, 67–81. doi: 10.1002/ana.20315

Yang, F., Li, A., Qin, Y., Che, H., Wang, Y., Lv, J., et al. (2019). A novel circular RNA mediates pyroptosis of diabetic cardiomyopathy by functioning as a competing endogenous RNA. *Mol. Ther. Nucleic Acids* 17, 636–643. doi: 10.1016/j.omtn.2019.06.026

Yi, Y., Liu, Y., Wu, W., Wu, K., and Zhang, W. (2019). Reconstruction and analysis of circRNAmiRNAmRNA network in the pathology of cervical cancer. *Oncol. Rep.* 41, 2209–2225. doi: 10.3892/or.2019.7028

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, S., Zhu, D., Li, H., Li, H., Feng, C., and Zhang, W. (2017). Characterization of circRNA-associated-ceRNA networks in a senescence-accelerated mouse prone 8 brain. *Mol. Ther.* 25, 2053–2061. doi: 10.1016/j.ymthe.2017.06.009

Zheng, W., Hu, Y., Chen, D., Li, Y., and Wang, S. (2019). Improvement of a mouse model of valproic acid-induced autism. *Nan Fang Yi Ke Da Xue Xue Bao* 39, 718–723. doi: 10.12122/j.issn.1673-4254.2019.06.14

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* 24:1714. doi: 10.3390/molecules24091714

**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: 09 March 2021
doi: 10.3389/fgene.2021.651610

# Identifying the Signatures and Rules of Circulating Extracellular MicroRNA for Distinguishing Cancer Subtypes

*Fei Yuan[1,2†], Zhandong Li[3†], Lei Chen[4†], Tao Zeng[5†], Yu-Hang Zhang[6], Shijian Ding[1], Tao Huang[5,7*] and Yu-Dong Cai[1*]*

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Department of Science and Technology, Binzhou Medical University Hospital, Binzhou, China, [3] College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [4] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [5] Bio-Med Big Data Center, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [6] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [7] CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

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

**Reviewed by:**
*Wenjin Li,*
*Shenzhen University, China*
*Xiao Chang,*
*Children's Hospital of Philadelphia, United States*

***Correspondence:**
*Tao Huang*
*huangtao@sibs.ac.cn*
*Yu-Dong Cai*
*cai_yud@126.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:** *10 January 2021*
**Accepted:** *10 February 2021*
**Published:** *09 March 2021*

**Citation:**
*Yuan F, Li Z, Chen L, Zeng T, Zhang Y-H, Ding S, Huang T and Cai Y-D (2021) Identifying the Signatures and Rules of Circulating Extracellular MicroRNA for Distinguishing Cancer Subtypes. Front. Genet. 12:651610. doi: 10.3389/fgene.2021.651610*

Cancer is one of the most threatening diseases to humans. It can invade multiple significant organs, including lung, liver, stomach, pancreas, and even brain. The identification of cancer biomarkers is one of the most significant components of cancer studies as the foundation of clinical cancer diagnosis and related drug development. During the large-scale screening for cancer prevention and early diagnosis, obtaining cancer-related tissues is impossible. Thus, the identification of cancer-associated circulating biomarkers from liquid biopsy targeting has been proposed and has become the most important direction for research on clinical cancer diagnosis. Here, we analyzed pan-cancer extracellular microRNA profiles by using multiple machine-learning models. The extracellular microRNA profiles on 11 cancer types and non-cancer were first analyzed by Boruta to extract important microRNAs. Selected microRNAs were then evaluated by the Max-Relevance and Min-Redundancy feature selection method, resulting in a feature list, which were fed into the incremental feature selection method to identify candidate circulating extracellular microRNA for cancer recognition and classification. A series of quantitative classification rules was also established for such cancer classification, thereby providing a solid research foundation for further biomarker exploration and functional analyses of tumorigenesis at the level of circulating extracellular microRNA.

**Keywords: circulating extracellular microRNA, signature, rule, cancer, subtype**

## INTRODUCTION

Cancer is one of the most threatening diseases to humans in the 21st century (Jemal et al., 2011; Siegel et al., 2019). Cancer is regarded as the second most deadly disease following cardiovascular diseases as it can invade multiple significant organs, including lung, liver, stomach, pancreas, and even brain. According to the World Health Organization's statistics in 2018 (Bray et al., 2018), more

than 18 million new cases and about 1 million deaths due to cancer exist globally. Accordingly, numerous studies have been conducted on the pathological mechanisms, clinical diagnosis, and treatment of cancer. Indeed, great achievements have been made in this field.

In particular, the identification of cancer biomarkers is regarded as one of the most significant parts of cancer studies as the foundation of clinical cancer diagnosis (Griffith et al., 2008; Ribaut et al., 2017) and related drug development (Jørgensen, 2019). Previously, researchers have revealed multiple cancer-subtype specific biomarkers by using genomics, transcriptomics, proteomics, or even multi-omic datasets (e.g., specific biomarkers of different cancer subtypes) at different biological omic levels. At the genomic level, specific biomarkers such as EGFR (Blakely et al., 2017) and KRAS (Arbour et al., 2018) exist for lung cancer, TP53 (Long et al., 2019) and LRP1B (Wang et al., 2019) for liver cancer, and BRAF (Ribas et al., 2019) and TP53 (Xiao et al., 2018) for skin melanoma. At the transcriptomic level, apart from the transcripts of already identified genomic biomarkers, multiple noncoding transcripts including microRNAs (e.g., hsa-miR-195-5p) (Li L. et al., 2020) and long non-coding RNAs (e.g., FOXE1 and HOXB13-AS1_2) for lung cancers have also been confirmed to be effective biomarkers for cancer diagnosis and classification (Li et al., 2019). With the development of biotechnology and biostatistics, cancer biomarkers at the proteomic level or even at the integrated multi-omic level have also been identified. For instance, in 2014, a systematic multi-omic analyses (Li et al., 2014) on lung cancer have revealed a group of potential multi-omic biomarkers for lung cancer, including EGFR and CCT6A. Analyzing data at different omics can improve accuracy and efficacy for potential biomarker identification. However, almost all such studies are based on cancer tissue *in situ*. In fact, during the large-scale screening for cancer prevention and early diagnosis, obtaining cancer-related tissues is impossible. To solve this problem, cancer-associated circulating biomarkers from liquid biopsy targeting have been presented, which has become one of the most important directions of clinical cancer diagnosis studies.

In the field of cancer-associated liquid biopsy, many research subdirections target biomarkers of different levels, such as cell-free DNA, plasma protein, and circulating RNAs. In particular, circulating RNAs have been extensively reported to be effective for cancer diagnosis or even classification. In 2004, researchers have shown that circulating plasma RNA may be a potential source of biomarkers for cancer screening (El-Hefnawy et al., 2004). In 2012, a systematic review has summarized the specificity and sensitivity of extracellular circulating RNAs to diagnosis and monitor different cancer subtypes (Zen and Zhang, 2012). In 2018, a study (Yokoi et al., 2018) integrating extracellular microRNA from serum for the diagnosis of ovarian cancer has demonstrated that extracellular microRNA biomarkers may distinguish one cancer subgroup from normal controls and contribute to the detailed cancer classification by comparing different cancer subgroups. These findings indicates that circulating extracellular microRNA may also be a specific "level/omics" of data that are sufficiently effective for cancer diagnosis and classification.

**TABLE 1 |** Statistic of samples used in this study.

| Index | Class | Sample size |
|---|---|---|
| 1 | Benign ovarian disease | 29 |
| 2 | Borderline ovarian tumor | 66 |
| 3 | Breast cancer | 115 |
| 4 | Colorectal cancer | 115 |
| 5 | Esophageal cancer | 88 |
| 6 | Gastric cancer | 115 |
| 7 | Hepatocellular carcinoma | 81 |
| 8 | Lung cancer | 115 |
| 9 | Non-cancer | 2759 |
| 10 | Ovarian cancer | 333 |
| 11 | Pancreatic cancer | 115 |
| 12 | Sarcoma | 115 |
| In total | | 4046 |

In the present study, based on shared data from a previous study (Yokoi et al., 2018), we performed an effective feature-selection procedure to identify candidate biomarkers for cancer recognition and classification by using multiple machine-learning models. The data was first analyzed by the Boruta (Kursa and Rudnicki, 2010) method to extract important microRNAs. Then, Max-Relevance and Min-Redundancy (mRMR) (Peng et al., 2005) feature selection method followed to evaluate the importance of each selected feature and ranked them in a feature list. Such list was fed into the incremental feature selection (IFS) (Liu and Setiono, 1998) method, incorporating one of the four classification algorithms, to extract latent microRNA biomarkers and build efficient classifiers. Additionally, a series of quantitative classification rules for cancer classification was established. This re-analysis on the extracellular microRNA dataset enabled the identification of a group of potential biomarkers for qualitative or quantitative cancer classification and laid a solid research foundation for further biomarker exploration and functional analyses of tumorigenesis at the circulating extracellular microRNA level.

# MATERIALS AND METHODS

## Data

We downloaded the extracellular microRNA profiles of various cancers and non-cancer samples from Gene Expression Omnibus with accession number GSE106817[1] (Yokoi et al., 2018); 4046 samples were included in such dataset and classified into 12 classes, including 11 cancer types and non-cancer class. The sample size of each class is given in **Table 1**. For each sample, the expression levels of 2565 microRNAs were measured with 3D-Gene Human miRNA V21_1.0.0. To accelerate the precision diagnosis of pan-cancer, we built a computational pipeline for extracellular microRNA-based cancer detection and classification.

[1]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106817

## Boruta Feature Filtering

In the investigated dataset, lots of microRNAs (features) were involved. Evidently, not all microRNAs are related to the investigated cancer types. It is necessary to extract important ones and discard others. Here, we employed Boruta (Kursa and Rudnicki, 2010) method to quickly select relevant features with particular class labels (e.g., cancer types or non-cancer class). This method has been applied to deal with different biological and medical problems (Pan et al., 2020; Yuan et al., 2020; Zhang et al., 2021a).

Boruta is a random forest (RF)-based feature filtering method. Its computation steps included the following steps: (1) creation of shuffled data with shuffling original features in the original dataset, (2) evaluation of feature importance by comparing the RF on the original and shuffled data, (3) calculation of Z score for each feature depending on the feature's importance score, (4) determination of the important feature by comparing its Z score with those of the shadow features, and (5) the above procedures stop until one of the following conditions was satisfied: (i) each feature is tagged as either "important" or "unimportant" and (ii) a predefined number of iterations is reached. The features tagged by "important" were kept for further analysis.

This study adopted the Boruta program obtained from https://github.com/scikit-learn-contrib/boruta_py, which was implemented by Python. For convenience, default parameters were used.

## Max-Relevance and Min-Redundancy Feature Selection

For the features selected by the Boruta method, mRMR (Peng et al., 2005) feature selection method was adopted to evaluate their importance. This method has wide applications in tackling several biological and medical problems (Chen et al., 2018, 2020; Zhao et al., 2018; Li M. et al., 2020; Pan et al., 2021).

mRMR method employed the Max-Relevance and Min-Redundancy to assess the importance of features. Features with high relevance to class labels and low redundancy to other features were termed to be important. To quantify the relevance and redundancy, it uses mutual information (MI). For two variables $x$ and $y$, the MI score between them is defined by:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \quad (1)$$

where $p(x)/p(y)$ and $p(x,y)$ represent the marginal probabilistic density of $x/y$ and joint probabilistic density of $x$ and $y$, respectively. The mRMR method evaluates the importance of features by listing them in a feature list. A loop procedure is performed to produce the list. At first, this list is empty. For each feature not in the list, calculate its relevance to class labels, measured by the MI score of it and class label variable, and its redundancy to features in the list, measured by the average MI scores between it and features in the list. The feature with highest difference of relevance and redundancy is picked up and added to the list. When all features are in the list, the loop stops. This list was called mRMR feature list in this study. The combination of some top features can be the optimum feature space for a given classification algorithm.

The current study adopted the mRMR program retrieved from http://penglab.janelia.org/proj/mRMR/. Likewise, default parameters were used.

## Incremental Feature Selection

mRMR method only provided a feature list. It was still a problem for selecting optimum features for a given classification algorithm. Thus, we employed the IFS method (Liu and Setiono, 1998; Zhang et al., 2021b).

Using the mRMR feature list from the above mRMR, a series of feature subsets can be produced with a step interval as one. For example, the first feature subset includes the first feature in the list, and the second feature subset includes the first two features, and so on. Each classifier is then trained on the training data, in which samples are represented by features in one feature subset. Then, each classifier is evaluated by 10-fold cross-validation (Kohavi, 1995). The classifier with the best performance is selected and termed as the optimum classifier. The corresponding feature subset is determined as the optimal one.

## Synthetic Minority Oversampling Technique

Considering the used extracellular microRNA dataset has remarkably different numbers of samples (see **Table 1**), synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) was performed to produce sufficient new samples for minor classes. When evaluating the performance of a classifier with ten-fold cross-validation, we used SMOTE to create a new dataset with an equal sample number of different classes. For this analysis, the "SMOTE" tool in Weka software[2] (Frank et al., 2004; Witten and Frank, 2005) was used.

## Classification Algorithm

To execute the IFS method, one classification algorithm is necessary. In this study, we tried four classification algorithms: RF (Breiman, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), k-nearest neighbor (kNN) (Cover and Hart, 1967), and decision tree (DT) (Safavian and Landgrebe, 1991). These algorithms have wide applications in tackling different problems (Ben-Hur et al., 2008; Ahmed et al., 2013; Chen et al., 2017; Sankari and Manimegalai, 2018; Baranwal et al., 2019; Jia et al., 2020; Liang et al., 2020; Liu H. et al., 2020; Zhou et al., 2020a,b; Zhu et al., 2021). For convenience, these algorithms were performed with their default parameters, which are set in the corresponding platform.

### RF

It is an assembly classification algorithm that contains several DTs. Each DT is built by randomly selecting samples and features from the original dataset. For a query sample, each DT provides the prediction class. RF integrates these prediction classes with majority voting, i.e., the class receiving most votes is the predicted class of RF. Although DT is a relatively weak classification

[2]https://www.cs.waikato.ac.nz/ml/weka

algorithm, RF is much stronger. The current study adopted the Scikit-learn package to implement RF.

### SVM

It can transform data with a nonlinear pattern from original low-dimensional data space to a new high-dimensional data space, where the data display a linear pattern. Then, it divides the data points in such high-dimensional space, requiring data-interval maximization among different data classes/groups. It could predict the class or group of a new sample by determining the interval to which this new sample data belongs. Here, the tool "SMO" in Weka was adopted to construct SVM classifiers. The training procedure of this SVM is optimized by the sequential minimal optimization algorithm (Platt, 1998).

### kNN

It is one of the most classic classification algorithms. For a test sample, it initially computes the distance between it and the training samples. Then, it ranks all training samples with the increasing order of the distances. Next, it selects the $k$ high-ranked training samples (i.e., nearest k neighbors) and further estimates the label distribution of these $k$ samples. The label distribution is then used to help predict the class of test sample, i.e., the class label with the highest frequency in the label distribution. The tool "IBk" in Weka was performed for kNN classifier building.

### DT

Different from the above three classification algorithms, which can only be used to construct black-box classifiers, DT can construct human understanding classification and regression models by using interpretative rules. Generally, it indicates individual features' roles and weights in classification or regression models by using the IF–TEHN format. Here, the CART algorithm with the Gini index in the Scikit-learn package was used for DT classifier construction.

## RESULTS AND DISCUSSION

In this study, we gave a computational investigation on the extracellular microRNA dataset of multiple cancer types. Some feature selection methods and classification algorithms were adopted. The entire procedures are illustrated in **Figure 1**. This section first introduced the results and then gave an extensive discussion.

### Results of Boruta and mRMR Methods

We first applied the Boruta method to the extracellular microRNA dataset for discarding non-essential features (microRNAs). As a result, 1849 features were excluded and 716 features were kept. These remaining features are provided in **Supplementary Table S1**.

For the remaining 716 features, they were further analyzed by the mRMR method. As mentioned in Section "Max-Relevance and Min-Redundancy Feature Selection", a feature list, mRMR feature list, was generated, in which features were ranked according to their importance. This list is also provided in **Supplementary Table S1**.

### Results of IFS Method With Different Classification Algorithms

The mRMR feature list generated by mRMR method was fed into the IFS method. Using an interval step of 1, many feature subsets were extracted, e.g., the first feature subset contained the top-ranked feature, and the second feature subset contained the two top-ranked features. For each feature subset and one of the four classification algorithms (SVM, RF, kNN, and DT), a classifier was built on samples represented by features in the subset. Ten-fold cross-validation (Kohavi, 1995) was adopted to evaluate the performance of each classifier. Notably, SMOTE was applied when assessing the performance of each classifier. Results were counted as the following measurements: accuracy on each class, overall accuracy (ACC) and Matthew correlation coefficient (MCC) (Matthews, 1975; Gorodkin, 2004). These measurements are available in **Supplementary Table S2**. For an easy observation, one IFS curve was plotted for each classification algorithm, in which MCC was set as the Y-axis and number of used features was set as the X-axis, which is shown in **Figure 2**. For kNN, the highest MCC was 0.957 when top 12 features were used. Accordingly, the optimum kNN classifier was built using these 12 features. The highest MCC of RF was 0.931, which was obtained by the top 14 features. The optimum RF classifier with these top 14 features can be set up. As for SVM, the highest MCC was 0.987 when top 552 features were adopted. It was higher than that of the optimum kNN or RF classifiers. The ACCs of above three optimum classifiers are listed in **Table 2**. The ACC of the optimum SVM classifier was also highest. The accuracies on 12 classes yielded by these optimum classifiers are illustrated in **Figure 3**. Evidently, the optimum SVM classifier was also best. Because the partition of the 10-fold cross-validation can influence the evaluation results, we further tested the performance of the optimum SVM classifier with ten-fold cross-validation 20 times. Obtained ACCs and MCCs are illustrated in **Figure 4**. The ACCs varied between 0.990 and 1.000, whereas MCCs were between 0.980 and 1.000, indicating that such optimum classifier was quite stable and above results can be believable.

In addition to three black-box classification algorithms, we also employed a white-box algorithm, DT, to do the same test. The IFS results are also provided in **Supplementary Table S2** and the IFS curve was plotted in **Figure 2**. The optimum DT classifier produced the MCC of 0.918, which was based on the top 74 features. The corresponding ACC was 0.955, which is listed in **Table 2**. The ACC and MCC were lower than those of the above-mentioned three optimum classifiers. Furthermore, the accuracies on 12 classes of the optimum DT classifier are shown in **Figure 3**. They were also lower than those of other three optimum classifiers. Although the performance of the optimum DT classifier was lower than other three optimum classifiers, it can provide a clear classification procedure, thereby providing more insights to investigate different cancer types. In view of this, we constructed a DT based on the top 74 features, which were used to build the optimum DT classifier. Then, 333 microRNA rules were extracted from such DT, which are available in **Supplementary Table S3**. Each class was assigned to some rules, where the number of rules (50) on "ovarian cancer" was most, followed by "non-cancer." The numbers of rules on "benign

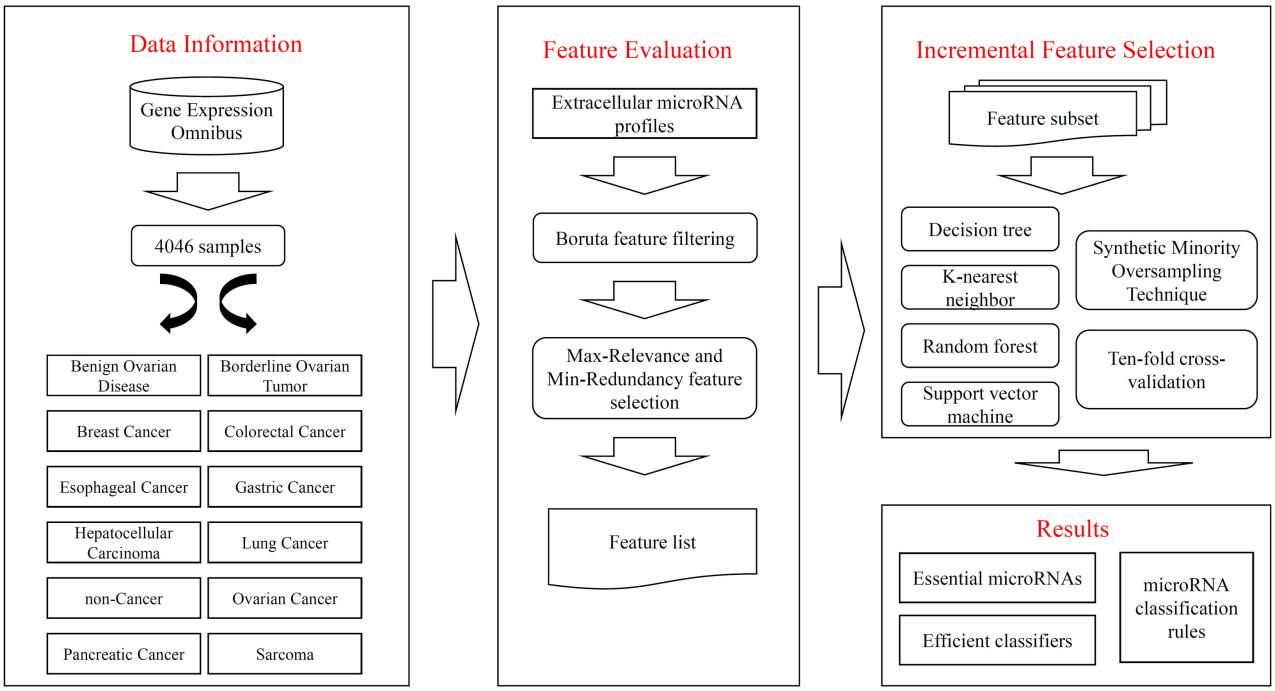


**FIGURE 1 |** Flow chart to show entire procedures. The original data are retrieved from Gene Expression Omnibus, which contains 4046 samples and 12 classes. The extracellular microRNA profiles are analyzed by Boruta and mRMR methods one by one, resulting in a feature list. The list is fed into the IFS method, incorporating four classification algorithms, SMOTE and ten-fold cross-validation, to extract essential microRNAs, construct efficient classifiers, and set up classification rules.

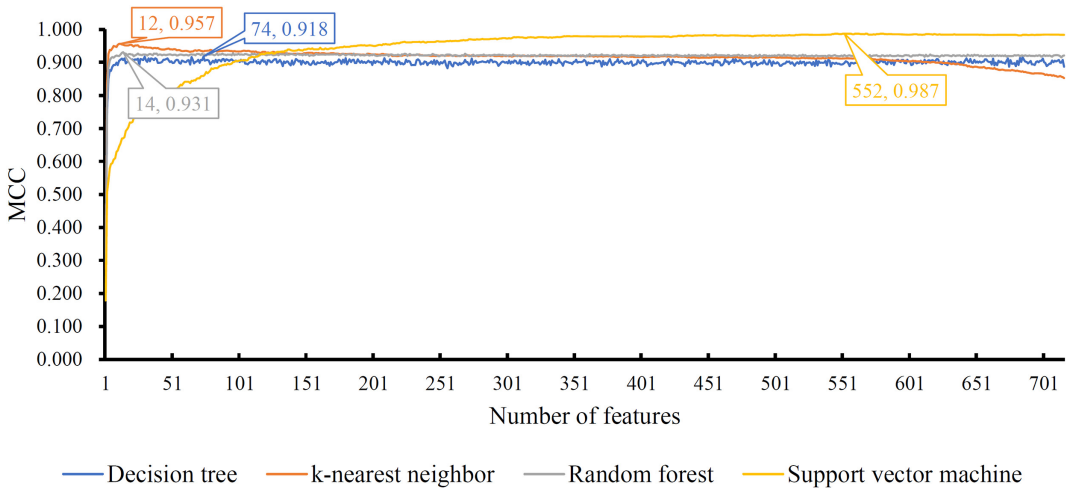


**FIGURE 2 |** IFS curves with different classification algorithms on different number of features. Four algorithms yielded the highest MCCs with top 74, 12, 14, and 552 features.

ovarian disease" and "gastric cancer" were least, which were only 17. The number of rules for each class is shown in **Figure 5**.

Here, a group of qualitative microRNAs (features) and quantitative microRNA rules were identified to contribute to detailed cancer-classification recognition. According to recent publications, the top-ranked optimal features and rules were supported and validated with the respective cancer-subtype specific pathological roles, which will be discussed in Sections "Optimal MicroRNAs Contributing to Cancer Classification" and "Optimal MicroRNA Rules Contributing to Cancer Classification".

## Optimal MicroRNAs Contributing to Cancer Classification

By analyzing the shared extracellular microRNA dataset, we identified a group of microRNAs that can effectively distinguish different cancer subtypes but not cancer or controls, reflecting

**TABLE 2 |** Performance of IFS with four different classification algorithms.

| Classification algorithm | Number of features | ACC | MCC |
|---|---|---|---|
| Decision tree | 74 | 0.955 | 0.918 |
| k-nearest neighbor | 12 | 0.976 | 0.957 |
| Random forest | 14 | 0.961 | 0.931 |
| Support vector machine | 552 | 0.993 | 0.987 |

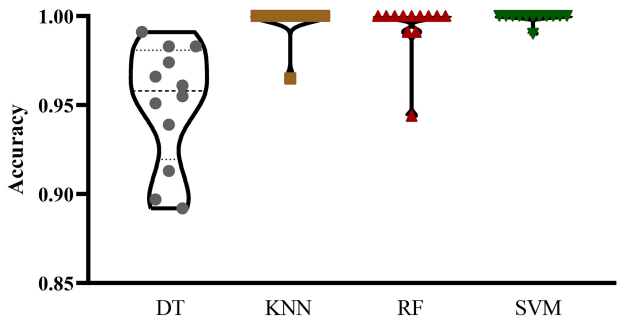


**FIGURE 3 |** Violin plot to show accuracies on 12 classes yielded by the optimum classifiers with four different classification algorithms. The optimum SVM classifier was best.

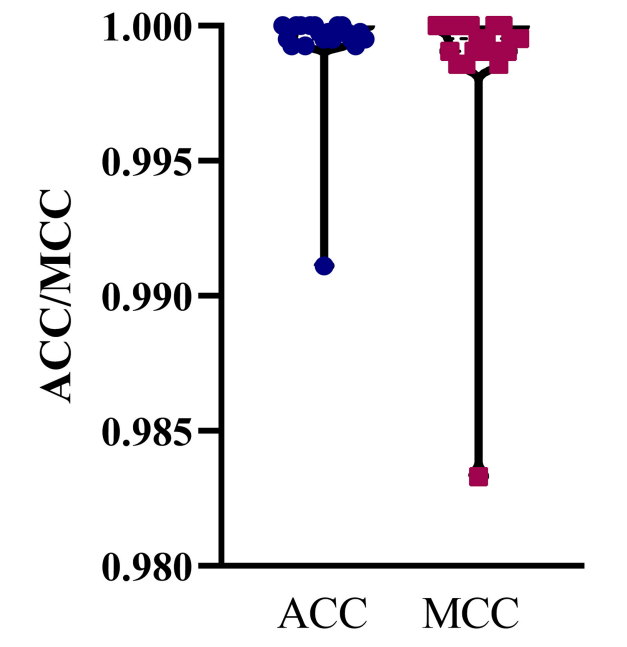


**FIGURE 4 |** Violin plot to show ACCs and MCCs yielded by the optimum SVM classifier under 10-fold cross-validation 20 times. ACC and MCC vary in a small interval, suggesting the stability of the optimum SVM classifier.

the internal differences among different cancer subtypes. This section selected the top 10 microRNAs in the mRMR feature list for detailed analysis, which are listed in **Table 3**.

The first identified microRNA was hsa-miR-5100 (MIMAT0022259). According to recent publications, this microRNA has been identified in multiple tumor-related studies and is functionally correlated with tumorigenesis (Tang et al., 2014; Wang et al., 2016; Jacob et al., 2018; Tian et al., 2020). However, it has been confirmed to have a specific expression level only in plasma in colon cancer (Jacob et al., 2018) and in extracellular matrix in oral carcinoma (Kawakubo-Yasukochi et al., 2018). Accordingly, predicting this microRNA to have discriminative capacity in 11 candidate cancer subtypes is reasonable.

The next predicted microRNA signature was miR-6088 (MIMAT0023713). It has also been identified in only three cancer subtypes, namely nasopharyngeal cancer (Li K. et al., 2020), ovarian cancer (Pandey et al., 2019), and melanoma (Wozniak et al., 2017), thereby confirming its classification capacity for ovarian cancer in our dataset. The third predicted signature, miR-4532 (MIMAT0019071), has also been regarded as a potential circulating extracellular cancer biomarker according to previous studies (Fiorino et al., 2016; Pascut et al., 2019; Zhao et al., 2019), including hepatocellular carcinoma (Fiorino et al., 2016) and leukemia (Zhao et al., 2019).

As regards the two microRNAs miR-6746 (MIMAT0027392) and miR-8073 (MIMAT0031000), both reportedly participate in specific cancer-associated tumorigenesis, corresponding with our prediction. For miR-6746, it has been shown to have specific expression level in the plasma of pancreatic cancer patients but not in those of other patients (Sheng et al., 2020). For miR-8073, it has been identified in both pancreatic (Shams et al., 2020) and breast (Cui et al., 2018) cancers, implying that such microRNA may distinguish two cancer subtypes from the other cancer subtypes and normal controls.

The next microRNA, miR-6800 (MIMAT0027500), is also reportedly a potential biomarker for prostate (Liu H.P. et al., 2020) and colorectal (Yan et al., 2017) cancers, confirming its capacity for distinguishing colorectal cancer from 11 other cancer subtypes and normal controls in this analysis.

The remaining microRNAs, namely miR-1343 (MIMAT0019776), miR-4783 (MIMAT0019947), miR-221 (MIMAT0000278), and miR-4787 (MIMAT0019957), have also been confirmed to contribute to specific cancer subtypes [e.g., lung adenocarcinoma correlated with miR-1343 (Zhang X. et al., 2020), rectal cancer correlated with miR-4783 (Mullany et al., 2016), prostate cancer correlated with miR-221 (Agaoglu et al., 2011), and pancreatic cancer correlated with miR-4787 (Mody et al., 2016)], thereby further validating the efficacy and accuracy of our newly established computational workflow.

## Optimal MicroRNA Rules Contributing to Cancer Classification

In addition to the above identified microRNA si+oldd gnatures, we recognized and established a series of quantitative classification rules for more interpretable cancer classification. Due to the limitation of the manuscript's length, we selected one representative rule for each specific cancer classification for subsequent detailed discussions, including 11 cancer subtypes and 1 normal control.

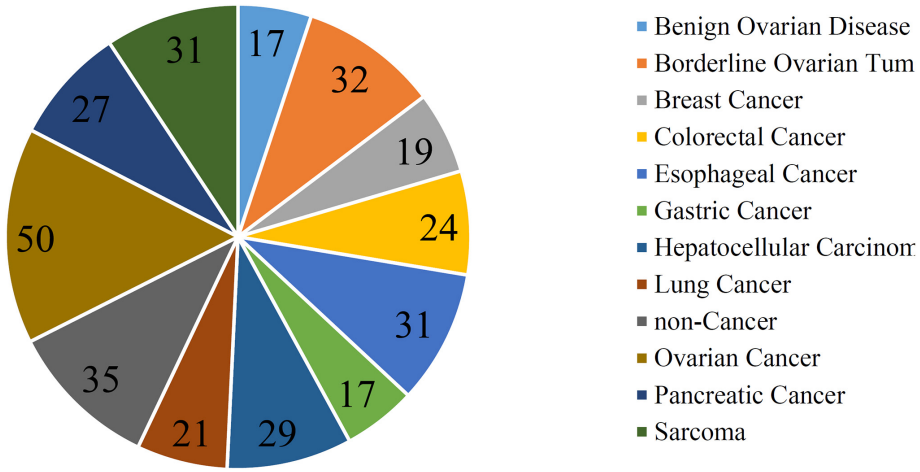


**FIGURE 5 |** Pie chart to show the number of rules for each class.

**TABLE 3 |** Top 10 microRNAs identified by Boruta and mRMR methods.

| Rank | miRbase accession number | microRNA (Full name) |
|---|---|---|
| 1 | MIMAT0022259 | hsa-miR-5100 |
| 2 | MIMAT0023713 | miR-6088 |
| 3 | MIMAT0019071 | miR-4532 |
| 4 | MIMAT0027392 | miR-6746 |
| 5 | MIMAT0031000 | miR-8073 |
| 6 | MIMAT0027500 | miR-6800 |
| 7 | MIMAT0019776 | miR-1343 |
| 8 | MIMAT0019947 | miR-4783 |
| 9 | MIMAT0000278 | miR-221 |
| 10 | MIMAT0019957 | miR-4787 |

The first rule for the identification of Benign Ovarian Disease is rule 58, involving 14 different microRNAs. Among these microRNAs, a specific microRNA named as miR-5100 (MIMAT0022259) has been detected in the plasma of benign ovarian cysts, which can be classified into benign ovarian disease, corresponding with our prediction (Zhang L. et al., 2020). As for Borderline Ovarian Tumor, rule 72 has been confirmed to contribute to the identification of patients with such disease. Among multiple microRNA biomarkers, the significant one is also miR-5100 (MIMAT0022259), indicating that it is still an ovarian-associated signature. Moreover, miR-296 (MIMAT0000690) has been predicted to be correlated with Borderline Ovarian Tumor, whose correlation has also been verified (Li Y. et al., 2020). For breast cancer, as discussed above, miR-8073 (MIMAT0031000) shown in rule 145 has been validated to be related to breast cancer with relatively high expression level (Cui et al., 2018). Similarly, miR-6800 (MIMAT0027500) of colorectal cancer shown in rule 274 has been discussed above (Yan et al., 2017), indicating a relatively low expression level of such microRNA compared with normal controls and other cancer subtypes.

For esophageal cancer and gastric cancer, the optimal quantitative microRNA features in the rules have also been validated. In esophageal cancer, as described in rule 13, miR-6784 (MIMAT0027468) has been shown to have a relatively high expression level and validated by recent publications (Fujihara et al., 2015). As for gastric cancer-associated signatures at the microRNA level, miR-3663 (MIMAT0018085) has been shown to be a potential biomarker for gastrointestinal tumors, including gastric cancer (Lee et al., 2016; Xu et al., 2018; Kubo et al., 2019). To specifically identify gastric cancer, another microRNA named miR-1343 (MIMAT0019776) has been shown to be a specific gastric cancer-associated microRNA by regulating TEAD4 (Zhou et al., 2017), thereby validating our prediction.

As regards class hepatocellular carcinoma, lung cancer, and ovarian cancer, we also identified specific classification rules with the specific microRNA signatures discussed above. For hepatocellular carcinoma, miR-4532 (MIMAT0019071) has been shown to be a decisive biomarker with a relatively low expression level (Fiorino et al., 2016) in rule 158, corresponding with our discussion above. In lung cancer-associated rules, a typical rule named rule 162 has been shown to have a relatively high expression level of miR-1343 (MIMAT0019776) in patients' plasma compared with normal controls and other patients with other cancer subtypes (Zhang X. et al., 2020). Similar rules have been established for ovarian cancer involving miR-6088 (Pandey et al., 2019), implying the reliability of our predicted rules.

For pancreatic cancer and sarcoma, miR-6746 (MIMAT0027392) shown as a significant parameter in rule 207 has also been confirmed to be correlated with and be specific for pancreatic cancer, as discussed above (Sheng et al., 2020), confirming the efficacy of our prediction. For sarcoma, miR-92B (MIMAT0004792) shown in rule 9 has been presented to be up-regulated in sarcoma compared with other cancer subtypes and no cancer controls. According to recent publications, in 2017, researchers already confirmed that miR-92B is a novel biomarker for carcinoma monitoring (Uotani et al., 2017),

corresponding with our prediction. Apart from the discussion above, individuals with extracellular microRNA profiling not satisfying either of the above rules may be classified into controls.

## CONCLUSION

As discussed above, our identified optimal microRNA signatures and related quantitative classification rules have all been verified by recent publications, helping us classify different cancer subgroups and non-cancer controls. For the first time, we integrated feature selection and machine-learning models with inherited information at the extracellular microRNA level to present a new workflow for cancer-classification recognition, early diagnosis, and monitoring with high prediction specificity. The promising results obtained in this study (microRNA signatures and rules) may validate the specific and diverse roles of extracellular microRNAs during tumorigenesis and may also lay a solid foundation for further studies on the potentials of extracellular microRNAs on tumor diagnosis and monitoring.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106817.

## ETHICS STATEMENT

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## AUTHOR CONTRIBUTIONS

TH and Y-DC designed the study. FY, LC, TZ, and SD performed the experiments. FY, ZL, TZ, and Y-HZ analyzed the results. FY, ZL, LC, and TZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

## FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200), National Key R&D Program of China (2017YFC1201200), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), and the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.651610/full#supplementary-material

**Supplementary Table 1 |** Features filtered by Boruta and their ranks generated by mRMR.

**Supplementary Table 2 |** Performance of IFS with different classifiers.

**Supplementary Table 3 |** Classification rules generated by decision tree.

## REFERENCES

Agaoglu, F. Y., Kovancilar, M., Dizdar, Y., Darendeliler, E., Holdenrieder, S., Dalay, N., et al. (2011). Investigation of miR-21, miR-141, and miR-221 in blood circulation of patients with prostate cancer. *Tumor Biol.* 32, 583–588. doi: 10.1007/s13277-011-0154-9

Ahmed, F., Kaundal, R., and Raghava, G. P. (2013). PHDcleav: a SVM based method for predicting human dicer cleavage sites using sequence and secondary structure of miRNA precursors. *BMC Bioinformatics* 14(Suppl. 14):S9. doi: 10.1186/1471-2105-14-S14-S9

Arbour, K. C., Jordan, E., Kim, H. R., Dienstag, J., Helena, A. Y., Sanchez-Vega, F., et al. (2018). Effects of co-occurring genomic alterations on outcomes in patients with KRAS-mutant non–small cell lung cancer. *Clin. Cancer Res.* 24, 334–340. doi: 10.1158/1078-0432.ccr-17-1841

Baranwal, M., Magner, A., Elvati, P., Saldinger, J., Violi, A., and Hero, A. O. (2019). A deep learning architecture for metabolic pathway prediction. *Bioinformatics* 36, 2547–2553. doi: 10.1093/bioinformatics/btz954

Ben-Hur, A., Ong, C. S., Sonnenburg, S. R., Lkopf, B. S., and Ra:Tsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* 4:e1000173. doi: 10.1371/journal.pcbi.1000173

Blakely, C. M., Watkins, T. B., Wu, W., Gini, B., Chabon, J. J., Mccoach, C. E., et al. (2017). Evolution and clinical impact of co-occurring genetic alterations in advanced-stage EGFR-mutant lung cancers. *Nat. Genet.* 49, 1693–1704. doi: 10.1038/ng.3990

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

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

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

Chen, L., Li, Z., Zeng, T., Zhang, Y.-H., Liu, D., Li, H., et al. (2020). Identifying robust microbiota signatures and interpretable rules to distinguish cancer subtypes. *Front. Mol. Biosci.* 7:604794. doi: 10.3389/fmolb.2020.604794

Chen, L., Pan, X., Hu, X., Zhang, Y.-H., Wang, S., Huang, T., et al. (2018). Gene expression differences among different MSI statuses in colorectal cancer. *Int. J. Cancer* 143, 1731–1740. doi: 10.1002/ijc.31554

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

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

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* 13, 21–27.

Cui, X., Li, Z., Zhao, Y., Song, A., Shi, Y., Hai, X., et al. (2018). Breast cancer identification via modeling of peripherally circulating miRNAs. *PeerJ* 6:e4551. doi: 10.7717/peerj.4551

El-Hefnawy, T., Raja, S., Kelly, L., Bigbee, W. L., Kirkwood, J. M., Luketich, J. D., et al. (2004). Characterization of amplifiable, circulating RNA in plasma and its potential as a tool for cancer diagnostics. *Clin. Chem.* 50, 564–573. doi: 10.1373/clinchem.2003.028506

Fiorino, S., Bacchi-Reggiani, M. L., Visani, M., Acquaviva, G., Fornelli, A., Masetti, M., et al. (2016). MicroRNAs as possible biomarkers for diagnosis and prognosis of hepatitis B-and C-related-hepatocellular-carcinoma. *World J. Gastroenterol.* 22, 3907–3936.

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Fujihara, S., Kato, K., Morishita, A., Iwama, H., Nishioka, T., Chiyo, T., et al. (2015). Antidiabetic drug metformin inhibits esophageal adenocarcinoma cell proliferation in vitro and in vivo. *Int. J. Oncol.* 46, 2172–2180. doi: 10.3892/ijo.2015.2903

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

Griffith, O. L., Chiu, C. G., Gown, A. M., Jones, S. J., and Wiseman, S. M. (2008). Biomarker panel diagnosis of thyroid cancer: a critical review. *Expert Rev. Anticancer Ther.* 8, 1399–1413. doi: 10.1586/14737140.8.9.1399

Jacob, H., Stanisavljevic, L., Storli, K. E., Hestetun, K. E., Dahl, O., and Myklebust, M. P. (2018). A four-microRNA classifier as a novel prognostic marker for tumor recurrence in stage II colon cancer. *Sci. Rep.* 8:6157.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90.

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-based machine learning model for predicting the metabolic pathways of compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/access.2020.3009439

Jørgensen, J. T. (2019). A paradigm shift in biomarker guided oncology drug development. *Ann. Transl. Med.* 7:148. doi: 10.21037/atm.2019.03.36

Kawakubo-Yasukochi, T., Morioka, M., Hazekawa, M., Yasukochi, A., Nishinakagawa, T., Ono, K., et al. (2018). miR-200c-3p spreads invasive capacity in human oral squamous cell carcinoma microenvironment. *Mol. Carcinog* 57, 295–302. doi: 10.1002/mc.22744

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International joint Conference on artificial intelligence* (New Jersey, NJ: Lawrence Erlbaum Associates Ltd), 1137–1145.

Kubo, H., Hiroshima, Y., Mori, R., Saigusa, Y., Murakami, T., Yabushita, Y., et al. (2019). MiR-194-5p in pancreatic ductal adenocarcinoma peritoneal washings is associated with peritoneal recurrence and overall survival in peritoneal cytology-negative patients. *Ann. Surg. Oncol.* 26, 4506–4514. doi: 10.1245/s10434-019-07793-y

Kursa, M., and Rudnicki, W. (2010). Feature selection with the boruta package. *J. Stat. Softw.* 36, 1–13.

Lee, A. R., Park, J., Jung, K. J., Jee, S. H., and Kim-Yoon, S. (2016). Genetic variation rs7930 in the miR-4273-5p target site is associated with a risk of colorectal cancer. *Onco Targets Ther.* 9, 6885–6895. doi: 10.2147/ott.s108787

Li, K., Zhu, X., Li, L., Ning, R., Liang, Z., Zeng, F., et al. (2020). Identification of non-invasive biomarkers for predicting the radiosensitivity of nasopharyngeal carcinoma from serum microRNAs. *Sci. Rep.* 10:5161.

Li, L., Feng, T., Zhang, W., Gao, S., Wang, R., Lv, W., et al. (2020). MicroRNA biomarker hsa-miR-195-5p for detecting the risk of lung cancer. *Int. J. Genomics* 2020:7415909.

Li, L., Wei, Y., To, C., Zhu, C.-Q., Tong, J., Pham, N.-A., et al. (2014). Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact. *Nat. Commun.* 5:5469.

Li, M., Pan, X. Y., Zeng, T., Zhang, Y. H., Feng, K. Y., Chen, L., et al. (2020). Alternative polyadenylation modification patterns reveal essential posttranscription regulatory mechanisms of tumorigenesis in multiple tumor types. *Biomed Res. Int.* 2020:6384120.

Li, R., Yang, Y.-E., Yin, Y.-H., Zhang, M.-Y., Li, H., and Qu, Y.-Q. (2019). Methylation and transcriptome analysis reveal lung adenocarcinoma-specific diagnostic biomarkers. *J. Transl. Med.* 17:324.

Li, Y., Wang, J., Zhu, Y., and Chen, Y. (2020). Prediction and analysis of hub genes in ovarian cancer based on network analysis. Research Square. Preprint.

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of drug side effects with a refined negative sample selection strategy. *Comput. Math. Methods Med.* 2020:1573543.

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230.

Liu, H., Hu, B., Chen, L., and Lu, L. (2020). Identifying protein subcellular location with embedding features learned from networks. *Curr. Proteomics* 17.

Liu, H.-P., Lai, H.-M., and Guo, Z. (2020). Prostate cancer early diagnosis: circulating microRNA pairs potentially beyond single microRNAs upon 1231 serum samples. *Brief. Bioinform.* bbaa111.

Long, J., Wang, A., Bai, Y., Lin, J., Yang, X., Wang, D., et al. (2019). Development and validation of a TP53-associated immune prognostic model for hepatocellular carcinoma. *EBioMedicine* 42, 363–374. doi: 10.1016/j.ebiom.2019.03.022

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

Mody, H. R., Hung, S. W., Alsaggar, M., Griffin, J., and Govindarajan, R. (2016). Inhibition of S-adenosylmethionine-dependent methyltransferase attenuates TGFβ1-induced EMT and metastasis in pancreatic cancer: putative roles of miR-663a and miR-4787-5p. *Mol. Cancer Res.* 14, 1124–1135. doi: 10.1158/1541-7786.mcr-16-0083

Mullany, L. E., Herrick, J. S., Wolff, R. K., Stevens, J. R., and Slattery, M. L. (2016). Association of cigarette smoking and microRNA expression in rectal cancer: insight into tumor phenotype. *Cancer Epidemiol.* 45, 98–107. doi: 10.1016/j.canep.2016.10.011

Pan, X. Y., Zeng, T., Zhang, Y. H., Chen, L., Feng, K. Y., Huang, T., et al. (2020). Investigation and prediction of human interactome based on quantitative features. *Front. Bioeng. Biotechnol.* 8:730. doi: 10.3389/fbioe.2020.00730

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021). Identification of protein subcellular localization with network and functional embeddings. *Front. Genetics* 11:626500. doi: 10.3389/fgene.2020.626500

Pandey, R., Woo, H.-H., Varghese, F., Zhou, M., and Chambers, S. K. (2019). Circulating miRNA profiling of women at high risk for ovarian cancer. *Transl. Oncol.* 12, 714–725. doi: 10.1016/j.tranon.2019.01.006

Pascut, D., Krmac, H., Gilardi, F., Patti, R., Calligaris, R., Crocè, L. S., et al. (2019). A comparative characterization of the circulating miRNome in whole blood and serum of HCC patients. *Sci. Rep.* 9:8265.

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

Platt, J. (1998). *Sequential Minimal Optimizaton: A Fast Algorithm for Training Support Vector Machines*. Washington, DC: Microsoft Research. *Technical Report MSR-TR-98-14*.

Ribas, A., Lawrence, D., Atkinson, V., Agarwal, S., Miller, W. H., Carlino, M. S., et al. (2019). Combined BRAF and MEK inhibition with PD-1 blockade immunotherapy in BRAF-mutant melanoma. *Nat. Med.* 25, 936–940. doi: 10.1038/s41591-019-0476-5

Ribaut, C., Loyez, M., Larrieu, J.-C., Chevineau, S., Lambert, P., Remmelink, M., et al. (2017). Cancer biomarker sensing using packaged plasmonic optical fiber gratings: towards in vivo diagnosis. *Biosens. Bioelectron.* 92, 449–456. doi: 10.1016/j.bios.2016.10.081

Safavian, S. R., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Syst. Man Cybern.* 21, 660–674. doi: 10.1109/21.97458

Sankari, E. S., and Manimegalai, D. (2018). Predicting membrane protein types by incorporating a novel feature set into Chou's general PseAAC. *J. Theor. Biol.* 455, 319–328. doi: 10.1016/j.jtbi.2018.07.032

Shams, R., Saberi, S., Zali, M., Sadeghi, A., Ghafouri-Fard, S., and Aghdaei, H. A. (2020). Identification of potential microRNA panels for pancreatic cancer diagnosis using microarray datasets and bioinformatics methods. *Sci. Rep.* 10:7559.

Sheng, L.-P., Han, C.-Q., Nie, C., Xu, T., Zhang, K., Li, X.-J., et al. (2020). *Identification of Potential Serum Exosomal microRNAs Involved in Acinar-Ductal Metaplasia That is A Precursor of Pancreatic Cancer Associated with Chronic Pancreatitis*. Durham: Reserach Square.

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA Cancer J. Clin.* 69, 7–34.

Tang, J.-F., Yu, Z.-H., Liu, T., Lin, Z.-Y., Wang, Y.-H., Yang, L.-W., et al. (2014). Five miRNAs as novel diagnostic biomarker candidates for primary nasopharyngeal carcinoma. *Asian Pac. J. Cancer Prev.* 15, 7575–7581. doi: 10.7314/apjcp.2014.15.18.7575

Tian, X., Liu, Y., Wang, Z., and Wu, S. (2020). lncRNA SNHG8 promotes aggressive behaviors of nasopharyngeal carcinoma via regulating miR-656-3p/SATB1 axis. *Biomed. Pharmacother.* 131:110564. doi: 10.1016/j.biopha.2020.110564

Uotani, K., Fujiwara, T., Yoshida, A., Iwata, S., Morita, T., Kiyono, M., et al. (2017). Circulating MicroRNA-92b-3p as a Novel Biomarker for Monitoring of Synovial Sarcoma. *Sci. Rep.* 7:14634.

Wang, L., Yan, K., Zhou, J., Zhang, N., Wang, M., Song, J., et al. (2019). Relationship of liver cancer with LRP1B or TP53 mutation and tumor mutation burden and survival. *J. Clin. Oncol.* 37, 1573–1573. doi: 10.1200/jco.2019.37.15_suppl.1573

Wang, Y., Chen, J., Lin, Z., Cao, J., Huang, H., Jiang, Y., et al. (2016). Role of deregulated microRNAs in non-small cell lung cancer progression using fresh-frozen and formalin-fixed, paraffin-embedded samples. *Oncol. Lett.* 11, 801–808. doi: 10.3892/ol.2015.3976

Witten, I. H., and Frank, E. (eds) (2005). *Data Mining:Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann.

Wozniak, M., Peczek, L., Czernek, L., and Düchler, M. (2017). Analysis of the miRNA profiles of melanoma exosomes derived under normoxic and hypoxic culture conditions. *Anticancer Res.* 37, 6779–6789.

Xiao, W., Du, N., Huang, T., Guo, J., Mo, X., Yuan, T., et al. (2018). TP53 mutation as potential negative predictor for response of anti-CTLA-4 therapy in metastatic melanoma. *EBioMedicine* 32, 119–124. doi: 10.1016/j.ebiom.2018.05.019

Xu, D., Guo, J., Zhu, G., Wu, H., Zhang, Q., and Cui, T. (2018). MiR-363-3p modulates cell growth and invasion in glioma by directly targeting pyruvate dehydrogenase B. *Eur. Rev. Med. Pharmacol. Sci.* 22, 5230–5239.

Yan, S., Han, B., Gao, S., Wang, X., Wang, Z., Wang, F., et al. (2017). Exosome-encapsulated microRNAs as circulating biomarkers for colorectal cancer. *Oncotarget* 8:60149. doi: 10.18632/oncotarget.18557

Yokoi, A., Matsuzaki, J., Yamamoto, Y., Yoneoka, Y., Takahashi, K., Shimizu, H., et al. (2018). Integrated extracellular microRNA profiling for ovarian cancer screening. *Nat. Commun.* 9:4319.

Yuan, F., Pan, X. Y., Zeng, T., Zhang, Y. H., Chen, L., Gan, Z. J., et al. (2020). Identifying cell-type specific genes and expression rules based on single-cell transcriptomic atlas data. *Front. Bioeng. Biotechnol.* 8:350. doi: 10.3389/fbioe.2020.00350

Zen, K., and Zhang, C. Y. (2012). Circulating microRNAs: a novel class of biomarkers to diagnose and monitor human cancers. *Med. Res. Rev.* 32, 326–348. doi: 10.1002/med.20215

Zhang, L., Liu, H., Zhu, L.-T., Luo, H.-J., Chen, X.-L., Yu, G.-Y., et al. (2020). Exosomal miRNAs as novel potential biomarkers for endometriosis. Research Square. Preprint.

Zhang, X., Du, L., Han, J., Li, X., Wang, H., Zheng, G., et al. (2020). Novel long non-coding RNA LINC02323 promotes epithelial-mesenchymal transition and metastasis via sponging miR-1343-3p in lung adenocarcinoma. *Thoracic Cancer* 11, 2506–2516. doi: 10.1111/1759-7714.13562

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying transcriptomic signatures and rules for SARS-CoV-2 infection. *Front. Cell Dev. Biol.* 8:627302. doi: 10.3389/fcell.2020.627302

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021b). Detecting the multiomics signatures of factor-specific inflammatory effects on airway smooth muscles. *Front. Genet.* 11:599970. doi: 10.3389/fgene.2020.599970

Zhao, C., Du, F., Zhao, Y., Wang, S., and Qi, L. (2019). Acute myeloid leukemia cells secrete microRNA-4532-containing exosomes to mediate normal hematopoiesis in hematopoietic stem cells by activating the LDOC1-dependent STAT3 signaling pathway. *Stem Cell Res. Ther.* 10, 1–12.

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

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2020a). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. *Bioinformatics* 36, 1391–1396.

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. *Bioinformatics* 36, 3568–3569. doi: 10.1093/bioinformatics/btaa166

Zhou, Y., Huang, T., Zhang, J., Wong, C. C., Zhang, B., Dong, Y., et al. (2017). TEAD1/4 exerts oncogenic role and is negatively regulated by miR-4269 in gastric tumorigenesis. *Oncogene* 36, 6518–6530. doi: 10.1038/onc.2017.257

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: a multi-label classifier for identifying metabolic pathway types of chemicals and enzymes with a heterogeneous network. *Comput. Math. Methods Med.* 2021:6683051.

**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: 24 March 2021
doi: 10.3389/fgene.2021.656526

# Identification of miRNA-Mediated Subpathways as Prostate Cancer Biomarkers Based on Topological Inference in a Machine Learning Process Using Integrated Gene and miRNA Expression Data

*Ziyu Ning[1,2†], Shuang Yu[1†], Yanqiao Zhao[1†], Xiaoming Sun[1], Haibin Wu[1] and Xiaoyang Yu[1*]*

[1] The Higher Educational Key Laboratory for Measuring and Control Technology and Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, Harbin, China, [2] School of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing, China

OPEN ACCESS

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

**Reviewed by:**
*Yang Chen,*
*Shantou University, China*
*Jian-Hua Zhang,*
*Peking University People's Hospital, China*

***Correspondence:**
*Xiaoyang Yu*
*xiyangyang_yu@163.com*

*† 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:** *21 January 2021*
**Accepted:** *02 March 2021*
**Published:** *24 March 2021*

**Citation:**
*Ning Z, Yu S, Zhao Y, Sun X, Wu H and Yu X (2021) Identification of miRNA-Mediated Subpathways as Prostate Cancer Biomarkers Based on Topological Inference in a Machine Learning Process Using Integrated Gene and miRNA Expression Data. Front. Genet. 12:656526. doi: 10.3389/fgene.2021.656526*

Accurately identifying classification biomarkers for distinguishing between normal and cancer samples is challenging. Additionally, the reproducibility of single-molecule biomarkers is limited by the existence of heterogeneous patient subgroups and differences in the sequencing techniques used to collect patient data. In this study, we developed a method to identify robust biomarkers (i.e., miRNA-mediated subpathways) associated with prostate cancer based on normal prostate samples and cancer samples from a dataset from The Cancer Genome Atlas (TCGA; $n$ = 546) and datasets from the Gene Expression Omnibus (GEO) database ($n$ = 139 and $n$ = 90, with the latter being a cell line dataset). We also obtained 10 other cancer datasets to evaluate the performance of the method. We propose a multi-omics data integration strategy for identifying classification biomarkers using a machine learning method that involves reassigning topological weights to the genes using a directed random walk (DRW)-based method. A global directed pathway network (GDPN) was constructed based on the significantly differentially expressed target genes of the significantly differentially expressed miRNAs, which allowed us to identify the robust biomarkers in the form of miRNA-mediated subpathways (miRNAs). The activity value of each miRNA-mediated subpathway was calculated by integrating multiple types of data, which included the expression of the miRNA and the miRNAs' target genes and GDPN topological information. Finally, we identified the high-frequency miRNA-mediated subpathways involved in prostate cancer using a support vector machine (SVM) model. The results demonstrated that we obtained robust biomarkers of prostate cancer, which could classify prostate cancer and normal samples. Our method outperformed seven other methods, and many of the identified biomarkers were associated with known clinical treatments.

**Keywords: machine learning, SVM, cancer, topological information, miRNA-mediated subpathway**

## INTRODUCTION

Prostate cancer is the second most commonly diagnosed cancer among males worldwide, and it is associated with miRNA dysfunction (Dankert et al., 2020). Prostate cancer is a highly heterogeneous disease with various mutations and tumor cell phenotypes (Peitzsch et al., 2020). The heterogeneity of prostate cancer causes difficulty regarding diagnosis and prognosis. Regarding treating prostate cancer patients, it is hoped that personalized medicine can be developed to mitigate the issues caused by the huge variations between different patient subgroups (Peng et al., 2017; Zhou et al., 2019).

The aim of this study was to identify robust biomarkers associated with prostate cancer, in the form of miRNA-mediated subpathways (miRNAs), and to evaluate the performance of our machine learning method based on other cancer datasets in addition to prostate cancer datasets. To identify cancer-related miRNAs to aid diagnosis and prognosis, high-throughput miRNA expression profiling has been used (Jay et al., 2007; Martens-Uzunova et al., 2012). Many studies have shown that miRNAs are stable not only in bodies but also in paraffin blocks (Baker, 2010). As miRNAs are promising biomarkers for cancer classification, several methods have been proposed to identify cancer biomarkers based on miRNA expression profiles, such as instance-based methods (Breiman et al., 1984; Breiman, 2001) and feature-based methods (Zararsiz et al., 2017; Peng et al., 2018). However, the performance of miRNA classification biomarkers in test sets varies greatly, even among patients with the same disease phenotype. Several factors, such as tissue heterogeneity, racial differences, and sequencing errors, contribute to this problem (Ning et al., 2019).

Many cancer-related pathways can be utilized as important classification biomarkers (Chen et al., 2021). For diagnosis prediction, pathway topological analysis can be used to identify risk classification biomarkers. Therefore, we integrated multiple types of data to identify the key miRNA-mediated subpathways of prostate cancer, which included data on the expression levels of miRNAs and their target genes and the topological weight of each gene in a global directed pathway network (GDPN). We employed a support vector machine (SVM)-based method to identify accurate risk biomarkers of prostate cancer based on the topological inference of miRNA-mediated subpathway activity. The method included five steps: merge pathways and construct network; perform directed random walk (DRW) (Liu et al., 2013); infer miRNA-mediated subpathway activity; select features and evaluate classification method; and obtain risk biomarkers. First, we obtained a dataset from The Cancer Genome Atlas (TCGA), a Gene Expression Omnibus (GEO) dataset, and a GEO cell line dataset, which together comprised 775 normal and human prostate cancer samples. Moreover, we also identified miRNA–target gene pairs in the TarBase v8.0 (Karagkouni et al., 2018) miRTarBase (Chou et al., 2018), and miRecords (Xiao et al., 2009) databases. Additionally, data on 4,090 samples in 10 other cancer TCGA datasets were downloaded from UCSC Xena. Thereafter, 343 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were merged into the GDPN, in which the nodes represented genes. Next, the method involved inferring the miRNA-mediated subpathway activity profile using a DRW-based method. Risk classification biomarkers (i.e., the high-frequency miRNA-mediated subpathways) were then identified using an SVM approach. Subsequently, we performed *within-dataset* analyses using the three prostate cancer datasets, and we identified the high-frequency miRNA-mediated subpathways in order to divide the samples into normal and cancer groups. We then evaluated the classification performance of these risk biomarkers in *cross-dataset* analyses using the prostate cancer datasets, followed by evaluating the performance in 10 other cancer datasets.

## MATERIALS AND METHODS

An overview of our biomarker identification method is shown in **Figure 1**. The method involves five major steps: merge pathways and construct network; perform DRW; infer miRNA-mediated subpathway activity; select features and evaluate classification method; and obtain risk markers. In addition, we transformed the gene expression profiles into an expression matrix, and we did the same for the miRNA expression profiles. In an expression matrix, each row refers to a miRNA/gene and each column refers to a sample. Next, we integrated the gene and miRNA expression data, the topological weights in the GDPN, and the miRNA–target gene pairs into an activity value. Consequently, we inferred an activity profile, in which each row and each column referred to one miRNA-mediated subpathway (miRNA) and one sample, respectively. After identifying biomarkers, the performance of our SVM model was evaluated using two validation GEO prostate cancer datasets and 10 other cancer datasets.

### Sample-Matched Datasets

We obtained three prostate cancer datasets, each of which included sample-matched gene and miRNA expression profiles. We downloaded one dataset ("PRAD-TCGA") from UCSC Xena[1], which involved sample-matched Illumina HiSeq level 3 gene and miRNA expression profiles. After removing the rows in which the expression values were equal to 0, we transformed all gene symbols to Entrez gene IDs. This resulted in 546 samples (52 normal and 494 cancer samples) with 12,118 genes and 209 miRNAs. To conduct an unbiased assessment of the performance of our method, we downloaded an independent dataset (GSE21036) (Taylor et al., 2010), which contained gene and miRNA expression profiles that were obtained using the GPL8227 microarray platform, from the GEO database[2]. We processed this dataset in the same way as the "PRAD-TCGA" dataset. It contained sample-matched data of 139 samples (28 normal and 111 cancer samples) with 18,941 genes and 373 miRNAs. To supplement this small independent validation GEO dataset, we downloaded a sample-matched cell line GEO dataset (GSE14794) on prostate cancer,

[1]https://xena.ucsc.edu/

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

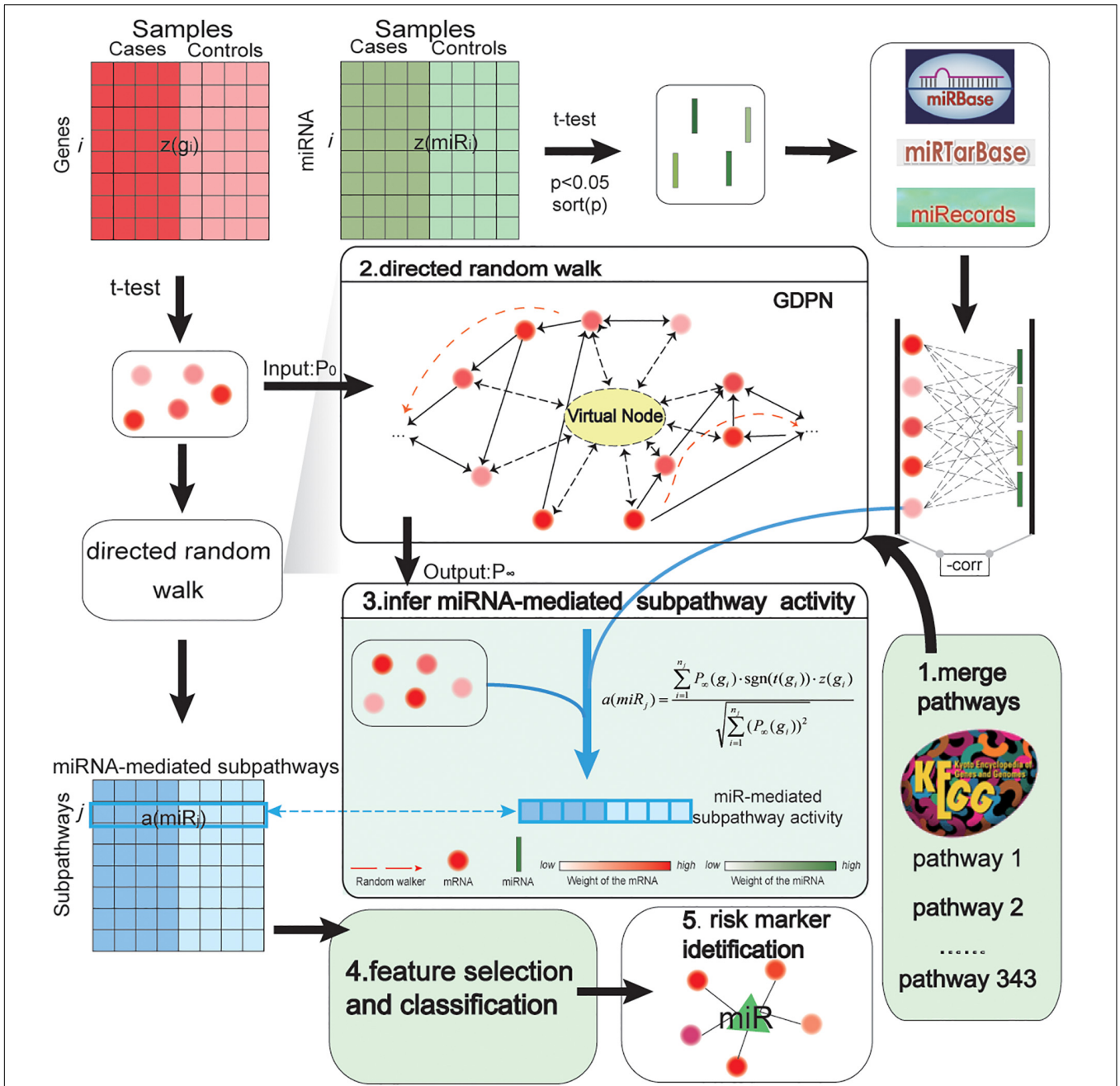


**FIGURE 1 |** Pipeline showing how miRNA-mediated subpathway activity profiles were inferred based on gene and miRNA expression and topological weights in a global directed pathway network (GDPN). The method to identify accurate risk biomarkers of prostate cancer involved five steps: merge pathways and construct network; perform directed random walk; infer miRNA-mediated subpathway activity; select features and evaluate classification method; and obtain risk biomarkers. $z(g_i)$ refers to a row-normalized vector of the expression of the $i$th gene over all samples, while $a(miR_j)$ refers to a row-normalized vector of the activity values of the $j$th miRNA (namely, the $j$th miRNA-mediated subpathway) over all samples. The middle part of the figure shows how the activity profiles were inferred. The 343 canonical KEGG pathways were merged into the GDPN. There are 39,930 directed edges and 7,159 nodes besides the virtual node in the GDPN. The virtual node is represented as a circle with a dotted line. $P_0$ refers to the matrix of the initial weights of all genes in the GDPN; $P_\infty$ refers to the final weights of all genes in the GDPN. To identify the important upstream genes, regarding assessing the direction of edges between genes, a gene was considered important if it influenced more downstream genes. The expression level and topological weights of the genes are integrated into $a(miR_j)$.

which contained data that were detected using the GPL6102 and GPL8178 platforms. We used 90 samples (45 control and 45 cancer samples) with 13,935 genes and 273 miRNAs (and no duplicates) from this dataset. Lastly, 10 other cancer TCGA datasets were downloaded from UCSC Xena, which involved 4,090 samples.

## miRNA–Target Genes Associated With Prostate Cancer

To identify the precise local subpathway regions associated with prostate cancer, we obtained reliable miRNA–target gene pairs from the following databases: TarBase v8.0 (Karagkouni et al., 2018), miRTarBase (Chou et al., 2018), and miRecords (Xiao et al., 2009). After removing the duplicates, there were 346,349 human-specific pairs, which consisted of 59 pairs from miRecords, 135,125 from TarBase, and 319,637 from miRTarBase. The specific interactions included two types of target relationships, which had been predicted based on calculations and verified by experiment.

## GDPN Construction

We obtained 343 canonical pathways from the KEGG database based on annotations of the differentially expressed target genes of the differentially expressed miRNAs. We then used the pathway interactions in the KEGG database to create a directed graph, and we merged this into a GDPN using "SubpathwayMiner" software[3] (Li et al., 2009). If genes appeared in diverse pathways, we merged them and kept the topological graphs. Finally, the GDPN included 39,930 directed edges and 7,159 gene nodes. Each edge direction could be traced back to the type of interaction between the pair of gene nodes according to the KEGG database, i.e., if gene P inhibited/activated gene Q, the edge direction pointed to gene Q. To ensure node weights flow in the network, we added a virtual node to the GDPN, with each node pointing to the virtual node and the virtual node pointing to all the nodes in the GDPN. We confirmed that the distributions of the GDPN node degree approximately followed power-law distributions, with $R^2 = 0.72$ (in-degree), 0.77 (out-degree), and 0.71 (total degree). Our method obeyed an important rule of the DRW algorithm, which involved having a low proportion of nodes that had higher degrees in the network (Watts and Strogatz, 1998).

## Performing DRW on the GDPN

The DRW algorithm simulated a walker that started at a source node and randomly stayed at the source or traveled to its neighbor node (Liu et al., 2013). New topological weights for the nodes in the GDPN were reassigned using the DRW algorithm, which is similar to the PageRank algorithm (Brin and Page, 1998). The PageRank algorithm is used by the Google search engine to search for related webpages; the higher the number of linkages that are directed toward a webpage, the more important it is. However, in our DRW algorithm, the direction of the linkages was reversed when compared to their direction in the PageRank algorithm, i.e., a gene that influenced more downstream genes was considered more important (Draghici et al., 2007). To calculate the new weights, the standard formula of the DRW algorithm was as follows:

$$P_{t+1} = (1 - r)M^T P_t + rP_0 \quad (1)$$

where $M^T$ is a row-normalized adjacency matrix (each element is divided by the sum of all elements in a row); $r \in [0, 1]$ is the restart probability ($r$ was set to 0.7), which slightly affected the result of the DRW algorithm (Kohler et al., 2008; Lv et al., 2015); and $P_0$ is a unit vector of the initial probabilities, which equaled $|t - score|$ (absolute value), generated based on $t$-test of normal vs. cancer samples. To start the process, $P_0$ was assigned to each GDPN node (the initial probability of the virtual node equaled 0) and several iterations were required until $|P_{t+1} - P_t| \leq 10^{-10}$. Eventually, $P_t$ converged to a stable state $P_\infty$, which was a vector of the new topological weights and was considered to represent the GDPN topological information.

## Inferring the Activity Profile From Gene Expression and GDPN Topological Information

For each differentially expressed miRNA between the normal and cancer samples, we determined their target genes and analyzed the differences in expression between the normal and cancer samples. Only target genes that were significantly differentially expressed ($t$-test $p$-value $< 0.05$) were used to infer the activity value of the miRNA-mediated subpathways. The significantly differentially expressed target genes $\{g_1, g_2, \cdots, g_{n_j}\}$ of miRNA $j$ ($miR_j$) were incorporated into an activity value, namely, miRNA-mediated subpathway activity $a(miR_j)$. In light of this, we have the following:

Constraint:

$$t(miR_j) \cdot t(g_i) < 0 \quad (2)$$

$$a(miR_j) = \frac{\sum_{i=1}^{n_j} P_\infty(g_i) \cdot sgn(t(g_i)) \cdot z(g_i)}{\sqrt{\sum_{i=1}^{n_j} (P_\infty(g_i))^2}} \quad (3)$$

where $t()$ is $|t - score|$of miRNAs or genes based on a $t$-test between the normal and cancer samples; $sgn()$ is a sign function {if $sgn[t(g_i)]$ is equal to a positive number, $sgn()$ returns +1, otherwise, it returns –1}; $z(g_i)$ is the normalized expression vector of gene $g_i$; $P_\infty(g_i)$ is the topological weight obtained by DRW; $a(miR_j)$ is the $j$th miRNA of the activity; and $n_j$ is the total number of significantly differentially expressed target genes. For Equation 3, Equation 2 is a constraint that ensures an inverse correlation between the expressions of miRNAs and their target genes. For example, to calculate the activity value of downregulated miRNAs, we integrated the expression of their upregulated target genes and the topological weights of their upregulated target genes into a special value. For upregulated miRNAs, we used the same method to calculate miRNA-mediated subpathway activity. Thus, the rows and columns indicated the miRNA-mediated subpathways (miRNAs) and samples, respectively and each value in the activity profile referred to the activity level of one miRNA in one sample.

## Evaluating Classification Performance

We performed fivefold cross-validation in *within-dataset* analyses of the "PRAD-TCGA," GSE21036, and GSE14794 datasets. In each of the three *within-dataset* analyses, we randomly split

[3]https://github.com/chunquanlipathway

the samples into five equal parts and selected four for training (*training* set) and one for testing (*test* set). Furthermore, the *training* set was randomly split into three equal parts, of which two (*training* subset) were used to build the classifiers and select candidate features, and the remaining one (*test* subset) was used to optimize the classifiers and select the risk biomarkers. First, we used a Student's *t*-test to obtain the *p*-values of the differences in the miRNA-mediated subpathway activities between the normal and prostate cancer samples in the *training* subset, and we sorted them by ascending *p*-value. We used the top 50 miRNA-mediated subpathways of the *training* subset as candidate biomarkers to establish the first classifier. The first classifier was built based on the candidate biomarker with the smallest *p*-value. Then, we added the candidate biomarker with the second-ranked *p*-value to it. If the area under the curve (AUC) increased, this candidate biomarker was kept in the risk biomarker set; otherwise, it was removed. We performed this process 50 times. We obtained the first optimized classifier and the average AUC from the first *test* subset. Thus, we could obtain three optimized classifiers from three *test* subsets. Then, we evaluated each of the three optimized classifiers by the five *test* sets in turn. We could obtain 15 AUCs. The experiment was repeated 10 times in each *within-dataset* analysis. We obtained the average AUC among the resulting 150 classifiers, which was used to represent the overall performance of our SVM-based method. Each SVM model was built using the "e1071" package in R (Meyer, 2013), which provides an R interface to libsvm. The functions "svm()" and "predict()" were used to build each SVM model and to predict the sample types, respectively. The "e1071" package was also used to perform the evaluation in the *cross-datasets* analyses.

Regarding the two *cross-datasets* analyses ("TCGA–GSE21036" and "TCGA–GSE14794"), we performed fivefold cross-validation, with the "PRAD-TCGA" dataset being used as the *training* set and GSE21036 or GSE14794 being used as the *test* set. We split the "PRAD-TCGA" samples into five equal parts and selected four for training (*training* subset) and the remaining one for testing (*test* subset). The validation process was similar to that in the *within-dataset* analyses. The *training* subsets were used to build the classifier and provide candidate biomarkers, and the *test* subset was used to optimize the classifier and select risk biomarkers. In each of the two *cross-datasets* analyses, five classifiers were optimized (optimizing the AUCs) by five *test* subsets in turn. The final performances of these classifiers were tested on the *test* set. For an unbiased assessment of the performance of our method, each validation experiment ("TCGA–GSE21036" and "TCGA–GSE14794") was repeated 10 times, and the average AUC was generated among the resulting 50 classifiers.

# RESULTS

## Inferred miRNA-Mediated Subpathway Activity Profile

Using the "PRAD-TCGA" dataset (before going on the do the same with the GSE21036 and GSE14794 GEO datasets), we employed Equation 1 to calculate the topological weight of each gene node in the GDPN. Equations 2, 3 were used to infer the miRNA-mediated subpathway (miRNA) activity profile. The rows and columns of the activity profile matrix refer to the miRNA-mediated subpathways (miRNAs) and samples, respectively. This matrix had 220 rows (miRNAs) and 546 columns (samples). We then computed the Student's *t*-test *p*-values for the miRNA-mediated subpathways in the activity profile and sorted them by ascending *p*-value. The top 50 miRNA-mediated subpathways were then used as candidate biomarkers and were subjected to SVM procedures using the "e1071" package in R. The number 50 was chosen based on the plateauing of the AUCs. To obtain validated risk biomarkers for prostate cancer and evaluate the performance of our method, fivefold cross-validation was performed 10 times. The AUC fluctuated between 0.848 and 0.998. We counted the frequency of each miRNA-mediated subpathway (miRNA) among the 150 SVM constructed classifiers and sorted by descending frequency. We only kept the miRNA-mediated subpathways with a frequency $> 50$ for further analysis, and we designated them as the high-frequency risk biomarkers. Thus, for the "PRAD-TCGA" dataset, 10 miRNA-mediated subpathways were identified as risk biomarkers. We used these 10 risk biomarkers to perform hierarchical clustering based on the miRNA-mediated subpathway activity profile (**Figure 2A**). Next, we obtained risk biomarkers based on the two GEO datasets and performed hierarchical clustering using their miRNA-mediated subpathway activity profiles (**Figures 2B,C**). **Figures 2A–C** show that the risk biomarkers of the various datasets could clearly separate normal and cancer samples.

The results indicated that our method could identify risk biomarkers that could be used to divide samples into normal and cancer groups (**Figures 2A–C**). Additionally, the risk biomarkers (miRNAs) were found to be related to prostate cancer, as indicated by data from the Human MicroRNA Disease Database (HMDD) (**Supplementary Table 1**-miRNAs).

## Integrating Topological Information Into the Activity Value

We identified risk biomarkers (miRNAs) based on the differences in miRNA-mediated subpathway activities between normal and cancer samples, which could help us to understand the biological mechanisms underlying cancer. We only considered target genes of miRNAs with $p < 0.05$ in the GDPN. The target genes of miRNAs with significantly differential expression were designated SDE target genes. To better understand the functions of the miRNA-mediated subpathways, these SDE target genes were annotated using KEGG pathways, and the pathways were sorted by ascending *p*-value. The pathways with a false discovery rate (FDR) $< 0.05$ and $p < 0.01$ (Benjamini and Hochberg method) were further analyzed. The SDE target genes of the 10 risk biomarkers (miRNA-mediated subpathways) in the "PRAD-TCGA" dataset were annotated with 299 KEGG pathways. Among these pathways, we selected the 41 pathways with 10 occurrences. Of these 41 pathways, 31 (82.93%) were related to prostate cancer, according to studies in PubMed (**Supplementary Table 1**-Pathways). The pathways associated with the SDE target genes of the 10 risk

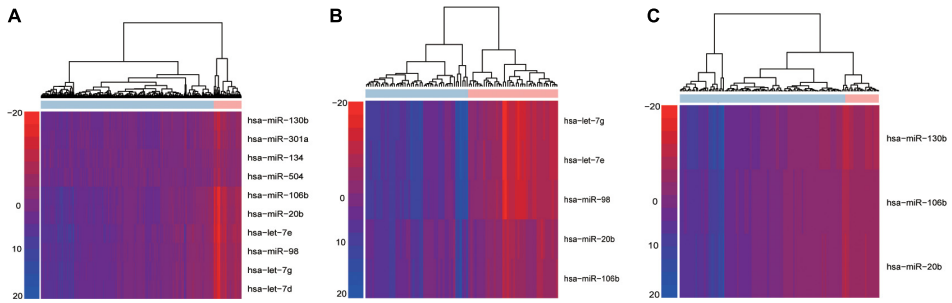


**FIGURE 2 |** Hierarchical cluster analysis of high-frequency miRNA-mediated subpathways in *within-dataset* analyses. Based on the results of 150 classifiers, the risk biomarkers identified in the **(A)** "PRAD-TCGA," **(B)** GSE14794, and **(C)** GSE21036 analyses were subjected to hierarchical cluster analysis. Rows and columns represent miRNA-mediated subpathways (miRNAs) and samples, respectively.

biomarkers included "Phosphatidylinositol-3-kinase (PI3K)-Akt signaling pathway" (hsa04151, **Figure 3A**), "mammalian target of rapamycin (mTOR) signaling pathway" (hsa04150), "mitogen-activated protein kinase (MAPK) signaling pathway" (hsa04010), and "cAMP signaling pathway" (hsa04024). The PI3K-Akt signaling pathway is a major research topic in prostate cancer treatment development, with more and more researchers paying close attention to it (Morgan et al., 2009; Toren and Zoubeidi, 2014). In prostate cancer, the activation of this pathway appears to be a characteristic of many aggressive cases, and this activation is observed more frequently as prostate cancer progresses to become a drug-resistant, metastatic disease (Toren and Zoubeidi, 2014). Many studies have shown that the PI3K-Akt signaling pathway plays a crucial role in prostate cancer metastasis and progression, along with the mTOR signaling pathway (Morgan et al., 2009; Bitting and Armstrong, 2013; Kim et al., 2017; Popolo et al., 2017). Moreover, the PI3K-Akt-mTOR and MAPK pathways can cooperate to facilitate prostate cancer growth and drug resistance (Shorning et al., 2020). Prostate cancer cell growth and invasion also involve the prostaglandin E2 receptor EP4 *via* the cAMP-PKA/PI3K-Akt signaling pathway (Xu et al., 2018). Thus, many studies have shown that oncogenic activation of the PI3K-Akt-mTOR pathway is a frequent event in prostate cancer that facilitates tumor formation, disease progression, and drug resistance (Shorning et al., 2020).

Moreover, we obtained new topological weights for the SDE target genes in the GDPN and sorted by descending weight. Among the top 100 SDE target genes (with topological importance), 22 genes were involved in the pathways with 10 occurrences regarding the SDE target genes of the 10 risk biomarkers (miRNA-mediated subpathways) in the "PRAD-TCGA" dataset. Several of the 22 genes are known to be important genes involved in prostate cancer. For example, coiled coil domain containing 6 (CCDC6) and DEAD-box RNA helicase p68 (DDX5) were annotated to "Pathways in cancer" (hsa05200) and "Transcriptional misregulation in cancer" (hsa05202), which are associated with cancer initiation and progression. Furthermore, eukaryotic translation initiation factor 4E (EIF4E) was annotated to the "PI3K-Akt signaling pathway" (hsa04151) and "mTOR signaling pathway" (hsa04150). CCDC6 protein turnover is regulated by the de-ubiquitinase USP7, which also controls androgen receptor (AR) stability (Criscuolo et al., 2019). Therefore, CCDC6 might be a predictive biomarker for the effectiveness of USP7 inhibitor and PARP inhibitor combination treatment in advanced prostate cancer (Criscuolo et al., 2019). DDX5 is an important AR transcriptional co-activator in prostate cancer and is overexpressed in late-stage disease (Clark et al., 2013). It is recruited to the AR transcriptional complex and required for the transcriptional regulation of AR-targeted genes (You et al., 2019). EIF4E plays a key role in protein synthesis and tumorigenesis (Xie et al., 2020). Regulation of EIF4E is partly achieved *via* phosphorylation (Furic et al., 2010). Ectopic expression of EIF4E prevented phenethyl isothiocyanate (PEITC)-induced translation inhibition and conferred significant protection against PEITC-induced apoptosis (Hu et al., 2007).

After discovering that hsa-miR-106b and hsa-miR-20b were the two biomarkers shared by the three datasets ("PRAD-TCGA," GSE21036, and GSE14794 datasets), we used KOBAS 3.0 to annotate their differentially expressed target genes with Gene Ontology (GO) terms. We then extracted the shared Gene Ontology (GO) terms (**Figure 3B**). The results showed that the differentially expressed target genes were associated with protein binding and cellular metabolic processes.

Finally, to assess the importance of the topological structure, 10, 20, 30, 40, and 50% of the miRNA–target gene pairs were randomly deleted. **Figure 3C** shows that the average AUC decreased as the percentage of deleted pairs increased. However, the average AUCs of our method remained stable when we deleted only 20% or only 30% in the "TCGA–GSE14794" and "TCGA–GSE21036" analysis, respectively. Lower stability was observed in the "TCGA–GSE14794" analysis, which might be caused by the fewer samples in the cell line dataset (GSE14794). The results indicate that stable performance might be best achieved by our method of integrating multi-omics data and topological weights, allowing risk biomarkers that can robustly classify samples to be identified.

## Our Method Applied to Prostate Cancer Datasets

To compare our method with other methods, we searched PubMed for studies involving similar methods, but there were no similar methods. However, we identified five classical methods

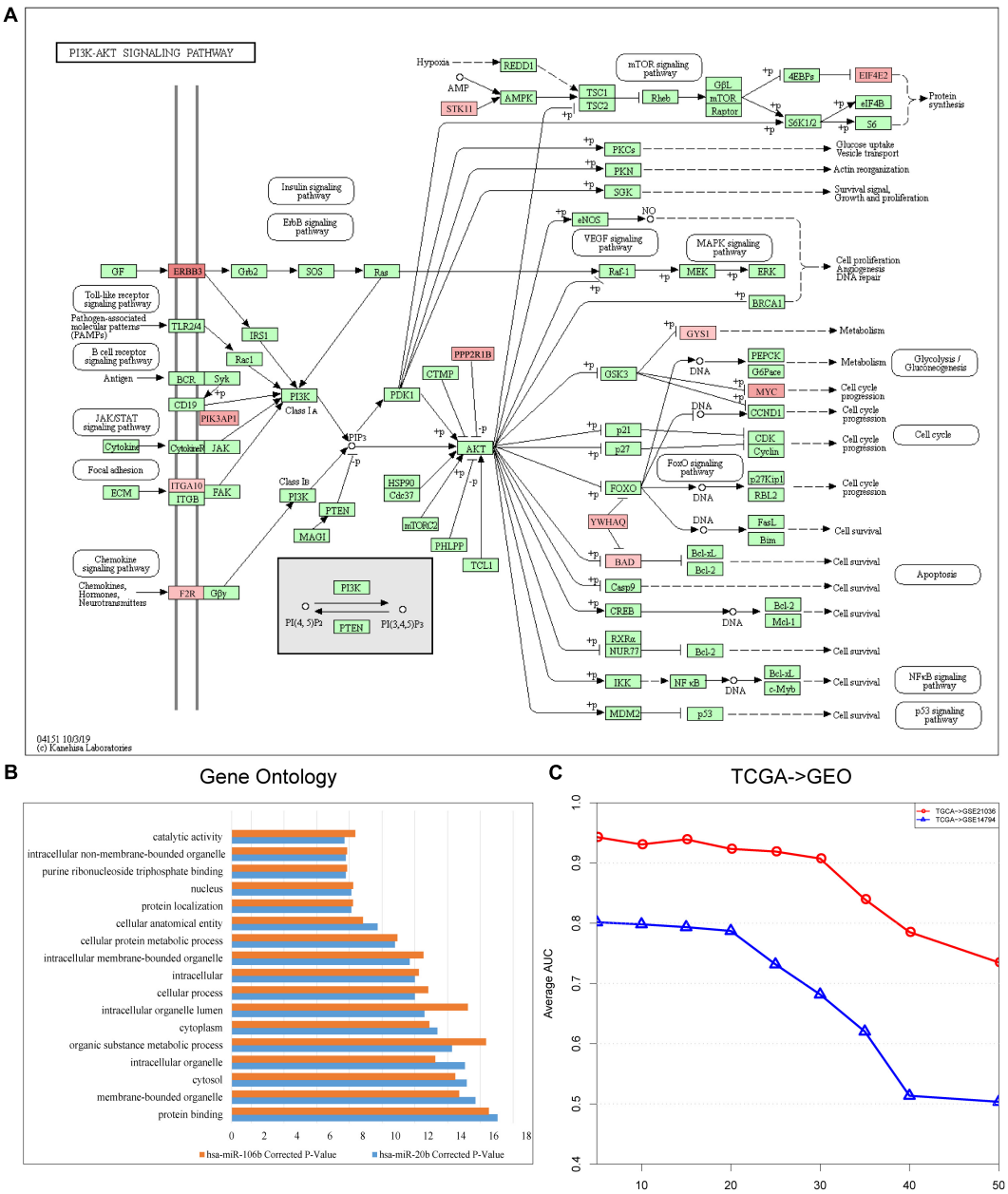


**FIGURE 3 | (A)** Landscape of PI3K-Akt signaling pathway (hsa04151). Red represents the target genes of hsa-miR-106b, and the intensity of red represents the level of differential expression. **(B)** Shared Gene Ontology (GO) terms of the target genes of hsa-miR-106b and hsa-miR-20b obtained by using KOBAS 3.0. **(C)** Line graph indicating the importance of the topological information. The x-axis refers to the percentage of deleted miRNA–target gene pairs; the y-axis refers to the corresponding AUCs.

to compare with our method, and these were pathway-based methods. They were the Mean method, Median method (Guo et al., 2005), component analysis (PCA) method (Bild et al., 2006), pathway activity inference using condition-responsive genes (PAC) method (Lee et al., 2008), and a previous DRW method (Liu et al., 2013). Furthermore, another two traditional methods, which classify samples based on single molecules (genes or miRNAs) were used in the comparison. The evaluation was performed similarly to the evaluation described by Lee et al. (Hu et al., 2007), who evaluated the classification performance of miRNA-mediated subpathways by fivefold cross-validation in a *within-dataset* analysis. To ensure an unbiased comparison, the SVM models were built based on the same datasets and evaluated based on the top 50 candidate biomarkers. As mentioned, in addition to pathway-based classification methods, two traditional classifiers were established based on genes and miRNAs, respectively; these classifiers were built with genes/miRNAs that belonged to the GDPN.

**Figure 4** depicts a summary of the average AUC and Accuracy in the *within-dataset* and *cross-datasets* analyses. The average AUC (0.9525) and Accuracy (0.9296) of our method in three *within-dataset* analyses ("PRAD-TCGA," GSE14794, and GSE21036 analyses) were calculated. We then examined the minimum standard deviation of the AUCs and Accuracies, which were 0.007 and 0.011 (**Figures 4A,B**), respectively. The average AUC and Accuracy outperformed the corresponding values for the pathway-based methods in the *within-dataset* analyses (**Supplementary Table 1**-Wilcoxon signed-rank test).

The results indicated that the miRNA-mediated subpathways could classify the sample phenotypes (normal vs. cancer). The results also revealed that our method was an effective strategy, integrating topological information into miRNA-mediated subpathway activities for sample classification. Thus, our classification biomarkers were more discriminative and stabler.

Furthermore, we performed *cross-datasets* analyses ("TCGA–GSE21036" and "TCGA–GSE14794") using the three prostate cancer datasets. The "PRAD-TCGA" dataset was used as the *training* set, and GSE21036 and GSE14794 were used as the *test* sets. The average AUC (Accuracy) in the "TCGA–GSE21036" and "TCGA–GSE14794" analyses were 0.9434 (0.9123) and 0.8015 (0.8903), respectively. For these *cross-datasets* analyses, the average AUC (Accuracy) of our SVM model was larger than corresponding values for the pathway-based, gene-based, and miRNA-based methods (**Figures 4C,D**). The average AUC (Accuracy) in the "TCGA–GSE14794" analyses compared to the "TCGA–GSE21036" analyses was slightly decreased, due to the imbalance between the *training* and *test* sets. Although the three prostate cancer datasets were obtained using different sequencing platforms and patient samples, the average AUC and Accuracy associated with our method were $> 0.80$. Therefore, our method detected accurate classification biomarkers (miRNA-mediated subpathways), which may be useful for diagnosis, and they had strong generalization ability and classification power.

## Evaluating Our SVM Model in Another 10 Cancer Datasets

So far, we have shown that the set of classification biomarkers (miRNA-mediated subpathways) work for prostate cancers. The consistency of these biomarkers was demonstrated by evaluating the SVM model both in *within-dataset* and *cross-datasets* analyses. Nevertheless, we had to consider whether the activity profile could be used to classify samples of other cancers. A total of 31 cancer datasets were downloaded from UCSC Xena (see text footnote 1). To avoid overfitting, we subjected 10 datasets to further analysis, each of which had $>200$ samples and sample-matched miRNA and gene expression profiles. We performed 10 fivefold cross-validation experiments on the 10 datasets (150 AUCs and Accuracies were calculated) and obtained the average AUCs and Accuracies (**Figures 5A,B**). To ensure an unbiased evaluation, the activity profiles in the 10 datasets were calculated to build the classifiers, and the frequency of each miRNA-mediated subpathway in each cancer was counted. The miRNA-mediated subpathways with frequency $>50$ were used as risk biomarkers for each cancer. A total of 56 risk biomarkers were used for further analysis, and most of them were related to specific cancers. Only eight miRNA-mediated subpathways occurred in $>5$ cancers, while 19 occurred in a single cancer only. The result implied that our method could detect cancer-specific risk biomarkers (miRNAs).

For example, hsa-let-7c, hsa-let-7i, hsa-let-7b, and hsa-let-7g occurred in seven, six, five, and five, respectively. Overwhelming evidence has demonstrated that the miRNA let-7 family (−a, −b, −c, −d, −e, −f, −g, and −i) plays regulatory roles at the transcriptional and post-transcriptional levels among various species, and their aberrant expression might be closely linked to the pathogenesis of cancers (Gibadulinova et al., 2020; Liu et al., 2020). According to our method, hsa-miR-191 appeared 130 times in the frequency list of liver hepatocellular carcinoma (LIHC), and it has been reported to play an important role in hepatocellular carcinoma (HCC) (Tian et al., 2019). Its overexpression reversed the anti-tumor effect of ANRIL on HepG2 cell proliferation, apoptosis, migration, and invasion (Huang et al., 2018). The hsa-miR-141 appeared in kidney renal clear cell carcinoma (KIRC) with high frequency, and it may play a crucial role in the diagnosis of kidney carcinoma. It also robustly discriminated between malignant and non-malignant tissues, and inhibiting it in normal renal proximal tubule epithelial cells (RPTEC) induced pro-cancerous characteristics (Dasgupta et al., 2020). It also acted as a potential biomarker for discriminating renal clear cell carcinoma (ccRCC) from normal tissues, and it acted as a crucial suppressor of ccRCC cell proliferation and metastasis by modulating the EphA2/p-FAK/p-AKT/MMPs signaling cascade (Chen et al., 2014).

The external evaluation in the 10 cancer datasets implied that practical risk biomarkers (miRNA-mediated subpathways) could be detected using our method. Moreover, the key target genes of the identified miRNAs were located in multi-pathway core regions. Identification of these regions provides opportunities to explore the interactions among genes, miRNAs, and pathways during cancer development.

## Case Study

Applying our method to the "PRAD-TCGA" dataset, we obtained 10 miRNA-mediated subpathways (miRNAs) and their 721 differently expressed target genes. The miRNA-mediated subpathway activity profile was inferred based on the SDE target gene expression and topological weights. Thus, the risk biomarkers (miRNA-mediated subpathways) had a stronger classification capacity.

We identified the shared miRNA-mediated subpathway biomarkers by comparing the results between the "PRAD-TCGA," GSE21036, and GSE14794 datasets. The hsa-miR-106b and hsa-miR-20b were the biomarkers that were shared among the three datasets. For the former miRNA, there were 273 miRNA–SDE target gene pairs, and for the latter, there were 277. The SDE target genes of these two miRNAs were annotated to several pathways, including "Metabolic pathways" (hsa01100) and "Protein processing in endoplasmic reticulum" (hsa04141). These pathways were ranked first and second, respectively, in the pathway list based on $p$-values (FDR $< 0.05$, Benjamini and Hochberg method). Previous gene expression research

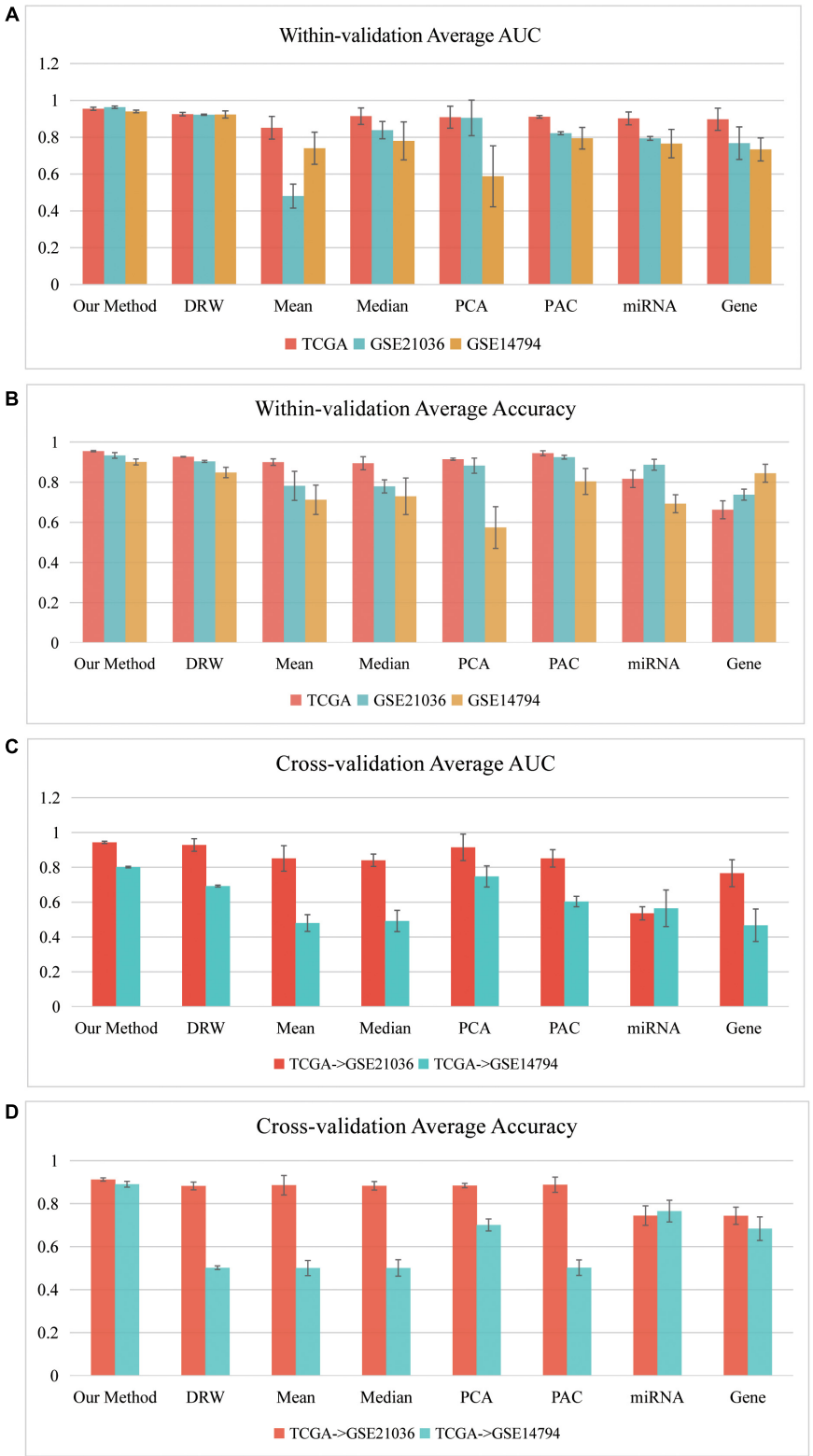


**FIGURE 4 |** Classification performances of our SVM model in *within-dataset* analyses. **(A)** Average AUC and **(B)** average Accuracy of the eight methods, including our method, which was calculated based on 150 SVM classifiers in each *within-dataset* analysis. **(C)** Average AUC and **(D)** average Accuracy of the eight methods, including our method, which was calculated based on 50 SVM classifiers in the "TCGA–GSE21036" and "TCGA–GSE14794" *cross-dataset* analyses.


**FIGURE 5 |** Classification performances of our SVM method based on 10 other cancers. **(A)** Average AUCs and **(B)** average Accuracies of the eight methods, including our method.

has shown that metastasis in prostate cancer is related to metabolic pathway dysregulation (Rhodes et al., 2002) and dysregulated transcriptional programs (LaTulippe et al., 2002). Research has also shown that, in androgen-independent prostate cancer cells, several small-molecule modulators of Sigma1 altered the endoplasmic reticulum (ER)-associated protein homeostasis pathways, including the unfolded protein response and autophagy (Maher et al., 2018). Furthermore, the SDE target genes of the two abovementioned miRNAs were also annotated to the "TGF-beta signaling pathway" (hsa04350), which was one of the top 10 pathways in the pathway list. Research has shown that induction of miR-106b plays a crucial role in the suppression of the proliferation of prostate cancer cells in a process that involves the TGF-beta signaling pathway (Zhang et al., 2012). Additionally, in human prostate cancer, miR-20b targets and downregulates TGFBR2, which in turn affects Smad2 activation and E2F1 expression, dysregulating the miR-20b-5p expression and contributing to TGF-β-induced epithelial-to-mesenchymal transition (Qi et al., 2019).

Notably, several miRNAs (such as hsa-miR-98, which was identified as a biomarker in the "PRAD-TCGA" analysis, and hsa-miR-301a, which was identified as a biomarker in both the "PRAD-TCGA" and GSE14794 analyses) play important roles in prostate cancer by regulating their target genes and thereby regulating pathways related to cancer. For example, hsa-miR-98 (which was annotated to 186 pathways) targeted nine differentially expressed genes in the "TGF-β signaling pathway" (hsa04350; E2F5, CDKN2B, MYC, BMP6, ACVR2B, SMAD7, ZFYVE16, RPS6KB2, and CHRD), and TGF-β affects multiple cellular responses *via* the canonical SMAD pathway and noncanonical pathways like the MAPK and PI3K-AKT pathways (Hamidi et al., 2017). Regarding E2F5, the E2F5/p38 axis plays a major role in uncontrolled prostate cancer cell proliferation *via* pSMAD3L activation, which provides strong support for using E2F5 as a biomarker for early detection of prostate cancer (Majumder et al., 2016). Additionally, regarding CDKN2B, upregulation of inhibitor of differentiation (Id1 and Id3) proteins attenuates all three cyclin-dependent kinase inhibitors (CDKN2B, -1A, and -1B), resulting in a more aggressive prostate cancer phenotype (Sharma et al., 2012). Moreover, regarding MYC, MYC-regulated fatty acid synthesis has been reported to be a valid target for treatment and/or prevention of prostate cancer. Not only are these three genes (E2F5, CDKN2B, and MYC) known to be associated with prostate cancer, but the six other target genes of hsa-miR-98 (BMP6, ACVR2B, SMAD7, ZFYVE16, RPS6KB2, and CHRD) have also been reported to be relevant to prostate cancer, according to previous studies (**Supplementary Table 1**-Genes). Moreover, hsa-miR-301a was annotated to the "p53 signaling pathway" (hsa04115), which can play a key role in the effectiveness of certain prostate cancer treatments. The ATM-CHEK2-p53 axis acts as a backbone for the DNA damage response (DDR) and is hypothesized to act as a barrier to cancer initiation (Stolarova et al., 2020). Significant associations with familial prostate cancer risk have been reported for both CHEK2 and ATM (Wokolorczyk et al., 2020).

The case study indicated that our SVM model could identify the key miRNAs, which may be useful as classification biomarkers for prostate cancer.

## DISCUSSION

Prostate cancer is a complicated cancer that has a high level of heterogeneity, many symptoms, and multiple subtypes.

In fact, prostate cancer has a lack of clear classification biomarkers because of its high heterogeneity and molecular instability. Clinically, age, prostate-specific antigen level, and Gleason score are generally used to diagnose this cancer among males. The miRNAs in blood and urine represent a convenient source of biomarkers for prostate cancer diagnosis and assessment of treatment efficacy due to their high stability and the low invasiveness of the sample collection process (Konoshenko et al., 2020).

Using machine learning to identify risk classification biomarkers of prostate cancer is a challenging task. With the increasing amount of high-throughput sequencing data, more and more miRNAs that are closely related to prostate cancer, such as hsa-miR-134 (Pelka et al., 2020), hsa-miR-504 (Katoh et al., 2013), and the hsa-let-7 family (Liu et al., 2012; Rong et al., 2020), have been discovered. Many researchers are now putting effort into identifying robust miRNA biomarkers of prostate cancer. However, there are no previously published sets of specific miRNA biomarkers for classifying samples into normal and prostate cancer groups, and the results of studies on miRNA biomarkers in other cancers often report conflicting results.

Due to advances in high-throughput multi-omics technologies, integrating multi-omics data into a special score is a promising approach to identifying biomarkers that can classify normal and cancerous samples. This integration strategy eliminates the dependence of machine learning on single types of data, such as gene or miRNA expression data. Thus, the approach provides an opportunity to detect robust classification biomarkers.

As is well known, the causes of cancers are complicated. Some researchers are convinced that biological pathways are disrupted by the target genes of miRNAs. Moreover, the target genes of single miRNAs can be found in several pathways. Pathways, miRNAs, and their target genes are involved in the occurrence, development, and metastasis of cancer, with the miRNAs playing a bridging role between pathways and genes. Also, many researchers believe that disease phenotypes are highly related to key local subpathways, rather than entire pathways (Li et al., 2013). We hold the opinion that our method is a promising way to detect classification biomarkers and to understand the biological mechanisms of cancer. The topological structure of biological networks should be considered when identifying risk biomarkers.

The purpose of our study was to identify robust classification biomarkers, and our method precisely classified samples. The method involved five steps: merge pathways and construct network; perform DRW; infer miRNA-mediated subpathway activity; select features and evaluate classification method; and obtain risk biomarkers. Each classification biomarker was composed of multiple types of integrated data, which included the GDPN topological information and the expression levels of miRNAs and their target genes. We reassigned new topological weights to the gene nodes using a DRW-based method. More topological weight was assigned to the gene nodes that exhibited topological importance. We amplified the signal of dysregulated hub genes and reassigned them larger topological weights because the expression of hub genes tends to vary only weakly between cases and controls (Lu et al., 2007). The miRNA-mediated subpathway activities were robust, and they may lead to better classification.

## 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

ZN and XY proposed and designed the method pipeline. ZN downloaded and filtered all datasets, performed the training process, and drafted the manuscript. SY, YZ, XS, HW, and XY revised the manuscript. All authors carefully examined and analyzed all datasets. The manuscript was reviewed by all authors.

## FUNDING

This work was supported by the National Natural Science Foundation of China (61571168 and 61671190), the Heilongjiang Postdoctoral Financial Assistance (LBH-Z19071), and the Fundamental Research Funds for the Universities in Heilongjiang Province (2018-KYYWF-1681).

## ACKNOWLEDGMENTS

We would like to thank all authors for their contributions to this project.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.656526/full#supplementary-material

**Supplementary Table 1 |** Details of results based on our SVM method. Genes: 22 key genes among the top 100 differentially expressed genes with high topological importance, and associated studies. miRNAs: High-frequency biomarkers (miRNA-mediated subpathways), and associated studies. Pathways: High-frequency pathways (with 10 occurrences) that the miRNA-mediated subpathways were annotated with, and associated studies. Wilcoxon signed-rank test: Wilcoxon signed-rank test results regarding the *within-dataset* and *cross-datasets* AUCs showing that our method outperformed the previous methods.

## REFERENCES


Baker, M. (2010). RNA interference: microRNAs as biomarkers. *Nature* 464:1227. doi: 10.1038/4641227a

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., et al. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 439, 353–357. doi: 10.1038/nature04296

Bitting, R. L., and Armstrong, A. J. (2013). Targeting the PI3K/Akt/mTOR pathway in castration-resistant prostate cancer. *Endocr. Relat. Cancer* 20, R83–R99. doi: 10.1530/ERC-12-0394

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32.

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

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comp. Netw. ISDN Syst.* 30, 107–117.

Chen, W., Jiang, J., Gong, L., Shu, Z., Xiang, D., Zhang, X., et al. (2021). Hepatitis B virus P protein initiates glycolytic bypass in HBV-related hepatocellular carcinoma via a FOXO3/miRNA-30b-5p/MINPP1 axis. *J. Exp. Clin. Cancer Res.* 40:1. doi: 10.1186/s13046-020-01803-8

Chen, X., Wang, X., Ruan, A., Han, W., Zhao, Y., Lu, X., et al. (2014). miR-141 is a key regulator of renal cell carcinoma proliferation and metastasis by controlling EphA2 expression. *Clin. Cancer Res.* 20, 2617–2630. doi: 10.1158/1078-0432.CCR-13-3224

Chou, C. H., Shrestha, S., Yang, C. D., Chang, N. W., Lin, Y. L., Liao, K. W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302. doi: 10.1093/nar/gkx1067

Clark, E. L., Hadjimichael, C., Temperley, R., Barnard, A., Fuller-Pace, F. V., and Robson, C. N. (2013). p68/DdX5 supports beta-catenin & RNAP II during androgen receptor mediated transcription in prostate cancer. *PLoS One* 8:e54150. doi: 10.1371/journal.pone.0054150

Criscuolo, D., Morra, F., Giannella, R., Cerrato, A., and Celetti, A. (2019). Identification of novel biomarkers of homologous recombination defect in DNA repair to predict sensitivity of prostate cancer cells to PARP-inhibitors. *Int. J. Mol. Sci.* 20:3100. doi: 10.3390/ijms20123100

Dankert, J. T., Wiesehofer, M., Wach, S., Czyrnik, E. D., and Wennemuth, G. (2020). Loss of RBMS1 as a regulatory target of miR-106b influences cell growth, gap closing and colony forming in prostate carcinoma. *Sci. Rep.* 10:18022. doi: 10.1038/s41598-020-75083-9

Dasgupta, P., Kulkarni, P., Majid, S., Hashimoto, Y., Shiina, M., Shahryari, V., et al. (2020). LncRNA *CDKN2B-AS1*/miR-141/cyclin D network regulates tumor progression and metastasis of renal cell carcinoma. *Cell Death Dis.* 11:660.

Draghici, S., Khatri, P., Tarca, L., Amin, K., Done, A., Voichita, C., et al. (2007). A systems biology approach for pathway level analysis. *Genome Res.* 17, 1537–1545. doi: 10.1101/gr.6202607

Furic, L., Rong, L., Larsson, O., Koumakpayi, I. H., Yoshida, K., Brueschke, A., et al. (2010). eIF4E phosphorylation promotes tumorigenesis and is associated with prostate cancer progression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 14134–14139. doi: 10.1073/pnas.1005320107

Gibadulinova, A., Bullova, P., Strnad, H., Pohlodek, K., Jurkovicova, D., Takacova, M., et al. (2020). CAIX-mediated control of LIN28/let-7 axis contributes to metabolic adaptation of breast cancer cells to hypoxia. *Int. J. Mol. Sci.* 21:4299. doi: 10.3390/ijms21124299

Guo, Z., Zhang, T., Li, X., Wang, Q., Xu, J., Yu, H., et al. (2005). Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* 6:58. doi: 10.1186/1471-2105-6-58

Hamidi, A., Song, J., Thakur, N., Itoh, S., Marcusson, A., Bergh, A., et al. (2017). TGF-beta promotes PI3K-AKT signaling and prostate cancer cell migration through the TRAF6-mediated ubiquitylation of p85alpha. *Sci. Signal.* 10:eaal4186. doi: 10.1126/scisignal.aal4186

Hu, J., Straub, J., Xiao, D., Singh, S. V., Yang, H. S., Sonenberg, N., et al. (2007). Phenethyl isothiocyanate, a cancer chemopreventive constituent of cruciferous vegetables, inhibits cap-dependent translation by regulating the level and phosphorylation of 4E-BP1. *Cancer Res.* 67, 3569–3573.

Huang, D., Bi, C., Zhao, Q., Ding, X., Bian, C., Wang, H., et al. (2018). Knockdown long non-coding RNA ANRIL inhibits proliferation, migration and invasion of HepG2 cells by down-regulation of miR-191. *BMC Cancer* 18:919. doi: 10.1186/s12885-018-4831-6

Jay, C., Nemunaitis, J., Chen, P., Fulgham, P., and Tong, A. W. (2007). miRNA profiling for diagnosis and prognosis of human cancer. *DNA Cell Biol.* 26, 293–300. doi: 10.1089/dna.2006.0554

Karagkouni, D., Paraskevopoulou, M. D., Chatzopoulos, S., Vlachos, I. S., Tastsoglou, S., Kanellos, I., et al. (2018). DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.* 46, D239–D245. doi: 10.1093/nar/gkx1141

Katoh, M., Igarashi, M., Fukuda, H., Nakagama, H., and Katoh, M. (2013). Cancer genetics and genomics of human FOX family genes. *Cancer Lett.* 328, 198–206. doi: 10.1016/j.canlet.2012.09.017

Kim, K. Y., Park, K. I., Kim, S. H., Yu, S. N., Park, S. G., Kim, Y. W., et al. (2017). Inhibition of autophagy promotes salinomycin-induced apoptosis via reactive oxygen species-mediated PI3K/AKT/mTOR and ERK/p38 MAPK-dependent signaling in human prostate cancer cells. *Int. J. Mol. Sci.* 18:1088. doi: 10.3390/ijms18051088

Kohler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* 82, 949–958. doi: 10.1016/j.ajhg.2008.02.013

Konoshenko, M. Y., Bryzgunova, O. E., Lekchnov, E. A., Amelina, E. V., Yarmoschuk, S. V., Pak, S. V., et al. (2020). The influence of radical prostatectomy on the expression of cell-free MiRNA. *Diagnostics* 10:600. doi: 10.3390/diagnostics10080600

LaTulippe, E., Satagopan, J., Smith, A., Scher, H., Scardino, P., Reuter, V., et al. (2002). Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.* 62, 4499–4506.

Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., and Lee, D. (2008). Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* 4:e1000217. doi: 10.1371/journal.pcbi.1000217

Li, C., Han, J., Yao, Q., Zou, C., Xu, Y., Zhang, C., et al. (2013). Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res.* 41:e101. doi: 10.1093/nar/gkt161

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., et al. (2009). SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.* 37:e131. doi: 10.1093/nar/gkp667

Liu, C., Kelnar, K., Vlassov, A. V., Brown, D., Wang, J., and Tang, D. G. (2012). Distinct microRNA expression profiles in prostate cancer stem/progenitor cells and tumor-suppressive functions of let-7. *Cancer Res.* 72, 3393–3404.

Liu, L., Wang, H., Yan, C., and Tao, S. (2020). An integrated analysis of mRNAs and miRNAs microarray profiles to screen miRNA signatures involved in nasopharyngeal carcinoma. *Technol. Cancer Res. Treat.* 19:1533033820956998. doi: 10.1177/1533033820956998

Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., et al. (2013). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* 29, 2169–2177. doi: 10.1093/bioinformatics/btt373

Lu, X., Jain, V. V., Finn, P. W., and Perkins, D. L. (2007). Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Mol. Syst. Biol.* 3:98. doi: 10.1038/msb4100138

Lv, Y., Wang, S., Meng, F., Yang, L., Wang, Z., Wang, J., et al. (2015). Identifying novel associations between small molecules and miRNAs based on integrated molecular networks. *Bioinformatics* 31, 3638–3644. doi: 10.1093/bioinformatics/btv417

Maher, C. M., Thomas, J. D., Haas, D. A., Longen, C. G., Oyer, H. M., Tong, J. Y., et al. (2018). Small-molecule Sigma1 modulator induces autophagic degradation of PD-L1. *Mol. Cancer Res.* 16, 243–255.

Majumder, S., Bhowal, A., Basu, S., Mukherjee, P., Chatterji, U., and Sengupta, S. (2016). Deregulated E2F5/p38/SMAD3 Circuitry reinforces the pro-tumorigenic switch of TGFbeta signaling in prostate cancer. *J. Cell. Physiol.* 231, 2482–2492. doi: 10.1002/jcp.25361

Martens-Uzunova, E. S., Jalava, S. E., Dits, N. F., van Leenders, G. J., Moller, S., Trapman, J., et al. (2012). Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene* 31, 978–991. doi: 10.1038/onc.2011.304

Meyer, D. (2013). Support Vector Machines the Interface to libsvm in package e1071. R News. 1, 1–3.

Morgan, T. M., Koreckij, T. D., and Corey, E. (2009). Targeted therapy for advanced prostate cancer: inhibition of the PI3K/Akt/mTOR pathway. *Curr. Cancer Drug Targets* 9, 237–249. doi: 10.2174/156800909787580999

Ning, Z., Feng, C., Song, C., Liu, W., Shang, D., Li, M., et al. (2019). Topologically inferring active miRNA-mediated subpathways toward precise cancer classification by directed random walk. *Mol. Oncol.* 13, 2211–2226. doi: 10.1002/1878-0261.12563

Peitzsch, C., Gorodetska, I., Klusa, D., Shi, Q., Alves, T. C., Pantel, K., et al. (2020). Metabolic regulation of prostate cancer heterogeneity and plasticity. *Semin Cancer Biol*. doi: 10.1016/j.semcancer.2020.12.002 [Epub ahead of print].

Pelka, K., Klicka, K., Grzywa, T. M., Gondek, A., Marczewska, J. M., Garbicz, F., et al. (2020). miR-96-5p, miR-134-5p, miR-181b-5p and miR-200b-3p heterogenous expression in sites of prostate cancer versus benign prostate hyperplasia-archival samples study. *Histochem. Cell Biol*. doi: 10.1007/s00418-020-01941-2 [Epub ahead of print].

Peng, L., Zhu, W., Liao, B., Duan, Y., Chen, M., Chen, Y., et al. (2017). Screening drug-target interactions with positive-unlabeled learning. *Sci. Rep.* 7:8087.

Peng, L. H., Yin, J., Zhou, L., Liu, M. X., and Zhao, Y. (2018). Human microbe-disease association prediction based on adaptive boosting. *Front. Microbiol.* 9:2440. doi: 10.3389/fmicb.2018.02440

Popolo, A., Pinto, A., Daglia, M., Nabavi, S. F., Farooqi, A. A., and Rastrelli, L. (2017). Two likely targets for the anti-cancer effect of indole derivatives from cruciferous vegetables: PI3K/Akt/mTOR signalling pathway and the aryl hydrocarbon receptor. *Semin. Cancer Biol.* 46, 132–137. doi: 10.1016/j.semcancer.2017.06.002

Qi, J. C., Yhang, Z., Zhang, Y. P., Lu, B. S., Yin, Y. W., Liu, K. L., et al. (2019). miR-20b-5p, TGFBR2, and E2F1 form a regulatory loop to participate in epithelial to mesenchymal transition in prostate cancer. *Front. Oncol.* 9:1535. doi: 10.3389/fonc.2019.01535

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.* 62, 4427–4433.

Rong, J., Xu, L., Hu, Y., Liu, F., Yu, Y., Guo, H., et al. (2020). Inhibition of let-7b-5p contributes to an anti-tumorigenic macrophage phenotype through the SOCS1/STAT pathway in prostate cancer. *Cancer Cell Int.* 20:470.

Sharma, P., Patel, D., and Chaudhary, J. (2012). Id1 and Id3 expression is associated with increasing grade of prostate cancer: Id3 preferentially regulates CDKN1B. *Cancer Med.* 1, 187–197. doi: 10.1002/cam4.19

Shorning, B. Y., Dass, M. S., Smalley, M. J., and Pearson, H. B. (2020). The PI3K-AKT-mTOR pathway and prostate cancer: at the crossroads of AR, MAPK, and WNT signaling. *Int. J. Mol. Sci.* 21:4507. doi: 10.3390/ijms21124507

Stolarova, L., Kleiblova, P., Janatova, M., Soukupova, J., Zemankova, P., Macurek, L., et al. (2020). CHEK2 Germline variants in cancer predisposition: stalemate rather than checkmate. *Cells* 9:2675. doi: 10.3390/cells9122675

Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., et al. (2010). Integrative genomic profiling of human prostate cancer. *Cancer Cell* 18, 11–22. doi: 10.1016/j.ccr.2010.05.026

Tian, F., Yu, C., Wu, M., Wu, X., Wan, L., and Zhu, X. (2019). MicroRNA-191 promotes hepatocellular carcinoma cell proliferation by has_circ_0000204/miR-191/KLF6 axis. *Cell Prolif.* 52:e12635. doi: 10.1111/cpr.12635

Toren, P., and Zoubeidi, A. (2014). Targeting the PI3K/Akt pathway in prostate cancer: challenges and opportunities (review). *Int. J. Oncol.* 45, 1793–1801. doi: 10.3892/ijo.2014.2601

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Wokolorczyk, D., Kluzniak, W., Huzarski, T., Gronwald, J., Szymiczek, A., Rusak, B., et al. (2020). Mutations in ATM, NBN and BRCA2 predispose to aggressive prostate cancer in Poland. *Int. J. Cancer* 147, 2793–2800. doi: 10.1002/ijc.33272

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.* 37, D105–D110. doi: 10.1093/nar/gkn851

Xie, J., Shen, K., Jones, A. T., Yang, J., Tee, A. R., Shen, M. H., et al. (2020). Reciprocal signaling between mTORC1 and MNK2 controls cell growth and oncogenesis. *Cell. Mol. Life Sci.* 78, 249–270.

Xu, S., Zhou, W., Ge, J., and Zhang, Z. (2018). Prostaglandin E2 receptor EP4 is involved in the cell growth and invasion of prostate cancer via the cAMPPKA/PI3KAkt signaling pathway. *Mol. Med. Rep.* 17, 4702–4712. doi: 10.3892/mmr.2018.8415

You, Z., Liu, C., Wang, C., Ling, Z., Wang, Y., Wang, Y., et al. (2019). LncRNA CCAT1 promotes prostate cancer cell proliferation by interacting with DDX5 and MIR-28-5P. *Mol. Cancer Ther.* 18, 2469–2479.

Zararsiz, G., Goksuluk, D., Korkmaz, S., Eldem, V., Zararsiz, G. E., Duru, I. P., et al. (2017). A comprehensive simulation study on classification of RNA-Seq data. *PLoS One* 12:e0182507. doi: 10.1371/journal.pone.0182507

Zhang, W., Edwards, A., Fan, W., Flemington, E. K., and Zhang, K. (2012). miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes. *PLoS One* 7:e40130. doi: 10.1371/journal.pone.0040130

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* 24:1714. doi: 10.3390/molecules24091714

**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: 13 April 2021
doi: 10.3389/fgene.2021.624128

# Deep Learning Enables Fast and Accurate Imputation of Gene Expression

*Ramon Viñas*[1]*, *Tiago Azevedo*[1], *Eric R. Gamazon*[2,3,4]* and *Pietro Liò*[1]*

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [2] Vanderbilt Genetics Institute and Data Science Institute, VUMC, Nashville, TN, United States, [3] MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom, [4] Clare Hall, University of Cambridge, Cambridge, United Kingdom

A question of fundamental biological significance is to what extent the expression of a subset of genes can be used to recover the full transcriptome, with important implications for biological discovery and clinical application. To address this challenge, we propose two novel deep learning methods, PMI and GAIN-GTEx, for gene expression imputation. In order to increase the applicability of our approach, we leverage data from GTEx v8, a reference resource that has generated a comprehensive collection of transcriptomes from a diverse set of human tissues. We show that our approaches compare favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime in two case studies and two imputation scenarios. In comparison conducted on the protein-coding genes, PMI attains the highest performance in inductive imputation whereas GAIN-GTEx outperforms the other methods in in-place imputation. Furthermore, our results indicate strong generalization on RNA-Seq data from 3 cancer types across varying levels of missingness. Our work can facilitate a cost-effective integration of large-scale RNA biorepositories into genomic studies of disease, with high applicability across diverse tissue types.

**Keywords: gene expression, transcriptomics, imputation, generative adversarial networks, machine learning, RNA-seq, GTEx, deep learning**

OPEN ACCESS

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

***Reviewed by:***
*Miguel Andrade,*
*Johannes Gutenberg University Mainz, Germany*
*Jidong Lang,*
*Geneis (Beijing) Co. Ltd, China*

***Correspondence:**
*Ramon Viñas*
*rv340@cam.ac.uk*
*Eric R. Gamazon*
*ericgamazon@gmail.com*
*Pietro Liò*
*pl219@cam.ac.uk*

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

***Received:*** *30 October 2020*
***Accepted:*** *12 March 2021*
***Published:*** *13 April 2021*

***Citation:***
*Viñas R, Azevedo T, Gamazon ER and Liò P (2021) Deep Learning Enables Fast and Accurate Imputation of Gene Expression. Front. Genet. 12:624128. doi: 10.3389/fgene.2021.624128*

## 1. INTRODUCTION

High-throughput profiling of the transcriptome has revolutionized discovery methods in the biological sciences. The resulting gene expression measurements can be used to uncover disease mechanisms (Emilsson et al., 2008; Cookson et al., 2009; Gamazon et al., 2018), propose novel drug targets (Evans and Relling, 2004; Sirota et al., 2011), provide a basis for comparative genomics (King and Wilson, 1975; Colbran et al., 2019), and motivate a wide range of fundamental biological problems. In parallel, methods that learn to represent the expression manifold can improve our mechanistic understanding of complex traits, with potential methodological and technological applications, including organs-on-chips (Low et al., 2020) and synthetic biology (Gupta and Zou, 2019), and the integration of heterogeneous transcriptomics datasets.

A question of fundamental biological significance is to what extent the expression of a subset of genes can be used to recover the full transcriptome with minimal reconstruction error. Genes that participate in similar biological processes or that have shared molecular function are likely to have similar expression profiles (Zhang and Horvath, 2005), prompting the question of gene expression prediction from a minimal subset of genes. Moreover, gene expression measurements may suffer

from unreliable values because some regions of the genome are extremely challenging to interrogate due to high genomic complexity or sequence homology (Conesa et al., 2016), further highlighting the need for accurate imputation. Moreover, most gene expression studies continue to be performed with specimens derived from peripheral blood or a convenient surrogate (e.g., lymphoblastoid cell lines; LCLs) due to the difficulty of collecting some tissues. However, gene expression may be tissue or cell-type specific, potentially limiting the utility of a proxy tissue.

The missing data problem can adversely affect downstream gene expression analysis. The simple approach of excluding samples with missing data from the analysis can lead to a substantial loss in statistical power. Dimensionality reduction approaches such as principal component analysis (PCA) and singular value decomposition (SVD) (Wall et al., 2003) cannot be applied to gene expression data with missing values. Clustering methods, a mainstay of genomics, such as *k*-means and hierarchical clustering may become unstable even with a few missing values (Troyanskaya et al., 2001).

To address these challenges, we develop two deep learning approaches to gene expression imputation. In both cases, we present an architecture that recovers missing expression data for multiple tissue types under different levels of missingness. In contrast to existing linear methods for deconfounding gene expression (Øystein Sørensen et al., 2018), our methods integrate covariates (global determinants of gene expression; Stegle et al., 2012) to account for their non-linear effects. In particular, a characteristic feature of our architectures is the use of word embeddings (Mikolov et al., 2013) to learn rich and distributed representations for the tissue types and other covariates. To enlarge the possibility and scale of a study's expression data (e.g., by including samples from highly inaccessible tissues), we train our model on RNA-Seq data from the Genotype-Tissue Expression (GTEx) project (The GTEx Consortium, 2015; GTEx Consortium, 2017), a reference resource (v8) that has generated a comprehensive collection of human transcriptome data in a diverse set of tissues.

We show that the proposed approaches compare favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime. In performance comparison on the protein-coding genes, GAIN-GTEx outperforms all the other methods in in-place imputation while PMI displays the highest performance in inductive imputation. Furthermore, we demonstrate that our methods are highly applicable across diverse tissues and varying levels of missingness. Finally, to analyse the cross-study relevance of our approach, we perform imputation on gene expression data from The Cancer Genome Atlas (TCGA; Weinstein et al., 2013) and show that our approach is robust when applied to independent RNA-Seq data.

# 2. METHODS

In this section, we introduce two deep learning approaches for gene expression imputation with broad applicability, allowing us to investigate their strengths and weaknesses in several realistic scenarios. Throughout the remainder of the paper, we use script letters to denote sets (e.g., $\mathcal{D}$), upper-case bold symbols to denote matrices or random variables (e.g., $\mathbf{X}$), and lower-case bold symbols to denote column vectors (e.g., $\mathbf{x}$ or $\bar{\mathbf{q}}_j$). The rest of the symbols (e.g., $\bar{q}_{jk}$, $G$, or $f$) denote scalar values or functions.

## 2.1. Problem Formulation

Consider a dataset $\mathcal{D} = \{(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$, where $\tilde{\mathbf{x}} \in \mathbb{R}^n$ represents a vector of gene expression values with missing components; $\mathbf{m} \in \{0,1\}^n$ is a mask indicating which components of the original vector of expression values $\mathbf{x}$ are missing or observed; $n$ is the number of genes; and $\mathbf{q} \in \mathbb{N}^c$ and $\mathbf{r} \in \mathbb{R}^k$ are vectors of $c$ categorical (e.g., tissue type or sex) and $k$ quantitative covariates (e.g., age), respectively. Our goal is to recover the original gene expression vector $\mathbf{x} \in \mathbb{R}^n$ by modeling the conditional probability distribution $P(\mathbf{X} = \mathbf{x} | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \mathbf{M} = \mathbf{m}, \mathbf{R} = \mathbf{r}, \mathbf{Q} = \mathbf{q})$, where the upper-case symbols denote the corresponding random variables.

## 2.2. Pseudo-Mask Imputation

We first introduce a novel imputation method named Pseudo-Mask Imputer (PMI).

**Formulation.** Let $\tilde{\mathbf{x}} = \mathbf{m} \odot \mathbf{x} \in \mathbb{R}^n$ be a vector of gene expression values whose missing components are indicated by a mask vector $\mathbf{m} \in \{0,1\}^n$. Our model is a function $f : \mathbb{R}^n \times \{0,1\}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that imputes the missing expression values $(\mathbf{1} - \mathbf{m}) \odot \mathbf{x}$ as follows:

$$\bar{\mathbf{x}} = f(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q}). \tag{1}$$

Here $\odot$ denotes element-wise multiplication. The recovered vector of gene expression values is then given by $\mathbf{m} \odot \tilde{\mathbf{x}} + (\mathbf{1} - \mathbf{m}) \odot \bar{\mathbf{x}}$.

**Optimization.** We optimize the model to maximize the imputation performance on a dynamic subset of observed, *pseudo-missing* components. In particular, we first generate a *pseudo-mask* $\tilde{\mathbf{m}}$ as follows:

$$\tilde{\mathbf{m}} = \mathbf{m} \odot \mathbf{b} \qquad \mathbf{b} \sim B(1, p) \qquad p \sim U(\alpha, \beta), \tag{2}$$

where $\mathbf{b} \in \{0,1\}^n$ is a vector sampled from a Bernoulli distribution $B$ and $\alpha \in [0,1]$ and $\beta \in [\alpha, 1]$ are hyperparameters that parameterize a uniform distribution $U$. Using the *pseudo-mask* $\tilde{\mathbf{m}}$, we split the observed expression values into a set of *pseudo-observed* components $\tilde{\mathbf{x}}$ and a set of *pseudo-missing* components $\tilde{\mathbf{y}}$:

$$\tilde{\mathbf{x}} = \mathbf{x} \odot \tilde{\mathbf{m}} \qquad \tilde{\mathbf{y}} = \mathbf{x} \odot \mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}}), \tag{3}$$

The imputed components are then given by $\bar{\mathbf{x}} = f(\tilde{\mathbf{x}}, \tilde{\mathbf{m}}, \mathbf{r}, \mathbf{q})$. We optimize our model to minimize the mean squared error between the ground truth and the imputed *pseudo-missing* components:

$$\mathcal{L}(\bar{\mathbf{x}}, \tilde{\mathbf{y}}, \mathbf{m}, \tilde{\mathbf{m}}) = \frac{1}{Z} \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)^\top (\bar{\mathbf{x}} - \tilde{\mathbf{y}})^2, \tag{4}$$

where $Z = \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)^\top \big(\mathbf{m} \odot (\mathbf{1} - \tilde{\mathbf{m}})\big)$ is a normalization term. We summarize our training algorithm in Algorithm 1.

Importantly, the *pseudo-mask* mechanism generates different sets of *pseudo-observed* components for each input example, effectively enlarging the number of training samples. Specifically, the hyperparameters $\alpha$ and $\beta$ control the fraction of *pseudo-observed* and *pseudo-missing* components through the probability $p \sim U(\alpha, \beta)$. On one hand, a low probability $p$ yields sparse *pseudo-observed* vectors $\hat{\mathbf{x}}$, resulting in fast convergence but high bias. On the other hand, a high probability $p$ yields denser *pseudo-observed* vectors $\hat{\mathbf{x}}$, resulting in low bias but slower convergence. At inference time, $p$ is set to 1 and the *pseudo-mask* $\tilde{\mathbf{m}}$ is equal to the input mask $\mathbf{m}$.

**Algorithm 1:** Training algorithm

**Input:** Input dataset $\mathcal{D} = \{(\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$, batch size $B$, hyperparameters $\alpha$ and $\beta$

- Initialise parameters of the model $f$

**while** *not convergence criteria reached* **do**

- Sample mini-batch: $\{(\mathbf{x}^{(i)}, \mathbf{m}^{(i)}, \mathbf{r}^{(i)}, \mathbf{q}^{(i)})\}_{i=1}^{B} \sim \mathcal{D}$
- Sample *pseudo-mask* for each example of the mini-batch: $p^{(i)} \sim U(\alpha, \beta)$; $\mathbf{b}^{(i)} \sim B(1, p^{(i)})$; $\tilde{\mathbf{m}}^{(i)} = \mathbf{m}^{(i)} \odot \mathbf{b}^{(i)}$
- Split components into *pseudo-observed* and *pseudo-missing*: $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}^{(i)} \odot \tilde{\mathbf{m}}^{(i)}$; $\tilde{\mathbf{y}}^{(i)} = \mathbf{x}^{(i)} \odot \mathbf{m}^{(i)} \odot (\mathbf{1} - \tilde{\mathbf{m}}^{(i)})$
- Impute *pseudo-missing* components: $\bar{\mathbf{x}}^{(i)} = f(\tilde{\mathbf{x}}^{(i)}, \tilde{\mathbf{m}}^{(i)}, \mathbf{r}^{(i)}, \mathbf{q}^{(i)})$
- Optimise the model by descending its stochastic gradient: $\nabla \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}(\bar{\mathbf{x}}^{(i)}, \tilde{\mathbf{y}}^{(i)}, \mathbf{m}^{(i)}, \tilde{\mathbf{m}}^{(i)})$

**end**

**Architecture.** We model the imputer $f$ as a neural network. We first describe how we use word embeddings, a distinctive feature that allows learning rich, dense representations for the different tissue types and, more generally, for all the covariates $\mathbf{q} \in \mathbb{N}^c$.

Formally, let $q_j$ be a categorical covariate (e.g., tissue type) with vocabulary size $v_j$, that is, $q_j \in \{1, 2, \ldots, v_j\}$, where each value in the vocabulary $\{1, 2, \ldots, v_j\}$ represents a different category (e.g., whole blood or kidney). Let $\bar{\mathbf{q}}_j \in \{0, 1\}^{v_j}$ be a one-hot vector such that $\bar{q}_{jk} = 1$ if $q_j = k$ and $\bar{q}_{jk} = 0$ otherwise. Let $d_j$ be the dimensionality of the embeddings for covariate $j$. We obtain a vector of embeddings $\mathbf{e}_j \in \mathbb{R}^{d_j}$ as follows:

$$\mathbf{e}_j = \bar{\mathbf{q}}_j^{\top} \mathbf{W}_j, \tag{5}$$

where each $\mathbf{W}_j \in \mathbb{R}^{v_j \times d_j}$ is a matrix of learnable weights. Essentially, this operation describes a lookup search in a dictionary with $v_j$ entries, where each entry contains a learnable $d_j$-dimensional vector of embeddings that characterize each of the possible values that $q_j$ can take. To obtain a global collection of embeddings $\mathbf{e}$, we concatenate all the vectors $\mathbf{e}_j$ for each categorical covariate $j$:

$$\mathbf{e} = \Big\|_{j=1}^{c} \mathbf{e}_j, \tag{6}$$

where $c$ is the number of categorical covariates and $\|$ represents the concatenation operator. We then use the learnable embeddings $\mathbf{e}$ in downstream tasks.

In terms of the architecture, we model $f$ as follows:

$$f(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q}) = \mathrm{MLP}(\tilde{\mathbf{x}} \| \mathbf{m} \| \mathbf{r} \| \mathbf{e}), \tag{7}$$

where MLP denotes a multilayer perceptron and $\tilde{\mathbf{x}} = \mathbf{x} \odot \mathbf{m} \in \mathbb{R}^n$ is the masked gene expression. **Figure 1** shows the architecture of the model.

## 2.3. Generative Adversarial Imputation Networks

The second method, which we call GAIN-GTEx, is based on Generative Adversarial Imputation Nets (GAIN; Yoon et al., 2018). Generative Adversarial Networks have previously been used to synthesize transcriptomics *in-silico* (Marouf et al., 2020; Viñas et al., 2021), but to our knowledge their applicability to gene expression imputation is yet to be studied. Similar to generative adversarial networks (GANs; Goodfellow et al., 2014), GAIN estimates a generative model via an adversarial process driven by the competition between two players, the *generator* and the *discriminator*.

**Generator.** The generator aims at recovering missing data from partial gene expression observations, producing samples from the conditional $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Formally, we define the generator as a function $G : \mathbb{R}^n \times \mathbb{R}^n \times \{0, 1\}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that imputes expression values as follows:

$$\bar{\mathbf{x}} = G(\mathbf{x} \odot \mathbf{m}, \mathbf{z} \odot (\mathbf{1} - \mathbf{m}), \mathbf{m}, \mathbf{r}, \mathbf{q}), \tag{8}$$

where $\mathbf{z} \in \mathbb{R}^n$ is a vector sampled from a fixed noise distribution. Similar to GAIN, we mask the $n$-dimensional noise vector as $\mathbf{z} \odot (\mathbf{1} - \mathbf{m})$, encouraging a bijective association between noise components and genes. Before passing the output $\bar{\mathbf{x}}$ to the discriminator, we replace the prediction for the non-missing components by the original, observed expression values:

$$\hat{\mathbf{x}} = \mathbf{m} \odot \tilde{\mathbf{x}} + (\mathbf{1} - \mathbf{m}) \odot \bar{\mathbf{x}}. \tag{9}$$

**Discriminator.** The discriminator takes the imputed samples $\hat{\mathbf{x}}$ and attempts to distinguish whether the expression value of each gene has been observed or produced by the generator. This is in contrast to the original GAN discriminator, which receives information from two input streams (generator and data distribution) and attempts to distinguish the true input source.

Formally, the discriminator is a function $D : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that outputs the probabilities $\hat{\mathbf{y}} \in \mathbb{R}^n$:

$$\hat{\mathbf{y}} = D(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{r}, \mathbf{q}), \tag{10}$$

where the $i$-th component $\hat{y}_i$ is the probability of gene $i$ being observed (as opposed to being imputed by the generator) for

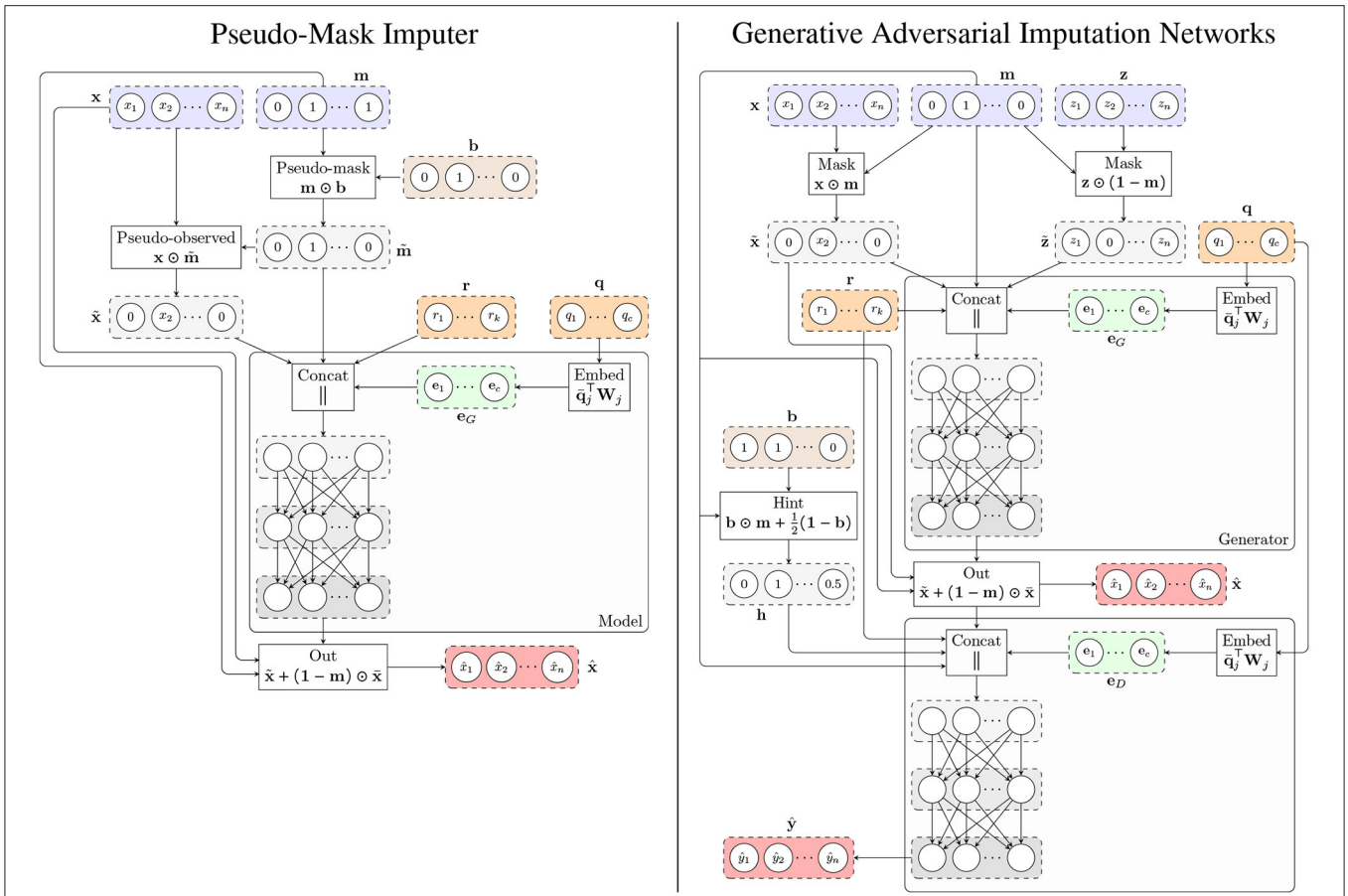


**FIGURE 1 |** Architecture of the proposed methods. **(Left)** Pseudo-Mask Imputer (PMI). **(Right)** Generative Adversarial Imputation Networks for GTEx (GAIN-GTEx). In both cases, the imputer takes gene expression values $\tilde{\mathbf{x}}$ with missing components according to a mask $\mathbf{m}$, and categorical (e.g., tissue type; $\mathbf{q}$) and numerical (e.g., age; $\mathbf{r}$) covariates, and outputs the imputed values $\bar{\mathbf{x}}$. The observed components of the imputer's output are then replaced by the actual observed expression values $\tilde{\mathbf{x}}$, yielding the imputed sample $\hat{\mathbf{x}}$. For PMI, the *pseudo-mask* mechanism masks out some of the observed components, which are then recovered at the output. For the adversarial model (right), we additionally train a discriminator that receives $\hat{\mathbf{x}}$, the sample covariates, and the hint vector $\mathbf{h}$, and produces the output $\hat{\mathbf{y}}$, whose $i$-th component $\hat{y}_i$ represents the probability of gene $i$ being observed as opposed to being imputed by the generator.

each $i \in \{1, \ldots, n\}$ and the vector $\mathbf{h} \in \mathbb{R}^n$ corresponds to the *hint* mechanism described in Yoon et al. (2018), which provides theoretical guarantees on the uniqueness of the global minimum for the estimation of $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Concretely, the role of the hint vector $\mathbf{h}$ is to *leak* some information about the mask $\mathbf{m}$ to the discriminator. Similar to GAIN, we define the hint $\mathbf{h}$ as follows:

$$\mathbf{h} = \mathbf{b} \odot \mathbf{m} + \frac{1}{2}(\mathbf{1} - \mathbf{b}) \qquad \mathbf{b} \sim B(1, p) \qquad p \sim U(\alpha, \beta), \quad (11)$$

where $\mathbf{b} \in \{0, 1\}^n$ is a binary vector that controls the amount of information from the mask $\mathbf{m}$ revealed to the discriminator. In contrast to GAIN, which discloses all but one components of the mask, we sample $\mathbf{b}$ from a Bernoulli distribution parametrized by a random probability $p \sim U(\alpha, \beta)$, where $\alpha \in [0, 1]$ and $\beta \in [\alpha, 1]$ are hyperparameters. This accounts for a high number of genes $n$ and allows to trade off the number of mask components that are revealed to the discriminator.

**Optimization.** Similarly to GAN and GAIN, we optimize the generator and discriminator adversarially, interleaving gradient updates for the discriminator and generator.

The discriminator aims at determining whether genes have been observed or imputed based on the imputed vector $\hat{\mathbf{x}}$, the covariates $\mathbf{q}$ and $\mathbf{r}$, and the hint vector $\mathbf{h}$. Since the hint vector $\mathbf{h}$ readily provides partial information about the ground truth $\mathbf{m}$ (Equation 11), we penalize $D$ only for genes $i \in \{1, 2, \ldots, n\}$ such that $h_i = 0.5$, that is, genes whose corresponding mask value is unavailable to the discriminator. We achieve this via the following loss function $\mathcal{L}_D : \{0, 1\}^n \times \mathbb{R}^n \times \{0, 1\}^n \to \mathbb{R}$:

$$\mathcal{L}_D(\mathbf{m}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z}(\mathbf{1} - \mathbf{b})^\top \big(\mathbf{m} \odot \log \hat{\mathbf{y}} + (\mathbf{1} - \mathbf{m}) \odot (\mathbf{1} - \log \hat{\mathbf{y}})\big), \quad (12)$$

where $Z = 1 + (\mathbf{1} - \mathbf{b})^\top(\mathbf{1} - \mathbf{b})$ is a normalization term. The only difference with respect to the binary cross entropy loss function is the dot product involving $(\mathbf{1} - \mathbf{b})$, which we

employ to ignore genes whose mask has been *leaked* to the discriminator through **h**.

The generator aims at implicitly estimating $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Therefore, its role is not only to impute the expression corresponding to missing genes, but also to reconstruct the expression of the observed inputs. Similar to GAIN, in order to account for this and encourage a realistic imputation function, we use the following loss function $\mathcal{L}_G : \{0,1\}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \{0,1\}^n \rightarrow \mathbb{R}$ for the generator:

$$\mathcal{L}_G(\mathbf{m}, \mathbf{x}, \bar{\mathbf{x}}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z_1}\big((\mathbf{1}-\mathbf{b})\odot(\mathbf{1}-\mathbf{m})\big)^\top \log \hat{\mathbf{y}} + \frac{\lambda}{Z_2}\mathbf{m}^\top(\mathbf{x}-\bar{\mathbf{x}})^2, \tag{13}$$

where $Z_1 = 1+(\mathbf{1}-\mathbf{b})^\top(\mathbf{1}-\mathbf{b})$ and $Z_2 = \mathbf{m}^\top\mathbf{m}$ are normalization terms, and $\lambda > 0$ is a hyperparameter. Intuitively, the first term in Equation (13) corresponds to the adversarial loss, whereas the mean squared error (MSE) term accounts for the loss that the generator incurs in the reconstruction of the observed gene expression values.

**Architecture.** We model the discriminator $D$ and the generator $G$ using neural networks. Similar to PMI, $D$ and $G$ leverage independent instances $\mathbf{e}^G$ and $\mathbf{e}^D$ of the categorical embeddings described in Equation (6). Specifically, we model the two players as follows:

$$\begin{aligned} G(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \mathbf{m}, \mathbf{r}, \mathbf{q}) &= \text{MLP}(\tilde{\mathbf{x}}\|\tilde{\mathbf{z}}\|\mathbf{m}\|\mathbf{r}\|\mathbf{e}^G) \\ D(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{r}, \mathbf{q}) &= \text{MLP}(\hat{\mathbf{x}}\|\mathbf{h}\|\mathbf{r}\|\mathbf{e}^D), \end{aligned} \tag{14}$$

where MLP denotes a multilayer perceptron and $\tilde{\mathbf{x}} = \mathbf{x}\odot\mathbf{m} \in \mathbb{R}^n$ and $\tilde{\mathbf{z}} = \mathbf{z}\odot(\mathbf{1}-\mathbf{m}) \in \mathbb{R}^n$ are the masked gene expression and noise input vectors, respectively. **Figure 1** shows the architecture of both players.

# 3. EXPERIMENTAL DETAILS

In this section, we provide an overview of the dataset and describe the experimental details, including all the different case studies and imputation scenarios that we considered. We also describe the implementation details of PMI (see **Supplementary Figure 6**) and GAIN-GTEx (see **Supplementary Figure 7**).

## 3.1. Materials

**Dataset.** The GTEx dataset is a public genomic resource of genetic effects on the transcriptome across a broad collection of human tissues, enabling linking of these regulatory mechanisms to trait and disease associations (Aguet et al., 2020). Our dataset contained 15,201 RNA-Seq samples collected from 49 tissues of 838 unique donors. We also selected the intersection of all the protein-coding genes among these tissues, yielding 12,557 unique genes. In addition to the expression data, we leveraged metadata about the sample donors, including sex, age, and cohort (post-mortem, surgical, or organ donor).

**Standardization.** A large proportion of gene expression data in public repositories contains normalized values. Thus, imputation in this context has practical utility. Imputing the relative expression levels (in normalized data) vs absolute levels (in non-normalized data) is also biologically meaningful, with important applications, e.g., differential expression analysis (between disease individuals and controls) that is robust to expression outliers. To this end, we normalized the expression data via the standard score, so that the standardized expression values have mean 0 and standard deviation 1 for each gene across all samples.

**Training, validation, and test splits.** To prevent any leakage of information between the training and test sets, we enforced all samples from the same donor to be within the same set. Concretely, we first flipped the GTEx donor identifiers (e.g., 111CU-1826 is flipped to 6281-UC111), we then sorted the reversed identifiers in alphabetical order, and we finally selected a suitable split point, forcing the two sets to be disjoint. After splitting the data, the training set, which we used to train the model, consisted of $\sim$ 60% of the total samples. The validation set, which we used to optimize the method, consisted of $\sim$ 20% of the total samples. The test set, on which we evaluated the final performance, contained the remaining $\sim$ 20% of the data.

## 3.2. Case Studies

We benchmarked the methods on two case studies:

**Case 1: Protein-coding genes.** As a first case study, we selected the intersection of all the protein-coding genes among the 49 GTEx tissues, resulting in a set of 12,557 unique genes. This case study is challenging for imputation methods that are not scalable across the number of input variables.

**Case 2: Genes in a pathway.** We selected a subset of 273 genes from the Alzheimer's disease pathway extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000). This case study allows to benchmark imputation methods that do not scale well with the number of variables.

## 3.3. Imputation Scenarios

We considered two realistic imputation scenarios:

**Scenario 1: In-place imputation.** Our goal is to impute the missing values of a dataset $\mathcal{D} = \{(\mathbf{m}\odot\mathbf{x}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$ without access to the ground truth missing values $(\mathbf{1}-\mathbf{m})\odot\mathbf{x}$. Importantly, for this scenario we assumed that the data is *missing completely at random* (MCAR; Little and Rubin, 2019), that is, the missingness does not depend on any of the observed nor unobserved variables.

**Scenario 2: Inductive imputation.** Given a training dataset $\mathcal{D}_{train} = \{(\mathbf{x}, \mathbf{1}, \mathbf{r}, \mathbf{q})\}$ where all expression values $\mathbf{x} \in \mathbb{R}^n$ are observed, our goal is to impute the missing expression values of an independent test dataset $\mathcal{D}_{test} = \{(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$. Methods trained in inductive mode (e.g., on comprehensive datasets such as GTEx) can be used to perform imputation on small, independent datasets where the small number of samples is insufficient to train a model in in-place mode.

## 3.4. Implementation

For both PMI and GAIN-GTEx, we included the donor's age as numerical covariate in **r** and the tissue type, sex and cohort as categorical covariates in **q**. We normalized the numerical variables via the standard score. For each categorical variable

**TABLE 1 |** Gene expression imputation performance with a missing rate of 50% across 3 runs (complete set of protein-coding genes).

| | Scenario 1: In-place imputation | | Scenario 2: Inductive imputation | |
|---|---|---|---|---|
| **Method** | $R^2$ | **Runtime (hours)** | $R^2$ | **Runtime (hours)** |
| MICE | – | – | – | – |
| MissForest | – | – | – | – |
| Blood surrogate | $-0.693 \pm 0.000$ | $0.000 \pm 0.000$ | $-0.952 \pm 0.000$ | $0.000 \pm 0.000$ |
| Median imputation | $0.000 \pm 0.000$ | $0.001 \pm 0.000$ | $-0.009 \pm 0.000$ | $0.001 \pm 0.000$ |
| 1-NN imputation | $0.179 \pm 0.000$ | $1.616 \pm 0.004$ | $0.203 \pm 0.000$ | $0.985 \pm 0.003$ |
| 5-NN imputation | $0.461 \pm 0.000$ | $2.224 \pm 0.107$ | $0.482 \pm 0.000$ | $1.441 \pm 0.096$ |
| 10-NN imputation | $0.468 \pm 0.000$ | $2.140 \pm 0.035$ | $0.495 \pm 0.000$ | $1.711 \pm 0.160$ |
| GAIN-MSE-GTEx | $0.637 \pm 0.005$ | $0.199 \pm 0.074$ | $0.638 \pm 0.003$ | $0.456 \pm 0.053$ |
| GAIN-GTEx | $0.638 \pm 0.007$ | $0.625 \pm 0.294$ | $0.636 \pm 0.001$ | $1.199 \pm 0.157$ |
| PMI | $0.479 \pm 0.003$ | $0.241 \pm 0.024$ | $0.707 \pm 0.001$ | $0.244 \pm 0.019$ |

*We do not report the $R^2$ scores for MICE and MissForest, because the runtime is longer than 7 days. Note GAIN-GTEx outperforms all the other methods in in-place imputation while PMI displays the highest performance in inductive imputation.*

$q_j \in \{1, 2, \ldots, v_j\}$, we used the rule of thumb $d_j = \lfloor \sqrt{v_j} \rfloor + 1$ to set all the dimensions of the categorical embeddings. We used ReLU activations for each hidden layer in the MLP architectures of both PMI and GAIN (see Equations 7 and 14).

We trained both models using the Adam optimizer (Kingma and Ba, 2014). We used batch normalization (Ioffe and Szegedy, 2015) in the hidden layers of the MLPs, which yielded a significant speed-up to the training convergence according to our experiments. We used early stopping with a patience of 30. The rest of parameters for each model, case study, and imputation scenario are presented in the **Supplementary Material**.

## 3.5. Baseline Methods

We compared PMI and GAIN-GTEx to several baseline methods:

**Common methods of imputation.** We considered two simple gene expression imputation approaches: blood surrogate and median imputation. The use of blood, an easily accessible tissue, as a surrogate for difficult-to-acquire tissues is done in studies of biomarker discovery, diagnostics, and eQTLs, and in the development of model systems (Gamazon et al., 2018; Kim et al., 2020). For blood surrogate imputation, we imputed missing gene expression values in any given tissue with the corresponding values in whole blood for the same donor. For median imputation, we imputed missing values with the median of the observed tissue-specific gene expression computed across donors.

***k*-Nearest Neighbours**. The *k*-Nearest Neighbours (*k*-NN) algorithm is an efficient method that is commonly used for imputation (Beretta and Santaniello, 2016). Here, we leveraged *k*-NN as a baseline for different values of *k*. This model estimates the missing values of a sample based on the values of the missing components in the *k* closest samples.

**State-of-the-art methods.** We considered two state-of-the-art imputation methods: Multivariate Imputation by Chained Equations (MICE; Buuren and Groothuis-Oudshoorn, 2010) and MissForest (Stekhoven and Bühlmann, 2012). MICE leverages chained equations to create multiple imputations of missing data. The hyperparameters of MICE include the minimum/maximum possible imputed value for each component and the maximum number of imputation rounds. MissForest (Stekhoven and Bühlmann, 2012) is a non-parametric imputation method based on random forests trained on observed values to impute the missing values. Among others, the hyperparameters of MissForest include the number of trees in the forest and the number of features to consider when looking for the optimal split.

# 4. RESULTS

Here we provide an overview of the imputation results, including a comparison with other imputation methods, an evaluation of the tissue-specific results, and an analysis of the cross-study relevance across different levels of missingness.

## 4.1. Comparison

**Tables 1**, **2** show a quantitative summary of the imputation performances for the two case-studies and the two imputation scenarios. In addition to the imputation scores, we report the runtime of all the methods. We labeled methods as computationally *unfeasible* when they took longer than 7 days to run on our server (CPU: Intel(R) Xeon(R) Processor E5-2630 v4. RAM: 125GB), after which we halted the execution. For example, MICE and MissForest were unfeasible for each imputation scenario on the complete set of protein-coding genes. An empirical study of the scalability of both methods (see **Supplementary Material**) showed that the runtime increases rapidly with the number of genes. However, on a smaller set of genes (i.e., 273 from the Alzheimer's disease pathway), evaluation of the performance was successfully obtained, with the runtime substantially higher for both methods than for the other methods. In addition, we included GAIN-MSE-GTEx as a baseline, consisting of a simplification of GAIN-GTEx that was optimized exclusively via the mean squared error term of the generator. GAIN-MSE-GTEx performed reasonably well relative to GAIN-GTEx, suggesting that the mean squared error term of the loss function was driving the learning (see **Supplementary Material**).

**TABLE 2 |** Gene expression imputation performance with a missing rate of 50% across 3 runs (for a subset of 273 genes from the Alzheimer's disease pathway).

| | Scenario 1: In-place imputation | | Scenario 2: Inductive imputation | |
|---|---|---|---|---|
| **Method** | $R^2$ | **Runtime (hours)** | $R^2$ | **Runtime (hours)** |
| MICE | 0.574 ± 0.001 | 2.062 ± 0.335 | 0.569 ± 0.001 | 2.252 ± 0.096 |
| MissForest (1 tree) | −0.147 ± 0.002 | 0.145 ± 0.002 | −0.042 ± 0.003 | 0.575 ± 0.167 |
| MissForest (10 trees) | 0.458 ± 0.001 | 0.839 ± 0.176 | 0.514 ± 0.001 | 3.220 ± 0.371 |
| MissForest (20 trees) | 0.478 ± 0.000 | 1.836 ± 0.068 | 0.540 ± 0.000 | 4.842 ± 0.495 |
| MissForest (100 trees) | 0.493 ± 0.000 | 6.438 ± 0.498 | 0.561 ± 0.001 | 16.186 ± 1.709 |
| Blood surrogate | −0.698 ± 0.002 | 0.000 ± 0.000 | −0.971 ± 0.002 | 0.000 ± 0.000 |
| Median imputation | 0.001 ± 0.000 | 0.000 ± 0.000 | −0.009 ± 0.000 | 0.000 ± 0.000 |
| 1-NN imputation | 0.186 ± 0.001 | 0.037 ± 0.001 | 0.301 ± 0.000 | 0.021 ± 0.001 |
| GAIN-MSE-GTEx | 0.519 ± 0.001 | 0.038 ± 0.002 | 0.533 ± 0.001 | 0.045 ± 0.004 |
| GAIN-GTEx | 0.533 ± 0.001 | 0.139 ± 0.041 | 0.527 ± 0.003 | 0.569 ± 0.017 |
| PMI | 0.536 ± 0.001 | 0.048 ± 0.002 | 0.630 ± 0.011 | 0.037 ± 0.002 |

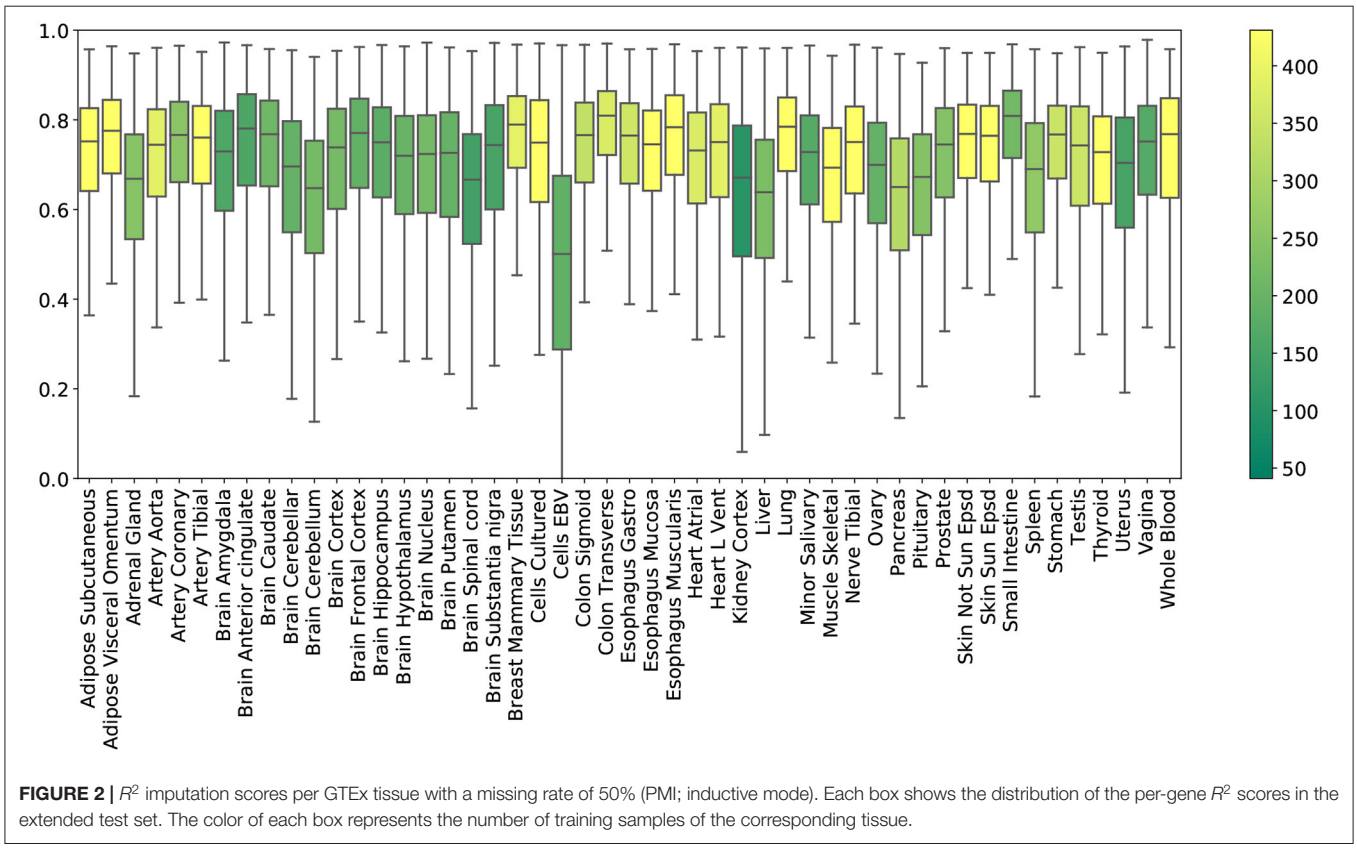


**FIGURE 2 |** $R^2$ imputation scores per GTEx tissue with a missing rate of 50% (PMI; inductive mode). Each box shows the distribution of the per-gene $R^2$ scores in the extended test set. The color of each box represents the number of training samples of the corresponding tissue.

In terms of the evaluation metrics, we report the coefficient of determination ($R^2$). This metric ranges from $-\infty$ to 1 and corresponds to the ratio of explained variance to the total variance. Negative scores indicate that the model predictions are worse than those of a baseline model that predicts the mean of the data. Here, to evaluate the performance, we generated random masks with a missing rate of 50% and computed the imputation $R^2$ per gene. We repeated the last step 3 times and reported the overall mean $R^2$ and the average per-gene standard deviation of the $R^2$ scores, averaged across the 3 runs. In inductive mode, blood surrogate and median imputation exhibited negative scores. Under in-place imputation on the protein-coding genes, GAIN-GTEx outperformed all the other methods ($0.638 \pm 0.007$). Under inductive imputation on the protein-coding genes, PMI showed the best overall performance ($0.707 \pm 0.001$) among all the methods.

## 4.2. Imputation Results

**Tissue-specific results.** **Figure 2** shows the $R^2$ scores achieved by PMI across all 49 tissue types. To obtain these results, we

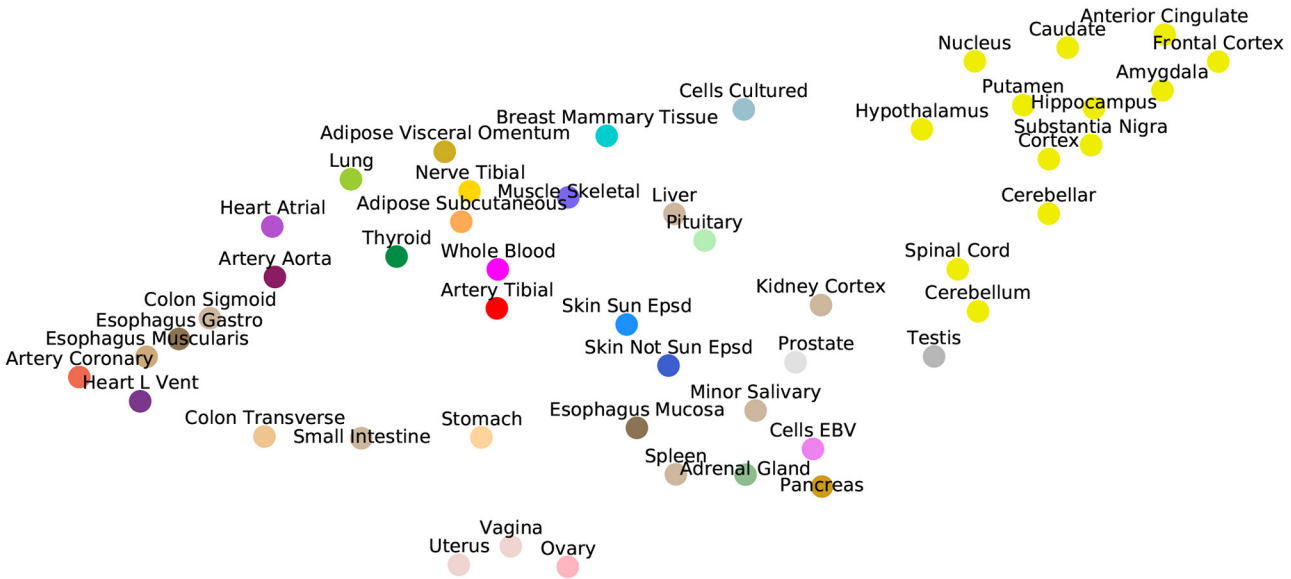


**FIGURE 3 |** UMAP visualization of the tissue embeddings from the generator. Colors are assigned to conform to the GTEx Consortium conventions. Note that the central nervous system, consisting of the 13 brain regions, clusters together on the top right corner.

**TABLE 3 |** Cross-study results for GAIN-GTEx and PMI trained on GTEx (inductive mode).

| GAIN-GTEx | | PMI | |
|---|---|---|---|
| **Tissue** | $R^2$ | **Tissue** | $R^2$ |
| TCGA LAML | $0.386 \pm 0.057$ | TCGA LAML | $0.394 \pm 0.065$ |
| TCGA BRCA | $0.408 \pm 0.023$ | TCGA BRCA | $0.427 \pm 0.023$ |
| TCGA LUAD | $0.439 \pm 0.034$ | TCGA LUAD | $0.451 \pm 0.050$ |
| GTEx Whole blood | $0.678 \pm 0.031$ | GTEx Whole blood | $0.709 \pm 0.034$ |
| GTEx Breast | $0.724 \pm 0.036$ | GTEx Breast | $0.751 \pm 0.039$ |
| GTEx Lung | $0.713 \pm 0.033$ | GTEx Lung | $0.744 \pm 0.035$ |

*We report the $R^2$ scores on data from 3 TCGA cancer types and their healthy counterpart on GTEx for a missing rate of 50%.*

generated random masks with a missing rate of 50% for the test set, performed imputation, and plotted the distribution of 12,557 gene $R^2$ scores for each tissue. Mean $R^2$ scores in the individual tissues ranged from $\sim$ 0.5 (Epstein Barr virus transformed lymphocytes; EBV) to $\sim$ 0.78 (small intestine). Kidney cortex, the tissue with the smallest sample size, had the highest variability in $R^2$ with an interquartile range of $Q_3 - Q_1 = 0.30$.

**Figure 3** illustrates the ability of GAIN-GTEx to learn rich tissue representations. Specifically, we plotted a UMAP representation (McInnes et al., 2018) of the learnt tissue embeddings $\mathbf{W}_j \in \mathbb{R}^{49\times 8}$ from the generator (see Equation 5), where $j$ indexes the tissue dimension. Strikingly, the tissue representations showed strong clustering of biologically-related tissues, including the central nervous system (i.e., the 13 brain regions), the gastrointestinal system (e.g., the esophageal and colonic tissues), and the female reproductive tissues (i.e., uterus, vagina, and ovary). The clustering properties were robust across UMAP runs and could be similarly appreciated using other dimensionality reduction algorithms such as tSNE (Maaten and Hinton, 2008).

**Cross-study results across missing rates.** To evaluate the cross-study relevance and generalizability of PMI and GAIN-GTEx, we leveraged the model trained on GTEx to perform imputation on The Cancer Genome Atlas (TCGA) gene expression data in acute myeloid leukemia (TCGA LAML; Cancer Genome Atlas Research Network et al., 2013), breast cancer (TCGA BRCA; Cancer Genome Atlas Network, 2012), and lung adenocarcinoma (TCGA LUAD; Cancer Genome Atlas Research Network, 2014). For each TCGA tissue and its *non-diseased* test counterpart on GTEx, we show the imputation quality in **Table 3** as well as the performance across varying missing rates in **Figure 4**.

**Imputation results on genes from the Alzheimer's disease pathway.** **Figure 5** shows the per-gene imputation scores for GAIN-GTEx trained on a subset of 273 genes corresponding to the Alzheimer's disease pathway. Amyloid-beta is a core element of senile plaques which are characteristic of the debilitating disease, with various pathophysiological consequences on cellular processes. The pathway consists of genes that are involved in a number of processes, including neuronal apoptosis, autophagy deficits, mitochondrial defect, and neurodegeneration. We observed that some genes in the pathway (e.g., PSMB6, COX6C, PSMD7, PSMA2, PSMD14, SDHB, TUBB1, TUBA8, FZD9, LPL, KIF5C, TUBB4A, TUBB2B, APOE) exhibited different distributions between brain and non-brain tissue types. The most highly imputed genes were enriched in known gene sets (see **Supplementary Figures 9, 10**).

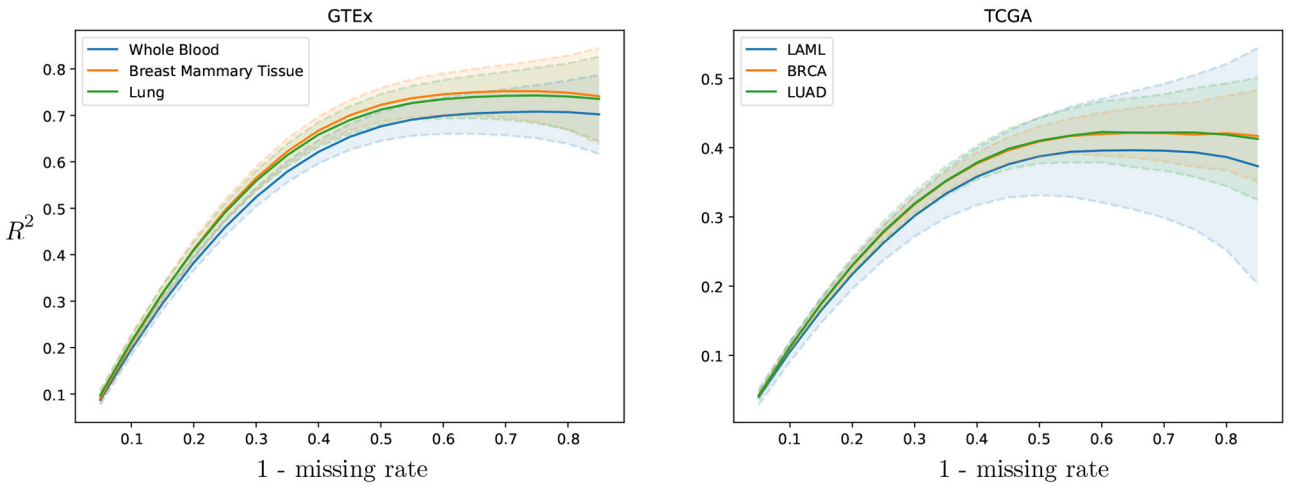


**FIGURE 4 |** GAIN-GTEx $R^2$ imputation scores per tissue across missing rate for 3 TCGA cancer types and their healthy counterpart in GTEx. The shaded area represents one standard deviation of the per-gene $R^2$ scores in the corresponding tissue. The greater the rate of missingness, the lower the performance.

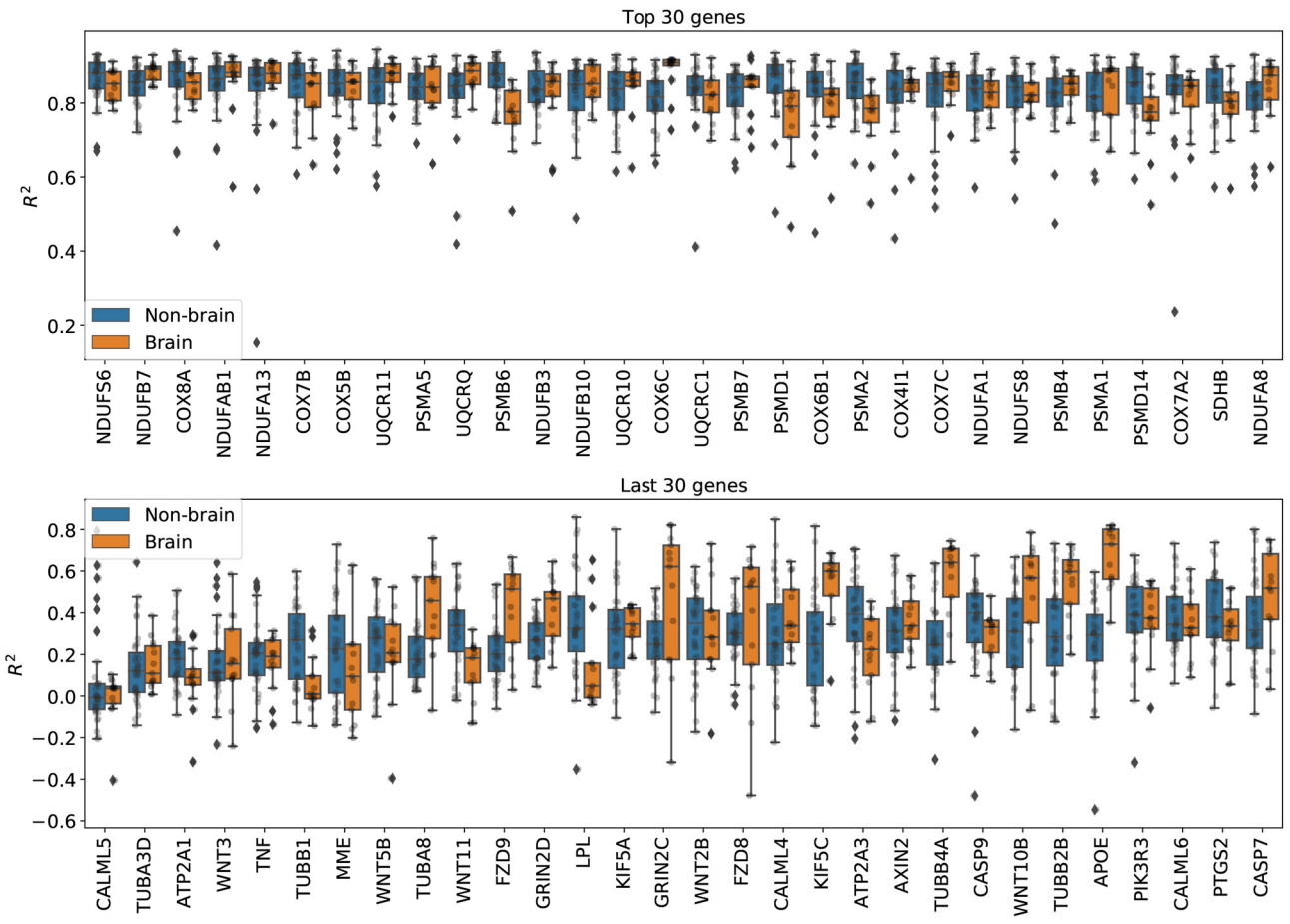


**FIGURE 5 |** Per-gene imputation $R^2$ scores on genes from the Alzheimer's disease pathway. Each point represents the average $R^2$ score in a tissue type. We note that some genes in the pathway (e.g., PSMB6, COX6C, PSMD7, PSMA2, PSMD14, SDHB, TUBB1, TUBA8, FZD9, LPL, KIF5C, TUBB4A, TUBB2B, APOE) exhibited different distributions between brain and non-brain tissue types.

## 5. DISCUSSION

We developed two imputation approaches to gene expression, facilitating the reconstruction of a high-dimensional molecular trait that is central to disease biology and drug target discovery. The proposed methods, which we called Pseudo-Mask Imputer (PMI) and GAIN-GTEx, were able to approximate the gene expression manifold from incomplete gene expression data and relevant covariates (potential global determinants of expression) and impute missing expression values. A characteristic feature of our architectures is the use of word embeddings, which enabled to learn distributed representations of the tissue types (see **Figure 3**). Importantly, this allowed to condition the imputation algorithms on factors that drive gene expression, endowing the architectures with the ability to represent them in a biologically meaningful way.

We leveraged the most comprehensive human transcriptome resource available (GTEx), allowing us to test the performance of our method in a broad collection of tissues (see **Figure 2**). The biospecimen repository includes commonly used surrogate tissues (whole blood and EBV transformed lymphocytes), central nervous system tissues from 13 brain regions, and a wide diversity of other primary tissues from *non-diseased* individuals. In particular, we observed that EBV transformed lymphocytes, an accessible and renewable resource for functional genomics, are a notable outlier in imputation performance. This is perhaps not surprising, consistent with studies about the transcriptional effect of EBV infection on the suitability of the cell lines as a model system for primary tissues (Carter et al., 2002). Interestingly, similar tissues exhibit similar $R^2$ scores (see **Supplementary Figure 12**).

We analyzed the performance of the proposed approaches and found that they compare favorably to several existing imputation methods in terms of imputation performance and runtime (see **Table 1**). We observed that standard approaches such as leveraging the expression of missing genes from a surrogate blood tissue yielded negative $R^2$ values and therefore did not perform well. Median imputation, although easy to implement, had a very limited predictive power. Imputation methods based on $k$-Nearest Neighbours were computationally feasible and yielded solid but poorer $R^2$ scores. In terms of state-of-the-art-methods, MICE and MissForest were computationally prohibitive given the high-dimensionality of the data and we halted the execution after running our experiments for 7 days. In particular, we performed an empirical study of the scalability of both methods (see **Supplementary Figures 1–5**) and observed that the runtime increases very rapidly with the number of genes. To alleviate this issue, we compared PMI and GAIN-GTEx with these methods on a subset of 273 genes from the Alzheimer's disease pathway (see **Table 2**). Under the in-place imputation scenario (Alzheimer's disease pathway), MICE performed better than PMI, GAIN-GTEx, and MissForest (100 trees). Under the inductive imputation setting, PMI outperformed all the other methods by a large margin.

In terms of the comparison between PMI and GAIN-GTEx, our experiments suggest that the latter is generally harder to optimize (see hyperparameter search in **Supplementary Material**). In particular, GAIN resembles a deep autoencoder in that the supervised loss penalizes the reconstruction error of the observed components. While this is a natural choice, autoencoder-like architectures are considerably sensitive to the user-definable bottleneck dimension. On one hand, a small number of units results in under-fitting. On the other hand, an excessively big bottleneck dimension allows the neural network to trivially *copy-paste* the observed components. In contrast, the loss function of PMI does not penalize the reconstruction error for the *pseudo-observed* components (e.g., the loss function of PMI penalizes the prediction error of the *pseudo-missing* components, which are not provided as input at training time). Together with the fact that the *pseudo-mask* mechanism dynamically enlarges the training size, this subtlety allows training considerably bigger networks without over-fitting. Finally, we observed that a simplification of GAIN-GTEx, GAIN-MSE-GTEx, performed similarly well, suggesting that the mean squared error term of the generator's loss function is driving the learning process. In **Supplementary Material**, we discuss our empirical findings about the adversarial loss of GAIN. For the purpose of reproducibility, as the gains of the adversarial loss appear to be small or negligible given our observations, we recommend training GAIN-GTEx without the adversarial term.

To evaluate the cross-study relevance of our method, we applied the trained models derived from GTEx (inductive mode) to perform imputation on The Cancer Genome Atlas gene expression data in acute myeloid leukemia, lung adenocarcinoma, and breast cancer. In addition to technical artifacts (e.g., batch effects), generalizing to this data is highly challenging because the expression is largely driven by features of the disease such as chromosomal abnormalities, genomic instabilities, large-scale mutations, and epigenetic changes (Baylin and Jones, 2011; Weinstein et al., 2013). Our results show that, despite these challenges, the methods were robust to gene expression from multiple diseases in different tissues (see **Table 3**), lending themselves to being used as tools to extend independent transcriptomic studies. Next, we evaluated the imputation performance of PMI and GAIN-GTEx for a range of values for the missing rate (see **Figure 4** and **Supplementary Figure 8**). We noted that the performance is stable and that the greater the proportion of missing values, the lower the prediction performance. Finally, we analyzed the imputation performance across genes from the Alzheimer's disease pathway (see **Figure 5**) and across all genes (see **Supplementary Figure 9**). We observed that the most highly imputed imputed genes are non-random and, indeed, cluster in some known pathways (see **Supplementary Figures 10, 11**).

**Broader Impact.** The study of the transcriptome is fundamental to our understanding of cellular and pathophysiological processes. High-dimensional gene expression data contain information relevant to a wide range of applications, including disease diagnosis (Huang et al., 2010), drug development (Sun et al., 2013), and evolutionary inference (Colbran et al., 2019). Thus, accurate and robust methods for imputation of gene expression have the enormous potential to enhance our molecular understanding of complex diseases, inform the search for novel drugs, and provide key insights into evolutionary processes. Here, we developed a

methodology that attains state-of-the art performance in several scenarios in terms of imputation quality and execution time. Our analysis showed that the use of blood as a surrogate for difficult-to-acquire tissues, as commonly practized in biomedical research, may lead to substantially degraded performance, with important implications for biomarker discovery and therapeutic development. Our method generalizes to gene expression in a disease class which has shown considerable health outcome disparities across population groups in terms of morbidity and mortality. Future algorithmic developments therefore hold promise for more effective detection, diagnosis, and treatment (Hosny and Aerts, 2019) and for improved implementation in clinical medicine (Char et al., 2018). Increased availability of transcriptomes in diverse human populations to enlarge our training data (a well-known and critical ethical challenge) should lead to further gains (i.e., decreased biases in results and reduced health disparities) (Wojcik et al., 2019). This work has the potential to catalyze research into the application of deep learning to molecular reconstruction of cellular states and downstream gene mapping of complex traits (Cookson et al., 2009; Zhou et al., 2020).

## 6. CONCLUSION

In this work, we developed two methods for gene expression imputation, which we named PMI and GAIN-GTEx. To increase the applicability of the proposed methods, we trained them on RNA-Seq data from the Genotype-Tissue Expression project, a reference resource that has generated a comprehensive collection of transcriptomes in a diverse set of tissues. A characteristic feature of our architectures is the use of word embeddings to learn distributed representations for the tissue types. Our approaches compared favorably to several standard and state-of-the-art imputation methods in terms of predictive performance and runtime, and generalized to transcriptomics data from 3 cancer types of the The Cancer Genome Atlas. PMI and GAIN-GTEx show optimal performance among the methods in inductive and in-place imputation, respectively, on the protein-coding genes. This work can facilitate the straightforward integration and cost-effective repurposing of large-scale RNA biorepositories into genomic studies of disease, with high applicability across diverse tissue types.

## DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found in the GTEx portal: https://gtexportal.org/. A detailed summary of the GTEx samples and donor information can be found at: https://gtexportal.org/home/tissueSummaryPage. Our code is publicly available at https://github.com/rvinas/GTEx-imputation.

## ETHICS STATEMENT

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

## AUTHOR CONTRIBUTIONS

RV and TA developed and trained the deep learning algorithm, generated all the results and figures. ERG provided the standardized RNA-seq data. ERG and PL supervised the study as joint senior authors. All the authors wrote and approved the manuscript.

## FUNDING

The project leading to these results has received funding from la Caixa Foundation (ID 100010434), under agreement LCF/BQ/EU19/11710059. This research was supported by the National Institutes of Health under award numbers R35HG010718 (ERG), R01HG011138 (ERG), R01GM140287 (ERG), and R01HL133559 (ERG). This research was also funded by the W. D. Armstrong Trust Fund, University of Cambridge, UK (TA) and the Engineering and Physical Sciences Research Council (R.V. EPSRC DTG 2018/19). PL was supported by MICA: Mental Health Data Pathfinder: University of Cambridge, Cambridgeshire and Peterborough NHS Foundation Trust, Microsoft, and the Medical Research Council (MC_PC_17213).

## ACKNOWLEDGMENTS

We thank Nikola Simidjievski, Cătălina Cangea, Ben Day, Cristian Bodnar, and Arian Jamasb for the helpful comments and discussion.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624128/full#supplementary-material

## REFERENCES

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., et al. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1101/787903

Baylin, S. B., and Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nat. Rev. Cancer* 11, 726–734. doi: 10.1038/nrc3130

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med. Inform. Decis. Mak.* 16:74. doi: 10.1186/s12911-016-0318-z

Buuren, S. V., and Groothuis-Oudshoorn, K. (2010). mice: multivariate imputation by chained equations in r. *J. Stat. Softw.* 45, 1–68. doi: 10.18637/jss.v045.i03

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490:61. doi: 10.1038/nature11412

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550. doi: 10.1038/nature13385

Cancer Genome Atlas Research Network, Ley, T. J., Miller, C., Ding, L., Raphael, B. J., Mungall, A. J., et al. (2013). Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.* 368, 2059–2074. doi: 10.1056/NEJMoa1301689

Carter, K. L., Cahir-McFarland, E., and Kieff, E. (2002). Epstein-barr virus-induced changes in b-lymphocyte gene expression. *J. Virol.* 76, 10427–10436. doi: 10.1128/JVI.76.20.10427-10436.2002

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. *N. Engl. J. Med.* 378:981. doi: 10.1056/NEJMp1714229

Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. (2019). Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat. Ecol. Evol.* 3, 1598–1606. doi: 10.1038/s41559-019-0996-x

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-1047-4

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194. doi: 10.1038/nrg2537

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., et al. (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423–428. doi: 10.1038/nature06758

Evans, W. E., and Relling, M. V. (2004). Moving towards individualized medicine with pharmacogenomics. *Nature* 429, 464–468. doi: 10.1038/nature02626

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., et al. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nat. Genet.* 50, 956–967. doi: 10.1038/s41588-018-0154-4

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems* - Volume 2 (Cambridge, MA: MIT Press), 2672-2680.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277

Gupta, A., and Zou, J. (2019). Feedback gan for DNA optimizes protein functions. *Nat. Mach. Intell.* 1, 105–111. doi: 10.1038/s42256-019-0017-4

Hosny, A., and Aerts, H. J. (2019). Artificial intelligence for global health. *Science* 366, 955–956. doi: 10.1126/science.aay5189

Huang, H., Liu, C.-C., and Zhou, X. J. (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6823–6828. doi: 10.1073/pnas.0912043107

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 448–456. Available online at: JMLR.org.

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

Kim, K., Kim, M.-J., Kim, D. W., Kim, S. Y., Park, S., and Park, C. B. (2020). Clinically accurate diagnosis of alzheimer's disease via multiplexed sensing of core biomarkers in human plasma. *Nat. Commun.* 11, 1–9. doi: 10.1038/s41467-019-13901-z

King, M.-C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv [Preprint]* arXiv:1412.6980.

Little, R. J., and Rubin, D. B. (2019). *Statistical Analysis With Missing Data*, Vol. 793. New York, NY: John Wiley & Sons.

Low, L. A., Mummery, C., Berridge, B. R., Austin, C. P., and Tagle, D. A. (2020). Organs-on-chips: into the next decade. *Nat. Rev. Drug Discov.*, 1–17. doi: 10.1038/s41573-020-0079-3

Maaten, L. V. D. and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9, 2579–2605. Available online at: https://www.jmlr.org/papers/v9/vandermaaten08a.html

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., and Bonn, S. (2020). Realistic *in silico* generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nat. Commun.* 11, 1–12. doi: 10.1038/s41467-019-14018-z

McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). Umap: uniform manifold approximation and projection. *J. Open Sour. Softw.* 3:861. doi: 10.21105/joss.00861

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, Vol. 26, eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Curran Associates, Inc.). Available online at: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

Øystein Sørensen, Hellton, K. H., Frigessi, A., and Thoresen, M. (2018). Covariate selection in high-dimensional generalized linear models with measurement error. *J. Comput. Graph. Stat.* 27, 739–749. doi: 10.1080/10618600.2018.1425626

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., et al. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3:96ra77. doi: 10.1126/scitranslmed.3001318

Stegle, O., Parts, L., Piipari, M., Winn, J., and Durbin, R. (2012). Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* 7:500. doi: 10.1038/nprot.2011.457

Stekhoven, D. J., and Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Sun, X., Vilar, S., and Tatonetti, N. P. (2013). High-throughput methods for combinatorial drug discovery. *Sci. Transl. Med.* 5:205rv1. doi: 10.1126/scitranslmed.3006667

The GTEx Consortium (2015). The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics* 17, 520–525. doi: 10.1093/bioinformatics/17.6.520

Viñas, R., Andrés-Terré, H., Liò, P., and Bryson, K. (2021). Adversarial generation of gene expression data. *Bioinformatics* btab035. doi: 10.1093/bioinformatics/btab035

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis*, eds D. P. Berrar, W. Dubitzky, and M. Granzow (Boston, MA: Springer), 91–109.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45:1113. doi: 10.1038/ng.2764

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570, 514–518. doi: 10.1038/s41586-019-1310-4

Yoon, J., Jordon, J., and Van Der Schaar, M. (2018). GAIN: missing data imputation using generative adversarial nets. *arXiv [Preprint]* arXiv:1806.02920.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4. doi: 10.2202/1544-6115.1128

Zhou, D., Jiang, Y., Zhong, X., Cox, N. J., Liu, C., and Gamazon, E. R. (2020). A unified framework for joint-tissue transcriptome-wide association and mendelian randomization analysis. *Nat. Genet.* 52, 1239–1246. doi: 10.1038/s41588-020-0706-2

**Conflict of Interest:** ERG receives an honorarium from the journal *Circulation Research* of the American Heart Association, as a member of the Editorial Board.

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: 29 April 2021
doi: 10.3389/fgene.2021.671523

# Bioinformatic Analysis of Crosstalk Between circRNA, miRNA, and Target Gene Network in NAFLD

*Cen Du[1†], Lan Shen[2†], Zhuoqi Ma[1], Jian Du[1*] and Shi Jin[1*]*

[1] *The Fourth Affiliated Hospital of China Medical University, Shenyang, China,* [2] *Shanghai Chest Hospital, Shanghai Jiaotong University, Shanghai, China*

**Background:** The majority of chronic liver disease is caused by non-alcoholic fatty liver disease (NAFLD), which is one of the highly prevalent diseases worldwide. The current studies have found that non-coding RNA (ncRNA) plays an important role in the NAFLD, but few studies on circRNA. In this study, genes, microRNA (miRNA), and circular RNA (circRNA) associated with NAFLD were found by bioinformatic methods, bringing a novel perspective for the prevention and treatment of NAFLD.

**Methods:** Expression data of GSE63067 was acquired from Gene Expression Omnibus (GEO) database. The liver samples were collected from the people diagnosed with NAFLD or not. Differentially expressed genes (DEGs) were obtained from the steatosis vs. the control group and non-alcoholic steatohepatitis (NASH) vs. the control group using the GEO2R online tool. The overlapping genes remained for further functional enrichment analysis and protein-protein interaction network analysis. MiRNAs and circRNAs targeting these overlapping DEGs were predicted from the databases. Finally, the GSE134146 dataset was used to verify the expression of circRNA.

**Results:** In summary, 228 upregulated and 63 downregulated differential genes were selected. The top 10 biological processes and relative signaling pathways of the upregulated differential genes were obtained. Also, ten hub genes were performed in the Protein-protein interaction (PPI) network. One hundred thirty-nine miRNAs and 902 circRNAs were forecast for the differential genes by the database. Ultimately, the crosstalk between hsa_circ_0000313, miR-6512-3p, and *PEG10* was constructed.

**Conclusion:** The crosstalk of hsa_circ_0000313-hsa-miR-6512-3p-*PEG10* and some related non-coding RNAs may take part in NAFLD's pathogenesis, which could be the potential biomarkers of NAFLD in the future.

**Keywords: NAFLD, non-coding RNAs, circRNA, miRNA, bioinformatic analysis**

OPEN ACCESS

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

***Reviewed by:***
*Tuba Denkçeken,*
*Sanko University, Turkey*
*Elif Pala,*
*Sanko University, Turkey*

****Correspondence:***
*Jian Du*
*dujianbox@126.com*
*Shi Jin*
*jinshi_1981@163.com*

*[†] 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:*** *24 February 2021*
***Accepted:*** *07 April 2021*
***Published:*** *29 April 2021*

***Citation:***
*Du C, Shen L, Ma Z, Du J and Jin S (2021) Bioinformatic Analysis of Crosstalk Between circRNA, miRNA, and Target Gene Network in NAFLD. Front. Genet. 12:671523. doi: 10.3389/fgene.2021.671523*

## INTRODUCTION

Currently, obesity takes center stage in a series of metabolic diseases. NAFLD is a manifestation of obesity in the liver. This general term describes a series of liver conditions ranging from steatosis to NASH, steatohepatitis with fibrosis, and cirrhosis. The worldwide incidence rate of NAFLD is increasing yearly. With a mean 25% prevalence globally, and the highest in the Middle East, 32%

and the lowest in Africa, 14% (Younossi et al., 2016b). The overall medical costs have exceeded $100 billion every year in the United States (Mundi et al., 2020) and €35 billion in European countries (Younossi et al., 2016a).

In the past, the pathogenesis of NAFLD was mainly based on the "*two-hit*" hypothesis. The first hit is insulin resistance giving rise to the liver fat accumulation. The second hit is caused by comprehensive effects of mitochondrial dysfunction, inflammatory cytokines, lipid peroxidation, and oxidative stress due to the damage of hepatocytes and inflammatory response. However, there has been a growing number of recently researched ncRNAs in NAFLD (Gjorgjieva et al., 2019; Chien et al., 2020). Studies on genome-wide transcriptome have shown that a large number of ncRNAs, such as miRNAs and circRNAs, can regulate the expression of human genome. MiRNA is a kind of small single-stranded RNA that can inhibit the expression of its target genes. CircRNA is a class of endogenous ncRNA as well, different from the traditional linear RNA formed by reverse splicing which has plentiful miRNA binding sites and can act as miRNA sponge. It is a circular closed structure without a 5-terminal cap and 3-terminal tail (Chen and Yang, 2015).

What's more critical, abnormal lipid metabolism in the liver is often accompanied by a disordered ncRNA expression (Singh et al., 2018). So far, the most accurate standard to diagnose NAFLD is liver biopsy. But for its invasiveness and expensiveness, it cannot be utilized widely. Non-invasive diagnostic methods should be considered for the continued investigation. In summary, the study on DEGs and ncRNA can explain the pathogenesis of NAFLD from another point of view. Also, it may be a non-invasive way of detection for NAFLD.

In our study, microarray data were obtained from the GEO database, and the DEGs were identified between individuals with or without NAFLD. Several databases predicted miRNAs and circRNAs targeting the DEGs to establish a circRNA-miRNA-mRNA network. In that case, some potentially therapeutic targets for NAFLD could be explored.

# MATERIALS AND METHODS

## Information of Microarray Data

Microarray data of GSE63067 was downloaded from the GEO database[1]. Three groups of participants were included in this study. Two subjects were diagnosed with steatosis and nine with NASH, and the other seven healthy controls. The microarray platform was GPL570[HG-U133_Plus_2] (Affymetrix Human Genome U133 Plus 2.0 Array).

## Screening of DEGs

Using the online tool GEO2R/R package limma, DEGs were screened from the microarray by the cut-off point of $P$-value $< 0.05$ and $|\log FC| > 0.5$. In this way, DEGs could be screened from the steatosis vs. healthy control, so did the non-alcoholic steatohepatitis vs. the control group. Only the overlapping genes in both of these groups could be selected as significant DEGs. Gene without a name should be excluded.

## Functional Enrichment Analysis of DEGs

The selected significant DEGs were uploaded to the Database for Annotation, Visualization, and Integrated Discovery (DAVID) version 6.8 Beta[2] for further analysis. In this study, we were committed to studying the GO annotation and KEGG pathways of DEGs in the DAVID database. $P$-value $< 0.05$ was chosen as the threshold.

## Protein-Protein Interaction Network

Protein-protein interaction (PPI) networks for the significant DEGs were constructed by the Search Tool for the Retrieval of Interacting Genes database (STRING version 11.0)[3] (Szklarczyk et al., 2011). The minimum required interaction score was 0.7. In this network, we hid the nodes without connections with others. Cytoscape3.7.2 (Doncheva et al., 2019)[4] was applied to display the relationship between proteins. The hub gene of PPI network can be analyzed by cytohubba (Chin et al., 2014) in Cytoscape.

## CircRNA-miRNA-mRNA Network Construction

We chose the top 10 differential genes for further analysis. MiRWalk[5], miRDB[6], and miRNet[7] databases were used to predict miRNA-targeted mRNA. The miRNA only in more than two of these databases would be retained. Starbase[8] database was applied to select the miRNA and their targeted circRNA. Finally, the network of circRNA-miRNA-mRNA was constructed.

## Expression Validation of circRNA

Certification of the expression of circRNA was undertaken by using another GEO dataset (GSE134146). The cut-off point of $P$-value and logFC were the same with the DEGs' selection standard.

# RESULTS

## Analysis of DEGs in Liver Sample With or Without NAFLD

Firstly, the data of GSE63067 was normalized. The expression of all genes in steatosis vs. Control (**Figure 1A**) and NASH vs. Control (**Figure 1B**) were exhibited in the volcano plot. There are 1362 differential genes in steatosis vs. Control: 857 upregulated genes and 505 downregulated genes. Simultaneously, 785 genes were found in NASH vs. Control: 617 upregulated genes and 168 downregulated genes. A Venn diagram was used to find the common genes in both groups: 228 upregulated genes

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

[2] https://david-d.ncifcrf.gov/

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

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

[5] http://mirwalk.umm.uni-heidelberg.de/

[6] http://www.mirdb.org/

[7] https://www.mirnet.ca/miRNet/home.xhtml

[8] http://starbase.sysu.edu.cn/

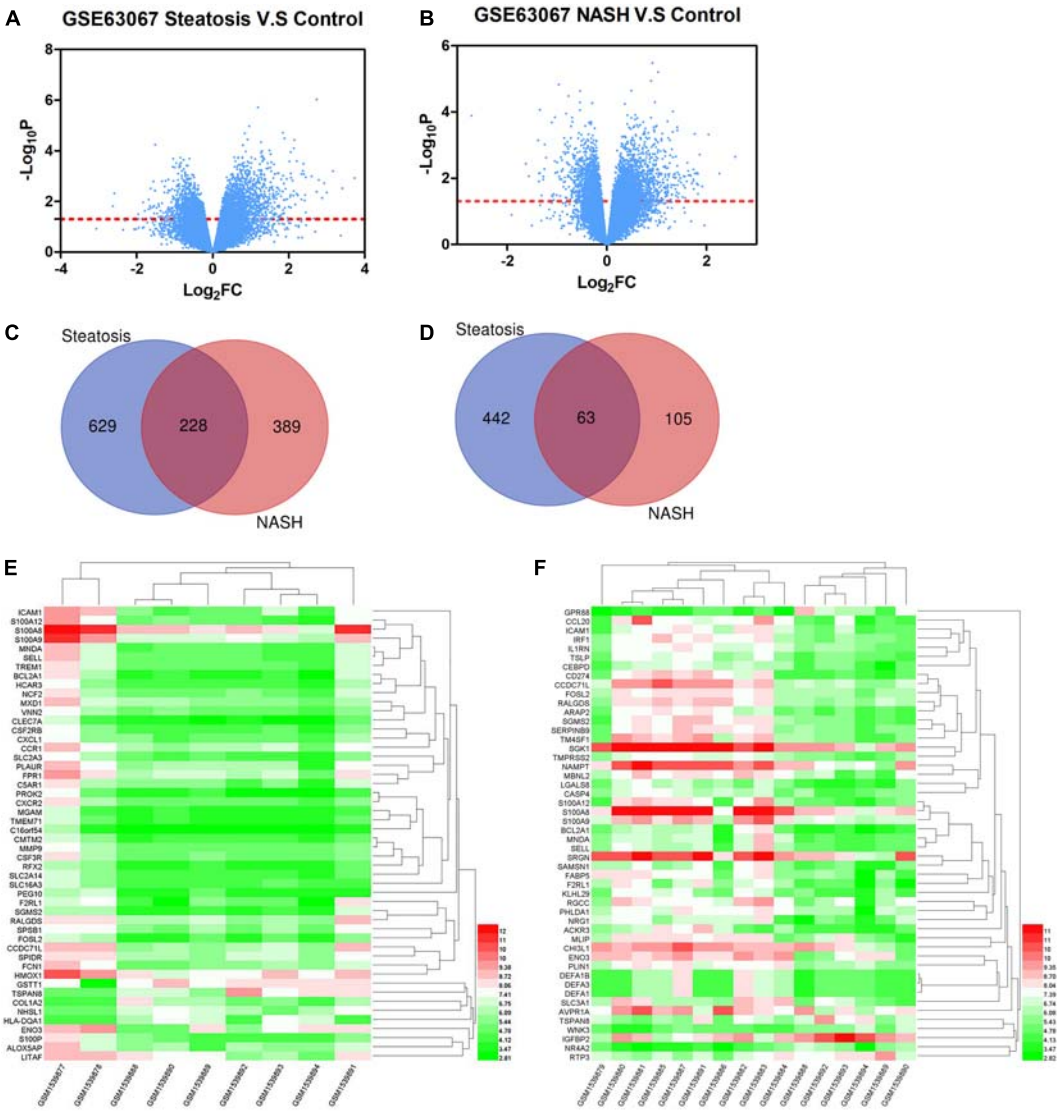


**FIGURE 1 | (A)** The expression of all genes in steatosis vs. Control. **(B)** The expression of all genes in NASH vs. Control. **(C)** Venn diagram of upregulated genes in steatosis and NASH. **(D)** Venn diagram of downregulated genes in steatosis and NASH. **(E)** Top 50 differential genes in steatosis vs. Control. **(F)** Top 50 differential genes in NASH vs. Control.

(**Figure 1C**) and 63 downregulated genes (**Figure 1D**). The top 10 differential genes were shown in **Table 1**. Heatmap was used to indicate the top 50 differential genes in steatosis vs. Control (**Figure 1E**) and NASH vs. Control (**Figure 1F**).

## Analysis of Signaling Pathway and Biological Functions for 228 Upregulated DEGs by DAVID Database

The biological processes of the 228 common upregulated genes were analyzed by the DAVID database (**Figure 2A**). The major biological processes were included as followed: leukocyte migration (GO:0050900, $P$ = 3.07E-11), inflammatory response (GO:0006954, $P$ = 3.82E-10), innate immune response (GO:0045087, $P$ = 2.16E-08), neutrophil chemotaxis (GO: 0030593, $P$ = 1.56E-06), chemotaxis (GO:0006935, $P$ = 2.95E-06), signal transduction (GO:0007165, $P$ = 6.90E-06), defense response to fungus (GO:0050832, $P$ = 1.84E-05), negative regulation of apoptotic process (GO:0043066, $P$ = 5.73E-05), immune response (GO:0006955, $P$ = 7.76E-05), response to lipopolysaccharide (GO:0032496, $P$ = 2.16E-04). The crucial signaling pathways were exhibited in **Figure 2B**.

## Protein-Protein Network Construction and Hub Genes Analysis

We used String to construct proteins' interaction and selected the network's essential genes (**Figure 3**). The top 10 of hub genes were *FPR2, SMAD3, CD53, FCER1G, SMURF2, FPR1, LYN, SOCS3,*

**TABLE 1 |** Top 10 differential genes in both steatosis vs. Control and NASH vs. Control.

| Gene symbol | Gene | LogFC | *P*-value |
|---|---|---|---|
| ICAM1 | intercellular adhesion molecule 1 | 3.73 | 1.19E-03 |
| S100A12 | S100 calcium binding protein A12 | 3.42 | 3.02E-03 |
| S100A9 | S100 calcium binding protein A9 | 3.17 | 6.31E-04 |
| S100A8 | S100 calcium binding protein A8 | 3.00 | 4.45E-03 |
| MNDA | myeloid cell nuclear differentiation antigen | 2.93 | 1.88E-03 |
| BCL2A1 | BCL2 related protein A1 | 2.88 | 7.28E-03 |
| SELL | selectin L | 2.79 | 1.42E-03 |
| PEG10 | paternally expressed 10 | 2.74 | 9.31E-07 |
| GPR88 | G protein-coupled receptor 88 | −2.72 | 1.30E-04 |
| F2RL1 | F2R like trypsin receptor 1 | 2.66 | 4.69E-02 |

*CD44, MMP9* (**Figure 4**). The score of each hub gene was shown in **Table 2**.

## Prediction of Target miRNA and circRNA

For those top 10 differential genes, we used databases (MiRWalk, miRDB, and miRNet) to forecast their target miRNAs. We found out one hundred thirty-nine miRNAs targeted these ten genes since one gene may target several miRNAs. For instance, gene GPR88 had five miRNAs as the potential targets, as miR-5591-5p, miR-181a-5p, miR-6507-5p, miR-920, and miR-628-5p. Meanwhile, several genes could target one miRNA. For example, miR-122-5p and miR-5004-5p could be targeted by more than one gene. Then we use the Starbase to predict the corresponding circRNAs for each of these 139 miRNAs. As a result, 902 circRNAs could be relevant to these miRNAs.

## Validation for Expression of circRNA in GSE134146

The selection criteria screened 752 differential circRNAs with 172 lowly expressed circRNAs and 580 highly expressed circRNAs in NAFLD. We took the interaction of these differential circRNAs with previous circRNAs. In the end, hsa_circ_0001453 and hsa_circ_0000313 could satisfy the above conditions. We could construct two circRNA-miRNA-mRNA as followed: hsa_circ_0001453-hsa-miR-27b-3p-*PEG10* and hsa_circ_0000313-hsa-miR-6512-3p-*PEG10*.

# DISCUSSION

In the past few years, overwhelming evidence has demonstrated that non-coding RNA may play a vital role in NAFLD progression. The Burgeoning high-throughput sequencing technologies make it possible for people to have a better understanding of non-coding transcripts. Starting from the differential genes on NAFLD, we speculated the miRNA and circRNA by online databases that they might bind to construct a circRNA-miRNA-mRNA interaction network to mirror NAFLD's molecular mechanism.

Two hundred ninety-one differential genes were obtained from the GSE63067. The most differential genes were *ICAM1, S100A12, S100A9, S100A8, MNDA, BCL2A1, SELL, PEG10, GPR88,* and *F2RL1. FPR2, SMAD3, CD53, FCER1G, SMURF2, FPR1, LYN, SOCS3, CD44*, and *MMP9* were the hub genes in the protein-protein network, which indicated a leading role in predicting the risk of NAFLD. *PEG10* was the only left differentially gene in the circRNA-miRNA-mRNA network after multiple filters among all of these genes.

Some of these genes could directly affect NAFLD development, and the others may influence NAFLD by some related effects. The level of *ICAM1* was markedly enhanced after the stimulation of lipopolysaccharide in the liver, and the blockade of it may destroy the cells' adhesion and expansion (Meng et al., 2019). *CD44*, is proved to be of great significance in non-alcoholic steatohepatitis. Patouraux found that *CD44* deficiency was strongly relevant to the activation of macrophages by lipopolysaccharide (LPS), hepatocyte damage-associated molecular patterns (DAMPs) and saturated fatty acids (Patouraux et al., 2017). *SMURF2*, a kind of E3 ubiquitin ligase, is able to attenuate liver fibrosis by inhibiting the hepatic stellate cell activation (Cai et al., 2018). These genes were closely connected with NAFLD (Chikada et al., 2020).

Additionally, apoptosis, inflammation, and insulin resistance were the critical processes in the pathogenesis of NAFLD. Inflammation is an inducing factor that can fuel the transition from steatosis to NASH (Schuster et al., 2018). Our study also discovered that inflammation attached great importance to NAFLD by analyzing the pathway and biological function of differential genes, which coincided with the previous studies. FPR are N-formylpeptide receptors involved in some pathogenic processes. Both *FPR1* and *FPR2* were found to have a relationship with NAFLD through influencing the inflammatory effect. Recently, a study indicated that *FPR2* deficiency can alleviate diet-induced insulin resistance, which is by weight loss and inhibiting inflammation (Chen et al., 2019). Some immune cells could also participate in many inflammatory processes. *S100A8, S100A9, and S100A12* were $Ca^{2+}$ binding proteins that were abundant in many immune cells. They can accelerate immune cells to release inflammatory factors to destroy the immunity homeostasis (Wang et al., 2018). A recent study indicated that silencing *S100A8* could alleviate inflammation and oxidative stress, along with the changes of corresponding proteins (Hu and Lin, 2021). As for *PEG10*, it was closely related to adipocyte differentiation (Hishida et al., 2007). The expression of *PEG10* was correlated positively with insulin resistance and physical activity in NASH (Arendt et al., 2019). It is worth studying the left genes which have not been found a relationship with NAFLD in the previous research. By analyzing the biological processes of these differential genes, we discovered that inflammation might be of great importance in NAFLD's pathogenesis. This conclusion was the same as the previous discoveries.

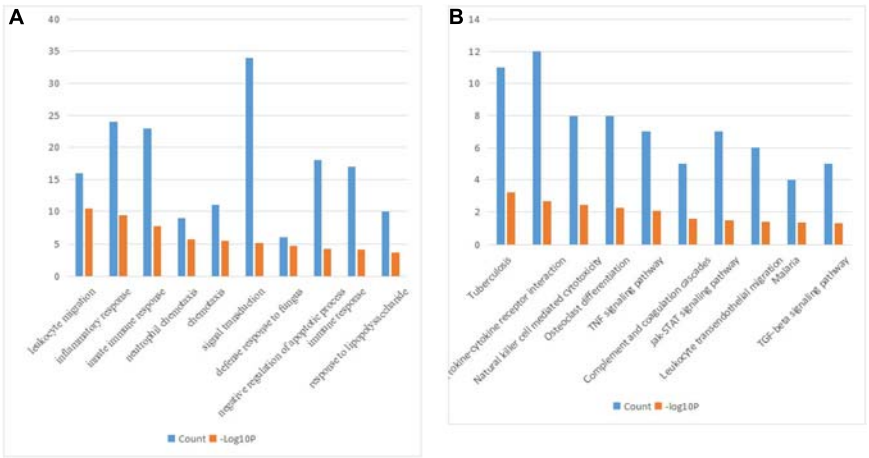


**FIGURE 2 | (A)** The biological processes of the 228 common upregulated genes. **(B)** KEGG pathways of the 228 common upregulated genes.

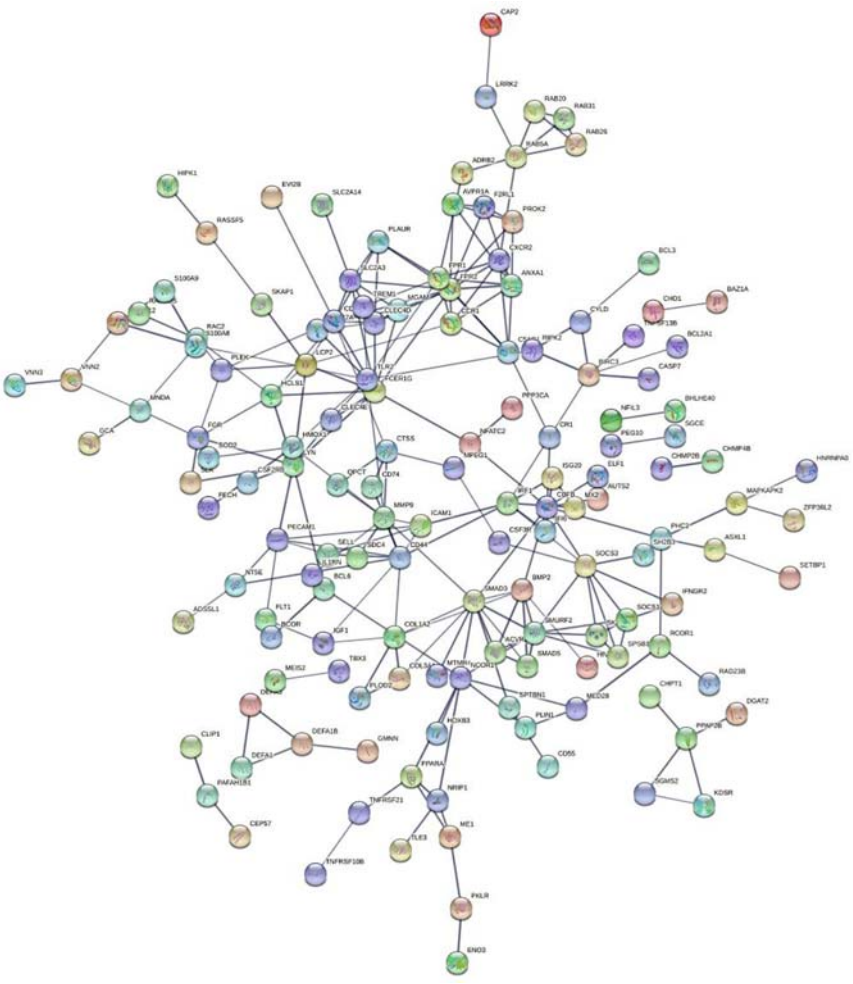

**FIGURE 3 |** Protein-protein networks.

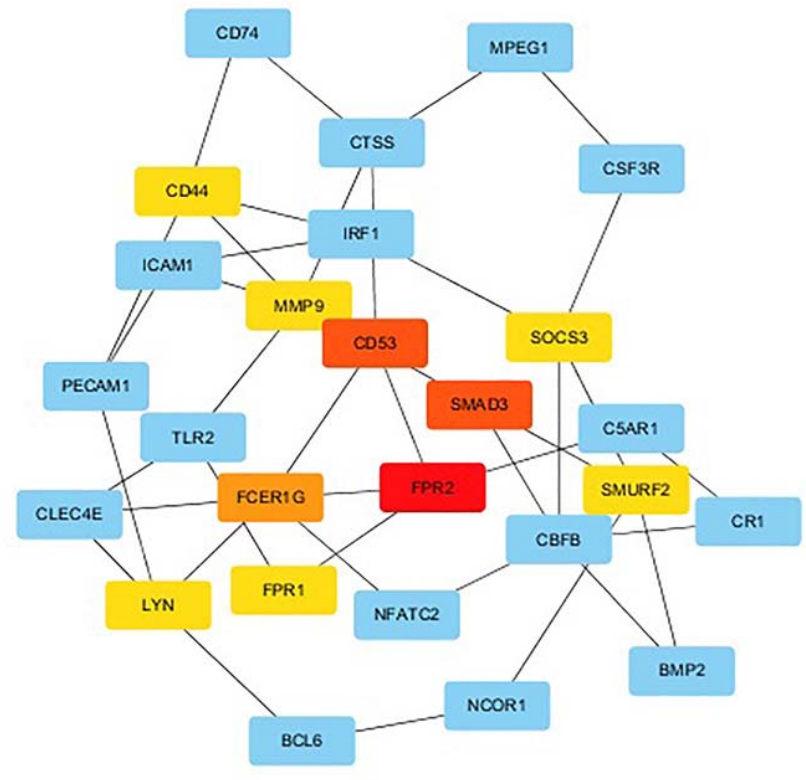


**FIGURE 4 |** The top 10 hub genes and expanded subnetwork.

**TABLE 2 |** The top 10 scores of hub genes.

| Gene symbol | Score |
|---|---|
| FPR2 | 14 |
| SMAD3 | 11 |
| CD53 | 11 |
| FCER1G | 10 |
| SMURF2 | 9 |
| FPR1 | 9 |
| LYN | 9 |
| SOCS3 | 9 |
| CD44 | 9 |
| MMP9 | 9 |

A great number of researches on differential genes of NAFLD have been heatedly discussed before. A study by Frades integrated genomic signatures of hepatocellular carcinoma derived from NAFLD (Frades et al., 2015). Compared with our study, the results they got integrate both human and mouse samples which maybe more universal. Another research by Feng et al. (2020) also used the GSE63067 to identify a total of 249 DEGs and one key gene (CCL20) for NAFLD. The DEGs they got may not be consistent with us because of the different screening criteria and grouping method. It is well known that NAFLD is comprised of four different conditions. In our study, we screened the DEGs in steatosis as well as in NASH, which makes the result more convinced. In terms of the miRNAs, we found miR-122-5p and miR-5004 had more than one target gene. MiR-122 is one of the most expressed MiRNAs in the liver (Yamada et al., 2015; Jampoka et al., 2018). In the different stages of NAFLD, the expression of miR-122 was different. *In vitro* and *Vivo* NAFLD model, miR-122 could promote the hepatic lipogenesis via suppressing the expression of *Sirt1* (Long et al., 2019). Reduced miR-122's expression can lighten the fatty deposits and inflammation (Hu et al., 2019). In the longitudinal evaluation of one patient from NAFLD's diagnosis until HCC, the expression of miR-122 may have a tendency to decrease before the progression of the fibrosis stage (Akuta et al., 2016). In our study, we speculated that miR-122-5p was downregulated for its opposite expression to the target gene. We summarized the possible reasons as followed. Firstly, we only chose one microarray to analyze the outcome. And it may have some incidental factors. Secondly, miRNA coming from a different source of samples may have an extra level of expression. Finally, the major subjects in this microarray were diagnosed with NASH. The miR-27-5p reported that miR-27-5p could increase lipid and TG contents (Murata et al., 2019) and it was an important adipogenic factor that can regulate adipogenesis in hyperlipidemia (Hsu et al., 2017). Still, we got the opposite outcome for the same reason

with miR-122-5p. No existing links were found in miR-5004 and miR-6512-3p with NAFLD. As a result, miR-5004 and miR-6512-3p deserve further investigation which may be the potential biomarkers for NAFLD. There is no thorough research for these two circRNAs, so that we couldn't validate our conjecture with previous research. All the results we got were based on the theoretical analysis. More experiments should be applied to verify these speculations.

In conclusion, we got some differential genes and constructed the hsa_circ_0000313-hsa-miR-6512-3p-*PEG10* network in NAFLD, which may be the underlying targets in diagnosis and treatment for NAFLD.

## 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

CD and LS contributed to the conception and design of the study, the data analysis, the data interpretation, the manuscript drafting, and the critical revision of the manuscript. ZM, JD, and SJ contributed to the data analysis, the data interpretation, the manuscript drafting, and the critical revision of the manuscript.

## FUNDING

This study was supported by the Provincial Science and Technology Department Natural Fund Guidance Project No. 2019-ZD-0774.

## REFERENCES

Akuta, N., Kawamura, Y., Suzuki, F., Saitoh, S., Arase, Y., Kunimoto, H., et al. (2016). Impact of circulating miR-122 for histological features and hepatocellular carcinoma of nonalcoholic fatty liver disease in Japan. *Hepatol. Int.* 10, 647–656. doi: 10.1007/s12072-016-9729-2

Arendt, B. M., Teterina, A., Pettinelli, P., Comelli, E. M., Ma, D. W. L., Fung, S. K., et al. (2019). Cancer-related gene expression is associated with disease severity and modifiable lifestyle factors in non-alcoholic fatty liver disease. *Nutrition* 62, 100–107. doi: 10.1016/j.nut.2018.12.001

Cai, Y., Huang, G., Ma, L., Dong, L., Chen, S., Shen, X., et al. (2018). Smurf2, an E3 ubiquitin ligase, interacts with PDE4B and attenuates liver fibrosis through miR-132 mediated CTGF inhibition. *Biochim. Biophys. Acta Mol. Cell Res.* 1865, 297–308. doi: 10.1016/j.bbamcr.2017.10.011

Chen, L. L., and Yang, L. (2015). Regulation of circRNA biogenesis. *RNA Biol.* 12, 381–388. doi: 10.1080/15476286.2015.1020271

Chen, X., Zhuo, S., Zhu, T., Yao, P., Yang, M., Mei, H., et al. (2019). Fpr2 deficiency alleviates diet-induced insulin resistance through reducing body weight gain and inhibiting inflammation mediated by macrophage chemotaxis and M1 polarization. *Diabetes* 68, 1130–1142. doi: 10.2337/db18-0469

Chien, Y., Tsai, P. H., Lai, Y. H., Lu, K. H., Liu, C. Y., Lin, H. F., et al. (2020). CircularRNA as novel biomarkers in liver diseases. *J. Chin. Med. Assoc.* 83, 15–17. doi: 10.1097/jcma.0000000000000230

Chikada, H., Ida, K., Nishikawa, Y., Inagaki, Y., and Kamiya, A. (2020). Liver-specific knockout of B cell lymphoma 6 suppresses progression of non-alcoholic steatohepatitis in mice. *Sci. Rep.* 10:9704.

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8(Suppl. 4):S11. doi: 10.1186/1752-0509-8-S4-S11

Doncheva, N. T., Morris, J. H., Gorodkin, J., and Jensen, L. J. (2019). Cytoscape stringapp: network analysis and visualization of proteomics data. *J. Proteome Res.* 18, 623–632. doi: 10.1021/acs.jproteome.8b00702

Feng, G., Li, X. P., Niu, C. Y., Liu, M. L., Yan, Q. Q., Fan, L. P., et al. (2020). Bioinformatics analysis reveals novel core genes associated with nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. *Gene* 742:144549. doi: 10.1016/j.gene.2020.144549

Frades, I., Andreasson, E., Mato, J. M., Alexandersson, E., Matthiesen, R., and Martínez-Chantar, M. L. (2015). Integrative genomic signatures of hepatocellular carcinoma derived from nonalcoholic fatty liver disease. *PLoS One* 10:e0124544. doi: 10.1371/journal.pone.0124544

Gjorgjieva, M., Sobolewski, C., Dolicka, D., Correia de Sousa, M., and Foti, M. (2019). miRNAs and NAFLD: from pathophysiology to therapy. *Gut* 68, 2065–2079. doi: 10.1136/gutjnl-2018-318146

Hishida, T., Naito, K., Osada, S., Nishizuka, M., and Imagawa, M. (2007). peg10, an imprinted gene, plays a crucial role in adipocyte differentiation. *FEBS Lett.* 581, 4272–4278. doi: 10.1016/j.febslet.2007.07.074

Hsu, C. C., Lai, C. Y., Lin, C. Y., Yeh, K. Y., and Her, G. M. (2017). MicroRNA-27b depletion enhances endotrophic and intravascular lipid accumulation and induces adipocyte hyperplasia in zebrafish. *Int. J. Mol. Sci.* 19:93. doi: 10.3390/ijms19010093

Hu, W., and Lin, C. (2021). S100a8 silencing attenuates inflammation, oxidative stress and apoptosis in BV2 cells induced by oxygen-glucose deprivation and reoxygenation by upregulating GAB1 expression. *Mol. Med. Rep.* 23:64.

Hu, Y., Du, G., Li, G., Peng, X., Zhang, Z., and Zhai, Y. (2019). The miR-122 inhibition alleviates lipid accumulation and inflammation in NAFLD cell model. *Arch. Physiol. Biochem.* [Epub ahead of print], 1–5. doi: 10.1080/13813455.2019.1640744

Jampoka, K., Muangpaisarn, P., Khongnomnan, K., Treeprasertsuk, S., Tangkijvanich, P., and Payungporn, S. (2018). Serum miR-29a and miR-122 as potential biomarkers for non-alcoholic fatty liver disease (NAFLD). *Microrna* 7, 215–222. doi: 10.2174/2211536607666180531093302

Long, J. K., Dai, W., Zheng, Y. W., and Zhao, S. P. (2019). miR-122 promotes hepatic lipogenesis via inhibiting the LKB1/AMPK pathway by targeting Sirt1 in non-alcoholic fatty liver disease. *Mol. Med.* 25:26.

Meng, D., Qin, Y., Lu, N., Fang, K., Hu, Y., Tian, Z., et al. (2019). Kupffer cells promote the differentiation of adult liver hematopoietic stem and progenitor cells into lymphocytes via ICAM-1 and LFA-1 interaction. *Stem Cells Int.* 2019:4848279.

Mundi, M. S., Velapati, S., Patel, J., Kellogg, T. A., Abu Dayyeh, B. K., and Hurt, R. T. (2020). Evolution of NAFLD and its management. *Nutr. Clin. Pract.* 35, 72–84. doi: 10.1002/ncp.10449

Murata, Y., Yamashiro, T., Kessoku, T., Jahan, I., Usuda, H., Tanaka, T., et al. (2019). Up-regulated microRNA-27b promotes adipocyte differentiation via induction of acyl-CoA thioesterase 2 expression. *Biomed. Res. Int.* 2019:2916243.

Patouraux, S., Rousseau, D., Bonnafous, S., Lebeaupin, C., Luci, C., Canivet, C. M., et al. (2017). CD44 is a key player in non-alcoholic steatohepatitis. *J. Hepatol.* 67, 328–338. doi: 10.1016/j.jhep.2017.03.003

Schuster, S., Cabrera, D., Arrese, M., and Feldstein, A. E. (2018). Triggering and resolution of inflammation in NASH. *Nat. Rev.*

*Gastroenterol. Hepatol.* 15, 349–364. doi: 10.1038/s41575-018-0009-6

Singh, A. K., Aryal, B., Zhang, X., Fan, Y., Price, N. L., Suárez, Y., et al. (2018). Posttranscriptional regulation of lipid metabolism by non-coding RNAs and RNA binding proteins. *Semin. Cell Dev. Biol.* 81, 129–140. doi: 10.1016/j.semcdb.2017.11.026

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* 39, D561–D568.

Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). S100A8/A9 in inflammation. *Front. Immunol.* 9:1298. doi: 10.3389/fimmu.2018.01298

Yamada, H., Ohashi, K., Suzuki, K., Munetsuna, E., Ando, Y., Yamazaki, M., et al. (2015). Longitudinal study of circulating miR-122 in a rat model of non-alcoholic fatty liver disease. *Clin. Chim. Acta.* 446, 267–271. doi: 10.1016/j.cca.2015.05.002

Younossi, Z. M., Blissett, D., Blissett, R., Henry, L., Stepanova, M., Younossi, Y., et al. (2016a). The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. *Hepatology* 64, 1577–1586. doi: 10.1002/hep.28785

Younossi, Z. M., Koenig, A. B., Abdelatif, D., Fazel, Y., Henry, L., and Wymer, M. (2016b). Global epidemiology of nonalcoholic fatty liver disease-meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* 64, 73–84. doi: 10.1002/hep.28431

**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: 10 May 2021
doi: 10.3389/fgene.2021.654955

# Transcriptome Analysis of Choroid and Retina From Tree Shrew With Choroidal Neovascularization Reveals Key Signaling Moieties

*Jie Jia[1,2†], Dandan Qiu[1,3†], Caixia Lu[1], Wenguang Wang[1], Na Li[1], Yuanyuan Han[1], Pinfen Tong[1], Xiaomei Sun[1], Min Wu[4]* and Jiejie Dai[1]**

[1]Institute of Medical Biology, Chinese Academy of Medical Science and Peking Union Medical College, Kunming, China, [2]Scientific Research Laboratory Center, The First Affiliated Hospital of Kunming Medical University, Kunming, China, [3]Kunming Medical University, Kunming, China, [4]Yunnan Eye Institute, The Second People's Hospital of Yunnan, Kunming, China

Pathological neovascularization in choroid, a leading cause of blindness, is a characteristic of many fundus diseases, such as diabetic retinopathy and age-related macular degeneration. The present study aimed to elucidate the key signaling pathways in choroidal neovascularization (CNV) by analyzing the mRNA profiles of choroid and retina in tree shrews with CNV. We induced choroidal angiogenesis by laser photocoagulation in 15 tree shrews and obtained mRNA profiles of their choroids and retinas by high-throughput transcriptome sequencing. Hierarchical cluster analysis, weighted gene co-expression network analysis (WGCNA), protein-protein interaction (PPI) network analysis, hematoxylin and eosin (HE) staining, CD31 immunohistochemistry (IHC), and reverse transcription quantitative PCR (RT-qPCR) were performed. After laser photocoagulation, we obtained a total of 350 differentially expressed genes (DEGs) in the choroid, including 59 genes in Module-FASN ("ME-FASN") module and 28 genes in Module-RPL ("ME-RPL") module. A total of 69 DEGs in retina, including 20 genes in Module-SLC ("ME-SLC") module. Bioinformatics analysis demonstrated that DEGs in choroid were mainly involved in membrane transport; DEGs in "ME-RPL" were prominent in pathways associated with IgA production, antigen presentation, and cell adhesion molecules (CAMs) signaling. DEGs in "ME-FASN" were involved in fatty acid metabolism and PPAR signaling pathway, while DEGs in "ME-SLC" were involved in GABAergic synapse, neuroactive life receptor interaction, cholinergic synapse, and retrograde endocannabinoid signaling pathway. PPI network analysis demonstrated that the ribosomal protein family genes (*RPL31, RPL7, RPL26L1,* and *RPL19*) are key factors of "ME-RPL," acyl-CoA superfamily genes (*ACACA, ACAT1, ACAA2,* and *ACACB*) and *FASN* are key factors of "ME-FASN" and superfamily of solid carrier genes (*SLC17A6, SLC32A1, SLC12A5,* and *SLC6A1*) and complement genes (*C4A, C3,* and *C2*) are key factors of "ME-SLC." In conclusion, the present study discovered the important signal transductions (fatty acid metabolic pathway and CAMs signaling) and genes (ribosomal protein family and the complement system) in tree shrew CNV. We consider that our findings hold implications in unraveling molecular mechanisms that underlie occurrence and development of CNV.

**Keywords: choroidal neovascularization, transcriptome sequencing, bioinformatics, tree shrew, signal transduction**

OPEN ACCESS

***Edited by:***
*Minxian Wallace Wang,*
*Broad Institute, United States*

***Reviewed by:***
*Anton Lennikov,*
*University of Missouri, United States*
*Lei Tian,*
*Stanford University, United States*

****Correspondence:***
*Jiejie Dai*
*djj@imbcams.com.cn*
*Min Wu*
*ynwumin@126.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:*** *18 January 2021*
***Accepted:*** *15 April 2021*
***Published:*** *10 May 2021*

***Citation:***
*Jia J, Qiu D, Lu C, Wang W, Li N, Han Y, Tong P, Sun X, Wu M and Dai J (2021) Transcriptome Analysis of Choroid and Retina From Tree Shrew With Choroidal Neovascularization Reveals Key Signaling Moieties. Front. Genet. 12:654955. doi: 10.3389/fgene.2021.654955*

## INTRODUCTION

Choroidal neovascularization (CNV) is the formation of new blood vessels in the choroid. The blood vessels form/develop between the retinal pigment epithelium (RPE) and Bruch's membrane. They extend through the retinal neuroepithelial layer eventually forming a fibrous vascular tissue. CNV often occurs in the macular area, thereby causing macular hemorrhage and serous exudation under the retina. CNV is a common characteristic of many fundus diseases, such as age-related macular degeneration, high myopic maculopathy, central exudative chorioretinopathy, and diabetic retinopathy; it renders grave vision impairment in the affected individuals.

Although the pathogenesis of CNV remains poorly understood, it is believed to involve a variety of cell growth factors, such as the vascular endothelial growth factor (VEGF), basic fibroblast growth factor (bFGF), and platelet-derived growth factor (PDGF), available in the local microenvironment (Ming et al., 2011). Moreover, inflammatory cells and cytokines are also involved in its pathogenesis, while tumor necrosis factor-α (TNF-α), interleukins 6 (IL6), and intercellular cell adhesion molecules 1 (ICAM-1) released by macrophages and neutrophils promote the occurrence and development of CNV (Jin et al., 2010).

The early stage of CNV is characterized by changes in the retinal microenvironment along with production of VEGF by the RPE and photoreceptors (PRs; Nagineni et al., 2012). Further, VEGF promotes the migration of macrophages to the Bruch's membrane, eventually leading to proteolytic degradation of the membrane (Grossniklaus et al., 2002). However, the key molecular mechanisms underlying CNV and its signaling between retina and choroid remain unclear.

Animal models are important tools in investigating pathogenesis of CNV. The most commonly used method to establish a CNV model is laser-induced selective destruction of photoreceptors, RPE cells, Bruch's membrane, and choroidal capillaries. Destruction of the Bruch's membrane stimulates a series of damage repair processes that leads to the formation of new blood vessels (Archer and Gardiner, 1981; ElDirini et al., 1991; Yang et al., 2006; Hoerster et al., 2012; Liu et al., 2013; Kunbei et al., 2014; Poor et al., 2014). Pathogenesis of CNV in humans is same as that in animals, wherein an injury to the Bruch's membrane-RPE-choroidal complex leads to an imbalance between angiogenic and inhibitory factors, resulting in a series of neovascularization processes, such as endothelial cell proliferation and migration, and lumen formation.

Animals that are most often used for CNV experiments are monkeys, rabbits, and mice. Although the monkey model manifests changes/characteristics similar to that by humans, their application in CNV studies is limited owing to their high cost of maintenance and low incidence of CNV (Archer and Gardiner, 1981; Kunbei et al., 2014). Further, the CNV rabbit model fails to exhibit typical leakage characteristics (less than 30%) as observed by FFA. In addition, retinal vascular circulation of mice and rabbits is disparate from that of humans; therefore, mice and rabbits are not ideal models for CNV studies (Zhu et al., 1989; ElDirini et al., 1991; Tamai et al., 2002; Yang et al., 2006; Hoerster et al., 2012; Liu et al., 2013; Poor et al., 2014). Tree shrew is a small mammal sharing evolutionary and anatomical similarities with primates. Tree shrews have a well-developed visual system, with cone cells accounting for 96% of all the photosensitive cells. These mammals have good color vision as well as stereoscopic vision. In recent years, tree shrews have been used for investigating visual development and ophthalmic diseases (Lin et al., 2013; Jia and Dai, 2019). The retina of tree shrews is mainly composed of cone cells, as mentioned above, and exhibits functions similar to that of human retina. The unique retinal structure of tree shrew provides a good basis for investigating the disease process of human CNV.

In the present study, we aimed to investigate the signal transduction in choroid and retina in a tree shrew CNV model. We found that the genes of ribosomal protein family, superfamily of acyl-CoA, and solute carrier family were differentially expressed. Our findings imply that fatty acid metabolic pathway and CAM pathway play key roles in tree shrew CNV.

## MATERIALS AND METHODS

### Experimental Animals

Fifteen adult tree shrews (*Tupaia belangeri chinensis*) with healthy eyes (seven females and eight males, aged 2–3 years, weighing 110–130 g) were collected from the Tree Shrew Germplasm Resource Center, Institute of Medical Biology, Chinese Academy of Medical Sciences, Kunming, China. The experimental animal production licenses were SCXK (Dian) K2018-0002 and use license SYXK (Dian) K2018-0002. All animal experiments were approved by the Animal Ethics Committee of the Institute of Medical Biology, Chinese Academy of Medical Science (approval number: DWSP201803019). The tree shrews were randomly divided into three groups (7, 21, and 30 days after laser photocoagulation) such that each group comprised five tree shrews (10 eyes per group, all left eyes were used as the experimental group and all right eyes as the control group).

### Laser-Induced CNV Model

The animals were anesthetized by intraperitoneal injection of 0.3% pentobarbital sodium (0.6 mg/kg), and the body temperature was maintained by a heating pad. Five minutes before laser photocoagulation, compound tropicamide eye drops were used to disperse the pupil of both eyes. Eyes were anesthetized with two drops of 4 mg/ml oxybuprocaine hydrochloride eye drops (Santen Pharmaceutical Co., Ltd., China) and carbomer eye drops (Dr. Gerhard Mann, Chem.-Pharm. Fabrik GmbH, Germany) was used to prevent corneal dryness. After the animals were fixed, laser photocoagulation (Lumenis, United States) was performed at 1 PD around the optic disc, 15 spots in total. The laser wavelength was 647.1 nm, diameter of the spots was 50 μm, and the exposure time was 0.01–0.05 s. Energy of the laser was adjusted according to the retinal reaction. The effective spots were marked by the formation of bubbles without bleeding after photocoagulation. Eyes with ruptured capillaries small blood vessels were not used for subsequent experiments (Fundus camera: TOPCON, United States).

## Hematoxylin and Eosin Staining

Eyeballs of tree shrew corneas were fixed with FAS (Servicebio, China) for 24 h, and 3-um-thick of were cut. The eyeball sections were dehydrated, cleared, embedded, and made into sections. The sections were used for hematoxylin and eosin (HE) staining and CD31 immunohistochemistry (IHC). The pathological changes were described by a section scanner (Mantra, United States).

## CD31 Immunohistochemistry

In order to confirm the occurrence of CNV after photocoagulation, IHC for CD31 was conducted.

Antigen repair was performed in citric acid sodium buffer solution (ph6.0; Servicebio, China). Sections were incubated in 3% hydrogen peroxide at room temperature for 25 min in dark to block endogenous peroxidase, and then blocked with 3% BSA.

Sections were incubated in 1:300 dilution of Anti-CD31 Rabbit pAb (Servicebio, China) at 4°C overnight, and then incubated in 1:200 dilution of HRP conjugated Goat Anti-Rabbit IgG (H + L; Servicebio, China) for 50 min at room temperature. Color development was performed using DAB (Servicebio, China). The result of CD31 IHC was observed under microscope (Mantra, United States), and inform was used for processing the image. Microvessel density (MVD) were counted refer to Weidner et al. (1993).

## RNA Sequencing of Retina and Choroid of Tree Shrew CNV Model

### Total RNA Extraction

Tree Shrews were euthanized by injection of 2% pentobarbital sodium (2 mg/kg). Tree shrews were disinfected and the eyeballs were rinsed with iodophor. The canthus and the conjunctiva were cut. The conjunctiva and fascia were separated so that the sclera was fully exposed. The optic nerve was cut and then the eyeball was removed.

The eyeballs were suspended in PBS in clean bench. After the cornea was cut, the lens and vitreum were separated. The retina-choroid-sclera was transferred to RNase-free water (TaKaRa, Japan) to isolate the retina (soft and transparent). Since the choroid was fragile and easy to fall off, the choroid-sclera was transferred to TRIzol and choroid (brown) was isolated.

Total RNA was extracted from the collected samples using TRIzol according to the standard protocol (Invitrogen, United States). Purity of the extracted RNA was spectrophotometrically quantified with NanoPhotometer (Implen, United States). RNA integrity was determined using Bioanalyzer 2100 system (Agilent Technologies, United States). In case the RNA content was insufficient (RNA < 2 μg), the sample was mixed with other samples of the same group.

### cDNA Library Preparation and Sequencing

First-strand DNA was synthesized using random primers and M-MuLV reverse transcriptase (RNase H-). Second-strand DNA was synthesized by DNA polymerase I and RNase H, wherein dTTPs were replaced by dUTPs. Further, the dU-containing second-strand cDNA was degraded by the USER enzyme, and the cDNA was PCR-amplified to obtain a library.

The cDNA was quantified using Qubit 2.0, and insert size of the library was determined by diluting the library to 1 ng/μl and subjecting to Agilent 2100 system. Effective concentration of the cDNA library was quantified accurately by qPCR (effective library concentration > 2 nM) to ensure quality of the cDNA library. Twenty-six DNA libraries were constructed. The libraries were sequenced on an Illumina HiSeq 4000 platform at the Novogene Bioinformatics Institute (Beijing, China).

## Differential Expression of mRNA in the Retina and Choroid

### Cluster Analysis

Subread package (Liao et al., 2014) was used to filter out the sequences with adapter, poly-n > 10%, and low quality of RNA sequencing to get clean data. TopHat v2.0.9 (Kim et al., 2013) was used to align clean data with tree shrew genome sequences (http://www.treeshrewdb.org/, Accession number: CRP000902), genome annotation in http://www.treeshrewdb.org/data/Chinese_treeshrew_function_annotation_information_2.0.txt.gz. The mapped fragments were spliced using the Cufflinks V2.1.1 (Trapnell et al., 2010). Gene expression level was quantitatively analyzed by Fragments Per Kilobase of exon model per Million (FPKM; mapped fragments; Bray et al., 2016). After statistical analyses (Anders and Huber, 2010; Robinson et al., 2010; Love et al., 2014), differentially expressed genes (DEGs) in the retina and choroid were selected at $|\log_2(\text{fold change})| > 1$ and Padj < 0.05.

We perform heatmap clustering with distance methods of pre-defined distance method (1-pearson), and clustering method of clustering function [cluster_rows = diana(mat) and cluster_columns = agnes *t*(mat)] from cluster package (Gu et al., 2016).

### GO and KEGG Analyses

In order to determine the functions of the DEGs, clusterProfiler was used to analyze gene ontoloagy (GO) function and kyoto encyclopedia of genes and genomes (KEGG) pathway enrichment. GO analyzes the biological processes, cellular components, and molecular functions. KEGG analysis determines the signal pathways enriched by the DEGs, and it is expected to find the biological pathways that play key roles in tree shrew CNV. Threshold of GO/KEGG enrichment was considered significant at Padj < 0.05.

### Protein-Protein Interaction Network Construction

STRING database[1] provides comprehensive information about interactions between proteins. In the present study, STRING was used to generate PPI networks among the differentially expressed mRNAs based on interactions with combined scores > 0.4. Additionally, Cytoscape was used to visualize the network, and PPI scores > 800 were considered for functional enrichment analysis of the modules.

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

### Weighted Gene Co-expression Network Analysis

Hierarchical cluster tree was established using WGCNA based on the correlation of gene expressions at the Pearson correlation coefficient 0.8 and soft threshold (power) 9. Functions were considered relevant when the genes demonstrated a high degree of co-expression correlation in a module. The modules were clustered and the correlation heatmap among modules was constructed to evaluate the connectivity between two genes within the module according to the module eigenvalues.

We analyzed the expression trends of DEGs in the modules during different courses of CNV.

### Reverse Transcription Quantitative PCR

The results of RNA sequencing were validated by reverse transcription quantitative PCR (RT-qPCR) using CFX96 PCR system (Bio-Rad, United States) and One Step TB Green PrimeScript PLUS RT-PCR Kit (TaKaRa, Japan). Five significantly differentially expressed transcripts were selected (*FASN, ACACB, ACAT1, ACAA2*, and *IL18*; **Table 1**). The reaction mixture (25 μl) contained 12.5 μl 2 × One Step TB Green RT-PCR Buffer, 0.5 μl PrimeScript PLUS RTase Mix, 1.5 μl TaKaRa Ex Taq HS Mix, 6 μl RNase-free dd$H_2O$, 2 μl RT-qPCR primers, 0.5 μl ROX reference dye, and 1 μg total RNA.

The reaction was performed at the following conditions: initial cDNA synthesis at 45°C for 5 min; initial denaturation at 95°C for 10 s; 40 cycles of denaturation at 95°C for 5 s, annealing and extension at 60–61°C for 30 s; 95°C for 15 s, 60°C for 60 s, and 95°C for 15 s for the dissolution curve. The internal reference gene was *GAPDH*. Each experiment was performed in triplicate. PCR primers and amplification conditions used in this study are shown in **Table 1**.

## Statistical Analysis

Statistical analyses were performed using SPSS 20.0. All data are expressed as mean ± SEM. $p < 0.05$ was considered statistically significant. Student's *t*-test was used for analyzing RT-qPCR and MVD results. GraphPad Prism 8.0 was used to prepare the graphs.

**TABLE 1** | Reverse transcription quantitative PCR (RT-qPCR) primer sequences and amplification conditions.

| Transcript | Primers | Sequence (5′ to 3′) | Annealing temperature (°C) | Product length (bp) |
|---|---|---|---|---|
| *ACACB* | Forward | CATCCGCGCCATTATCAG | 60 | 171 |
| | Reverse | GGCACGAAGTTGAGGAAG | | |
| *ACAT1* | Forward | CACTGCCAGCCACTAAACTTG | 60 | 119 |
| | Reverse | TCCTTCGCCTCCTTGTAGAAC | | |
| *ACAA2* | Forward | CACTGCTACTGACTTGAC | 60 | 103 |
| | Reverse | CTGCATGACATTGCCTAC | | |
| *IL18* | Forward | TTAGAGGTCTGGCTGTAG | 61 | 111 |
| | Reverse | CGCTGATATTGTCAGGAG | | |
| *FASN* | Forward | CTGCTGCTGAAACCTACC | 60 | 200 |
| | Reverse | CAGCCGTCTGTGTTAGTG | | |
| *GAPDH* | Forward | CTTCAACTCTGGCAAGGT | 60 – 61 | 279 |
| | Reverse | AAGATGGTGATGGACTTCC | | |

# RESULTS

## Pathology of the Tree Shrews CNV

In the control groups, tree shrews have intact retina (**Figure 1L**). After 7 days of photocoagulation, retinal edema (**Figures 1B,F**), damaged haller's layer in choroid (**Figures 1A,E**) was observed. After 14 days of photocoagulation, neovascularization in retina leads to destruction of ganglion cell (GCL) layer and disorder of nerve fiber layer (NFL) layer. CNV was developed, and neovascularization surrounded by neutrophils, macrophages, and fibroblasts was observed found in choroid (**Figures 1C,G**). Haller's layer and sattler's layer was loose in choroid (**Figures 1D,H**). After 30 days of photocoagulation, retinal neovascularization was developed (**Figure 1I**). Neutrophils and macrophages were seen around the neovascularization in choroid. Damaged choroidal structure was not recover (**Figures 1J,K**).

## IHC for CD31

The normal tree shrews have complete retina, choroid and sclera with distinct layers, and no neovascularization was observed (**Figure 2A**). After 7 days of photocoagulation, retinal edema was observed in the photocoagulation sites, and the choroid structure was destroyed. Neovascularization was observed in the subchoroid (**Figure 2C**) and scleral (**Figure 2B**), but no neovascularization was observed in the choroid and retina. After 14 days of photocoagulation, retinal edema was still observed in the photocoagulation sites. Neovascularization with small size was obvious in the retina (**Figures 2D,F**) and choroid (**Figure 2D**), while the neovascularization in the sclera was larger than that in 7 days of photocoagulation group (**Figure 2E**). After 21 days of photocoagulation, increased retinal (**Figure 2I**) and choroid (**Figure 2H**) neovascularization in the photocoagulation sites. The lumen of retinal neovascularization was larger than that in 7 days of photocoagulation group (**Figure 2G**).

Microvessel density in laser photocoagulation group was higher than that in control group ($p < 0.05$). MVD in 14 days of photocoagulation group (7.00 ± 1.00) were higher than that in 7 days of photocoagulation group (3.33 ± 0.58). However, MVD were no significant difference between 14 days of photocoagulation group and 21 days of photocoagulation group (5.33 ± 2.31).

## RNA Sequencing

Raw reads were generated from 36 samples, including 18 choroids and 18 retinas. The average clean data of each sample were not less than 7 GB. The average number of clean reads was 46,719,819; Q20 was higher than 97%; Q30 was higher than 94%.

## Quantitative Assessment by PCR

Relative expressions of *FASN, ACACB, ACAT1, ACAA2*, and *IL18* are shown in **Figure 3**. The expression of *ACAA2* was significantly upregulated after 7 days ($p < 0.001$) and downregulated after 30 days ($p = 0.022$) of photocoagulation

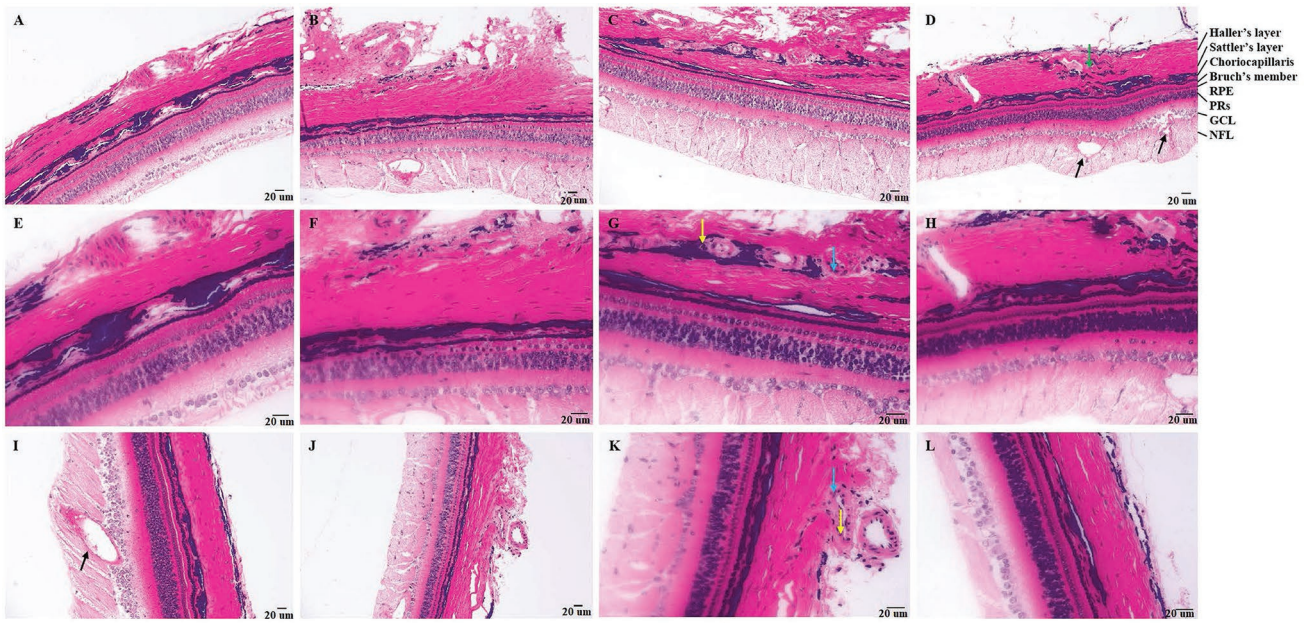


**FIGURE 1 |** Pathology of tree shrews with choroidal neovascularization (CNV). [**A** and **B** (200×), **E** and **F** (400×)] Pathology of tree shrew CNV after 7 days of laser photocoagulation. [**C** and **D** (200×), **G** and **H** (400×)] Pathology of tree shrew after CNV 14 days of laser photocoagulation. [**J** (200×), **I** and **K** (400×)] Pathology of tree shrew CNV after 21 days of laser photocoagulation. [**L** (400×)] Intact retina and choroid with clear structure of normal tree shrew. The yellow arrow indicates neutrophils, the blue arrow indicates macrophages, the black arrow indicates neovascularization, and the green arrow indicates loose haller's and sattler's layer in choroid. RPE, Retinal pigment epithelium; PRs, Photoreceptor; GCL, ganglion cell; and NFL, nerve fiber layer.

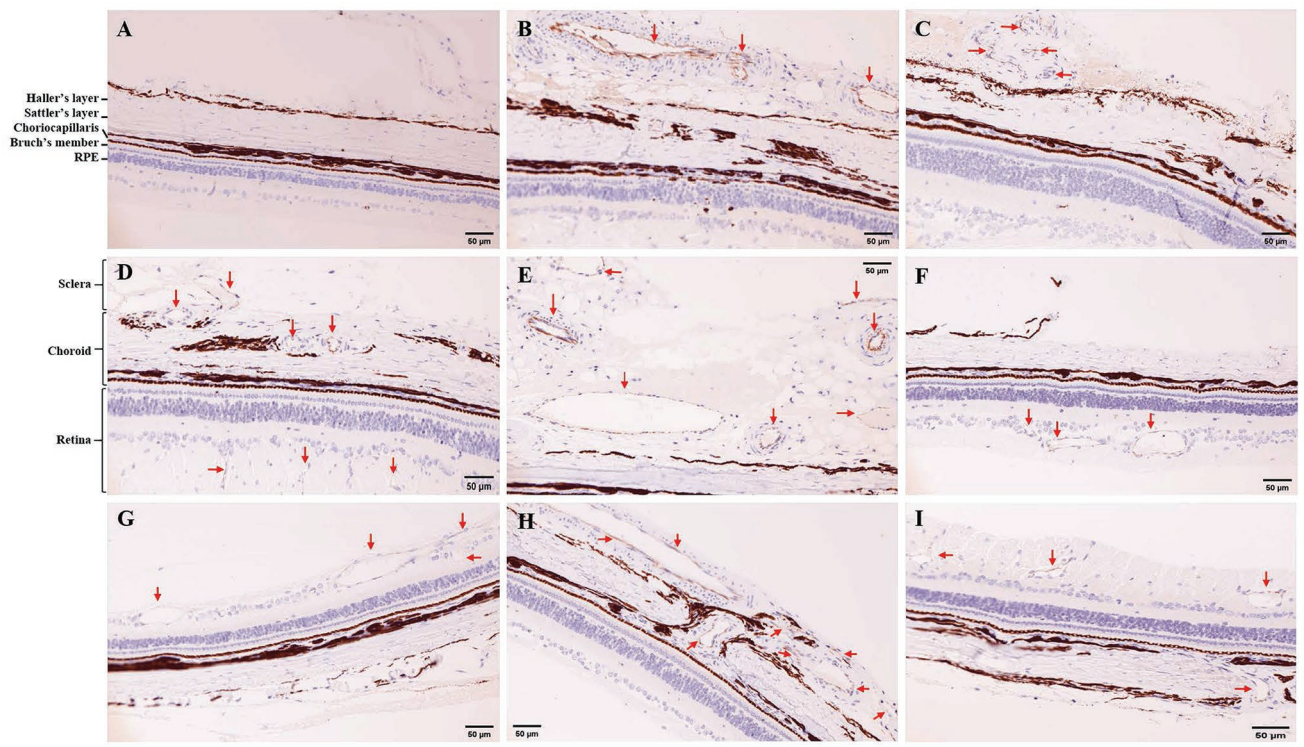


**FIGURE 2 |** Immunohistochemistry (IHC) for CD31 in tree shrew retina and choroid (200×). **(A)** IHC in control group. **(B,C)** IHC in 7 days of photocoagulation group. **(D,E,F)** IHC in 14 days of photocoagulation group. **(G,H,I)** IHC in 21 days of photocoagulation group. The red arrow indicates neovascularization. RPE, Retinal pigment epithelium.

in choroid (**Figure 3A**), whereas it was upregulated after 21 days ($p < 0.001$) and downregulated after 30 days ($p = 0.007$) of photocoagulation in the retina (**Figure 3B**).

The expression of *ACACB* was significantly upregulated after 21 days ($p = 0.027$) of photocoagulation in choroid (**Figure 3C**), and upregulated after 21 days ($p = 0.006$), but downregulated after 30 days ($p < 0.001$) of photocoagulation in retina (**Figure 3D**).

The expression of *FASN* was upregulated after 7 days ($p = 0.048$) in the choroid (**Figure 3F**), but downregulated after 21 days ($p < 0.001$) and 30 days ($p < 0.001$) of photocoagulation in the retina (**Figure 3G**).

The expression of *IL18* was upregulated after 21 days ($p < 0.022$) and 30 days ($p = 0.022$) of photocoagulation in choroid (**Figure 3H**), but downregulated after 7 days ($p < 0.025$) and 21 days ($p < 0.001$) of photocoagulation in the retina (**Figure 3I**).

The expression of *ACAT1* was upregulated after 7 days ($p = 0.039$) and 21 days ($p = 0.039$) of photocoagulation in the choroid (**Figure 3E**), but downregulated after 21 days ($p = 0.001$) of photocoagulation in the retina (**Figure 3J**). The qPCR results were consistent with the RNA sequencing data.

## Cluster Analysis of DEGs

The mRNA expression levels were estimated with FPKM. In the choroid, 335 mRNAs, 9 mRNAs, and 6 mRNAs were differentially expressed after 7, 21, and 30 days of laser photocoagulation, respectively. In the retina, 6, 56, and 7 mRNAs were differentially expressed after 7, 21, and 30 days of photocoagulation, respectively. A cluster analysis of the differentially expressed mRNAs was conducted and the results are shown as heatmaps in **Figure 4**.

## WGCNA

### Hierarchical Cluster Tree

We obtained a total of 18 gene modules and found that the DEGs were mainly clustered in the Module-RPL ("ME-RPL"), Module-SLC ("ME-SLC"), and Module-FASN ("ME-FASN") modules. These modules contained 37, 184, and 91 significant DEGs, respectively (**Figure 5**).

### Heatmap of Inter-Module Correlation

Correlation analysis between the modules and samples showed that "ME-RPL," "ME-FASN," and choroid are highly (and positively) correlated, while "ME-SLC" and retina are highly (and positively) correlated, as depicted in **Figure 6**.

### Cluster Heatmap of the Gene Modules

Analysis of interactions among the gene modules revealed that when the color of the region within and between the three modules is evident, the genes in these modules have a high degree of association. It is suggested that "ME-RPL," "ME-SLC," and "ME-FASN" modules interact closely with each other. The results in the form of heatmap visualization are shown in **Figure 7**.

### Expression Patterns of the Gene Modules

In order to analyze the expression trend of the genes in different modules, the module gene expression pattern was constructed. The results showed that "ME-RPL" genes were downregulated in the retina but upregulated in choroid; "ME-SLC" genes were upregulated in the retina but downregulated in choroid; and "ME-FASN" genes were downregulated in retina but upregulated in choroid, as shown in **Figure 8**.

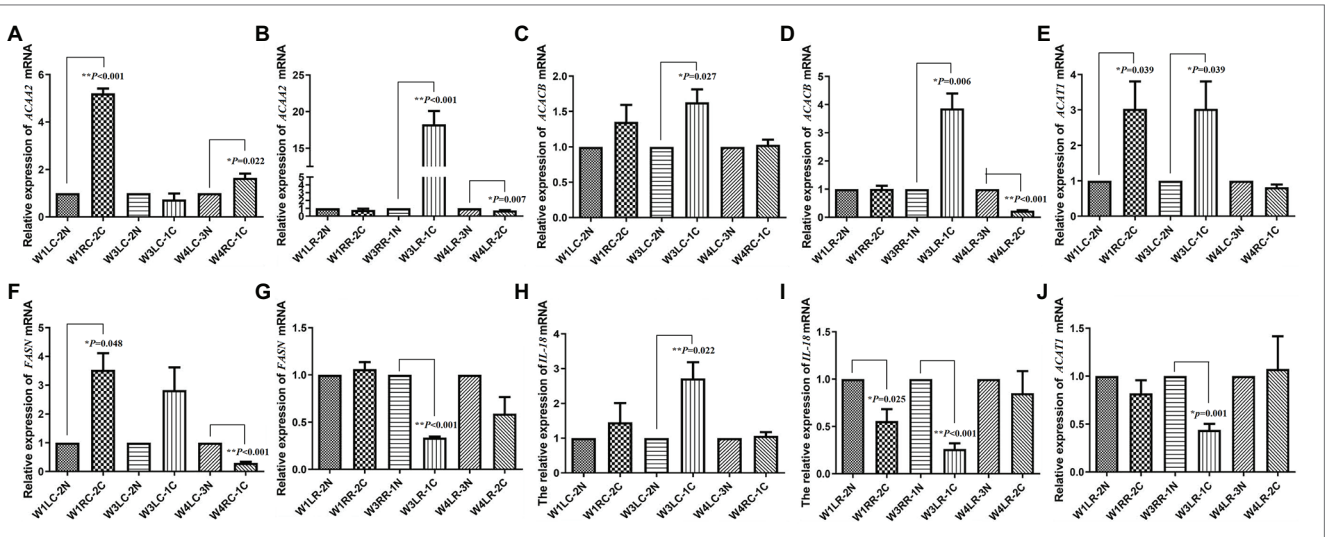


**FIGURE 3 |** Validation of transcript expression by quantitative PCR (qPCR). **(A,B)** The expression of *ACAA2* mRNA in tree shrew choroid and retina, respectively; **(C,D)** The expression of *ACACB* mRNA in tree shrew choroid and retina, respectively; **(E,J)** The expression of *ACAT1* mRNA in tree shrew choroid and retina, respectively; **(F,G)** The expression of *FASN* mRNA in tree shrew choroid and retina, respectively; **(H,I)** The expression of *IL18* mRNA in tree shrew choroid and retina, respectively. Data are presented as mean ± SEM ($n = 3$). *GAPDH* was used as a housekeeping internal control. Transcript expression was quantified relative to the expression level of *GAPDH* by $2^{-\Delta\Delta CT}$ method. $0.001 \leq {}^{*}p < 0.05$; ${}^{**}p < 0.001$.

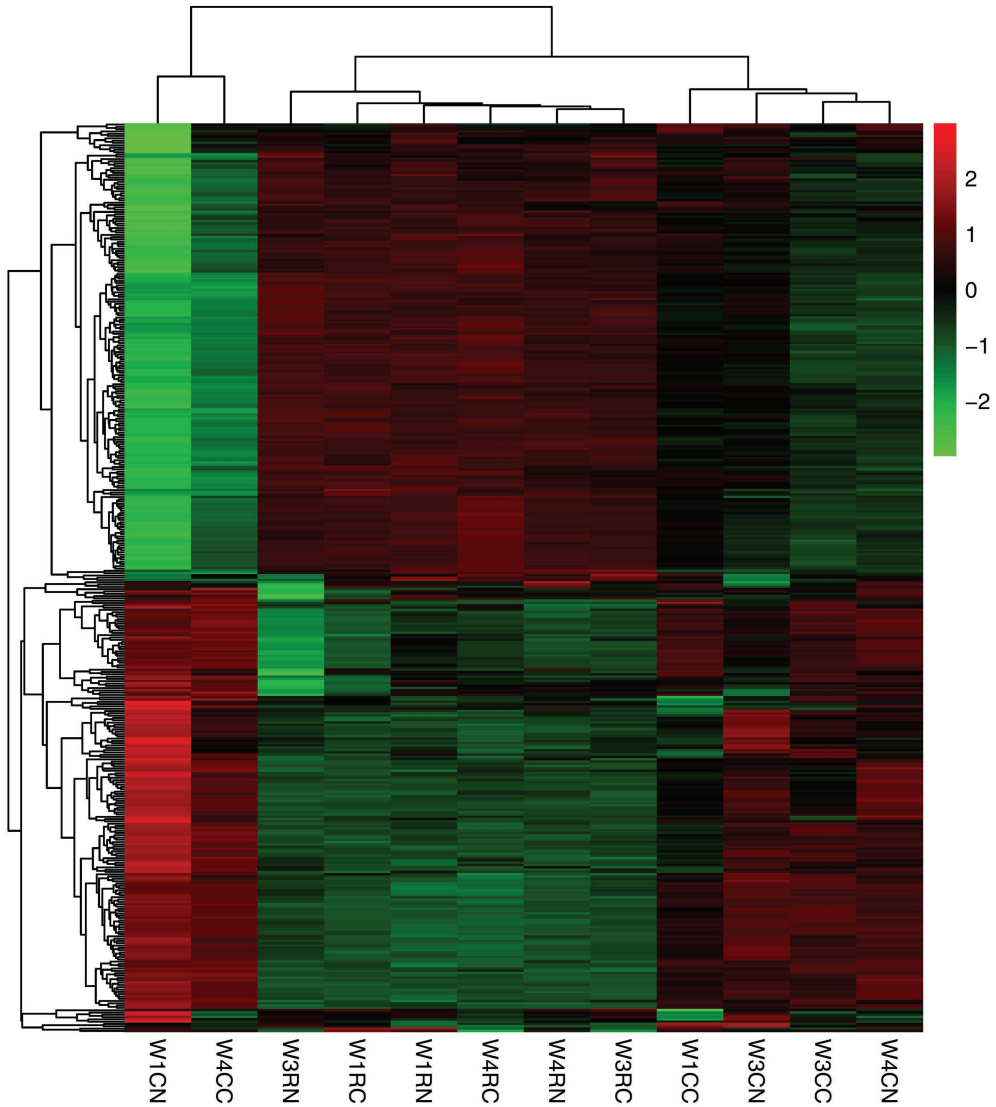


**FIGURE 4 |** Cluster analysis of differentially expressed mRNAs. Red indicates upregulated mRNAs; green indicates downregulated mRNAs. W1CC is choroid after 7 days of laser photocoagulation, W1CN is another choroid collected from the same tree shrew as W1CC. W3CC and W4CC are choroids after 21 and 30 days of laser photocoagulation, respectively. W3CN and W4CN are choroids collected from the same tree shrews as W3CC and W4CC, respectively. W1RC is retina after 7 days of laser photocoagulation, W1RN is another retina collected from the same tree shrew as W1RC. W3RC and W4RC are retina after 21 and 30 days of laser photocoagulation, respectively. W3RN and W4RN are retina collected from the same tree shrews as W3RC and W4RC, respectively.

## GO and KEGG Analyses of DEGs

GO analysis showed that the differentially expressed mRNAs in the choroid were mainly involved in membrane transport processes, while there was no significant enrichment for differentially expressed mRNA in the retina (**Figure 9**).

KEGG analysis showed that "ME-RPL" genes were enriched in IgA production, antigen presentation, and CAMs pathways. "ME-FASN" genes are enriched in fatty acid metabolism and PPAR signaling pathway, while "ME-SLC" genes were enriched in GABAergic synapse formation/development, neuroactive ligand-receptor interaction, cholinergic synapse, and retrograde endocannabinoid signaling pathway (**Figure 10**).

## PPI Network Analysis

The PPI network of "ME-FASN" consisted of 69 genes and 102 interactions, including superfamily of acyl-CoA (*ACACA, ACAT1, ACAA2, ACACB, ACADS,* and *ACADVL*) and superfamily of solid carrier genes (*SLC6A3, SLC25A20,* and *SLC27A2*; **Figure 11**). "ME-RPL" consisted of 28 genes and 23 interactions, including ribosomal protein family (*RPL31, RPL7, RPL26L1,* and *RPL19*), complement genes (*C4A, C1S,* and *C1R*), and integrin genes (*ITGA5, ITGB6,* and *ITGAV*; **Figure 11**). "ME-SLC" consisted of 20 genes and 17 interactions, including superfamily of solid carrier genes (*SLC17A6, SLC32A1, SLC12A5,* and *SLC6A1*) and complement genes (*C4A, C3,* and *C2*; **Figure 12**).

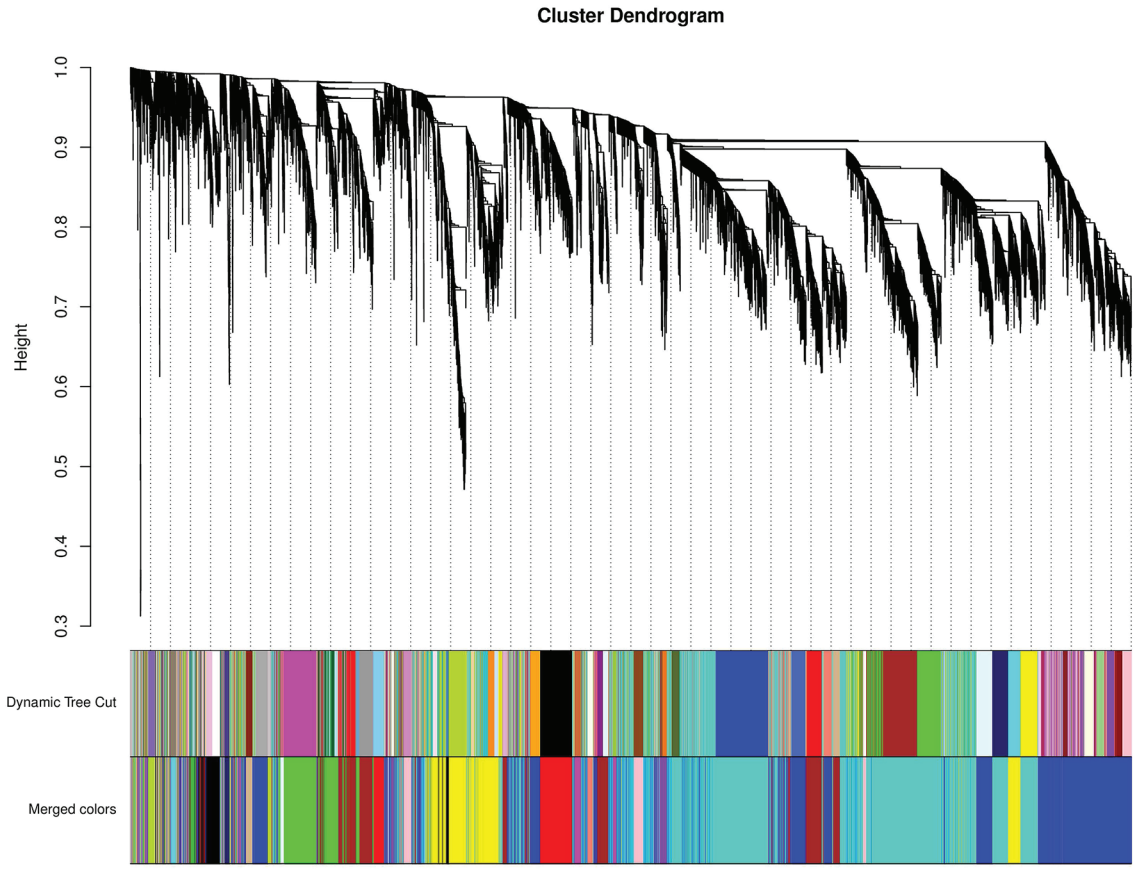


**FIGURE 5 |** Cluster dendrogram of differentially expressed mRNAs. Upper portion of the figure is the gene cluster tree. A leaf represents a gene while the branches represent different gene modules. Different colors in the middle portion of the dynamic tree cut represent different modules. Merged colors (lowermost portion) represent the merged modules calculated when the dissimilarity coefficient < 0.25. Yellow tree represents Module-FASN ("ME-FASN") module; Blue tree represents Module-RPL ("ME-RPL") module; Turquoise tree represents Module-SLC ("ME-SLC") module.

## DISCUSSION

Pathogenesis of CNV is influenced by the choroid microenvironment that is composed of extracellular matrix, cytokines, PDGF, fibroblast growth factors, tumor necrosis factor-α (TNF-α), pigment epithelium-derived growth factor, interleukins, the complement system, ephrins, and angiopoietins (Lambert et al., 2016). Reportedly, CNV-related signaling pathways mainly include the VEGF pathway, transforming growth factor (TGF)-β/Smad pathway (Wang et al., 2017), Wnt pathway (Zhou et al., 2010), sonic hedgehog (Shh) pathway (Yang et al., 2012; Holliday et al., 2013), and notch pathway (Dou et al., 2016). VEGF is the most important cytokine for neovascularization that can stimulate the proliferation and migration of vascular endothelial cells and promote capillary lumen formation (Jie et al., 2004). TNF-α contributes to CNV by upregulating VEGF production in RPE cells through ROS-dependent activation of β-catenin signaling (Haibo et al., 2016). Wnt3a has been shown to activate β-catenin, upregulate VEGF and TNF-α, and increase oxidative stress (Zhou et al., 2010) while WNT7A/B promotes CNV (Joseph et al., 2018). FGF21 inhibits TNF-α expression but does not alter VEGFA expression in neovascular eyes, and FGF21 administration suppresses pathological retinal neovascularization and CNV in mice (Fu et al., 2017). Following Bruch's membrane injury, RPE cells release IL6, IL8, and growth factors (TGF-β) thereby exerting strong chemotactic effects on macrophages and neutrophils involved in the neovascularization (Hollyfield et al., 2008; Zarranz-Ventura et al., 2013). It is reported that the lack of TGF-β signaling in retinal microglia can cause retinal degeneration and aggravate CNV (Ma et al., 2019). Several studies have found that TGF-β significantly enhances VEGF secretion, vascular permeability, and extracellular matrix remodeling on its own or in concert with other cytokines, such as TNF-α (Walshe et al., 2009; Suzuki et al., 2012).

In the present study, we showed that the expression of *IL18* increased significantly after 21 days of laser photocoagulation in the choroid, although it decreased significantly after 7 and 21 days of laser photocoagulation in the retina. In mouse model of CNV, the expression of *IL18* decreased significantly in RPE-choroid tissue (Choudhary et al., 2015). It has been reported that *IL18* regulates pathological retinal neovascularization. In addition, IL18 inhibits the

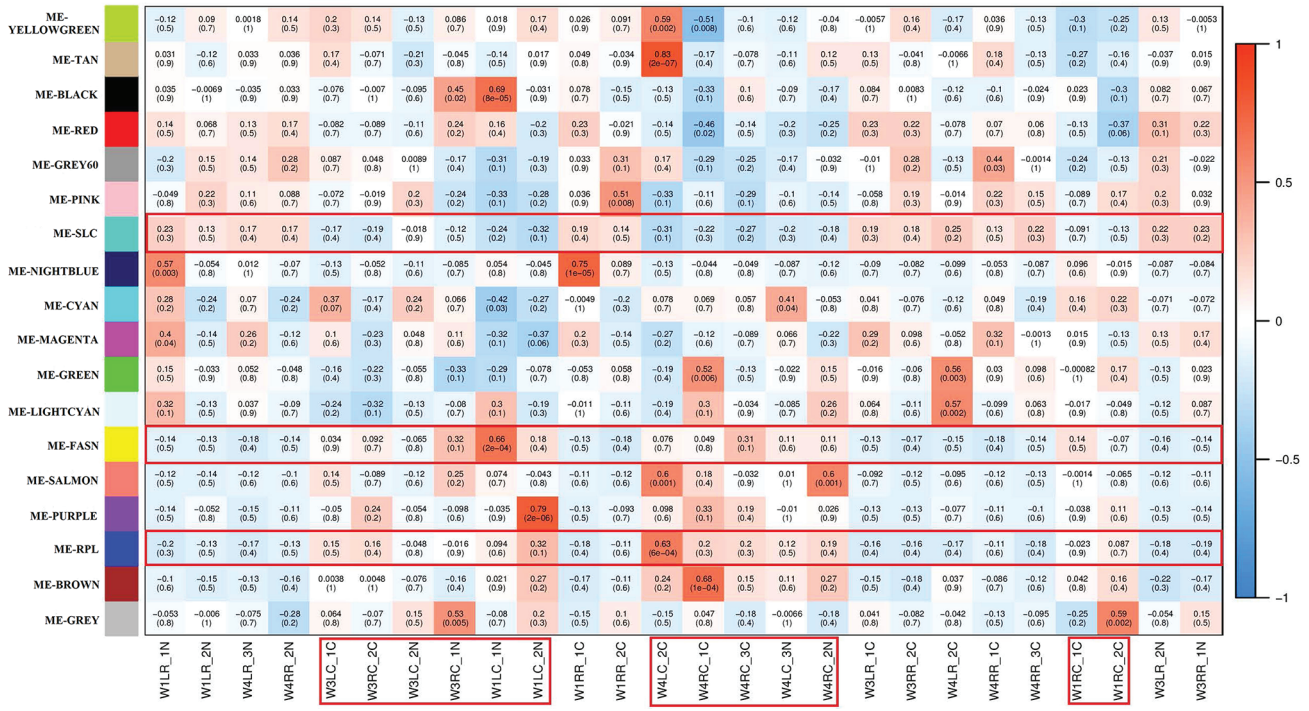


**FIGURE 6 |** Module – Sample relationships. The abscissa is the sample and the ordinate is the module. The numbers represent correlation between the module and the sample. Closer the value is to 1, stronger the positive correlation between the module and the sample; closer to −1, stronger the negative correlation. The number in brackets represents value of *p*, smaller the value, and higher the significance. The red specimen is choroid of the tree shrews.

formation of experimental CNV through the stimulation of interferon-γ and thrombospondin (Xiao and Wu, 2016). IL18 reduce CNV development in the nonhuman primate eye, and inhibits vascular leakage in a mouse model of spontaneous neovascularization (Doyle et al., 2015). However, the mechanism of IL18 inhibiting CNV formation remains unclear and needs to be further investigated.

In present study, genes in "ME-RPL" had a high correlation with choroid. Moreover, KEGG analysis showed that "ME-RPL" genes were significantly enriched in CAMs pathway. CAMs are associated with neovascularization (Dong et al., 2018), wherein ICAM-1 is closely related to the occurrence and development of CNV (Gao and He, 2015). However, the role of CAMs in CNV remains to be fully understood.

There were significant differences in the expression of ribosomal protein family genes in the "ME-RPL" module, and the interaction among proteins in "ME-RPL" was close. It has been reported that RPL17 inhibits the growth of vascular smooth muscle cells and silencing *RPL17* promotes proliferation of these cells (Smolock et al., 2012). It is suggested that the ribosomal protein family may be involved in the regulation of vascular growth. Concurrent with previous studies, we found that the ribosomal protein family may be involved in the regulation of CNV. However, the underlying mechanism requires further in-depth research.

In the "ME-FASN" module, fatty acid synthase (*FASN*) was the center of PPI network. KEGG analysis showed that many "ME-FASN" DEGs were enriched in the fatty acid metabolic, GABAergic synapse, neuroactive life receptor interaction, cholinergic synapse, and retrograde endocannabinoid pathways. *FASN* regulates tumor angiogenesis by altering secretion and activity of VEGF (Seguin et al., 2012). This study suggests that the acylation involved in fatty acid metabolism may be related to CNV. The acyl-CoA superfamily genes (*ACACA*, *ACACB*, *ACAA2*, and *ACADVL*) in "ME-FASN" module were significantly differentially expressed and had strong interactions with *FASN*. In mouse model of CNV, the *ACACB* and *ACADVL* gene was significantly differentially expressed (Choudhary et al., 2015), which correlate with present study. The present study indicates that fatty acid metabolic pathway and acyl-CoA superfamily genes may play an important role in the occurrence and development of CNV. It is reported that FASN knockdown in endothelial cells elevated malonyl-CoA levels, and reduced pathological ocular neovascularization (Bruning et al., 2018). The expression levels of acyl-CoA oxidase 1 (Acox1), fatty acid binding protein 4 (Fabp4) is associated with the progression of retinal pathological neovascularization in a murine model (Tomita et al., 2019). *SLC27A2* and *C3* gene in "ME-SLC" were significantly differentially expressed in CNV tree shrew models and CNV mouse model (Choudhary et al., 2015).

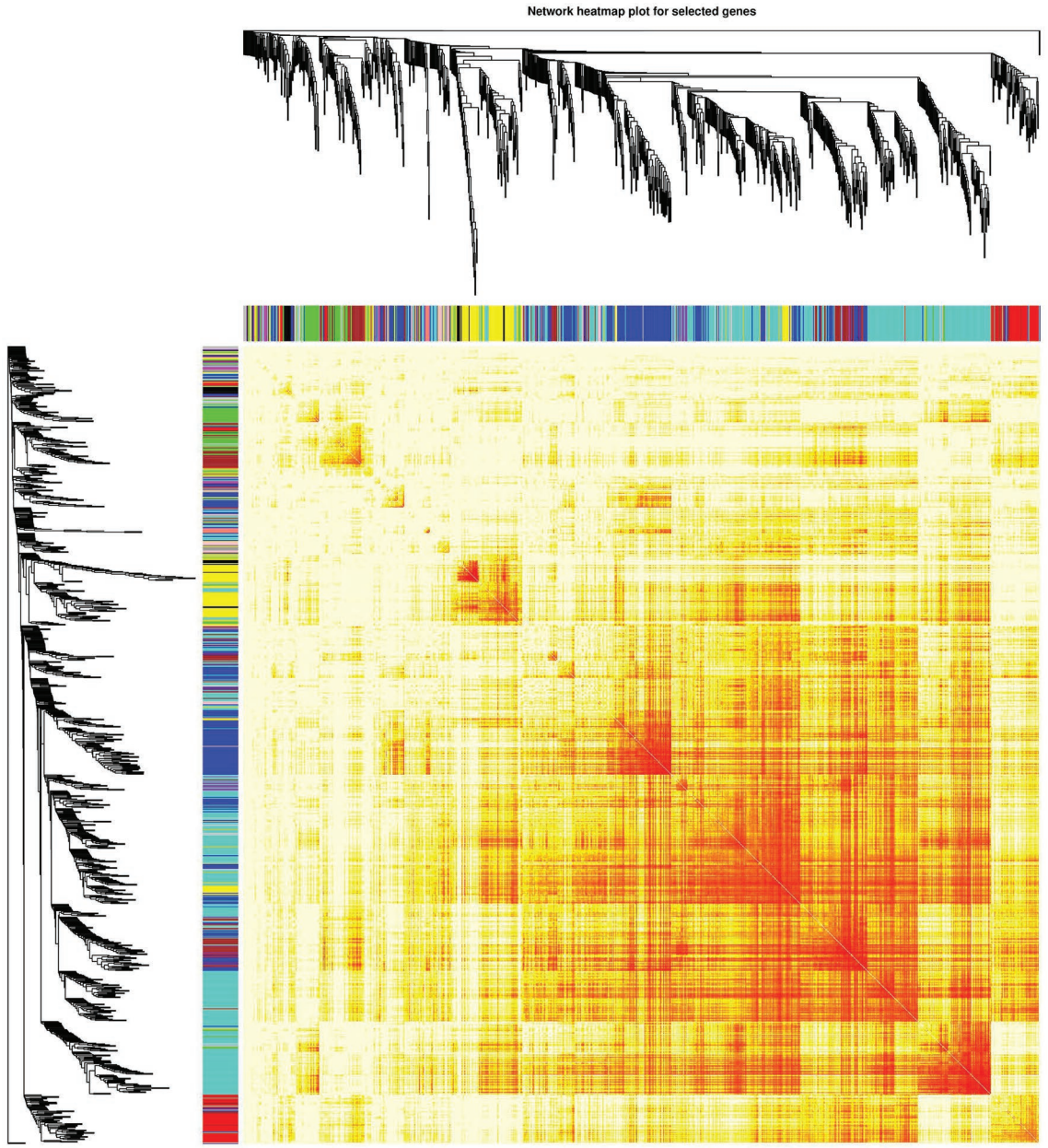


**FIGURE 7 |** Network heatmap of the gene modules. The tree represents a module (top and left), and the branch represents a gene. Darker the color of the dot (white → yellow → red), stronger the connectivity between the two genes corresponding to the row and column. Yellow tree represents "ME-FASN" module; Blue tree represents "ME-RPL" module; Turquoise tree represents "ME-SLC" module.

The result above demonstrated that the transcriptome profile in CNV tree shrew model and in CNV mouse model is consistency. In addition, the result of huvec cell (human umbilical vein endothelial cell) high throughput sequencing in NCBI GEO database (Popovic et al., 2020) showed that *IL18*, *FASN*, *ACADS*, *SLC25A20*, *RPL7*, and *RPL26L1* gene was significantly differentially expressed after treated with DAND5, it indicated that these gene may related to the developmental and pathological ocular angiogenesis. Furthmore, there are 12 genes (*ACACB*, *IL18*, *RPS20*, *ACSL1*, *ACADVL*, *HMGCS1*, *ECI1*, *TKT*, *EIF3I*, *PYCARD*, *CASP4*, and *CFB*) were differentially expressed in tree shrew CNV, mouse CNV (Choudhary et al., 2015), and huvec cell neovascularization models (Popovic et al., 2020).

In conclusion, we constructed the interaction network of "ME-RPL" involved in CAMs pathway and "ME-FASN" genes involved in fatty acid metabolism in choroid, "ME-SLC" genes involved in GABAergic synapse in retina during CNV in tree shrews. Our findings hold implications in unraveling molecular mechanisms that underlie occurrence and

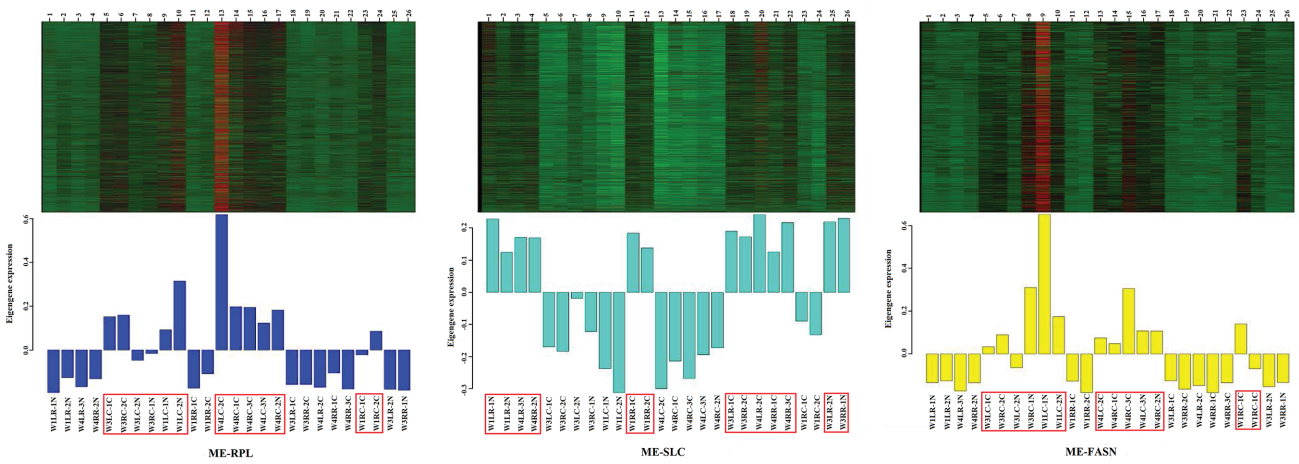


**FIGURE 8 |** Heatmap visualization of eigenvalues of "ME-RPL," "ME-SLC," and "ME-FASN" modules. Upper portion of the figure shows the gene expression heatmaps. Red indicates upregulated; green indicates downregulated. Lower portion of the figure shows the module eigenvalues in different samples, the displayed gene is upregulated/downregulated in the respective module.

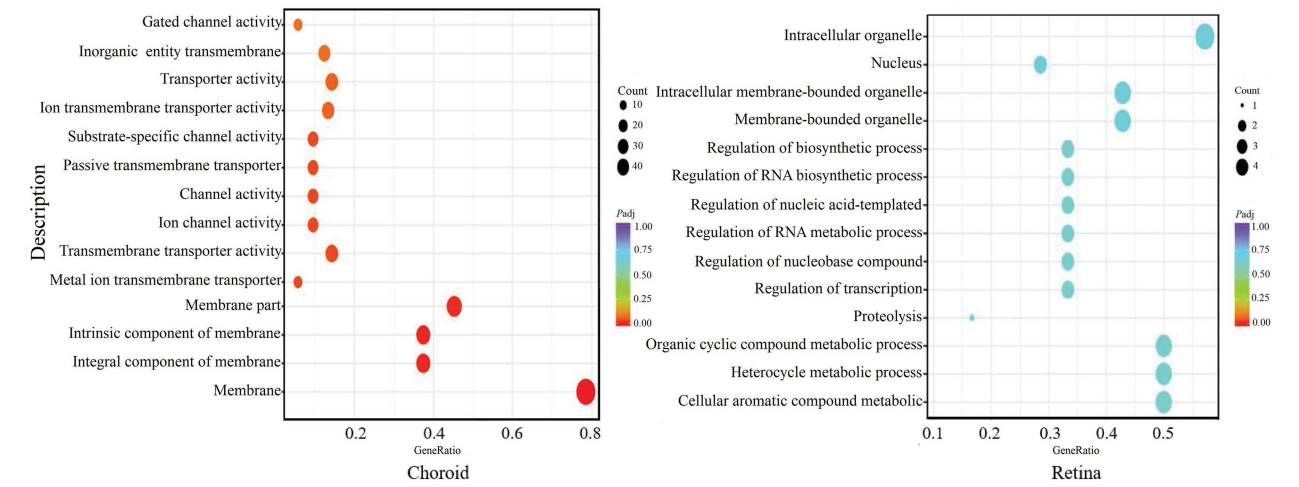


**FIGURE 9 |** GO enrichment of the significant differentially expressed genes (DEGs) in the choroid and retina. The abscissa shows the rich factors and the ordinate represents the pathways. Size of each point indicates the number of candidate target genes in the GO term, and the color of each point corresponds to the value of *p* after correction.

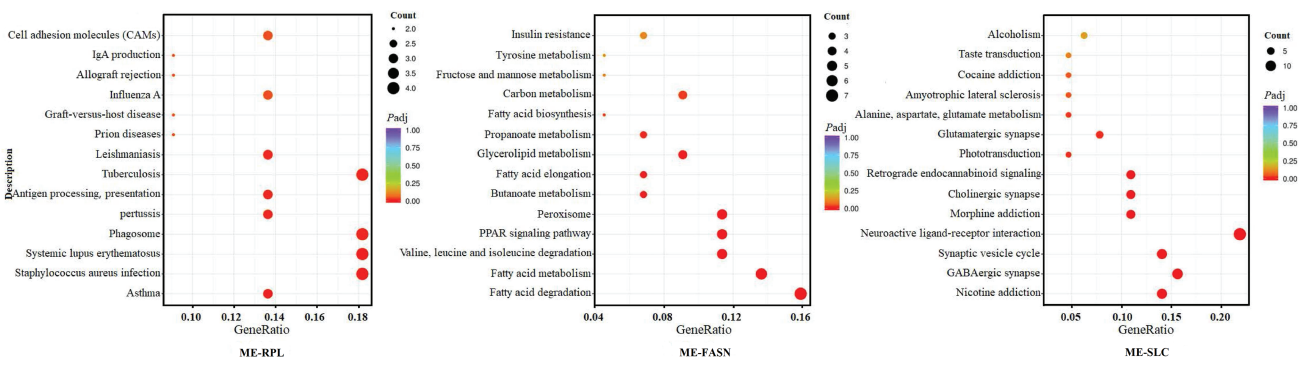


**FIGURE 10 |** KEGG enrichment of the significant differentially expressed mRNAs in "ME-RPL," "ME-FASN," and "ME-SLC." The abscissa shows the rich factors and the ordinate represents the pathways. Size of each point indicates the number of candidate target genes in the pathway, and the color of each point corresponds to the value of *p* after correction.

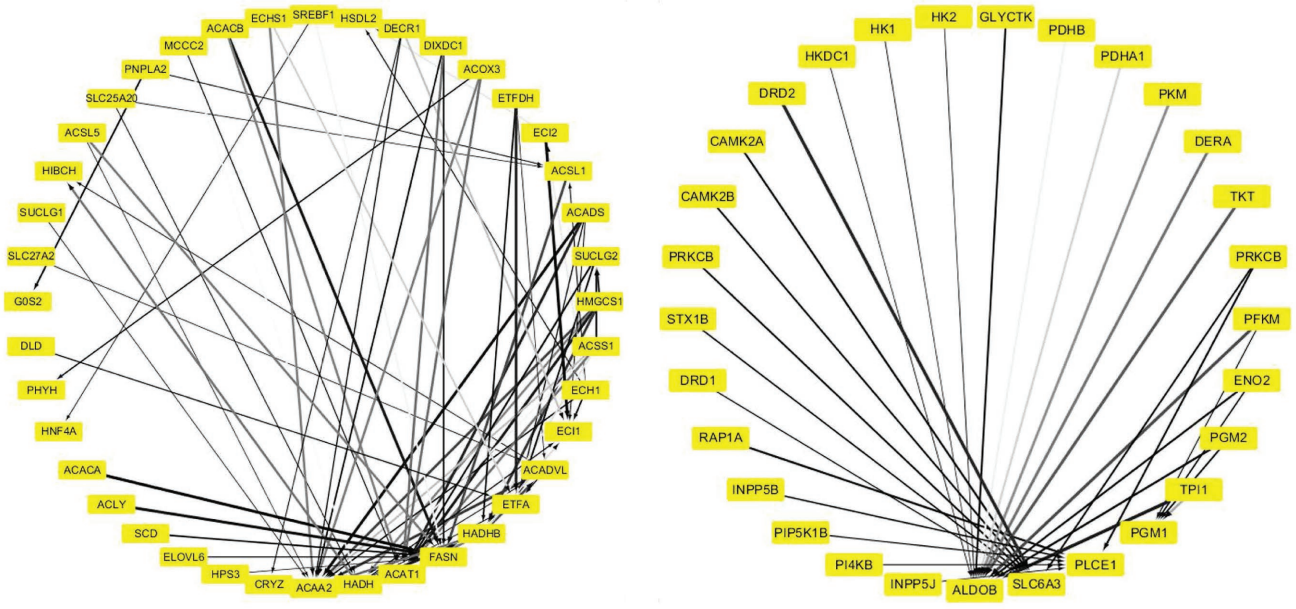


**FIGURE 11 |** PPI network of the DEGs in "ME-FASN." Grayscale value of each line correlates with the strength of the interaction between the proteins.

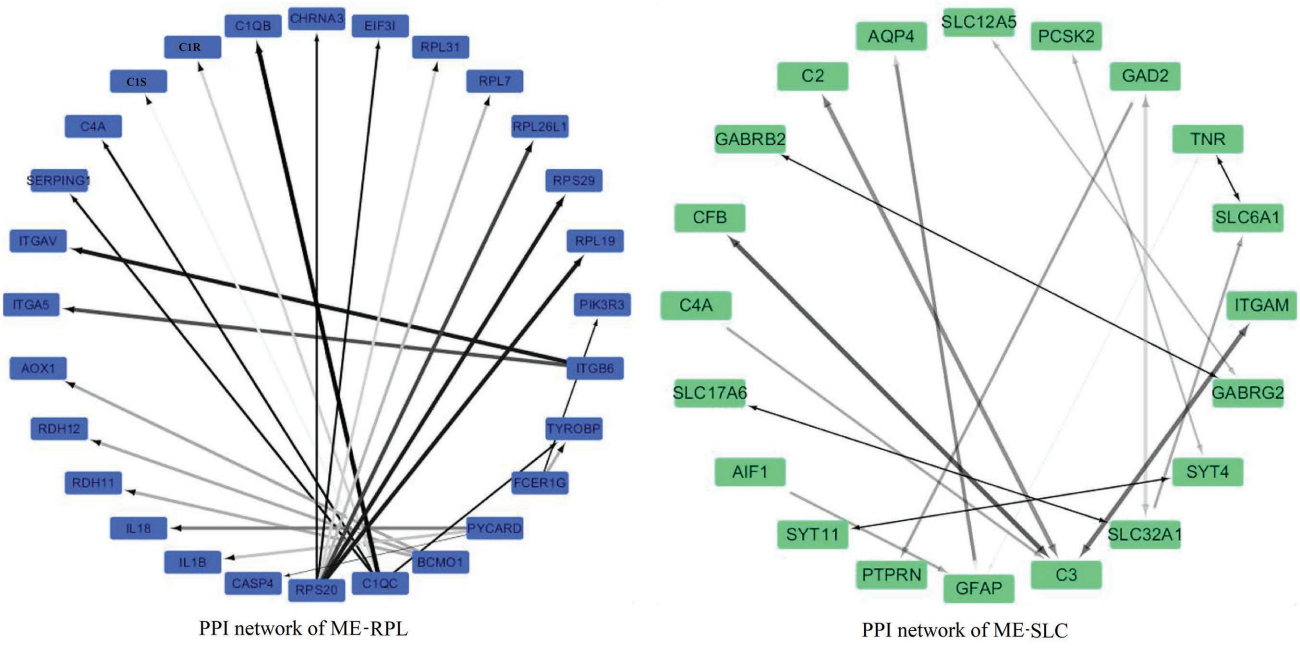


**FIGURE 12 |** PPI networks of the DEGs in "ME-RPL" and "ME-SLC." Grayscale value of each line correlates with the strength of the interaction between the proteins.

development of CNV. Although with cone cells accounting for 96% of all the photosensitive cells, tree shrew not have discernible macula. The problem of lack established transgenic technology for tree shrews need to be solved, and the role of CAMs and fatty acid metabolic pathway needs further investigations to elucidate CNV pathogenesis at the molecular level.

## DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found below: The European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) ArrayExpress, https://www.ebi.ac.uk/arrayexpress/, E-MTAB-10198.

## ETHICS STATEMENT

The animal study was reviewed and approved by Animal Ethics Committee of the Institute of Medical Biology, Chinese Academy of Medical Science (Ethics approval number: DWSP201803019).

## AUTHOR CONTRIBUTIONS

JJ, DQ, and JD conceived and designed this study. JJ, DQ, and NL performed the experiments. XS, WW, CL, YH, and PT collected the samples. JJ and DQ interpreted the data and drafted and edited the manuscript. The study was performed under the supervision of JD and MW. All authors contributed to the article and approved the submitted version.

## FUNDING

This work was supported by Yunnan Key Project of Science and Technology Plan (2019FA028), Yunnan Key Laboratory of Ophthalmic Research and Disease Control (2017DG008), Yunnan Science and Technology Talent and Platform Program (2017HC019 and 2018HB071), Yunnan Health Training Project of High Level Talents (D-2018026), and Kunming Science and Technology Innovation Team (2019-1-R-24483).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.654955/full#supplementary-material

**Supplementary Figure 1 |** Symptoms of tree shrews with choroidal neovascularization (CNV) after laser photocoagulation. **(A)** Representative fundus angiograph of normal tree shrew. **(B)** Representative fundus angiograph of tree shrew with laser photocoagulation. **(C,D)** FFA/ICGA of normal tree shrew. **(E,F)** FFA/ICGA of tree shrew after 7 days of laser photocoagulation. **(G,H)** FFA/ICGA of tree shrew after 30 days of laser photocoagulation.

## REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Archer, D. B., and Gardiner, T. A. (1981). Morphologic flurescein angiographic, and light microscopic features of experimental choroidal neovascularization. *Am J. Ophthalmol.* 91, 297–311. doi: 10.1016/0002-9394(81)90281-6

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Bruning, U., Morales-Rodriguez, F., Kalucka, J., Goveia, J., Taverna, F., Queiroz, K. C. S., et al. (2018). Impairment of angiogenesis by fatty acid synthase inhibition involves mTOR Malonylation. *Cell Metab.* 28, 866–880.e15. doi: 10.1016/j.cmet.2018.07.019

Choudhary, M., Kazmin, D., Hu, P., Thomas, R. S., McDonnell, and D. P., Malek, G. (2015). Aryl hydrocarbon receptor knock-out exacerbates choroidal neovascularization via multiple pathogenic pathways. *J. Pathol.* 235, 101–102. doi: 10.1002/path.4433

Dong, Z. J., Shi, Y. N., Zhao, H. J., Li, N., Ye, L., Zhang, S., et al. (2018). Sulphonated formononetin induces angiogenesis through vascular endothelial growth factor/cAMP response element-binding protein/early growth response 3/vascular cell adhesion molecule 1 and Wnt/β-catenin signaling pathway. *Pharmacology* 101, 76–85. doi: 10.1159/000480662

Dou, G. R., Li, N., Chang, T. F., Zhang, P., Gao, X., Yan, X. C., et al. (2016). Myeloid-specific blockade of notch signaling attenuates choroidal neovascularization through compromised macrophage infiltration and polarization in mice. *Sci. Rep.* 6:28617. doi: 10.1038/srep28617

Doyle, S. L., López, F. J., Celkova, L., Brennan, K., Mulfaul, K., Ozaki, E., et al. (2015). IL-18 immunotherapy for neovascular AMD: tolerability and efficacy in nonhuman primates. *Invest. Ophthalmol. Vis. Sci.* 56, 5424–5430. doi: 10.1167/iovs.15-17264

ElDirini, A. A., Ogden, T. E., and Ryan, S. J. (1991). Subretinal endophotocoagulation. A new model of subretinal neovascularization in the rabbit. *Retina* 11, 244–249. doi: 10.1097/00006982-199111020-00010

Fu, Z. J., Gong, Y., Liegl, R., Wang, Z. X., Liu, C. H., Meng, S. S., et al. (2017). FGF21 administration suppresses retinal and choroidal neovascularization in mice. *Cell Rep.* 18, 1606–1613. doi: 10.1016/j.celrep.2017.01.014

Gao, X. Y., and He, S. Z. (2015). Dynamic expression of intercellular adhesion molecule-1 in laser-induced choroidal neovascularization in brown Norway rats. *Chin. J. Exp. Ophthalmol.* 33, 1103–1107. doi: 10.3760/ema.j.issn.2095-0160.2015.12.011

Grossniklaus, H. E., Ling, J. X., Wallace, T. M., Dithmar, S., Lawson, D. H., Cohen, C., et al. (2002). Macrophage and retinal pigment epithelium expression of angiogenic cytokines in choroidal neovascularization. *Mol. Vis.* 8, 119–126.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Haibo, W., Xiaokun, H., Erika, S., and Hartnett, M. E. (2016). TNF-α mediates choroidal neovascularization by upregulating VEGF expression in RPE through ROS-dependent β-catenin activation. *Mol. Vis.* 22, 116–128.

Hoerster, R., Muether, P. S., Vierkotten, S., Schröder, S., Kirchhof, B., and Fauser, S. (2012). In-vivo and ex-vivo characterization of laser-induced choroidal neovascularization variability in mice. *Graefes Arch. Clin. Exp. Ophthalmol.* 250, 1579–1586. doi: 10.1007/s00417-012-1990-z

Holliday, E. G., Smith, A. V., Cornes, B. K., Buitendijk, G. H. S., Jensen, R. A., Sim, X. L., et al. (2013). Insights into the genetic architecture of early stage age-related macular degeneration: a genomewide association study meta-analysis. *PLoS One* 8:e53830. doi: 10.1371/journal.pone.0053830

Hollyfield, J. G., Bonilha, V. L., Rayborn, M. E., Yang, X. P., Shadrach, K. G., Lu, L., et al. (2008). Oxidative damage-induced inflammation initiates age-related macular degeneration. *Nat. Med.* 14, 194–198. doi: 10.1038/nm1709

Jia, J., and Dai, J. J. (2019). Advantages and challenges of tree shrews in biomedical research. *Lab. Anim. Comp. Med.* 39, 3–8. doi: 10.26914/c.cnkihy.2019.068189

Jie, Z., Yusheng, W., and Yannian, H. (2004). Formation of choroidal neovascularization and its inhibition. *Rec. Adv. Ophthalmol.* 24, 57–60. doi: 10.13389/j.cnki.rao.2004.01.028

Jin, L., Yuhua, H., and Xin, Z. (2010). Relation between inflammation and choroidal neovascularization. *Rec. Adv. Ophthalmol.* 30, 293–296. doi: 10.13389/j.cnki.rao.2010.03.008

Joseph, B. L., Abdoulaye, S., Luke, A. W., Santeford, A., Nudleman, E., Nakamura, R., et al. (2018). WNT7A/B promote choroidal neovascularization. *Exp. Eye Res.* 174, 107–112. doi: 10.1016/j.exer.2018.05.033

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Kunbei, L., Chenjin, J., Shu, T., Yunfan, X., Rui, H., and Jian, G. (2014). The study of laser-induced choroidal neovascularization in rhesus monkeys. *Chin. J. Ophthalmol.* 50, 203–208. doi: 10.3760/cma.j.issn.0412-4081.2014.03.010

Lambert, N. G., ElShelmani, H., Singh, M. K., Mansergh, F. C., Wride, M. A., Padilla, M., et al. (2016). Risk factors and biomarkers of age-related macular degeneration. *Prog. Retin. Eye Res.* 54, 64–102. doi: 10.1016/j.preteyeres.2016.04.003

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Lin, X., Yun, Z., Bin, L., Lü, L. B., Chen, C. S., Chen, Y. B., et al. (2013). Tree shrews under the spot light: emerging model of human diseases. *Zool. Res.* 34, 59–69. doi: 10.3724/SP.J.1141.2013.02059

Liu, T., Hui, L., Wang, Y. S., Guo, J. Q., Li, R., Su, J. B., et al. (2013). In-vivo investigation of laser-induced choroidal neovascularization in rat using spectral-domain optical coherence tomography (SD-OCT). *Graefes Arch. Clin. Exp. Ophtalmol.* 251, 1293–1301. doi: 10.1007/s00417-012-2185-3

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, 1–21. doi: 10.1186/s13059-014-0550-8

Ma, W. X., Silverman, S. M., Lian, Z., Villasmil, R., Campos, M., Amaral, J., et al. (2019). Absence of TGFb signaling in retinal microglia induces retinal degeneration and exacerbates choroidal neovascularization. *Elife Sci.* 8:e42049. doi: 10.7554/eLife.42049

Ming, A., Fang, Y., and Dai, L. (2011). Recent advance on pathogenesis of choroidal neovascularization. *J. Clin. Ophthalmol.* 19, 351–354. doi: 10.3969/j.issn.1006-8422.2011.04.038

Nagineni, C. N., Kommineni, V. K., William, A., Detrick, B., and Hooks, J. J. (2012). Regulation of VEGF expression in human retinal cells by cytokines: implications for the role of inflammation in age-related macular degeneration. *J. Cell. Physiol.* 227, 116–126. doi: 10.1002/jcp.22708

Poor, S. H., Qiu, Y., Fassbender, E. S., Shen, S., Woolfenden, A., Delpero, A., et al. (2014). Reliability of the mouse model of choroidal neovascularization induced by laser photocoagulation. *Invest. Ophthalmol. Vis. Sci.* 55, 6525–6534. doi: 10.1167/iovs.14-15067

Popovic, N., Hooker, E., Barabino, A., Flamier, A., Provost, F., Buscarlet, M., et al. (2020). COCO/DAND5 inhibits developmental and pathological ocular angiogenesis. *EMBO Mol. Med.* 13:e12005. doi: 10.15252/emmm.202012005

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Seguin, F., Carvalho, M. A., Bastos, D. C., Agostini, M., Zecchin, K. G., Alvarez-Flores, M. P., et al. (2012). The fatty acid synthase inhibitor orlistat reduces experimental metastases and angiogenesis in B16-F10 melanomas. *Br. J. Cancer* 107, 977–987. doi: 10.1038/bjc.2012.355

Smolock, E. M., Korshunov, V. A., Glazko, G., Qiu, X., Gerloff, J., and Berk, B. C. (2012). Ribosomal protein L17, RpL17, is an inhibitor of vascular smooth muscle growth and carotid intima formation. *Circulation* 126, 2418–2427. doi: 10.1161/CIRCULATIONAHA.112.125971

Suzuki, Y., Ito, Y., Mizuno, M., Kinashi, H., Sawai, A., Noda, Y., et al. (2012). Transforming growth factor-beta induces vascular endothelial growth factor-C expression leading to lymph angiogenesis in rat unilateral ureteral obstruction. *Kidney Int.* 81, 865–879. doi: 10.1038/ki.2011.464

Tamai, K., Spaide, R. F., Ellis, E. A., Iwabuchi, S., Ogura, Y., and Armstrong, D. (2002). Lipid hydroperoxi destimulates subretinal choroidal neovascularization in the rabbit. *Exp. Eye Res.* 74, 301–308. doi: 10.1006/exer.2001.1121

Tomita, Y., Ozawa, N., Miwa, Y., Ishida, A., Ohta, M., Tsubota, K., et al. (2019). Pemafibrate prevents retinal pathological neovascularization by increasing FGF21 level in a murine oxygen-induced retinopathy model. *Int. J. Mol. Sci.* 23:5878. doi: 10.3390/ijms20235878

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Walshe, T. E., Saint-Geniez, M., Maharaj, A. S., Sekiyama, E., Maldonado, A. E., D'Amore, P. A., et al. (2009). TGF-beta is required for vascular barrier function, endothelial survival and homeostasis of the adult microvasculature. *PLoS One* 4:e5149. doi: 10.1371/journal.pone.0005149

Wang, X., Ma, W., Han, S., Meng, Z. Y., Zhao, L., Yin, Y., et al. (2017). TGF-β participates choroid neovascularization through Smad2/3-VEGF/TNF-α signaling in mice with laser-induced wet age-related macular degeneration. *Sci. Rep.* 7:9672. doi: 10.1038/s41598-017-10124-4

Weidner, N., Carroll, P. R., Flax, J., Blumenfeld, W., and Folkman, J. (1993). Tumor angiogenesis correlates with metastasis in invasive prostate carcinoma. *Am. J. Pathol.* 143, 401–409.

Xiao, Y. L., and Wu, Y. Y. (2016). Interleukin-18 inhibits experimental choroidal neovascularization and its potential therapeutic applications. *Chin. J. Ocul. Fundus Dis.* 32, 457–459. doi: 10.3760/cma.j.issn.1005-1015.2016.04.030

Yang, C., Chen, W., Chen, Y., and Jiang, J. (2012). Smoothened transduces hedgehog signal by forming a complex with Evc/Evc2. *Cell Res.* 22, 1593–1604. doi: 10.1038/cr.2012.134

Yang, X. M., Wang, Y. S., Xu, J. F. X., and Zhang, P. (2006). Characteristics of choroidal neovascularization induced by laser in pigmented rats. *Rec. Adv. Ophthalmol.* 26, 161–166. doi: 10.13389/j.cnki.rao.2006.03.001

Zarranz-Ventura, J., Fernández-Robredo, P., Recalde, S., Salinas-Alamán, A., Borrás-Cuesta, F., Dotor, J., et al. (2013). Transforming growth factor-beta inhibition reduces progression of early choroidal neovascularization lesions in rats: P17 and P144 peptides. *PLoS One* 8:e65434. doi: 10.1371/journal.pone.0065434

Zhou, T., Hu, Y., Chen, Y., Zhou, K. K., Zhang, B., Gao, G. Q., et al. (2010). The pathogenic role of the canonical Wnt pathway in age - related macular degeneration. *Invest. Opthalmol. Vis. Sci.* 51, 4371–4379. doi: 10.1167/iovs.09-4278

Zhu, Z. R., Goodnight, R., Sorgente, N., Ogden, T. E., and Ryan, S. J. (1989). Experimental subretinal neovascularization in the rabbit. *Graefes Arch. Clin. Exp. Ophthalmol.* 227, 257–262. doi: 10.1007/BF02172759

**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 June 2021
doi: 10.3389/fgene.2021.686298

# Correlations Between the Characteristics of Alternative Splicing Events, Prognosis, and the Immune Microenvironment in Breast Cancer

*Youyuan Deng[1], Hongjun Zhao[1], Lifen Ye[1], Zhiya Hu[2], Kun Fang[3] and Jianguo Wang[1]**

*[1] Department of General Surgery, Xiangtan Central Hospital, Xiangtan, China, [2] Department of Pharmacy, Third Hospital of Changsha, Changsha, China, [3] Department of Surgery, Yinchuan Maternal and Child Health Hospital, Yinchuan, China*

OPEN ACCESS

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

***Reviewed by:***
*Ashish Misra,*
*Indian Institute of Technology Hyderabad, India*
*Tuba Denkçeken,*
*Sanko University, Turkey*

***Correspondence:**
*Jianguo Wang*
*wangjianguoxtszxyy@163.com*

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

***Received:*** *26 March 2021*
***Accepted:*** *17 May 2021*
***Published:*** *14 June 2021*

***Citation:***
*Deng Y, Zhao H, Ye L, Hu Z, Fang K and Wang J (2021) Correlations Between the Characteristics of Alternative Splicing Events, Prognosis, and the Immune Microenvironment in Breast Cancer. Front. Genet. 12:686298. doi: 10.3389/fgene.2021.686298*

**Objective:** Alternative splicing (AS) is the mechanism by which a few genes encode numerous proteins, and it redefines the concept of gene expression regulation. Recent studies showed that dysregulation of AS was an important cause of tumorigenesis and microenvironment formation. Therefore, we performed a systematic analysis to examine the role of AS in breast cancer (Breast Cancer, BrCa) progression.

**Methods:** The present study included 993 BrCa patients from The Cancer Genome Atlas (TCGA) database in the genome-wide analysis of AS events. We used differential and prognostic analyses and found differentially expressed alternative splicing (DEAS) events and independent prognostic factors related to patients' overall survival (OS) and disease-free survival (DFS). We divided the patients into two groups based on these AS events and analyzed their clinical features, molecular subtyping and immune characteristics. We also constructed a splicing factor (SF) regulation network for key AS events and verified the existence of AS events in tissue samples using real-time quantitative PCR.

**Results:** A total of 678 AS events were identified as differentially expressed, of which 13 and 10 AS events were independent prognostic factors of patients' OS and DFS, respectively. Unsupervised clustering analysis based on these prognostic factors indicated that the Cluster 1 group had a better prognosis and more immune cell infiltration. SFs were significantly related to the expression of AS events, and AA-RPS21 was significantly upregulated in tumors.

**Conclusion:** Alternative splicing expands the mechanism of breast cancer progression from a new perspective. Notably, alternative splicing may affect the patient's prognosis by affecting the infiltration of immune cells. Our research provides important guidance for subsequent studies of AS in breast cancer.

**Keywords: breast cancer, alternative splicing, immune cell infiltration, the cancer genome atlas, splicing factor**

## INTRODUCTION

Breast cancer (BrCa) is the most common malignant tumor in women. There were approximately 2.1 million newly diagnosed BrCa cases worldwide in 2018, which accounts for approximately 25% of the total number of female malignancies and poses a serious threat to women's health and a heavy burden to public health (Schneider et al., 2014; Bray et al., 2018). Early diagnosis significantly improves the survival rate of patients. The 5-year survival rate of patients diagnosed with BrCa before metastasis is 99% (Siegel et al., 2018), and the 5-year survival rate of patients whose tumor has spread to distant organs is only 26% (Koual et al., 2019). Due to the heterogeneity and complexity of BrCa, traditional inspection methods, such as immunohistochemical testing, do not identify effective biomarkers to screen and evaluate BrCa (Network, 2012). Researchers examined the mechanisms of the occurrence and development of BrCa from various aspects, such as gene expression disorders (Cruz et al., 2018), copy number variation (Long et al., 2018; Yang et al., 2019) and DNA methylation (Kresovich et al., 2019; Yari and Rahimi, 2019). Although these studies achieved promising results, they were primarily limited to the transcriptional level, and the post-transcriptional level, such as alternative splicing (AS), was neglected.

AS is an important way to generate greater transcriptome and proteomic diversity in a limited genome (Cieply and Carstens, 2015). Pre-mRNAs may be spliced into mature mRNAs by retaining specific intron regions or excluding specific exons in multi-exon genes (Kelemen et al., 2013), which generates structural and functional protein variants, and further promotes protein diversity and phenotypic complexity (Leoni et al., 2011). The genetic similarity between human and chimpanzee DNA is 99%, but the homology of mature mRNA is less than 60% (Barbosa-Morais et al., 2012). Genome-wide studies show that 92–95% of human exons have undergone alternative splicing (Feng et al., 2013), and the expression of most cellular transcripts have spatial and temporal differences (Wang et al., 2008). The importance of AS in the development of tumors was recognized recently (Srebrow and Kornblihtt, 2006). AS dysfunction leads to changes in the biological behavior of tumor cells, including cell proliferation (Endo, 2019), cell apoptosis (Pal et al., 2019), tumor angiogenesis (Pentheroudakis et al., 2019) and immune escape (Yao et al., 2016). Increasing evidence shows that the unbalanced expression or mis-expressed isomers of splicing variants is another feature of cancer (Ladomery, 2013). Therefore, cancer-specific splicing variants may be used as diagnostic, prognostic, and therapeutic targets. AS events were first reported as a prognostic biomarker for non-small cell lung cancer in 2017 (Li et al., 2017), and subsequent studies were performed in thyroid cancer (Lin et al., 2019), colorectal cancer (Xiong et al., 2018), pancreatic cancer (Yu et al., 2019) and other tumors. However, there are no related reports on the possibility of differentially expressed alternative splicing (DEAS) events as a biomarker for predicting the prognosis of BrCa.

The splicing of pre-mRNA is regulated by cis-acting elements and trans-acting factors. According to different positions and different effects on splicing sites, cis-acting elements are divided into exon splicing enhancers, exon splicing silencers, intron splicing enhancers and intron splicing silencers, which determine their affinity with homologous splicing factors (SFs). Trans-acting factors, including Ser/Arg-rich members and heterologous ribonucleoprotein family members, work by combining with exon splicing enhancers and silencers to further activate or inhibit specific splicing sites (Kornblihtt et al., 2013). According to their sequence, SFs affect the splicing site selection of the splicing regulatory complex (spliceosome) by binding to pre-mRNAs on exon splicing enhancers or silencers (Ule et al., 2006). Recent studies showed that abnormal AS events are caused by a maladjustment of SFs (Anczuków and Krainer, 2016). Therefore, the identification of splicing factors that are responsible for AS in breast cancer must be further studied.

The present study used the RNA sequence data from The Cancer Genome Atlas (TCGA) to perform a genome-wide systemic analysis of the AS events and SFs of BrCa. We combined the DEAS with clinical data to screen for biomarkers associated with survival and recurrence. Our results provide new insights and potential mechanisms for predicting the prognosis and evaluating clinical outcomes for BrCa patients.

## MATERIALS AND METHODS

### Data Acquisition and Processing

We downloaded and integrated the RNA sequence data, gene expression data, methylation data and corresponding clinical information of BrCa patients from the TCGA data portal website (access date December 20, 2020[1]; Colaprico et al., 2016). To quantify the AS events for each BrCa patient, we used a Java application called SpliceSeq to explicitly quantify the splicing pattern and percent-spliced-in (PSI) for each AS event (Ryan et al., 2012). To generate a set of AS event data as reliably as possible, a strict filter condition of samples with PSI values $\geq 0.75$, average PSI value $\geq 0.05$ was used. The following inclusion criteria were used: (1) female; (2) patients diagnosed with BrCa using pathology; (3) the patient did not receive neoadjuvant therapy; (4) complete clinical characteristics, including age, histological classification, pathological stage, Stage T, Stage N, and stage M; (5) the patient survived at least 30 days after the surgery; and (6) corresponding mRNA splicing variant data were full-scale. A total of 993 patients were enrolled in our study cohort. For maximizing the value of the data and minimizing the possible bias caused by the deletion of the missing values, the IterativeImputer package in Python was used to perform multiple interpolations of the missing values (White et al., 2011). The UpSet diagram, created with UpSetR (version: 1.3.3), shows the interaction sets in 7 types of AS. Circos graphs were generated from Circlize (version 0.4.5) to visualize AS events and details of genes in chromosomes. The Pam50 subtype data of BrCa were obtained from Berger et al. (2018). We quantified the infiltration level of ssGSEA immune cell types in the R language Gsva package (Hänzelmann et al., 2013).

[1]https://portal.gdc.cancer.gov/

To accurately describe AS events, a unique annotation was assigned to each AS event by combining the splicing type, the ID number in the SpliceSeq database, and the corresponding gene symbol. For example, in the annotation "ME_HLCS_ID_96019," the mutually exclusive exon (ME) represents the splicing type, ID_96019 represents the specific ID of the splicing variant, and HLCS is the corresponding genetic symbol.

## Identification and Functional Enrichment Analysis of DEAS Events

To identify DEAS between BrCa tissues and adjacent normal tissues, PSI values for each AS event were measured from the TCGA BrCa cohort (993 BrCa tissues and 113 adjacent normal tissues). Expression differences were characterized as difference multiples (log2FoldChange, log2FC) and adj.P value (Bonferroni corrected *P*-value). $|\log2FC| > 1$ and adj.P $< 0.05$ indicate that the corresponding AS events were up-regulated or down-regulated, respectively. The heatmap and the volcano map were used to show the AS events expressed differently. The DEAS event parent genes were subjected to biological functional enrichment analysis to examine the potential mechanism of BrCa. The "Enrichplot" software package (version 1.6.0) of the R software (version 3.6.1) was used to perform the gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses. A False Discovery Rate (FDR, *P*-value corrected by Benjamini-Hochberg method) less than 0.05 was considered significant.

## Building a Splicing Related Network

Seventy-one splicing factors (SFs) were verified in experiments, and these SFs belonged to two main families, Ser/ARG-rich (SR) protein and heteroribonucleoprotein (hnRNPs). The correlation between SF expression and the PSI value of DEAS events was analyzed ($|R| > 0.5$, $P < 0.05$), and the correlation graph was generated using Cytoscape (version 3.7.1).

## Investigating the Prognostic Value of DEAS Events

Based on the DEAS event, univariate Cox regression analysis and LASSO (Least Absolute selection Operator) regression analysis were performed on overall survival (OS) and disease-free survival (DFS) to determine independent prognostic factors in BrCa. The relationship between clinicopathological data and prognosis was further examined.

## Evaluation of the Correlation With Clinical, Molecular and Immune Characteristics

Based on the total independent risk factors, the "Consent ClusterPlus" package was used to classify the TCGA BrCa cohort in an unbiased and unsupervised manner (Wilkerson and Hayes, 2010; de Lena et al., 2017).

The following clustering settings were used in our study: maxK = 9; Clustering algorithm = PAM; and Correlation method = Euclidean. The relative change in area under the cumulative distribution function curve was used to determine the optimal clustering number K. Differences in clinicopathological information (T, N, M, and TNM stages), survival status (OS and DFS), immune cell content and molecular typing were analyzed between groups, and a violin diagram was used to visualize the different immune cell contents between different subgroups. We performed a survival analysis on the above-listed groups to verify the impact of these events on patient prognosis.

## Real-Time Fluorescence Quantitative PCR to Validate DEAS Events

The Medical Ethics Committee of Xiangtan Central Hospital approved the use of patient samples, and the study complied with the provisions of the Declaration of Helsinki (amended in 2013). Twenty cases of frozen BrCa tissue and paired adjacent tissue were collected from patients who received treatment at the General Surgery Department of Xiangtan Central Hospital from June 1, 2019 to December 1, 2019. Real-time fluorescence quantitative PCR was performed to verify DEAS events. Total RNA was isolated from frozen tissue using TRIzol (Invitrogen, ThermoFisher, CA, United States), and it was reverse transcribed into cDNA using the PrimeScript$^{TM}$ RT kit (TaKaRa, Otsu, Japan). SYBR Premix Ex-Taq$^{TM}$ (TaKaRa, Otsu, Japan) was performed in an FTB-3000P PCR system (Funglyn Biotech, Shanghai, China). The $2^{-\Delta\Delta CT}$ method was used to calculate the expression quantity of relative splicing variants. The corresponding PCR primers are listed in **supplementary Table 1**.

# RESULTS

## Overview of BrCa AS Events

The flow diagram of the research is presented in **Figure 1**. A total of 993 patients who met the screening criteria were included in the study cohort, and the baseline characteristics of these patients are summarized in **Supplementary Table 2**. At a median follow-up of 30.2 months (range 1–286.8), 107 (10.8%) patients relapsed, and 136 (13.7%) patients died. The 5-year mortality and recurrence rates were 8.7 and 9.2%, respectively, with recurrences within 5 years accounting for 66.9% of the total deaths. We detected 35,520 AS events from 10,279 genes using rigorous filtering criteria. A total of 1,455,675 missing values were detected in the data integrity check, which accounted for approximately 4.1% of the total data volume. To maximize the statistical power of the data, we used multiple interpolation to estimate missing values. These AS events were classified into seven modes of splicing, including 3′ Alternate Acceptor site (AA), 5′ Alternate Donor site (AD), Alternate Promoter (AP), Alternate Terminator (AT), Exon Skip (ES), Mutually Exclusive Exons (ME), and Retained Intron (RI), as shown in **Figure 2A**. Among these splicing patterns, ES occurred most frequently (42.3%) (**Figure 2B**). A single gene may have multiple splicing patterns. Therefore, an upset map was created to analyze the interaction sets of seven types of AS events. **Figure 2D** shows that a single gene may have up to four different splicing patterns. A Circos diagram showed the position of AS events on the

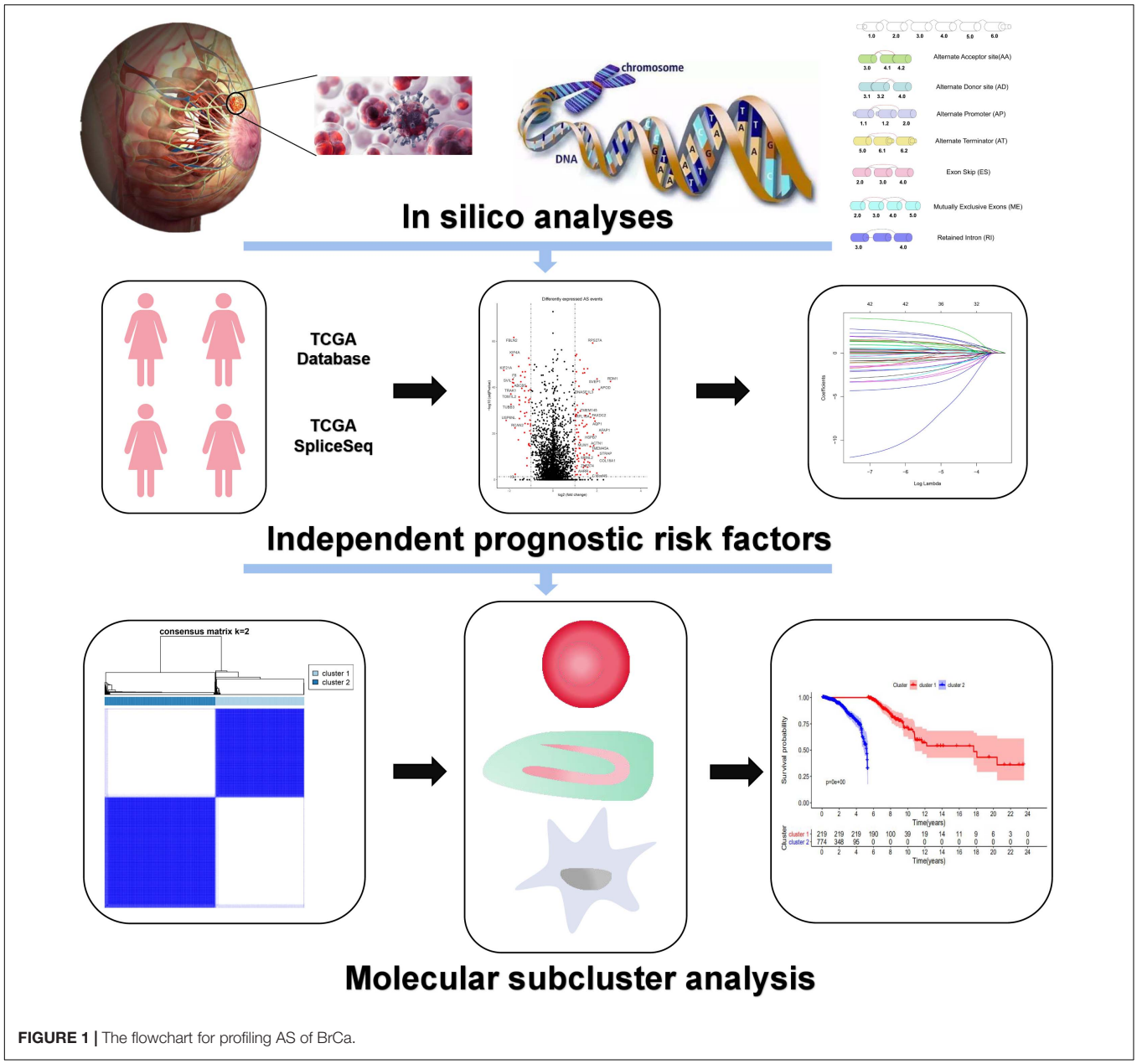


**FIGURE 1 |** The flowchart for profiling AS of BrCa.

chromosome and the possible interactions between their parent genes (**Figure 2E**).

## Identification and Enrichment Analysis of DEAS Events

A total of 678 DEAS events of 564 genes were significantly different between 993 BrCa tissues and 113 adjacent normal tissues. Tumor and normal samples were clearly divided into two discrete groups using unsupervised hierarchical clustering based on DEAS events, which indicated that the screened DEAS events were credible (**Figure 3A**). The volcanogram showed high- and low-expressed DEAS events (**Figure 3B**). Notably, ES events had the highest frequency of all AS events, but the situation changed in DEAS events, and AP events accounted for the highest proportion (**Figure 2C**). The pattern of splicing was not evenly distributed, which suggests a different role in tumor progression. Abnormal AS events may directly affect the expression of its parent RNA.

The results showed that the parent genes of the DEAS events were enriched in pathways that were closely related to the BrCa process in GO analysis (**Figure 3C**). The pathways primarily included protein aggregation (FDR = 0.004), cell cycle G2/M phase transition (FDR = 0.028), G2/M transition mitotic cell cycle (FDR = 0.029), and integrin-mediated signaling pathways (FDR = 0.031). The KEGG pathways associated with BrCa tumorigenesis (**Figure 3D**) were enriched, such as the cancer pathway (FDR = 0.001), MAPK signaling pathway

(FDR = 0.001), ECM-receptor interaction (FDR = 0.004), and PI3K-Akt signaling pathway (FDR = 0.005). Notably, the discovery of immune-related pathways, such as leukocyte migration across the endothelium (FDR = 0.002), aroused our interest, and we further examined the impact of variable splicing on tumor immunity.

## Construction of DEAS Events and SFs Regulatory Network

A splicing regulation network was established, and the expression of 10 SFs significantly correlated with the PSI of 27 DEAS events (including 19 positive correlations and 8 negative correlations). The network is shown in **Figure 4C**. Notably, four different SFs controlled a DEAS event, which reflects the complex cooperative and competitive relationships between SFs and partially explains the diversity of splice homologous patterns caused by only a few factors. For the AA splicing event of RPS21, its PSI value is positively correlated with the splicing factor (RBM5), but the expression of its parent gene is negatively correlated with the RBM5, which suggests that our previous mRNA detection methods ignore the diversity of gene transcripts (**Figures 4A,B**). Therefore, specific detection of gene transcripts can provide a more detailed description of tumor characteristics.

## Use of DEAS Events to Construct a Prognostic Risk Model

Univariate COX analysis showed that 48 and 49 DEAS events significantly correlated with OS and DFS, respectively. LASSO regression was used to filter variables and prevent over-fitting of the proportional hazards model (**Supplementary Figure 1**). These filtered DEAS events were included in the multivariate Cox regression analysis. 13 and 10 DEAS events were identified as independent prognostic factors for OS and DFS, respectively. Twenty-one independent prognostic risk factors were identified after the removal of repeats and were named total independent prognostic factors. The detailed results of these prognostic AS events are shown in the **Supplementary Table 3**.

## Classification, Prognosis, Molecular and Immune Characteristics Analysis Based on AS Events

To understand the impact of independent prognostic AS events on patients, we performed an unsupervised cluster analysis. According to the consensus matrix heat map, patients were divided into two subclusters (**Figure 5A**): Cluster 1 ($n$ = 219, 22.1%) and Cluster 2 ($n$ = 774, 77.9%). We analyzed the clinicopathological characteristics and immune cell infiltration of the two subclusters. Kaplan-Meier survival analysis showed that the prognosis of Cluster 1 was significantly better than Cluster 2, regardless of OS or DFS (**Figures 5C,D**). As shown in the **Figure 5E**, T stage, M stage, TNM stage, age, and survival status (OS and DFS) were not randomly distributed between different clusters ($P < 0.05$). Immune cells in Cluster 1 were significantly higher than Cluster 2 (**Figure 5B**). These immune components included APC coinhibition, APC costimulation and $CD8^+$ T cells, and these results are consistent with the previous conclusion that Cluster 1 has a better prognosis in OS and DFS. Briefly, the subgroup based on the independent prognostic AS events distinguished the immunophenotype and prognosis of BrCa patients, which indicates that this method has good clinical practical value. **Figure 5E** shows the cluster analysis of clinicopathological information and immune components. (* denotes $P < 0.05$ and $\geq 0.01$; ** denotes $P < 0.01$ and $\geq 0.001$; *** denotes $P < 0.001$).

## Real-Time Quantitative PCR Verification of DEAS Events in Tissues

To verify the accuracy of the bioinformatics analysis, we collected pairs of BrCa tissue and adjacent tissue samples for further verification. We reviewed studies of parental genes involved in all splicing events and selected one DEAS events for further validation. Two primers were designed for each gene to maintain consistency in the experimental approach. One primer was located on the splice sequence of the DEAS event, and the other primer was located on the CDS sequence of all transcripts. We used qPCR to obtain the splicing event and CD region expression of each gene, and the ratio of the two expression levels is the percent-spliced-in (PSI) value. A box diagram was generated to illustrate the qPCR results (**Supplementary Figures 2A,B**). The ratio of this DEAS event was significant upregulation in tumor tissue, which suggests that the increase in these DEAS events affects tumorigenesis. Notably, these findings provide important guidance for more detailed functional testing.

# DISCUSSION

Current applications of targeted therapy are primarily focused at the gene level. However, studies at the post-transcriptional level found that the same gene had different functions. For example, Marina found that high expression of a new E-cadherin splice variant was associated with BrCa progression (Rosso et al., 2019). Catherine showed that the use of selective promoters leads to the overexpression of specific subtypes of ELF5, which may be a feature in the occurrence of BrCa (Piggin et al., 2016). Changes in splicing type lead to changes in the expression of certain AS events, which are of great significance in the occurrence and development of tumors (Chang and Lin, 2019).

Previous data on the function of BrCa AS events primarily focused on one or several genes, and there were no studies on the prognostic value of AS. The rapid development of high-throughput sequencing technology in recent years created favorable conditions for comprehensive discussions of the prognostic value of AS in BrCa.

Tumorigenesis of BrCa is a complex regulatory network. Compared to a single intuitive clinical indicator, the integration of multiple biomarkers into an aggregation model improves the ability of a model to predict prognosis. Over the past decade, great efforts were made to integrate genome-wide prognostic biomarkers to guide doctors in the early diagnosis and prognosis of BrCa patients. However, the focus of previous studies is limited to transcriptome level analyses and the mining of prognostic mRNA, lncRNA, or miRNA markers

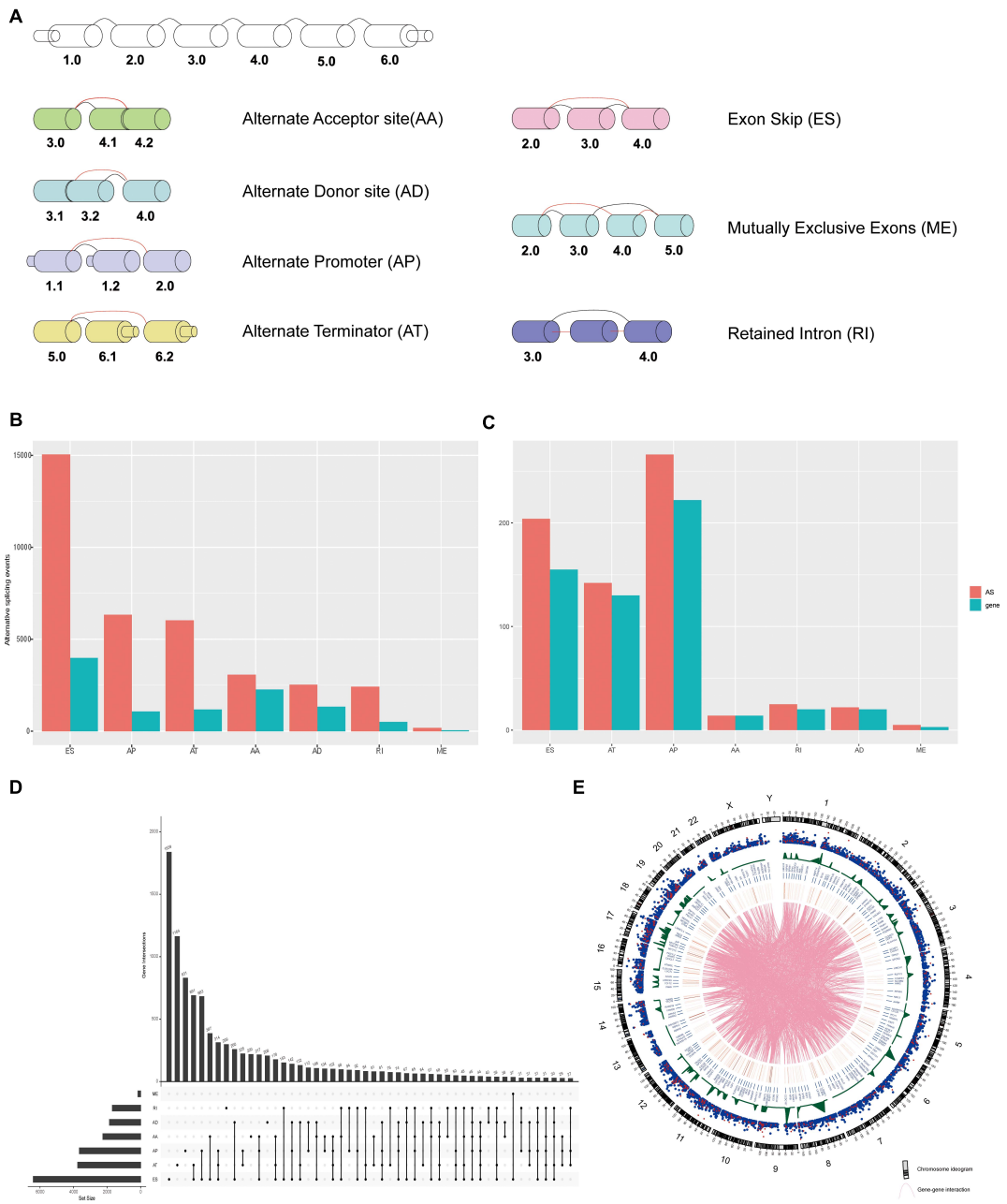


**FIGURE 2 | (A)** The schematic diagram of the seven AS event patterns. **(B)** The number of AS events and their parental genes in BrCa patients. **(C)** The number of DEAS events and their parental genes in BrCa patients. **(D)** The upset diagram shows the intersection of the seven types of splicing events in BrCa. **(E)** Circos of AS events on chromosomes and their parent gene annotations. The outer circle consists of a point map representing the detected AS events, which is linked to the position of the parent gene in the chromosome, and the red dots represent the DEAS events. The blue dots represent 454 the non-differentially expressed AS events. The gene in the middle is called the DEAS event parent. 455 Lines represent potential interactions between parental genes of the DEAS events.

(Meng et al., 2014; Bing et al., 2016). Zhang et al. recently obtained RNA-seq data and corresponding clinical information of BrCa patients from the TCGA. The impact of AS events on the prognosis of patients was analyzed, and a survival prediction model was constructed (Zhang et al., 2019). Because the driving factors of tumorigenesis and development are generally differentially expressed in tumors and normal tissues (Meng et al., 2019), this article is limited because the data of adjacent normal tissues were not included in the study cohort. Therefore, we analyzed the differences in AS events between tumor tissues and adjacent normal tissues in our study and performed prognostic analyses based on DEAS events. Activation of immune-related pathways was found in functional enrichment analyses.

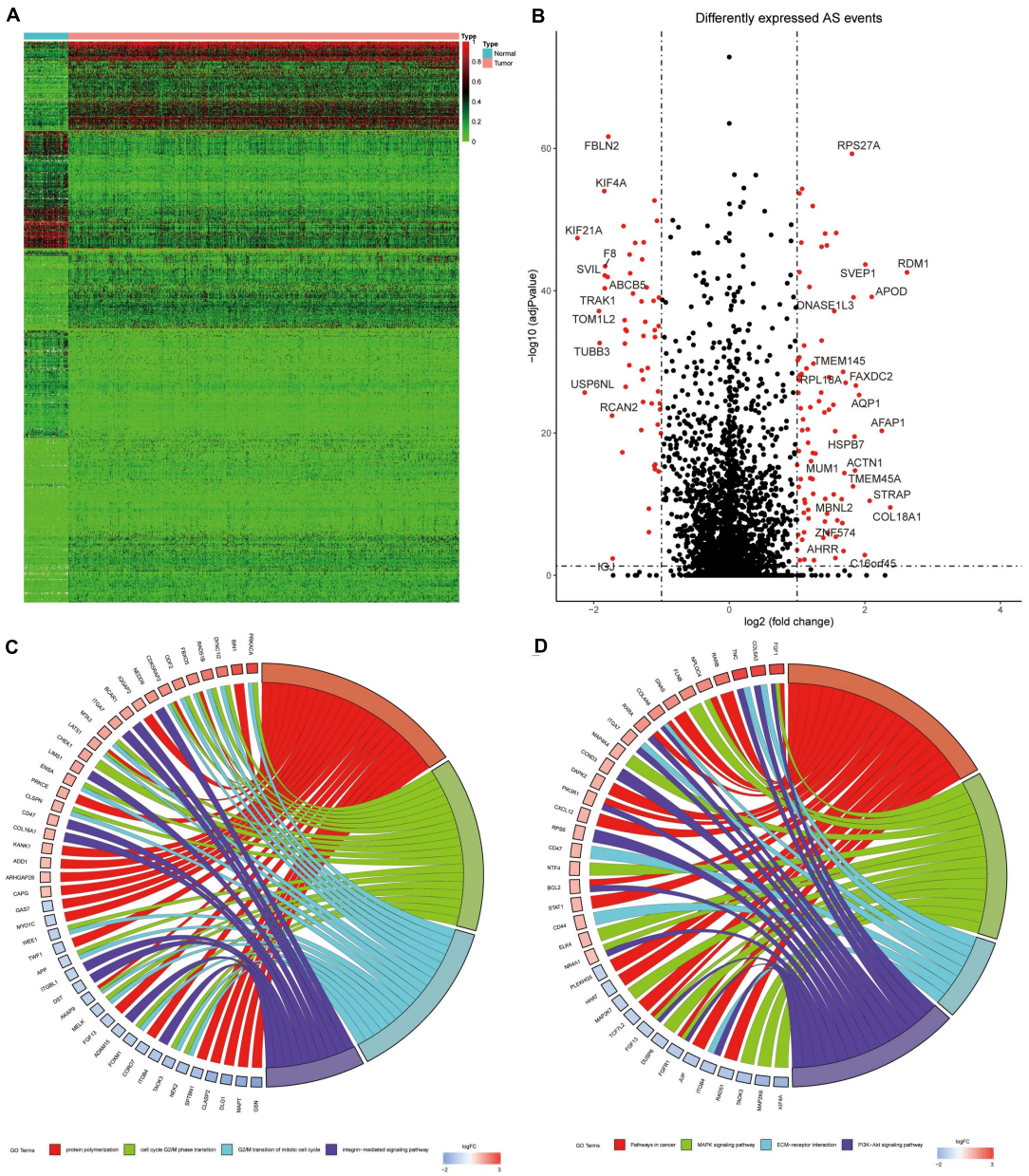


**FIGURE 3 | (A)** Heat map of BrCa tissues and adjacent normal tissues. **(B)** Volcanic diagram of BrCa 458 DEAS events, with red dots representing low-expressed or high-expressed AS events. **(C)** The GO 459 analysis of BrCa DEAS event parent genes. **(D)** the KEGG analysis of BrCa DEAS event parent genes. In **(C,D)** the blue to red code next to the selected gene represents logFC, and the GO term represents the biological behavior related to the parent gene.

Twenty-one AS events were independent prognostic risk factors for OS or DFS. Several genes in these AS events play a vital role in tumor biology. For example, NFIB inhibits the growth of glioma cells, and its low expression may be an important reason for the occurrence of glioma (Li et al., 2019). Therefore, study of the functions and potential mechanisms of these AS events may be of great significance for the development of new therapeutic strategies. Due to the high degree of tumor heterogeneity, the clinical treatment of BrCa is significantly distinct in patients with different molecular subtypes. The present study identified two molecular clusters (Clusters 1, 2) based on total independent prognosis AS events. Stage T, stage M, stage TNM, age, and survival status (OS and DFS) were unevenly distributed between the molecular clusters. Notably, there were also significant differences in the infiltration of several important immune cells between the two groups, such as $CD8^+$ T cells, NK cells and Treg cells. The content of $CD8^+$ T and NK cells in Cluster 1 was higher than Cluster 2, which was also related to the good prognosis of OS and DFS. This conclusion is consistent with Peng GL, who suggested that $CD8^+$ T cells were associated

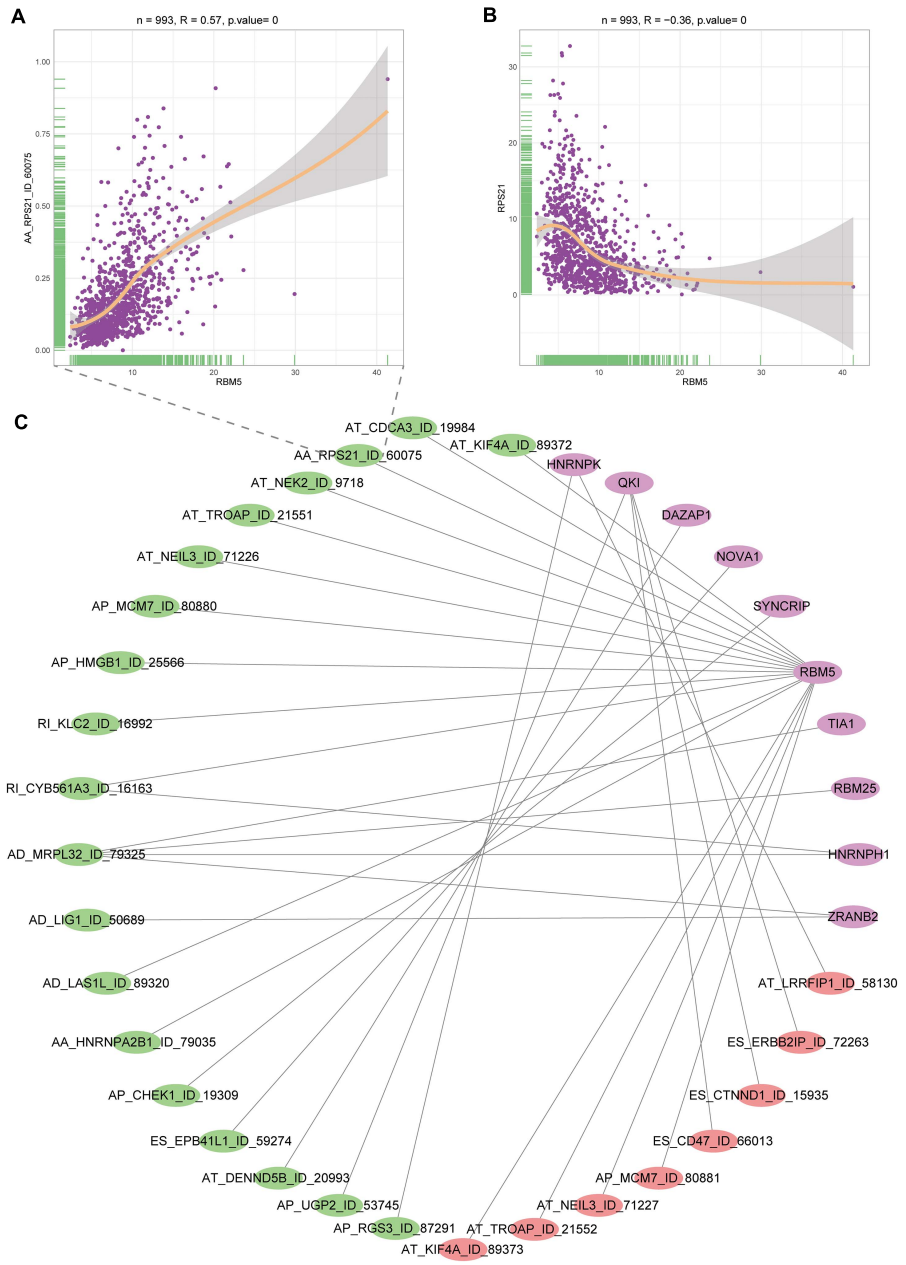


**FIGURE 4 |** Interaction diagram between SFs and DEAS events. **(A)** Correlation diagram of the SF named RBMS and alternate splicing event of RPS21. **(B)** Correlation diagram of the SF named RBMS and parent gene expression of RPS21. **(C)** Network between SFs and AS events. Purple represents SFs, green represents positive correlation, and red represents negative correlation.

with a favorable prognosis in BrCa patients (Peng et al., 2019). NK cells are valuable cells in immunotherapy. The activation of these cells is driven by the balance between activation and inhibitory signals, and NKs exert anti-tumor functions without preactivation (Terrén et al., 2019). Treg cells participate in immunosuppression via a variety of mechanisms, and their high infiltration is related to poor survival in many of tumors. Our findings provide theoretical guidance for examining the relationships between alternative splicing and immunotherapy.

To examine the mechanisms of the effects of upstream factors on alternative splicing and the effects on immunity, we analyzed the relationships between 71 known SFs and splicing events. Our results showed that the abnormal expression of SF was closely related to the expression of DEAS events. A single SF may regulate multiple AS events, and an AS event may also be regulated by multiple SFs. Ribosomal protein s21 (RPS21) is a ribosome-related protein, which has been shown to play an important role in ribosomal biogenesis and may affect

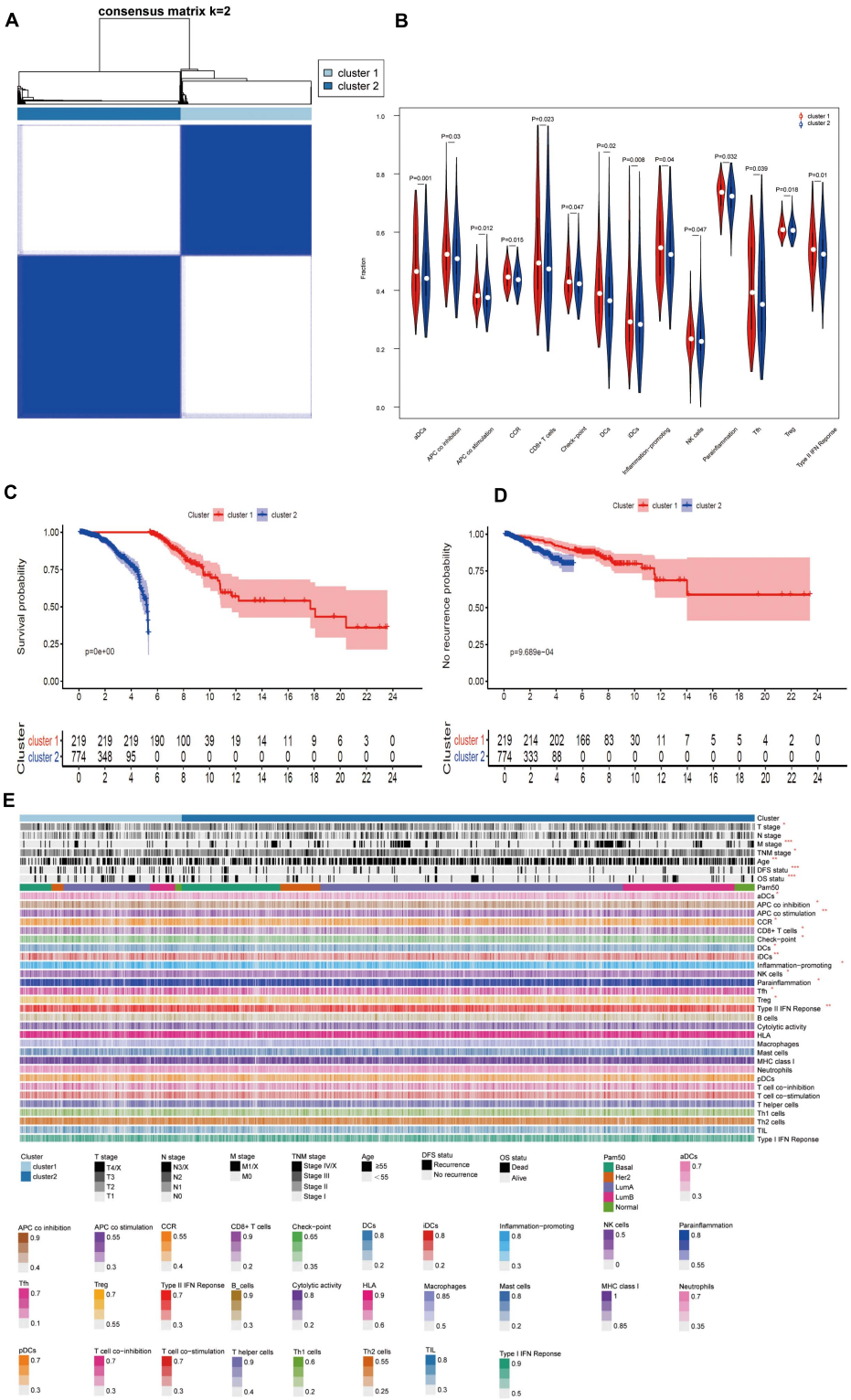


**FIGURE 5 | (A)** Consensus cluster analysis identified two clusters (sample, $n$ = 993). Heat maps of white (consensus = 0, samples never come together) and blue (consensus = 1, samples always come together) show sample consensus. **(B)** Violin diagram of immune-related components between; cluster 1 and 2. **(C,D)** Kaplan-Meier survival analysis for different molecular clusters based on OS or DFS; **(E)** Cluster analysis of clinicopathological information and immune components. (*$P$ < 0.05 and =0.01; **$P$ < 0.01 and =0.001; ***$P$ < 0.001).

the occurrence and development of osteosarcoma and prostate cancer through the RAS/MAPK pathway (Arthurs et al., 2017; Fan et al., 2018; Liang et al., 2019; Sawyer et al., 2020; Wang et al., 2020). Through the query of the database GEPIA2, the expression of RPS21 gene is not significantly different in tumor and normal tissues (**Supplementary Figure 2C**), so its potential function in breast cancer has not been studied so far, but according to our research, one of the transcripts of RPS (AA-RPS21) is differentially expressed in cancer and normal tissues of breast cancer patients, which suggests that this transcript may play a tumor-driving role in breast cancer, so it is worthy of further investigation.

## CONCLUSION

In summary, the present study discovered that 21 independent prognostic AS events played important roles in the occurrence and development of breast cancer. These AS events may affect the prognosis and recurrence of patients via immune-related pathways, and we also found their potential upstream splicing mechanism. Our research provides directions for future research on the three levels of splicing factors, AS events and immune infiltration, and these new findings urgently need follow-up studies to support their clinical value.

## DATA AVAILABILITY STATEMENT

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

## ETHICS STATEMENT

This investigation was approved by the Ethics Committee of the Central Hospital of Xiangtan City. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

JW conceived and directed the project. YD designed the study and analyzed the data. YD, HZ, and LY wrote the manuscript. ZH and KF reviewed the data. All authors have read and approved the final manuscript for publication.

## FUNDING

This work was supported by the Xiangtan Medical Research Project (No. 2020xtyx-3).

## ACKNOWLEDGMENTS

We thank the contributors of TCGA databases for the availability of the data.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.686298/full#supplementary-material

## REFERENCES

Anczuków, O., and Krainer, A. R. (2016). Splicing-factor alterations in cancers. *RNA* 22, 1285–1301. doi: 10.1261/rna.057919.116

Arthurs, C., Murtaza, B. N., Thomson, C., Dickens, K., Henrique, R., Patel, H. R. H., et al. (2017). Expression of ribosomal proteins in normal and cancerous human prostate tissue. *PLoS One* 12:e0186047. doi: 10.1371/journal.pone.0186047

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338, 1587–1593. doi: 10.1126/science.12 30612

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 33, 690.e–705.e.

Bing, Z., Tian, J., Zhang, J., Li, X., Wang, X., and Yang, K. (2016). An integrative model of miRNA and mRNA expression signature for patients of breast invasive carcinoma with radiotherapy prognosis. *Cancer Biother. Radiopharmaceut.* 31, 253–260. doi: 10.1089/cbr.2016.2059

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

Chang, H.-L., and Lin, J.-C. (2019). SRSF1 and RBM4 differentially modulate the oncogenic effect of HIF-1α in lung cancer cells through alternative splicing mechanism. *Biochim. Biophys. Acta* 1866:118550. doi: 10.1016/j.bbamcr.2019.118550

Cieply, B., and Carstens, R. P. (2015). Functional roles of alternative splicing factors in human disease. *Wiley Interdiscipl. Rev. RNA* 6, 311–326. doi: 10.1002/wrna.1276

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–e71.

Cruz, C., Castroviejo-Bermejo, M., Gutiérrez-Enríquez, S., Llop-Guevara, A., Ibrahim, Y., Gris-Oliver, A., et al. (2018). RAD51 foci as a functional biomarker of homologous recombination repair and PARP inhibitor resistance in germline BRCA-mutated breast cancer. *Ann. Oncol.* 29, 1203–1210. doi: 10.1093/annonc/mdy099

de Lena, P. G., Paz-Gallardo, A., Paramio, J. M., and García-Escudero, R. (2017). Clusterization in head and neck squamous carcinomas based on lncRNA expression: molecular and clinical correlates. *Clin. Epigenet.* 9:36.

Endo, T. (2019). Dominant-negative antagonists of the Ras–ERK pathway: DA-Raf and its related proteins generated by alternative splicing of Raf. *Exp. Cell Res.* 387:111775. doi: 10.1016/j.yexcr.2019.111775

Fan, S., Liang, Z., Gao, Z., Pan, Z., Han, S., Liu, X., et al. (2018). Identification of the key genes and pathways in prostate cancer. *Oncol. Lett.* 16, 6663–6669. doi: 10.3892/ol.2018.9491

Feng, H., Qin, Z., and Zhang, X. (2013). Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer Lette.* 340, 179–191. doi: 10.1016/j.canlet.2012.11.010

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

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., et al. (2013). Function of alternative splicing. *Gene* 514, 1–30. doi: 10.1002/9783527678679.dg00350

Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E., and Muñoz, M. J. (2013). Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* 14:153. doi: 10.1038/nrm3525

Koual, M., Cano-Sancho, G., Bats, A.-S., Tomkiewicz, C., Kaddouch-Amar, Y., Douay-Hauser, N., et al. (2019). Associations between persistent organic pollutants and risk of breast cancer metastasis. *Environ. Int.* 132:105028. doi: 10.1016/j.envint.2019.105028

Kresovich, J. K., Xu, Z., O'Brien, K. M., Weinberg, C. R., Sandler, D. P., and Taylor, J. A. (2019). Epigenetic mortality predictors and incidence of breast cancer. *Aging* 11, 11975–11987. doi: 10.18632/aging.102523

Ladomery, M. (2013). Aberrant alternative splicing is another hallmark of cancer. *Int. J. Cell Biol.* 2013:463786.

Leoni, G., Le Pera, L., Ferrè, F., Raimondo, D., and Tramontano, A. (2011). Coding potential of the products of alternative splicing in human. *Genome Biol.* 12:R9.

Li, Y., Sun, N., Lu, Z., Sun, S., Huang, J., Chen, Z., et al. (2017). Prognostic alternative mRNA splicing signature in non-small cell lung cancer. *Cancer Lett.* 393, 40–51. doi: 10.1016/j.canlet.2017.02.016

Li, Y., Xu, J., Zhang, J., Zhang, J., Zhang, J., and Lu, X. (2019). MicroRNA-346 inhibits the growth of glioma by directly targeting NFIB. *Cancer Cell Int.* 19:294.

Liang, Z., Mou, Q., Pan, Z., Zhang, Q., Gao, G., Cao, Y., et al. (2019). Identification of candidate diagnostic and prognostic biomarkers for human prostate cancer: RPL22L1 and RPS21. *Med. Oncol,.* 36:56. doi: 10.1007/s12032-019-1283-z

Lin, P., He, R.-Q., Huang, Z.-G., Zhang, R., Wu, H.-Y., Shi, L., et al. (2019). Role of global aberrant alternative splicing events in papillary thyroid cancer prognosis. *Aging* 11:2082. doi: 10.18632/aging.101902

Long, J., Evans, T. G., Bailey, D., Lewis, M. H., Gower−Thomas, K., and Murray, A. (2018). Uptake of risk−reducing surgery in BRCA gene carriers in Wales, UK. *Breast J.* 24, 580–585. doi: 10.1111/tbj.12978

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.

Meng, X., Zhao, Y., Liu, J., Wang, L., Dong, Z., Zhang, T., et al. (2019). Comprehensive analysis of histone modification-associated genes on differential gene expression and prognosis in gastric cancer. *Exp. Therapeut. Med.* 18, 2219–2230.

Network, C. G. A. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490:61. doi: 10.1038/nature11412

Pal, S., Medatwal, N., Kumar, S., Kar, A., Komalla, V., Yavvari, P. S., et al. (2019). A Localized Chimeric Hydrogel Therapy Combats Tumor Progression through Alteration of Sphingolipid Metabolism. *ACS Central Sci.* 5, 1648–1662. doi: 10.1021/acscentsci.9b00551

Peng, G.-L., Li, L., Guo, Y.-W., Yu, P., Yin, X.-J., Wang, S., et al. (2019). CD8+ cytotoxic and FoxP3+ regulatory T lymphocytes serve as prognostic factors in breast cancer. *Am. J. Translat. Res.* 11:5039.

Pentheroudakis, G., Mavroeidis, L., Papadopoulou, K., Koliou, G.-A., Bamia, C., Chatzopoulos, K., et al. (2019). Angiogenic and Antiangiogenic VEGFA Splice Variants in Colorectal Cancer: Prospective Retrospective Cohort Study in Patients Treated With Irinotecan-Based Chemotherapy and Bevacizumab. *Clin. Colorectal Cancer* 18, e370–e384.

Piggin, C. L., Roden, D. L., Gallego-Ortega, D., Lee, H. J., Oakes, S. R., and Ormandy, C. J. (2016). ELF5 isoform expression is tissue-specific and significantly altered in cancer. *Breast Cancer Res.* 18:4.

Rosso, M., Lapyckyj, L., Besso, M. J., Monge, M., Reventós, J., Canals, F., et al. (2019). Characterization of the molecular changes associated with the overexpression of a novel epithelial cadherin splice variant mRNA in a breast cancer model using proteomics and bioinformatics approaches: identification of changes in cell metabolism and an increased expression of lactate dehydrogenase B. *Cancer Metabol.* 7:5.

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., and Weinstein, J. N. (2012). SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics* 28, 2385–2387. doi: 10.1093/bioinformatics/bts452

Sawyer, J. K., Kabiri, Z., Montague, R. A., Allen, S. R., Stewart, R., Paramore, S. V., et al. (2020). Exploiting codon usage identifies intensity-specific modifiers of Ras/MAPK signaling *in vivo*. *PLoS Genet.* 16:e1009228. doi: 10.1371/journal.pgen.1009228

Schneider, A. P., Zainer, C. M., Kubat, C. K., Mullen, N. K., and Windisch, A. K. (2014). The breast cancer epidemic: 10 facts. *Linacre Quart.* 81, 244–277. doi: 10.1179/2050854914y.0000000027

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

Srebrow, A., and Kornblihtt, A. R. (2006). The connection between splicing and cancer. *J. Cell Sci.* 119, 2635–2641. doi: 10.1242/jcs.03053

Terrén, I., Orrantia, A., Vitallé, J., Zenarruzabeitia, O., and Borrego, F. (2019). NK Cell Metabolism and Tumor Microenvironment. *Front. Immunol.* 10:2278. doi: 10.3389/fimmu.2019.02278

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., et al. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature* 444:580. doi: 10.1038/nature05304

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470. doi: 10.1038/nature07509

Wang, T., Wang, Z. Y., Zeng, L. Y., Gao, Y. Z., Yan, Y. X., and Zhang, Q. (2020). Down-regulation of ribosomal protein RPS21 inhibits invasive behavior of osteosarcoma cells through the inactivation of MAPK pathway. *Cancer Manag. Res,.* 12, 4949–4955. doi: 10.2147/cmar.S246928

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statist. Med.* 30, 377–399. doi: 10.1002/sim.4067

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Xiong, Y., Deng, Y., Wang, K., Zhou, H., Zheng, X., Si, L., et al. (2018). Profiles of alternative splicing in colorectal cancer and their clinical significance: A study based on large-scale sequencing data. *EBioMedicine* 36, 183–195. doi: 10.1016/j.ebiom.2018.09.021

Yang, Y., Li, F., Luo, X., Jia, B., Zhao, X., Liu, B., et al. (2019). Identification of LCN1 as a potential biomarker for breast cancer by bioinformatic analysis. *DNA Cell Biol.* 38, 1088–1099. doi: 10.1089/dna.2019.4843

Yao, J., Caballero, O. L., Huang, Y., Lin, C., Rimoldi, D., Behren, A., et al. (2016). Altered expression and splicing of ESRP1 in malignant melanoma correlates with epithelial–mesenchymal status and tumor-associated immune cytolytic activity. *Cancer Immunol. Res.* 4, 552–561. doi: 10.1158/2326-6066.cir-15-0255

Yari, K., and Rahimi, Z. (2019). Promoter Methylation Status of the Retinoic Acid Receptor-Beta 2 Gene in Breast Cancer Patients: A Case Control Study and Systematic Review. *Breast Care* 14, 117–123. doi: 10.1159/000489874

Yu, M., Hong, W., Ruan, S., Guan, R., Tu, L., Huang, B., et al. (2019). Genome-Wide Profiling of Prognostic Alternative Splicing Pattern in Pancreatic Cancer. *Front. Oncol.* 9:773. doi: 10.3389/fonc.2019.00773

Zhang, D., Duan, Y., and Cun, J. (2019). Identification of prognostic alternative splicing signature in breast carcinoma. *Front. Genet.* 10:278. doi: 10.3389/fgene.2019.00278

**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: 15 July 2021
doi: 10.3389/fgene.2021.680617

# Development of a Four-mRNA Expression-Based Prognostic Signature for Cutaneous Melanoma

*Haiya Bai[1†], Youliang Wang[2†], Huimin Liu[1] and Junyang Lu[1*]*

[1] Department of Female Plastic Surgery, Gansu Provincial Maternity and Child-Care Hospital, Lanzhou, China, [2] Department of Pediatric Surgery, Gansu Provincial Maternity and Child-Care Hospital, Lanzhou, China

We aim to find a biomarker that can effectively predict the prognosis of patients with cutaneous melanoma (CM). The RNA sequencing data of CM was downloaded from The Cancer Genome Atlas (TCGA) database and randomly divided into training group and test group. Survival statistical analysis and machine-learning approaches were performed on the RNA sequencing data of CM to develop a prognostic signature. Using univariable Cox proportional hazards regression, random survival forest algorithm, and receiver operating characteristic (ROC) in the training group, the four-mRNA signature including CD276, UQCRFS1, HAPLN3, and PIP4P1 was screened out. The four-mRNA signature could divide patients into low-risk and high-risk groups with different survival outcomes (log-rank $p < 0.001$). The predictive efficacy of the four-mRNA signature was confirmed in the test group, the whole TCGA group, and the independent GSE65904 (log-rank $p < 0.05$). The independence of the four-mRNA signature in prognostic prediction was demonstrated by multivariate Cox analysis. ROC and timeROC analyses showed that the efficiency of the signature in survival prediction was better than other clinical variables such as melanoma Clark level and tumor stage. This study highlights that the four-mRNA model could be used as a prognostic signature for CM patients with potential clinical application value.

**Keywords: cutaneous melanoma, prognostic signature, random survival forest, MRNA expression data, machine learning**

OPEN ACCESS

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

***Reviewed by:***
*Xinyi Liu,*
*University of Illinois at Chicago, United States*
*Yuhua Yin,*
*Shanghai Jiao Tong University, China*

***Correspondence:**
*Junyang Lu*
*lujy10@lzu.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:*** *15 March 2021*
***Accepted:*** *17 May 2021*
***Published:*** *15 July 2021*

***Citation:***
*Bai H, Wang Y, Liu H and Lu J (2021) Development of a Four-mRNA Expression-Based Prognostic Signature for Cutaneous Melanoma. Front. Genet. 12:680617. doi: 10.3389/fgene.2021.680617*

## INTRODUCTION

Cutaneous melanoma (CM) is a highly aggressive and heterogeneous skin malignancy. In recent years, the morbidity and mortality of CM have increased significantly (Dimitriou et al., 2018), with approximately 232,100 new cases and 55,500 death each year (Schadendorf et al., 2018). Although CM is usually detected in T1 stage and the corresponding patients' 5-year survival exceeds 90%, the rapid progression of melanoma and the failure to detect thin melanoma in time lead to the progression and metastasis, which worsen the prognosis of CM patients [the 5-year survival rate is reduced to 62.9% (Dinnes et al., 2018) for regional lymph nodes spread and 19% (Miller et al., 2019) for distant metastasis]. Classic prognostic factors, including age and American Joint Committee on Cancer (AJCC) stage, have been proven to be effective indicators for melanoma (Kretschmer et al., 2011; Shields et al., 2012, 2015). Melanoma-specific indicators, including Clark level and Breslow thickness (Morton et al., 1993; Panda et al., 2018), are also used to assess the survival of CM

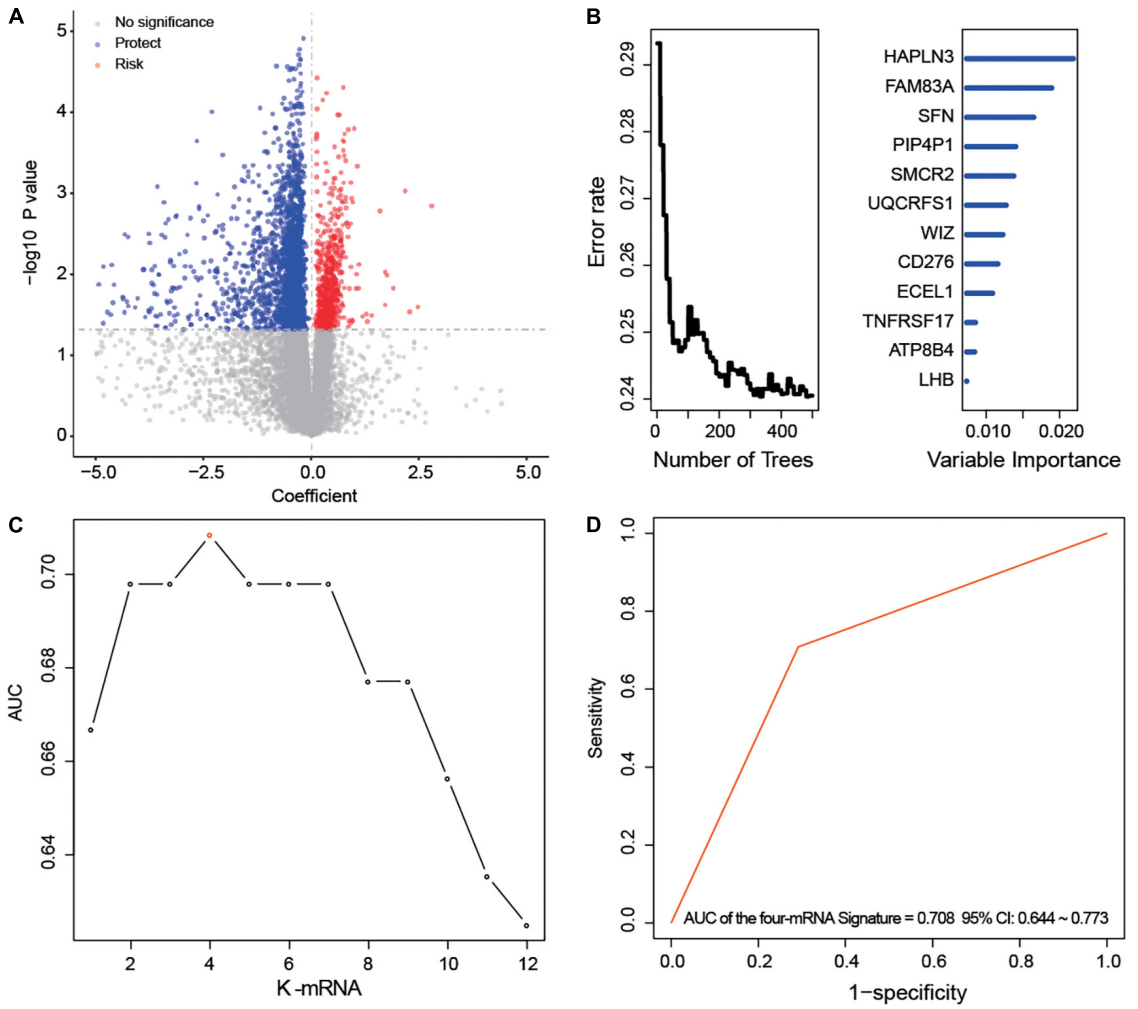


**FIGURE 1 |** Development of the prognostic messenger RNA (mRNA) signature. **(A)** The survival-associated mRNAs obtained from Cox analysis are displayed on the volcano plot. **(B)** After random forest classification algorithm, the prognosis-associated mRNAs were decreased to 12. **(C,D)** The prognostic four-mRNA signature was selected because its area under the curve (AUC) was the largest (AUC = 0.708) among the $2^{12}-1$ = 4,095 signatures.

patients. However, these clinicopathological indicators cannot reflect the molecular heterogeneity of melanoma (Palmieri et al., 2018), nor can they accurately predict the clinical outcome. Thus, novel prognostic biomarkers are extremely necessary for CM patients.

The development of sequencing technology and bioinformatics tools has promoted the discovery of new tumor biomarkers and the study of tumor molecular mechanisms.

**TABLE 1 |** Survival analysis of the messenger RNAs (mRNAs) in the prognostic signature.

| geneID | HR | Right | Left | COX P | KM P |
|---|---|---|---|---|---|
| CD276 | 1.42 | 1.07 | 1.88 | 0.01 | 0.03 |
| HAPLN3 | 0.66 | 0.55 | 0.81 | <0.001 | <0.001 |
| PIP4P1 | 0.47 | 0.30 | 0.74 | <0.001 | <0.001 |
| UQCRFS1 | 2.13 | 1.43 | 3.16 | <0.001 | <0.001 |

Based on the analysis of public messenger RNA (mRNA) expression data, studies have shown that gene signature could be prognostic marker for different types of tumors. For instance, a nine-gene signature can reliably predict the overall survival of patients with pancreatic cancer (Wu et al., 2019). An eight-gene signature showed a robust prognostic performance in early-stage non-small cell lung cancer (He et al., 2019). A 15-gene signature has been found to divide colon cancer patients into two groups with different prognosis (Xu et al., 2017). The prognostic two-gene signature has presented good predictive ability in the prognosis of GBM patients (Pan et al., 2020). For melanoma, Wang et al. (2020) identified that an eight-gene signature could independently predict the poor clinical outcome of melanoma patients.

It is well known that signatures with fewer genes have better clinical significance. In this study, through mining the mRNA expression profile and clinical information of 385 CM patients by statistical and machine-learning analyses, we aim to evaluate

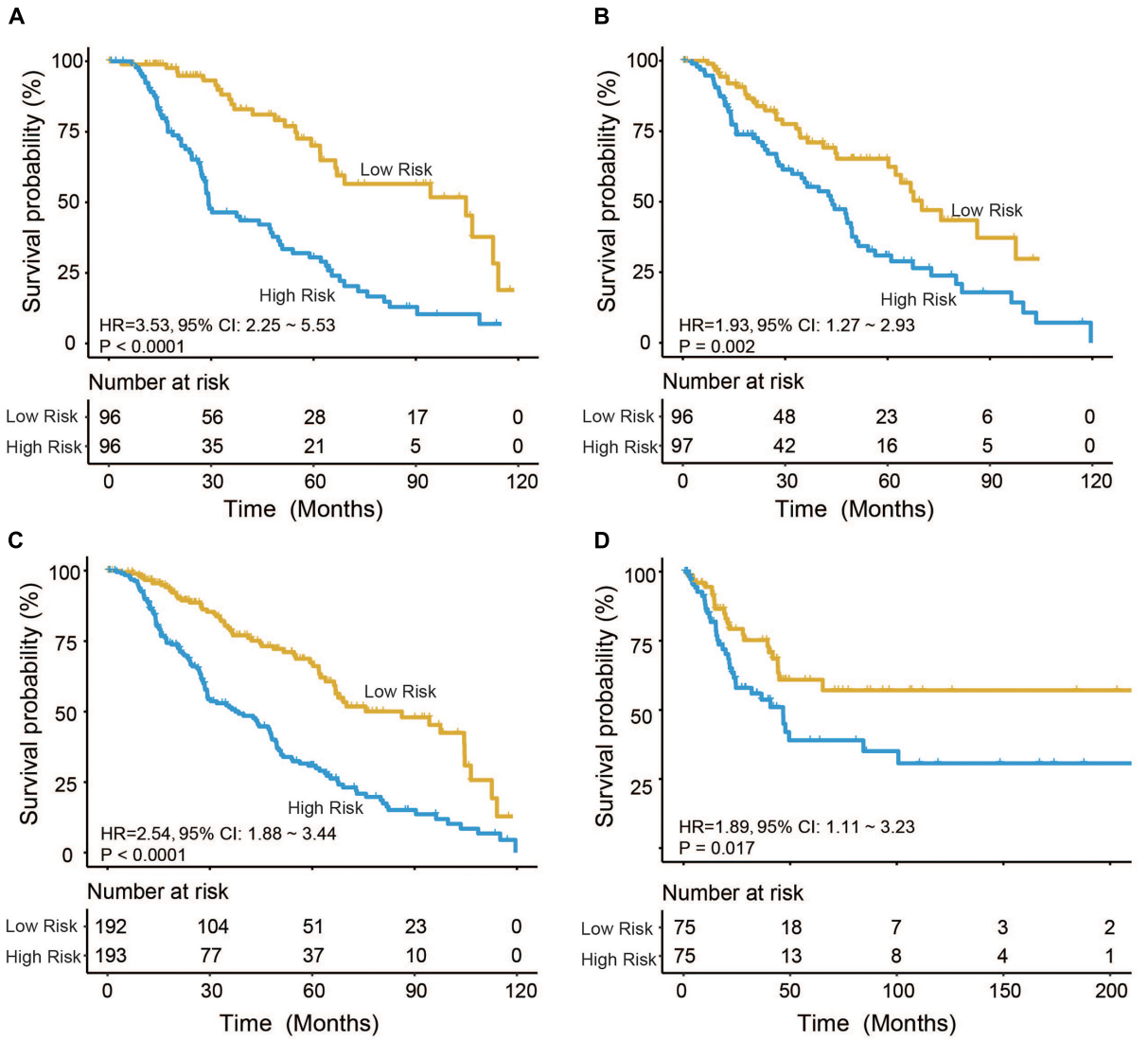


**FIGURE 2 |** Cutaneous melanoma patients were divided by the four-messenger RNA (four-mRNA) signature into two risk groups with significantly different survival outcomes in the **(A)** training, **(B)** test, **(C)** entire The Cancer Genome Atlas (TCGA), and **(D)** GSE65904 datasets.

the prognostic significance of all expressed mRNAs and construct an effective prognostic signature for CM patients.

## MATERIALS AND METHODS

### Expression Profile of CM Patients

The clinical parameters and mRNA expression data of CM in The Cancer Genome Atlas (TCGA) database[1] were from UCSC Xena[2]. The CM cases with clinical survival follow-up information were selected to establish a prognostic model, which were randomly divided into training and test groups. The independent validation set (GSE65904) was obtained from Gene Expression Omnibus (GEO) database. The clinical details of CM patients including age; gender; pathological T, N, and M stage; and tumor stage are displayed and summarized in **Supplementary Table 1**. We discarded genes whose expression values were missing in more than 20% of CM samples. All expression values were log2 transformed.

### The Process of Developing Prognostic Models Through Statistics and Machine-Learning Methods

Univariable Cox analysis was applied to seek mRNAs significantly correlated with CM patients' OS in the training group. Then, random survival forest (RSFVH, a machine learning approach) was performed. In RSFVH, an iteration procedure was implemented to reduce the node set in which the one-third

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

[2]https://xenabrowser.net/datapages/

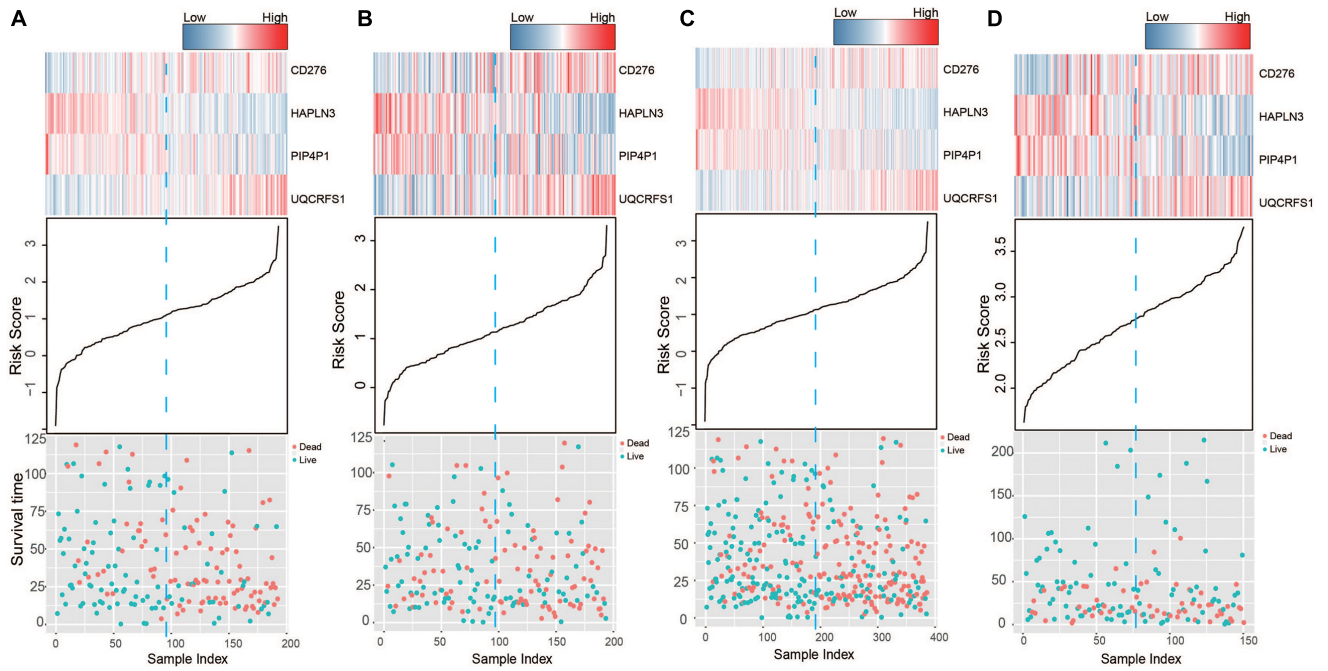


**FIGURE 3 |** The risk score distribution, survival status, and messenger RNA (mRNA) expression patterns of cutaneous melanoma patients in the **(A)** training, **(B)** test, **(C)** entire The Cancer Genome Atlas (TCGA), and **(D)** GSE65904 datasets.

least important mRNAs were discarded at each iteration step (Li et al., 2014). As a result, we obtained a set of prognosis-related mRNAs that contained a relatively small number. Subsequently, we constructed prognostic combination models as the following formula:

$$\textbf{Risk score (RS)} = \Sigma_{i=1}^{N} (coefficient_i^* \ \textbf{mRNA}Exp_i)$$

where N is the number of prognosis-related mRNAs, **mRNA*Exp*** is the expression value of prognosis-related mRNA, and the coefficient of prognosis-related mRNA is derived from Cox regression. The prognostic RS model was selected for its largest area under the curve (AUC) value from all the combinations (Zhang et al., 2020).

## Statistical Analysis

Log rank test and Kaplan–Meier (KM) analysis were performed to analyze the difference in survival between the two groups separated by the median risk score. Chi-square test was performed to test the association of the selected signature with other clinical parameters. The predictive performance of the signature in survival was tested by receiver operating characteristic (ROC) and time-dependent ROC. The R program (**??**) performed the above analyses using R packages named pROC, randomForestSRC, and survival. The prognostic mRNAs were analyzed by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis through the Cluego plug-in of Cytoscape software (Bindea et al., 2009).

# RESULTS

## Development of the Prognostic Four-mRNA Signature

From the TCGA database, we obtained 385 patients diagnosed with CM and their mRNA expression profiles including a total of 18,496 expressed mRNAs. After summarizing the clinical characteristics of CM patients, we found that the median age was 58 years (15–87 years), and a large proportion of patients were men, indicating that CM was more common in male adults. Additionally, we found that the median survival time of CM patients was 37.5 months, and only a small number of people were still alive, confirming the poor prognosis of CM (**Supplementary Table 1**).

After univariable Cox regression, we discovered 3,058 mRNAs that were significantly related to survival (red and blue dots in **Figure 1A**, $p < 0.05$). Subsequently, we screened out 12 prognostic mRNAs based on the importance score of RSFVH analysis (**Figure 1B**). When these 12 prognostic mRNAs were incorporated into the risk prediction model in different combinations, we got a total of $2^{12}-1 = 4{,}095$ possible signatures. After performing ROC analysis on all 4,095 signatures, we found a four-mRNA signature with the largest AUC value (AUCsignature = 0.708; **Figures 1C,D** and **Table 1**). The prognostic four mRNAs from the signature were CD276, HAPLN3, PIP4P1, and UQCRFS1. The selected RS model formula was as follows: RS = (0.35 × CD276 expression level) + (0.75 × UQCRFS1 expression level) + (-0.41 × HAPLN3 expression level) + (-0.75 × PIP4P1 expression level).

**TABLE 2 |** Association of the messenger RNA (mRNA) signature with clinical characteristics in cutaneous melanoma (CM) patients.

| Variables | Training group | | *p* | Test group | | *p* | TCGA group | | *p* |
|---|---|---|---|---|---|---|---|---|---|
| | Low risk* | High risk* | | Low risk* | High risk* | | Low risk* | High risk* | |
| Age (years) | | | 0.89 | | | 0.27 | | | 0.57 |
| ≤ 58 | 48 | 46 | | 38 | 47 | | 86 | 93 | |
| >58 | 48 | 50 | | 58 | 50 | | 106 | 100 | |
| Gender | | | 0.38 | | | 1.00 | | | 0.50 |
| Female | 43 | 36 | | 33 | 33 | | 76 | 69 | |
| Male | 53 | 60 | | 63 | 64 | | 116 | 124 | |
| Radiotherapy | | | <0.001 | | | 0.13 | | | <0.001 |
| Unknown | 25 | 50 | | 29 | 41 | | 54 | 91 | |
| No | 63 | 42 | | 57 | 51 | | 120 | 93 | |
| Yes | 8 | 4 | | 10 | 5 | | 18 | 9 | |
| Pathological M | | | 0.33 | | | 0.52 | | | 0.21 |
| M0 | 84 | 89 | | 82 | 86 | | 166 | 175 | |
| M1 | 6 | 5 | | 5 | 6 | | 11 | 11 | |
| Pathological N | | | 0.56 | | | 0.31 | | | 0.19 |
| N0 | 48 | 44 | | 42 | 37 | | 90 | 81 | |
| N1 | 14 | 19 | | 19 | 16 | | 33 | 35 | |
| N2 | 11 | 9 | | 12 | 12 | | 23 | 21 | |
| N3 | 11 | 16 | | 10 | 18 | | 21 | 34 | |
| Pathological T | | | 0.17 | | | 0.31 | | | 0.27 |
| T0 | 5 | 1 | | 6 | 7 | | 11 | 8 | |
| T1 | 12 | 9 | | 8 | 1 | | 20 | 10 | |
| T2 | 16 | 17 | | 12 | 14 | | 28 | 31 | |
| T3 | 16 | 25 | | 17 | 17 | | 33 | 42 | |
| T4 | 30 | 37 | | 35 | 41 | | 65 | 78 | |
| Tumor stage | | | 0.31 | | | 0.87 | | | 0.38 |
| Stage 0 | 0 | 1 | | 1 | 0 | | 1 | 1 | |
| Stage I | 20 | 14 | | 11 | 9 | | 31 | 23 | |
| I/II Nos | 2 | 0 | | 1 | 0 | | 3 | 0 | |
| Stage II | 29 | 31 | | 29 | 31 | | 58 | 62 | |
| Stage III | 32 | 42 | | 41 | 44 | | 73 | 86 | |
| Stage IV | 6 | 5 | | 5 | 6 | | 11 | 11 | |
| Race demographic | | | 0.22 | | | 0.03 | | | 0.51 |
| Asian | 3 | 3 | | 2 | 4 | | 5 | 7 | |
| White | 93 | 90 | | 88 | 93 | | 181 | 183 | |

**Low risk ≤ median of risk score, high risk > median of risk score.*

## The Survival Prediction Performance and Validation of the Four-mRNA Signature

Based on the model constructed above, the risk scores of CM patients were calculated, and the median risk score was obtained as the cutoff. In the training dataset, the median RS divided patients into a high-risk group ($n$ = 96) or a low-risk group ($n$ = 96). Kaplan–Meier analysis showed that there were a significant difference in survival time between patients in the high- and the low-risk group (median survival time: 29.2 vs. 104.7 months, $p < 0.001$; **Figure 2A**). Then, Kaplan–Meier analysis was performed in the test ($n$ = 193) and entire TCGA datasets ($n$ = 385). The four-mRNA signature could distinguish the CM patients into two risk groups with different survival outcomes in the test group (median survival time: 43.8 vs. 70.0 months, $p < 0.001$; **Figure 2B**). The same performance for survival prediction was shown in the entire TCGA dataset (median survival time: 38.5 vs. 86.3 months, $p < 0.001$; **Figure 2C**). In addition, we also verified its survival prediction performance in an independent set (GSE65904, $n$ = 150) from GEO database. The median RS value also classified patients from GSE65904 into high- or low-risk group significantly ($p$ = 0.017, **Figure 2D**). Moreover, when the patient's mRNA expression, survival time, and risk score were displayed on the same chart, we found that CM patients with higher risk mRNAs expression and higher risk scores had poorer survival outcomes in the training (**Figure 3A**), test (**Figure 3B**), entire TCGA datasets (**Figure 3C**), and GSE65904 (**Figure 3D**).

## Independent Prognostic Value of the Four-mRNA Signature

Chi-square test did not detect the relationship between the four-mRNA signature and other clinical variables (**Table 2**). Then, we

tested the independence of the four-mRNA model by performing Cox regression analysis (including univariate and multivariable Cox, **Table 3**). Multivariable Cox regression confirmed the independence of four-mRNA signature in prognosis prediction in the training, test, entire TCGA, or GSE65904 datasets (HR training = 3.00, $p < 0.001$, $n = 192$; HR test = 1.72, $p < 0.05$, $n = 193$; HR entire = 2.00, $p < 0.001$, $n = 385$; HRGSE65904 = 2.25, $p < 0.001$, $n = 150$, **Table 3**).

## The Comparison of the Performance in Survival Prediction Between the Four-mRNA Signature With Melanoma Clark Level and Tumor Stage

We compared the performance of the four-mRNA signature with other clinical prognostic markers (including melanoma Clark level and tumor stage) in predicting survival. **Figure 4A** shows that the four-mRNA signature was better than other clinical variables in survival prediction of the entire TCGA set (AUC signature = 0.67 vs. AUCtumor stage = 0.52 vs. AUCmelanoma Clark level = 0.55). TimeROC analysis found that the AUC values of the four-mRNA signature within 1–12 years were greater than that of the Clark level or tumor stage (**Figure 4B**). All these suggest that the four-mRNA signature have better performance in survival prediction of CM.

## Function Enrichment Analysis of the Four Selected mRNAs

To explore the biological roles of the four selected mRNAs from the signature in the CM, we conducted Pearson correlation test and obtained a total of 533 coexpressed genes in the TCGA dataset (coefficient > 0.5/ < -0.5, $p < 0.001$). Then, KEGG and GO analysis were performed on the above 533 coexpressed genes. We found that these coexpressed genes were significantly enriched in 739 GO terms and 45 KEGG pathways ($p < 0.05$), such as regulation of immune system process, leukocyte activation, T-cell activation, Toll-like receptor signaling pathway, Jak-STAT signaling pathway, nuclear factor (NF)-kappa B signaling pathway, suggesting the four mRNAs may influence the immune function of CM patients (top 30 are shown, **Figure 5**).

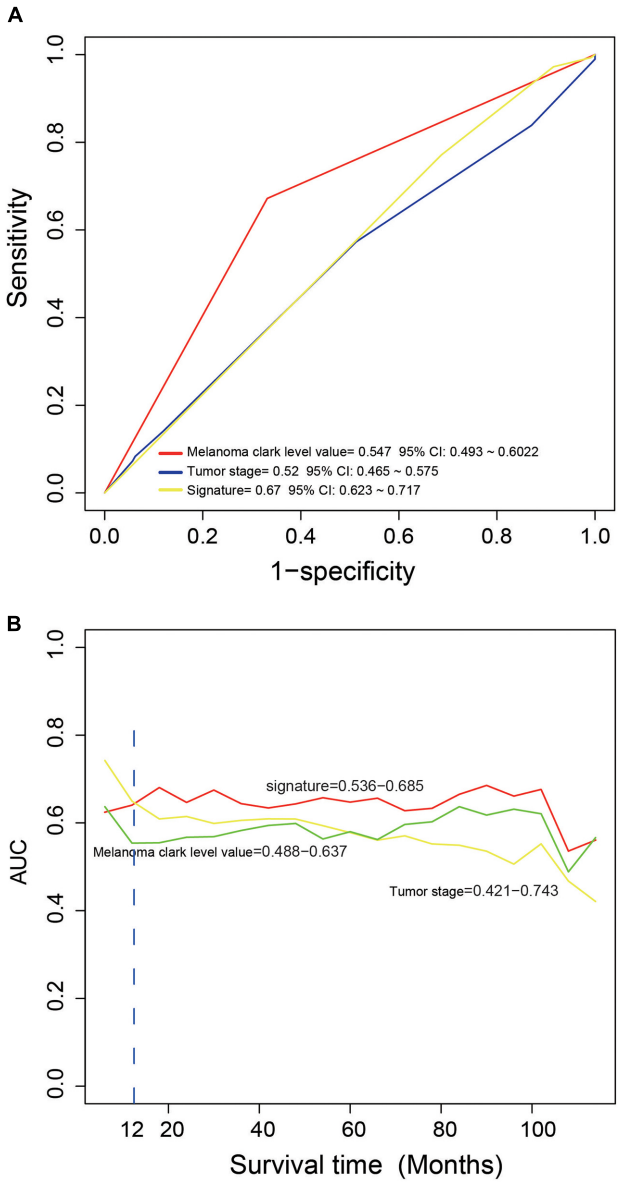


**FIGURE 4 |** The comparison of the performance in survival prediction between the four-messenger RNA (four-mRNA) signature with tumor stage and Clark level. **(A)** ROC analysis was performed to compare the performance of the four-mRNA signature with that of tumor stage and Clark level. **(B)** TimeROC analysis was conduct to compare the performance of the four-mRNA signature with that of tumor stage and Clark level.

# DISCUSSION

Cutaneous melanoma is a highly malignant disease with large difference in prognosis, lacking effective biomarkers for accurate survival prediction or reliable prognostic indicators. The application of precision medicine in the field of oncology highlights the prediction of individual prognosis based on gene expression profiles. Through analyzing gene expression profiles, gene expression signatures have been used to predict the prognosis of patients in different types of cancer (Burska et al., 2014), such as glioblastoma, esophageal squamous cell carcinoma, breast cancer, lung cancer, hepatocellular carcinoma, and bladder carcinoma. Therefore, using statistics and machine learning approaches, we analyzed the clinical survival information and mRNA expression of 385 melanoma patients and developed the four-mRNA signature, which could be a good prognostic biomarker for patients with CM.

After performing a variety of bioinformatics analysis methods in the TCGA group, a prognostic risk model based on the expression of four mRNAs was established to distinguish CM patients with different prognosis. This model has two advantages in predicting the prognosis of CM patients: first, it is an independent prognostic biomarker, which means that

**TABLE 3 |** Cox regression analysis of the signature with the survival of cutaneous melanoma (CM).

| Variables | | Univariable analysis | | | | Multivariable analysis | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | HR | 95% CI of HR | | *P* | HR | 95% CI of HR | | *p* |
| | | | lower | upper | | | lower | upper | |
| **The training group** | | | | | | | | | |
| Age | >58 vs. ≤58 | 1.47 | 0.98 | 2.22 | 0.07 | 1.24 | 0.77 | 2.01 | 0.38 |
| Gender | Male vs. female | 0.79 | 0.52 | 1.20 | 0.27 | 1.03 | 0.64 | 1.67 | 0.89 |
| Tumor stage | III, IV vs. I, II | 1.01 | 0.98 | 1.03 | 0.60 | 1.03 | 0.97 | 1.08 | 0.36 |
| Melanoma Clark level | IV, V vs. I, II, III | 1.87 | 1.16 | 3.02 | 0.01 | 1.40 | 1.01 | 1.94 | 0.04 |
| Signature | High risk vs. low risk | 3.53 | 2.25 | 5.53 | < 0.001 | 3.00 | 1.77 | 5.06 | < 0.001 |
| **The test group** | | | | | | | | | |
| Age | >58 vs. ≤58 | 0.87 | 0.58 | 1.31 | 0.51 | 0.85 | 0.51 | 1.41 | 0.52 |
| Gender | Male vs. female | 1.01 | 0.65 | 1.58 | 0.96 | 0.75 | 0.43 | 1.31 | 0.31 |
| Tumor stage | III, IV vs. I, II | 0.99 | 0.97 | 1.02 | 0.56 | 0.98 | 0.94 | 1.01 | 0.24 |
| Melanoma Clark level | IV, V vs. I, II, III | 1.35 | 0.86 | 2.11 | 0.20 | 1.33 | 0.94 | 1.89 | 0.11 |
| Signature | High risk vs. low risk | 1.93 | 1.27 | 2.93 | <0.001 | 1.72 | 1.02 | 2.90 | 0.04 |
| **The TCGA dataset** | | | | | | | | | |
| Age | >58 vs. ≤58 | 1.13 | 0.85 | 1.50 | 0.40 | 0.90 | 0.50 | 1.64 | 0.74 |
| Gender | Male vs. female | 0.90 | 0.67 | 1.22 | 0.51 | 0.98 | 0.52 | 1.85 | 0.95 |
| Tumor stage | III, IV vs. I, II | 1.00 | 0.98 | 1.01 | 0.80 | 1.04 | 0.99 | 1.10 | 0.14 |
| Melanoma Clark level | IV,V vs. I, II,III | 1.55 | 1.13 | 2.13 | 0.01 | 1.88 | 1.19 | 2.96 | 0.01 |
| Radiation therapy | Yes vs. no | 0.74 | 0.35 | 1.55 | 0.42 | 0.68 | 0.23 | 2.06 | 0.50 |
| Signature | High risk vs. low risk | 2.54 | 1.88 | 3.44 | <0.001 | 2.00 | 1.08 | 3.69 | 0.03 |
| **The GSE65904 dataset** | | | | | | | | | |
| Age | >58 vs. ≤58 | 1.61 | 0.91 | 2.85 | 0.10 | 1.52 | 1.00 | 2.29 | 0.05 |
| Gender | Male vs. female | 2.56 | 1.36 | 4.85 | <0.001 | 0.95 | 0.62 | 1.45 | 0.82 |
| Signature | High risk vs. low risk | 1.89 | 1.11 | 3.23 | 0.02 | 3.55 | 2.25 | 5.58 | < 0.001 |

known clinical predicators such as melanoma Clark level and tumor grade will not affect its prognosis prediction. Second, the AUC value of the four-mRNA signature is greater than that of melanoma Clark and tumor grade, indicating that the four-mRNA signature has the best survival predictive performance.

The high expression of CD276 and UQCRFS1 in the four-mRNA signature was associated with short OS (Cox regression coefficient > 0), suggesting that they were risk factors for prognostic. Meanwhile, the high expression of HAPLN3 and PIP4P1 were associated with long OS (Cox regression coefficient < 0), which indicates that these two genes were beneficial factors for prognosis. Among the candidate genes, CD276 (B7-H3) is an important component of the B7 family, which can provide stimulus or inhibitory signals to enhance or weaken T-cell immune response. CD276 at the mRNA level is widely expressed in normal tissues. In tumors, CD276 or B7-H3 is reported to be highly expressed in tumors of various tissue types including melanoma, which is closely related to the poor clinical outcome of tumor patients. Studies have shown that B7-H3 plays an important role in tumor immune escape and can also affect tumor proliferation, invasion, and migration (Castellanos et al., 2017; Dong et al., 2018). In accordance with the results of this article, researchers found that B7-H3 is a significant factor in tumor progression and poor prognosis for CM patients (Tekle et al., 2012; Wang et al., 2013). HAPLN3 is a member of the hyaluronan and proteoglycan binding link protein gene family with a length of 2.1 kb and is widely expressed in various tissues (Spicer et al., 2003). HAPLN3 has been reported to play an important role in maintaining the stability of the extracellular matrix, thereby regulating the mobility and migration of tumor cells. Kuo et al. (2010) found that HAPLN3 was significantly overexpressed in breast cancer tissues, but there was no correlation between HAPLN3 gene expression and overall survival. Ubiquinol cytochrome c reductase (UQCRFS1, also named as Rieske Fe–S protein) is a catalytic submit of complex III and plays an important role in the mitochondrial respiratory chain. Researchers have found that UQCRFS1 is overexpressed in ovarian carcinoma and gastric cancer and may promote tumor development (Kaneko et al., 2003; Jun et al., 2012). Phosphatidylinositol-4, 5-bisphosphate 4-phosphatase (PIP4P1, known as TMEM55B) is an enzyme that influences cholesterol homeostasis (Medina et al., 2014) and regulates lysosomal positioning (Willett et al., 2017). However, there is no report about the role of TMEM55B in tumors.

In recent years, the role of immunity and inflammation in tumor progression has been gradually discovered (Coussens and Werb, 2002; Keibel et al., 2009). To explore the function of genes in the four-mRNA signature, GO and KEGG analyses were performed and identified that these genes were enriched in several immune and inflammation-related pathways, such as

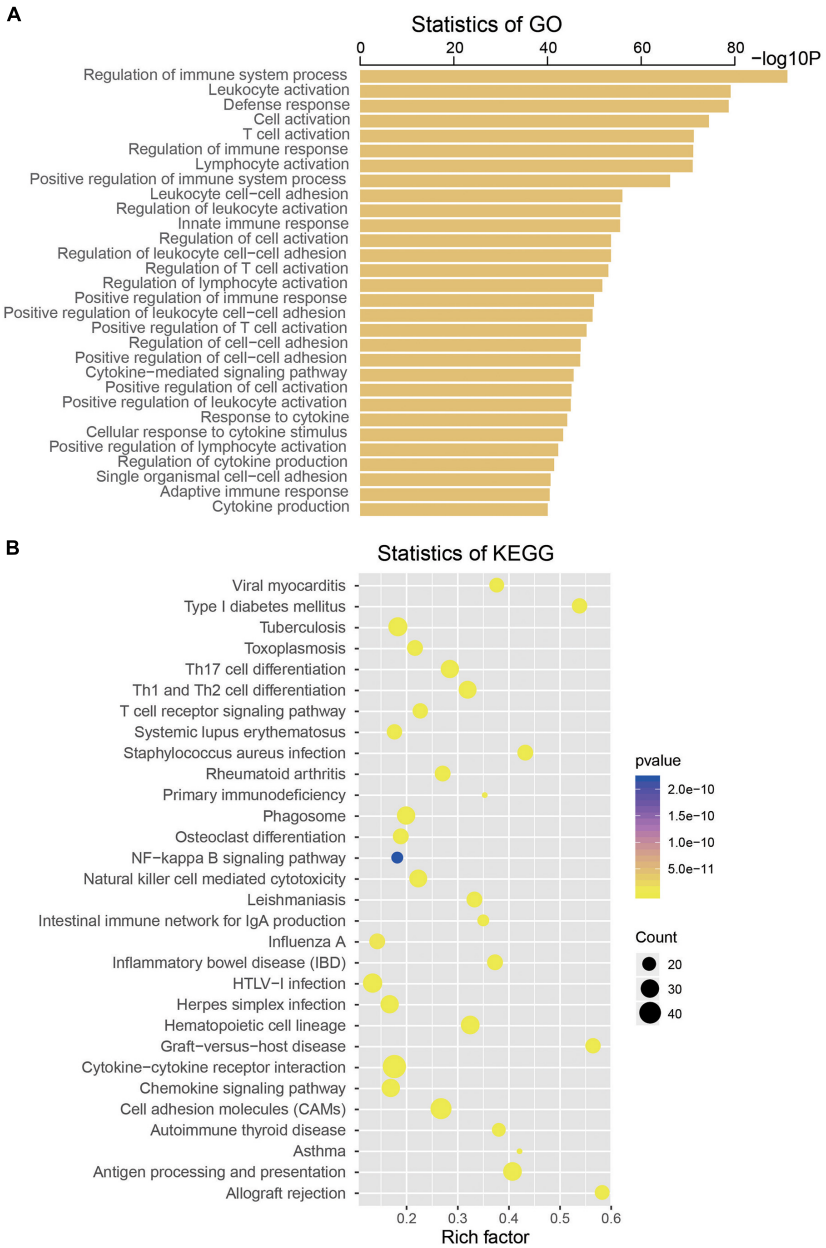


**FIGURE 5 |** Functional enrichment analysis of the four messenger RNAs (mRNAs) in the signature by Gene Ontology **(A)** and Kyoto Encyclopedia of Genes and Genomes **(B)**.

regulation of immune system process, regulation of inflammatory response and Toll-like receptor signaling pathway, Jak-STAT signaling pathway, and NF-kappa B signaling pathway, which suggests that the four-mRNA signature might influence the survival of patients with CM through regulating immune and inflammation-related pathways (Pansky et al., 2000; Janostiak et al., 2017; Luo et al., 2018; Rathore et al., 2019).

In summary, our study developed a prognostic four-mRNA signature (CD276, HAPLN3, PIP4P1, UQCRFS1) for CM, which can predict the clinical outcome of patients. Since its prognostic ability is better than the current markers (Clark level or tumor stage), the four-mRNA signature would have stronger clinical application value.

## DATA AVAILABILITY STATEMENT

Publicly available datasets analyzed in this study can be found here: https://xenabrowser.net/datapages/?cohort=TCGA%20Melanoma%20(SKCM)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65904.

## AUTHOR CONTRIBUTIONS

JL designed the study. HB and YW performed all analyses. HL checked the final manuscript. All authors reviewed the manuscript and approved the manuscript for publication.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.680617/full#supplementary-material

## REFERENCES


Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101

Burska, A. N., Roget, K., Blits, M., Soto Gomez, L., van de Loo, F., Hazelwood, L. D., et al. (2014). Gene expression analysis in RA: towards personalized medicine. *Pharmacogenomics J.* 14, 93–106. doi: 10.1038/tpj.2013.48

Castellanos, J. R., Purvis, I. J., Labak, C. M., Guda, M. R., Tsung, A. J., Velpula, K. K., et al. (2017). B7-H3 role in the immune landscape of cancer. *Am. J. Clin. Exp. Immunol.* 6, 66–75.

Coussens, L. M., and Werb, Z. (2002). Inflammation and cancer. *Nature* 420, 860–867. doi: 10.1038/nature01322

Dimitriou, F., Krattinger, R., Ramelyte, E., Barysch, M. J., Micaletto, S., Dummer, R., et al. (2018). The world of Melanoma: epidemiologic, genetic, and anatomic differences of Melanoma across the Globe. *Curr. Oncol. Rep.* 20:87. doi: 10.1007/s11912-018-0732-8

Dinnes, J., Deeks, J. J., Saleh, D., Chuchu, N., Bayliss, S. E., Patel, L., et al. (2018). Skin Cancer Diagnostic Test Accuracy, Reflectance confocal microscopy for diagnosing cutaneous melanoma in adults. *Cochrane Database Syst. Rev.* 12:CD013190. doi: 10.1002/14651858.CD013190

Dong, P., Xiong, Y., Yue, J., Hanley, S. J. B., and Watari, H. (2018). B7H3 As a Promoter of Metastasis and Promising Therapeutic Target. *Front. Oncol.* 8:264. doi: 10.3389/fonc.2018.00264

He, R., Zuo, S., and Robust, A. (2019). 8-Gene Prognostic Signature for Early-Stage Non-small Cell Lung Cancer. *Front. Oncol.* 9:693. doi: 10.3389/fonc.2019.00693

Janostiak, R., Rauniyar, N., Lam, T. T., Ou, J., Zhu, L. J., Green, M. R., et al. (2017). Growth by Stimulating the NF-kappaB Pathway. *Cell Rep.* 21, 2829–2841. doi: 10.1016/j.celrep.2017.11.033

Jun, K. H., Kim, S. Y., Yoon, J. H., Song, J. H., and Park, W. S. (2012). Amplification of the UQCRFS1 Gene in Gastric Cancers. *J. Gastric Cancer* 12, 73–80. doi: 10.5230/jgc.2012.12.2.73

Kaneko, S. J., Gerasimova, T., Smith, S. T., Lloyd, K. O., Suzumori, K., and Young, S. R. (2003). CA125 and UQCRFS1 FISH studies of ovarian carcinoma. *Gynecol. Oncol.* 90, 29–36. doi: 10.1016/S0090-8258(03)00144-6

Keibel, A., Singh, V., and Sharma, M. C. (2009). Inflammation, microenvironment, and the immune system in cancer progression. *Curr. Pharm. Des.* 15, 1949–1955. doi: 10.2174/138161209788453167

Kretschmer, L., Starz, H., Thoms, K. M., Satzger, I., Volker, B., Jung, K., et al. (2011). Age as a key factor influencing metastasizing patterns and disease-specific survival after sentinel lymph node biopsy for cutaneous melanoma. *Int. J. Cancer* 129, 1435–1442. doi: 10.1002/ijc.25747

Kuo, S. J., Chien, S. Y., Lin, C., Chan, S. E., Tsai, H. T., and Chen, D. R. (2010). Significant elevation of CLDN16 and HAPLN3 gene expression in human breast cancer. *Oncol. Rep.* 24, 759–766. doi: 10.3892/or_00000918

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

Luo, N., Formisano, L., Gonzalez-Ericsson, P. I., Sanchez, V., Dean, P. T., Opalenik, S. R., et al. (2018). Melanoma response to anti-PD-L1 immunotherapy requires JAK1 signaling, but not JAK2. *Oncoimmunology* 7:e1438106. doi: 10.1080/2162402X.2018.1438106

Medina, M. W., Bauzon, F., Naidoo, D., Theusch, E., Stevens, K., Schilde, J., et al. (2014). Transmembrane protein 55B is a novel regulator of cellular cholesterol metabolism. *Arterioscler. Thromb. Vasc. Biol.* 34, 1917–1923. doi: 10.1161/ATVBAHA.113.302806

Miller, K. D., Nogueira, L., Mariotto, A. B., Rowland, J. H., Yabroff, K. R., Alfano, C. M., et al. (2019). Cancer treatment and survivorship statistics, 2019. *CA Cancer J. Clin.* 69, 363–385. doi: 10.3322/caac.21565

Morton, D. L., Davtyan, D. G., Wanek, L. A., Foshag, L. J., and Cochran, A. J. (1993). Multivariate analysis of the relationship between survival and the microstage of primary melanoma by Clark level and Breslow thickness. *Cancer* 71, 3737–3743. doi: 10.1002/1097-0142(19930601)71:11<3737::AID-CNCR2820711143>3.0.CO;2-7

Palmieri, G., Colombino, M., Casula, M., Manca, A., Mandala, M., Cossu, A., et al. (2018). Molecular pathways in Melanomagenesis: what we learned from Next-Generation sequencing approaches. *Curr. Oncol. Rep.* 20:86. doi: 10.1007/s11912-018-0733-7

Pan, Y., Zhang, J. H., Zhao, L., Guo, J. C., Wang, S., Zhao, Y., et al. (2020). A robust two-gene signature for glioblastoma survival prediction. *J. Cell. Biochem.* 121, 3593–3605. doi: 10.1002/jcb.29653

Panda, S., Dash, S., Besra, K., Samantaray, S., Pathy, P. C., and Rout, N. (2018). Clinicopathological study of malignant melanoma in a regional cancer center. *Indian J. Cancer* 55, 292–296. doi: 10.4103/ijc.IJC_612_17

Pansky, A., Hildebrand, P., Fasler-Kan, E., Baselgia, L., Ketterer, S., Beglinger, C., et al. (2000). Defective Jak-STAT signal transduction pathway in melanoma cells resistant to growth inhibition by interferon-alpha. *Int. J. Cancer* 85, 720–725. doi: 10.1002/(SICI)1097-0215(20000301)85:5<720::AID-IJC20>3.0.CO;2-O

Rathore, M., Girard, C., Ohanna, M., Tichet, M., Ben Jouira, R., Garcia, E., et al. (2019). Cancer cell-derived long pentraxin 3 (PTX3) promotes melanoma migration through a toll-like receptor 4 (TLR4)/NF-kappaB signaling pathway. *Oncogene* 38, 5873–5889. doi: 10.1038/s41388-019-0848-9

Schadendorf, D., van Akkooi, A. C. J., Berking, C., Griewank, K. G., Gutzmer, R., Hauschild, A., et al. (2018). Melanoma. *Lancet* 392, 971–984. doi: 10.1016/S0140-6736(18)31559-9

Shields, C. L., Kaliki, S., Furuta, M., Fulco, E., Alarcon, C., and Shields, J. A. (2015). American Joint Committee on Cancer Classification of Uveal Melanoma (Anatomic Stage) Predicts Prognosis in 7,731 Patients: the 2013 Zimmerman Lecture. *Ophthalmology* 122, 1180–1186. doi: 10.1016/j.ophtha.2015.01.026

Shields, C. L., Kaliki, S., Furuta, M., Mashayekhi, A., and Shields, J. A. (2012). Clinical spectrum and prognosis of uveal melanoma based on age at presentation in 8,033 cases. *Retina* 32, 1363–1372. doi: 10.1097/IAE.0b013e31824d09a8

Spicer, A. P., Joo, A., and Bowling, R. A. Jr. (2003). A hyaluronan binding link protein gene family whose members are physically linked adjacent to chondroitin sulfate proteoglycan core protein genes: the missing links. *J. Biol. Chem.* 278, 21083–21091. doi: 10.1074/jbc.M213100200

Tekle, C., Nygren, M. K., Chen, Y. W., Dybsjord, I., Nesland, J. M., Maelandsmo, G. M., et al. (2012). B7-H3 contributes to the metastatic capacity of melanoma cells by modulation of known metastasis-associated genes. *Int. J. Cancer* 130, 2282–2290. doi: 10.1002/ijc.26238

Wang, J., Chong, K. K., Nakamura, Y., Nguyen, L., Huang, S. K., Kuo, C., et al. (2013). B7-H3 associated with tumor progression and epigenetic regulatory activity in cutaneous melanoma. *J. Investig. Dermatol.* 133, 2050–2058. doi: 10.1038/jid.2013.114

Wang, J., Kong, P. F., Wang, H. Y., Song, D., Wu, W. Q., Zhou, H. C., et al. (2020). Identification of a Gene-Related Risk Signature in Melanoma Patients Using Bioinformatic Profiling. *J. Oncol.* 2020:7526204. doi: 10.1155/2020/7526204

Willett, R., Martina, J. A., Zewe, J. P., Wills, R., Hammond, G. R. V., and Puertollano, R. (2017). TFEB regulates lysosomal positioning by modulating TMEM55B expression and JIP4 recruitment to lysosomes. *Nat. Commun.* 8:1580. doi: 10.1038/s41467-017-01871-z

Wu, M., Li, X., Zhang, T., Liu, Z., and Zhao, Y. (2019). Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. *Front. Oncol.* 9:996. doi: 10.3389/fonc.2019.00996

Xu, G., Zhang, M., Zhu, H., and Xu, J. (2017). A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. *Gene* 604, 33–40.

Zhang, J. H., Hou, R., Pan, Y., Gao, Y., Yang, Y., Tian, W., et al. (2020). A five-microRNA signature for individualized prognosis evaluation and radiotherapy guidance in patients with diffuse lower-grade glioma. *J Cell. Mol. Med.* 24, 7504–7514. doi: 10.1111/jcmm.15377

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 December 2021
doi: 10.3389/fgene.2021.675865

# Establishment of A Nomogram for Predicting the Prognosis of Soft Tissue Sarcoma Based on Seven Glycolysis-Related Gene Risk Score

*Yuhang Liu, Changjiang Liu, Hao Zhang, Xinzeyu Yi and Aixi Yu**

*Department of Trauma and Microsurgery Orthopedics, Zhongnan Hospital of Wuhan University, Wuhan, China*

**Background:** Soft tissue sarcoma (STS) is a group of tumors with a low incidence and a complex type. Therefore, it is an arduous task to accurately diagnose and treat them. Glycolysis-related genes are closely related to tumor progression and metastasis. Hence, our study is dedicated to the development of risk characteristics and nomograms based on glycolysis-related genes to assess the survival possibility of patients with STS.

**Methods:** All data sets used in our research include gene expression data and clinical medical characteristics in the Genomic Data Commons Data Portal (National Cancer Institute) Soft Tissue Sarcoma (TCGA SARC) and GEO database, gene sequence data of corresponding non-diseased human tissues in the Genotype Tissue Expression (GTEx).Next, transcriptome data in TCGA SARC was analyzed as the training set to construct a glycolysis-related gene risk signature and nomogram, which were confirmed in external test set.

**Results:** We identified and verified the 7 glycolysis-related gene signature that is highly correlated with the overall survival (OS) of STS patients, which performed excellently in the evaluation of the size of AUC, and calibration curve. As well as, the results of the analysis of univariate and multivariate Cox regression demonstrated that this 7 glycolysis-related gene characteristic acts independently as an influence predictor for STS patients. Therefore, a prognostic-related nomogram combing 7 gene signature with clinical influencing features was constructed to predict OS of patients with STS in the training set that demonstrated strong predictive values for survival.

**Conclusion:** These results demonstrate that both glycolysis-related gene risk signature and nomogram were efficient prognostic indicators for patients with STS. These findings may contribute to make individualize clinical decisions on prognosis and treatment.

**Keywords: glycolysis-related gene, prognostic model, bioinformatics analysis, soft tissue sarcoma, biomarker, nomogram**

OPEN ACCESS

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

***Reviewed by:***
*Gary S. Stein,*
*University of Vermont, United States*
*Xiaoxu Yang,*
*University of California, San Diego, United States*

****Correspondence:***
*Aixi Yu*
*yuaixi@whu.edu.cn*

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

***Received:*** *04 March 2021*
***Accepted:*** *16 November 2021*
***Published:*** *02 December 2021*

***Citation:***
*Liu Y, Liu C, Zhang H, Yi X and Yu A (2021) Establishment of A Nomogram for Predicting the Prognosis of Soft Tissue Sarcoma Based on Seven Glycolysis-Related Gene Risk Score. Front. Genet. 12:675865. doi: 10.3389/fgene.2021.675865*

## INTRODUCTION

Soft tissue sarcoma (STS) accounts for 1% of adult cancer. Although its occurrence is relatively rare, STS highly heterogeneous. In the United States, approximately 13,500 new people were diagnosed with STS in 2019 (Siegel et al., 2019). There are more than 100 subtypes of STS, and the clinical characteristics of each subtype are different (Choi and Ro, 2020). In general, even if the primary tumor is removed, 25% of the patients will develop distant metastasis (Brennan et al., 2014). The ability to precisely predict the outcome on the basis of every patient's clinical information, pathological and molecular features have attracted growing attention especially in the era of precision tumor treatment. A number of reports in literature have predicted the survival status of STS patients (Mariani et al., 2005; Cahlon et al., 2012; Callegaro et al., 2016; Gamboa et al., 2020). However, nomograms in these studies were all limited to be based on the clinical features that can only be determined after surgery; in consequence, their clinical applications might be restricted. The ideal nomograms should also include biomarkers, molecular signatures, and genomic expression to help predict survival status more accurately (Gamboa et al., 2020). With the rapid development of bioinformatics tools, multiple biomarkers for clinical diagnosis, treatment, and prognosis prediction have been identified (Wu et al., 2018; Li et al., 2019a; Long et al., 2019). The heterogeneity of the genome and the low response to traditional therapies warrant the development of effective therapeutic targets. Therefore, it is essential to determine potential precise new clinical prognostic biomarkers and therapeutic targets.

Metabolic pattern changes are one of the hallmarks of tumor cells. They tend to have a higher glycolysis efficiency, which is called the Warburg effect (Ganapathy-Kanniappan and GeschwindGeschwind, 2013). Glycolysis is essential for the development, invasion, metastasis, and drug resistance of tumor cells (Gatenby and Gillies, 2004). Moreover, previous studies have shown that targeting glycolysis-related metabolic pathways can effectively inhibit the growth of tumor cells (Abdel-Wahab et al., 2019). However, only few studies have investigated the role of glycolysis-related genes in STS. Huangyang et al. reported that the expression level of gluconeogenic isozyme fructose-1,6-bisphosphatase 2 (FBP2) was down-regulated in most subtypes of STS, and the re-expression of FBP2 significantly inhibited tumor growth (Huangyang et al., 2020). Mao et al. (2016) found that melatonin could inhibit the Warburg effect and directly inhibit the growth of leiomyosarcoma tumors. However, further studies on the mechanism of glycolysis-related genes are still necessary to develop more effective treatment for STS patients.

The Cancer Genome Atlas (TCGA) project aims to provide comprehensive transcriptome data and corresponding clinical information for various cancer patients (Jia et al., 2018). The GTEx database provides transcriptome sequencing data of 54 normal tissue samples from nearly 1,000 individuals (GTEx Consortium, 2015).

In our work, we comprehensively analyzed the RNA sequencing profile and clinical characteristics of the TCGA-SARC data set and the GTEx data set to screen out 7 glycolysis-related genes that are related to the prognosis of STS patients, and established a 7 gene prognostic signature. Furthermore, we constructed and verified a predictive nomogram for STS patients based on the 7-gene signature, which may aid in determining the prognostic status of STS patients and guiding tumor therapy and postoperative monitor. The workflow of our study is shown in **Figure 1**.

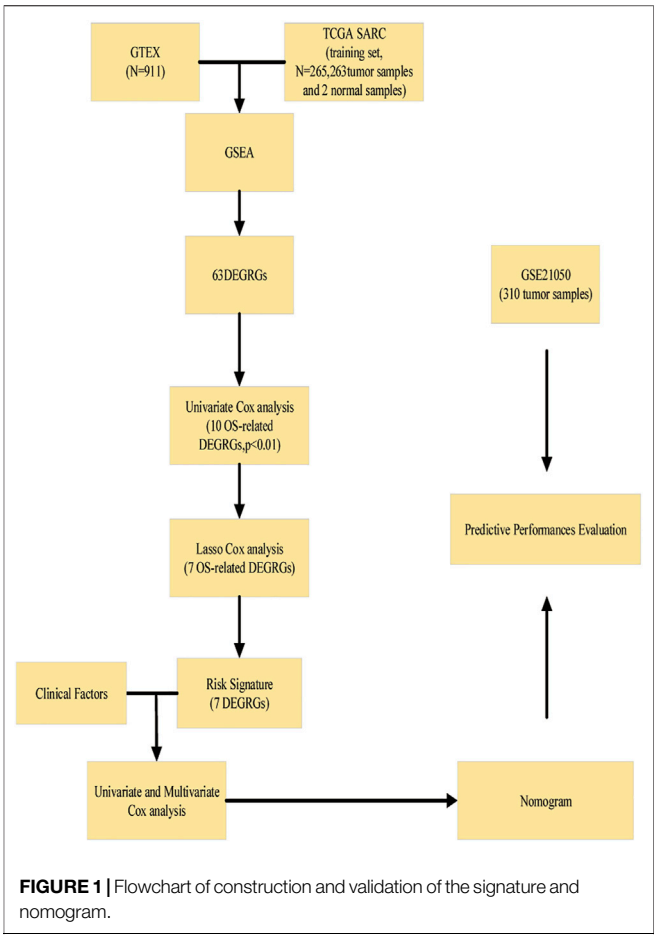


**FIGURE 1 |** Flowchart of construction and validation of the signature and nomogram.

## MATERIALS AND METHODS

### Data Obtainment and Preliminary Collation

We downloaded the RNA expression profiles with corresponding clinical features from the TCGA-SARC and GTEx databases at the UCSC Xena website (https://xena.ucsc.edu/). The TCGA database includes the results of large-scale sequencing of 33 human tumors and the corresponding clinical information, which helps study the molecular mechanism of tumors. Sequencing data (RNA FPKM values) of 265 samples (including 263 sarcoma samples and 2 normal tissue samples) were obtained and transformed by $\log_2$(FPKM+1). The GTEx database includes RNA transcriptome data of 54 normal tissue samples from healthy individuals. We obtained the RNA sequencing data (FPKM value) of corresponding muscle and adipose tissues from the GTEx database and used them as a control for additional matches. Correspondingly, the RNA

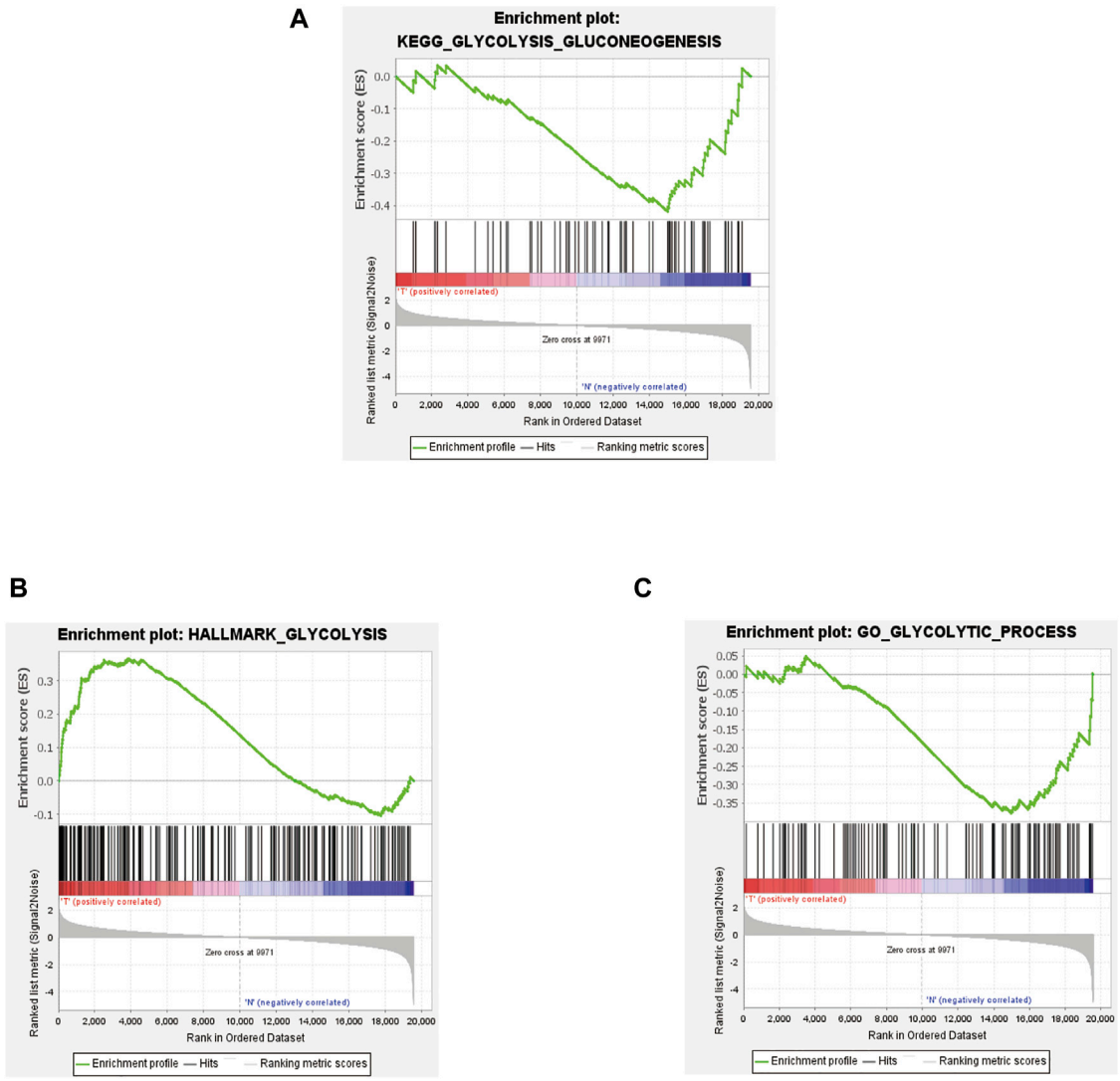


**FIGURE 2 |** Results of GSEA based on glycolysis-related gene sets. KEGG_GLYCOLYSIS_GLUCONEOGENESIS **(A)**, HALLMARK_GLYCOLYSIS **(B)**, GO_GLYCOLYTIC_PROCESS **(C)**.

sequencing data in GTEx were also converted by $\log_2$ (FPKM+1) for comparison with those in TCGA.

The gene expression data of GSE21050 were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). The GSE21050 gene sequencing data were based on the GPL570 platform, which contains 310 sarcoma samples. The samples of TCGA-SARC were determined as the training set, while the GSE21050 dataset was identified as an external validation.

## Identification of Differentially Expressed Genes Related to Glycolysis

First, we merged the transcriptome data in TCGA-SARC and GTEx, including a total of 913 normal samples and 263 tumor samples. In order to screen out glycolysis-related genes, we performed Gene Set Enrichment Analysis (GSEA; version 4.1) based on the gene set downloaded from Molecular Signatures Database v5.1: BIOCARTA_GLYCOLYSIS_PATHWAY, GO_GLYCOLYTIC_PROCESS, HALLMARK_GLYCOLYSIS, KEGG_GLYCOLYSIS_GLUCONEOGENESIS, and REACTOME_GLYCOLYSIS. Then, the 63 differentially expressed glycolysis-related genes (DEGRGs) were identified based on the criteria of $|\log_2 FC|>1$ and the adjusted $p$ value $< 0.05$.

## Identification of DEGRGs Related to Overall Survival and Establishment of Prognostic Signatures

To identify the prognostic DEGRGs, the 263 sarcoma samples were analyzed as the training set, and the matched TCGA-SARC clinical features were acquired from UCSC Xena (https://xenabrowser.net/). The following data analyses were completed in R software (version 3.6.2). Firstly, the "survival" package (version 3.2.7) in R software was used to perform univariate Cox regression to analyze the 63 DEGRGs

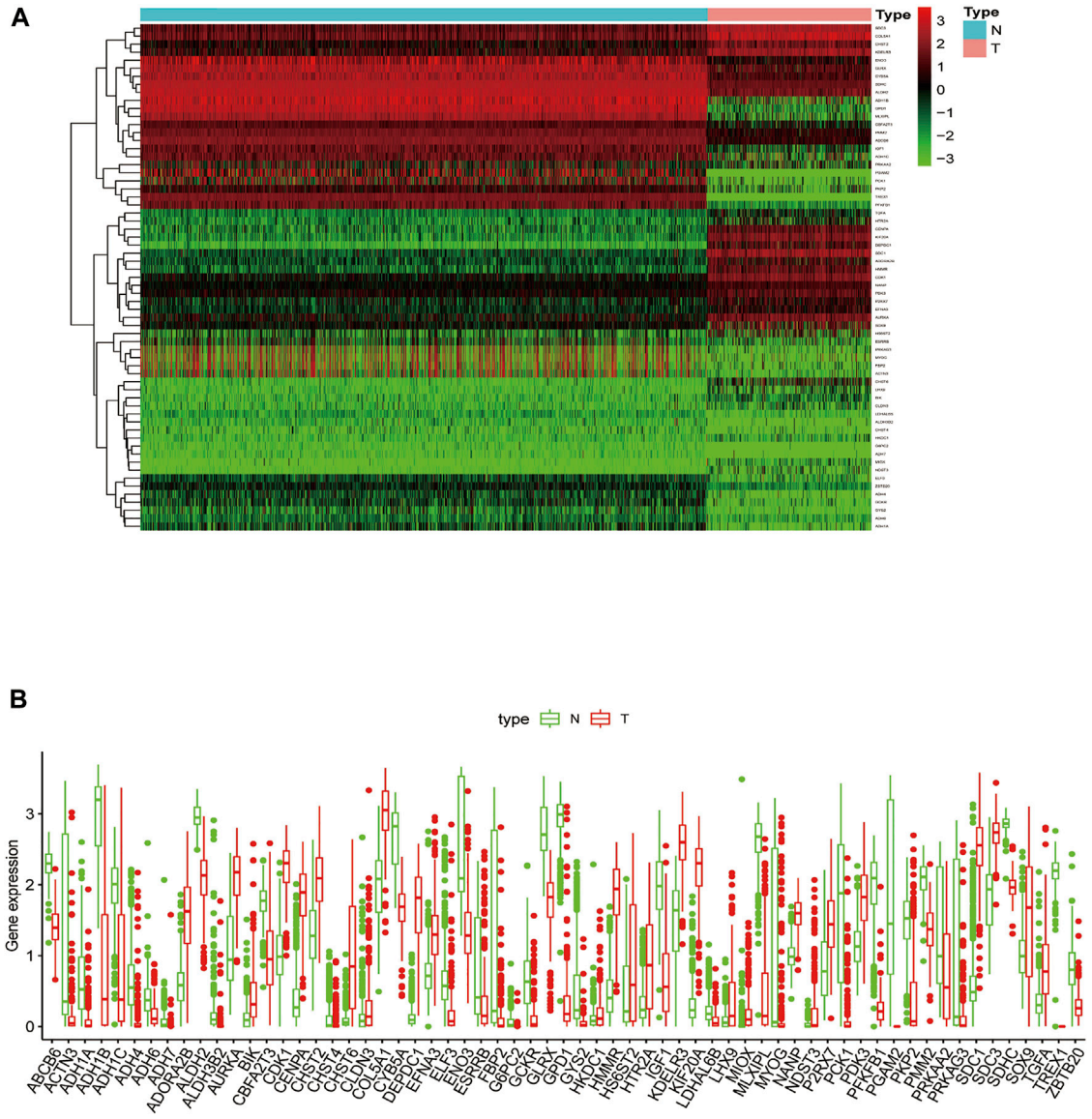


**FIGURE 3 |** Differently expressed glycolysis-related genes (DEGRGs) between sarcoma samples and normal tissues. The screening criteria was based on the $|\log_2 FC| \geq 1$ and $p<0.05$. **(A)** Heatmap of DEGRGs **(B)** Box plot of DEGRGs. FC, fold change. Color images are available online.

identified above. Then, DEGRGs with $p$ value < 0.05 were selected to perform the least absolute shrinkage and selection operator regression (LASSO) analysis based on the "glmnet" (version 4.0.2) and "survival" (version 3.2.7) packages. LASSO is a biased estimate for processing data with multicollinearity, which identifies an optimum lambda value. Finally, a 7 DEGRGs signature related to the prognosis of STS patients was developed, and the risk score of each patient with STS was generated and calculated as follows: Risk Score = $\sum_{i=0}^{n} \boldsymbol{\beta}_i^* \boldsymbol{G}_i$,where $\boldsymbol{\beta}_i$ is defined as the coefficient of gene $\boldsymbol{i}$ of the LASSO analysis; $\boldsymbol{G}_i$ presents the expression level of each gene. Based on this gene signature, STS patients in the training set were divided into high-risk and low-risk groups on account of the critical value (i.e., the median risk score). A total of 309 samples in GSE21050 (survival data missing in 1 sample) were identified as the test set. To assess the accuracy of results, we analyzed the data in the test set at the same level. To assess the availability of the signature, we conducted overall survival analysis to evaluate the overall survival differences in the high-risk and low-risk patients. Receiver operating characteristic (ROC) curves were obtained using the "survivalROC" package (version 1.0.2). The ROC curves at 3 and 5 years were also generated to evaluate the credibility and accuracy of the risk signature.

## Analysis of the Clinical Characteristics Associated With Prognosis

Univariate Cox regression and multivariate Cox regression analyses were conducted to evaluate the influences of clinical features on the prognosis of patients. We eliminated the samples

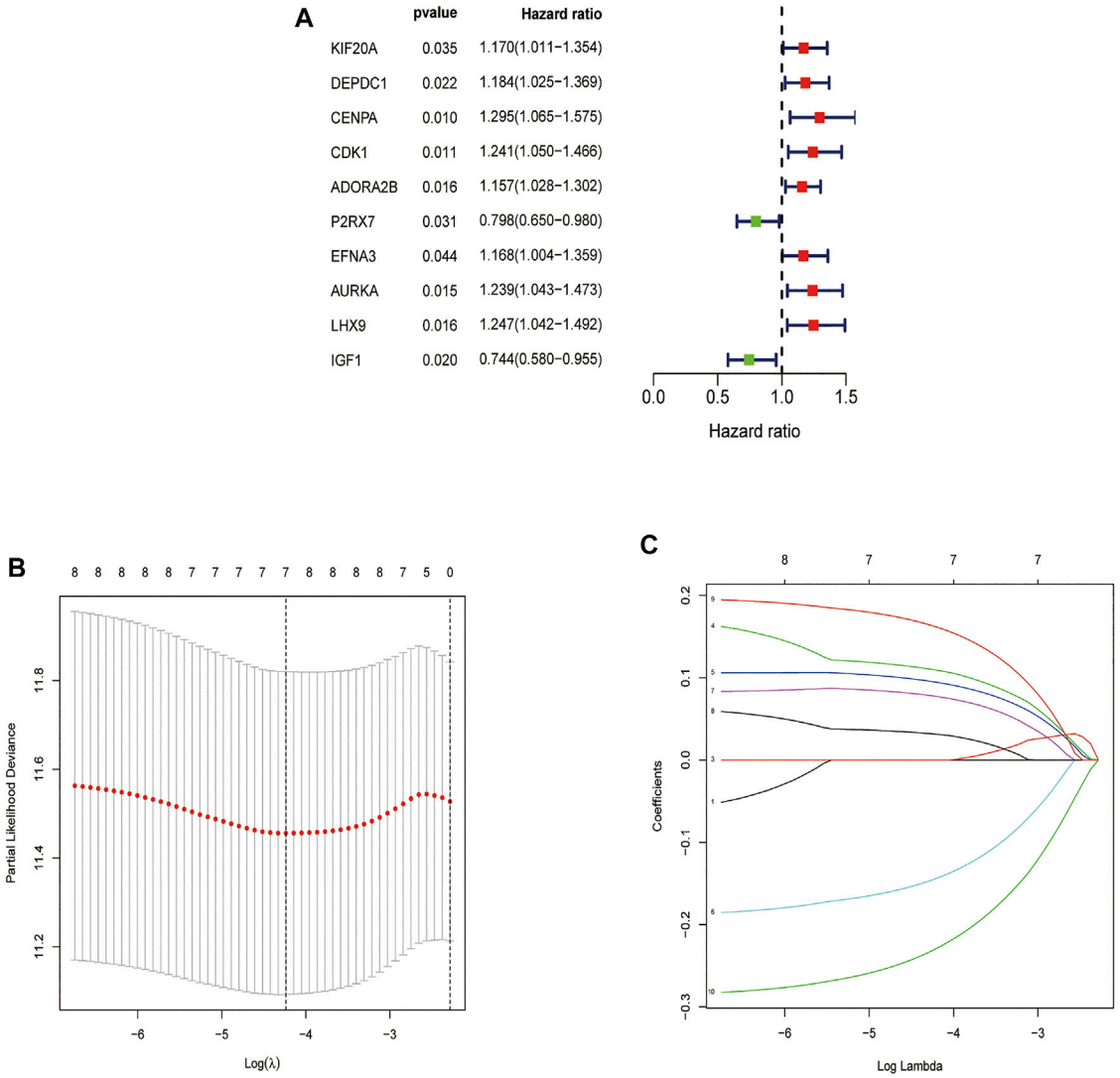


**FIGURE 4 |** Feature selection using the Univariate Cox analysis and Lasso regression model. **(A)** Forest map of DEGRGs associated with STS survival, univariate Cox regression, $p < 0.01$. **(B)** LASSO coefficient spectrum of 7 DEGRGs. **(C)** On account of 1000 cross-validation for tuning parameter selection via LASSO.

with incomplete clinical information in the training set, and those with complete information were selected for the next processing. The Cox regression analyses included the following factors: age, margin, metastasis, depth, ethnicity, gender, race, diagnoses, and the risk score identified above. Multiple ROC curves were constructed using the "survivalROC" package in R software.

## Establishment of Prognosis-Related Nomogram and Validation

We combined the risk signature with the factors identified above to build a nomogram for prognostic prediction of sarcoma patients in the training set, using the "rms" package (version 6.0.1) in R software. Meanwhile, the 1, 3, and 5 years calibration curves were created to assess the consistency between the realistic results and the results demonstrated by the nomogram in the training set.

# RESULTS

## Patient Characteristics and Transcriptome Expression Level

TCGA-SARC RNA sequence expression data and the corresponding clinical information were obtained from UCSC Xena. There were 265 sequence profiles in TCGA-SARC, with 263 tumor samples and 2 normal samples. The 263 STS samples were contained, including 105 leiomyosarcomas (LMS), 56 dedifferentiated liposarcomas (DL), 34 undifferentiated sarcomas (US), 25 myxofibrosarcomas (MF), 12 malignant fibrous histiocytomas (MFH), 10 malignant peripheral nerve sheath tumors (MPNST), 10 synovial sarcomas (SS), 3 myxoid leiomyosarcomas, 3 giant cell sarcomas, 2 pleomorphic liposarcomas, and 3 other sarcoma samples. The matched normal samples obtained from the GTEx database included 911 normal tissue samples (396 muscle and 515 adipose samples). Finally, we identified 263 tumor

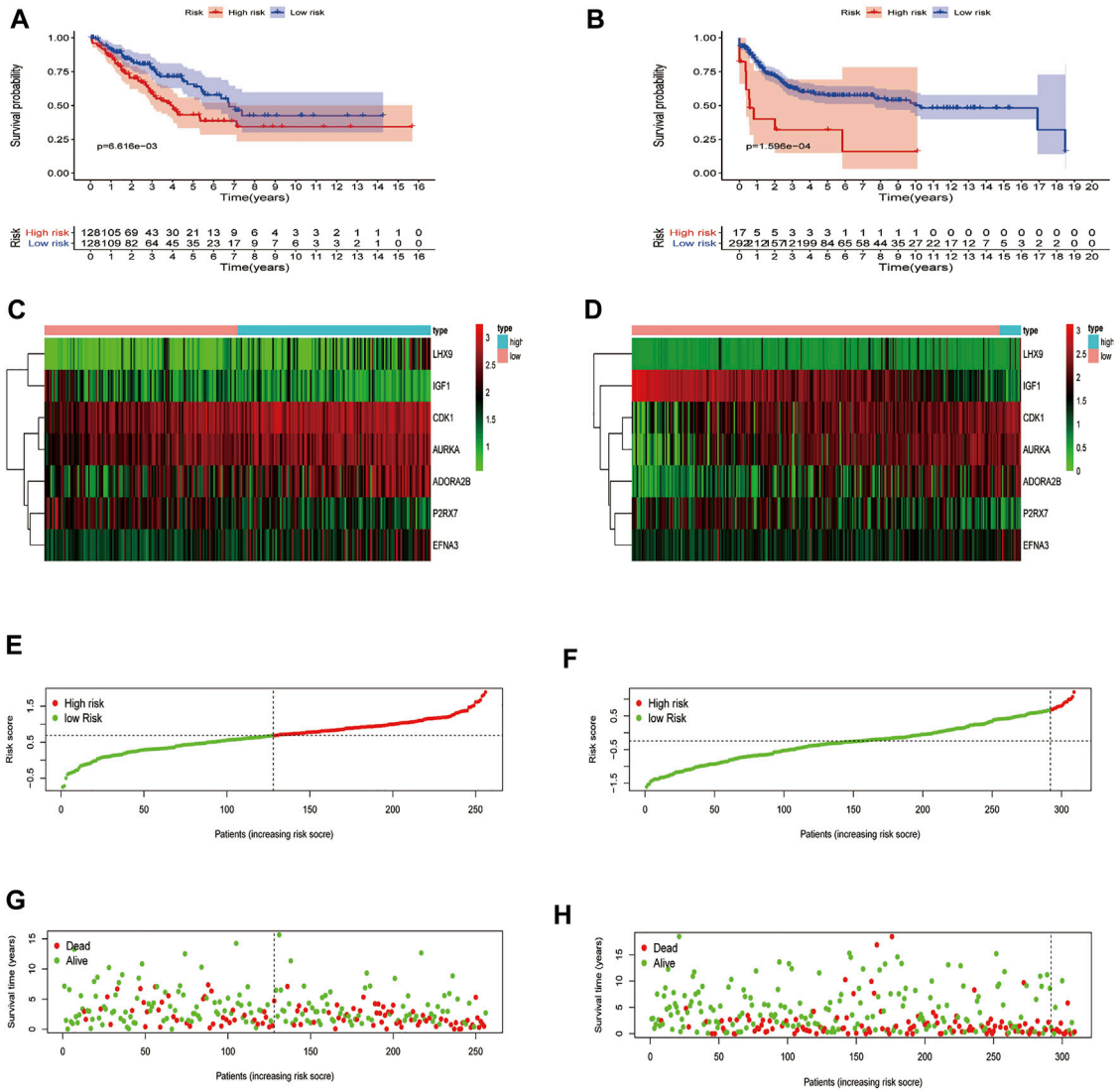


**FIGURE 5 |** Risk signature development. Survival analysis of the training **(A)**, test **(B)**. The upper part demonstrates the KM(Kaplan-Meier) curves for the high and low risk groups. The number shows the living patients with time in the two groups. Differences in gene expression level between two groups, risk distribution of per samples, and relationship between survival status and survival times, training set **(C,E,G)**, test set **(D,F,H)**. The dark line shows the cut-off point dividing the two groups.

samples and 913 healthy tissue samples. Eventually, the expression levels of 19,532mRNAs were identified.

## Screening of DEGRGs

Five gene sets related to glycolysis were downloaded from the Molecular Signatures Database v5.1, and a total of 326 gene expression data were generated. In order to analyze the difference levels of 5 glycolysis-related gene sets in STS and normal samples, we performed GSEA analysis (version 4.1). The analysis results of DEGRGs are shown in **Figure 2**.

Next, we obtained the expression levels of a total of 326 glycolysis-related genes from these 5 gene sets in the training set. Based on the criteria of $|\log_2 FC| \geq 1$ and adjusted $p < 0.05$, 63 DEGRGs were identified, including 34 down-regulated and 29 up-regulated genes. The heatmap (**Figure 3A**) and boxplot (**Figure 3B**) of these 63 DEGRGs were generated in R software. The glycolysis-related gene expression matrix are shown in the **Supplementary Materials**.

## Risk Signature Construction

In order to determine the overall survival-related DEGRGs, univariate Cox regression analysis was applied to analyze the above identified 63 DEGRGs in the training set, and 10 genes were selected (**Figure 4A**). Then, using the "glmnet" package, DEGRGs with $p < 0.01$ were further used for the LASSO regression analysis to establish a gene signature (**Figures 4B,C**). The risk score of every patient was acquired by multiplying the gene level (X) with the regression coefficient (⊠). Finally, 7 DEGRGs highly related to prognosis were utilized to build a prognostic-related model: risk score = ($X_{CDK1}$ * 0.1079) + ($X_{ADORA2B}$ * 0.0936) + (X

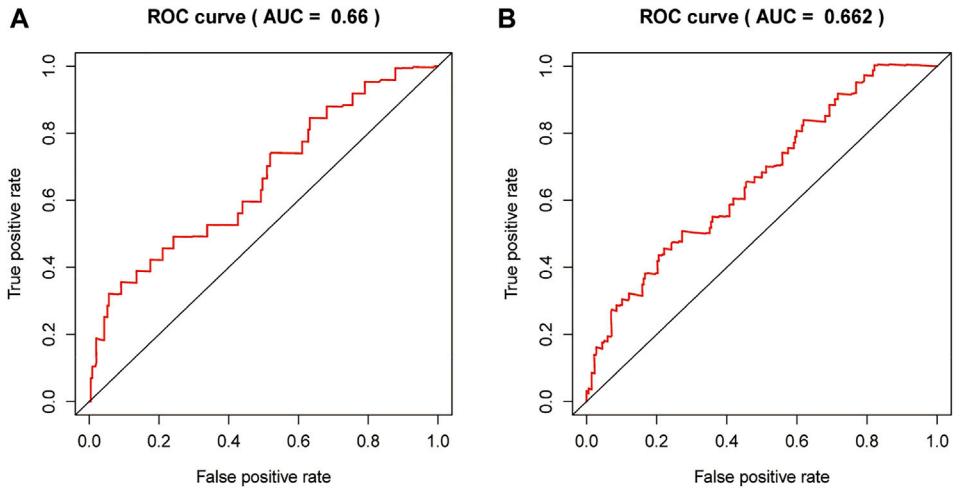


**FIGURE 6 |** Receiver operating characteristic (ROC) of 7 DEGRGs model in the training (I), test (J).

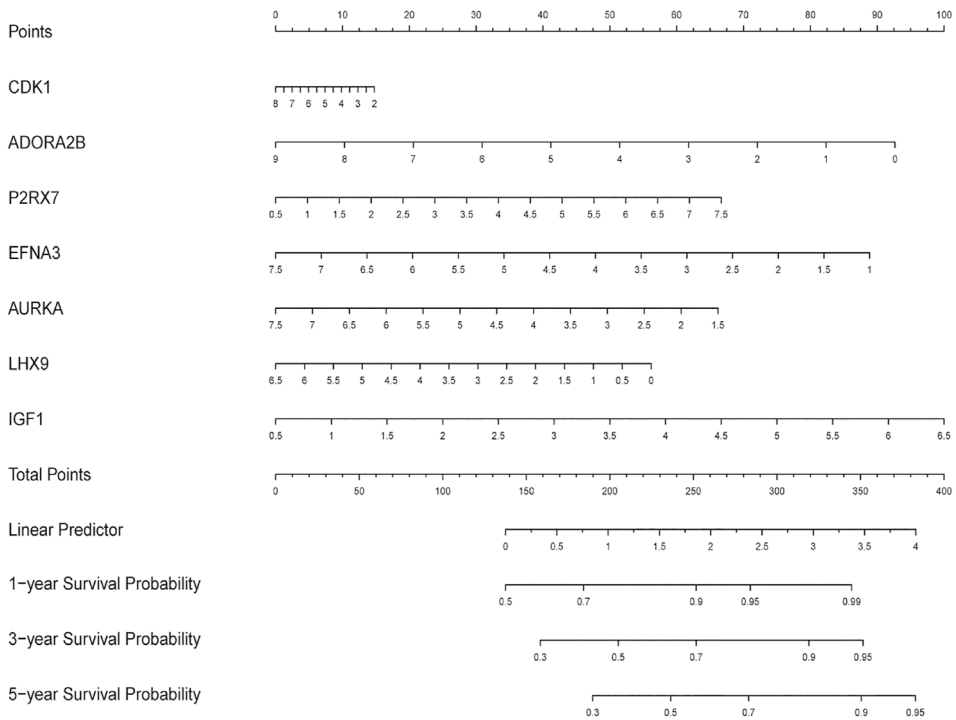


**FIGURE 7 |** The nomogram predicting the survival status of STS patients on the base of the expression level of 7 DEGRGs.

$_{P2RX7}$ * -0.1412) + (X $_{EFNA3}$ * 0.0764) + (X $_{AURKA}$ * 0.0311) + (X$_{LHX9}$ * 0.1598) + (X $_{IGF1}$ * −0.2258).

On account of the median risk score, the patients with STS in the training set were divided into high-risk and low-risk groups. Expression heatmaps, risk distribution plots, and survival status profiles of the 7 identified DEGRGs were constructed, and the survival difference between the two groups in training set (**Figures 5A,C,E,G**). Similar differences were also observed in the test group, which verified the prognostic model (**Figures 5B,D,F,H**). As shown in **Figures 6A,B**, the characteristics of the 7 DEGRGs can satisfactorily predict the survival status of STS patients, with AUC = 0.66 (test set: 0.662).

## Evaluation of the DEGRGs Signature

First, we constructed a nomogram according to the expressions of the 7 DEGRGs to predict the prognosis status of STS patients (**Figure 7**).

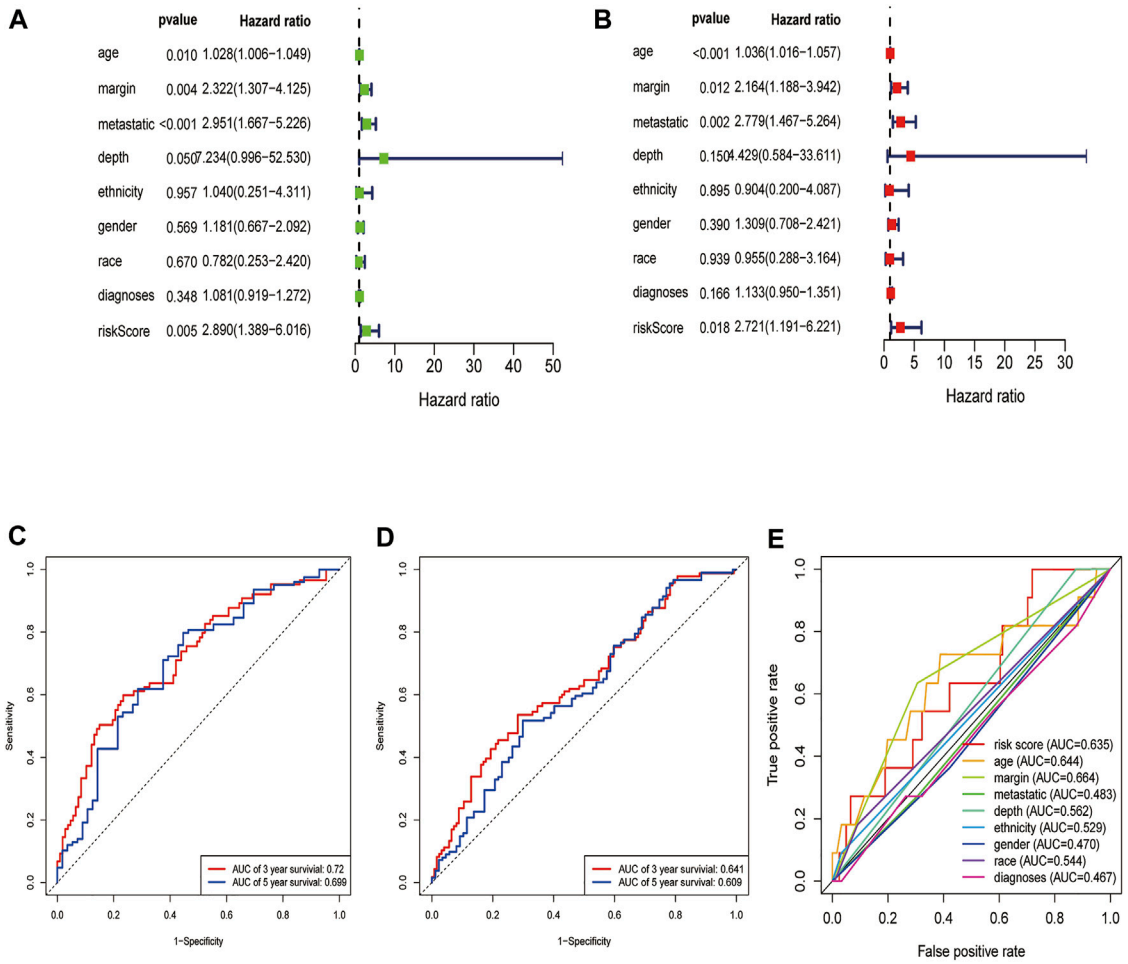


**FIGURE 8 |** Evaluation of the DEGRGs signature. The Result of univariate and multivariate Cox analyses **(A,B)**. The AUC curves to predict the survival status of STS patients at the 3- and 5-year survival time in train set **(C)**, test set **(D)**.The multi ROC curves of risk model and other clinical characteristic.

The results from univariate and multivariate independent prognostic analyses showed that the risk characteristics of the 7 DEGRGs were significantly ($p < 0.05$) related to the survival status of STS patients (**Figures 8A,B**). Analysis of multiple ROC curves showed that risk score signature had the largest AUC area (**Figure 8E**). The AUC size represents the prognostic efficiency of the 7-DEGRGs model. The larger the area, the better the predictive effect on patient's prognosis. In addition, based on the "timeROC" package (version 0.4) in R software, curves were plotted to evaluate the predictive value (**Figures 8C,D**). Our results showed that the 7 DEGRGs prognostic model could predicted both 3-year survival rate (AUC = 0.72) and 5-year survival rate (AUC = 0.699). These results demonstrated the excellent accuracy and sensitivity of the model.

## Construction and Evaluation of a Nomogram Incorporating the DEGRGs Signature With Clinical Factors

Based on multiple Cox regression, we constructed a prognostic nomogram to predict 1-year, 3-year, and 5-year survival possibility (**Figure 9**). Furthermore, calibration plots of the 1-year, 3-year, and 5-year survival prediction were used to assess the predictive ability of the nomogram, as shown in **Figures 10A–C**. The calibration curve showed that the nomogram had a high consistency between the actual and prediction results of survival state in the training set.

# DISCUSSION

Glycolysis-related genes have been revealed to play an important role in the occurrence and development of tumors (Yu et al., 2017; Yang et al., 2020). YAP1 affects the glycolytic metabolism of undifferentiated pleomorphic sarcoma through the NF-κB pathway (Rivera-Reyes et al., 2018). Lactate dehydrogenase inhibitors reduce the production of lactic acid and inhibit glycolysis, thereby inhibiting the proliferation of A673 sarcoma cells (Rai et al., 2017). Phosphoglycerate dehydrogenase (PHGDH) is highly expressed in Ewing's sarcoma and is associated with poor patient survival. PHGDH knockdown or

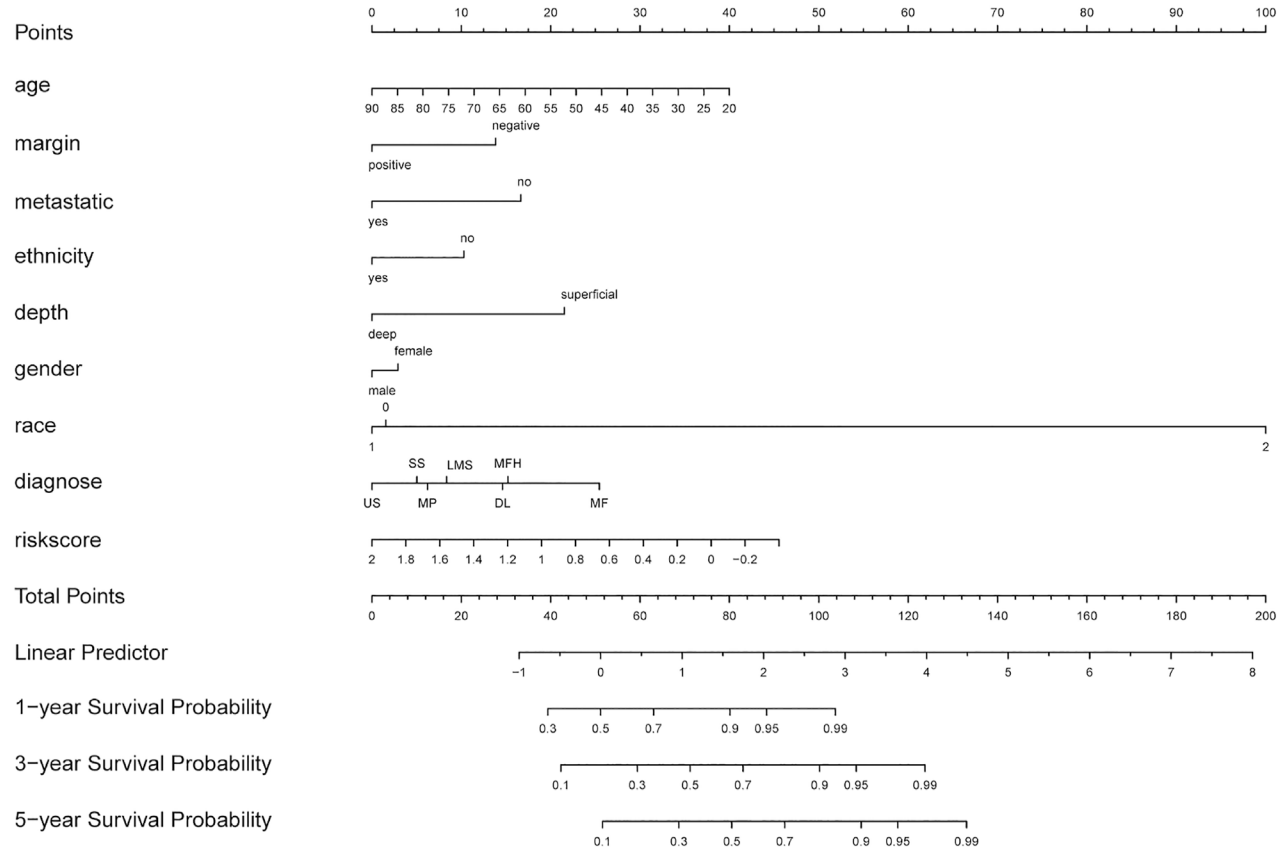


**FIGURE 9 |** Nomogram combing risk score and clinical factors for prediction of the individualized survival probability of STS patients. In the "diagnose" row, LMS (leiomyosarcomas), DL ( dedifferentiated liposarcomas), US( undifferentiated sarcomas), MF(25 myxofibrosarcomas), MFH(malignant fibrous histiocytomas), MPNST (malignant peripheral nerve sheath tumors), SS( synovial sarcomas).

*in vitro* pharmacological inhibition lead to decreased cell proliferation and cell death (Tanner et al., 2017). These studies indicate that glycolysis-related genes may play an important role in STS.

STS is a rare cancer, including more than 100 subtypes, with different pathological characteristics, molecular changes, and various prognosis of the patients. The efficient diagnosis and effective treatment of STS is difficult on account of the rarity and complex subtypes (Meyer and Seetharam, 2019). With the availability of more gene databases, novel analytic tools can be developed to explore biomarkers for rare tumors (e.g., STS) from existing data (van IJzendoorn et al., 2019). With the development of bioinformatics, a growing number of studies have proved that processing gene databases is an effective method to assess the transcriptome characteristics associated with prognosis, which can help identify new serum biomarkers for clinical diagnosis, prognosis prediction, as well as postoperative treatment (Ouyang et al., 2019).

Cancer cells usually have more vigorous metabolism than normal cells, which is characterized by aerobic glycolysis and anabolic cycles to support tumor metastasis and proliferation (Ancey et al., 2018; Orang et al., 2019). In recent years, an increased number of tumor-related studies have focused on investigating the glycolytic process (Abbaszadeh et al., 2020). According to these studies, multiple genes and pathways related to glycolysis have been discovered. Analogs and blockers of these genes have also been developed, involving a variety of molecules, chemical drugs, and nano-drugs (AkinsAkins et al., 2018). Several studies have also reported the mechanism of glycolysis in STS. For example, Duan et al. found that glycolysis inhibitor 2-deoxyglucose can induce alveolar rhabdomyosarcoma cell apoptosis by regulating the expression level of Noxa (Ramírez-Peinado et al., 2011).

We were committed to identifying potential glycolysis-related gene biomarkers for assessing the risk and prognosis of STS patients. In this study, we screened out 7 glycolysis-related genes and established a prognostic nomogram by combined the model with several clinical features, which was effective in predicting STS. Through the joint analysis of the TCGA and GTEx databases, 63 DEGRGs were identified, which might serve as potential biomarkers for STS. Univariate Cox regression analyses were conducted and filtered out 10 prognosis-related DEGRGs. Subsequently, the LASSO regression was performed to further analyze these 10 genes in the training set, based on which the 7 GEGRGs (CDK1, ADORA2B, P2RX7, EFNA3, AURKA, LHX9, andIGF1) model was finally established.

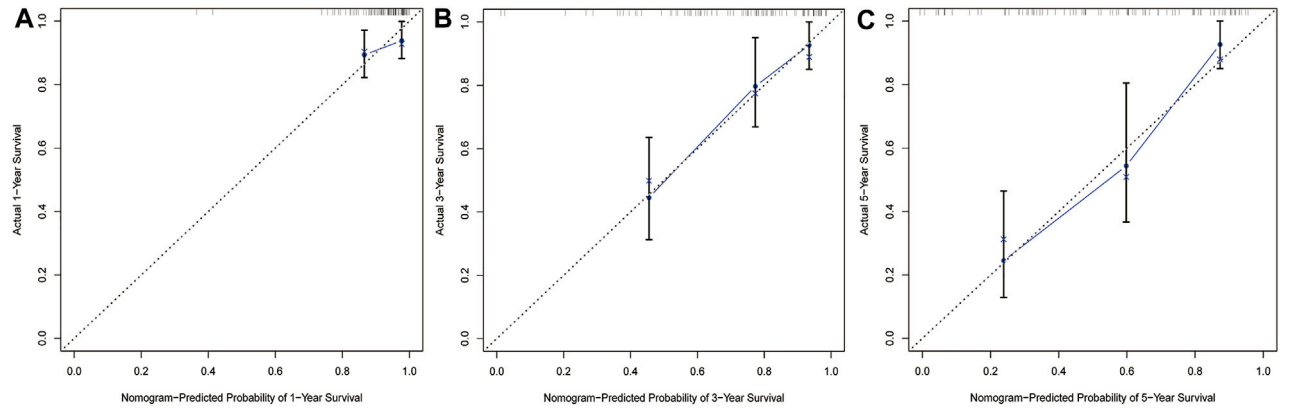


**FIGURE 10 |** Calibration curves of nomogram in train set. The curve shows the agreement between nomogram -predicted and observed 1**(A)**-,3**(B)**-and 5**(C)**-year survival outcomes of STS patient. The closer the blue line in the figure is to the 45-degree dotted line, the better the prediction performance of the nomogram.

The functions of the identified 7 DEGRGs have been previously reported. For instance, Menon et al. demonstrated that CDK1 was up-regulated in melanoma cells and interacted with Sox2 to promote the proliferation of melanoma (Ravindran Menon et al., 2018). Desmet et al. found that blocking ADORA2B inhibited the invasive activity of breast cancer cells and reduced their ability to metastasize (Wilkat et al., 2020). ADORA2B knockdown reduced tumor vascularization and thus inhibited the growth of head and neck squamous cell carcinomas (Desmet et al., 2013). Furthermore, Wang et al. proved that P2RX7 was overexpressed in gastric cancer tissues, promoting tumor proliferation through ERK1/2 pathway and Akt pathway, which was also correlated with poor prognosis (Lili et al., 2019). Wang et al. reported that cathelicidin inhibited colon cancer metastasis through a P2RX7-dependent pathway (Wang et al., 2020). EFNA3, as an Eph receptor ligand, affected the migration and proliferation of human umbilical cord endothelial cells through the PI3K/AKT pathway (Cheng et al., 2019). Chen et al. found that AURKA directly promoted the Warburg effect by phosphorylating lactate dehydrogenase B (LDHB), thereby promoting tumor growth (Li et al., 2019b). It has also been confirmed that the expression level of LHX9 was significantly up-regulated in osteosarcoma, and inhibiting LHX9 reduced the ability of cell growth and invasion (Li et al., 2019c). Li et al. proved that the levels of IGF-1 and IGF-1R in osteosarcoma were elevated, and their overexpression promoted the invasion and resistance of osteosarcoma cells (Yu et al., 2020).

Recently, with the development of bioinformatics tools, multiple glycolysis-related gene models have been developed to assess the survival status of cancer patients (Cai et al., 2020). To our knowledge, our study is the first to screen out DEGRGs by analyzing data from the public TCGA database to predict the survival status of STS patients. Moreover, based on these 7 DEGRGs, through combining the risk score and clinical characteristics, we constructed a nomogram to assess the prognosis of STS patients. We found that glycolysis-related genes and STS prognosis were closely correlated, which may provide us with a novel strategy for the treatment of STS.

This work has some limitations. First of all, the number of STS samples in TCGA-SARC data set was relatively small, and that of normal samples was insufficient though GTEx database was also involved. Second, several important clinical features (e.g., tumor stage) of the patients in the TCGA database were not sufficiently detailed, which may affect the treatment and prognosis of STS patients. Finally, more independent external queues need to be analyzed on the basis of our model to ensure the predictive performance of the nomogram.

## CONCLUSION

By using the high-throughput sequencing data in the TCGA database, we performed a variety of high-dimensional regression analyses (LASSO and Cox regression models) to identify the prognostic DEGRG markers for STS patients. The 7 gene prognostic signature is an effective predictor of STS. Through combining the 7 DEGRGs and clinical characteristics of STS patients, we established a prognostic nomogram that has superior efficacy in STS risk and patient survival prediction. The significant effectiveness of this model may be helpful for decision-making in clinical treatment, and further study is warranted to reveal the biological and molecular roles of these DEGRGs in STS.

## 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

YL designed the study. YL, CL, and HZ generated the figures and table. YL conducted data processing. YL and XY wrote the manuscript. AY supervised the research.

## FUNDING

This work was supported by the Hubei Province Scientific and Technical Innovation Key Project (No. 2019ACA136).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.675865/full#supplementary-material

## REFERENCES

Abbaszadeh, Z., Çeşmeli, S., and Biray Avcı, Ç. (2020). Crucial Players in Glycolysis: Cancer Progress. *Gene* 726, 144158. doi:10.1016/j.gene.2019.144158

Abdel-Wahab, A. F., Mahmoud, W., and Al-Harizy, R. M. (2019). Targeting Glucose Metabolism to Suppress Cancer Progression: Prospective of Anti-glycolytic Cancer Therapy. *Pharmacol. Res.* 150, 104511. doi:10.1016/j.phrs.2019.104511

AkinsAkins, N. S., Nielson, T. C., and Le, H. V. (2018). Inhibition of Glycolysis and Glutaminolysis: An Emerging Drug Discovery Approach to Combat Cancer. *Ctmc* 18 (6), 494–504. doi:10.2174/1568026618666180523111351

Ancey, P-B., Contat, C., and Meylan, E. (2018). Glucose Transporters Incancer - from Tumor Cells to the Tumor Microenvironment. *FEBS J.* 285 (16), 2926–2943. doi:10.1111/febs.14577

Brennan, M. F., R Antonescu, C., Moraco, N., and Singer (2014). Lessons Learned from the Study of 10,000 Patients with Soft Tissue Sarcoma. *Ann. Surg.* 260 (3), 416–421. doi:10.1097/SLA.0000000000000869

Cahlon, O., Brennan, M. F., Jia, X., Qin, L.-X., Singer, S., and Alektiar, K. M. (2012). A Postoperative Nomogram for Local Recurrence Risk in Extremity Soft Tissue Sarcomas after Limb-Sparing Surgery without Adjuvant Radiation. *Ann. Surg.* 255 (2), 343–347. doi:10.1097/SLA.0b013e3182367aa7

Cai, L., Hu, C., Yu, S., Liu, L., Yu, X., Chen, J., et al. (2020). Identification and Validation of a Six-Gene Signature Associated with Glycolysis to Predict the Prognosis of Patients with Cervical Cancer. *BMC Cancer* 20 (1), 1133. doi:10.1186/s12885-020-07598-3

Callegaro, D., Miceli, R., Bonvalot, S., Ferguson, P., Strauss, D. C., Levy, A., et al. (2016). Development and External Validation of Two Nomograms to Predict Overall Survival and Occurrence of Distant Metastases in Adults after Surgical Resection of Localised Soft-Tissue Sarcomas of the Extremities: a Retrospective Analysis. *Lancet Oncol.* 17 (5), 671–680. doi:10.1016/S1470-2045(16)00010-3

Cheng, A., Zhang, P., Wang, B., Yang, D., Duan, X., Jiang, Y., et al. (2019). Aurora-A Mediated Phosphorylation of LDHB Promotes Glycolysis and Tumor Progression by Relieving the Substrate-Inhibition Effect. *Nat. Commun.* 10 (1), 5566. doi:10.1038/s41467-019-13485-8

Choi, J. H., and Ro, J. Y. (2020). The 2020 WHO Classification of Tumors of Soft Tissue: Selected Changes and New Entities. *Adv. Anat. Pathol.* 28, 44–58. doi:10.1097/PAP.0000000000000284

Desmet, C. J., Gallenne, T., Prieur, A., Reyal, F., Visser, N. L., WittnerWittner, B. S., et al. (2013). Identification of a Pharmacologically Tractable Fra-1/ADORA2B axis Promoting Breast Cancer Metastasis. *Proc. Natl. Acad. Sci.* 110 (13), 5139–5144. doi:10.1073/pnas.1222085110

Gamboa, A. C., Gronchi, A., and Cardona, K. (2020). Soft-tissue Sarcoma in Adults: An Update on the Current State of Histiotype-specific Management in an Era of Personalized Medicine. *CA A. Cancer J. Clin.* 70 (3), 200–229. doi:10.3322/caac.21605

Ganapathy-Kanniappan, S., and GeschwindGeschwind, J.-F. H. (2013). Tumor Glycolysis as a Target for Cancer Therapy: Progress and Prospects. *Mol. Cancer* 12, 152. doi:10.1186/1476-4598-12-152

Gatenby, R. A., and Gillies, R. J. (2004). Why Do Cancers Have High Aerobic Glycolysis? *Nat. Rev. Cancer* 4 (11), 891–899. doi:10.1038/nrc1478

GTEx Consortium (2015). Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* 348 (6235), 648–660. doi:10.1126/science.1262110

Huangyang, P., Li, F., Lee, P., Nissim, I., WeljieWeljie, A. M., Mancuso, A., et al. (2020). Fructose-1,6-Bisphosphatase 2 Inhibits Sarcoma Progression by Restraining Mitochondrial Biogenesis. *Cel Metab.* 31 (1), 174–188. doi:10.1016/j.cmet.2019.10.012

Jia, D., Li, S., Li, D., Xue, H., Yang, D., and Liu, Y. (2018). Mining TCGA Database for Genes of Prognostic Value in Glioblastoma Microenvironment. *Aging* 10 (4), 592–605. doi:10.18632/aging.101415

Li, G., Xu, W., Zhang, L., Liu, T., Jin, G., Song, J., et al. (2019). Development and Validation of a CIMP-Associated Prognostic Model for Hepatocellular Carcinoma. *EBioMedicine* 47, 128–141. doi:10.1016/j.ebiom.2019.08.064

Li, S-Q., Tu, C., Lu, W., ChenDuan, R-Q., and Ren, Xiao-Lei. (2019). FGF-induced *LHX9* Regulates the Progression and Metastasis of Osteosarcoma via FRS2/TGF-B/β-Catenin Pathway. *Cell Div* 14, 13. doi:10.1186/s13008-019-0056-6

Li, Y-S., Liu, Q., He, H-B., and Luo, W. (2019). The Possible Role of Insulin-like Growth Factor-1 in Osteosarcoma. *Curr. Probl. Cancer* 43 (3), 228–235. doi:10.1016/j.currproblcancer.2018.08.008

Lili, W., Yun, L., Tingran, W., Xia, W., and Yanlei, S. (2019). P2RX7 Functions as a Putative Biomarker of Gastric Cancer and Contributes to Worse Prognosis. *Exp. Biol. Med. (Maywood)* 244 (9), 734–742. doi:10.1177/1535370219846492

Long, J., Wang, A., Bai, Y., Lin, J., Yang, X., Wang, D., et al. (2019). Development and Validation of a TP53-Associated Immune Prognostic Model for Hepatocellular Carcinoma. *EBioMedicine* 42, 363–374. doi:10.1016/j.ebiom.2019.03.022

Mao, L., Dauchy, R. T., Blask, D. E., Dauchy, E. M., Slakey, L. M., Brimer, S., et al. (2016). Melatonin Suppression of Aerobic Glycolysis (Warburg Effect), Survival Signalling and Metastasis in Human Leiomyosarcoma. *J. Pineal Res.* 60 (2), 167–177. doi:10.1111/jpi.12298

Mariani, L., Miceli, R., Kattan, M. W., Brennan, M. F., Colecchia, M., Fiore, M., et al. (2005). Validation and Adaptation of a Nomogram for Predicting the Survival of Patients with Extremity Soft Tissue Sarcoma Using a Three-Grade System. *Cancer* 103 (2), 402–408. doi:10.1002/cncr.20778

Meyer, M., and Seetharam, M. (2019). First-Line Therapy for Metastatic Soft Tissue Sarcoma. *Curr. Treat. Options. Oncol.* 20 (1), 6. doi:10.1007/s11864-019-0606-9

Orang, A. V., Petersen, J., McKinnon, R. A., and Michael, M. Z. (2019). Micromanaging Aerobic Respiration and Glycolysis in Cancer Cells. *Mol. Metab.* 23, 98–126. doi:10.1016/j.molmet.2019.01.014

Ouyang, D., Li, R., Li, Y., and Zhu, X. (2019). A 7-lncRNA Signature Predict Prognosis of Uterine Corpus Endometrial Carcinoma. *J. Cel Biochem* 120 (10), 18465–18477. doi:10.1002/jcb.29164

Rai, G., Brimacombe, K. R., MottMott, B. T. D. J. U., Hu, D. J. X., Lee, S.-M. T. D., CheffLee, D. M., et al. (2017). Discovery and Optimization of Potent, Cell-Active Pyrazole-Based Inhibitors of Lactate Dehydrogenase (LDH). *J. Med. Chem.* 60 (22), 9184–9204. doi:10.1021/acs.jmedchem.7b00941

Ramírez-Peinado, S., Alcázar-Limones, F., Lagares-Tena, L., El Mjiyad, N., Caro-Maldonado, A., TiradoTirado, O. M., et al. (2011). 2-deoxyglucose Induces Noxa-dependent Apoptosis in Alveolar Rhabdomyosarcoma. *Cancer Res.* 71 (21), 6796–6806. doi:10.1158/0008-5472.CAN-11-0759

Ravindran Menon, D., Luo, Y., Liu, J. J. S., KrishnanKutty, L. N., Li, D. G. Y., Samson, J. M., et al. (2018). CDK1 Interacts with Sox2 and Promotes Tumor Initiation in Human Melanoma. *Cancer Res.* 78 (23), 6561–6574. doi:10.1158/0008-5472.CAN-18-0330

Rivera-Reyes, A., Ye, S., E. Marino, G. S., E. Ciotti, G. S., E Ciotti, Y., Posimo, J. M. P. M. C., et al. (2018). YAP1 Enhances NF-κb-dependent and Independent Effects on Clock-Mediated Unfolded Protein Responses and Autophagy in Sarcoma. *Cell Death Dis* 9 (11), 1108. doi:10.1038/s41419-018-1142-4

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019. *CA A. Cancer J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Tanner, J. M., Bensard, C., Wei, P., Krah, N. M., Schell, J. C., Gardiner, J., et al. (2017). EWS/FLI Is a Master Regulator of Metabolic Reprogramming in Ewing Sarcoma. *Mol. Cancer Res.* 15 (11), 1517–1530. doi:10.1158/1541-7786.MCR-17-0182

van IJzendoorn, D. G. P., Szuhai, K., Briaire-de BruijnBriaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., and Bovée, J. V. M. G. (2019). Machine Learning Analysis of Gene Expression Data Reveals Novel Diagnostic and Prognostic

Biomarkers and Identifies Therapeutic Targets for Soft Tissue Sarcomas. *Plos Comput. Biol.* 15 (2), e1006826. doi:10.1371/journal.pcbi.1006826

Wang, Hui., Wang, Lin., Zhou, X., Luo, X., Liu, K., Jiang, E., et al. (2020). OSCC Exosomes Regulate miR-210-3p Targeting EFNA3 to Promote Oral Cancer Angiogenesis through the PI3K/AKT Pathway. *Biomed. Res. Int.*, 20202125656. doi:10.1155/2020/2125656

Wilkat, M., Bast, H., Drees, R., Dünser, J., Mahr, A., Azoitei, N., et al. (2020). Adenosine Receptor 2B Activity Promotes Autonomous Growth, Migration as Well as Vascularization of Head and Neck Squamous Cell Carcinoma Cells. *Int. J. Cancer* 147 (1), 202–217. doi:10.1002/ijc.32835

Wu, S.-X., Huang, J., Liu, Z.-W., Chen, H.-G., Guo, P., Cai, Q.-Q., et al. (2018). A Genomic-Clinicopathologic Nomogram for the Preoperative Prediction of Lymph Node Metastasis in Bladder Cancer. *EBioMedicine* 31, 54–65. doi:10.1016/j.ebiom.2018.03.034

Yang, J., Ren, B., Yang, G., Wang, H., Chen, G., You, L., et al. (2020). The Enhancement of Glycolysis Regulates Pancreatic Cancer Metastasis. *Cell. Mol. Life Sci.* 77 (2), 305–321. doi:10.1007/s00018-019-03278-z

Yu, L., Lu, M., Jia, D., Ma, J., Ben-Jacob, E., Levine, H., et al. (2017). Modeling the Genetic Regulation of Cancer Metabolism: Interplay between Glycolysis and Oxidative Phosphorylation. *Cancer Res.* 77 (7), 1564–1574. doi:10.1158/0008-5472.CAN-16-2074

Yu, S., Hu, C., Cai, L., Du, X., Fan, L., Yu, Q., et al. (2020). Seven-Gene Signature Based on Glycolysis Is Closely Related to the Prognosis and Tumor Immune Infiltration of Patients with Gastric Cancer. *Front. Oncol.* 10, 1778. doi:10.3389/fonc.2020.01778

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

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

# Advantages of publishing in Frontiers

**OPEN ACCESS**
Articles are free to read for greatest visibility and readership

**FAST PUBLICATION**
Around 90 days from submission to decision

**HIGH QUALITY PEER-REVIEW**
Rigorous, collaborative, and constructive peer-review

**TRANSPARENT PEER-REVIEW**
Editors and reviewers acknowledged by name on published articles

**Frontiers**
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

**Visit us:** www.frontiersin.org
**Contact us:** frontiersin.org/about/contact

**REPRODUCIBILITY OF RESEARCH**
Support open data and methods to enhance research reproducibility

**DIGITAL PUBLISHING**
Articles designed for optimal readership across devices

**FOLLOW US**
@frontiersin

**IMPACT METRICS**
Advanced article metrics track visibility across digital media

**EXTENSIVE PROMOTION**
Marketing and promotion of impactful research

**LOOP RESEARCH NETWORK**
Our network increases your article's readership