BIOMARKER DETECTION ALGORITHMS AND TOOLS FOR MEDICAL IMAGING OR OMIC DATA

EDITED BY: Fengfeng Zhou, William C. Cho, Lin Hua, Jie Li and Feng Liu

PUBLISHED IN: Frontiers in Genetics and Frontiers in Molecular Biosciences

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88976-570-6
DOI 10.3389/978-2-88976-570-6

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

BIOMARKER DETECTION ALGORITHMS AND TOOLS FOR MEDICAL IMAGING OR OMIC DATA

Topic Editors:
Fengfeng Zhou, Jilin University, China
William C. Cho, QEH, Kowloon, SAR China
Lin Hua, Capital Medical University, China
Jie Li, Harbin Institute of Technology, China
Feng Liu, Wuhan University, China

Citation: Zhou, F., Cho, W. C., Hua, L., Li, J., Liu, F., eds. (2022). Biomarker Detection Algorithms and Tools for Medical Imaging or Omic Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-570-6

Table of Contents

frontiers | Frontiers in Genetics

EDITORIAL
published: 25 May 2022
doi: 10.3389/fgene.2022.919390

Editorial: Biomarker Detection Algorithms and Tools for Medical Imaging or Omics Data

William C. Cho[1]*, *Fengfeng Zhou*[2], *Jie Li*[3], *Lin Hua*[4] and *Feng Liu*[5]

[1]Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong SAR, China, [2]College of Computer Science and Technology, Jilin University, Changchun, China, [3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [4]Institute of Biomedical Engineering, Capital Medical University, Beijing, China, [5]School of Mechanic and Power Engineering, Wuhan University, Wuhan, China

Keywords: biomarker, detection algorithm, medical imaging, omics, artificial intelligence

Editorial on the Research Topic

Biomarker Detection Algorithms and Tools for Medical Imaging or Omics Data

Biomarkers are characteristics that can be objectively detected and assessed and can be used as indicators of normal biological processes, pathological processes, or pharmacological responses to therapeutic interventions. In the clinical aspect, biomarkers play a crucial role in the early diagnosis and classification of diseases, the judgment of disease degree, test of treatment effect, and prevention of disease. Therefore, some biomarker detection algorithms based on statistical models and artificial intelligence models have been constructed. However, there are still many issues in the existing algorithms, especially the high-performance algorithms to detect biomarkers of complex disease, such as cancer.

Traditional biomarker detection methods based on manual experimental methods are compl ex, inefficient, and costly. With the wide application of sequencing technology and digital imaging technology in biomarker detection, digital multi-omics data and medical images can be obtained rapidly and massively, providing the possibility for systematically detecting the characterization of disease, pathological causes, and data basis for algorithm-based automated biomarker detection. It is particularly important to combine multi-omics data with medical imaging, design algorithms that can efficiently identify biomarkers, discover more valuable biomarkers, and through the systematic combination of these new technologies and traditional biotechnology systems, ultimately provide a research basis for researchers, and doctors. However, how the construction of novel biomarker detection algorithms and identification of high-performance and robust biomarkers are still challenging problems.

In order to further promote the development of biomarker detection algorithms and develop more innovative algorithms, we proposed this Research Topic, which provided a platform for collecting recent discoveries in new feature extraction and feature selection algorithms for machine learning and deep learning models based on medical imaging and/or omics (genome, transcriptome, epigenome, proteome, and metabolome) data.

In structural biology and computer science, the image processing step is to traditionally cluster 2D cryo-electron microscopy (cryo-EM) images according to projection angle. Lei and Yang designed a new model, cascade of denoising autoencoders (CDAE), which was an efficient cryo-EM image denoising model. The model consisted of stacked deep neural network blocks that progressively reduced noise. When comparing state-of-the-art image denoising methods with significantly enhanced clustering results, they achieved a very competitive peak signal-to-noise ratio. Furthermore, the quantification and visualization of CDAE showed good noise reduction performance in clustered single-particle cryo-EM images.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
William C. Cho
chocs@ha.org.hk

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

Received: *13 April 2022*
Accepted: *18 April 2022*
Published: *25 May 2022*

Citation:
Cho WC, Zhou F, Li J, Hua L and Liu F (2022) Editorial: Biomarker Detection Algorithms and Tools for Medical Imaging or Omics Data. Front. Genet. 13:919390. doi: 10.3389/fgene.2022.919390

Conventional computed tomography (CT) is an important imaging technique for establishing disease diagnosis. Huang and Lu provided a case report using CT findings and histopathological features of primary liver carcinosarcoma (PLC). CT scans and three-stage enhanced scans were performed on the patients. Pathological features were analyzed. They concluded that the CT features observed in this study were very beneficial for the diagnosis of PLCs.

In recent years, exploring the diagnostic value of CT imaging and radiomics features in diseases has become a hotspot (Feng et al., 2022). For the classification of lung adenocarcinomas presenting as ground-glass nodules (GGNs) on CT, Zheng et al. studied 312 GGNs. Univariate and multivariate logistic regression was used to establish clinical models, minimum redundancy maximum relevance, least absolute shrinkage, and selection operator (LASSO) algorithms were used to select radiomics features, and construct radiomics models. A combined nomogram was developed based on the combined model and evaluated using its calibration curves and concordance indices. They found that the area under the curve (AUC) value was higher in both models compared to the individual clinical or radiomic models. They claimed that the nomogram served as a non-invasive and accurate predictive tool to help judge the aggressiveness of GGN before surgery and to help clinicians develop personalized treatment strategies.

It is well known that ultrasonography is an important step in ultrasound-guided diagnosis and treatment, but it is difficult to develop an ideal segmentation method due to strong imaging artifacts. Wu et al. proposed a novel boundary-guided multi-scale network to improve the performance of breast lesion segmentation in ultrasound images based on a feature pyramid network (FPN). First, they developed a boundary-guided feature enhancement module to enhance the feature maps of each FPN layer by learning the boundary maps of breast lesion regions. They then devised a multi-scale scheme to exploit information from different image scales to deal with ultrasound artifacts. The segmentation results were then generated by fusing the fine and coarse segmentation maps to accurately segment the breast lesion area from the ultrasound image and effectively remove the false detections due to boundary feature enhancement and multi-scale image information. Finally, they found that their results outperformed state-of-the-art methods.

Chronic rhinosinusitis with nasal polyps (CRSwNP) is a complex multifactorial disease with significant public health concerns, but its pathogenesis is still unclear. Noncoding RNAs have been reported to be promising biomarkers for various diseases. Among them, circular RNAs (circRNAs) are associated with inflammatory diseases. Therefore, Yu et al. studied the expression of circRNAs and microRNAs (miRNAs) in the CRSwNP group and control group. The biological functions of predicted abnormally expressed circRNAs and miRNAs were verified by qRT-PCR using Gene Ontology enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway analysis. Differentially expressed circRNAs and miRNAs between CRSwNP and controls were found. Among them, the altered expressions of hsa-circ-0031593 and hsa-miR-145-5p are the strongest evidence for involvement in the occurrence and development of CRSwNP, as their AUCs were higher than other molecules tested in this study.

In diabetic patients with and without ischemic stroke, Abdelaleem et al. found high expression levels of LINC00657 and miR-9 in serum and significantly lower serum miR-106a in the diabetic patients without stroke compared to the control participants. They claimed that serum noncoding RNAs (TUG1, LINC00657, miR-9, and miR-106a) might serve as potential novel biomarkers for stroke in diabetes. Their research may reveal new therapeutic targets for treating diabetic patients with stroke.

Multi-omics data are often measured to enrich the understanding of the biological mechanisms of certain phenotypes. However, due to the complex relationships and high dimensionality of multi-omics data, it is difficult to relate omics features to certain biological features of interest. Below are some diseases that use multi-omics data/methods for biomarker discovery.

Hepatocellular carcinoma (HCC), the third leading cause of cancer-related death worldwide, is a heterogeneous tumor with a complex tumor microenvironment (TME). TME refers to the microenvironment formed by immune cells and their products in tumor tissues (Fu et al., 2019). Bai et al. constructed a novel risk scoring model with prognostic value to elucidate the tumor immune microenvironment of HCC. ESTIMATE algorithm, single-sample gene set enrichment analysis (GSEA), and CIBERSORT analysis were used to reveal the characteristics of the HCC tumor immune microenvironment. After multiple analyses, four glycolysis-related long noncoding RNAs (lncRNAs) were obtained. The risk scores constructed with the four lncRNAs were found to be significantly associated with the prognosis of the patients. Besides, the risk scores were significantly correlated with immune scores, immune-related features, infiltrating immune cells (such as B cells), and key immune checkpoint blockade (ICB) molecules (such as CTLA4). Furthermore, they showed that MIR4435-2HG had a significant effect on the overall survival of the samples and was strongly associated with ICB treatment in HCC patients.

On the other hand, increasing evidence suggests that the abnormal expression of autophagy-related genes (ARGs) plays an important role in the occurrence and development of HCC. Luo et al. studied the ARGs in HCC. They constructed ARG pairs using ARGs extracted from the Human Autophagy Database and Molecular Signatures Database. They then developed a prognostic model based on ARG pairs, using LASSO Cox regression to assess the prognosis of patients after hepatectomy. Finally, they combined the signatures with independent prognostic factors to construct a nomogram. Based on ARG pair signatures, they could classify patients into high- or low-risk groups. Survival analysis and receiver operating characteristic (ROC) curve analysis verified the validity of the signature (AUC: 0.786–0.828). This model has a more accurate predictive effect than most HCC prognostic models. Their study provides evidence for the importance of autophagy in the occurrence and development of HCC, as well as a potential biomarker for targeted therapy.

For the poor prognosis of HCC, the development of prognostic prediction models is of great significance. Zhang et al. have

identified seven gene signatures associated with pyroptosis (BAK1, CHMP4B, GSDMC, NLRP6, NOD2, PLCG1, and SCAF11) to predict the prognosis of HCC patients. They constructed a novel LASSO Cox regression pyroptosis-related gene signature that could predict the prognosis of HCC patients. GSEA analysis further revealed novel signature-related mechanisms of immune responses in high-risk populations. Furthermore, they found that the expression of immune checkpoints was enhanced in the high-risk group, while m6A-related modifications were differentially expressed between the low- and high-risk groups.

In addition to autophagy and pyroptosis, recent studies have identified ferroptosis as a programmed cell death involved in regulating tumor biological behavior. Song et al. investigated the association between ferroptosis-related gene (FRG) expression profile and prognosis in esophageal squamous cell carcinoma (ESCC) patients based on The Cancer Genome Atlas and Gene Expression Omnibus (GEO). They developed a novel signature of FRGs, including ALOX12, ALOX12B, ANGPTL7, DRD4, MAPK9, SLC38A1, and ZNF419. A prognostic nomogram was then constructed combining clinical factors and risk scores. Their study demonstrates that ferroptosis-related features are a factor independently predicting ESCC risk and their prognostic risk models can predict ESCC prognosis.

Breast cancer subtypes are well-defined at the molecular level but difficult to classify using gene expression data. Jung et al. proposed a multi-omics analysis method, called multi-omics non-negative tensor decomposition for integrative analysis (MONTI), which aimed to select multi-omics features that could represent trait-specific features. They formed a three-dimensional tensor from the multi-omics data. They found that MONTI could well explain certain clinical attributes using multi-omics data. Furthermore, MONTI could detect subtype-specific genomes that were strongly regulated by certain omics, from which correlations between omics types could be inferred.

Various technological revolutions have occurred in recent years. Molecular assays based on transcriptome data are developing rapidly. Clinically, distinguishing benign from malignant thyroid nodules is challenging. Yang and Gong combined five independent transcriptomic studies to discover genetic signatures between benign and malignant thyroid nodules. Hundreds of differentially expressed genes were discovered by feature selection methods and weighted gene co-expression network analysis was performed to identify the modules of highly co-expressed genes. Ultimately, they identified four key genes (ST3GAL5, NRCAM, MT1F, and PROS1) involved in the pathogenesis of malignant thyroid.

Single-cell RNA sequencing (scRNA-seq) is emerging as one of the most powerful tools for uncovering disease complexity. scRNA-seq performs high-throughput sequencing analysis of epigenetics, transcriptomes, and genomes at the single-cell level, with the advantages of high-throughput and high resolution. The revelation of new cell subsets can focus disease initiation and progression on specific biological activities of specific cells. Regarding the complexity of the retina, Ying et al. reviewed the novel retinal cell subtypes and some specific gene markers discovered by scRNA-seq. Since the batch effects in scRNA-seq data are known to remain a hindrance when integrating disparate datasets, Zou et al. proposed a new deep learning-based method, deep mutual nearest neighbor (deepMNN), to correct for batch effects in scRNA-seq data. They searched for MNN pairs across different batches in a principal component analysis subspace. A batch correction network was then constructed by stacking the two residual blocks and further applied to remove batch effects. They demonstrated that deepMNN achieved better or comparable performance in qualitative analysis using uniform manifold approximation and projection plots and quantitative metrics (such as batch and cell entropy). Furthermore, deepMNN ran much faster than other methods for large-scale datasets. With these properties, deepMNN may be well suited for large-scale single-cell gene expression data analysis.

Absorption contrast between the terahertz (THz) frequency range of adipose and cancerous tissue allows the diagnosis of cancer by THz imaging. Even without external comparisons, Chen et al. have successfully demonstrated the ability of THz imaging to measure the volume of small breast cancers in a subcutaneous xenograft mouse model. They estimated the volumetric detection limit of a fiber-based THz scanning imaging system using a highly sensitive cryogenically operated Schottky diode detector to be less than 1 mm^3, thus showing the potential application of this technique in early cancer diagnosis.

Pulmonary hypertension (PH) affects the normal function of human pulmonary arteries. Peripheral blood mononuclear cells are an ideal source for minimally invasive disease diagnosis. Liu et al. proposed an ensemble feature selection algorithm (EnRank) to integrate the ranking information of popular feature selection algorithms, including T-test, chi-squared test, ridge regression, and LASSO. Using PH patient data, the biomarkers detected by EnRank provided useful information from these four feature selection algorithms and achieved very good predictive accuracy in predicting PH patients.

Epilepsy is a complex chronic neurological disorder that affects the health of approximately 70 million patients worldwide. About one-third of people with epilepsy develop drug resistance. Han et al. performed bioinformatic analysis to explore potential diagnostic markers and the mechanisms of drug-resistant epilepsy. Weighted correlation network analysis was applied to genes in epilepsy patients downloaded from the GEO database to identify key modules. Genes resistant to carbamazepine, phenytoin, and valproate were screened using LASSO regression and support vector machine (SVM) recursive feature elimination algorithms. Finally, ingenuity pathway analysis (IPA) was used for disease and functional pathway and network analysis. They found that the joint analysis yielded 17 resistance genes to construct a three-class classification SVM model. ROC analysis showed that the model could accurately predict patient resistance. Protein-protein interaction (PPI) revealed that six resistance genes (CD247, CTSW, IL2RB, MATK, NKG7, and PRF1) might play a central role in drug resistance in epilepsy patients. Finally, IPA revealed that resistance genes (PRKCH and S1PR5) were involved in CREB signaling in neurons.

PPI networks are critical for predicting essential proteins. The fusion of multiple biological information can reduce the impact of false data in PPI, but inevitably generates more noisy data. Zhang et al. proposed a new non-negative matrix tri-factorization (NMTF)-based model to predict essential proteins. A weighted PPI network was first built using the topological features of the network. The NMTF technique was then performed to reconstruct the optimized PPI network with more potential PPIs. A final ranking score for each protein was calculated using the PageRank algorithm, where the protein's subcellular localization and homology information were used to calculate the initial score. This study demonstrates that introducing NMTF can effectively improve the condition of PPI network and reduce the impact of noise on predictions.

The sparse canonical correlation analysis (SCCA) model is a well-known tool for identifying meaningful biomarkers in imaging genetics. However, most SCCA models contain only diagnostic status information, which poses challenges in finding disease-specific biomarkers. To overcome this obstacle, Ke et al. proposed a multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model to reveal disease-specific associations between single nucleotide polymorphisms and quantitative traits derived from multimodal neuroimaging data in the Alzheimer's disease (AD) Neuroimaging Initiative cohort. MT-SCCAR used complementary information carried by multi-perspective cognitive scores and encouraged the population sparsity of genetic variation. This study used MT-SCCAR to identify major genetic risk factors for AD, including rs429358. They found some patterns of association between genetic variants and brain regions.

Deciphering the effects of epigenetic alterations on regulatory elements requires innovative computational approaches that can benefit from massive epigenomic datasets, such as roadmaps or blueprints. Wang et al. developed a software named Integrative Ranking of Epigenetic Network of Enhancers to enable quantitative analyses of differential epigenetic modifications through an integrated network-based approach. The additive effects of alterations on multiple regulatory elements of the gene were considered. Using this tool, the authors have successfully identified many known cancer genes from publicly available cancer epigenome datasets.

The omics dataset has high dimensionality, and the relationship between omics features is very complex. Yao et al. proposed a method based on integrated swarm intelligence to identify key biomarkers and effectively reduce the feature dimension. It was an end-to-end method that only relied on the rules of the algorithm itself, without presets such as the number of filtered features. Furthermore, this method achieved good classification accuracy without excessive consuming computational resources.

With the development of multi-omics algorithms and the application of artificial intelligence, the automatic identification and classification of biomarkers have made great progress and have been widely used in biomarker detection research (Marshall et al., 2022).

AUTHOR CONTRIBUTIONS

WCC wrote the editorial, which was edited by JL, and endorsed by FZ, LH, and FL.

REFERENCES

Feng, L., Liu, Z., Li, C., Li, Z., Lou, X., Shao, L., et al. (2022). Development and Validation of a Radiopathomics Model to Predict Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer: A Multicentre Observational Study. *Lancet Digit. Health* 4 (1), E8–E17. doi:10.1016/S2589-7500(21)00215-6

Fu, Y., Liu, S., Zeng, S., and Shen, H. (2019). From Bench to Bed: The Tumor Immune Microenvironment and Current Immunotherapeutic Strategies for Hepatocellular Carcinoma. *J. Exp. Clin. Cancer Res.* 38, 396. doi:10.1186/s13046-019-1396-4

Marshall, J. L., Peshkin, B. N., Yoshino, T., Vowinckel, J., Danielsen, H. E., Melino, G., et al. (2022). The Essentials of Multiomics. *Oncologist* 27 (4), 272–284. doi:10.1093/oncolo/oyab048

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

frontiers in Genetics

METHODS
published: 20 January 2021
doi: 10.3389/fgene.2020.627746

CDAE: A Cascade of Denoising Autoencoders for Noise Reduction in the Clustering of Single-Particle Cryo-EM Images

Houchao Lei[1] *and Yang Yang*[1,2]*

[1] Center for Brain-Like Computing and Machine Intelligence, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, [2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai, China

As an emerging technology, cryo-electron microscopy (cryo-EM) has attracted more and more research interests from both structural biology and computer science, because many challenging computational tasks are involved in the processing of cryo-EM images. An important image processing step is to cluster the 2D cryo-EM images according to their projection angles, then the cluster mean images are used for the subsequent 3D reconstruction. However, cryo-EM images are quite noisy and denoising them is not easy, because the noise is a complicated mixture from samples and hardware. In this study, we design an effective cryo-EM image denoising model, CDAE, i.e., a cascade of denoising autoencoders. The new model comprises stacked blocks of deep neural networks to reduce noise in a progressive manner. Each block contains a convolutional autoencoder, pre-trained by simulated data of different SNRs and fine-tuned by target data set. We assess this new model on three simulated test sets and a real data set. CDAE achieves very competitive PSNR (peak signal-to-noise ratio) in the comparison of the state-of-the-art image denoising methods. Moreover, the denoised images have significantly enhanced clustering results compared to original image features or high-level abstraction features obtained by other deep neural networks. Both quantitative and visualized results demonstrate the good performance of CDAE for the noise reduction in clustering single-particle cryo-EM images.

Keywords: cryo-EM, autoencoder, image denoising, clustering, deep learning

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Qi Zhao,
University of Science and Technology Liaoning, China
Xiangxiang Zeng,
Hunan University, China
Fei Guo,
Tianjin University, China

***Correspondence:**
Yang Yang
yangyang@cs.sjtu.edu.cn

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

Received: *10 November 2020*
Accepted: *21 December 2020*
Published: *20 January 2021*

Citation:
Lei H and Yang Y (2021) CDAE: A Cascade of Denoising Autoencoders for Noise Reduction in the Clustering of Single-Particle Cryo-EM Images. Front. Genet. 11:627746. doi: 10.3389/fgene.2020.627746

1. INTRODUCTION

Recent progress of cryo-electron microscopy (cryo-EM) has revolutionized the field of structural biology (Cheng et al., 2015). Thanks to this technology, more and more spatial structures of biomolecules with nearly atomic-resolution have been solved. In order to obtain the 3D structure of a macromolecular, a large amount of 2D projection images with various orientations are captured, processed and averaged for reconstruction. At present, there are some softwares to realize the whole 3D reconstruction process, such as SPREAD (Xie et al., 2020). The whole pipeline involves quite a few scientific problems with great challenges in computation and algorithms.

During the preprocessing steps of images before 3D reconstruction, there are some major computational tasks listed in the following:

1. Estimation of the contrast transfer function (CTF) induced by the underfocus issue (Penczek et al., 1997). Specialized image processing algorithms such as phase flipping and amplitude correction/wiener filtering can or partially correct the CTF (Downing and Glaeser, 2008);
2. Automatic particle picking, i.e., recognizing and extraction of the particles from micrographs. Some popular software packages, like XMIPP (de la Rosa-Trevín et al., 2013), provide GUI programs to help pick projection images semi-automatically;
3. Clustering images by their projection angles. The images within clusters are averaged for 3D reconstruction, In addition to the common clustering methods such as kmeans, IterVM (Ji et al., 2018) proposes an iterative clustering model based on convolutional autoencoder model to solve the single particle clustering problem in cryo electron microscopy;
4. Identification of structural heterogeneity. The raw images often exhibit different conformations due to various reasons. In order to obtain high-resolution structures, different conformations should be distinguished and classified into homogeneous groups.

Solving the last two tasks largely relies on unsupervised learning algorithms, since in the real cryo-EM images, each particle's orientation is random and unknown, and the conformation information is also absent. The clustering result has a substantial impact on the sub-sequent reconstruction quality, as the projection images with dissimilar angles will dramatically decrease the qualities of class average images, which are the reconstruction inputs. Due to the low electron dose limitation (to prevent radiation damage), the cryo-EM images usually have too much noise, leading to extremely low signal-to-noise ratios (SNRs), which greatly increases the complexity of particle picking and clustering of images. However, the existing clustering algorithms are general-purpose methods, few of them are designed for such low-SNR scenario. Besides, denoising is not easy for cryo-EM images because the noise is a complicated mixture from samples and hardware. Therefore, how to reduce noise and improve the clustering performance has become a crucial problem for the structure reconstruction.

In this paper, we focus on noise reduction for the clustering of cryo-EM images. Especially, we design an image denoising model, CDAE, which is a cascade of denoising autoencoders to reduce noise in a progressive manner. The model comprises 3 blocks, each of which is pre-trained by a simulated data set and fine-tuned by the target data set. We evaluate the performance of the new model on both simulated and real data sets. The results show that CDAE achieves much higher PSNR (peak signal-to-noise ratio) than the state-of-the-art denoising methods, and it significantly improves the performance of conventional clustering methods compared with the clustering based on original images or feature representations yielded by other models.

To summarize, the contributions of this study are two folds:

1. In order to deal with the extremely low signal-to-ratio in cryo-EM images, we propose a cascade architecture, which consists of a stack of autoencoders, for denoising in a progressive manner.
2. In order to address the unsupervised denoising problem, we propose a two-phase learning strategy, including pre-training using simulated data and fine-tuning using real data. The strategy improves the denoising performance of autoencoders effectively.

2. RELATED WORK

2.1. Autoencoders for Feature Learning and Denoising

Autoencoder is a kind of unsupervised neural network, which comprises two parts, namely encoder and decoder. Encoder defines a parameterized function to extract features while decoder attempts to reconstruct original data from encoded features. The basic idea is to extract features through minimizing the reconstruction error.

Till now, various variants have been proposed to regularize the model. For instance, sparse autoencoder imposes a sparsity penalty on the latent layer to enforce sparsity of the features (Lee et al., 2007; Scholkopf et al., 2007). Instead of adding a penalty to the cost function, denoising autoencoder (DAE) (Vincent et al., 2008) attempts to reconstruct the original data from corrupted ones, which promotes the model to learn more useful and robust features. Following the DAE, contractive autoencoder (CAE) (Rifai et al., 2011) adds an analytic contractive penalty, which is a generalization of DAE. More recently, variational autoencoder (VAE)(Kingma and Welling, 2014) and adversarial autoencoders (AAE) (Makhzani et al., 2015) were designed to constraint the distribution of hidden variables. Most of these models aim to provide latent feature representations (dimensionality reduction) for subsequent learning, and some of them have been directly used for unsupervised clustering. For instance, GMVAE (Dilokthanakul et al., 2016) models the latent feature distribution as a Gaussian mixture distribution to cluster the latent vectors, and AAE could also serve as a clustering method when modeling the latent variables as a categorical distribution (Makhzani et al., 2015).

Besides, autoencoders have also been introduced in the denoising tasks. LeCun and Gallinari (Gallinari et al., 1987; Le Cun, 1987) pioneered the studies using autoencoders for noise reduction, and (Memisevic, 2007) designed a gated autoencoders for denoising. Note that denoising autoencoder (DAE) (Vincent et al., 2008) gets the name because its inputs are corrupted data, while its training objective is obtaining robust features rather than denoising.

2.2. Clustering of Cryo-EM Images

In recent years, various software packages for cryo-EM image processing have been released, many of which contain the clustering function. Some of them use *k*-means based clustering

algorithm, such as XMIPP (Scheres et al., 2008). The clustering module of XMIPP is an implementation of CL2D algorithm (Sorzano et al., 2010), which is a modified K-means method. CL2D uses cross-entropy as the measurement of image similarity and proposes a new clustering criterion to address the varied SNR issue. Another well-known package, Spider (Frank et al., 1996), implements hierarchical clustering. These methods perform distance calculation directly using raw images.

Besides the conventional clustering methods, new algorithms specialized for cryo-EM images have also emerged. Relion (Scheres, 2012) developed a maximum likelihood (ML) based approach, aiming to find the optimal probability estimation, which is more robust to the influence of noise than traditional methods, but it is incompetent in differentiating subtle structural heterogeneity. Recently, a new software package ROME (Wu et al., 2016) was proposed, which introduces a new kind of clustering method based on statistical manifold learning (SML). The basic idea is to map the original data space into a lower dimensional latent space by a non-linear transformation, and then optimize the parameters by expectation-maximization (EM) algorithm.

3. METHODOLOGY

3.1. Problem Description

In a basic autoencoder model, the input and target output are the same; while our goal is noise reduction, thus the input and target output in our model are different. Let X and Y denote the sets of original noisy images and target clean images, respectively. We want to find a mapping function $f : \mathcal{X} \mapsto \mathcal{Y}$, as formulated in Equation (3),

$$z = EC(x), \tag{1}$$

$$y = DC(z), \tag{2}$$

$$y = f(x) = DC(EC(x)), \tag{3}$$

where $x \in X$, $y \in Y$, and z is the latent representation. EC is an encoder, and DC is a decoder.

In a supervised learning scenario, the mapping function f can be learned from training data, but our task is unsupervised, because real cryo-EM images have no clean targets. In order to address this problem, we convert the original task into a supervised learning problem and adopt a two-phase learning strategy as shown in **Figure 1**. First, we pre-train the autoencoders with simulated paired cryo-EM data, which has the clean target image for training, and then we fine-tune the model with real data to transfer knowledge from simulated cryo-EM data to real data. These two phases are described in sections 3.2, 3.3, respectively.

3.2. Pre-training

Let X_{tr} and Y_{tr} denote the sets of the corrupted images and target images of the simulated training data, respectively. And $x_{tr}^{(i)} \in X_{tr}$ is an input image for the encoder, where $i \in \{1, 2, \ldots, n\}$ and n is the number of training images. The parameters, $\theta = \{W, b\}$ for EC and $\phi = \{W', b'\}$ for DC, are optimized to minimize the average reconstruction error as shown in Equation (4),

$$\theta^*, \phi^* = \arg\min_{\theta,\phi} \frac{1}{n} \sum_{i=1}^{n} L(y_{tr}^{(i)}, DC_\phi(EC_\theta(x_{tr}^{(i)}))), \tag{4}$$

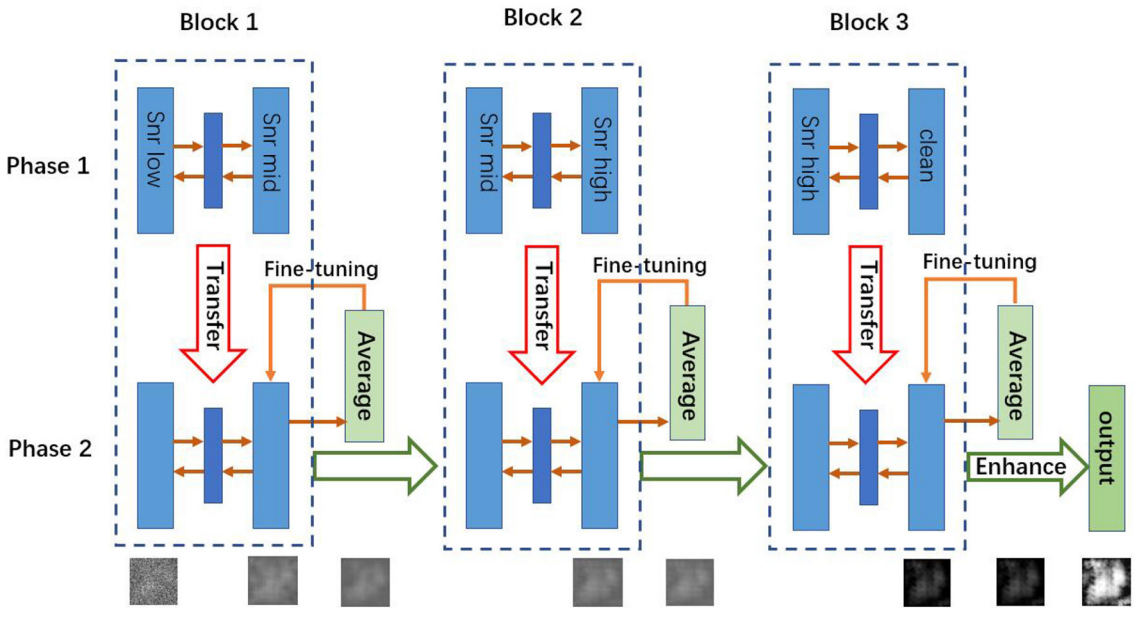


FIGURE 1 | Architecture of CDAE. The algorithm has two phases. In Phase 1, for each block, the autoencoder is pre-trained by using the training data. Then the learned weights are transferred into Phase 2, where the autoencoders are fine-tuned by test data sets. Note that in Phase 1, the three blocks can be trained parallely, while in Phase 2, they are fine-tuned sequentially.

where L is the loss function, such as mean-square-error.

3.3. Fine-Tuning

Let X_{te} denote the sets of the images of test dataset, i.e., real data, and $x_{te}^{(i)} \in X_{te}$, where $i \in \{1, 2, \ldots, m\}$ and m is the number of test images. DC' and EC' are pre-trained decoder and encoder, respectively. The parameters, θ' of EC' and ϕ' of DC', are further optimized to minimize the average reconstruction error as shown in Equation (5),

$$\theta'^{*}, \phi'^{*} = \arg\min_{\theta', \phi'} \frac{1}{m} \sum_{i=1}^{m} L(\overline{x_{te}^{(i)'}}, DC'_{\phi'}(EC'_{\theta'}(x_{te}^{(i)}))), \quad (5)$$

$$x_{te}^{(i)'} = DC'_{\phi'}(EC'_{\theta'}(x_{te}^{(i)})), \quad (6)$$

where $x_{te}^{(i)'}$ is the corresponding output of $x_{te}^{(i)}$ by using EC' and DC' (Equation 6), and $\overline{x_{te}^{(i)'}}$ is the mean image of $x_{te}^{(i)'}$ averaged over its neighborhood, which is determined by a certain similarity metric and a threshold. Since there is no known clean data for test data, the mean images are used as target output instead. We use mean images as the targets because images of close orientations or conformations have similar features, but the noises mostly due to random events are not similar in these images. Thus the mean images will weaken the influence of noise and it could be regarded as a substitute for the target images without noise.

It is worth noting that we use the same data set in the fine-tuning stage and the test stage. However, in the fine-tuning stage, we only use the images of the test dataset, but not the targets of the test dataset. We use the mean images averaged over each image's neighbors as the target for training; while in the test stage, we use images and targets of test dataset to calculate the corresponding quantitative metrics.

3.4. The Cascade Design

The proposed CDAE model is a cascade of denoising autoencoders, which aims to reduce noise in a progressive manner for the images with very low SNR. As shown in **Figure 1**, CDAE has three blocks, each of which contains a convolutional autoencoder. During the pre-training phase, the first block learns the mapping from the images with a low SNR (SNR_{low}) to images with a medium SNR (SNR_{mid}), the second block learns from data of SNR_{mid} to data of SNR_{high}, and the last layer learns from data of SNR_{high} to clean data. Then, we fine-tune the blocks sequentially from Block 1 to Block 3. The outputs of the fine-tuned blocks are fed to the next block. Finally, we make a histogram equalization enhancement to the output images of the last block. The procedure is described in Algorithm 1.

3.5. Architecture of the Model Components

The proposed CDAE model comprises three components/blocks. Considering the advantages of convolutional neural networks in representing image features, we build a convolutional autoencoder in each block. The three autoencoders use the same parameters as listed in **Figure 2**. The encoder consists of 3 modules, each of which contains 2 convolutional layers and a pooling layer; while the decoder consists of 4 layers, including 3 deconvolutional layers and a convolutional layer. The function of the last convolution layer is to combine 32 channels into one channel as output. In order to avoid overfitting, we use dropout in the encoder and decoder and set dropout rate to 0.5.

Algorithm 1: The CDAE Algorithm

Input: The training data sets: $X_{tr,i} (i \in \{1, 2, 3, 4\})^a$, and the test data set X_{te};

Output: Denoised image X_{te}^{*};

1: Train the three blocks separately and obtain the mapping function f_j from $X_{tr,j}$ to $X_{tr,j+1}$, i.e.,
$f_j : X_{tr,j} \mapsto X_{tr,j+1}, j \in \{1, 2, 3\}$
2: $X_{te}^{1} = X_{te}$
3: Fine-tune Block 1 and obtain the updated mapping function f_1', i.e.,
$f_1' : X_{te} \mapsto \overline{f(X_{te}^{1})}$
4: **for** $j \in \{2, 3\}$ **do**
5: $\quad X_{te}^{j} = f_{j-1}'(X_{te}^{j-1})$
6: $\quad f_j' : X_{te}^{j} \mapsto \overline{f(X_{te}^{j})}$ (f_j' is initialized by f_j)
7: **end for**
8: $X_{te}^{*} = Enhance(f_3'(X_{te}^{3}))$
9: **return** X_{te}^{*};

[a] $X_{tr,1}$, $X_{tr,2}$, $X_{tr,3}$ and $X_{tr,4}$ denote the training sets with SNR_{low}, SNR_{mid}, SNR_{high} and no noise, respectively.

4. EXPERIMENTAL RESULTS

4.1. Dataset and Experimental Setup

We collect molecular structure data from the Electron Microscopy Data Bank (EMDB) at PDBe (Sameer et al., 2016),and prepare two kinds of data, including the data simulated by ourselves and real data downloaded from EMDB. For the simulated data, we extract the 3D structures of 4 proteins from the EMDB database, whose PDB IDs are 5wth, 5k0y, 5flc, and 5gjq, and their real structures are present in **Figure 3**. We simulate their 2D EM projection images by using the cryo-EM data processing suitcase software, XMIPP (de la Rosa-Trevín et al., 2013), which has been widely used in cryo-EM data processing and protein reconstruction task. In our experiment, we take the 2D images of 5flc as the training data (for pre-training the model), and images of the other 3 proteins as the test data. For 5flc, we simulate images with 4 different noise ratios (SNR_{low}, SNR_{mid}, SNR_{high} and no noise) and 4 orientations. The number of images with the same orientation and SNR is 1,000. Thus, there is a total of $4 \times 4 \times 1,000 = 16,000$ pictures; while for the other 3 proteins, we only simulate the images with SNR_{low} at four orientations, thus each of which has 4,000 pictures. In addition, the SNR_{low}, SNR_{mid} and SNR_{high} used for simulation are set to 0.1, 0.4, 0.6, respectively. And, the number of closest neighbors (k) for obtaining mean images is set to 30.

Beside the simulated data, we also retrieve a real data set from EMDB, the cryo-EM images of GroEL (PDB entry 10029), where the simulation condition is 300 kV acceleration voltage. Since there is no orientation or conformation information in the data set, here we only show the visualized results (see section 4.5), i.e., the mean images from the clusters of denoised images.

4.2. Evaluation Criteria

In order to assess the new model, we provide both quantitive results (denoising and clustering experiments) and visualized results. The measurement of denoising performance lies in the similarity between reconstruction data and the clean data, while the clustering performance is evaluated via the following criteria, F_1, *Precision*, and *Recall*. The visualized results provide a comparison between the denoised images and ground truth structure, which can be observed directly.

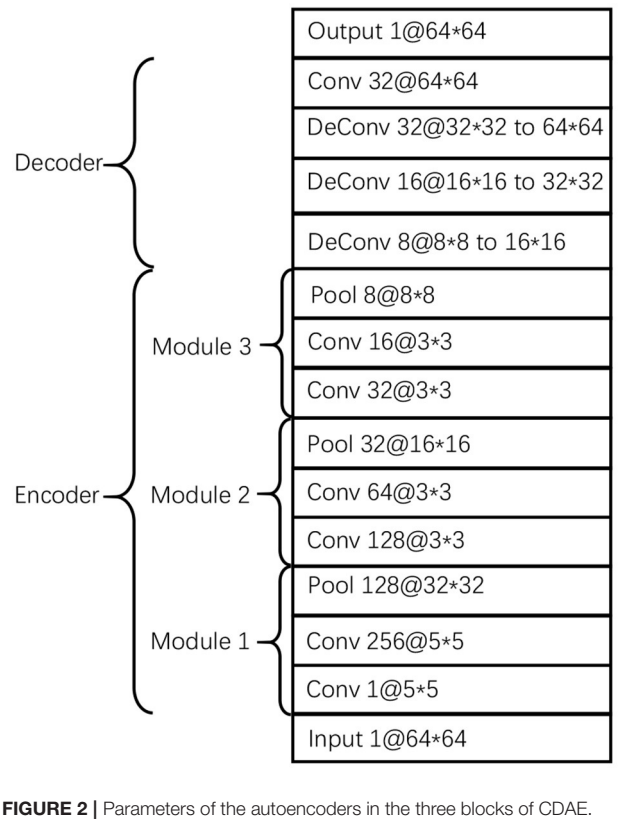


FIGURE 2 | Parameters of the autoencoders in the three blocks of CDAE.

4.3. Denoising Performance

We first compare the denoising performance of the new model with the state-of-the-art denoising methods in terms of PSNR (peak signal-to-noise ratio), a common criterion for measuring the denoising quality. The higher the PSNR, the better the quality of the reconstructed image. In this experiment, we use the simulated images of 5wth, 5gjq, and 5k0y (SNR = 0.1) as the test datasets, and compare the PSNR scores obtained by CDAE and the following 7 methods:

- Filter-based denoising, including NLFMT (Kumar, 2013), BM3D (Dabov et al., 2008) and PID (Knaus and Zwicker, 2014);
- Sparse coding-based denoising, NCSR (Dong et al., 2013);
- Effective priori-based method, PCLR (Xu et al., 2017)
- Deep learning-based method, DnCNN (Zhang et al., 2017) and a single denoising autoencoder, namely single DAE, which has the same model architecture as the autoencoder used in each block of CDAE.

The results are listed in **Table 1**, and the denoised images are shown in **Supplementary Table 1**. We use histogram equalization enhancement (HEE) in our method because the output gray values concentrate in a narrow range and the output is sparse. Specifically, the gray values of our model outputs concentrate in a narrow range, and HE can help remap the gray values to a wider range. HEE is commonly used in signal processing and does not modify the main property and features of denoised images. In order to examine the effect of histogram equalization enhancement, we consider two versions of the 6 existing methods, i.e., with and without HEE. Among the 8 methods, CDAE achieves the highest PSNR on 5wth, which is the hardest one among the three proteins, because protein 5wth is small and it has no distinct structural characteristics (as can be seen in **Figure 3**). For 5k0y, CDAE performs very close to the best method, NLFMT (8.2143 vs. 8.2640); and for

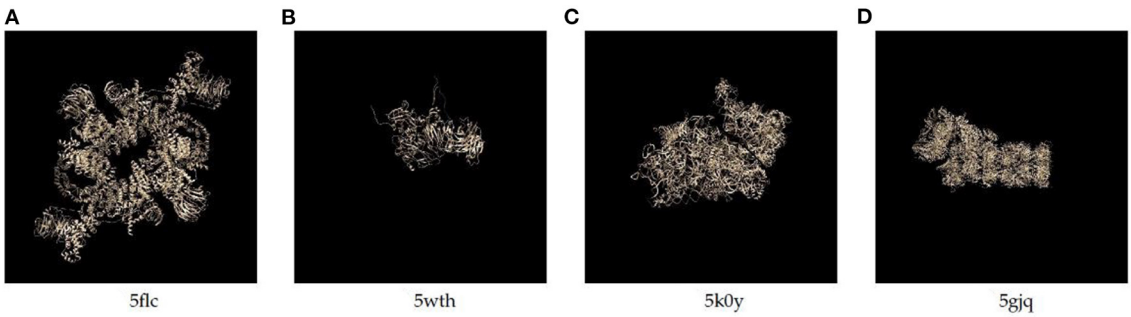


FIGURE 3 | Structures of the 4 proteins used in our data sets.

5gjq, CDAE ranks the third place. The histogram equalization enhanced NLFMT achieves the best results on both 5k0y and 5gjq. However, its HEE version performs not stable, as the PSNR values decreases dramatically on 5wth. For most of the methods, HEE leads to reduced PSNRs. Overall, CDEA is a very competative method compared with the existing image denoising methods. Also, through the denoised images, we find that CDAE gets more sparse images than others, thus the specific structural features will be enhanced. Interestingly, our cascade model outperforms the single denoising autoencoder on all the three data sets, indicating that reducing noise progressively would be a practical strategy for handling very-low-SNR images.

4.4. Clustering Performance

Since our ultimate goal is to improve the clustering performance, so as to get better mean images for 3D structure reconstruction, we cluster the denoised images with some conventional unsupervised algorithms, i.e., *k*means and hierarchical clustering (HC), and compare the accuracy with 6 other methods, which fall into two categories:

1. Traditional methods: *k*means (working on original images), HC (working on original images), PCA+*k*means (working on principle components of the original images) and CL2D (implemented in XMIPP);
2. Deep model based methods: CAE+*k*means (convolutional autoencoder with *k*means), AAE+*k*means (adversarial autoencoder with *k*means, the generator of AAE is a convolutional autoencoder, Makhzani et al., 2015), and DAE+*k*means (denoising autoencoder with *k*means). For the first two methods, latent representations extracted from the middle layer of the convolutional autoencoder are used for clustering, and both inputs and outputs are the original test images; while for DAE, the mean image (averaged over 30 nearest neighbors) for each original image serves as target output, and the outputs of decoder are used in clustering (note that it is different from the original denoising autoencoder proposed by Vincent et al. (2008) as there is no clean target for test data).

TABLE 1 | Denoising result comparison for eight methods.

Method	5k0y	5wth	5gjq
PCLR	7.22/8.18	6.78/5.66	6.85/8.21
PID	6.87/5.33	6.48/4.95	6.50/5.22
NLFMT	7.15/**8.26**	6.89/5.64	6.74/**8.55**
BM3D	7.05/5.13	6.66/5.22	6.56/6.87
NCSR	7.09/5.37	6.58/4.92	6.55/5.29
DnCNN	7.02/5.38	6.58/4.99	6.52/5.32
Single DAE	7.55	6.77	6.88
CDAE	8.21	**6.92**	7.07

The numbers before and after "/" denote the PSNR values without and with histogram equalization enhancement (HEE). Both single DAE and CDAE include the HEE step, thus their PSNRs are obtained after HEE. The best values are in bold.

All the convolutional autoencoders in the compared deep models (CAE, DAE, and the generator of AAE) have almost the same architecture as the single blocks in our model. We use rmsprop optimizer and train the model by 20 epochs, while in AAE we add extra GAN training procedure to set constraints on latent variables. We also use rmsprop optimizer and train the model by 1500 iterations.

Table 2 shows that our model outperforms other methods at all of the three datasets, indicating that deep-models have great potential serving as image denoising tools. The detailed discussions are as follows.

Among the first four traditional methods, PCA obtains the best results on both 5gjq and 5k0y. Although it is a simple linear transformation, PCA captures the key features that are helpful for clustering the images.

The last five methods are all based on autoencoders, while their performance differs a lot. AAE does not perform well in

TABLE 2 | Clustering result comparison*.

Method	Measure	5gjq	5wth	5k0y
*k*means	F_1	0.76	0.29	0.54
	Precision	0.68	0.25	0.43
	Recall	0.80	0.34	0.59
HC	F_1	0.79	0.29	0.56
	Precision	0.74	0.27	0.51
	Recall	**0.84**	0.33	0.65
PCA+*k*means	F_1	0.76	0.29	0.72
	Precision	0.67	0.26	0.63
	Recall	0.78	0.34	0.72
CL2D	F_1	0.30	0.28	0.29
	Precision	0.29	0.27	0.27
	Recall	0.34	0.33	0.30
CAE+*k*means	F_1	0.77	0.3	0.54
	Precision	0.7	0.26	0.42
	Recall	0.8	0.39	0.59
DAE+*k*means	F_1	0.40	0.34	0.59
	Precision	0.36	0.32	0.47
	Recall	0.45	0.37	0.75
AAE+*k*means	F_1	0.4	0.29	0.46
	Precision	0.26	0.26	0.41
	Recall	0.32	0.35	0.47
CDAE+HC	F_1	**0.81**	0.94	0.75
	Precision	**0.79**	0.94	0.73
	Recall	0.80	0.93	**0.77**
CDAE+*k*means	F_1	0.76	**0.95**	0.76
	Precision	0.76	**0.95**	**0.75**
	Recall	0.78	**0.95**	0.77

**HC and kmeans denote hierarchical clustering and kmeans method working with the raw images, respectively; PCA+kmeans denotes clustering of principle components via kmeans; CAE and AAE denote the conventional convolutional autoencoder and adversarial autoencoder, respectively; DAE denotes the convolutional autoencoder with the original test images as input and their mean images within the neighborhood as output (no pre-training), and CDAE denotes our model. Bold values means that they are the maximum metrics value in this dataset.*

this task, mainly due to the intrinsic difficulties in the training of the model, which restricts its applications. AAE obtains a lower accuracy even than the traditional methods. As the latent feature vector is a compact representation for the image with much lower dimensionality, if the representation is not good, the clustering performance may be even worse than using original images.

According to the accuracy of CAE, the latent representations could also be useful in the clustering of cryo-EM images, but they also try to reconstruct the noisy patterns, thus may not yield a satisfying result.

DAE has much lower accuracy than CDAE, suggesting that the average images of the original images may not be a good choice for the reconstruction target. By contrast, CDAE adopts a two-phase learning strategy and a cascade structure, which both contribute to the good performance.

CDAE+*k*means and CDAE+HC have very close performance, indicating the robustness of the extracted representations. An interesting result is that our model achieves significantly better accuracies on 5wth. We find that this molecule is relatively small compared to two others, and presents as a denser form in the central area of the images, which may increase the difficulty in clustering. Except CDAE, all the other methods almost group the images into one cluster. The results demonstrate that CDAE captures the discriminant features rightly, thus greatly enhances the performance.

4.5. Visualized Results

As mentioned in section 4.1, we download a data set of protein GroEL from EMDB without corresponding clean images or orientation information. Therefore, clustering or denoising performance can not be evaluated, thus we present the visualized result. **Figure 4** shows some examples of the denoised images. It can be observed that the images are consistent with the true structure, and can differentiate between the projection angles.

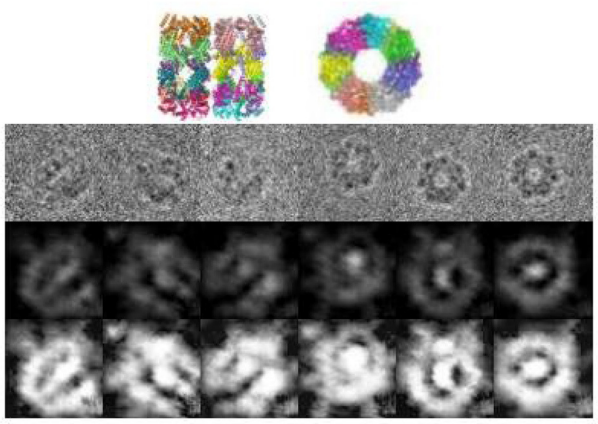


FIGURE 4 | Denoised images of GroEL in different orientations. The 1st row shows the original images, The 2nd row shows the projections of the ground truth structure, the 3rd row shows the denoised images, and the last row shows the denoised images processed by histogram equalization.

5. DISCUSSION

The proposed CDAE model involves both pre-training and fine-tuning. Benefitting from the abundance of 3D structure simulation software, it is convenient to generate projection images from pre-defined orientations for a certain biomolecule. Therefore, the simulated cryo-EM images could serve as a kind of supervision in the learning algorithms. Furthermore, the mean images can be used for fine-tuning, because the averaging operation can effectively reduce random noisy, and many cryo-EM data processing algorithms use it to enhance the image features, like EMAN2 (Tang et al., 2016). We also design the denoising model in a cascade structure based on the following concern. The cryo-EM images often have a high noise ratio. During the pre-training phase, if we choose a low SNR for the simulated data, apparently the input and target output differ a lot, and it is hard for the layers to adapt the noise; but if we set a high SNR, although the deep network could easily learn the noisy pattern, it does not accord with the real case, and the quality of learning would be affected. Therefore, we want to reduce noisy in a progressive manner and design a cascade of denoising autoencoders to reduce the noise step by step.

The quantitative and visualized experimental results in the previous sections demonstrate the good performance of CDAE, which is attributed to the advantages on model design. Comparing with the DnCNN model, our model has a deeper network architecture, which may have greater capacity on feature representation; and comparing with the single DAE model, our model benefits from the cascade design, which can gradually and smoothly guide the denoising process, thus making the denoising process more controllable and leading to better denoising effect.

The proposed model is closely related with denoising autoencoder (DAE) (Vincent et al., 2008) and Stacked Denoising Autoencoders (SdA) (Vincent et al., 2010). Actually, the components of our model, the autoencoder in each block, has the same architecture of DAE, and both of them are fed with corrupted images and rendered to reconstruct clean images. However, the objectives of these two methods are fundamentally different. Unlike our model, DAE aims to learn robust features, and use the pre-trained autoencoder as an initialization for subsequence supervised learning tasks. Therefore, the DAE model is fine-tuned by training data in a supervised manner, while our model is fine-tuned in a pseudo-supervised manner, in which the mean images are assumed to be the reconstruction targets.

Besides, our model also looks similar with the SdA model (Vincent et al., 2010). However, the architecture of these two models are very different. Our model consists of three blocks, each block has the same component autoencoder. And for Blocks 2 and 3, they are red by the outputs (denoised images) from previous blocks; while in SdA, it is the latent representation rather than the output being passed to the next autoencoder. And, SdA has the same object of DAE and also receives a supervised fine-tuning.

Although CDAE achieves a good performance on PSNR metric and visual results, there is still a big gap between the denoised images and the ground truth clean images. There are

two possible reasons. First, the added noise in simulated data may be very different from true noise. The noise in real cryo-EM images usually has complex sources, while the simulated images are added with Gaussian noise or noise with single types of distribution. Second, the neighboring images that are used for computing mean images may be selected inaccurately, as the images are extremely noisy and it is difficult to measure image similarity. Therefore, our future work will focus on the generation of noisy images to improve the pre-training process and investigate the similarity metric of images.

6. CONCLUSION

In this study, we propose a cascade of denoising autoencoders to reduce noise in cryo-EM images and enhance the clustering performance. This model contains 3 denoising blocks, and each block contains a denoising autoencoder. The 3 blocks learn simulated images from low SNR to medium SNR, medium SNR to high SNR, high SNR to clean data, respectively. After the pre-training, each autoencoder is fine-tuned by using the mean images. We provide both quantitative and visualized results on both simulated and real data sets. In the quantitative experiments, we compare the PSNR values with other denoising algorithms and evaluate the clustering performance, while in visualization evaluation, we compare the denoised images with the ground truth protein structure. The experiments show that our method achieves significant better performance of denoising and clustering than the state-of-the-art methods on the highly noisy cryo-EM images.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: ebi.ac.uk/pdbe/emdb/empiar/entry/10029/.

AUTHOR CONTRIBUTIONS

HL and YY designed the model, analyzed the results, and wrote the manuscript. HL conducted the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Program of China (No. 2018YFC0910500), and the National Natural Science Foundation of China (No. 61972251).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.627746/full#supplementary-material

REFERENCES

Cheng, Y., Grigorieff, N., Penczek, P., and Walz, T. (2015). A primer to single-particle cryo-electron microscopy. *Cell* 161, 438–449. doi: 10.1016/j.cell.2015.03.050

Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. O. (2008). "Image restoration by sparse 3D transform-domain collaborative filtering," in *Image Processing: Algorithms and Systems VI* (San Jose, CA). doi: 10.1117/12.766355

de la Rosa-Trevín, J., Otón, J., Marabini, R., Zaldívar, A., Vargas, J., Carazo, J. M., et al. (2013). XMIPP 3.0: an improved software suite for image processing in electron microscopy. *J. Struct. Biol.* 184, 321–328. doi: 10.1016/j.jsb.2013.09.015

Dilokthanakul, N., Mediano, P. A. M., Garnelo, M., Lee, M. C. H., Salimbeni, H., Arulkumaran, K., et al. (2016). Deep unsupervised clustering with Gaussian mixture variational autoencoders. *arXiv*. arXiv:1611.02648.

Dong, W., Zhang, L., Shi, G., and Li, X. (2013). Nonlocally centralized sparse representation for image restoration. *IEEE Trans. Image Process.* 22, 1620–1630. doi: 10.1109/TIP.2012.2235847

Downing, K. H., and Glaeser, R. M. (2008). Restoration of weak phase-contrast images recorded with a high degree of defocus: the twin image problem associated with ctf correction. *Ultramicroscopy* 108, 921–928. doi: 10.1016/j.ultramic.2008.03.004

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., et al. (1996). Spider and web: processing and visualization of images in 3d electron microscopy and related fields. *J. Struct. Biol.* 116, 190–199. doi: 10.1006/jsbi.1996.0030

Gallinari, P., Lecun, Y., Thiria, S., and Fogelman-Soulie, F. (1987). "Memoires associatives distribuees: une comparaison (distributed associative memories: a comparison)," in *Proceedings of COGNITIVA* 87.

Ji, G., Yang, Y., and Shen, H. (2018). "IterVM: an iterative model for single-particle cryo-em image clustering based on variational autoencoder and multi-reference alignment," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 999–1002. doi: 10.1109/BIBM.2018.8621474

Kingma, D. P., and Welling, M. (2014). Auto-encoding variational bayes. *arXiv*. arXiv:1312.6114.

Knaus, C., and Zwicker, M. (2014). Progressive image denoising. *IEEE Trans. Image Process.* 23, 3114–3125. doi: 10.1109/TIP.2014.2326771

Kumar, B. K. S. (2013). Image denoising based on non-local means filter and its method noise thresholding. *Signal Image Video Process.* 7, 1211–1227. doi: 10.1007/s11760-012-0389-y

Le Cun, Y. (1987). Modeles connexionnistes de l'apprentissage. *These De Doctorat Universite Paris* 15, 1–9. doi: 10.3406/intel.1987.1804

Lee, H., Ekanadham, C., and Ng, A. Y. (2007). "Sparse deep belief net model for visual area v2," in *Conference on Advances in Neural Information Processing Systems*.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2015). Adversarial autoencoders. Computerence. *arXiv*. arXiv:1511.05644.

Memisevic, R. (2007). Non-linear latent factor models for revealing structure in high-dimensional data. Available online at: http://hdl.handle.net/1807/11118

Penczek, P. A., Zhu, J., Schroder, R., and Frank, J. (1997). Three dimensional reconstruction with contrast transfer compensation from defocus series. *Scan. Microsc.* 118, 147–154.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011). "Contractive auto-encoders: explicit invariance during feature extraction," in *ICML*.

Sameer, V., van Ginkel, G., Younes, A., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., et al. (2016). PDBE: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.* 44, D385–D395. doi: 10.1093/nar/gkv1047

Scheres, S. H., Núñez-Ramírez, R., Sorzano, C. O., Carazo, J. M., and Marabini, R. (2008). Image processing for electron microscopy single-particle analysis using XMIPP. *Nat. Protoc.* 978. doi: 10.1038/nprot.2008.62

Scheres, S. H. W. (2012). RELION: Implementation of a bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180, 519–530. doi: 10.1016/j.jsb.2012.09.006

Scholkopf, B., Platt, J., and Hofmann, T. (2007). Efficient learning of sparse representations with an energy-based model. in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, 1137–1144.

Sorzano, C. O. S., Bilbao-Castro, J. R., Shkolnisky, Y., Alcorlo, M., Melero, R., Caffarena-Fernandez, G., et al. (2010). A clustering approach to multireference alignment of single-particle projections in electron microscopy. *Journal of Structural Biology* 171, 197–206. doi: 10.1016/j.jsb.2010.03.011

Tang, G., Peng, L., Mann, D., Yang, C., Penczek, P., Goodyear, G., et al. (2016). Eman2: Software for image analysis and single particle reconstruction. *Microscopy Microanal.* 12, 388–389. doi: 10.1017/S1431927606067699

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. in *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103. doi: 10.1145/1390156.1390294

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A. (2010). Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, 3371–3408. doi: 10.1016/j.mechatronics.2010.09.004

Wu, J., Ma, Y. B., Congdon, C., Brett, B., Chen, S., Ouyang, Q., et al. (2016). Unsupervised single-particle deep classification via statistical manifold learning. *arXiv*. doi: 10.1371/journal.pone.0182130

Xie, R., Chen, Y.-X., Cai, J.-M., Yang, Y., and Shen, H.-B. (2020). Spread: a fully automated toolkit for single-particle cryogenic electron microscopy data 3d reconstruction with image-network-aided orientation assignment. *J. Chem. Inform. Model.* 60, 2614–2625. doi: 10.1021/acs.jcim.9b01099

Xu, J., Zhang, L., and Zhang, D. (2017). External prior guided internal prior learning for real-world noisy image denoising. *IEEE Trans. Image Process.* 27, 2996–3010. doi: 10.1109/TIP.2018.2811546

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a Gaussian denoiser: residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* 26, 3142–3155. doi: 10.1109/TIP.2017.2662206

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 22 April 2021
doi: 10.3389/fmolb.2021.645084

Identification of Prognostic Glycolysis-Related lncRNA Signature in Tumor Immune Microenvironment of Hepatocellular Carcinoma

Yang Bai[1,2], *Haiping Lin*[1], *Jiaqi Chen*[3], *Yulian Wu*[2]* *and Shi'an Yu*[1]*

1Department of Hepatobiliary and Pancreatic Surgery, Affiliated Jinhua Hospital, Zhejiang University School of Medicine, Jinhua, China, [2]Department of Surgery, the Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China, [3]The Affiliated Hospital of Stomatology, School of Stomatology, Zhejiang University School of Medicine, and Key Laboratory of Oral Biomedical Research of Zhejiang Province, Hangzhou, China

Purpose: The purpose of this study was to construct a novel risk scoring model with prognostic value that could elucidate tumor immune microenvironment of hepatocellular carcinoma (HCC).

Samples and methods: Data were obtained through The Cancer Genome Atlas (TCGA) database. Univariate Cox analysis, least absolute shrinkage and selection operator (LASSO) analysis, and multivariate Cox analysis were carried out to screen for glycolysis-related long noncoding RNAs (lncRNAs) that could provide prognostic value. Finally, we established a risk score model to describe the characteristics of the model and verify its prediction accuracy. The receiver operating characteristic (ROC) curves of 1, 3, and 5 years of overall survival (OS) were depicted with risk score and some clinical features. ESTIMATE algorithm, single-sample gene set enrichment analysis (ssGSEA), and CIBERSORT analysis were employed to reveal the characteristics of tumor immune microenvironment in HCC. The nomogram was drawn by screening indicators with high prognostic accuracy. The correlation of risk signature with immune infiltration and immune checkpoint blockade (ICB) therapy was analyzed. After enrichment of related genes, active behaviors and pathways in high-risk groups were identified and lncRNAs related to poor prognosis were validated *in vitro*. Finally, the impact of MIR4435-2HG upon ICB treatment was uncovered.

Results: After screening through multiple steps, four glycolysis-related lncRNAs were obtained. The risk score constructed with the four lncRNAs was found to significantly correlate with prognosis of samples. From the ROC curve of samples with 1, 3, and 5 years of OS, two indicators were identified with high prognostic accuracy and were used to draw a nomogram. Besides, the risk score significantly correlated with immune score, immune-

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Sayan Chakraborty,
*Institute of Molecular and Cell Biology (A*STAR), Singapore*
Vinay Kumar Mittal,
Other, Singapore

***Correspondence:**
Shi'an Yu
ysa513513@gmail.com
Yulian Wu
yulianwu@zju.edu.cn

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *22 December 2020*
Accepted: *19 February 2021*
Published: *22 April 2021*

Citation:
Bai Y, Lin H, Chen J, Wu Y and Yu S (2021) Identification of Prognostic Glycolysis-Related lncRNA Signature in Tumor Immune Microenvironment of Hepatocellular Carcinoma. Front. Mol. Biosci. 8:645084. doi: 10.3389/fmolb.2021.645084

Abbreviations: DMEM, Dulbecco's minimum essential media; GSEA, gene set enrichment analysis; HCC, hepatocellular carcinoma; HR, hazard ratio; KEGG, kyoto encyclopedia of genes and genomes; lncRNAs, long noncoding RNAs; LASSO, least absolute shrinkage and selection operator; OS, overall survival; qRT-PCR, quantitative real-time polymerase chain reaction; ROC, receiver operating characteristic; TCGA, The Cancer Genome Atlas; ICB, Immune Checkpoint Blockade; TIME, Tumor Immune Microenvironment; TMB, Tumor Mutation Burden; MAF, Mutation Annotation Format; MsigDB, Molecular Signatures Database; TIMER, Tumor Immune Estimation Resource.

related signature, infiltrating immune cells (i.e. B cells, etc.), and ICB key molecules (i.e. CTLA4,etc.). Gene enrichment analysis indicated that multiple biological behaviors and pathways were active in the high-risk group. *In vitro* validation results showed that MIR4435-2HG was highly expressed in the two cell lines, which had a significant impact on the OS of samples. Finally, we corroborated that MIR4435-2HG had intimate relationship with ICB therapy in hepatocellular carcinoma.

Conclusion: We elucidated the crucial role of risk signature in immune cell infiltration and immunotherapy, which might contribute to clinical strategies and clinical outcome prediction of HCC.

Keywords: hepatocellular carcinoma, glycolysis, prognostic model, tumor immune environment, immune checkpoint blockade, bioinformatics analysis

INTRODUCTION

Liver cancer is one of the most common malignant tumors with a high rate of metastasis and high mortality (Siegel et al., 2020). With the development of modern medicine, the comprehensive treatment strategy has greatly improved the prognosis of samples with liver cancer (Anwanwan et al., 2020). However, due to the high recurrence rate of liver cancer, the long-term prognosis of samples remains poor (Dufour et al., 2013). Currently, the administrations of immune checkpoint blockade inhibitors have revolutionized antitumor treatment in wide range of cancers. According to preclinical trials, about 20% of samples were observed for objective response, indicating immune checkpoint inhibitors may contribute novel insight into clinical intervention and decision-making of HCC (Cheng et al., 2019). The immune cells function as tumor inhibitor or tumor promoter and may act as important players in the tumor immune microenvironment (TIME) (Lei et al., 2020). Due to characteristics of the immune contexture significantly influencing immune therapy outcome (Zhang et al., 2019), it is worth identifying immune indicators which could predict treatment efficacy and prognosis. At present, the prognosis of samples is typically judged by the grade and stage of tumors (Hu et al., 2019). Tumor mutation burden (TMB), which represents the somatic coding errors such as base substitutions, insertions, or deletions across per million bases, has been termed as a promising indicator for predicting responsiveness to ICB based on numerous researches (Snyder et al., 2014; Rizvi et al., 2015; Chan et al., 2019). Exploring new ways to judge prognosis and clinical outcome is helpful to the survival evaluation and disease treatment of samples.

Long noncoding RNAs (lncRNAs) are similar to mRNA in structure, with a length of more than 200 nucleotides, though they do not have the ability to encode proteins (Kopp and Mendell, 2018). Earlier views believed that lncRNAs were a byproduct of translation and generally did not have a function. At the present time, increasing studies have provided evidence to support that lncRNAs act as a vital regulator in immune response, such as immune activation and antigen release (Carpenter and Fitzgerald, 2018; Denaro et al., 2019). An independent research pointed out that lncRNA GAS5 was downexpressed in HCC tumor compared with normal tissue and interference of lncRNA GAS5 accelerated tumor cell migration by reducing NK cell cytotoxicity (Fang et al., 2019). Likewise, lncRNA TCONS_00019715 could promote antitumor response via harnessing macrophage transformation into the M1 phenotype (Huang et al., 2016). Some studies reported that lncRNAs could serve as novel indicators for disease diagnosis, treatment monitoring, and prognostic prediction in HCC (DiStefano, 2017; Wei et al., 2019). However, with increasing research, it has been found that lncRNAs play an important role in cell growth, differentiation and regulation of gene expression (Schmitt and Chang, 2016). It has been reported that a variety of lncRNAs are stably expressed in HCC tissues and that specific lncRNAs play a significant role in the occurrence and development of HCC (Yuan et al., 2016).

The energy supply of human cells mainly comes from mitochondrial oxidative phosphorylation and glycolysis (Lu et al., 2015). Compared to normal cells, tumor cells choose glycolysis as the main method to supply energy, even under aerobic conditions. This abnormal energy metabolism is an important feature of tumor tissue (Ganapathy-Kanniappan, 2018). In this study, we used a variety of statistical methods to identify glycolysis-related lncRNAs to construct a prognostic risk score model, which provides a novel idea for the TIME characterization and ICB treatment of HCC, contributing to clinical management and decision-making of samples with liver cancer.

MATERIAL AND METHODS

Multiomic Data Collection

Gene expression profiling for HCC sample compared with normal tissues were obtained from the TCGA-LIHC project (**Supplementary Table S6**). The corresponding clinical profiles (**Supplementary Table S7**) were also downloaded from the TCGA portal as described previously. Four categories of somatic mutation data of HCC samples were downloaded from TCGA database (https://portal.gdc.cancer.gov/). We singled out the mutation data files which were obtained through the "SomaticSniper variant aggregation and masking" platform for subsequent analysis (**Supplementary Material** in MAF form). We prepared the

Mutation Annotation Format (MAF) of somatic variants and implemented the "maftools" (Mayakonda et al., 2018) R package which provides a multitude of analysis modules to perform the visualization process. HCC samples were randomly divided into the training set and verification set at a ratio of 1:1. The clinical characteristics of samples within and across groups were similar. All data were obtained from the TCGA public database, and therefore, there was no need for ethics committee approval.

Patient Data and Tissue Specimens

Five pairs of HCC tissues and adjacent liver tissues were acquired from samples that underwent surgical resection. Corresponding adjacent tissues were harvested 3 cm from the edges of the tumor lesion. Tissue specimens were immediately put into liquid nitrogen postoperation. The tissues were then stored in a −80°C refrigerator for total RNA extraction. To control the potential confounding factors, all samples were diagnosed with HCC by histopathological examination, while the samples that received chemotherapy or radiotherapy were excluded from the study. All participants have signed the written informed consent form.

Glycolysis-Related Long Noncoding RNAs

RNA sequencing data of HCC samples were obtained from the TCGA-LIHC project, and noncoding genes were identified according to RefSeq IDs or Ensembl IDs. LncRNAs were retained with reference to NetAffx Annotation files. Glycolysis-related genes were obtained from the gene set "HALLMARK_GLYCOLYSIS" in Molecular Signatures Database (MsigDB) (Liberzon et al., 2015). Pearson correlation analysis was performed on the acquired lncRNAs, as well as glycolysis-related genes. When the correlation coefficient $|R| > 0.4$ and $p < 0.005$, the two genes were considered to be related. The obtained lncRNA was regarded as glycolysis-related lncRNA. Then, it was visualized using Cytoscape. The processing flow of the data conforms to the relevant policies of NIH TCGA human subject protection.

Prognostic Risk Score Calculation

Using the training set, we conducted a univariate Cox proportional hazard regression analysis, LASSO regression analysis, and two-step multivariate Cox proportional hazard regression analysis on the glycolysis-related lncRNAs. Finally, we selected four glycolysis-related lncRNAs for incorporation into the risk score. The expression of lncRNAs between normal and cancer tissues was compared. The regression coefficient β of multivariate Cox regression model and lncRNA expression were used to construct risk score formula as follows:

$$\begin{aligned}\text{Risk score} = {} & \beta\,\text{lncRNA1} \times \text{LncRNA1 Expression} + \beta\,\text{lncRNA2} \\ & \times \text{lncRNA2 Expression} + \cdots + \beta\,\text{lncRNA}\,n \\ & \times \text{lncRNA}\,n\,\text{Expression}.\end{aligned}$$

Prognostic Characteristics of Risk Score

Using the training set, validation set, and all samples, we sorted the samples according to the size of the risk score. The samples were divided into high- or low-risk groups depending on the average risk score. Additionally, we drew the lncRNA expression heat map, risk score distribution map, and risk score and survival relationship map. The Kaplan–Meier method was utilized to draw the survival curve and ROC curve of high- and low-risk samples. In order to determine whether the risk score is an independent prognostic factor, the univariate and multivariate Cox regression analysis was conducted on the risk score and some clinical indicators.

Nomograph Drawing

In order to construct a quantitative scoring system for prognostic evaluation of HCC samples, a ROC curve was drawn with risk score and partial clinical features. Furthermore, the appropriate indicators were selected to construct a nomogram. Subsequently, we analyzed the calibration curve which showed the prognostic value of as-constructed nomogram.

Enrichment Analysis of Gene Set Enrichment Analysis

We utilized the "h.all.v7.2. symbols.gmt [cancer hallmarks]" and "c2. cp.kegg.v7.2. symbols.gmt [Curated]" gene sets from the MsigDB of the GSEA (version 4.0) to analyze the risk score and explore the possible cellular pathways.

Assessment of Correlation of Risk Score With Tumor Immune Environment Characterization

To distinguish TIME difference between low-/high-risk subgroups, we employed several analyses as follows. R package "ESTIMATE" was utilized to estimate tumor purity and the extent and level of infiltrating cells (stromal cell and immune cell), which reflected the characteristics of tumor immune microenvironment. Subsequently, single-sample gene set enrichment analysis was conducted via the R package "GSEAbase" to elucidate the enrichment of 29 immune function–related gene sets. The subpopulation of 22 immune cells in each tumor sample was explored through immune cell subtype identification by using CIBERSORT (https://cibersort.stanford.edu/). Furthermore, we compared the expression levels of 46 immune checkpoint blockade–related genes, (i.e. CD274, etc.) between low-risk samples and high-risk samples.

Assessment of Correlation of Signature With Tumor Immune Infiltration

Immune infiltration information contains each tumor sample's immune cell fraction (i.e. B cells, CD4+T-cells, CD8+T-cells, dendritic cells, macrophages, and neutrophils), which were obtained from Tumor Immune Estimation Resource (TIMER) (https://cistrome.shinyapps.io/timer/). The correlation of tumor immune cell infiltrating with prognostic risk signature was analyzed to explore whether risk signature could act as a novel and reliable indicator in tumor of immune microenvironment of HCC.

Assessment of Role of Risk Signature in Immune Checkpoint Blockade Treatment

Based on reported researches, immune checkpoint blockade key targets expression level might be closely associated with clinical outcome of immune checkpoint inhibitors (Goodman et al., 2017). Herein, we selected six key genes of immune checkpoint blockade–related genes: programmed death ligand 1 (PD-L1, namely CD274), programmed death ligand 2 (PD-L2, namely PDCD1LG2), programmed death 1 (PD-1, namely PDCD1), cytotoxic T-lymphocyte antigen 4 (CTLA-4), indoleamine 2,3-dioxygenase 1 (IDO1), and T-cell immunoglobulin domain and mucin domain-containing molecule-3 (TIM-3, namely HAVCR2) in HCC (Kim et al., 2017; Nishino et al., 2017; Zhai et al., 2018). To investigate the potential role of lncRNA-based signature in ICB therapy of HCC, we correlated risk signature with expression level of six immune checkpoint blockade key targets.

Cell Lines and Culture

One human normal hepatocyte cell line (HL-7702) and two human HCC cell lines (HepG2 and MHCC97H) were cultured in Dulbecco's Modified Eagle Medium (DMEM, Gibco, United States) containing 10% fetal bovine serum (FBS, Gibco, United States) in a humidified atmosphere at 37°C, containing 5% CO_2.

Quantitative Real-Time PCR

For specific qPCR steps, please refer to previous literature (Zhang et al., 2016). The primer sequences used in this study were as follows: MIR4435-2HG forward, 5′-GACTCTCCTACTGGTGCTTGGT-3′ and reverse 5′-CACTGCCTGGTGAGCCTGTT-3′; glyceraldehyde-3-phosphate dehydrogenase (GAPDH) forward, 5′-CAGGAGGCATTGCTGATGAT-3′ and reverse 5′-GAAGGCTGGGGCTCATTT-3′. The relative gene expression levels were calculated by normalizing to GAPDH.

Statistical Analysis

Statistical analysis was performed by R software (version 4.0.2; R Foundation). Comparisons between multiple groups were analyzed using a one-way analysis of variance (ANOVA) and comparisons between the two groups were analyzed by Student's t-test. Construction of the glycolysis-related lncRNA co-expression network was carried out with Cytoscape software (version 3.7.2; The Cytoscape Consortium). $p < 0.05$ was considered as significant difference.

RESULTS

Multiple lncRNAs Are Associated With Glycolysis-Related Genes

Overall, 14,142 lncRNAs were identified using the TCGA-LIHC database, and glycolysis-related genes were identified using the Molecular Signatures Database. To identify glycolysis-related lncRNAs, Pearson's correlation test was performed. lncRNAs with Pearson's correlation coefficient with an absolute value of >0.4 and $p < 0.005$ were set for further analysis. Finally, 1,699 glycolysis-related lncRNAs were obtained (**Supplementary Table S1**).

TABLE 1 | Baseline data of all HCC samples.

Characteristic	Type	n	Proportion (%)
Age	≤65	235	62.33
	>65	141	37.40
	Unknown	1	0.27
Gender	Female	122	32.36
	Male	255	67.64
Grade	G1-2	235	62.33
	G3-4	137	36.34
	Unknown	5	1.33
Stage	Stage I–II	262	69.50
	Stage III–IV	91	24.14
	Unknown	24	6.37
T Stage	T1–2	280	74.27
	T3–4	94	24.93
	Unknown	3	0.80
M Stage	M0	272	72.15
	M1	4	1.06
	Unknown	101	26.79
N stage	N0	257	68.17
	N1	4	1.06
	Unknown	116	30.77

LASSO Regression Analysis Was Able to Accurately Identify Long Noncoding RNAs With Prognostic Value

According to the process shown in **Supplementary Figure S1**, 377 HCC samples were obtained using the TCGA database, and seven samples with incomplete information were excluded from the study. In total, 370 samples were selected for further research. The basic clinicopathological information of samples is shown in **Table 1**. A detailed description was recorded in **Supplementary Table S7**. A total of 22 glycolysis-related lncRNAs were identified using univariate Cox analysis, with results shown in **Supplementary Table S4**. In order to exclude the overfitting, LASSO regression analysis was conducted on 22 lncRNAs, and a total of five glycolysis-related lncRNAs were identified. The screening process and results are shown in **Figures 1A,B**, and **Supplementary Table S5**. These five lncRNAs were analyzed using a two-step multivariate Cox regression analysis. Finally, four glycolysis-related lncRNAs were found to be associated with prognosis of HCC samples (**Figure 1C**). Among them, AL031985.3, AL365203.2, and MIR4435-2HG were found to be poor prognostic factors (hazard ratio, HR > 1), and their expression was upregulated in HCC samples. On the other hand, AC015908.3 was a protective factor (HR < 1), and its expression was found to be decreased in HCC samples. The results are shown in **Figures 1D–G** and **Table 2**. Four lncRNAs were used to construct the co-expression network, the results of which are shown in **Supplementary Figures S1B,C**. According to expression of lncRNAs and multivariate Cox regression coefficient, the prognosis risk score of

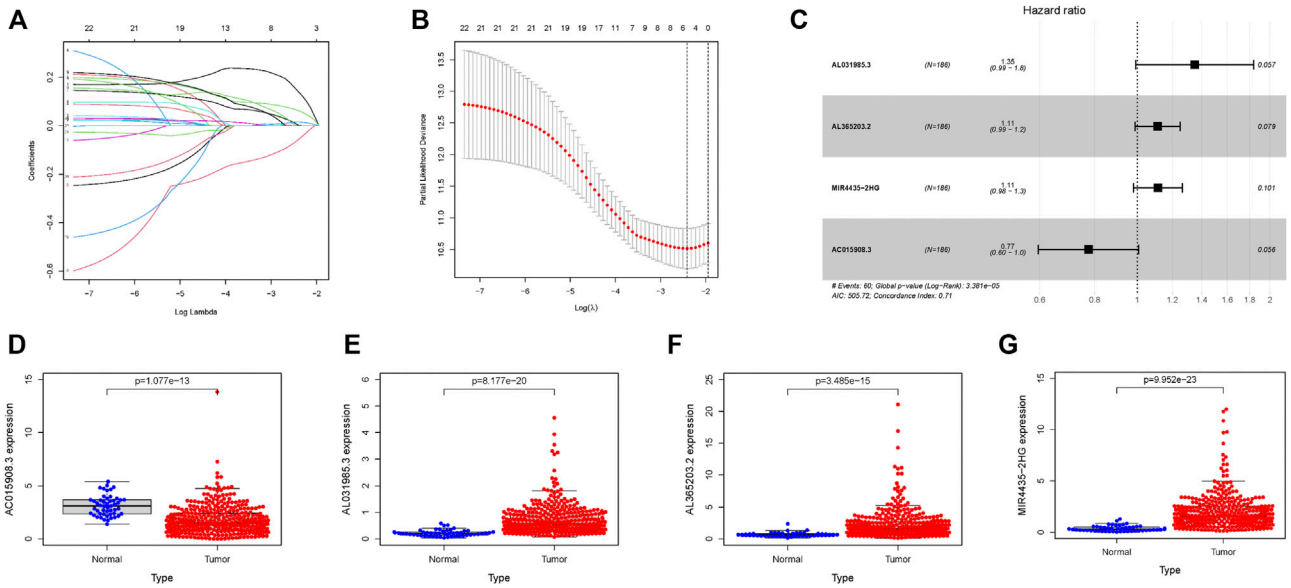


FIGURE 1 | Four glycolysis-related lncRNAs with prognostic value in the training set. **(A)** Plots for LASSO expression coefficients of 22 glycolysis-related lncRNAs. **(B)** Cross-validation plot for the penalty term. **(C)** Relationship between four glycolysis-related lncRNAs and prognosis of HCC patients **(D–G)** Expression of four glycolysis-related lncRNAs in tumors and normal tissues; the data come from TCGA database, where all *p* values < 0.05.

TABLE 2 | Multivariate Cox results of lncRNAs based on TCGA-LIHC data.

Id	Coef	HR	HR.95 L	HR.95H	*p* value
AL031985.3	0.299,987	1.349,841	0.991,382	1.837,909	0.05678
AL365203.2	0.105,369	1.111,121	0.987,831	1.249,799	0.079101
"MIR4435-2HG"	0.107,428	1.113,411	0.979,232	1.265,977	0.101,078
AC015908.3	-0.25568	0.774,388	0.595,609	1.006829	0.056244

glycolysis-related lncRNAs was calculated as follows (0.299987 × AL031985.3 expression) + (0.105369 × AL365203.2 expression) + (0.107428 × MIR4435-2HG expression) − (0.25568×AC015908.3 expression). Samples were equally and randomly divided into training set and verification set, including 186 cases in the training set and 184 cases in the verification set. The results of random grouping are shown in **Supplementary Tables S2, S3**.

The Risk Score Is Significantly Related to Patient Prognosis

According to this scoring system, the prognostic risk score of each patient was calculated and samples were arranged from left to right according to their score level. The heat map distribution of four lncRNAs is shown in **Figure 2A**. With increasing risk score, the number of surviving samples decreased and the amounts of dead samples increased. The prognosis of samples in the low-risk group was significantly better than that in the high-risk group (**Figures 2B,C**). The Kaplan–Meier survival curve shows that the 5-year survival rate of samples in the low-risk group is significantly higher than that in the high-risk group (**Figure 2D**, $p = 3.819e - 05$). Moreover, these four lncRNAs were used to construct a prognosis scoring system with high accuracy (**Figure 2E**, AUC = 0.763). Consistent with these results, univariate and multivariate Cox regression analysis showed that the increased risk score indicates the higher the risk score, the poorer the prognosis (**Figures 2F,G**).

Validation of Prognostic Risk Score

The risk scoring system was validated using an internal validation set, as well as all samples. The four lncRNAs had similar distributions in the heat map, as well as risk score distribution (**Figures 3A,B**; **Supplementary Figure S2A,B**). The higher the risk score, the fewer samples survived and the more deaths that occurred (**Figure 3C**; **Supplementary Figure S2C**). The 5-year survival rate in the low-risk group was significantly higher (**Figure 3D**; **Supplementary Figure S2D**). The risk scoring system in the validation set, as well as overall samples, has the same degree of predictive accuracy as the training set (**Figure 3E**; **Supplementary Figure S2E**). Consistent with results from the training set, a risk score can be used as an independent prognostic factor to judge patient

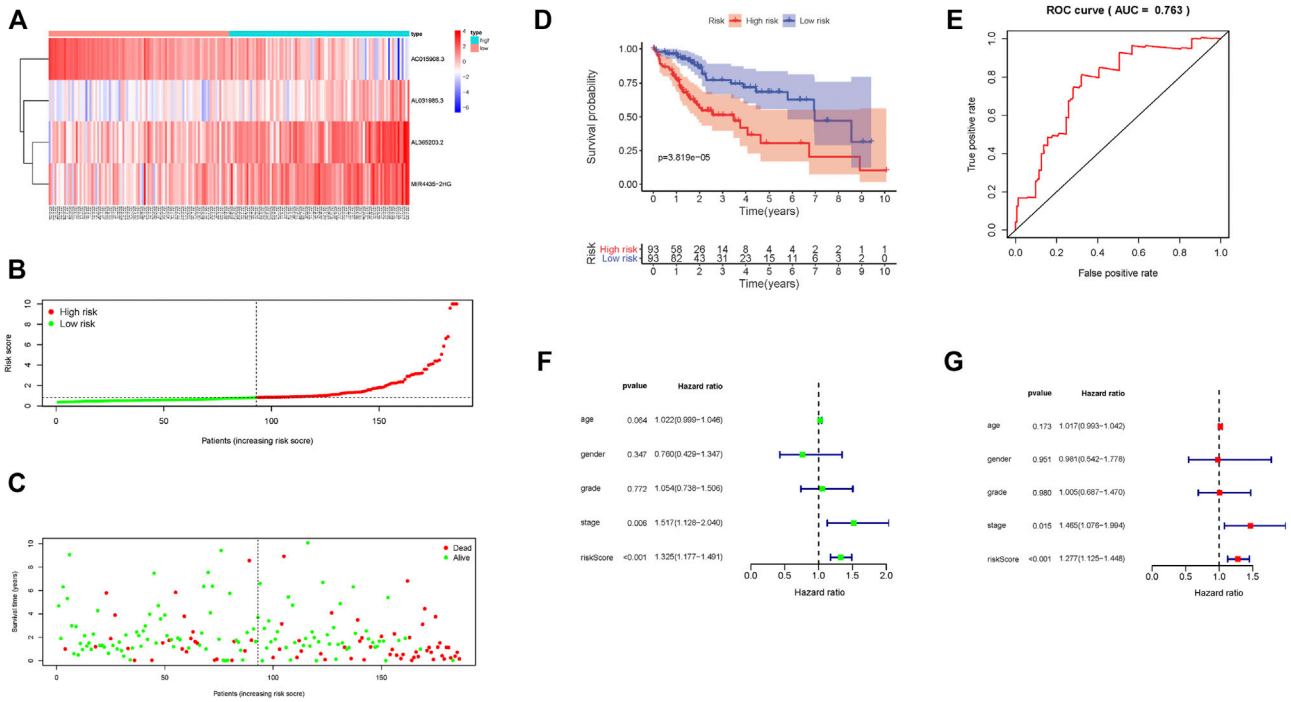


FIGURE 2 | Prognostic risk score characteristics of glycolysis-associated lncRNAs in the training set. **(A)** Heat map of the expression of four lncRNAs in HCC samples. The color from green to red indicates a trend from low expression to high expression. **(B)**. Distribution of risk scores for HCC samples. **(C)** The relationship between survival time and status of HCC samples and risk score. **(D)** Kaplan–Meier survival curve of samples in high- and low-risk groups. **(E)** ROC curve of risk score in samples with HCC. **(F)** Univariate Cox regression analysis of clinical features and risk score in HCC samples. **(G)** Multivariate Cox regression analysis of clinical features and risk score in HCC samples.

prognosis. The higher the risk score, the worse the prognosis (**Figures 3F,G**; **Supplementary Figures S2F,G**), the more serious the tumor grade (**Figure 3H**).

Close Correlation of Risk Score With Tumor Immune Environment Characterization of Hepatocellular Carcinoma

To further uncover the potential role of prognostic risk score in TIME of HCC, we investigated the relationship between risk score and immune-related score (calculated with the R package "ESTIMATE"), immune signature (via ssGSEA analysis) and Tumorinitiating cell subtypes and level (assessed by CIBERSORT method), and the 46 immune checkpoint blockade–related genes expression level.

These results indicated that samples with low risk had a higher estimate score, stromal score, immune score but lower tumor purity compared with high-risk samples (**Figures 4A–D**). Then, we examined whether there was distinction of immune signatures between groups low/high risk. From the ssGSEA results, we found that the infiltrating levels of aDCs, DCs, iDCs, macrophages, pDCs, Tfh, Th1 cells, Th2 cells, and Tregs were remarkably elevated and some immune signatures (i.e. APC costimulation, checkpoint, parainflammation, HLA molecule, IFN response type II, and MCH class I) were significantly activated with the increased risk score (**Figure 4E**; **Supplementary Figure S3A**). **Supplementary Figure S3B** shows each patient immune-related signature with corresponding immune-related scores in groups low/high risk. The CIBERSORT analysis results pointed out that the more the fraction of regulatory T cells, the higher the risk score (**Figure 4F**). Further correlation analysis presented that 40 of 46 (i.e. CD274, IDO1, etc.) immune check blockade–related genes expression levels were significantly different between two risk groups (**Figure 4G**). These results suggested that lncRNA-based risk signature may contribute a novel insight into TIME feature and immune response of HCC.

The Predictive Power of Risk Score was Significantly Better Compared to Other Clinical Characteristics

The prognostic risk score, combined with age, gender, and tumor grade and stage, were used to draw ROC survival curve. The results indicated that compared to other clinical traits, the glycolysis-related lncRNA prognostic risk scoring system was more accurate at predicting the 1-, 3-, and 5-year survival rate of HCC samples (**Figures 5A–C**, AUC = 0.747, 0.660, and 0.656, respectively). The prognostic factors with AUC >0.6 were identified in ROC curve, and the nomogram was drawn. The results are shown in **Figure 5D**. The 1-, 3-, and 5-year survival rates were calculated quantitatively according to the tumor stage and risk score. We corroborated that our nomograph had great prognostic predictive performance of 1-, 2-, and 3-year survival time by employing calibrate curves (**Figures 5E–G**).

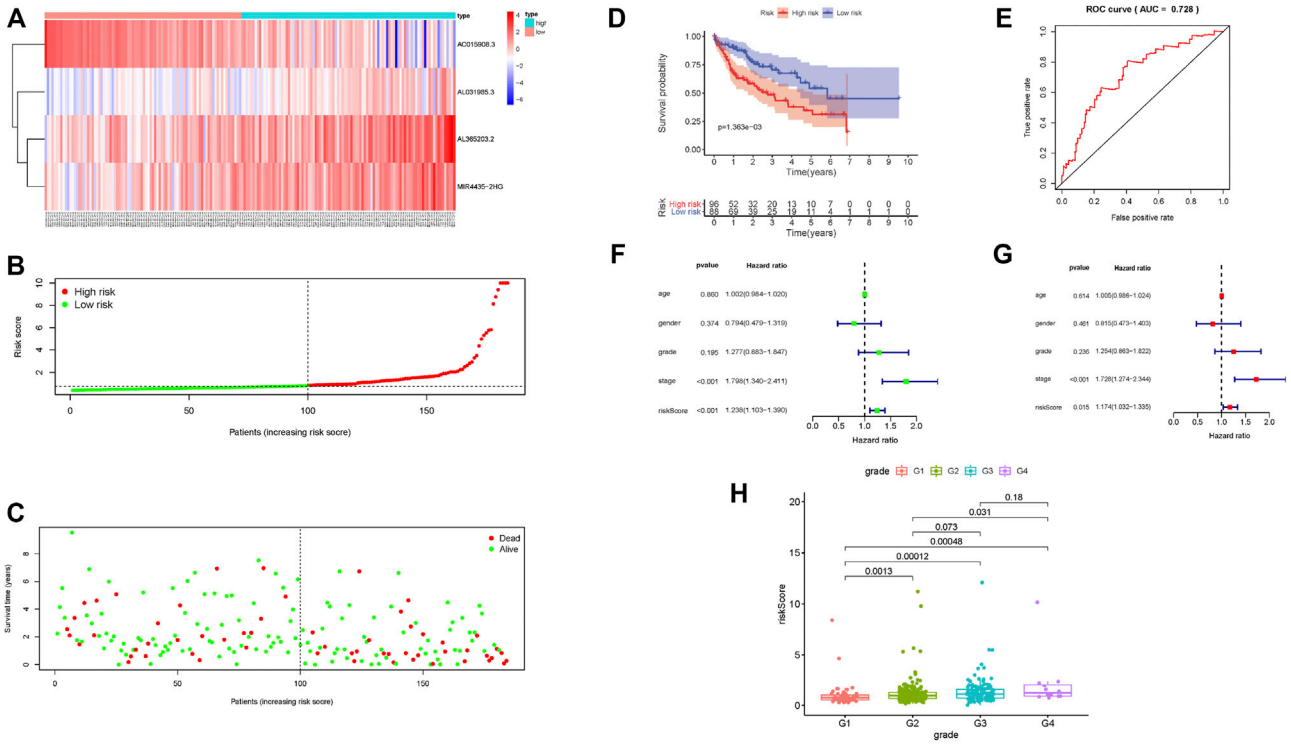


FIGURE 3 | Prognostic risk score characteristics of glycolysis-related lncRNA in the validation set. **(A–C)** Heat map of the expression of four glycolysis-related lncRNAs in HCC samples, distribution map of risk score, relationship map of survival status and risk score. **(D–E)** Kaplan–Meier survival curve and ROC curve of high- and low-risk group. **(F–G)**. Univariate and multivariate Cox regression analysis of clinical features and risk score in HCC samples. **(H)** Relationship between tumor grade and risk score; risk score significantly increased for advanced grade cases.

To validate whether lncRNA risk signature remained with excellent prognostic predictive performance in different clinicopathological subgroups, furthermore, we performed a stratification analysis. Regardless of young or old, the risk signature could further distinguish low-risk group and high-risk group with significantly distinct survival time (**Supplementary Figures S5A,B**). Likewise, risk signature presented powerful prognosis prediction ability for samples in grade 1–2 or 3–4 (**Supplementary Figures S5C,D**), early stage or late stage (**Supplementary Figures S5E,F**), T status one to two or 3–4 (**Supplementary Figures S5G,H**), N0 status (**Supplementary Figure S5I**), M0 status (**Supplementary Figures S5J**) ,and male gender (**Supplementary Figure S5K**). We found that *p*-value was 0.081, however, female samples' survival time shortened with the increase of risk score (**Supplementary Figure S5L**). These results suggested that it can be an outstanding predictor in samples with HCC.

Risk Score Affects the Results of Gene Enrichment

Hallmark enrichment analysis indicated that apoptosis and glycolysis were active in high-risk group, while being silent in the low-risk group. Additionally, multiple pathways, including IL/STAT5 and NOTCH, were active in the high-risk group and silent in the low-risk group (**Supplementary Figure S4A**). Finally, Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis suggested that bladder cancer and colorectal cancer were active in the high-risk group but silent in the low-risk group (**Supplementary Figure S4B**).

Correlation of Risk Signature With Infiltrating Immune Cell and Immune Checkpoint Blockade Key Molecules

To further explore the influence of lncRNA-based signature upon TIME of HCC, we analyzed correlation of risk signature with immune cell infiltration type and level. We observed that the risk signature significantly correlated with infiltrating B cells (r = 0.191; p = 2.171e − 04), infiltrating CD4+T cells (r = 0.212; p = 3.918e − 05), infiltrating CD8+T cells (r = 0.305; p = 2.139e − 09), infiltrating dendritic cells (r = 0.361; p = 8.239e − 13), infiltrating macrophages (r = 0.411; p = 1.665e − 16), and infiltrating neutrophils (r = 0.353; p = 2.856e − 12; **Figures 6A–F**). These results suggested that prognostic risk signature was closely correlated with immune infiltration in HCC.

Next, we singled out six key immune checkpoint inhibitor genes (PDCD1, CD274, PDCD1LG2, CTLA-4, HAVCR2, and IDO1) for further research (Vidyasagar, 2015; Chen et al., 2018; Bejani and Ghatee, 2020). We performed the correlation analysis of ICB key gene expression with risk signature to investigate the potential role of signature in the ICB therapy of HCC

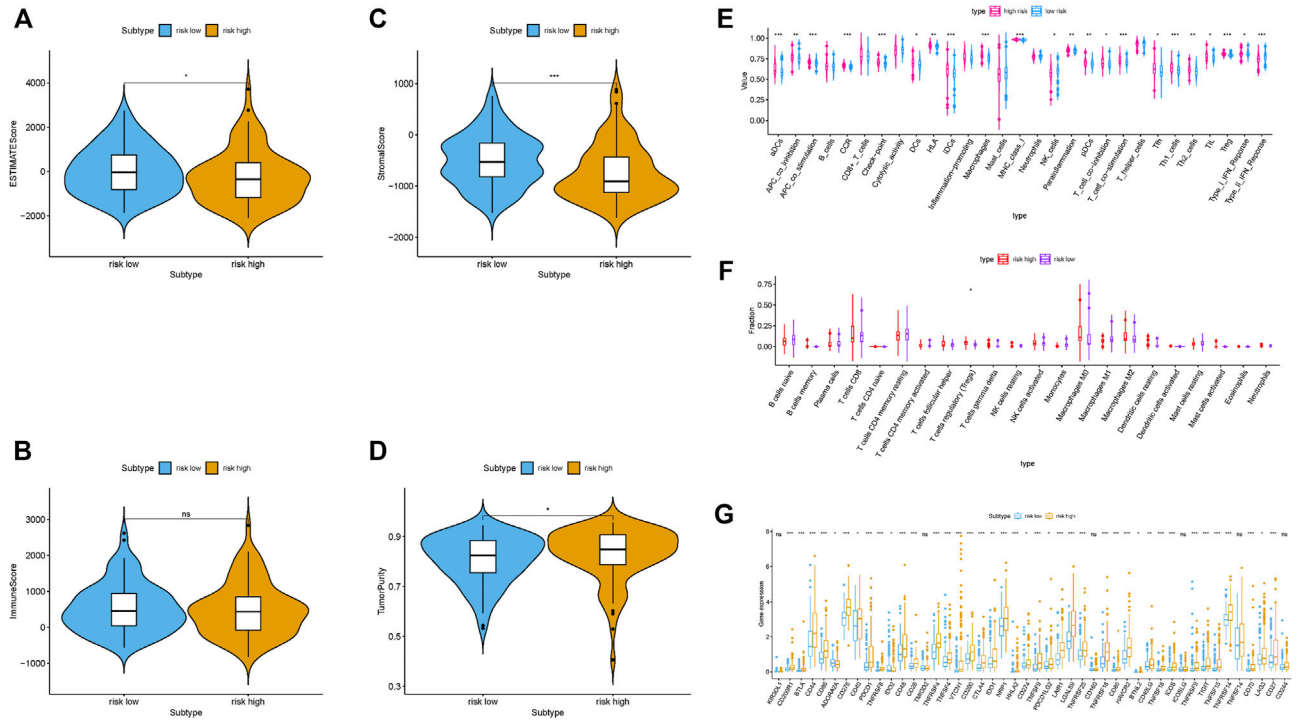


FIGURE 4 | Correlation of prognostic risk score with TIME characterization of HCC **(A–D)** Comparison of estimate score, stromal score, immune score, and tumor purity between these two subtypes. **(E)** Distinction of enrichment of immune-related signatures between low- and high-risk groups. **(F)** Difference of infiltrating immune cell subpopulations and levels between low-/high-risk groups. **(G)** Comparison of 46 immune checkpoint blockade–related genes expression levels in two risk score subgroups.

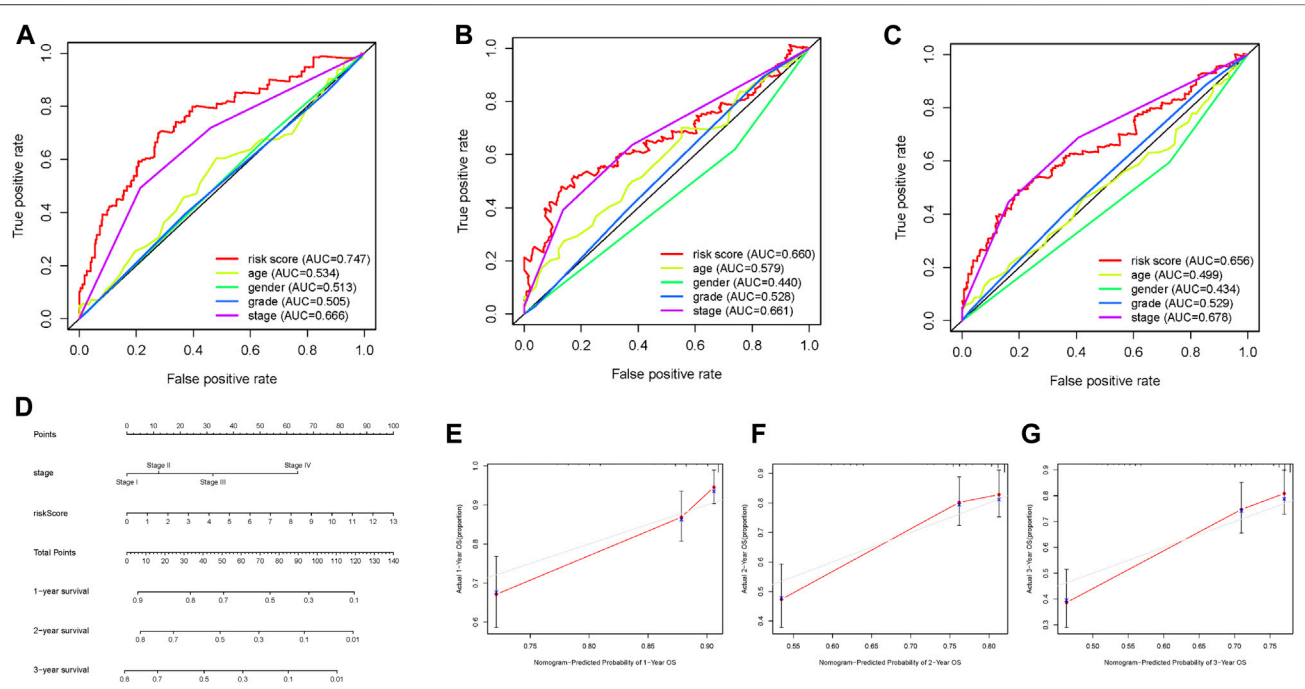


FIGURE 5 | Screening prognostic indicators and nomogram. **(A–C)** ROC curve of 1-, 2-, and 3-year OS for multiple prognostic indicators of HCC samples. **(D)** The nomogram was drawn using tumor grade and risk score. **(E)** One-year nomogram calibration curves of entire TCGA cohort. **(F)** Two-year nomogram calibration curves of entire TCGA cohort. **(G)** Three-year nomogram calibration curves of entire TCGA cohort.

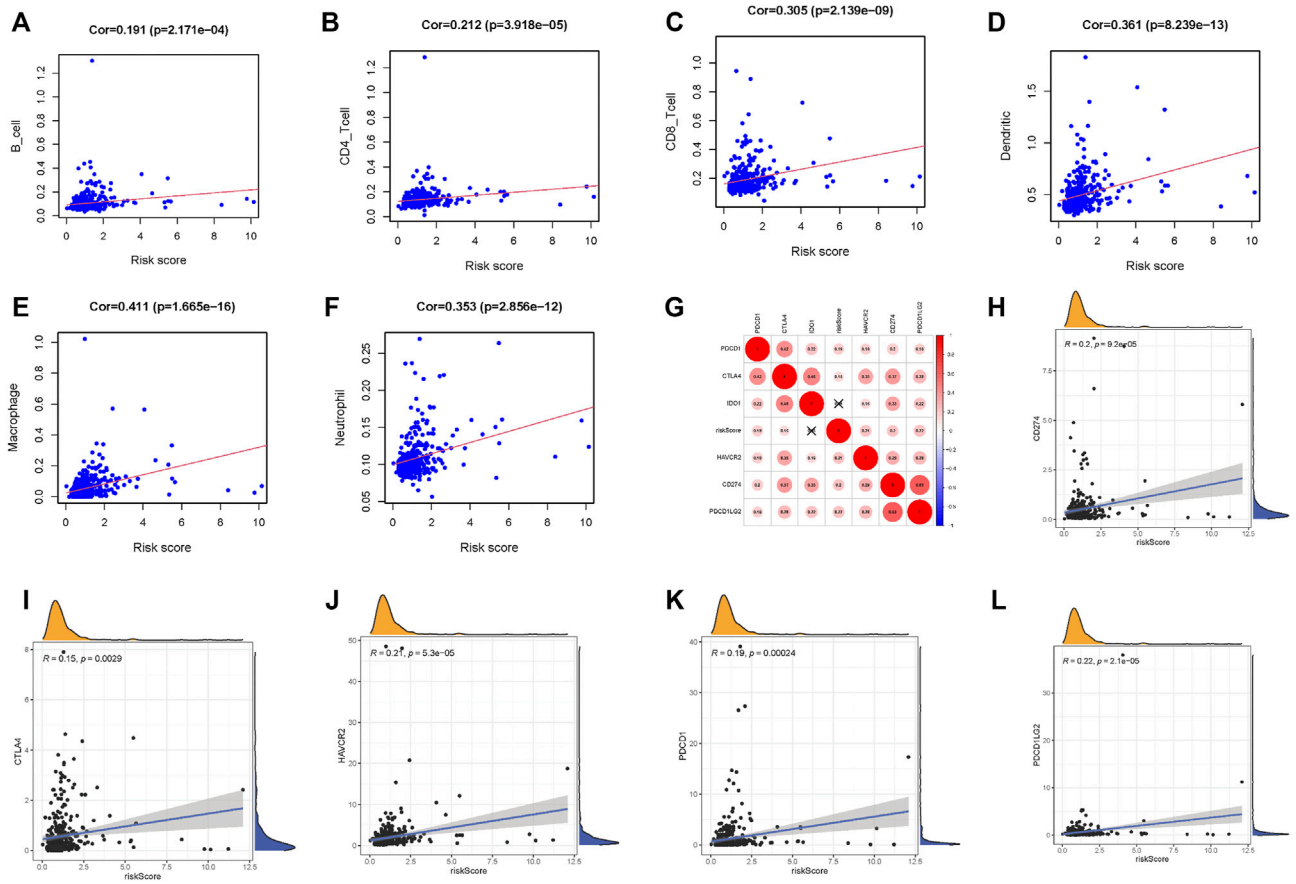


FIGURE 6 | Correlation between tumor immune infiltration and this immune-related lncRNA signature. **(A)** Association between this signature and B cells. **(B)** Association between this signature and CD4$^+$ T cells. **(C)** Association between this signature and CD8$^+$ T cells. **(D)** Association between this signature and dendritic cells. **(E)** Association between this signature and macrophages cells. **(F)** Association between this signature and neutrophil cells. **(G)** Association analyses between immune checkpoint inhibitors CD274, PDCD1, PDCD1LG2, CTLA4, HAVCR2, and IDO1 and lncRNAs signature. **(H)** Association between our risk model and CD274. **(I)** Association between our risk model and CTLA4. **(J)** Association between our risk model and HAVCR2. **(K)** Association between our risk model and PDCD1. **(L)** Association between our risk model and PDCD1LG2.

(**Figure 6G**). The analysis result pointed out that risk signature had close relationship with CD274 ($r = 0.2$; $p = 9.2e-05$), CTLA4 ($r = 0.15$; $p = 0.0029$), HAVCR2 ($r = 0.21$; $p = 5.3e-05$), PDCD1 ($r = 0.19$; $p = 0.00024$), and PDCD1LG2 ($r = 0.22$; $p = 2.1e-05$; **Figures 6H–L**), indicating risk signature might exert a nonnegligible player in ICB treatment outcome prediction in HCC.

High Expression of MIR4435-2HG in Hepatocellular Carcinoma Suggests Poor Prognosis

We evaluated the expression of MIR4435-2HG in cell lines and tissues. The results showed that in comparison to normal liver cell lines, the expression of MIR4435-2HG in hepatoma cell lines was significantly increased (**Figure 7A**, $p < 0.05$). Likewise, MIR4435-2HG was upregulated in tumor tissue relative to normal samples. Limited by number of samples, we observed no statistical difference (**Figure 7B**). Consistent with the results of *in vitro* experiments, the OS of samples with low expression of MIR4435-2HG was significantly longer than that of samples with high expression (**Figure 7C**, $p = 0.0018$), suggesting that MIR4435-2HG is a poor prognostic factor for HCC samples. The expression level analysis among major clinical stages shown that MIR4435-2HG expressed significantly differently among distinct clinicopathological stages (**Figures 7D**, F value = 5.48 and $p = 0.0011$).

MIR4435-2HG Correlates With Immune Checkpoint Blockade Therapy Key Genes in Hepatocellular Carcinoma

Then we analyzed the correlation between the MIR4435-2HG and ICB-related key genes to elucidate the impact of MIR4435-2HG on the ICB therapy of HCC. The results presented that MIR4435-2HG was significantly positively correlated to CD274 ($r = 0.12$; $p = 0.014$), CTLA4 ($r = 0.27$; $p = 1.4e-08$), HAVCR2 ($r = 0.19$; $p = 6.4e-05$), IDO1 ($r = 0.13$; $p = 0.0075$), PDCD1 ($r = 0.13$; $p = 0.0086$), and PDCD1LG2 ($r = 0.16$; $p = 0.001$; **Figures 7E–J**), suggesting MIR4435-2HG may be a novel and potential target in ICB treatment in HCC.

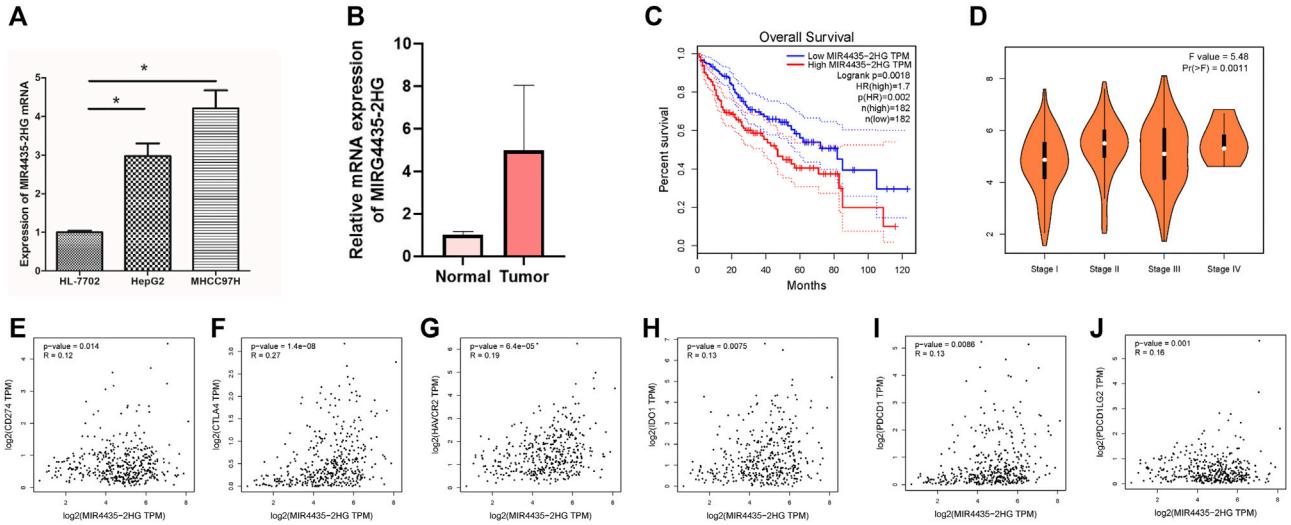


FIGURE 7 | High expression of MIR4435-2HG indicates poor prognosis. **(A)**. qPCR results indicate that MIR4435-2HG is highly expressed in HepG2 and MHCC97H cell lines, $^*p < 0.05$. Each experiment is repeated at least three times. **(B)** qPCR results indicate that MIR4435-2HG is highly expressed in tumor tissue. **(C)** The prognosis of samples with high or low expression of MIR4435-2HG is significantly different. **(D)** The expression of MIR4435-2HG had significant difference between major pathological stages. **(E)** The mRNA expression between MIR4435-2HG and CD274 had more similar pattern in HCC and normal tissues. **(F)** The mRNA expression between MIR4435-2HG and CTLA4 had more similar pattern in HCC and normal tissues. **(G)** The mRNA expression between MIR4435-2HG and HAVCR2 had more similar pattern in HCC and normal tissues. **(H)** The mRNA expression between MIR4435-2HG and IDO1 had more similar pattern in HCC and normal tissues. **(I)** The mRNA expression between MIR4435-2HG and PDCD1 had more similar pattern in HCC and normal tissues. **(J)** The mRNA expression between MIR4435-2HG and PDCD1LG2 had more similar pattern in HCC and normal tissues.

Role of MIR4435-2HG in Tumor Immune Environment Characterization

To further examine whether MIR4435-2HG can act as immune indicators, we performed the correlation analysis of MIR4435-2HG expression level with immune infiltration. HCC samples were classified into high/low MIR4435-2HG subtypes based on the median MIR4435-2HG expression value. ESTIMATE results indicated that samples with higher MIR4435-2HG expression had a significant higher stromal score, immune score, and ESTIMATE score but lower tumor purity relative to samples in high MIR4435-2HG group (**Figures 8A,B**). Subsequently, we identified difference of enrichment in immune-related signatures between two different subgroups. The subjects in MIR4435-2HG high group remarked as high infiltration of aDCs, DCs, iDCs, pDCs, macrophages, Tfh, Th1 cells, Th2 cells, and Tregs and enrichment of T cell costimulation, APC costimulation, CCR, checkpoint, HLA, inflammation promoting, parainflammation, and class I MHC, which suggested immune-activated phenotype (**Figure 8C**). The CIBERSORT result presented that expression level of MIR4435-2HG was positively correlated with M0 and M2 macrophage infiltration, whereas negatively correlated with plasma cells, CD8 T cells, and Tfhs (**Figure 8D**). In summary, these results pointed out that MIR4435-2HG may serve as a key indicator in TIME characterization and immunological reaction in HCC.

Correlation of Mutation of TP53 With Risk Score

Based on previous research (Ruan et al., 2016), CTSB played a pivotal role in HCC initiation and progression. According to results of somatic mutation data, TP53 was the genes with highest mutation frequency (**Supplementary Figure S6A**). Thus, we proposed to uncover the role of gene mutation in risk score and analyzed the proportion of mutation gene in both low- and high-risk groups. We observed that mutation of TP53 was significantly correlated with risk score (**Figures 8E,F**; **Supplementary Figure S6B**; training set, testing set, and whole cohort, respectively), whereas mutation of CTSB was similar between low- and high-risk groups (**Supplementary Figure S6C**). These results indicated that mutation of TP53 may contribute to HCC development.

DISCUSSION

The pathogenesis of HCC is very complex as it involves cell cycle and apoptosis, transcriptional regulation disorder (Lin et al., 2014), and energy metabolism abnormalities (Hsu et al., 2015). LncRNAs affect tumorigenesis and development in many ways, including regulating cell proliferation and migration (Shen et al., 2019), influencing epigenetic regulation (Miao et al., 2019) and regulating energy metabolism rate-limiting enzymes. Glycolysis is an inefficient method of energy production, but this process produces a reduction equivalent (Terabe et al., 2019) and biosynthetic substrate necessary for tumor cell proliferation (Liang et al., 2019). In this study, we obtained clinical and transcriptomic data of HCC from the TCGA database and successively applied univariate Cox analysis, LASSO analysis, and two-step multivariate Cox analysis to identify glycolysis-related lncRNAs. Additionally, abnormal energy metabolism and

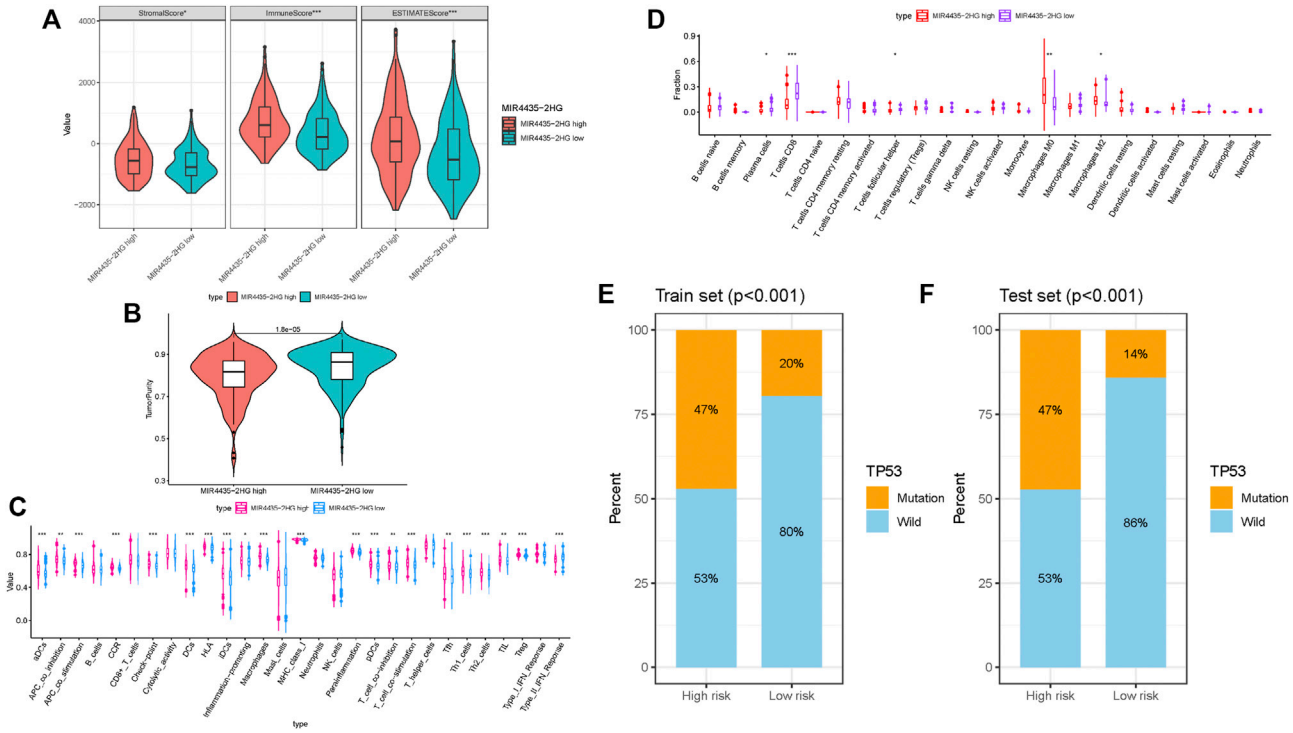


FIGURE 8 | Correlation of MIR4435-2HG in TIME characterization. **(A–B)** Comparison of the immune score (ESTIMATE algorithm) between MIR4435-2HG low/high groups. **(C)** Difference of immune-related signatures between low and high MIR4435-2HG subgroups. **(D)** Distinction of infiltrating immune cell subpopulations and levels between low/high MIR4435-2HG groups. **(E)** Proportion of mutation of TP53 in both low-/high-risk groups form the training set. **(F)** Proportion of mutation of TP53 in both low-/high-risk groups form the testing set.

lncRNAs were combined to construct a risk score model with prognostic value. The model was verified across different groups so that the prognostic judgment of HCC could be quantified and specific and provides guidance for survival prediction of samples.

When selecting specific variables to build a model, there is often overfitting (Dawes et al., 2019). This problem often occurs when there are too many variables. With regard to human genes, only 2% can encode proteins, and 98% of them are noncoding sequences, which constitute a complex regulatory network (Boon et al., 2016). In our study, we observe that there are still 22 lncRNAs that are related to the prognosis of samples after screening by univariate COX analysis, and excessive lncRNAs involved in constructing can cause the risk scoring model to lead to overfitting. An important method to solve overfitting is regularization (Bejani and Ghatee, 2020). LASSO regression constructs a penalty function and adds L1 regularization after the loss function to obtain a more accurate model with fewer variables (Vidyasagar, 2015). After LASSO regression analysis of 22 lncRNAs, only five were found to be related to patient prognosis. Even after two-step multivariate Cox regression, only one lncRNA was identified. The final remaining four lncRNAs indicated high accuracy in the validation set, as well as overall prognosis for samples.

The ROC curves of OS of samples with liver cancer were constructed by combining several clinical characteristics of samples with a prognostic risk score. Indicators with AUC >0.6 were selected to draw a nomogram, which made the judgment of survival rate of samples with liver cancer visualized and more specific. From our results, we are able to see that the risk score based on glycolysis-related lncRNA construction shows high accuracy in predicting the survival rate of samples. The reason is that abnormal energy metabolism plays an important role in metabolomics and epigenetics of liver tumors, and glycolysis-related pathways are significantly related to survival and prognosis of samples (Chen et al., 2018). Furthermore, 90% of energy in normal tissues comes from tricarboxylic acid cycle in mitochondria (Anderson et al., 2018), while more than 50% of the energy depends on glycolysis, which is known as the "Warburg effect" (Pascale et al., 2020). At present, it is believed that the main mechanisms of Warburg effect include mitochondrial dysfunction (Riera Leal et al., 2020), tumor adaptation (Ždralević et al., 2018), microenvironment changes (Sun et al., 2018), oncogene (Banks, 2013), and related signal pathway disorders. According to the results of GSEA enrichment analysis, we found that Notch, p53, Wnt, and other signaling pathways are active in the high-risk group whether we use the Hallmark dataset or KEGG dataset. These pathways are closely related to the recurrence of liver cancer (Invalidcitation, 2018). In addition, we found that glycolysis is active in the high-risk group in the hallmark dataset, which is consistent with our results.

According to published works, we observed that more and more researches focusing on TIME have revealed the potential implication of lncRNAs upon infiltrating immune cells. Peng Lirong et al. reported

that LncRNA MIAT was significantly correlated with immune cell infiltration and may exert an important player in the immune escape of HCC (Peng et al., 2020). The study of Ji Jie et al. demonstrated that Lnc-Tim3 was involved in the survival of the exhausted CD8+T cells and facilitating CD8+T exhaustion (Ji et al., 2018). Consequently, we speculated that the subtype of infiltrating immune cells had close connection with lncRNAs. Herein, we corroborated that lncRNA-based risk signature was significantly correlated with immune cell infiltration, (i.e. macrophages, dendritic cells, neutrophils, B cells, CD4+T cells, and CD8+T cells). ESTIMATE results presented that risk score was negatively correlated with estimate score, stromal score, and immune score but positively with tumor purity, suggesting risk signature could serve as a novel immune indicator in HCC. Besides, ssGSEA analysis pointed out that the infiltrating immune cells (i.e. DCs, macrophages, Th1 cells, and Tregs) were significantly increased and immune signatures (i.e. APC costimulation, checkpoint, parainflammation, IFN response type II, and MCH class I) were remarkably activated when risk score elevated. Finally, CIBERSORT algorithm results showed that risk score elevated when the fraction of regulatory T cells increased, indicating that as-constructed signature works through regulating Tregs infiltration and might have an undeniable role in tumor immune microenvironment of HCC. The immune-activated condition in the high-risk group was associated with high ICB-relevant genes expression, suggesting samples in with low risk score might respond to immunotherapy.

With the emergence of immune checkpoint blockade (ICB) treatment, immune checkpoint inhibitors have considerably transformed clinical decision-making in cancer oncology (Pitt et al., 2016; Llovet et al., 2018; Salik et al., 2020). Immune-checkpoint blockade treatment has contributed a novel insight into clinical management in samples with HCC(Ng et al., 2020). Nevertheless, HCC samples obtained relatively few benefits from ICB therapy and less than one in three samples were observed for objective response to immune checkpoint inhibitors treatment (Liu et al., 2020). Such biomolecules as immune checkpoint blockade–related gene expression level and tumor mutational burden were unable to accurately predict clinical outcome of ICB treatment. It is therefore urgent to identify indicators that can precisely forecast responsiveness to ICB treatment for further individualized treatment and advance precision immunotherapy (Nishino et al., 2017; Ng et al., 2020; Mushtaq et al., 2018). Recently, accumulating evidences have supported that numerous lncRNAs possess key roles in regulating immunity, such as immune cell infiltration, antigen presentation, and so on (Carpenter and Fitzgerald, 2018; Denaro et al., 2019). In this study, the correlation analysis showed that PDCD1, CD274, PDCD1LG2, CTLA-4, IDO1, and HAVCR2 were coexpressed. Furthermore, our risk signature was significantly associated with the ICB treatment key target genes (i.e. PDCD1LG2, PDCD1, CD274, HAVCR2, and CTLA4), and the expression level of immune checkpoint blockade–related genes (i.e. IDO1 and TIGHT) increased significantly with increased risk scores. Due to no ICB treatment dataset in HCC cohort, we were unable to explore the correlation between risk score and ICB immunotherapy response. These findings indicated that our signature may possess the ability to predict clinical outcome of ICB therapy in HCC samples.

It has been reported that MIR4435-2HG is associated with prognosis of HCC (Kong et al., 2019). Overexpression of MIR4435-2HG can promote proliferation of HCC cells, which is consistent with our experimental results. However, previous literature has only described this phenomenon. MIR4435-2HG expression was significantly positively associated with ICB immunotherapy key genes (i.e. CD274, CTLA4, HAVCR2, IDO1, PDCD1LG2, and PDCD1). We also demonstrated that MIR4435-2HG expression had close relationship with high infiltration of immune cells (i.e. macrophages) in HCC. These findings indicated that high MIR4435-2HG expression level was associated with a poor prognosis that could facilitate immune evasion and immunotherapy resistance. Our results first linked the mechanism of MIR4435-2HG with immune infiltration and immunotherapy, which provides a new rationale for further research. However, our experiment lacks verification results of clinical samples and only obtains clinical information from the database in order to verify expression of MIR4435-2HG, which is a limitation in our experiment.

CONCLUSION

In our study, the LASSO regression method helped identify glycolysis-related lncRNAs to construct a risk score model. This model can quantitatively and accurately judge the prognosis of HCC samples. Moreover, as-constructed lncRNAs signature was significantly correlated to not only immune cell infiltration but also responsiveness to ICB treatment key genes in HCC. Conclusively, this research provided a promising avenue to facilitate the individualized survival prediction and reveal landscape of tumor immune environment of HCC, further contributing valuable clinical applications in HCC ICB therapy. Notwithstanding, our findings should be validated in further researches which explore HCC tumorigenesis and progression mechanisms and the implication of these 4 glycolysis-related lncRNAs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/repository?facetTab=files&filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.primary_site%22%2C%22value%22%3A%5B%22liver%20and%20intrahepatic%20bile%20ducts%22%5D%7D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.project.program.name%22%2C%22value%22%3A%5B%22TCGA%22%5D%7D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.project.project_id%22%2C%22value%22%3A%5B%22TCGA-LIHC%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.data_category%22%2C%22value%22%3A%5B%22transcriptome%20profiling%22%5D%7D%7D%2C%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22files.data_type%22%2C%22value

%22%3A%5B%22Gene%20Expression%20Quantification%22%5D%7D%7D%5D%7D.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

SY and YW conceived and designed the study. YB conducted the study. HL and JC contributed to the acquisition of data. YB analyzed and interpreted the data. SY, YW, and YB reviewed and edited the manuscript. All authors read and gave final approval of the manuscript.

FUNDING

This study was supported by the Research Project of Zhejiang Provincial Public Welfare Fund Project in the Field of Social Development (No. LGF20H160028), and the Major Projects of Jinhua Science and Technology Plan Project (No. 2018-3-001a).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.645084/full#supplementary-material.

REFERENCES

Anderson, N. M., Mucka, P., Kern, J. G., and Feng, H. (2018). The emerging role and targetability of the TCA cycle in cancer metabolism. *Protein Cell* 9 (2), 216–237. doi:10.1007/s13238-017-0451-1

Anwanwan, D., Singh, S. K., Singh, S., Saikam, V., and Singh, R. (2020). Challenges in liver cancer and possible treatment approaches. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1873 (1), 188314. doi:10.1016/j.bbcan.2019.188314

Banks, R. E. (2013). Oncogene-induced cellular senescence elicits an anti-Warburg effect. *Proteomics* 13 (17), 2542–2543. doi:10.1002/pmic.201300335

Bejani, M. M., and Ghatee, M. (2020). Theory of adaptive SVD regularization for deep neural networks. *Neural Networks* 128, 33–46. doi:10.1016/j.neunet.2020.04.021

Boon, R. A., Jaé, N., Holdt, L., and Dimmeler, S. (2016). Long noncoding RNAs. *J. Am. Coll. Cardiol.* 67 (10), 1214–1226. doi:10.1016/j.jacc.2015.12.051

Carpenter, S., and Fitzgerald, K. (2018). Cytokines and long noncoding RNAs. *Cold Spring Harbor Perspect. Biol.* 10 (6). doi:10.1101/cshperspect.a028589

Chan, T. A., Yarchoan, M., Jaffee, E., Swanton, C., Quezada, S. A., Stenzinger, A., et al. (2019). Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Ann. Oncol.* 30 (1), 44–56. doi:10.1093/annonc/mdy495

Chen, R., Zhu, S., Fan, X.-G., Wang, H., Lotze, M. T., Zeh, H. J., et al. (2018). High mobility group protein B1 controls liver cancer initiation through yes-associated protein -dependent aerobic glycolysis. *Hepatology* 67 (5), 1823–1841. doi:10.1002/hep.29663

Cheng, H., Sun, G., Chen, H., Li, Y., Han, Z., Li, Y., et al. (2019). Trends in the treatment of advanced hepatocellular carcinoma: immune checkpoint blockade immunotherapy and related combination therapies. *Am. J. Cancer Res.* 9 (8), 1536–1545.

Dawes, A. J., Sacks, G. D., Needleman, J., Brook, R. H., Mittman, B. S., Ko, C. Y., et al. (2019). Injury-specific variables improve risk adjustment and hospital quality assessment in severe traumatic brain injury. *J. Trauma Acute Care Surg.* 87 (2), 386–392. doi:10.1097/ta.0000000000002297

Denaro, N., Merlano, M. C., and Lo Nigro, C. (2019). Long noncoding RNA s as regulators of cancer immunity. *Mol. Oncol.* 13 (1), 61–73. doi:10.1002/1878-0261.12413

DiStefano, J. K. (2017). Long noncoding RNAs in the initiation, progression, and metastasis of hepatocellular carcinoma. *Non-coding RNA Res.* 2, 129–136. doi:10.1016/j.ncrna.2017.11.001

Dufour, J. F., Bargellini, I., De Maria, N., De Simone, P., Goulis, I., and Marinho, R. T. (2013). Intermediate hepatocellular carcinoma: current treatments and future perspectives. *Ann. Oncol.* 24 Suppl 2, ii24–9. doi:10.1093/annonc/mdt054

Fang, P., Xiang, L., Chen, W., Li, S., Huang, S., Li, J., et al. (2019). LncRNA GAS5 enhanced the killing effect of NK cell on liver cancer through regulating miR-544/RUNX3. *Innate Immun.* 25 (2), 99–109. doi:10.1177/1753425919827632

Ganapathy-Kanniappan, S. (2018). Molecular intricacies of aerobic glycolysis in cancer: current insights into the classic metabolic phenotype. *Crit. Rev. Biochem. Mol. Biol.* 53 (6), 667–682. doi:10.1080/10409238.2018.1556578

Goodman, A., Patel, S. P., and Kurzrock, R. (2017). PD-1-PD-L1 immune-checkpoint blockade in B-cell lymphomas. *Nat. Rev. Clin. Oncol.* 14 (4), 203–220. doi:10.1038/nrclinonc.2016.168

Hsu, C.-C., Wu, L.-C., Hsia, C.-Y., Yin, P.-H., Chi, C.-W., Yeh, T.-S., et al. (2015). Energy metabolism determines the sensitivity of human hepatocellular carcinoma cells to mitochondrial inhibitors and biguanide drugs. *Oncol. Rep.* 34 (3), 1620–1628. doi:10.3892/or.2015.4092

Hu, K. S., Tang, B., Yuan, J., Lu, S. X., Li, M., Chen, R. X., et al. (2019). A new substage classification strategy for Barcelona clinic liver cancer stage B patients with hepatocellular carcinoma. *J. Gastroenterol. Hepatol.* 34 (11), 1984–1991. doi:10.1111/jgh.14673

Huang, Z., Luo, Q., Yao, F., Qing, C., Ye, J., Deng, Y., et al. (2016). Identification of differentially expressed long non-coding RNAs in polarized macrophages. *Scientific Rep.* 6, 19705. doi:10.1038/srep19705

Invalidcitation (2018). *Invalidcitation*, 29–31.

Ji, J., Yin, Y., Ju, H., Xu, X., Lin, W., Fu, Q., et al. (2018). Long non-coding RNA Lnc-Tim3 exacerbates CD8 T cell exhaustion via binding to Tim-3 and inducing nuclear translocation of Bat3 in HCC. *Cel Death andDis.* 9 (5), 478. doi:10.1038/s41419-018-0528-7

Kim, J. E., Patel, M. A., Mangraviti, A., Kim, E. S., Theodros, D., Velarde, E., et al. (2017). Combination therapy with anti-PD-1, anti-TIM-3, and focal radiation results in regression of murine gliomas. *Clin. Cancer Res.* 23 (1), 124–136. doi:10.1158/1078-0432.ccr-15-1535

Kong, Q., Liang, C., Jin, Y, Pan, Y, Tong, D, Kong, Q, et al. (2019). The lncRNA MIR4435-2HG is upregulated in hepatocellular carcinoma and promotes cancer cell proliferation by upregulating miRNA-487a. *Cell andMol. Biol. Lett.* 24, 26. doi:10.1186/s11658-019-0148-y

Kopp, F., and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding RNAs. *Cell* 172 (3), 393–407. doi:10.1016/j.cell.2018.01.011

Lei, X., Lei, Y., Li, J.-K., Du, W.-X., Li, R.-G., Yang, J., et al. (2020). Immune cells within the tumor microenvironment: biological functions and roles in cancer immunotherapy. *Cancer Lett.* 470, 126–133. doi:10.1016/j.canlet.2019.11.009

Liang, W., Zhang, Y., Song, L., and Li, Z. (2019). 2,3′4,4′,5-Pentachlorobiphenyl induces hepatocellular carcinoma cell proliferation through pyruvate kinase M2-dependent glycolysis. *Toxicol. Lett.* 313, 108–119. doi:10.1016/j.toxlet.2019.06.006

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cel Syst.* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004

Lin, L., Yao, Z., Bhuvaneshwar, K., Gusev, Y., Kallakury, B., Yang, S., et al. (2014). Transcriptional regulation of STAT3 by SPTBN1 and SMAD3 in HCC through cAMP-response element-binding proteins ATF3 and CREB2. *Carcinogenesis* 35 (11), 2393–2403. doi:10.1093/carcin/bgu163

Liu, M., Zhou, J., Liu, X., Feng, Y., Yang, W., Wu, F., et al. (2020). Targeting monocyte-intrinsic enhancer reprogramming improves immunotherapy efficacy in hepatocellular carcinoma. *Gut* 69 (2), 365–379. doi:10.1136/gutjnl-2018-317257

Llovet, J. M., Montal, R., Sia, D., and Finn, R. S. (2018). Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat. Rev. Clin. Oncol.* 15 (10), 599–616. doi:10.1038/s41571-018-0073-4

Lu, J., Tan, M., and Cai, Q. (2015). The Warburg effect in tumor progression: mitochondrial oxidative metabolism as an anti-metastasis mechanism. *Cancer Lett.* 356, 156–164. doi:10.1016/j.canlet.2014.04.001

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., and Koeffler, H. P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* 28 (11), 1747–1756. doi:10.1101/gr.239244.118

Miao, H., Wang, L., Zhan, H., Dai, J., Chang, Y., Wu, F., et al. (2019). A long noncoding RNA distributed in both nucleus and cytoplasm operates in the PYCARD-regulated apoptosis by coordinating the epigenetic and translational regulation. *PLoS Genet.* 15 (5), e1008144. doi:10.1371/journal.pgen.1008144

Mushtaq, M., Papadas, A., Pagenkopf, A., Flietner, E., Morrow, Z., Chaudhary, S. G., et al. (2018). Tumor matrix remodeling and novel immunotherapies: the promise of matrix-derived immune biomarkers. *J. Immunother. Cancer* 6 (1), 65. doi:10.1186/s40425-018-0376-0

Ng, H., Lee, R. Y., Goh, S., Lim, X., Lee, B., Chew, Y., et al. (2020). Immunohistochemical scoring of CD38 in the tumor microenvironment predicts responsiveness to anti-PD-1/PD-L1 immunotherapy in hepatocellular carcinoma. *J. Immunother. Cancer* 8 (2). doi:10.1136/jitc-2020-000987

Nishino, M., Ramaiya, N. H., Hatabu, H., and Hodi, F. S. (2017). Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nat. Rev. Clin. Oncol.* 14 (11), 655–668. doi:10.1038/nrclinonc.2017.88

Pascale, R., Calvisi, D. F., Simile, M. M., Feo, C. F., and Feo, F., (2020). The Warburg effect 97 Years after its discovery. *Cancers* 12 (10), 2819. doi:10.3390/cancers12102819

Peng, L., Chen, Y., Ou, Q., Wang, X., and Tang, N. (2020). LncRNA MIAT correlates with immune infiltrates and drug reactions in hepatocellular carcinoma. *Int. immunopharmacology* 89, 107071. doi:10.1016/j.intimp.2020.107071

Pitt, J. M., Vétizou, M., Daillère, R., Roberti, M. P., Yamazaki, T., Routy, B., et al. (2016). Resistance mechanisms to immune-checkpoint blockade in cancer: tumor-intrinsic and -extrinsic factors. *Immunity* 44 (6), 1255–1269. doi:10.1016/j.immuni.2016.06.001

Riera Leal, A., Ortiz-Lazareno, P. C., Jave-Suárez, L. F., Ramírez De Arellano, A., Aguilar-Lemarroy, A., Ortiz-García, Y. M., et al. (2020). 17β-estradiol-induced mitochondrial dysfunction and Warburg effect in cervical cancer cells allow cell survival under metabolic stress. *Int. J. Oncol.* 56 (1), 33–46. doi:10.3892/ijo.2019.4912

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., et al. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348 (6230), 124–128. doi:10.1126/science.aaa1348

Ruan, J., Zheng, H, Rong, X, Rong, X, Zhang, J, Fang, W, et al. (2016). Over-expression of cathepsin B in hepatocellular carcinomas predicts poor prognosis of HCC patients. *Mol. Cancer* 15, 17. doi:10.1186/s12943-016-0503-9

Salik, B., Smyth, M., and Nakamura, K. (2020). Targeting immune checkpoints in hematological malignancies. *J. Hematol. andOncol.* 13 (1), 111. doi:10.1186/s13045-020-00947-6

Schmitt, A. M., and Chang, H. Y. (2016). Long noncoding RNAs in cancer pathways. *Cancer cell* 29 (4), 452–463. doi:10.1016/j.ccell.2016.03.010

Shen, S. N., Li, K., Liu, Y., Yang, C. L., He, C. Y., and Wang, H. R. (2019). Down-regulation of long noncodingRNA PVT1 inhibits esophageal carcinoma cell migration and invasion and promotes cell apoptosis via microRNA-145-mediated inhibition ofFSCN1. *Mol. Oncol.* 13 (12), 2554–2573. doi:10.1002/1878-0261.12555

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA A. Cancer J. Clin.* 70 (1), 7–30. doi:10.3322/caac.21590

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371 (23), 2189–2199. doi:10.1056/nejmoa1406498

Sun, L., Suo, C., Li, S.-t., Zhang, H., and Gao, P. (2018). Metabolic reprogramming for cancer cells and their microenvironment: beyond the Warburg Effect. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1870 (1), 51–66. doi:10.1016/j.bbcan.2018.06.005

Terabe, K., Ohashi, Y., Tsuchiya, S., Ishizuka, S., Knudson, C. B., and Knudson, W. (2019). Chondroprotective effects of 4-methylumbelliferone and hyaluronan synthase-2 overexpression involve changes in chondrocyte energy metabolism. *J. Biol. Chem.* 294 (47), 17799–17817. doi:10.1074/jbc.ra119.009556

Vidyasagar, M. (2015). Identifying predictive features in drug response using machine learning: opportunities and challenges. *Annu. Rev. Pharmacol. Toxicol.* 55, 15–34. doi:10.1146/annurev-pharmtox-010814-124502

Wei, L., Wang, X., Lv, L., Liv, J., Xing, H., Song, Y., et al. (2019). The emerging role of microRNAs and long noncoding RNAs in drug resistance of hepatocellular carcinoma. *Mol. Cancer* 18 (1), 147. doi:10.1186/s12943-019-1086-z

Yuan, S.-x., Zhang, J., Xu, Q.-g., Yang, Y., and Zhou, W.-p. (2016). Long noncoding RNA, the methylation of genomic elements and their emerging crosstalk in hepatocellular carcinoma. *Cancer Lett.* 379 (2), 239–244. doi:10.1016/j.canlet.2015.08.008

Ždralević, M., Brand, A., Lanni, L. D., Dettmer, K., Peter, K., Reinders, J., et al. (2018). Double genetic disruption of lactate dehydrogenases A and B is required to ablate the "Warburg effect" restricting tumor growth to oxidative metabolism. *J. Biol. Chem.* 293 (41), 15947–15961. doi:10.1074/jbc.RA118.004180

Zhai, L., Ladomersky, E., Lenzen, A., Nguyen, B., Patel, R., Lauing, K. L., et al. (2018). Ido1 in cancer: a Gemini of immune checkpoints. *Cell Mol Immunol* 15 (5), 447–457. doi:10.1038/cmi.2017.143

Zhang, L., He, T., Yan, Y, Zhang, Y., Zhou, X, Kong, Y., et al. (2016). Expression and clinical significance of the novel long noncoding RNA znf674-AS1 in human hepatocellular carcinoma. *Biomed. Research International* 2016, 3608914. doi:10.1155/2016/3608914

Zhang, Q., He, Y., Luo, N., Patel, S. J., Han, Y., Gao, R., et al. (2019). Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* 179 (4), 829–845. doi:10.1016/j.cell.2019.10.003

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 April 2021
doi: 10.3389/fgene.2021.636429

EnRank: An Ensemble Method to Detect Pulmonary Hypertension Biomarkers Based on Feature Selection and Machine Learning Models

*Xiangju Liu[1], Yu Zhang[1], Chunli Fu[1], Ruochi Zhang[2] and Fengfeng Zhou[2]**

[1]*Department of Geriatric Medicine & Shandong Key Laboratory Cardiovascular Proteomics, Qilu Hospital of Shandong University, Jinan, China,* [2]*College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China*

OPEN ACCESS

Edited by:
Robert Friedman, Retired, Columbia, SC, United States

Reviewed by:
Wanjun Gu, Southeast University, China
Jianbo Pan, Johns Hopkins Medicine, United States
Weiqun Peng, George Washington University, United States

***Correspondence:**
Fengfeng Zhou
fengfengzhou@gmail.com; ffzhou@jlu.edu.cn

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

Received: *05 January 2021*
Accepted: *30 March 2021*
Published: *27 April 2021*

Citation:
Liu X, Zhang Y, Fu C, Zhang R and Zhou F (2021) EnRank: An Ensemble Method to Detect Pulmonary Hypertension Biomarkers Based on Feature Selection and Machine Learning Models. Front. Genet. 12:636429. doi: 10.3389/fgene.2021.636429

Pulmonary hypertension (PH) is a common disease that affects the normal functioning of the human pulmonary arteries. The peripheral blood mononuclear cells (PMBCs) served as an ideal source for a minimally invasive disease diagnosis. This study hypothesized that the transcriptional fluctuations in the PMBCs exposed to the PH arteries may stably reflect the disease. However, the dimension of a human transcriptome is much higher than the number of samples in all the existing datasets. So, an ensemble feature selection algorithm, EnRank, was proposed to integrate the ranking information of four popular feature selection algorithms, i.e., T-test (Ttest), Chi-squared test (Chi2), ridge regression (Ridge), and Least Absolute Shrinkage and Selection Operator (Lasso). Our results suggested that the EnRank-detected biomarkers provided useful information from these four feature selection algorithms and achieved very good prediction accuracy in predicting the PH patients. Many of the EnRank-detected biomarkers were also supported by the literature.

Keywords: EnRank, ensemble feature selection, filter, pulmonary hypertension, biomarker detection

INTRODUCTION

Pulmonary hypertension (PH) shows the symptom of high blood pressure in the lung arteries which impedes the delivery of blood from the heart to the lungs (Mandras et al., 2020). PH is diagnosed by at least 20 mmHg (millimeter of mercury) of the rest-state mean pulmonary arterial pressure (mPAP) and the right-sided heart catheterization (Simonneau et al., 2019). Although PH may be caused by various factors, the PH patients painfully suffer from shortness of breath and increased mortality (Mandras et al., 2020). As many as, 10% of people over age 65 are affected by PH, and more than half of them develop heart failure (Hoeper et al., 2016). Detection of novel transcriptomic biomarkers may facilitate the understanding of the PH molecular mechanisms and serve as candidates for investigation and prognosis of the disease (Jardim and Souza, 2015; Swaminathan et al., 2015).

The high throughput DNA sequencing technology generates the expression levels of all human protein-coding and non-coding genes (Jandl et al., 2019; Tzimas et al., 2019), and machine learning methods rely on this data (Stephens et al., 2015; Mirza et al., 2019). The sequenced samples may be lesion tissues, e.g., the endothelial cells or the small remodeled arteries (Jandl et al., 2019; Tzimas et al., 2019). The peripheral blood mononuclear cells (PBMCs) serve ideally as the targets for the transcriptome sequencing because it is less invasive than use of lesion tissue (Tzouvelekis et al., 2018).

Transcriptome based disease prediction is often limited by sufficient sampling toward detection of disease features. This is mostly caused by the high cost of sequencing, a transcriptome and the difficulty in recruiting a cohort of individuals with and without the disease (Diao and Vidyashankar, 2013). Building a prediction model using all features may lead to overfitting and loss of applicability to a non-training data set (Schinkel et al., 2019). This problem is addressed by an algorithm for selecting a subset of the features in building the disease prediction model (Schinkel et al., 2019; Shi et al., 2019; McCabe et al., 2020).

There are two main categories of feature selection algorithms, filter and wrapper (Ye et al., 2017). A filter feature selection algorithm evaluates the association of each feature with a class label and then ranks features based on the significance of this association (Hall and Smith, 1999). These filter algorithms are commonly used for detection of biomarkers since their time complexity is linear. However, the filters ignore the inter-feature correlations and cannot detect the subset of low ranked features with good prediction performance (Ye et al., 2017). A wrapper feature selection algorithm heuristically generates the feature subsets and evaluates the prediction performance of a given feature subset using a user defined classifier (Das, 2001). The wrappers usually have higher time complexities than the filters and tend to deliver the feature subsets with better prediction performance than the filters (Ge et al., 2016).

This study proposes an ensemble feature selection algorithm, EnRank, to take advantage of both filters and wrappers. The main idea of EnRank is to integrate the ranks of multiple feature selection algorithms and verify that the final feature subset is efficient for use in a prediction model. A comprehensive evaluation was carried out to test which classifier achieved the best prediction performance. Our experimental data suggested that different feature selection algorithms may contribute complementary information to each other and the orchestration of the features selected by these algorithms are efficient for use in a predictive model.

MATERIALS AND METHODS

Collection of Data

This study used the transcriptome dataset GSE33463 of pulmonary hypertension patients and controls (Cheadle et al., 2012). The gene expression levels were profiled from the PBMCs of the recruited participants on the microarray platform GPL6947 (Illumina HumanHT-12 V3.0 expression beadchip). Each sample had 49,576 transcriptomic features, and the feature annotations were retrieved from the platform definition file.

This dataset consisted of 140 samples in total. There were 30 idiopathic pulmonary arterial hypertension (PAH) patients, 19 patients with systemic sclerosis (SSc) without pulmonary hypertension, 42 scleroderma-associated PAH patients, and eight patients with SSc complicated by interstitial lung diseases and pulmonary hypertension. The remaining 41 samples were non-disease controls. This study investigated the binary classification problem between the 99 patients (positive samples) and 41 non-disease controls (negative samples).

Feature Selection Algorithms

Feature selection algorithms were used to find the biomarkers with the best disease detection performance. Each sample had 49,576 transcriptomic features, and the overall dataset had 140 samples in total. A classification model may have a large chance of overfitting for this "large p small n" situation (Keel et al., 2019; Ren et al., 2020). A feature selection algorithm may be used to find a subset of features for building an accurate and stable classification model. This would also make the model easier to be interpreted along with better performance during the training step. The following four feature selection algorithms were utilized to find a good subset of features.

The Chi-squared test (Chi2) helps to test the relationships or dependence between two variables. Chi2 may be used to remove the features without dependency on the class labels (Xiao et al., 2020). In other words, these removed features will have a small contribution to any classification model.

The T-test (Ttest) is widely used to evaluate the statistical significance (p value) of the null hypothesis that a feature of the positive samples has the same normal distribution as that of the negative samples (Govindan et al., 2019; Soh et al., 2020). A feature with the value of $p < 0.05$ is typically considered a candidate for differential expression between the positive and negative samples. In addition, a feature with a lower p value is considered to have increased power for binary classification.

The ridge regression (Ridge) evaluates a subset of features for their connections with the class labels (Gao et al., 2020; Xu et al., 2020). Ridge provides a model-based trade-off between the fitting and complexity of the features by adding the L2 regularization to the regression model.

The Least Absolute Shrinkage and Selection Operator (Lasso) algorithm adds the L1 regularization to the regression model along with a penalty value for number of features (Deshpande et al., 2019). So, Lasso tends to select a small subset of features and weights them for building a robust regression model.

Binary Classification Methods

The models for predicting disease were trained using five binary classifiers.

Logistic regression (LR) is a statistics model using a logistic function to model a binary classification problem (Cuadrado-Godia et al., 2019; Khandezamin et al., 2020).

The logistic model calculates the log-odds for the class label by a linear combination of one or multiple predictors.

Support vector machine (SVM) is a supervised machine learning algorithm originally designed for binary classification (Jin et al., 2019; Wang et al., 2019). SVM searches for a hyperplane to separate two classes of samples with the maximal margins. It enriches the feature space through a kernel function to quantify the inter-sample similarities.

A simple algorithm K Nearest Neighbor (KNN) is a popular supervised machine learning framework for both classification and regression tasks (Wang et al., 2020; Yuan et al., 2020). KNN determines the class label of a query sample through the majority voting strategy of the KNNs of the query sample.

Decision tree (DT) uses a tree structure to solve the classification problem (Prieto-Gonzalez et al., 2020). Each node except for the leaves in a DT classifier exerts a feature evaluation, and the evaluation result determines which sub-branch of this current node to follow. DT is a simple and easy-to-interpret classifier.

The adaptive boosting tree (AdaBoost) is an integrated machine learning technique (Qiao and Xie, 2019; Dou et al., 2020). The weight of a sample will be increased if this sample leads to a misclassified base classifier. Each iteration will add new base classifiers. The final goal is to find a strong classifier with sufficiently small error rate.

Performance Evaluation Metrics

The supervised machine learning algorithms were evaluated by the following performance metrics. These metrics are essential to measure a prediction model from different aspects. This study used specificity (Sp), sensitivity (Sn), accuracy (Acc), and the area under the receiver operating characteristics curve (AUC). The number of correctly predicted positive samples was defined as the true positive (TP) and that of the incorrectly predicted positives was the false negative (FN). The true negative (TN) and the false positive (TP) defined the numbers of correctly and incorrectly predicted negative samples, respectively.

The overall accuracy is calculated as the number of all the correct predictions divided by the total number of samples in the dataset. That is to say, Acc = (TP + TN)/(TP + FN + TN + FP). The value of Acc is between 0.0 and 1.0. The two metrics Sp and Sn describe the ratios of correctly predicted negative and positive samples, respectively. So Sp = TN/(TN + FP) and Sn = TP/(TP + FN). Both metrics are between 0.0 and 1.0. A larger value of the three metrics Acc/Sp/Sn suggests a better prediction performance. The Matthews' Correlation Coefficient (MCC; Matthews, 1975) was introduced by the biochemist Brian W. Matthews in 1975 and MCC is generally regarded as a balanced measurement which can be used even if the classes are of very different sizes. The metric AUC is a parameter independent metric for the prediction model and shows a trade-off between Sp and Sn (Shao et al., 2020).

The Proposed Feature Ranking Algorithm, EnRank

This study proposed the ensemble feature selection algorithm, EnRank, by calculating the weighted ranks of the four feature selection algorithms, i.e., Ttest, Chi2, Ridge, and Lasso. The two filter algorithms Ttest and Chi2 rank the features by their individual association values of p with the class labels. The two linear fitting algorithms Ridge and Lasso rank the features based on the absolute values of the fitted model's coefficients. The values of the feature ranks start from 1, i.e., the best ranked feature has the rank 1. Each feature selection algorithm selects top-ranked pTopK = 100 features for further screening.

The proposed algorithm EnRank defines a weight Aim_i for each feature selection algorithm, where $i \in$ {Ttest, Chi2, Ridge, Lasso}. The pTopK features selected by each algorithm were loaded into the five classification algorithms, i.e., LR, SVM, KNN, DT, and AdaBoost. The stratified 5-fold cross validation (S5FCV) strategy was used to calculate the metric AUC, and each feature selection algorithm received five AUC values. This study aimed to find a feature subset with stably high AUC values for five classification algorithms, and defined $Aim_i = Avg_i/Var_i$, where Avg_i and Var_i were the averaged value and variance of the five AUC values of the feature selection algorithm I, respectively.

Finally, EnRank generated an integrated rank for each feature f. To avoid the case of very low ranking features, the rank of feature $Rank_i(f)$ was redefined as the penalization rank pPenaltyRank = 1,000, if $Rank_i(f)$ > pTopK. The integrated rank EnRank(f) = Average($Rank_i(f) \times Aim_i$) was defined as the EnRank metric, where the function Average() is the averaged value, and $i \in$ {Ttest, Chi2, Ridge, Lasso}.

Then, any filter-based feature selection frameworks, e.g., the incremental feature selection (IFS), may be used to find the best subset of top-ranked features generated by EnRank.

Workflow of This Study

This study proposed an ensemble feature selection algorithm, EnRank, by integrating the feature ranks from different algorithms (**Figure 1**). The experimental data in the following section suggested that different feature selection algorithms performed differently, and it is necessary to integrate the ranking information calculated by different feature selection algorithms.

RESULTS AND DISCUSSION

Comparison of the Feature Ranks by Ttest and Chi2

Table 1 illustrated the top-ranked 10 features delivered by the two filter algorithms Ttest and Chi2. Firstly, the statistical significance p values of the two algorithms Ttest and Chi2 were different to each other. The minimum p value of Ttest was 2.83e-19 while Chi2 only calculated the minimum p value 4.08e-4 for its null hypothesis. Actually, even the rank-100 feature ILMN_1698668 by Ttest had value of p = 2.28e-12, which was much smaller than the minimum value of p = 4.08e-4 of the algorithm Chi2.

And there was only one feature ILMN_1789074 shared among the top-ranked 10 features by Ttest and Chi2. The p value for

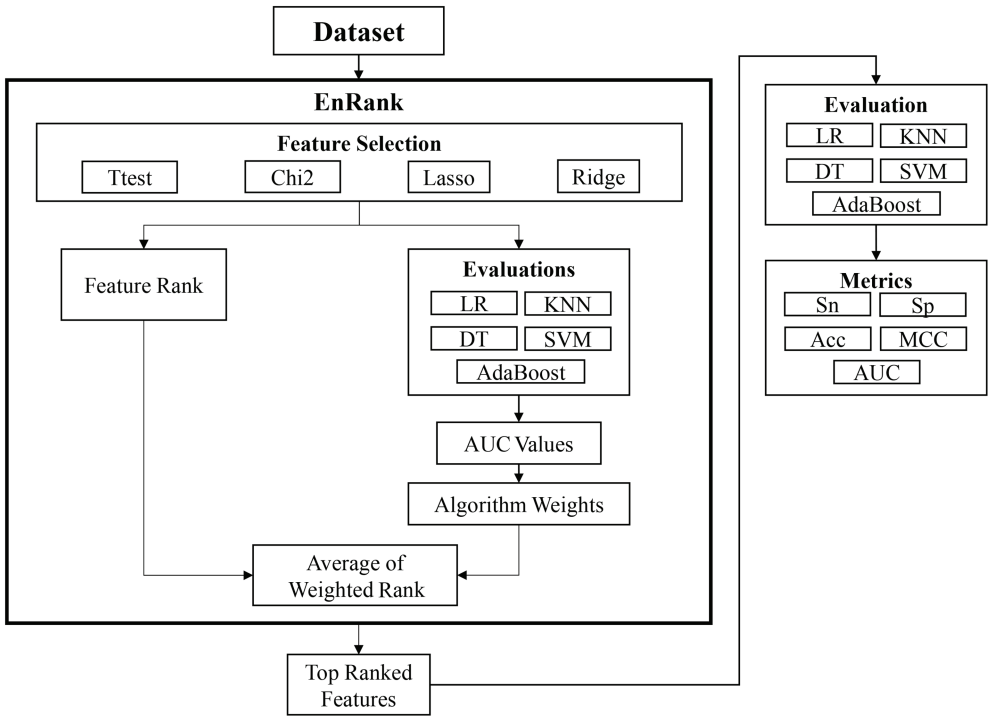


FIGURE 1 | Workflow of this study. The proposed algorithm EnRank integrates the ranks of both feature selection algorithms and classification algorithms. The finally generated feature subset is further evaluated by five different classification algorithms.

TABLE 1 | The top-10 features ranked by Ttest and Chi2. The two columns "Ttest" and "Chi2" gave the names of the ranked features.

Ttest	Ttest-*p* value	Rank	Chi2	Chi2-*p* value
ILMN_1812970	2.83E-19	1	ILMN_1806023	4.08E-04
ILMN_1875248	9.71E-19	2	ILMN_1656011	8.59E-04
ILMN_1704335	2.28E-18	3	ILMN_1702691	1.48E-03
ILMN_1794233	6.85E-18	4	ILMN_2367126	2.42E-03
ILMN_1765725	1.06E-17	5	ILMN_2339955	3.21E-03
ILMN_1758687	1.83E-17	6	ILMN_1815527	3.44E-03
ILMN_1687526	5.60E-17	7	**ILMN_1789074**	5.21E-03
ILMN_1767168	7.40E-17	8	ILMN_1782305	5.49E-03
ILMN_2159384	1.38E-16	9	ILMN_2088437	8.62E-03
ILMN_1789074	5.47E-16	10	ILMN_1751607	9.14E-03

Column "Rank" provided the rank values. "Ttest-p value" and "Chi2-p value" provided the statistical p values as calculated by the two algorithms Ttest and Chi2. The feature was highlighted in bold if it was among the top-ranked 10 features of both algorithms.

the Ttest null hypothesis was 5.47e-16 for feature ILMN_1789074 (Ttest rank 10), while Chi2 recommended ILMN_1789074 as the rank 7 feature with p value 5.21e-3.

So, the statistical tests Ttest and Chi2 generated significantly different p values for the features, and we had to integrate the features by their rank values.

Comparison of the Feature Ranks by Ridge and Lasso

Only six out of the top-10 ranked features by the absolute values of their model coefficients were shared by the two algorithms Ridge and Lasso (**Table 2**). This study assumed that both positive and negative correlations of the features with the class labels were important, and the absolute values of the model correlation coefficients of these features were used to rank the features in descending order. The feature ILMN_1697499 was the best ranked feature by Ridge, but it was not even within the top-10 ranked features by Lasso. Actually, the feature ILMN_1697499 was ranked 26 by Lasso. And the best ranked feature ILMN_1678859 by Lasso was only the ninth ranked feature by Ridge.

Venn diagram (**Figure 2**) shows that very few features were shared by these four feature selection algorithms, i.e., Ttest, Chi2, Lasso, and Ridge, except between Lasso and Ridge.

The data in **Tables 1, 2** suggested that the top-ranked features by the four algorithms Ttest, Chi2, Ridge, and Lasso described

TABLE 2 | The top-10 features ranked by the model coefficients of the regression models Ridge and Lasso.

Ridge	RidgeCoef	Rank	Lasso	LassoCoef
ILMN_1697499	0.0671	1	**ILMN_1678859**	0.1663
ILMN_2165753	0.0563	2	**ILMN_1781236**	0.1589
ILMN_1807491	0.0484	3	ILMN_2058782	0.1362
ILMN_1781236	0.0435	4	**ILMN_2083066**	0.1307
ILMN_1806023	0.0435	5	**ILMN_1807491**	0.1269
ILMN_2083066	0.0428	6	**ILMN_1806023**	0.1142
ILMN_1721113	0.0425	7	ILMN_1801216	0.1120
ILMN_2229649	0.0408	8	ILMN_1822671	0.1078
ILMN_1678859	0.0398	9	ILMN_1789074	0.1077
ILMN_2323933	0.0395	10	**ILMN_2165753**	0.1072

The two columns "Ridge" and "Lasso" identified the top-10 ranked features. Column "Rank" provided the rank values. The two columns "RidgeCoef" and "LassoCoef" gave the absolute values of the model coefficients calculated by the two algorithms Ridge and Lasso. The features were highlighted in bold if they were among the top-ranked 10 features of both algorithms.

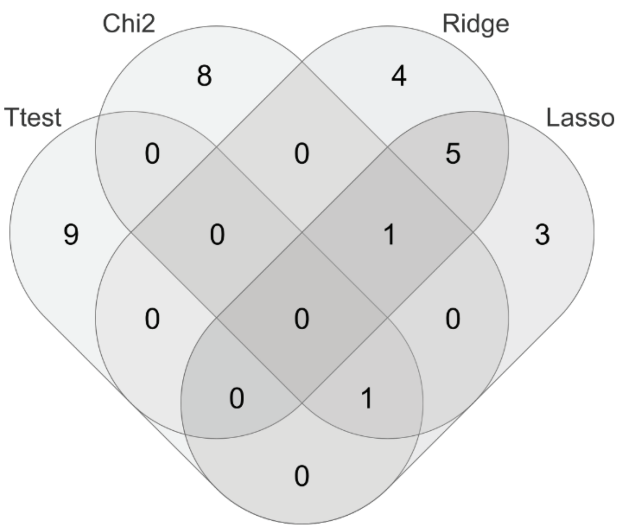


FIGURE 2 | Venn diagram of the top-10 features ranked by the four feature selection algorithms. The feature selection algorithms were T-test (Ttest), Chi-squared test (Chi2), Least Absolute Shrinkage and Selection Operator (Lasso), and ridge regression (Ridge).

the class correlations of the features from different aspects. **Figure 3** evaluated different value choices of the parameter pTopK. Both pTopK = 75 and 100 achieved the best averaged AUC = 0.9446. In order to introduce more feature diversity, this study focused on the four lists of top-ranked pTopK = 100 features by the above four algorithms, and their union consisted of 269 features.

Evaluation of the Four Feature Selection Algorithms

Figure 4A demonstrated that the classification algorithm DT had low performance on all four feature lists. And the other four classification algorithms achieved at least 0.9000 in the metric AUC for all four feature lists. Although the Lasso-selected 100 features achieved the best mean AUC value 0.9571 by the five classification algorithms, its SD 0.0701 was larger than that (0.0594) of another algorithm Ridge. So the Lasso's Aim 13.6620 was slightly larger than that (15.9508) of Ridge, as shown in **Figure 4B**. The filter Ttest was assigned the Aim 10.6090 due to its largest SD 0.0877.

Distribution of the Calculated EnRank Metrics

The ranking metric EnRank was defined in the above subsection "The proposed feature ranking algorithm, EnRank." EnRank used the EnRank metrics to rank the features in ascending order, and the features were roughly separated into four groups, as shown in **Figure 5**. The EnRank metrics of the ordered features were within these four ranges, i.e., [1, 1,000], [2,500, 3,300], [5,000, 5,700], and [7,500, 7,900]. The experimental data suggested that these four groups of features consisted of features recommended by four, three, two, and one feature selection algorithms, respectively. That is to say, a feature recommended by four feature selection algorithms was not penalized by the penalization rank pPenaltyRank, and algorithm aims were between 10 and 16. Such a feature had an EnRank smaller than 1,000. So the metric EnRank reasonably described how each feature was ranked by multiple feature selection algorithms.

Literature Supportive of the EnRank-Detected 50 Biomarkers

The metric Literature Support (LR) of a feature was defined by the number of PubMed (Fiorini et al., 2017) publications matching the gene symbol of this feature and the key word "pulmonary disease" in both title and abstract. The query term was "term={}[tiab] AND pulmonary disease [tiab]," and the queried date was November 16, 2020. The cumulative LR (CLR) of the top-k ranked features was defined as the sum of the LR values of these k features.

In order to compare with the biomarkers selected by EnRank, we randomly selected the same number of genes among the remaining genes as a control group, and then compared the metrics CLR and LR in the two groups. **Figure 6** illustrated that the EnRank-detected top-ranked features were investigated for their roles in pulmonary diseases many more times than the randomly-chosen features. The randomly-chosen features were supported

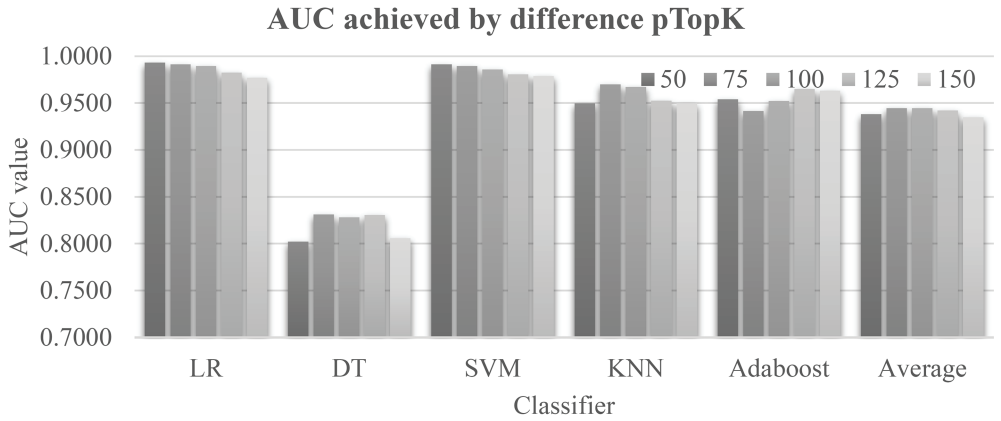
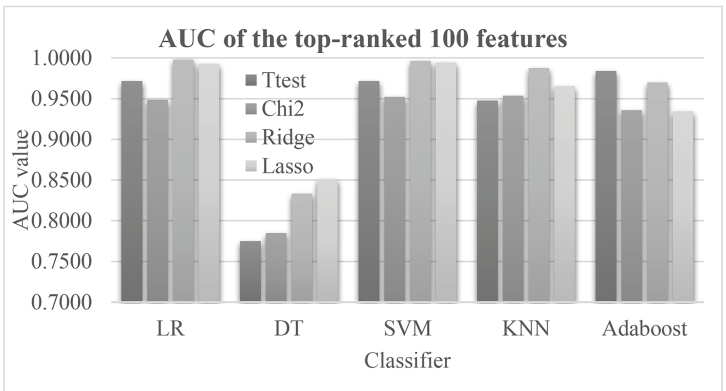


FIGURE 3 | Evaluation of the parameter pTopK of EnRank. The horizontal axis listed the five classifiers and the averaged area under the receiver operating characteristics curve (AUC) values by pTopK value. The five values, 50/75/100/125/150, are from EnRank.

A



B

Algorithm	Mean	Std	Aim
Ttest	0.9301	0.0877	10.6090
Chi2	0.9152	0.0731	12.5239
Ridge	0.9476	0.0594	15.9508
Lasso	0.9571	0.0701	13.6620

FIGURE 4 | The model performances and the weights of the four feature selection algorithms. **(A)** The AUC values of the top-100 features ranked by the four feature ranking algorithms using the five classification algorithms. Each of the four feature ranking algorithms Ttest, Chi2, Ridge, and Lasso selected the top-ranked 100 features. The AUC values of the feature lists were calculated by the stratified 5-fold cross validation (S5FCV) strategy of the five popular classification algorithms. **(B)** Calculation of the algorithm weight "Aim" for each of the four feature selection algorithms. The columns "Mean" and "Std" were the mean values and the SDs of the five classification algorithms. And the column "Aim" was defined as Mean/Std.

by at most two PubMed publications, and only four out of the 50 randomly-selected features had literature support. And the EnRank-detected top-ranked 50 features were more significantly supported by the scientific literature. Some features were supported by as many as nine PubMed publications, and 14 out of the 50 features had literature support. So the EnRank-detected features were consistently supported by the literature.

Model Evaluation Based on the EnRank-Detected Biomarkers

A comparative study was carried out to evaluate whether the proposed algorithm EnRank recommended features with good prediction performance of pulmonary hypertension (**Figure 7**). The baseline models in **Figure 7A** showed that the classifier DT achieved the worst PH prediction accuracy (Acc = 0.7545), while the classifier LR achieved the best Acc = 0.9000. SVM achieved the same Sn = 0.9560 as LR, but much worse Sp = 0.5361 than that (Sp = 0.8056) of LR. So, it is necessary to find a subset of biomarker features with a better PH prediction accuracy.

Figure 7B showed that the 50 EnRank-detected biomarkers improved the prediction accuracies of all five classification algorithms. The largest improvement in Acc (0.1364) was achieved for both SVM and KNN. The classification algorithm LR achieved the best Acc = 0.9545 again using the 50 EnRank-detected biomarkers. The parameter-independent metric AUC = 0.9894 of LR was also the best among the five classification algorithms.

So this study delivered a PH prediction model using the 50 EnRank-detected biomarkers and the LR classification algorithm.

Further Validation of the Proposed PH Biomarkers

Firstly, the proposed PH biomarkers were validated using an independent dataset GSE22356 (Risbano et al., 2010)

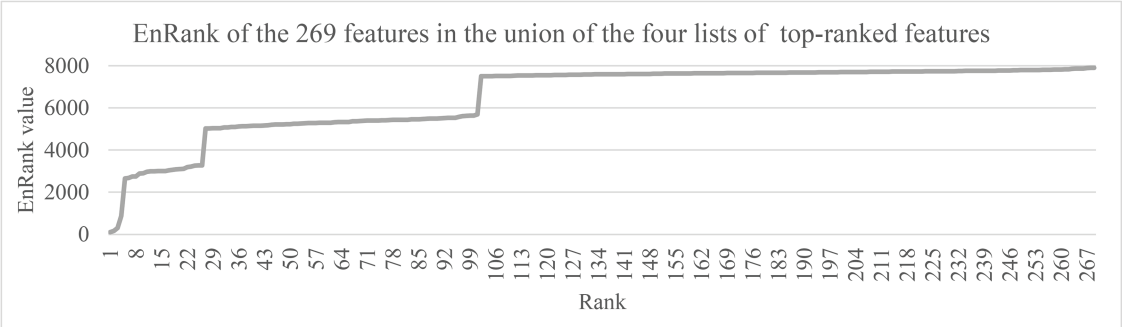


FIGURE 5 | The EnRank metrics of the 269 features in the union of the four lists of top-100 ranked features. The horizontal axis gave the feature ranks ordered by the EnRank metric, and the vertical axis gave the EnRank metrics of the top-ranked 269 features. These features were among the union of the top-100 ranked features recommended by the four algorithms, Ttest, Chi2, Ridge, and Lasso.

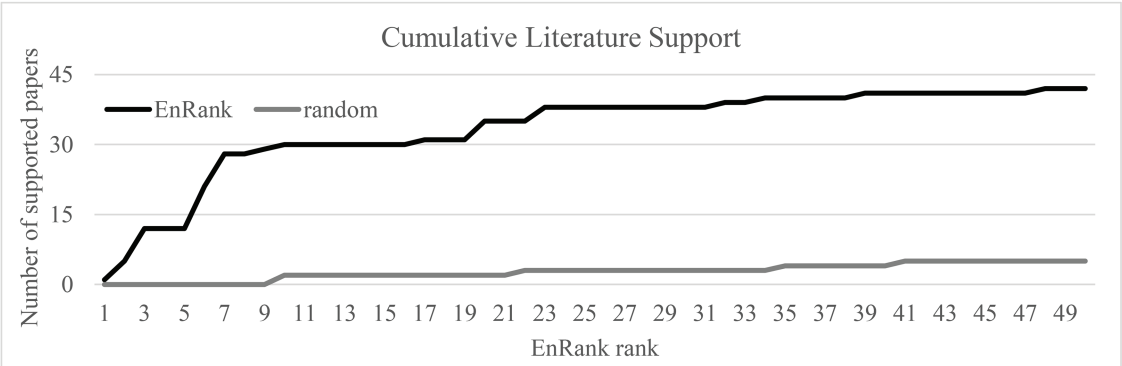


FIGURE 6 | Evaluation of the cumulative literature support LR (CLR) of the top-50 EnRank-ranked features. The horizontal axis gave the EnRank-recommended ranks and the vertical axis shows the metric CLR.

from the Gene Expression Omnibus (GEO) database (Clough and Barrett, 2016). This dataset consists of 38 PBMC samples profiled by the Affymetrix Human Genome U133 Plus 2.0 Array (GPL570), and investigates the altered immune phenotypes of the scleroderma-associated pulmonary hypertension (Risbano et al., 2010). The normal controls were assumed as the negative samples, and the other samples were regarded as the positive ones. The 50 EnRank-recommended features matched 236 features through 36 unique genes in the independent dataset. The same settings of training and evaluation as EnRank were used. **Figure 8A** showed that four of the five classifiers achieved AUC values at least 0.8000. The classifier LR achieved the largest AUC = 0.8433, and the largest Acc = 0.8893. Considering that this independent dataset was profiled using a different transcriptome platform than our original dataset GSE33463, the independent validation results supported the robustness of the EnRank-recommended PH biomarkers.

We searched the literature database PubMed using the keywords "pulmonary hypertension" and "biomarker" in the titles, and only 41 publications were detected. Most of them focused on the protein (Wu et al., 2020), vocal (Sara et al., 2020), and imaging (Jivraj et al., 2017; Jose et al., 2020) data. So we collected the PH marker genes from the recently updated database MarkerDB (Wishart et al., 2021). Three unique genes were annotated as the PH biomarkers, including Bone Morphogenetic Protein Receptor Type 2 (BMPR2), Activin A Receptor Like Type 1 (ACVRL1), and Endoglin (ENG). Four features were associated with these three genes. The prediction performances of these four biomarker features were shown in **Figure 8B**. Unfortunately, no classifiers showed larger than 0.7000 in either AUC or Acc using these biomarkers. This should be due to that the existing biomarkers were screened for their individual associations with the phenotype PH, and their combined PH prediction performances were not investigated in the existing studies.

Further Evaluation of Other Feature Selection Combinations

The proposed algorithm EnRank is a feature selection framework that may integrate the ranking data of multiple feature selection algorithms. The above sections integrated four feature selection algorithms, i.e., Ttest, Chi2, Ridge, and Lasso. **Figure 9A** evaluated the proposed ensembled algorithm EnRank and its

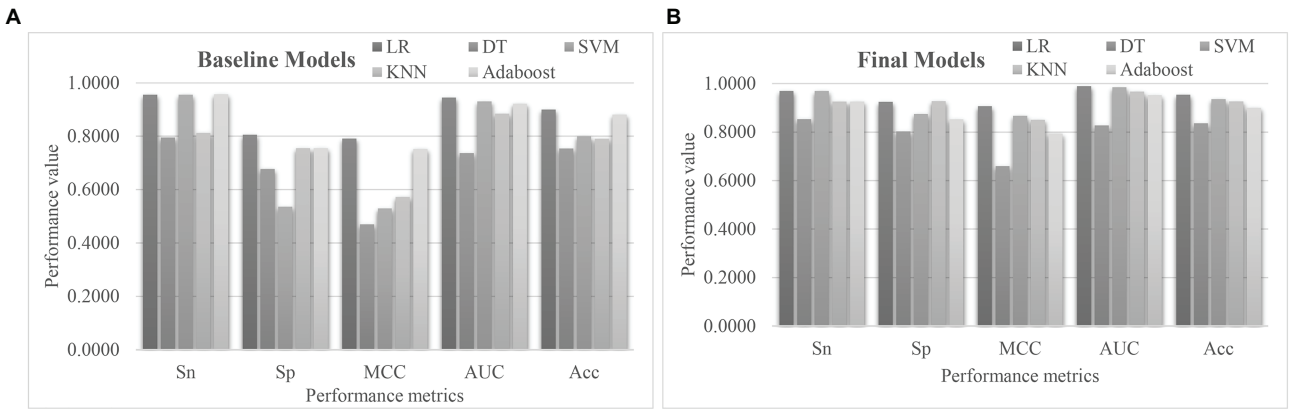


FIGURE 7 | Performance comparison of the five classification algorithms. The S5FCV strategy was used to train the five classification algorithms using **(A)** all the transcriptomic features, and **(B)** the 50 EnRank-detected biomarkers. The horizontal axis gave the performance metrics sensitivity (Sn), specificity (Sp), Matthews' Correlation Coefficient (MCC), AUC, and accuracy (Acc). The vertical axis gave the values of these performance metrics.

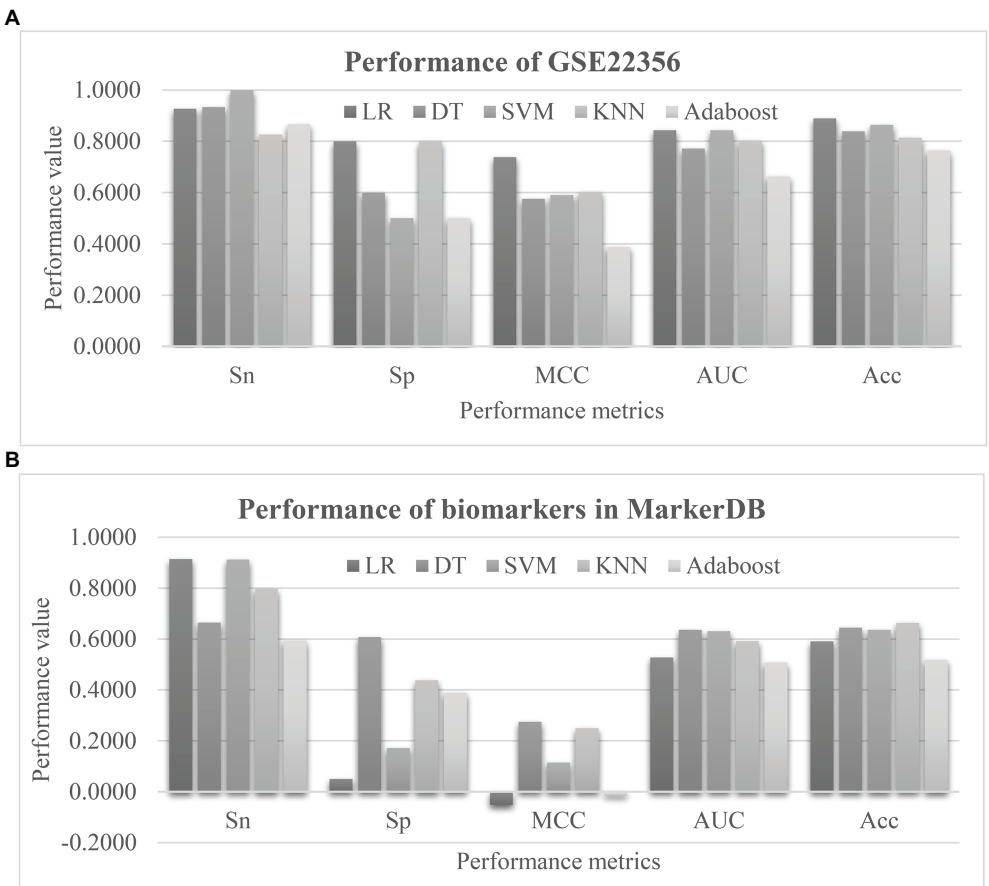


FIGURE 8 | Evaluation of the pulmonary hypertension (PH) detection model. **(A)** Validation of the 50 EnRank-recommended PH biomarkers in the dataset GSE22356. **(B)** Performances of the existing PH biomarkers. The horizontal axis listed the performance metrics Sn/Sp/MCC/ROC/Acc and the five classifiers were given as data series. The vertical axis gave the values of the performance metrics.

four individual feature selection algorithms using the same training and testing settings. The parameter-independent metric AUC was used to compare the performances of the feature selection algorithms. EnRank achieved the best AUC values using three out of the five classifiers. The Lasso-recommended features achieved the best AUC = 0.9946 while the

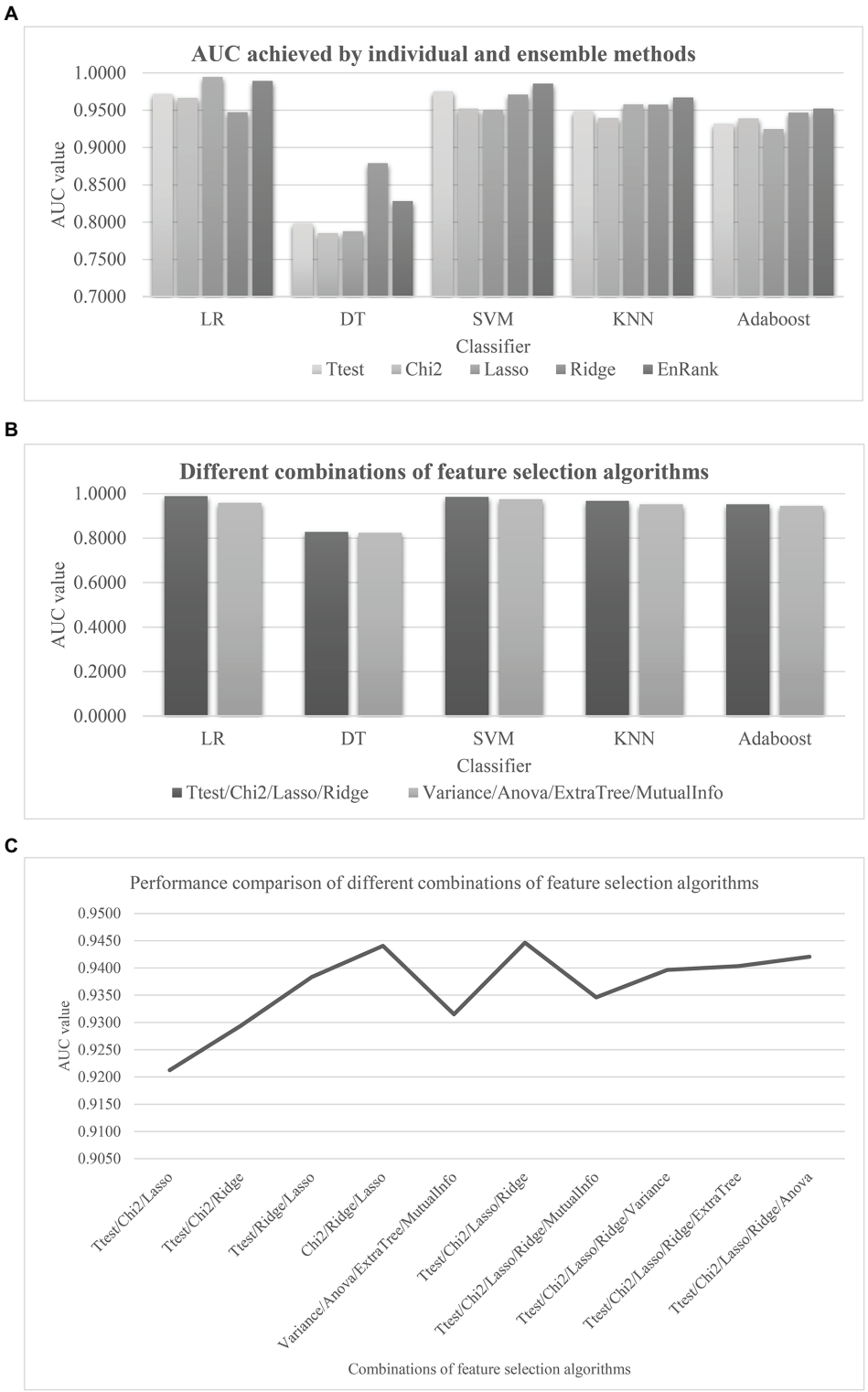


FIGURE 9 | Comparison of EnRank with the other feature selection algorithms and their combinations. **(A)** Evaluation of EnRank and its individual feature selection algorithms. **(B)** Two groups of feature selection algorithms were integrated by EnRank. The original version of EnRank was "Ttest/Chi2/Lasso/Ridge," and the new version was "Variance/Anova/ExtraTree/MutualInfo." The horizontal axis listed the classifiers and the vertical axis gave the AUC values of the evaluated models. The ensembled algorithm EnRank and its four individual feature selection algorithms. **(C)** The AUC values of different combinations of feature selection algorithms averaged over the five classifiers logistic regression (LR)/decision tree (DT)/support vector machine (SVM)/k nearest neighbor (KNN)/adaptive boosting tree (AdaBoost). The horizontal axis listed the algorithm combinations, and the vertical axis gave the AUC values.

EnRank-recommended features achieved the second best AUC = 0.9894. EnRank achieved the second best AUC = 0.8283 using the classifier DT, while Ridge-recommended features achieved the slightly better AUC = 0.8790.

The original version of EnRank integrated four feature selection algorithms Ttest/Chi2/Lasso/Ridge, which was compared with the new version integrating four new feature selection algorithms, as shown in **Figure 9B**. The four new feature selection algorithms were Variance Threshold (Variance), Mutual Information (MutualInfo), Extra Trees (ExtraTree), and ANOVA (Anova). The same model training and testing setting were carried out. The original version of EnRank outperformed the new version for all five classifiers. The best classifier LR was even improved by 0.0302 in the parameter-independent performance metric AUC.

The EnRank's performance relied on the including efficient feature selection algorithms (**Figure 9C**). So a comparison was carried out for the performances of different combinations of feature selection algorithms. Here, we investigated the combinations of three or five algorithms. **Figure 9C** showed that the original version of EnRank achieved the best AUC value = 0.9446, although a slightly worse AUC = 0.9441 was achieved by removing Ttest.

Biological Involvement of the EnRank-Detected Biomarkers

Table 3 listed the 50 EnRank-detected biomarkers and their corresponding gene information. Many transcriptomic biomarkers are from chromosomes 19 and 2. And two biomarkers ILMN_1807491 and ILMN_2323933 are from the same gene Leukocyte Associated Immunoglobulin Like Receptor 2 (LAIR2). Limited knowledge was known about the roles of LAIR2 in the PH patients, based on the information from PubMed (Fiorini et al., 2017) and MalaCards (Rappaport et al., 2017). There were five transcriptomic biomarkers with unknown chromosomal locations.

The feature ILMN_2088437 was from the gene C-X3-C Motif Chemokine Receptor 1 (CX3CR1), which was known to be involved in HIV proliferation (Mhandire et al., 2014; Guo et al., 2020). The absence of CX3CR1 was observed to provide protection from tissue destruction from chronic obstructive pulmonary disease (COPD; Lee, 2012). And the gene CX3CR1 also demonstrated differential expressions in the COPD patients (Huang et al., 2019). Another feature ILMN_1740875 was within the gene Formyl Peptide Receptor 2 (FPR2) encoded on chromosome 19, which was actively involved in the mononuclear phagocyte responses in Alzheimer disease (Iribarren et al., 2005). FPR2 also demonstrated its capability of promoting the chemotaxis and survival of neutrophils in the COPD patients (Iribarren et al., 2005).

The EnRank-recommended genes were analyzed using the online tool DAVID version 6.8 (Jiao et al., 2012). The list of genes was annotated to cover the top 50 EnRank-recommended features and was screened against the human genome. The statistical significance p values were adjusted by the multi-test Benjamini corrections, and only the functional terms with the Benjamini-corrected values of $p < 0.05$ were kept for further analysis. It is interesting to observe that no GO terms were significantly enriched in PH biomarkers; while seven KEGG pathways were enriched

TABLE 3 | Detailed information of the 50 EnRank-detected biomarkers.

Rank	Feature	Gene	Chr	Strand	Rank	Feature	Gene	Chr	Strand
41	ILMN_1804350	LOC644852	1	+	39	ILMN_1711786	NFE2	12	–
1	ILMN_1806023	JUN	1	–	20	ILMN_2207291	IFNG	12	–
15	ILMN_1723912	IFI44L	1	+	25	ILMN_2388547	EPSTI1	13	–
43	ILMN_2339955	NR4A2	2	–	40	ILMN_2229649	KCTD12	13	–
11	ILMN_1782305	NR4A2	2	–	5	ILMN_2058782	IFI27	14	+
23	ILMN_1800602	GCA	2	+	38	ILMN_1763364	WHDC1	15	+
8	ILMN_1733998	DHRS9	2	+	10	ILMN_2057836	RNU2	17|NT_113932.1	–
44	ILMN_1755643	MGAT4A	2	–	12	ILMN_1772796	DYNLL2	17	+
22	ILMN_1801307	TNFSF10	3	–	47	ILMN_1742618	XAF1	17	+
3	ILMN_2088437	CX3CR1	3	–	26	ILMN_1749722	RNF213	17	+
7	ILMN_1745788	CX3CR1	3	–	24	ILMN_2413331	TMEM107	17	–
33	ILMN_1801216	S100P	4	+	18	ILMN_1775304	DNAJB1	19	–
48	ILMN_1745522	PF4V1	4	+	46	ILMN_2302757	FCGBP	19	–
49	ILMN_1710734	GZMK	5	+	16	ILMN_1751607	FOSB	19	+
42	ILMN_1779147	ENC1	5	–	6	ILMN_1740875	FPR2	19	+
9	ILMN_1702691	TNFAIP3	6	+	28	ILMN_1807491	LAIR2	19	+
32	ILMN_1721113	HLA-C	6	–	45	ILMN_2323933	LAIR2	19	+
2	ILMN_1789074	HSPA1A	6	+	17	ILMN_1664861	ID1	20	+
34	ILMN_1697499	HLA-DRB5	6	–	29	ILMN_2083066	IGLL3	22	+
4	ILMN_1748473	GIMAP4	7	+	35	ILMN_1796830	UBE2L3	22	+
14	ILMN_1799467	SAMD9L	7	–	13	ILMN_1852793	UniGene|BC067908		
21	ILMN_1684982	PDK4	7	–	27	ILMN_1781236	RefSeq|XR_001116.1		
37	ILMN_1716733	MYOM2	8	+	30	ILMN_1678859	RefSeq|XM_938277.1		
50	ILMN_1773313	USMG5	10	–	31	ILMN_2165753	RefSeq|NM_001080840.1		
19	ILMN_1674063	OAS2	12	+	36	ILMN_1822671	UniGene|BC020840		

Column "Gene" gave the gene symbol for each biomarker. Some biomarkers may not reside in a protein-coding gene, and they may have no annotated gene information.

TABLE 4 | Enriched functional terms of the 50 EnRank-detected PH biomarkers.

KEGG	Term	*p* values	Benjamini
hsa05164	Influenza A	3.50E-06	2.80E-04
hsa05162	Measles	3.30E-04	1.30E-02
hsa04612	Antigen processing and presentation	9.40E-04	2.20E-02
hsa05168	Herpes simplex infection	1.10E-03	2.20E-02
hsa05332	Graft-versus-host disease	3.30E-03	4.70E-02
hsa05169	Epstein-Barr virus infection	3.70E-03	4.70E-02
hsa05330	Allograft rejection	4.10E-03	4.70E-02

The screening was carried out using the online tool DAVID in the Gene Ontology (GO) terms and KEGG terms. The first two columns gave the KEGG IDs and the corresponding terms. The last two columns gave the statistical significance p values and the Benjamini-corrected p values. Only the terms with the Benjamini-corrected p values < 0.05 were kept for further analysis.

with PH biomarkers. Many of these KEGG pathways were associated with antiviral immunity. The most significant KEGG pathway was hsa05164 (Influenza A) with the Benjamini-corrected p value = 2.80e-4. The infection of influenza A caused a patient's death after 3 months of treatment with the popular drug bosentan for pulmonary hypertension in a clinical trial (Hoeper et al., 2005). As of now, no direct link was presented in the literature. But virus infection is known to be closely connected with pulmonary hypertension (Kimura et al., 2019; Miyasaka et al., 2020; **Table 4**).

CONCLUSION

This study proposed a novel ensemble filter feature selection algorithm EnRank by the weighted integration of four popular filter algorithms. Five classification algorithms were used to evaluate the filter algorithms. The EnRank-detected biomarkers demonstrated very good performances on the PH prediction problem. And most of these biomarkers also demonstrated close connections with the disease PH from the literature.

The proposed algorithm EnRank is a feature selection framework, and may integrate feature selection algorithms with feature weights. The main limitation of EnRank is the choices of feature selection algorithms to be integrated. The parameter pTopK may also impact the final model performances. Others may want to carry out a series of comparable experiments to find the best parameters for their own datasets.

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=GSE33463 [The dataset has the accession GSE33463 in the Gene Expression Omnibus (GEO) database] and https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22356 [The dataset has the accession GSE22356 in the Gene Expression Omnibus (GEO) database].

AUTHOR CONTRIBUTIONS

FZ, XL, and RZ designed the project, carried out the experiments, and drafted the manuscript. XL, YZ, and CF were involved in the clinical annotations and results discussion. RZ carried out the coding of the computational analysis. RZ and FZ revised and polished the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Natural Science Foundation of the Shandong Province (ID-ZR2017MH122).

ACKNOWLEDGMENTS

We appreciate the constructive comments from the handling editor and the three reviewers that have substantially improved this manuscript.

REFERENCES

Cheadle, C., Berger, A. E., Mathai, S. C., Grigoryev, D. N., Watkins, T. N., Sugawara, Y., et al. (2012). Erythroid-specific transcriptional changes in PBMCs from pulmonary hypertension patients. *PLoS One* 7:e34951. doi: 10.1371/journal.pone.0034951

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Cuadrado-Godia, E., Jamthikar, A. D., Gupta, D., Khanna, N. N., Araki, T., Maniruzzaman, M., et al. (2019). Ranking of stroke and cardiovascular risk factors for an optimal risk calculator design: logistic regression approach. *Comput. Biol. Med.* 108, 182–195. doi: 10.1016/j.compbiomed.2019.03.020

Das, S. (2001). "Filters, wrappers and a boosting-based hybrid for feature selection." 74–81.

Deshpande, S., Shuttleworth, J., Yang, J., Taramonli, S., and England, M. (2019). PLIT: An alignment-free computational tool for identification of long non-coding RNAs in plant transcriptomic datasets. *Comput. Biol. Med.* 105, 169–181. doi: 10.1016/j.compbiomed.2018.12.014

Diao, G., and Vidyashankar, A. N. (2013). Assessing genome-wide statistical significance for large p small n problems. *Genetics* 194, 781–783. doi: 10.1534/genetics.113.150896

Dou, L., Li, X., Zhang, L., Xiang, H., and Xu, L. (2020). iGlu_AdaBoost: identification of lysine glutarylation using the Adaboost classifier. *J. Proteome Res.* 20, 191–201. doi: 10.1021/acs.jproteome.0c00314

Fiorini, N., Lipman, D. J., and Lu, Z. (2017). Towards PubMed 2.0. *elife* 6:e28801. doi: 10.7554/eLife.28801

Gao, X., Liu, S., Song, H., Feng, X., Duan, M., Huang, L., et al. (2020). AgeGuess, a methylomic prediction model for human ages. *Front. Bioeng. Biotechnol.* 8:80. doi: 10.3389/fbioe.2020.00080

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., et al. (2016). McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC bioinformatics* 17:142. doi: 10.1186/s12859-016-0990-0

Govindan, R. B., Massaro, A., Vezina, G., Chang, T., and du Plessis, A. (2019). Identifying an optimal epoch length for spectral analysis of heart rate of critically-ill infants. *Comput. Biol. Med.* 113:103391. doi: 10.1016/j.compbiomed.2019.103391

Guo, N., Chen, Y., Su, B., Yang, X., Zhang, Q., Song, T., et al. (2020). Alterations of CCR2 and CX3CR1 on three monocyte subsets during HIV-1/treponema pallidum coinfection. *Front. Med.* 7:272. doi: 10.3389/fmed.2020.00272

Hall, M.A., and Smith, L.A. (1999). "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper." 235–239.

Hoeper, M. M., Humbert, M., Souza, R., Idrees, M., Kawut, S. M., Sliwa-Hahnle, K., et al. (2016). A global view of pulmonary hypertension. *Lancet Respir. Med.* 4, 306–322. doi: 10.1016/S2213-2600(15)00543-3

Hoeper, M. M., Kramm, T., Wilkens, H., Schulze, C., Schafers, H. J., Welte, T., et al. (2005). Bosentan therapy for inoperable chronic thromboembolic pulmonary hypertension. *Chest* 128, 2363–2367. doi: 10.1378/chest.128.4.2363

Huang, X., Li, Y., Guo, X., Zhu, Z., Kong, X., Yu, F., et al. (2019). Identification of differentially expressed genes and signaling pathways in chronic obstructive pulmonary disease via bioinformatic analysis. *FEBS Open Bio.* 9, 1880–1899. doi: 10.1002/2211-5463.12719

Iribarren, P., Zhou, Y., Hu, J., Le, Y., and Wang, J. M. (2005). Role of formyl peptide receptor-like 1 (FPRL1/FPR2) in mononuclear phagocyte responses in Alzheimer disease. *Immunol. Res.* 31, 165–176. doi: 10.1385/IR:31:3:165

Jandl, K., Thekkekara Puthenparampil, H., Marsh, L. M., Hoffmann, J., Wilhelm, J., Veith, C., et al. (2019). Long non-coding RNAs influence the transcriptome in pulmonary arterial hypertension: the role of PAXIP1-AS1. *J. Pathol.* 247, 357–370. doi: 10.1002/path.5195

Jardim, C., and Souza, R. (2015). Biomarkers and prognostic indicators in pulmonary arterial hypertension. *Curr. Hypertens. Rep.* 17:556. doi: 10.1007/s11906-015-0556-y

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806. doi: 10.1093/bioinformatics/bts251

Jin, H., Titus, A., Liu, Y., Wang, Y., and Han, A. Z. (2019). Fault diagnosis of rotary parts of a heavy-duty horizontal lathe based on wavelet packet transform and support vector machine. *Sensors* 19:4069. doi: 10.3390/s19194069

Jivraj, K., Bedayat, A., Sung, Y. K., Zamanian, R. T., Haddad, F., Leung, A. N., et al. (2017). Left atrium maximal axial cross-sectional area is a specific computed tomographic imaging biomarker of World Health Organization Group 2 pulmonary hypertension. *J. Thorac. Imaging* 32, 121–126. doi: 10.1097/RTI.0000000000000252

Jose, A., Kher, A., O'Donnell, R. E., and Elwing, J. M. (2020). Cardiac magnetic resonance imaging as a prognostic biomarker in treatment-naive pulmonary hypertension. *Eur. J. Radiol.* 123:108784. doi: 10.1016/j.ejrad.2019.108784

Keel, B. N., Snelling, W. M., Lindholm-Perry, A. K., Oliver, W. T., Kuehn, L. A., and Rohrer, G. A. (2019). Using SNP weights derived from gene expression modules to improve gwas power for feed efficiency in pigs. *Front. Genet.* 10:1339. doi: 10.3389/fgene.2019.01339

Khandezamin, Z., Naderan, M., and Rashti, M. J. (2020). Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier. *J. Biomed. Inform.* 111:103591. doi: 10.1016/j.jbi.2020.103591

Kimura, D., McNamara, I. F., Wang, J., Fowke, J. H., West, A. N., and Philip, R. (2019). Pulmonary hypertension during respiratory syncytial virus bronchiolitis: a risk factor for severity of illness. *Cardiol. Young* 29, 615–619. doi: 10.1017/S1047951119000313

Lee, J. S. (2012). Heterogeneity of lung mononuclear phagocytes in chronic obstructive pulmonary disease. *J. Innate Immun.* 4, 489–497. doi: 10.1159/000337434

Mandras, S. A., Mehta, H. S., and Vaidya, A. (2020). Pulmonary hypertension: a brief guide for clinicians. *Mayo Clin. Proc.* 95, 1978–1988. doi: 10.1016/j.mayocp.2020.04.039

Matthews, B. W. (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

McCabe, S. D., Lin, D. Y., and Love, M. I. (2020). Consistency and overfitting of multi-omics methods on experimental data. *Brief. Bioinform.* 21, 1277–1284. doi: 10.1093/bib/bbz070

Mhandire, K., Duri, K., Kandawasvika, G., Chandiwana, P., Chin'ombe, N., Kanyera, R. B., et al. (2014). CCR2, CX3CR1, RANTES and SDF1 genetic polymorphisms influence HIV infection in a Zimbabwean pediatric population. *J. Infect. Dev. Ctries.* 8, 1313–1321. doi: 10.3855/jidc.4599

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Gene* 10:87. doi: 10.3390/genes10020087

Miyasaka, A., Yoshida, Y., Suzuki, A., Ueda, H., Morino, Y., and Takikawa, Y. (2020). A case of suspected portal-pulmonary hypertension due to hepatitis C virus infection. *Clin. J. Gastroenterol.* 13, 90–96. doi: 10.1007/s12328-019-01016-3

Prieto-Gonzalez, D., Castilla-Rodriguez, I., Gonzalez, E., and Couce, M. L. (2020). Automated generation of decision-tree models for the economic assessment of interventions for rare diseases using the RaDiOS ontology. *J. Biomed. Inform.* 110:103563. doi: 10.1016/j.jbi.2020.103563

Qiao, L., and Xie, D. (2019). MIonSite: ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. *Anal. Biochem.* 566, 75–88. doi: 10.1016/j.ab.2018.11.009

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi: 10.1093/nar/gkw1012

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., et al. (2020). Semiparametric Bayesian variable selection for gene-environment interactions. *Stat. Med.* 39, 617–638. doi: 10.1002/sim.8434

Risbano, M. G., Meadows, C. A., Coldren, C. D., Jenkins, T. J., Edwards, M. G., Collier, D., et al. (2010). Altered immune phenotype in peripheral blood cells of patients with scleroderma-associated pulmonary hypertension. *Clin. Transl. Sci.* 3, 210–218. doi: 10.1111/j.1752-8062.2010.00218.x

Sara, J. D. S., Maor, E., Borlaug, B., Lewis, B. R., Orbelo, D., Lerman, L. O., et al. (2020). Non-invasive vocal biomarker is associated with pulmonary hypertension. *PLoS One* 15:e0231441. doi: 10.1371/journal.pone.0231441

Schinkel, M., Paranjape, K., Nannan Panday, R. S., Skyttberg, N., and Nanayakkara, P. W. B. (2019). Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput. Biol. Med.* 115:103488. doi: 10.1016/j.compbiomed.2019.103488

Shao, Y., Nir, G., Fazli, L., Goldenberg, L., Gleave, M., Black, P., et al. (2020). Improving prostate cancer classification in H&E tissue micro arrays using Ki67 and P63 histopathology. *Comput. Biol. Med.* 127:104053. doi: 10.1016/j.compbiomed.2020.104053

Shi, L., Westerhuis, J. A., Rosen, J., Landberg, R., and Brunius, C. (2019). Variable selection and validation in multivariate modelling. *Bioinformatics* 35, 972–980. doi: 10.1093/bioinformatics/bty710

Simonneau, G., Montani, D., Celermajer, D. S., Denton, C. P., Gatzoulis, M. A., Krowka, M., et al. (2019). Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *Eur. Respir. J.* 53:1801913. doi: 10.1183/13993003.01913-2018

Soh, D. C. K., Ng, E. Y. K., Jahmunah, V., Oh, S. L., San, T. R., and Acharya, U. R. (2020). A computational intelligence tool for the detection of hypertension using empirical mode decomposition. *Comput. Biol. Med.* 118:103630. doi: 10.1016/j.compbiomed.2020.103630

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195

Swaminathan, A. C., Dusek, A. C., and McMahon, T. J. (2015). Treatment-related biomarkers in pulmonary hypertension. *Am. J. Respir. Cell Mol. Biol.* 52, 663–673. doi: 10.1165/rcmb.2014-0438TR

Tzimas, C., Rau, C. D., Buergisser, P. E., Jean-Louis, G. Jr., Lee, K., Chukwuneke, J., et al. (2019). WIPI1 is a conserved mediator of right ventricular failure. *JCI Insight* 5:e122929. doi: 10.1172/jci.insight.122929

Tzouvelekis, A., Herazo-Maya, J. D., Ryu, C., Chu, J. H., Zhang, Y., Gibson, K. F., et al. (2018). S100A12 as a marker of worse cardiac output and mortality in pulmonary hypertension. *Respirology* 23, 771–779. doi: 10.1111/resp.13302

Wang, W., Ding, M., Duan, X., Feng, X., Wang, P., Jiang, Q., et al. (2019). Diagnostic value of plasma microRNAs for lung cancer using support vector machine model. *J. Cancer* 10, 5090–5098. doi: 10.7150/jca.30528

Wang, C., Long, Y., Li, W., Dai, W., Xie, S., Liu, Y., et al. (2020). Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics. *Sci. Rep.* 10:5880. doi: 10.1038/s41598-020-62803-4

Wishart, D. S., Bartok, B., Oler, E., Liang, K. Y. H., Budinski, Z., Berjanskii, M., et al. (2021). MarkerDB: an online database of molecular biomarkers. *Nucleic Acids Res.* 49, D1259–D1267. doi: 10.1093/nar/gkaa1067

Wu, X., You, W., Wu, Z., Ye, F., and Chen, S. (2020). Serum biomarker analysis at the protein level on pulmonary hypertension secondary to old anterior myocardial infarction. *Pulm. Circ.* 10:2045894020969079. doi: 10.1177/2045894020969079

Xiao, M., Ma, F., Li, Y., Li, Y., Li, M., Zhang, G., et al. (2020). Multiparametric MRI-based radiomics nomogram for predicting lymph node metastasis in

early-stage cervical cancer. *J. Magn. Reson. Imaging* 52, 885–896. doi: 10.1002/jmri.27101

Xu, W., Liu, X., Leng, F., and Li, W. (2020). Blood-based multi-tissue gene expression inference with Bayesian ridge regression. *Bioinformatics* 36, 3788–3794. doi: 10.1093/bioinformatics/btaa239

Ye, Y., Zhang, R., Zheng, W., Liu, S., and Zhou, F. (2017). RIFS: a randomly restarted incremental feature selection algorithm. *Sci. Rep.* 7:13013. doi: 10.1038/s41598-017-13259-6

Yuan, Z., Zha, X., and Zhang, X. (2020). Adaptive multi-type fingerprint indoor positioning and localization method based on multi-task learning and weight coefficients k-nearest neighbor. *Sensors* 20:5416. doi: 10.3390/s20185416

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fmolb.2021.670241

Autophagy-Related Gene Pairs Signature for the Prognosis of Hepatocellular Carcinoma

Yiming Luo[1,2†], Furong Liu[1,2†], Shenqi Han[1,2], Yongqiang Qi[1,2], Xinsheng Hu[1,2], Chenyang Zhou[1,2], Huifang Liang[1,2] and Zhiwei Zhang[1,2,3*]*

[1]Hepatic Surgery Center, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Hubei Key Laboratory of Hepato-Pancreato-Biliary Diseases, Wuhan, China, [3]Key Laboratory of Organ Transplantation, Ministry of Education, NHC Key Laboratory of Organ Transplantation, Key Laboratory of Organ Transplantation, Chinese Academy of Medical Sciences, Wuhan, China

Hepatocellular carcinoma (HCC) has been recognized as the third leading cause of cancer-related deaths worldwide. There is increasing evidence that the abnormal expression of autophagy-related genes plays an important role in the occurrence and development of HCC. Therefore, the study of autophagy-related genes can further elucidate the genetic drivers of cancer and provide valuable therapeutic targets for clinical treatment. In this study, we used 232 autophagy-related genes extracted from the Human Autophagy Database (HADb) and Molecular Signatures Database (MSigDB) to construct 1884 autophagy-related gene pairs. On this basis, we developed a prognostic model based on autophagy-related gene pairs using least absolute shrinkage and selection operator (LASSO) Cox regression to evaluate the prognosis of patients after liver cancer resection. We then used 845 liver cancer samples from three different databases to test the reliability of the risk signature through survival analysis, receiver operating characteristic (ROC) curve analysis, univariate and multivariate analysis. To further explore the underlying biological mechanisms, we conducted an enrichment analysis of autophagy-related genes. Finally, we combined the signature with independent prognostic factors to construct a nomogram. Based on the autophagy-related gene pair (ARGP) signature, we can divide patients into high- or low-risk groups. Survival analysis and ROC curve analysis verified the validity of the signature (AUC: 0.786—0.828). Multivariate Cox regression showed that the risk score can be used as an independent predictor of the clinical outcomes of liver cancer patients. Notably, this model has a more accurate predictive effect than most prognostic models for hepatocellular carcinoma. Moreover, our model is a powerful supplement to the HCC staging indicator, and a nomogram comprising both indicators can provide a better prognostic effect. Based on pairs of multiple autophagy-related genes, we proposed a prognostic model for predicting the overall survival rate of HCC patients after surgery, which is a promising prognostic indicator. This study confirms the importance of autophagy in the occurrence and development of HCC, and also provides potential biomarkers for targeted treatments.

Keywords: hepatocellular carcinoma, autophagy-related gene, gene pairs, prognosis signature, nomogram

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Xiaoping Li,
Jiangmen Central Hospital, China
Sayan Chakraborty,
*Institute of Molecular and Cell Biology (A*STAR), Singapore*

***Correspondence:**
Huifang Liang
lianghuifang1997@126.com
Zhiwei Zhang
zwzhang@tjh.tjmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *20 February 2021*
Accepted: *05 May 2021*
Published: *20 May 2021*

Citation:
Luo Y, Liu F, Han S, Qi Y, Hu X, Zhou C, Liang H and Zhang Z (2021) Autophagy-Related Gene Pairs Signature for the Prognosis of Hepatocellular Carcinoma. Front. Mol. Biosci. 8:670241. doi: 10.3389/fmolb.2021.670241

INTRODUCTION

Hepatocellular carcinoma, the predominant primary tumor of the liver, has been recognized as the third leading cause of cancer-related death worldwide (Forner et al., 2018). Many patients are diagnosed when the cancer has already metastasized and a series of severe complications have occurred, indicating that the liver cancer has reached an advanced stage (Cabibbo et al., 2010). In site of recent advances in surgical resection or liver transplantation, the 5-year survival rate of HCC patients remains relatively low (Bosetti et al., 2014; Singal and El-Serag, 2015). Therefore, extensive analysis is urgently needed to identify reliable prognostic biomarkers and develop therapies that can target the major oncogenes of HCC.

Autophagy is an important intracellular selective recycling mechanism through which cell components are transported to lysosomes for degradation to recover materials and provide energy (Mizushima, 2018). Due to its unique functions, autophagy is closely related to many human diseases, including immune diseases (Gukovskaya et al., 2017; Yang et al., 2017), neurodegenerative diseases (Hu et al., 2017; Moloudizargari et al., 2017) and different types of cancer (White, 2015; Gugnoni et al., 2016). A large number of studies have shown that autophagy has two opposite effects during the occurrence of common cancers, especially in HCC (Czaja et al., 2013). At the same time, there is increasing evidence that abnormal expression of autophagy-related genes plays a pathogenic role in the development of multiple human diseases, including cancer (Mizushima, 2018). As autophagy plays a key role in hepatocellular carcinoma, prognostic signatures based on autophagy-related genes can help us explore the genetic control mechanism of hepatocellular carcinoma and provide valuable therapeutic targets (Lin et al., 2018). However, few studies have used autophagy-related genes to construct prognostic signatures for HCC.

In this study, we developed and validated a promising prognostic model for HCC based on autophagy-related gene pairs. First, we collected sequencing data of autophagy-related genes from three independent groups to screen for candidate gene pairs. We screened out nine gene pairs that are closely related to the patients' prognosis and used them to construct a gene-pair model. After calculating the risk scores of the patients using the model, we divided the patients into two groups with significant differences in prognosis. In a series of subsequent verifications, our model showed a good prognostic ability for HCC patients. Our promising prognostic model confirms the important role of autophagy in HCC and provides potential therapeutic targets.

MATERIALS AND METHODS

Data Sources

We obtained an RNA-seq dataset (n = 377), which was used as a training set to build the model, and the corresponding clinical information of HCC patients from The Cancer Genome Atlas (TCGA) using the UCSC Xena browser (https://xenabrowser.net/). The validation set was based on a second RNA-seq dataset (n = 243) downloaded from the International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org) and a microarray dataset (GSE14520, n = 225) from GEO database (http://www.ncbi.nlm.nih.gov/geo). We extracted 232 autophagy-related genes from the Human Autophagy Database (HADb, http://www.autophagy.lu/index.html) and 394 from the GO_AUTOPHAGY gene set in the Molecular Signatures Database v7.1 (MSigDB, http://software.broadinstitute.org/gsea/msigdb). Our autophagy-related gene set was formed by the integration of these two gene sets.

Data Preprocessing

Specimens of HCC patients who survived less than one month or whose clinical data were incomplete were not included. We removed the data of normal tissue samples and only kept the data of the primary tumor. When multiple specimens were taken from the same patient, the average gene expression value was used to represent the patient's gene expression level. Only the sequencing data of autophagy-related genes were retained. When the same gene was matched by multiple probes, we used the average expression value of multiple probes to indicate the expression level of the gene. For the RNA-seq data from TCGA database, we excluded HCC samples in which more than half of the gene probe expression values were zero. The expression profiles of common autophagy-related genes were screened from the three data sets.

Establishment of the Prognostic Model Based on Autophagy-Related Genes

We compared the expression values of autophagy-related genes in each sample to obtain the score of each ARGP. The expression values of autophagy-related genes in each sample were compared in pairs to calculate the score of each ARGP. In a pairwise comparison, if the previous value is greater than the next value, the output is 1, and if it is not, the output is 0. We excluded ARGPs that scored 0 or 1 in more than 90% of the samples in each dataset, and the remaining ARGPs were used to establish the prognostic model for HCC. First, we performed univariate Cox regression analysis using the R package "survival" to select gene pairs that are related to the overall survival of HCC patients in TCGA. Differences with $p < 0.001$ were considered statistically significant. To minimize the risk of overfitting, we used "glmnet" R package to conduct LASSO penalized Cox regression (3,000 iterations) to calculate the frequency of the models. The gene pair model with the highest frequency among the iterations was used to establish a prognostic model. Stepwise multivariate Cox regression analysis was performed.

Validation and Assessment of the Autophagy-Related Gene Pair Signature

After calculating the risk score in every dataset, the patients were classified into high-risk and low-risk groups according to the median value of the risk score. Kaplan–Meier survival analysis ($p < 0.05$) was used to analyze the over survival (OS) of the high-risk and low-risk groups. After drawing ROC curves for 1, 3 and

TABLE1 | Clinical and pathologic factors of the datasets used in this study.

	TCGA(n,%)	ICGC (n,%)	GSE14520 (n,%)
Total	346	241	221
Age			
Median age (years)	61	69	50
Mean age (years)	59.44	67.49	50.819
Gender			
Female	110 (31.79%)	61 (25.31%)	30 (13.57%)
Male	236 (68.21%)	180 (74.69%)	191 (86.43%)
TNM stage			
Stage I	163 (47.11%)	36 (14.94%)	93 (42.08%)
Stage II	78 (22.54%)	110 (45.64%)	77 (34.84%)
Stage III	80 (23.12%)	74 (30.71%)	49 (22.17%)
Stage IV	3 (0.87%)	21 (8.71%)	0
NA	22 (6.36%)	0	2 (0.90%)
Grade			
G1	53 (15.32%)		
G2	162 (46.82%)		
G3	114 (32.95%)		
G4	12 (3.47%)		
NA	5 (1.45%)		
Survival status			
Alive	222 (64.16%)	199 (82.57%)	136 (61.54%)
Dead	124 (35.84%)	44 (18.26%)	85 (38.46%)
Median follow-up time (days)	632.5	780	1,569

Abbreviations: TCGA, TCGA LIHC dataset; ICGC, ICGC LIHC dataset;

5 years, we used the area under curve (AUC) value to verify the accuracy and sensitivity of this model. The closer the AUC value is to 1, the better the predictive effect of the prognostic model. To perform multivariate Cox regression analysis, available clinical and pathological data were integrated with the ARGP signature. Tumor stage, grade, age, and sex were regarded as continuous variables. In addition, we selected three representative prognostic gene models for HCC. Our model was compared with these existing models using the 5-year multiple ROC curves. The respective AUC values were used to estimate the prognostic accuracy of each signature.

Gene Set Enrichment Analysis

In order to further reveal the biological mechanisms through which the identified autophagy-related genes contribute to the development of HCC, we used the MSigDB hallmark gene set (h.all.v7.1.symbols.gmt) to run gene set enrichment analysis (GSEA). We used an FDR value < 0.25, a nominal (NOM) $p < 0.05$, and $|$ NES $|$ > 1 as the screening criteria to identified signaling pathways that are highly related to the model genes.

Construction of a Nomogram

Independent prognostic factors that are highly correlated with OS in HCC patients ($p < 0.05$) were screened out using univariate and multivariate Cox regression analyses. We then integrated these independent prognostic factors using the R package "RMS" and constructed the predictive nomogram and corresponding calibration diagram for 1, 3, and 5 years. The calibration maps were verified by calibration and discrimination. The expected possibility of collinearity was plotted graphically as an observable indicator to assess the alignment of the nomogram. The closer the calibration curve was to the reference line (diagonal line), the better the predictive effect of the nomogram.

Statistical Analysis

All statistical analyses were performed using R software (version 3.6.3, https://www.r-project.org/). The OS of the HCC patients in the low- and high-risk groups was compared using the log-rank test, and the Kaplan–Meier survival curves were drawn using the R package "survminer" (version: 0.4.6). The gene pair prognostic signature was established based on the LASSO Cox regression algorithm using the R package "glmnet" (version: 3.0.2). ROC curves and multiple ROC curves were drawn using the R packages "survivalROC" and "timeROC", respectively.

RESULTS

Construction of the Autophagy-Related Gene Pair Signature

After eliminating duplicate genes from HADb and MSigDB, an autophagy-related gene set comprising 527 genes was obtained. As shown in **Table 1**, this study included 808 HCC patients from three cohorts. ARGPs were constructed using a total of 269 autophagy-related genes that are represented in all three data sets.

We removed ARGPs with a score of 0 or 1 in more than 90% of the samples in all datasets, resulting in 1885 ARGPs. We used univariate Cox regression analysis to screen 117 prognostic ARGPs that were significantly associated with overall survival ($p < 0.001$), and established a prognostic gene model of ARGP using Lasso penalty score Cox regression in the TCGA dataset. After multivariate Cox regression analysis, 9 ARGPs were selected

TABLE 2 | Univariate and multivariate analyses of prognostic factors in terms of OS.

datasets	Variable	Univariate		Multivariate	
		HR (95%CI)	p-value	HR (95%CI)	p-value
TCGA	Risk score (low risk vs high risk)	4.77 (3.46–6.57)	6.16E-24	1.25 (1.19–1.31)	7.98E-18
	Gender (male vs female)	0.75 (0.51–1.1)	0.141,902	0.82 (0.55–1.23)	0.338,054
	Grade (G1 and G2 vs G3 and G4)	1.12 (0.86–1.45)	0.4012	1.16 (0.89–1.51)	0.271,925
	Age (<60 vs ≥60)	1.01 (0.99–1.02)	0.492,155	1.01 (0.99–1.02)	0.228,735
	Stage (I and II vs III and IV)	1.80 (1.46–2.22)	4.72E-08	1.59 (1.27–1.99)	6.14E-05
ICGC	Risk score (low risk vs high risk)	1.11 (1.06–1.16)	1.43E-05	1.11 (1.06–1.17)	4.88E-05
	Gender (male vs female)	0.46 (0.24–0.86)	0.014548	0.34 (0.17–0.65)	0.001113
	Age (<70 vs ≥70)	1 (0.97–1.03)	0.914,617	0.99 (0.96–1.03)	0.764,419
	Stage (I and II vs III and IV)	2 (1.38–2.91)	0.000268	2.15 (1.48–3.13)	5.49E-05
GSE14520	Risk score (low risk vs high risk)	1.05 (1.02–1.09)	0.001776	1.03 (1–1.07)	0.048104
	Gender (male vs female)	1.66 (0.80–3.45)	0.172,844	1.30 (0.62–2.73)	0.487,211
	Age (<50 vs ≥ 50)	0.99 (0.97–1.01)	0.356,607	1.00 (0.97–1.02)	0.741,149
	Stage (I and II vs III and IV)	2.38 (1.78–3.17)	3.40E-09	2.23 (1.66–3.00)	9.76E-08

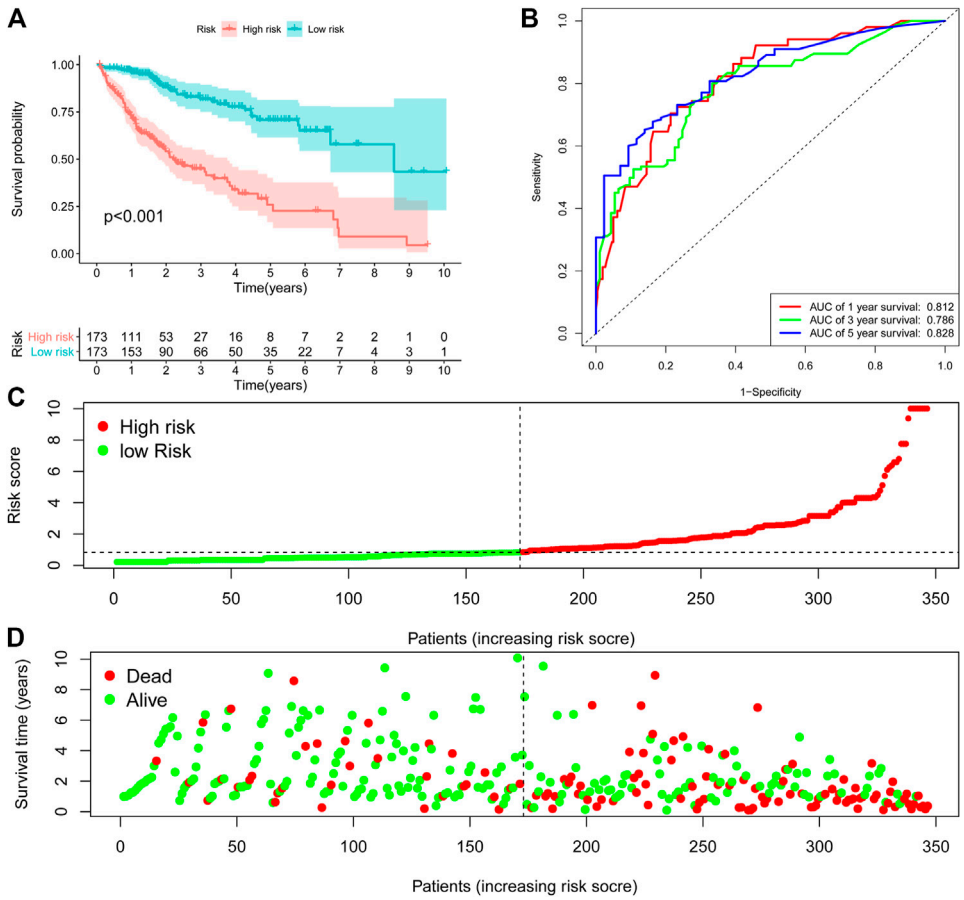


FIGURE 1 | Establishment 9-ARGP signature in the TCGA database. **(A)** Kaplan-Meier survival curves showed the prognostic value of the risk signature between low-risk group (n = 263) and high-risk group (n = 262). **(B)** ROC curves were used to assess the efficiency of the risk signature for predicting 1-, 3- and 5-y survival. **(C)** The risk scores distribution in the TCGA database. **(D)** The patients survival status in the TCGA database.

to construct the most stable prognostic signatures, and the corresponding coefficients were used to calculate the risk score for our datasets. Details of the 9-ARGP prognostic model are listed in **Table 2**.

The 9-ARGP signature in the TCGA dataset reflected the postoperative prognosis of patients very well (**Figure 1**). We calculated the risk score of each HCC patient in TCGA according to the prognostic characteristics of autophagy, and then divided

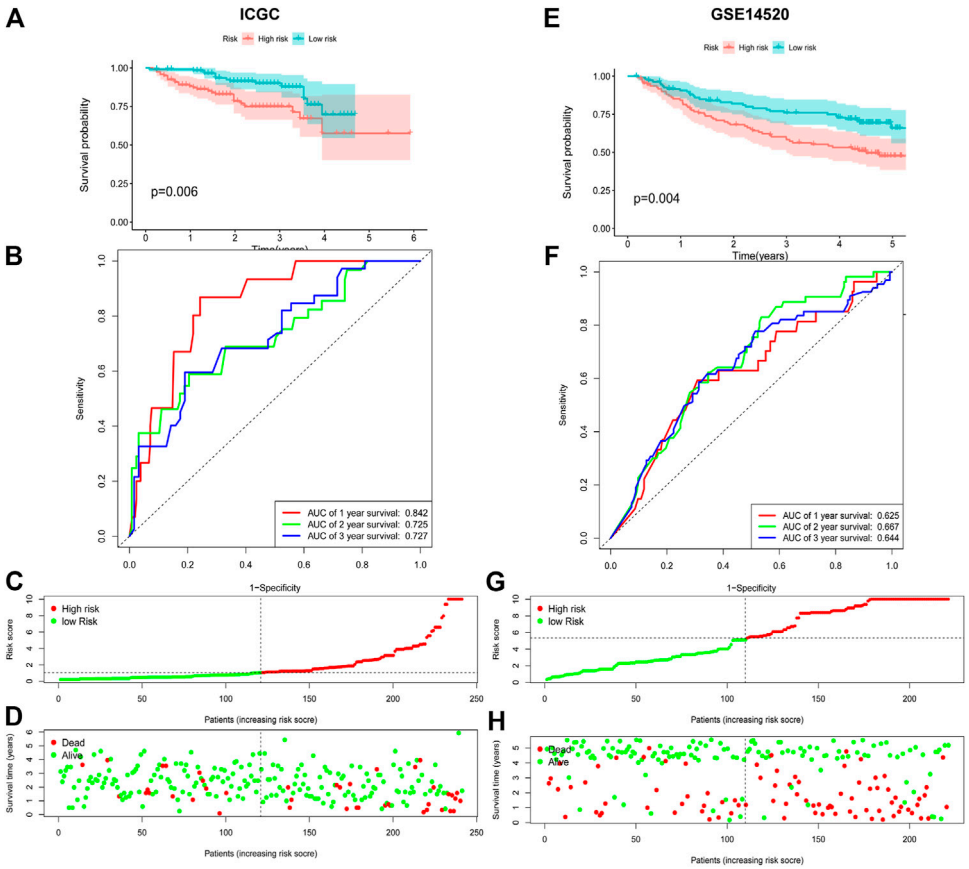


FIGURE 2 | Evaluating the efficiencies of the risk signature in the ICGC and GSE14520 data sets. **(A,E)**, Kaplan-Meier survival curves showed the prognostic value of the risk signature in ICGC data set **(A)**. low-risk group, $n = 120$; high-risk group, $n = 121$; $p < 0.05$) and GSE14520 database **(E)**. low-risk group, $n = 111$; high-risk group, $n = 110$; $p < 0.001$). **(B, F)**, ROC curves evaluated the efficiency of the risk signature for predicting 1-, 2- and 3-y survival in ICGC data set **(B)** and GSE14520 database **(F)**. **(C,G)**, The risk scores distribution in the ICGC data set **(C)** and GSE14520 database **(G)**. **(D,H)**, The patients' survival status in the ICGC data set **(D)** and GSE14520 database **(H)**.

the 346 cases into high- and low-risk groups based on the median score. As shown in **Figure 1A**, the low-risk group had a significantly lower mortality rate than the high-risk group (95%CI: 8.09–21.07, $p < 0.0001$). We evaluated the specificity and sensitivity of the prognostic model using time-dependent ROC curve analysis. The AUC values for 1, 3, and 5 years after surgery reached 0.812, 0.786, and 0.828, respectively, which demonstrated that our ARGP prognostic signature has a promising predictive ability (**Figure 1B**). The distribution of the autophagy-related prognostic model for patients in the TCGA data set is shown in **Figure 1C,D**.

Validation of the Autophagy-Related Gene Pair Signature

In order to further verify its predictive power, we applied the prognostic signature to the ICGC database (containing 243 HCC cases) and the GSE14520 database (containing 225 HCC cases) for analysis. According to the median risk value calculated using the 9-ARGP prognostic signature, HCC patients in the two databases were assigned into high- and low autophagy-based risk groups, respectively. Consistent with the conclusions obtained using the training set, the OS of the high-risk groups in the two validation datasets was significantly lower than that of the low-risk group ($p < 0.05$) (**Figure 2A,E**). In the ICGC cohort, the AUC values of the prognostic model were 0.842 at 1 year, 0.725 at 2 years, and 0.727 at 3 years (**Figure 2B**), while in the GSE14520 cohort they were 0.625 at 1 year, 0.667 at 2 years, and 0.644 at 3 years (**Figure 2F**). **Figure 2C,D** show the distribution of risk scores corresponding to gene expression levels in the ICGC cohort, while **Figure 2G,H** shows the corresponding data for the GSE14520 cohort. In univariate Cox regression analysis, TNM staging and the ARGP signature risk score were significantly related to the OS in the three cohorts (HR > 1.00, $p < 0.05$). After correcting for age, gender, grade and TMN staging in multivariate Cox regression analysis, the ARGP signature risk score was still significantly associated with the OS as an independent prognostic factor in the TCGA dataset (HR: 1.25, 95% CI: 1.19–1.31, $p < 0.0001$), ICGC dataset (HR: 1.11, 95%CI: 1.06–1.17, $p< 0.001$) and GSE14520 dataset (HR: 1.03, 95%CI: 1–1.17, $p = 0.07$) (**Table 2**).

The gene pairs with the largest coefficients were STAM/TP53, PLOD2/CDKN1B and NTHL1/BLCR5, since the large

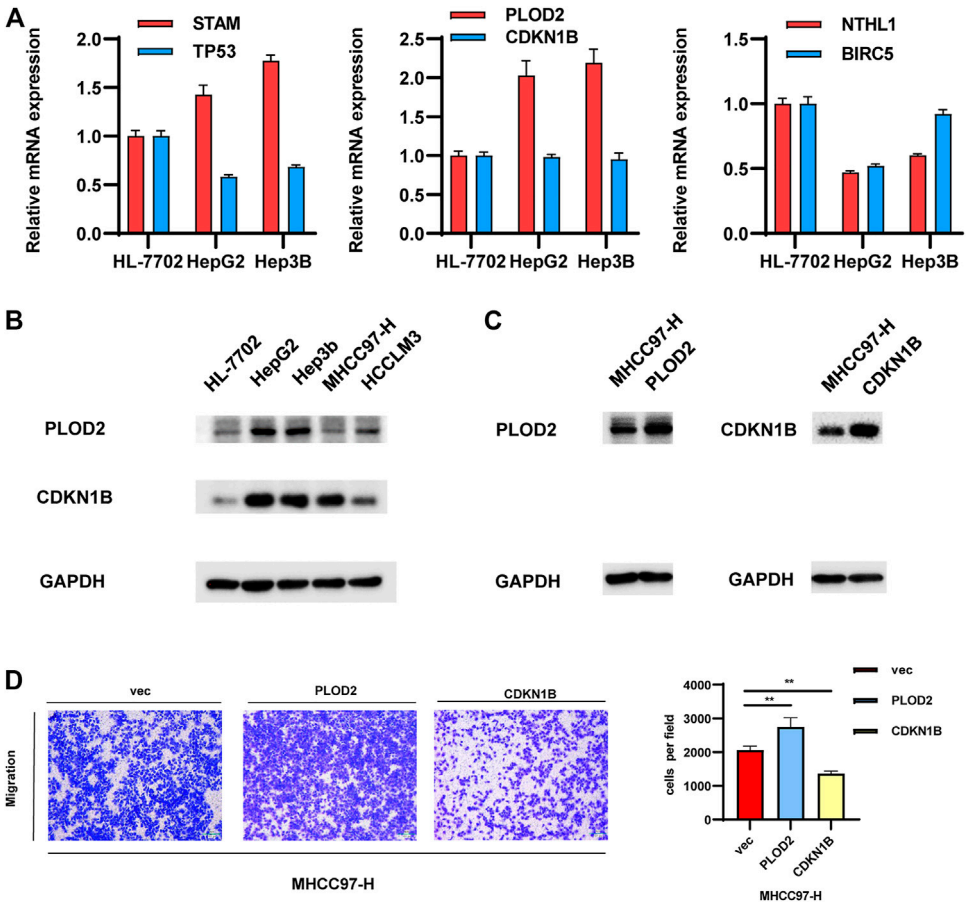


FIGURE 3 | Validation of the gene pairs that make up the model. **(A,B)** Examining the expression levels of representative gene pairs with the large coefficient in normal-liver cell line and in several common HCC cell lines. **(C,D)**. Validate the effects of representative gene pair in HCC cell lines.

coefficients indicate that they have the greatest influence on the model. We examined the expression levels of these genes in a normal liver cell line and in several common HCC cell lines. As shown in **Figure 3A**, the qPCR results showed that the values of STAM/TP53 and PLOD2/CDKN1B in the HCC cell lines HepG2 and Hep3B were significantly higher than in the normal liver cell line HL7702, which further supported the significance of our model. The coefficient of the gene pair NTHL1/BLCR5 was −0.52, and was significantly lower in the HCC cell lines HepG2 and Hep3B than in the normal liver cell line HL7702, which is also consistent with this conclusion. At the protein level, we found that the expression of PLOD2 in liver cancer cell lines was higher than in normal liver cells, and it was also higher in non-invasive liver cancer cells (**Figure 3B**).

Further experiments were performed on the representative gene pair Plod2/CDKN1B. We overexpressed these genes in MHCC97-H cells, and the protein levels of PLOD2/CDKN1B were confirmed to be increased after transfection with pcDNA-PLOD2 (oe-PLOD2) and pcDNA-CDKN1B (oe-CDKN1B) (**Figure 3C**). The results of the transwell assay showed that PLOD2 can promote HCC migration, while CDKN1B had the opposite effect. Since their coefficient is greater than zero, this experimental result was consistent with the previous conclusion (**Figure 3D**).

We verified the expression levels of the representative gene pair PLOD2 /CDKN1B in 45 liver cancer samples, and combined with the OS and RFS of the corresponding patients, we found that PLOD2 /CDKN1B can better predict the prognosis of the patients. The results are shown in **Figure 7**. Samples were collected from surgical biopsies of patients who underwent radical resection of liver cancer without preoperative treatment at Tongji Hospital in Wuhan, China, between 2015 and 2018. The Ethics Committee of Wuhan Tongji Hospital authorized this study on patient tissues with written informed consent of the patients.

Comparison With Representative Published Prognostic Models for Hepatocellular Carcinoma

Our ARGP prognostic marker was compared with three published representative gene prognostic markers (Lin et al., 2018; Long et al., 2018; Wang et al., 2020) using ROC curves for 5-year OS. All the data for validation were derived from TCGA.

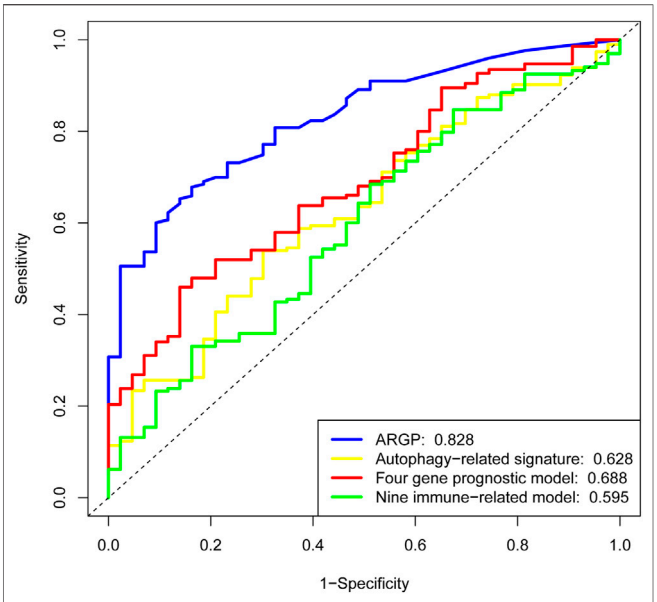


FIGURE 4 | Determination of the receiver operating characteristic (ROC) for different prognostic signatures. The AUC values for the ARGP model, Autophagy-related signature model, four prognostic lncRNA model, and nine immune-related model were 0.828, 0.628, 0.688, and 0.595, respectively. This result indicates that our signature possesses a higher predictive efficacy and accuracy than the other models.

As shown in **Figure 4**, the AUC value was 0.828 for our prognostic signature, which was obviously more predictive and accurate than the existing autophagy-related signature (AUC = 0.628), the four-gene prognostic model (AUC = 0.688) and the nine immune-related gene model (AUC = 0.595).

Construction of a Nomogram for Predicting the Over Survival

Multivariate Cox regression analysis showed that only TNM stage and the ARGP signature were significant independent prognostic factors for OS (**Figure 5A**). We attempted to provide a method to more intuitively and accurately predict the survival of HCC patient, which can aid individual clinical decision-making and selection of treatment options. Therefore, a predictive nomogram was constructed based on multivariate Cox regression analysis and combined with two independent prognostic factors (**Figure 5B**). The scores of each independent prognostic factor were calculated according to the different degree of influence of each independent prognostic factor on the clinical outcomes of the patients, after which the scores were summed up to obtain the total score. Finally, the 1, 3, and 5-year survival rates were predicted based on the functional relationship between the total score and the survival rate. According to the calibration curves of the 1-, 3-, and 5-year nomograms, which were all close to the optimal prediction curve, the predicted OS rate was highly consistent with the actual observed values (**Figure 5C–E**).

Physiological Signal Channel Correlated With the Autophagy-Related Gene Pair Model

We performed GSEA in the high- and low-risk groups from the TCGA cohort, divided according to the median risk value. A total of 12 cancer hallmark gene sets were identified in the high-risk group (**Figure 6A**). Some of these pathways are "MYC_TARGETS", "GLYCOLYSIS" and "DNA_REPAIR", indicating that these signaling pathways are closely related to the progression of HCC. To make the results of enrichment analysis more intuitive, we visualized the significance, the number of included genes and the enrichment score in a bubble chart with different colors, sizes and locations (**Figure 6B**).

DISCUSSION

Although many environmental or genetic risk factors associated with the occurrence of HCC have been elucidated, the molecular mechanisms underlying the metastasis and recurrence of HCC remain unclear. Consequently, hepatocellular carcinoma remains one of the deadliest malignancies in the world, with exceptionally high recurrence and low survival. In recent years, the application of high-throughput technology and the emergence of large-scale cancer gene expression databases have deepened our understanding of the characteristics of liver cancer and provided the possibility for us to predict postoperative survival rates based on the genetic phenotypes of the individual tumor. Based on gene expression profiles, some studies have established prognostic markers for predicting the survival after liver cancer surgery, while others have explored molecular subtypes of liver cancer based on multi-group analysis (Liu et al., 2020). However, these results are far from clinical application. Due to the diversity of data types among different databases, gene expression levels of different sequencing platforms need to be appropriately standardized before use, but it is still difficult to completely overcome biological heterogeneity and eliminate the technical bias of cross-sequencing platforms. Thus, improving the genetic models and selecting stable specific prognostic markers is still the main task of current liver cancer research.

In this study, we established a prognostic model for HCC based on 9 autophagy-related gene pairs (ARGPs) and validated it across different platforms using the independent datasets ICGC and GSE14520. Our 9-ARGP signature proved to be a significant and excellent predictor in a series of validation analyses, successfully dividing patients into high and low-risk groups with significantly different prognostic outcomes. Compared with three other existing prognostic models for hepatocellular carcinoma, our model showed a more accurate predictive power. We further combined the model with selected significant pathological features. According to the results, the 9-ARGP signature is a powerful complement to HCC staging indicators, and their combination provides a better prognostic performance.

ARGPs were generated based on the pairwise comparison of gene pairs, so there is no need to consider batch differences among different databases. Furthermore, the correlation

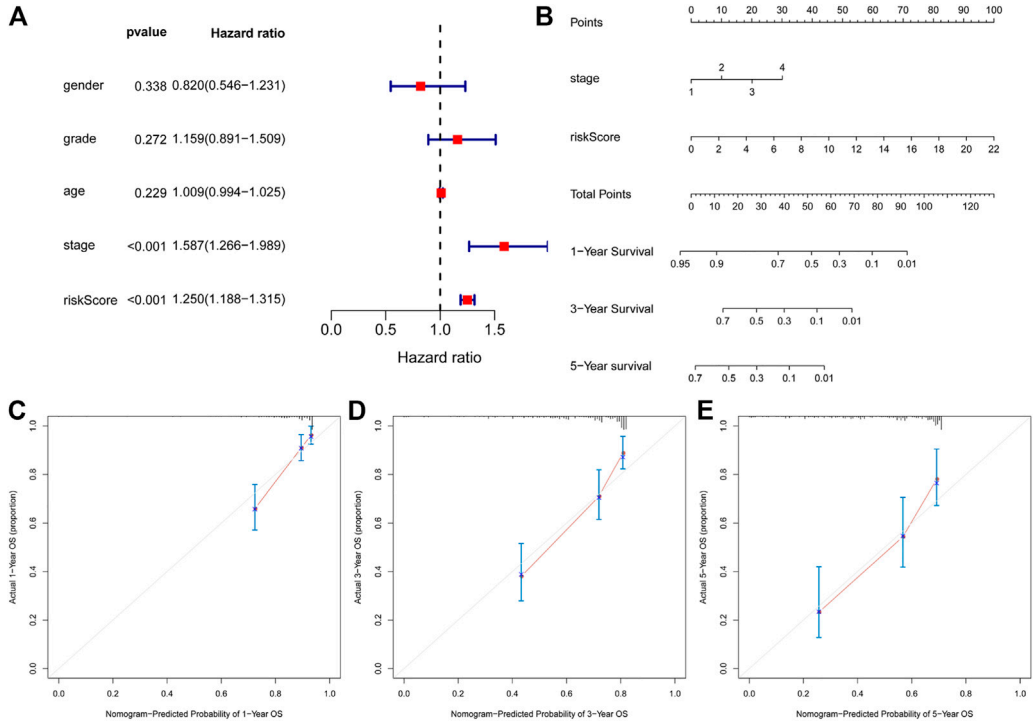


FIGURE 5 | Construction of a nomogram for predicting 1-, 3- and 5-y survival of HCC. **(A)**, Multivariate Cox regression analyses evaluated the contribution of each variable to HCC survival in the TCGA cohort. **(B)** nomogram for predicting 1-, 3- and 5-y survival rate of HCC patients was established. **(C–E)**, Calibration curves showed the probability of 1- **(C)**, 3- **(D)** and 5-y survival **(E)** between the prediction and the observation in the TCGA cohort.

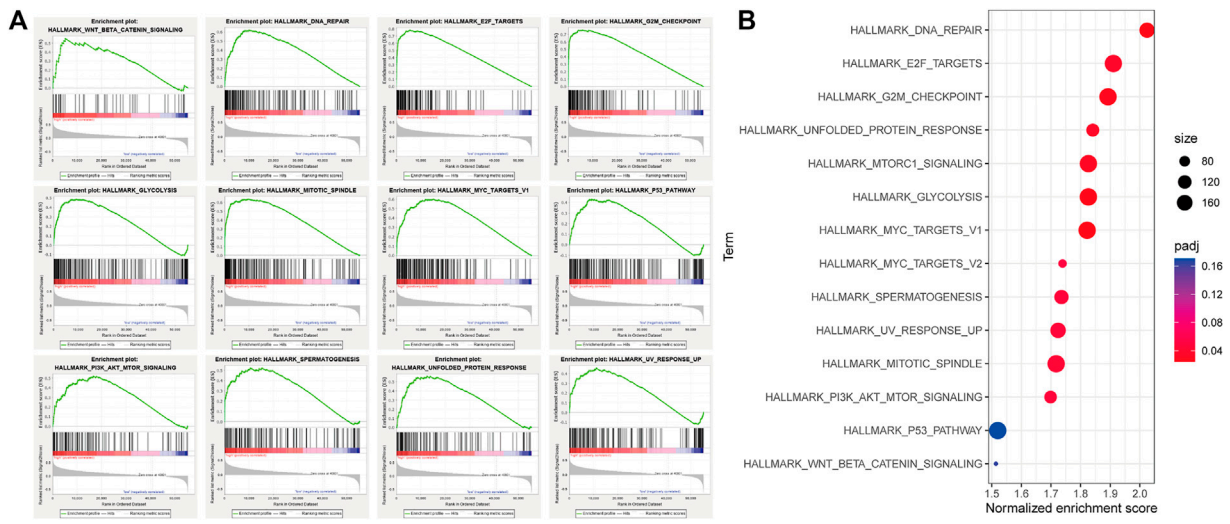


FIGURE 6 | Gene set enrichment analysis (GSEA) between high and low autophagy risk groups. **(A)**, Twelve cancer hallmark gene sets are enriched in the high autophagy risk group in patients with HCC ($p < 0.05$, FDR < 0.25, |NES| > 1). **(B)** bubble chart for visualizing the GSEA result.

coefficients of ARGP were calculated based on the gene expression in the same sample, so the data does not need to be standardized. Therefore, our prognostic signature can be used for the individualized prediction of postoperative survival in HCC patients, and it can be more easily applied in clinical practice.

Autophagy has been reported to play a key role in promoting the formation of liver cancer (Czaja et al., 2013). Our 9-ARGP signature contains 17 autophagy-related genes in total. These genes are directly or indirectly related to the occurrence and prognosis of HCC, which has been described in many studies. To provide

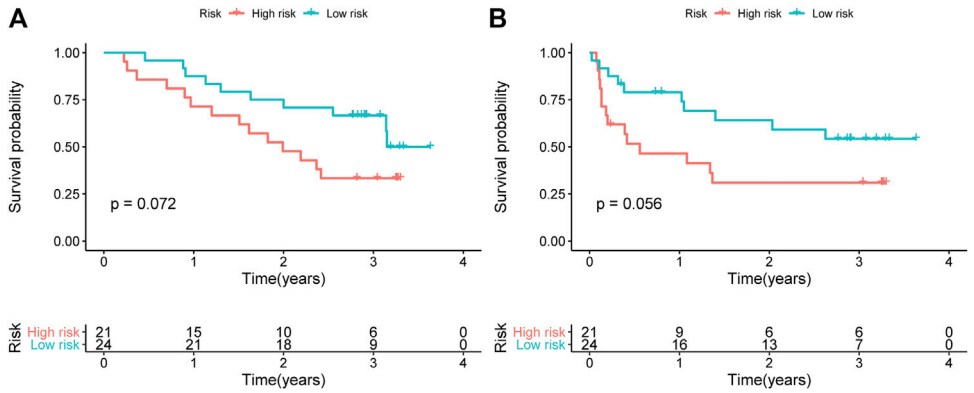


FIGURE 7 | Verified the expression level of the representative PLOD2/CDKN1B in 45 liver cancer samples. Kaplan-Meier curves of **(A)** OS and **(B)** RFS for HCC patients stratified by the signature risk groups in the Tongji cohorts.

theoretical support for this statistical signature, we further explored the genes included in the model. The following studies support a mechanistic link between our model and HCC.

CDKN1A/p21 and CDKN1B/p27 are members of a family of cyclin-dependent kinase inhibitors that act as tumor suppressors and inhibit cell proliferation. CDKN1A expression is strictly controlled by the tumor suppressor protein p53 (TP53), which mediates G1 phase arrest in response to various stress factors. CDKN1A can be considered as an independent factor for the development of liver cancer, and in patients with cirrhosis, high expression of CDKN1A may be associated with the occurrence of liver cancer (Wagayama et al., 2002). However, CDKN1A expression is beneficial in patients after hepatectomy and may be an independent prognostic factor for patient survival (Kao et al., 2007). Interruption of the P53-CDKN1A cell cycle pathway may lead to further tumor progression (Lee et al., 2004). The activation of CDKN1A gene expression induced by RNA may have a significant potential for the treatment of HCC and other cancers (Wu et al., 2011). In addition, the subcellular localization of CDKN1A was found to contribute to the development of HCC (Qiu et al., 2011). CDKN1B shares a limited similarity with CDKN1A. Furthermore, reduced CDKN1B expression often predicts poor clinical outcomes in HCC (Huang et al., 2011; Matsuda et al., 2013), and CDKN1B silencing increases the viability of HCC cells (Xu et al., 2019). This is consistent with previous findings that CDKN1B potentially plays an active role as a negative regulator in the early stages of HCC progression (Ito et al., 1999). The risk of HCC is increased by CDKN1A polymorphisms, alone or in combination with CDKN1B polymorphisms (Liu et al., 2013). These studies indicate that both CDKN1A and CDKN1B are closely related with the occurrence of liver cancer and can be used as prognostic biomarkers. BIRC5, a member of the inhibitor of apoptosis (IAP) gene family, promotes cancer development by inhibiting the apoptosis of HCC cells (Zhang et al., 2014), promoting cell proliferation (Sun et al., 2013), enhancing chemoradiotherapy resistance (Liu et al., 2013b) and inducing stromal angiogenesis in the tumor (Fernandez et al., 2014). Similarly, BIRC5 was reported to be directly associated with autophagosome formation and contribute to the survival of HCC cells (Chang et al., 2014). DLC-1 is a GTPase-activating protein that targets Rho (Kim et al., 2007), and as a tumor suppressor, DLC-1 is not only involved in hepatocarcinogenesis, but also inhibits the cancer progression and oncogenic autophagy of hepatocellular carcinoma (Wu et al., 2018) (Zhou et al., 2004). The protein encoded by Fas is a member of the TNF receptor superfamily. It plays a central role in the physiological regulation of programmed cell death and is associated with various malignancies and immune system diseases. Fas stimulation may contribute to the survival or proliferation of HCC cells (Okano et al., 2003). However, downregulation of Fas expression by HBV might inhibit the apoptosis of HCC cells (Zou et al., 2015).

The remaining genes in the signature are also associated with liver cancer in different ways and play a role in our signature together with these genes in the form of gene pairs. Some of these genes may have a more important effect on the expression imbalance than a single gene with abnormal expression. GSEA was used to analyze the differential expression of genes in the high- and low-risk groups. Consistent with previous reports, the expression of genes related to a number of signaling pathways was significantly different in the high-risk group, including "PI3K-AKT-mTOR signaling" (Zhou et al., 2011; Wang et al., 2017), "DNA_REPAIR" (Lin et al., 2016), "G2M checkpoint" (Yin et al., 2017), and "GLYCOLYSIS" (Qin et al., 2018). In addition, we also found that "UNFOLDED_PROTEIN_RESPONSE", "E2F_TARGETS", "MTORC1_signaling" and other hallmarks were also enriched in the high-risk group. As a central tumor suppressor, p53 protects the genome by coordinating multiple DNA damage response (DDR) mechanisms (Williams and Schumacher, 2016). Many mechanisms of DNA repair in cells are influenced by p53. The coordination of DNA repair is an important process through which p53 inhibits tumor development (Janic et al., 2018). It is therefore perhaps unsurprising that the p53 pathway was also enriched in the high-risk group according to the GSEA analysis. Next, we collected the information of patients with P53 mutant HCC from the TCGA data set in the CBioPortal database. We found that P53 mutations were present in 32% of HCC cases, and the risk score of in the mutant group was significantly higher than that in the non-mutant group by calculating the levels of autophagy-related genes.

The corresponding results are provided in **Supplementary Figure S3**. Besides, we found that β-catenin mutations were present in 26% of HCC cases, but we were unable to draw meaningful conclusions by calculating the levels of autophagy-related genes. The corresponding results are provided in **Supplementary Figure S4**.

In spite of the exciting finding, this study also has several limitations. First, the data were sourced from a limited number of databases, and are not sufficiently broad to prove the universality of the signature. Secondly, the training dataset samples used to establish the autophagy characteristics were derived from previous retrospective studies, and we also need a prospective cohort to verify the results. Prospective studies are needed to further verify the clinical use and biological function of the signature. Future studies will incorporate more datasets and integrate other clinical and pathological indicators, which may provide more useful and accurate results.

CONCLUSION

Based on multiple pairs of autophagy-related genes, we proposed a prognostic model for predicting the overall survival of HCC patients after surgery. The gene-air signature is a promising prognostic indicator. The credibility of the model was verified using two unrelated verification sets. Compared with most other existing prognostic models, our model shows a more accurate prediction effect. At the same time, this study further proves the importance of autophagy in the occurrence and development of HCC, and also provides potential therapeutic targets.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was approved by the ethics review board of the Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology and conforms to the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

YL and FL performed the bioinformatics analysis, cell experiments, wrote the manuscript and designed the figures. SH, FL, YQ, XH and CZ collected the related references and participated in discussion. HL and ZZ provided guidance and revised this manuscript. All authors read and approved the final manuscript.

FUNDING

Clinical medicine research plan of Tongji Hospital (No. 2019CR202); Chen Xiao-ping Foundation for the Development of Science and Technology of Hubei province (CXPJJH11800001-2018104); Hubei Natural Science Foundation of China (No. 2015CFB462).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.670241/full#supplementary-material

REFERENCES

Bosetti, C., Turati, F., and La Vecchia, C. (2014). Hepatocellular Carcinoma Epidemiology. *Best Pract. Res. Clin. Gastroenterol.* 28, 753–770. doi:10.1016/j.bpg.2014.08.007

Cabibbo, G., Enea, M., Attanasio, M., Bruix, J., Craxì, A., and Cammà, C. (2010). A Meta-Analysis of Survival Rates of Untreated Patients in Randomized Clinical Trials of Hepatocellular Carcinoma. *Hepatology* 51, 1274–1283. doi:10.1002/hep.23485

Chang, Y.-J., Li, L.-T., Chen, H.-A., Hung, C.-S., and Wei, P.-L. (2014). Silencing Survivin Activates Autophagy as an Alternative Survival Pathway in HCC Cells. *Tumor Biol.* 35, 9957–9966. doi:10.1007/s13277-014-2257-6

Czaja, M. J., Ding, W.-X., Donohue, T. M., Jr., Friedman, S. L., Kim, J.-S., Komatsu, M., et al. (2013). Functions of Autophagy in Normal and Diseased Liver. *Autophagy* 9, 1131–1158. doi:10.4161/auto.25063

Fernandez, J. G., Rodriguez, D. A., Valenzuela, M., Calderon, C., Urzua, U., Munroe, D., et al. (2014). Survivin Expression Promotes VEGF-Induced Tumor Angiogenesis via PI3K/Akt Enhanced Beta-catenin/Tcf-Lef Dependent Transcription. *Mol. Cancer* 13, 209. doi:10.1186/1476-4598-13-209

Forner, A., Reig, M., and Bruix, J. (2018). Hepatocellular Carcinoma. *Lancet* 391, 1301–1314. doi:10.1016/S0140-6736(18)30010-2

Gugnoni, M., Sancisi, V., Manzotti, G., Gandolfi, G., and Ciarrocchi, A. (2016). Autophagy and Epithelial-Mesenchymal Transition: an Intricate Interplay in Cancer. *Cell Death Dis* 7, e2520. doi:10.1038/cddis.2016.415

Gukovskaya, A. S., Gukovsky, I., Algül, H., and Habtezion, A. (2017). Autophagy, Inflammation, and Immune Dysfunction in the Pathogenesis of Pancreatitis. *Gastroenterology* 153, 1212–1226. doi:10.1053/j.gastro.2017.08.071

Hu, Z. Y., Chen, B., Zhang, J. P., and Ma, Y. Y. (2017). Up-regulation of Autophagy-Related Gene 5 (ATG5) Protects Dopaminergic Neurons in a Zebrafish Model of Parkinson's Disease. *J. Biol. Chem.* 292, 18062–18074. doi:10.1074/jbc.M116.764795

Huang, X., Qian, X., Cheng, C., He, S., Sun, L., Ke, Q., et al. (2011). Expression of Pirh2, a p27Kip1 Ubiquitin Ligase, in Hepatocellular Carcinoma: Correlation with p27Kip1 and Cell Proliferation. *Hum. Pathol.* 42, 507–515. doi:10.1016/j.humpath.2010.04.021

Ito, Y., Matsuura, N., Sakon, M., Miyoshi, E., Noda, K., Takeda, T., et al. (1999). Expression and Prognostic Roles of the G1-S Modulators in Hepatocellular Carcinoma: P27 Independently Predicts the Recurrence. *Hepatology* 30, 90–99. doi:10.1002/hep.510300114

Janic, A., Valente, L. J., Wakefield, M. J., Di Stefano, L., Milla, L., Wilcox, S., et al. (2018). DNA Repair Processes Are Critical Mediators of P53-dependent Tumor Suppression. *Nat. Med.* 24, 947–953. doi:10.1038/s41591-018-0043-5

Kao, J.-T., Chuah, S.-K., Huang, C.-C., Chen, C.-L., Wang, C.-C., Hung, C.-H., et al. (2007). P21/WAF1 Is an Independent Survival Prognostic Factor for Patients with Hepatocellular Carcinoma after Resection. *Liver Int.* 27, 772–781. doi:10.1111/j.1478-3231.2007.01499.x

Kim, T. Y., Lee, J. W., Kim, H.-P., Jong, H.-S., Kim, T.-Y., Jung, M., et al. (2007). DLC-1, a GTPase-Activating Protein for Rho, Is Associated with Cell

Proliferation, Morphology, and Migration in Human Hepatocellular Carcinoma. *Biochem. Biophysical Res. Commun.* 355, 72–77. doi:10.1016/j.bbrc.2007.01.121

Lee, T. K., Man, K., Poon, R. T., Lo, C. M., Ng, I. O., and Fan, S. T. (2004). Disruption of P53-p21/WAF1 Cell Cycle Pathway Contributes to Progression and Worse Clinical Outcome of Hepatocellular Carcinoma. *Oncol. Rep.* 12, 25–31. doi:10.3892/or.12.1.25

Lin, P., He, R. Q., Dang, Y. W., Wen, D. Y., Ma, J., He, Y., et al. (2018). An Autophagy-Related Gene Expression Signature for Survival Prediction in Multiple Cohorts of Hepatocellular Carcinoma Patients. *Oncotarget. Apr* 3 (9), 17368–17395. doi:10.18632/oncotarget.24089

Lin, Z., Xu, S. H., Wang, H. Q., Cai, Y. J., Ying, L., Song, M., et al. (2016). Prognostic Value of DNA Repair Based Stratification of Hepatocellular Carcinoma. *Sci. Rep.* 6 (6), 25999. doi:10.1038/srep25999

Liu, F., Qin, L., Liao, Z., Song, J., Yuan, C., Liu, Y., et al. (2020). Microenvironment Characterization and Multi-Omics Signatures Related to Prognosis and Immunotherapy Response of Hepatocellular Carcinoma. *Exp. Hematol. Oncol.* 9, 10. doi:10.1186/s40164-020-00165-3

Liu, F., Wei, Y.-G., Luo, L.-M., Wang, W.-T., Yan, L.-N., Wen, T.-F., et al. (2013a). Genetic Variants of P21 and P27 and Hepatocellular Cancer Risk in a Chinese Han Population: a Case-Control Study. *Int. J. Cancer* 132, 2056–2064. doi:10.1002/ijc.27885

Liu, W., Zhu, F., Jiang, Y., Sun, D., Yang, B., and Yan, H. (2013b). siRNA Targeting Survivin Inhibits the Growth and Enhances the Chemosensitivity of Hepatocellular Carcinoma Cells. *Oncol. Rep. Mar.* 29, 1183–1188. doi:10.3892/or.2012.2196

Long, J., Zhang, L., Wan, X., Lin, J., Bai, Y., Xu, W., et al. (2018). A Four-Gene-Based Prognostic Model Predicts Overall Survival in Patients with Hepatocellular Carcinoma. *J. Cel Mol Med.* 22, 5928–5938. doi:10.1111/jcmm.13863

Matsuda, Y., Wakai, T., Hirose, Y., Osawa, M., Fujimaki, S., and Kubota, M. (2013). p27 Is a Critical Prognostic Biomarker in Non-alcoholic Steatohepatitis-Related Hepatocellular Carcinoma. *Int. J. Mol. Sci. Nov.* 29 (14), 23499–23515. doi:10.3390/ijms141223499

Mizushima, N. (2018). A Brief History of Autophagy from Cell Biology to Physiology and Disease. *Nat. Cel Biol.* 20, 521–527. doi:10.1038/s41556-018-0092-5

Moloudizargari, M., Asghari, M. H., Ghobadi, E., Fallah, M., Rasouli, S., and Abdollahi, M. (2017). Autophagy, its Mechanisms and Regulation: Implications in Neurodegenerative Diseases. *Ageing Res. Rev.* 40, 64–74. doi:10.1016/j.arr.2017.09.005

Okano, H., Shiraki, K., Inoue, H., Kawakita, T., Saitou, Y., Enokimura, N., et al. (2003). Fas Stimulation Activates NF-kappaB in SK-Hep1 Hepatocellular Carcinoma Cells. *Oncol. Rep.* 10, 1145–1148. doi:10.3892/or.10.5.1145

Qin, X.-Y., Suzuki, H., Honda, M., Okada, H., Kaneko, S., Inoue, I., et al. (2018). Prevention of Hepatocellular Carcinoma by Targeting MYCN-Positive Liver Cancer Stem Cells with Acyclic Retinoid. *Proc. Natl. Acad. Sci. USA* 115, 4969–4974. doi:10.1073/pnas.1802279115

Qiu, R., Wang, S., Feng, X., Chen, F., Yang, K., and He, S. (2011). Effect of Subcellular Localization of P21 on Proliferation and Apoptosis of HepG2 Cells. *J. Huazhong Univ. Sci. Technol. [Med. Sci.* 31, 756–761. doi:10.1007/s11596-011-0672-0

Singal, A. G., and El-Serag, H. B. (2015). Hepatocellular Carcinoma from Epidemiology to Prevention: Translating Knowledge into Practice. *Clin. Gastroenterol. Hepatol.* 13, 2140–2151. doi:10.1016/j.cgh.2015.08.014

Sun, B., Xu, H., Zhang, G., Zhu, Y., Sun, H., and Hou, G. (2013). Basic Fibroblast Growth Factor Upregulates Survivin Expression in Hepatocellular Carcinoma Cells via a Protein Kinase B-dependent Pathway. *Oncol. Rep. Jul* 30, 385–390. doi:10.3892/or.2013.2479

Wagayama, H., Shiraki, K., Sugimoto, K., Ito, T., Fujikawa, K., Yamanaka, T., et al. (2002). High Expression of p21WAF1/CIP1 Is Correlated with Human Hepatocellular Carcinoma in Patients with Hepatitis C Virus-Associated Chronic Liver Diseases. *Hum. Pathol.* 33, 429–434. doi:10.1053/hupa.2002.124724

Wang, S. S., Chen, Y. H., Chen, N., Wang, L. J., Chen, D. X., Weng, H. L., et al. (2017). Hydrogen Sulfide Promotes Autophagy of Hepatocellular Carcinoma Cells through the PI3K/Akt/mTOR Signaling Pathway. *Cel Death Dis* 8 (8), e2688. doi:10.1038/cddis.2017.18

Wang, Z., Zhu, J., Liu, Y., Liu, C., Wang, W., Chen, F., et al. (2020). Development and Validation of a Novel Immune-Related Prognostic Model in Hepatocellular Carcinoma. *J. Transl Med.* 18, 67. doi:10.1186/s12967-020-02255-6

White, E. (2015). The Role for Autophagy in Cancer. *J. Clin. Invest.* 125, 42–46. doi:10.1172/jci73941

Williams, A. B., and Schumacher, B. (2016). p53 in the DNA-Damage-Repair Process. *Cold Spring Harb Perspect. Med.* 6, 6. doi:10.1101/cshperspect.a026070

Wu, H.-T., Xie, C.-R., Lv, J., Qi, H.-Q., Wang, F., Zhang, S., et al. (2018). The Tumor Suppressor DLC1 Inhibits Cancer Progression and Oncogenic Autophagy in Hepatocellular Carcinoma. *Lab. Invest.* 98, 1014–1024. doi:10.1038/s41374-018-0062-3

Wu, Z.-m., Dai, C., Huang, Y., Zheng, C.-f., Dong, Q.-z., Wang, G., et al. (2011). Anti-cancer Effects of p21WAF1/CIP1 Transcriptional Activation Induced by dsRNAs in Human Hepatocellular Carcinoma Cell Lines. *Acta Pharmacol. Sin* 32, 939–946. doi:10.1038/aps.2011.28

Xu, K., Zhang, Z., Qian, J., Wang, S., Yin, S., Xie, H., et al. (2019). LncRNA FOXD2-AS1 Plays an Oncogenic Role in Hepatocellular Carcinoma through Epigenetically Silencing CDKN1B(p27) via EZH2. *Exp. Cel Res* 380, 198–204. doi:10.1016/j.yexcr.2019.04.016

Yang, R., Zhang, Y., Wang, L., Hu, J., Wen, J., Xue, L., et al. (2017). Correction: Increased Autophagy in Fibroblast-like Synoviocytes Leads to Immune Enhancement Potential in Rheumatoid Arthritis. *Oncotarget. Aug* 22 (8), 57906. doi:10.18632/oncotarget.20371

Yin, L., Chang, C., and Xu, C. (2017). G2/M Checkpoint Plays a Vital Role at the Early Stage of HCC by Analysis of Key Pathways and Genes. *Oncotarget. Sep.* 29 (8), 76305–76317. doi:10.18632/oncotarget.19351

Zhang, W., Lu, Z., Kong, G., Gao, Y., Wang, T., Wang, Q., et al. (2014). Hepatitis B Virus X Protein Accelerates Hepatocarcinogenesis with Partner Survivin through Modulating miR-520b and HBXIP. *Mol. Cancer* 13, 128. doi:10.1186/1476-4598-13-128

Zhou, Q., Lui, V. W., and Yeo, W. (2011). Targeting the PI3K/Akt/mTOR Pathway in Hepatocellular Carcinoma. *Future Oncol.* 7, 1149–1167. doi:10.2217/fon.11.95

Zhou, X., Thorgeirsson, S. S., and Popescu, N. C. (2004). Restoration of DLC-1 Gene Expression Induces Apoptosis and Inhibits Both Cell Growth and Tumorigenicity in Human Hepatocellular Carcinoma Cells. *Oncogene* 23, 1308–1313. doi:10.1038/sj.onc.1207246

Zou, C., Chen, J., Chen, K., Wang, S., Cao, Y., Zhang, J., et al. (2015). Functional Analysis of miR-181a and Fas Involved in Hepatitis B Virus-Related Hepatocellular Carcinoma Pathogenesis. *Exp. Cel Res* 331, 352–361. doi:10.1016/j.yexcr.2014.11.007

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 28 May 2021
doi: 10.3389/fmolb.2021.643504

Gene Expression Profiles of Circular RNAs and MicroRNAs in Chronic Rhinosinusitis With Nasal Polyps

*Jieqing Yu[1,2†], Xue Kang[1,3†], Yuanping Xiong[1], Qing Luo[1], Daofeng Dai[1] and Jing Ye[1,2]**

[1]Department of Otorhinolaryngology Head and Neck Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, China, [2]Jiangxi Otorhinolaryngology Head and Neck Surgery Institute, Nanchang, China, [3]Department of Otorhinolaryngology Head and Neck Surgery, Jiangxi Provincial Children's Hospital, Nanchang, China

Introduction: Chronic rhinosinusitis (CRS) is often classified primarily on the basis of the absence or presence of nasal polyps (NPs), that is, as CRS with nasal polyps (CRSwNP) or CRS without nasal polyps (CRSsNP). Additionally, according to the percentage of eosinophils, CRSwNP can be further divided into eosinophilic CRSwNP (ECRSwNP) and non-ECRSwNP. CRSwNP is a significant public health problem with a considerable socioeconomic burden. Previous research reported that the pathophysiology of CRSwNP is a complex, multifactorial disease. There have been many studies on its etiology, but its pathogenesis remains unclear. Dysregulated expression of microRNAs (miRNAs) has been shown in psoriasis, rheumatoid arthritis, pulmonary fibrosis, and allergic asthma. Circular RNAs (circRNAs) are also involved in inflammatory diseases such as rheumatoid arthritis, septic acute kidney injury, myocardial ischemia/reperfusion injury, and sepsis-induced liver damage. The function of miRNAs in various diseases, including CRSwNP, is a research hotspot. In contrast, there have been no studies on circRNAs in CRSwNP. Overall, little is known about the functions of circRNAs and miRNAs in CRSwNP. This study aimed to investigate the expression of circRNAs and miRNAs in a CRSwNP group and a control group to determine whether these molecules are related to the occurrence and development of CRSwNP.

Methods: Nine nasal mucosa samples were collected, namely, three ECRSwNP samples, three non-ECRSwNP samples, and three control samples, for genomic microarray analysis of circRNA and microRNA expression. All of the tissue samples were from patients who were undergoing functional endoscopic sinus surgery in our department. Then we selected some differentially expressed miRNAs and circRNAs for qPCR verification. Meanwhile, GO enrichment analysis and KEGG pathway analysis were applied to predict the biological functions of aberrantly expressed circRNAs and miRNAs based on the GO and KEGG databases. Receiver operating characteristic (ROC) curve analysis and principal component analysis (PCA) were performed to confirm these molecules are involved in the occurrence and development of CRSwNP.

Results: In total, 2,875 circRNAs showed significant differential expression in the CRSwNP group. Specifically, 1794 circRNAs were downregulated and 1,081 circRNAs were upregulated. In the CRSwNP group, the expression of 192 miRNAs was significantly

OPEN ACCESS

Edited by:
Jie Li,
Harbin Institute of Technology, China

Reviewed by:
Sayan Chatterjee,
Guru Gobind Singh Indraprastha University, India
Frederico Marianetti Soriani,
Federal University of Minas Gerais, Brazil

***Correspondence:**
Jing Ye
yjholly@email.ncu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *18 December 2020*
Accepted: *26 April 2021*
Published: *28 May 2021*

Citation:
Yu J, Kang X, Xiong Y, Luo Q, Dai D and Ye J (2021) Gene Expression Profiles of Circular RNAs and MicroRNAs in Chronic Rhinosinusitis With Nasal Polyps. Front. Mol. Biosci. 8:643504. doi: 10.3389/fmolb.2021.643504

downregulated, and none of the miRNAs were significantly upregulated. GO and KEGG analysis showed differential circRNAs and miRNAs were enriched in "amoebiasis," "salivary secretion," "pathways in cancer," and "endocytosis." Through qRT-PCR verification, the expression profiles of hsa-circ-0031593, hsa-circ-0031594, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p were shown to have statistical differences. In addition, ROC curve analysis showed that the molecules with the two highest AUCs were hsa-circ-0031593 with AUC 0.8353 and hsa-miR-145-5p with AUC 0.8690. Through PCA with the six ncRNAs, the first principal component explained variance ratio was 98.87%. The AUC of the six ncRNAs was 0.8657.

Conclusion: In our study, the expression profiles of ECRSwNP and non-ECRSwNP had no statistical differences. The differentially expressed circRNAs and miRNAs between CRSwNP and control may play important roles in the pathogenesis of CRSwNP. Altered expression of hsa-circ-0031593 and hsa-miR-145-5p have the strongest evidence for involvement in the occurrence and development of CRSwNP because their AUCs are higher than the other molecules tested in this study.

Keywords: chronic rhinosinusitis with nasal polyps, circular RNA, micro RNA, microarray analysis, gene express profile

INTRODUCTION

Chronic rhinosinusitis with nasal polyps (CRSwNP) is a significant public health problem with a considerable socioeconomic burden. Chronic rhinosinusitis (CRS), which is characterized by persistent mucosa inflammation of the sinuses, is one of the most common chronic diseases, and its pathophysiology remains unclear (AI-Sayed et al., 2017).

According to whether nasal polyps exist or not, CRS is often classified as CRSwNP or CRS without nasal polyps (CRSsNP) (Cho et al., 2017). CRSwNP remains a challenging clinical problem due to its propensity for recurrence. According to the percentage of eosinophils, CRSwNP can be classified as eosinophilic CRSwNP (ECRSwNP), with an eosinophil count ≥10%, and non-eosinophilic CRSwNP (non-ECRSwNP), with an eosinophil count <10%. There are differences between the two subtypes of CRSwNP (Cao et al., 2009). Compared with non-ECRSwNP, ECRSwNP is characterized by more eosinophils infiltrating the nasal mucosa, and it has a worse prognosis and higher recurrence rate (Shi et al., 2013). ECRSwNP and non-ECRSwNP have different clinical symptoms, recurrence rates, and responses to drugs and endoscopic surgery (Lou et al., 2015). ECRSwNP is a hard-to-treat subtype of CRS.

To discover the pathogenesis of and better treatment for CRSwNP, more research is needed to further explore the different molecular and cytological mechanisms of the subtypes that lead to the different clinical and pathophysiological characteristics between the two subtypes.

With advancements in genomic microarray technology, a revolutionary change has taken place in the field of genetic analysis, that makes it possible to quantify thousands of gene expressions simultaneously. Studies have shown that the human genome can be widely transcribed into a large amount of non-coding RNAs (ncRNAs) that are closely related to the initiation as well as progression of diseases (Beermann et al., 2016). Genomic microarray technology has been widely used in the field of biomedicine to explore the occurrence and development of human diseases, including CRSwNP, at the genetic level (Plager et al., 2010; Yao et al., 2019).

CircRNAs (circular RNAs) are a type of ncRNA with important functions that have tissue specificity and disease specificity (Xia et al., 2017). Unlike linear RNAs (containing 5′ and 3′ ends), circRNAs are closed continuous loops that are free from exonuclease-mediated degradation and are more stable than most linear RNAs (Jeck et al., 2013). It has been found that circRNAs, acting as miRNA sponges, are rich in miRNA binding sites and increase the expression of target genes by mitigating the inhibition of miRNAs on their target genes (Kulcheski et al., 2016). CircRNAs represent a class of naturally occurring endogenous ncRNAs that have recently been recognized as important regulators of gene expression networks (Oude Voshaar et al., 2019). In recent years, researchers have explored the expression profiles of circRNAs in different diseases. For example, one study showed that oxidized low-density lipoprotein accelerates the injury of endothelial cells via the circ-USP36/miR-98-5p/VCAM1 axis (Peng et al., 2021). Another study found that circRNA_09505 aggravates inflammation and joint damage in RA via the miR-6089/AKT1/NF-κB axis (Yang et al., 2020).

MiRNAs (microRNAs) are another group of ncRNAs that are involved in many pathologic and physiological processes, such as proliferation, differentiation, and tumorigenesis (Zhang X.-H. et al., 2012; Ferreira et al., 2018; Martínez-Rivera et al., 2018). Research has shown that the expression of miR-125 b is increased in ECRSwNP, which may lead to mucosal eosinophilia (Zhang Y.-N. et al., 2012). In addition, miR-1 can regulate the transport of eosinophils in CRS. Overexpression of miR-1 inhibits the increase of airway eosinophils and inhibits the binding of eosinophils and

endothelial cells induced by IL-13 (Korde et al., 2020). The interaction between circRNAs and miRNAs plays an important role in inflammation and immune responses. In psoriasis, circRNA-0061012 enhances GAB1 expression through spongy miR-194-5p, thereby promoting the proliferation, migration, and invasion of keratinocytes induced by IL-22 (He et al., 2021). CircRNA-WBSCR17 aggravates the inflammatory response of human renal tubular epithelial cells induced by high glucose by targeting the miR-185-5p/SOX6 axis (Li et al., 2020). In osteoarthritis, circRNA-9119 blocks the miR-26a/PTEN axis to protect IL-1β-treated chondrocytes from apoptosis (Chen et al., 2020). These reports show that circRNAs and miRNAs are related to inflammation.

To date, dysregulated expression of miRNAs has been shown in psoriasis, rheumatoid arthritis, pulmonary fibrosis, and allergic asthma. CircRNA is also involved in inflammatory diseases, but there is no research of circRNA in CRSwNP (Zhang Y.-N. et al., 2012). In our study, we aimed to compare the microarray expression profiles of miRNAs and circRNAs in nasal polyps of CRSwNP and normal nasal mucosa from control subjects. However, the results of RNA-seq genomic microarray analysis of ECRSwNP and non-ECRSwNP had no statistical differences, so, we combined ECRSwNP and non-ECRSwNP into a single group denoted as CRSwNP. Then, we validated the abnormally expressed circRNAs and miRNAs by qRT-PCR (quantitative real time polymerase chain reaction). In particular, we explored the potentially biological functions and involved signaling pathways of these ncRNAs by using the GO and KEGG databases. We concluded that the differentially expressed circRNAs and miRNAs may play important roles in the pathogenesis of CRSwNP. Based on ROC (receiver operating characteristic) curve analysis and principal component analysis (PCA), the altered expressions of hsa-circ-0031593, and hsa-miR-145-5p have the most evidence supporting their involvement in the occurrence and development of CRSwNP.

MATERIALS AND METHODS

Subjects and Samples

Nasal polyp specimens were collected from CRSwNP patients undergoing functional endoscopic sinus surgery. The middle turbinate mucosae of the control group were obtained from patients undergoing optic nerve decompression and nasal bone fracture surgery. Controls with nasal inflammation or upper respiratory tract infection were excluded. Subjects using corticosteroids or other immune-modulating drugs within 1 month, and subjects with antrochoanal polyps or fungal sinusitis were excluded. All of the participants were enrolled from the Department of Otorhinolaryngology Head and Neck Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, China in 2017 (detailed information is shown in **Supplementary Table S1**). The study was approved by the Medical Research Ethics Committee of The First Affiliated Hospital of Nanchang University (2017080).

Study Process

The quantity of eosinophils in each specimen was observed by hematoxylin-eosin staining in three random microscopic high power fields (HPFs, ×400 magnification). CRSwNP patients were classified according to the percentage of eosinophils in nasal polys. Nasal polyps from three ECRSwNP, three non-ECRSwNP, and three control individuals were collected for total RNA extraction and microarray analysis. We found that ECRSwNP and non-ECRSwNP tissues shared similar expression patterns of circRNAs and miRNAs. Therefore, we combined ECRSwNP and non-ECRSwNP into a single group denoted as CRSwNP. Then, the aberrant circRNAs and miRNAs with fold change > ± 2.5 and $p < 0.05$ were validated in an independent cohort (control group, $n = 5$; CRSwNP, $n = 5$) by qRT-PCR. Next, we expanded the sample size (control group, $n = 25$; CRSwNP group, $n = 29$) to conduct further research on the quantitative expression of the selected ncRNAs in the third cohort.

Hematoxylin-Eosin Staining for Eosinophils

The quantity of eosinophils was analyzed by hematoxylin-eosin staining. Specimens of nasal polyps were fixed in 10% formaldehyde and placed in low to high concentration alcohol to remove water from the tissues. Then the specimens were embedded in paraffin wax, sliced by a microtome into sections no more than 0.5 μm thick, and deparaffinized to yield tissue sections. After rehydration, the tissue sections were stained with hematoxylin for 10 min and eosin for 3 min. In order to classify CRSwNP into ECRSwNP and non-ECRSwNP, we calculated the percentage of eosinophils in all of the inflammatory cells through five random high-power fields. Specimens with eosinophils ≥10% were defined as ECRSwNP, and those with eosinophils <10% as non-ECRSwNP (AI-Sayed et al., 2017).

CircRNA Microarray Analysis

The total RNAs were extracted from the subjects for microarray analysis. The purity and concentration of RNA were determined by the OD260/280 readings of a spectrophotometer (NanoDrop ND-1000). The integrity of RNAs was detected by standard denaturing agarose gel electrophoresis (Bioanalyzer 2100, Agilent Technologies, United States). The results are shown in **Supplementary Figure S1**. The digestion, amplification, and labeling of RNAs were performed based on the protocol provided by the manufacturer. The labeled RNAs were hybridized onto the microarray (Agilent-084217) after purification. The circRNA array data were analyzed by GeneSpring software V13.0 (Agilent). In order to select the differential expression of circRNAs, we used the threshold ≥ ±2.5 fold change and $p < 0.05$.

MiRNA Microarray Analysis

MiRNA expression profile microarrays of these specimens were performed by CapitalBio. Procedures are described in detail on the CapitalBio website (http://www.capitalbio.com). Briefly, the procedure included total RNA extraction, quality control, miRNA isolation, FlashTag biotin labeling of miRNAs, hybridization to an Affymetrix GeneChip microarray (Affymetrix miRNA 4.0), and microarray washing, staining,

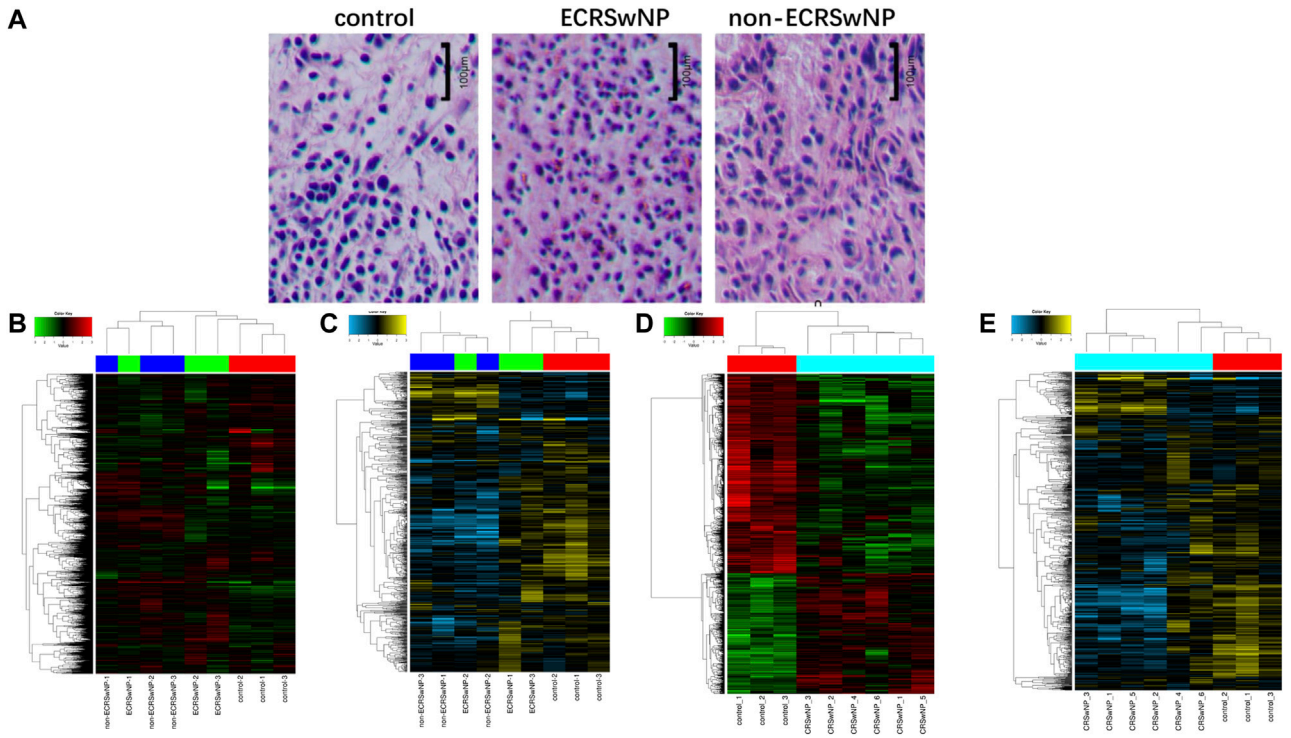


FIGURE 1 | Hematoxylin-eosin staining of control, ECRSwNP, and non-ECRSwNP groups, and differential expression of circRNAs and miRNAs in the three groups. **(A)** Representative images of hematoxylin-eosin staining. The scale bar represents 100 μm. **(B)** Hierarchical clustering analysis heat map showing significantly changed of circRNAs and **(C)** of miRNAs with fold change ≥2.0 ($p < 0.05$) in the three groups (ECRSwNP, non-ECRSwNP, and control). The expressions of circRNAs and miRNAs were significantly different between CRSwNP and control groups, but ECRSwNP and non-ECRSwNP groups shared similar gene expression profiles. Then, they were combined as the CRSwNP group. Hierarchical clustering analysis heat map showing significantly changed of circRNAs **(D)** and of miRNAs **(E)** with fold change ≥2.0 ($p < 0.05$) in the two groups (CRSwNP and control).

and scanning. If the miRNAs expression changed by at least ±2.5-fold ($p < 0.05$), this was considered a significant difference.

Correlation and Co-Expression Analysis

CircRNAs bind with miRNAs competitively, which inhibits the negative regulation of miRNAs on target genes and leads to the increase of the functional activity and expression of target genes (Fokkens et al., 2012). We constructed co-expression networks to predict the target genes of circRNAs and miRNAs of CRSwNP. The co-expression analysis was based on miRanda-3.3 software, with entropy values below 20. The top 40 circRNA-miRNA networks, $p < 0.05$, were selected for analysis. In the network analysis, each point represents a gene, and two points connected by a line represent two closely related genes.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Analyses

We identified functional categories that were significantly enriched relative to the background reference by GO enrichment analysis. The related pathways and gene interactions associated with the abnormal expression of circRNA and microRNAs were found based on the latest KEGG pathway enrichment database. The significant GO terms and pathways were determined by Fisher's exact test, and the false discovery rate was utilized to correct the p-values.

Quantitative Real-Time Polymerase Chain Reaction for Validation of circRNAs and miRNAs

The significant differential expression of circRNAs and miRNAs was quantified by qRT-PCR. Total RNA was isolated from nasal polyps with TRIzol reagent. The cDNAs were synthesized by reverse transcription with a PrimeScript RT reagent kit with random primers. Then, qRT-PCR was conducted by SYBR Premix Ex Taq II (Tli RNaseH Plus; TaKaRa). Primers for selected ncRNAs and house-keeping genes were synthesized by Sangon Biotech (Shanghai, China). The primers used are shown in **Supplementary Tables S2, S3**.

Statistical Analysis

SPSS 22.0 was used in this study. The Mann-Whitney U-test was used to calculate the differences of the expression of circRNAs and miRNAs between groups, and $p < 0.05$ was considered to be statistically significant. The functional values of the selected circRNAs and miRNAs for CRSwNP were evaluated by conducting ROC curve analysis and PCA.

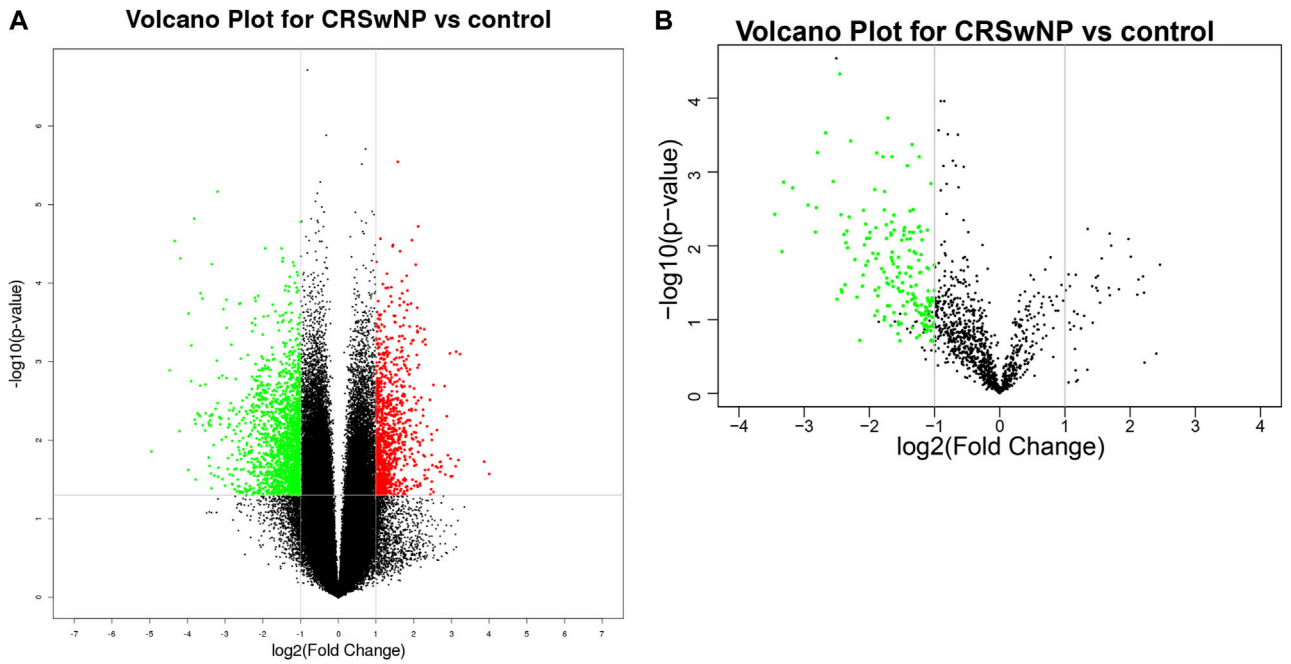


FIGURE 2 | Different expressions of circRNAs and miRNAs shown in volcano plots. The circRNAs and miRNAs were widely distributed in all chromosomes. Each column corresponds to a circRNA or a miRNA, the green column toward the center represents significantly downregulated and the red one toward the outside means upregulated. **(A)** Differential expression of circRNAs. **(B)** Differential expression of miRNAs.

RESULTS

ECRSwNP and Non-ECRSwNP Patients Shared Similar Gene Expression Profiles

In this study, we used hematoxylin-eosin staining for eosinophil counts (**Figure 1A**). Hierarchical clustering analysis was used for evaluating gene expression differences among groups. As shown in **Figures 1B,C**, the ECRSwNP and non-ECRSwNP groups shared similar gene expression profiles, so we could not distinguish ECRSwNP from non-ECRSwNP by hierarchical clustering analysis. Then, we combined the ECRSwNP group and non-ECRSwNP group into the CRSwNP group. The results showed that the expression profiles of circRNAs (**Figure 1D**) and miRNAs (**Figure 1E**) were significantly different between the CRSwNP and control groups.

Differential Expression of circRNAs and miRNAs in CRSwNP

Volcano plots were used to assess the locations of circRNAs (**Figure 2A**) and miRNAs (**Figure 2B**). These ncRNAs were widely distributed in all of the chromosomes. Circos plots and scatter plots were also applied to analyze the gene expression differences between the CRSwNP and control groups (**Supplementary Figures S2, S3**). CircRNAs and miRNAs downregulated or upregulated with fold change $\geq \pm 2.5$ ($p < 0.05$) in both the ECRSwNP group and non-ECRSwNP group were considered to have significant differential expression and were selected for further research. A total of 2,875 circRNAs showed significant differential expression in the CRSwNP group, including 1,794 downregulated circRNAs and 1,081 upregulated circRNAs. Additionally, 192 miRNAs were significantly downregulated and no miRNAs were significantly upregulated in the CRSwNP group.

Co-Expression Network in CRSwNP

From gene co-expression network analysis, we found that there was a difference in the co-expression of circRNAs and miRNAs between the CRSwNP group and control group, which revealed the underlying molecular mechanism of the pathogenesis of CRSwNP. We selected 40 circRNAs differentially expressed between the CRSwNP and control groups. The top 40 circRNAs-miRNAs networks are shown in **Figure 3**. Hsa-circ-0031593, hsa-circ-0031594, and hsa-miR-27b-3p are present in **Figure 3**. One circRNA can be associated with multiple miRNAs, and one miRNA can be related to multiple circRNAs, resulting in complex functional connections. In **Figure 3**, the darker and larger nodes indicate the higher fold change of circRNAs, purple indicates upregulation, and blue indicates downregulation. There were many regulatory relationships between the circRNAs and miRNAs in the networks, which further indicated that the networks of regulatory relationships were ubiquitous.

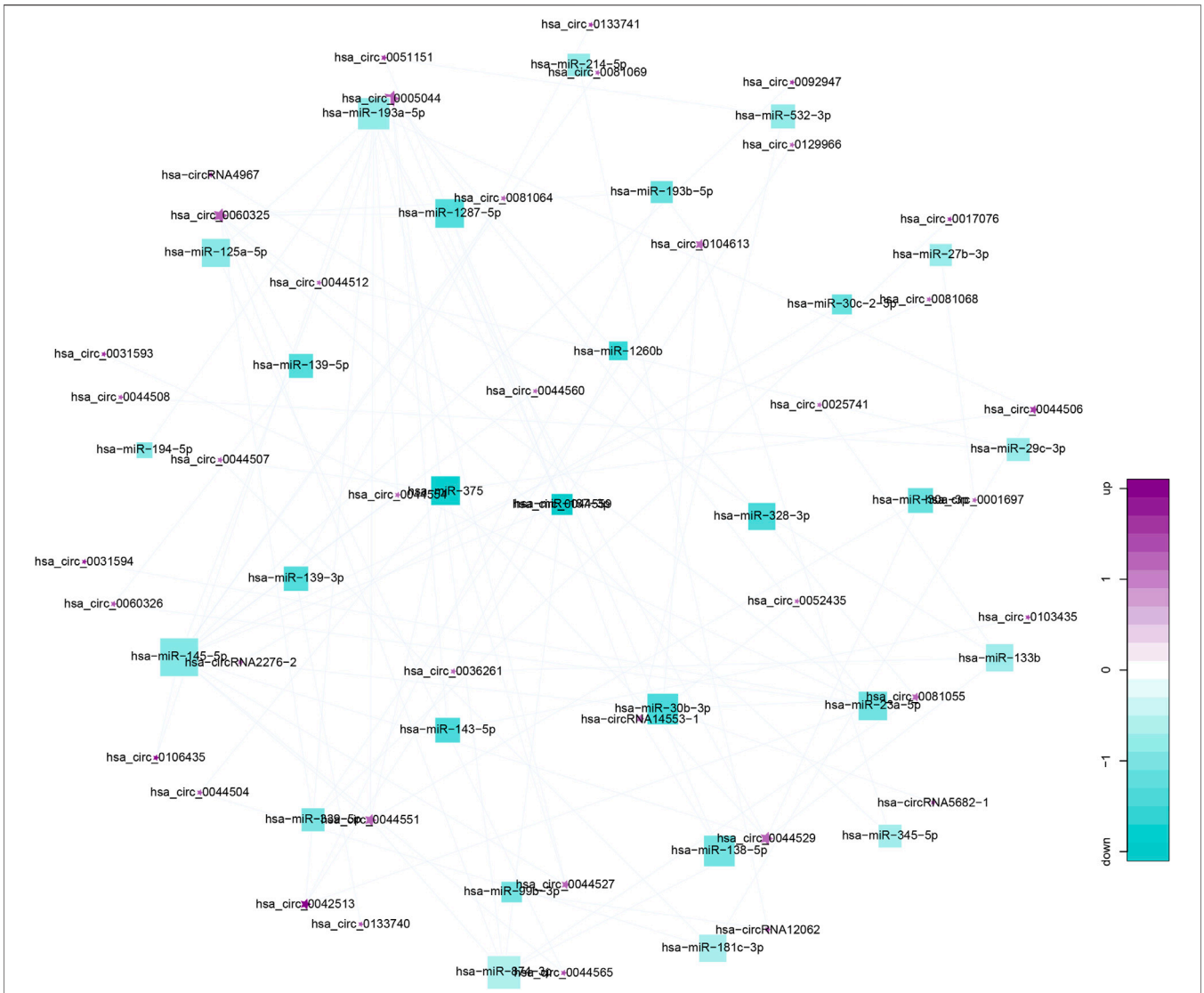


FIGURE 3 | Top 40 circRNA-miRNA co-expression networks. Co-expression networks were analyzed by miRanda-3.3 software, combined with entropy values below 20. Stars represent circRNAs; squares represent miRNA; lines represent correlative relationships. Different sizes and colors represent the corresponding (up/down) relationship.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Analyses

We predicted the biological functions of aberrantly expressed circRNAs and miRNAs through GO enrichment analysis and KEGG pathway analysis. The enriched GO terms ($p < 0.05$) could reflect the function of aberrantly expressed circRNAs and miRNAs. Each plot represents a specific biological function, and a larger size means a larger gene number. For circRNAs, compared with the normal control subjects, the significantly over-presented terms were mainly involved in cell adhesion, membrane, and receptor activity. In contrast, miRNAs were mainly involved in cellular component organization, intracellular parts, protein binding, and so on. The details of the top 10 significantly enriched GO terms of circRNAs are presented in **Figure 4**. According to GO analysis of circRNAs in the biological process (BP) category, the differentially expressed circRNAs were chiefly enriched in "single-organism process" and "single-organism cell process" (**Figure 4**). GO analysis related to cellular components (CC) showed that they were mainly enriched in "membrane" and "membrane part" (**Figure 4**). GO analysis related to molecular functions (MF) demonstrated that they were involved in "molecular transducer activity" and "receptor activity" (**Figure 4**). The GO analysis of miRNAs is shown in **Figure 5**. The biological process category was enriched in "positive regulation of biological process" and "anatomical structure development" (**Figure 5**). Cellular components (CC) were mostly enriched in "intracellular" and "intracellular part" (**Figure 5**). Molecular functions (MF) were involved in "binding" and "protein binding" (**Figure 5**). In

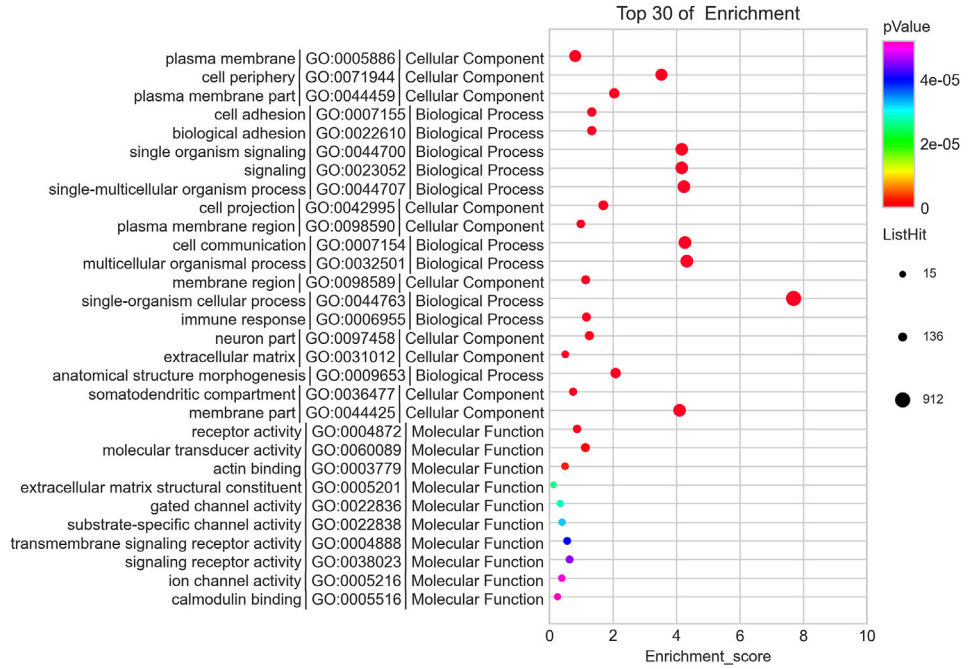


FIGURE 4 | Top 10 significant enriched GO terms analysis of circRNAs. Differentially expressed circRNAs were mainly involved in biological process, cellular component, and molecular function. Each plot represents a specific biological function, and the larger the size, the stronger the circRNAs expression. Compared with the normal subjects, the significantly over-presented circRNAs in the CRSwNP group were mainly involved in cell adhesion, membrane, and receptor activity.

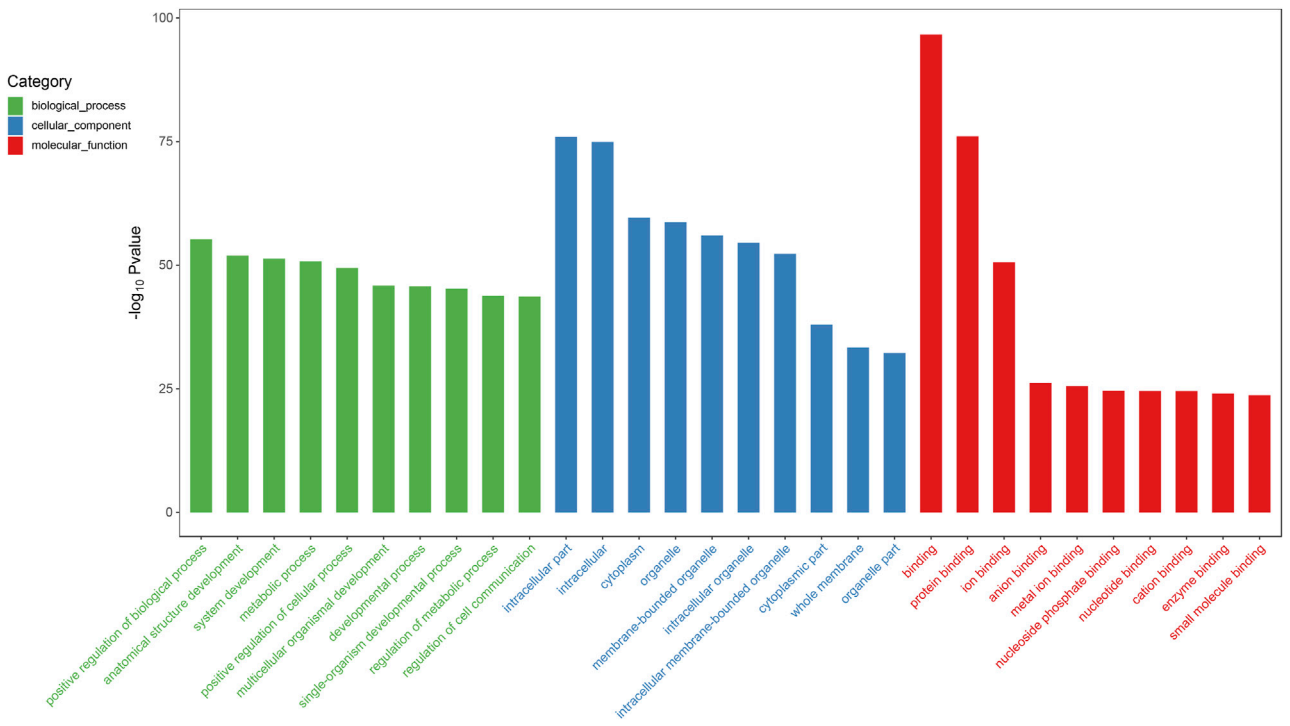


FIGURE 5 | Top 10 significant enriched GO terms analysis of miRNAs. Differentially expressed miRNAs were mainly involved in biological process, cellular component, and molecular function. Compared with the normal subjects, the significantly over-presented miRNAs in the CRSwNP group were mainly involved in positive regulation of biological process, intracellular part, and binding.

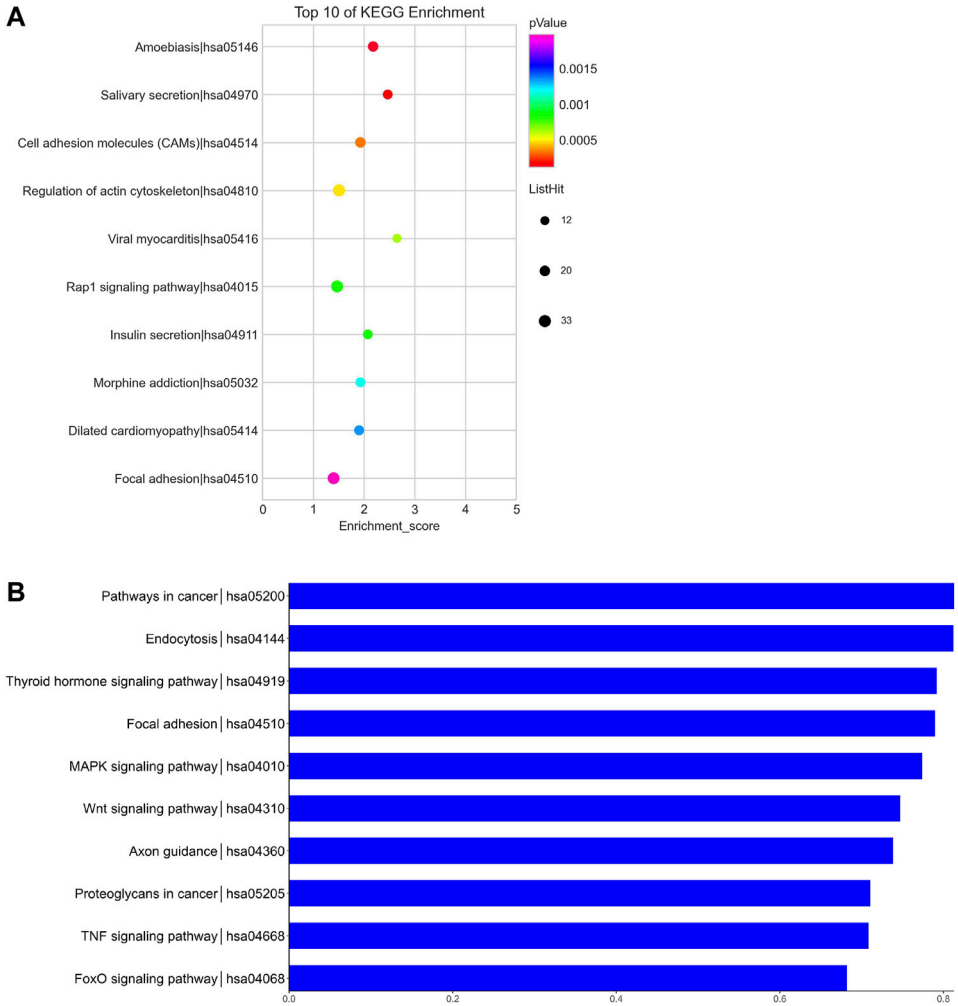


FIGURE 6 | KEGG pathway analysis of circRNAs and miRNAs. Biological roles of the differentially expressed circRNAs **(A)** and miRNAs **(B)**. Biological roles of the differentially expressed circRNAs include amebiasis, salivary secretion, cell adhesion molecules (CAMs), and so on. Biological roles of the differentially expressed miRNAs include pathways in cancer, endocytosis, thyroid hormone signaling pathway, and so on.

addition, from KEGG pathway analysis, we found that differentially expressed circRNAs were annotated to amebiasis, salivary secretion, and others (**Figure 6A**); and miRNAs were annotated to pathways in cancer, endocytosis, and so on (**Figure 6B**).

Confirmation of the Expression of circRNAs and miRNAs Using RNA-Sequencing and Quantitative Real-Time Polymerase Chain Reaction

The results of RNA-seq analysis showed that hsa-circ-0031593 and hsa-circ-0031594 were expressed to a significantly higher degree in the CRSwNP group than in the control group. Hsa-circ-0109623, hsa-circ-0000736, hsa-miR-125a-5p, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p were to a significantly lower degree in the CRSwNP group than in the control group (**Figure 7A**). To confirm the reliability of the microarray results, circRNAs and miRNAs downregulated and upregulated with fold change ≥ ± 2.5 ($p < 0.05$) in five CRSwNP subjects and five control subjects were selected for qRT-PCR in order to analyze the expression levels of these ncRNAs. The results showed that hsa-circ-0031593 and hsa-circ-0031594 were expressed significantly more in the CRSwNP group than in the control group. Hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p were expressed significantly less in the CRSwNP group than in the control group. There were no statistical differences among the expressions of hsa-circ-0109623, hsa-circ-0000736, or hsa-miR-125a-5p (**Figure 7B**).

After that, more samples (29 CRSwNP subjects and 25 control subjects) were used for further research of the quantitative expressions of these circRNAs and miRNAs. The detailed expression levels of these ncRNAs are shown in **Figure 8**. The results in **Figure 8** are in accordance with the results in **Figure 7B**. The expressions of hsa-circ-0031593 and hsa-circ-0031594 in the CRSwNP group were significantly higher than those in the

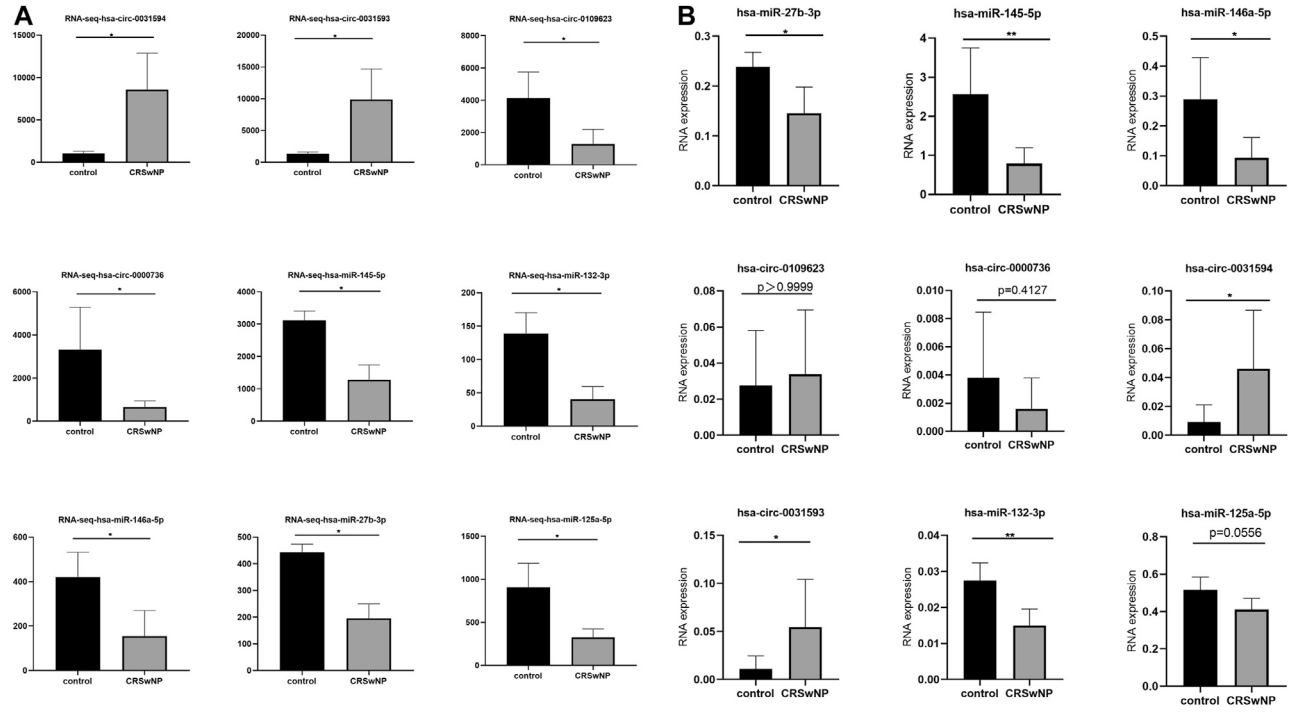


FIGURE 7 | Confirmation of the expression of circRNAs and miRNAs using RNA-sequencing and quantitative real-time polymerase chain reaction (qRT-PCR). CircRNAs and miRNAs were validated by RNA-sequencing (CRSwNP(n = 6); control subjects (n = 3) **(A)**; and qRT-PCR in an independent cohort of five CRSwNP subjects and five control subjects **(B)**. Each sample was detected in triplicate. GAPDH was used as the reference gene for circRNAs, and hsa-miR-16 was used as the reference gene for miRNAs. The relative expression levels of hsa-circ-0031593, hsa-circ-0031594, hsa-circ-0109623, hsa-circ-0000736, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, hsa-miR-125a-5p, and hsa-miR-27b-3p by qRT-PCR. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

control group. The expressions of hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p in the CRSwNP group were significantly lower than those in the control group. There were no statistical differences among the expressions of hsa-circ-0109623, hsa-circ-0000736, or hsa-miR-125a-5p.

ROC Curve Analysis and PCA of Selected circRNAs and miRNAs

ROC curve analysis was carried out to estimate the functional values of the selected circRNAs and miRNAs in the occurrence and development of CRSwNP (**Figure 9**). The sensitivity and specificity for the values of CRSwNP are shown in **Table 1**. The PCA method was used in this research. The data of PCA with combinations of hsa-circ-0031593, hsa-circ-0031594, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p are displayed in **Table 2**. The first principal component explained variance ratio was 98.87%. We used the first principal component 1 of these six ncRNAs to carry out ROC curve analysis, and the AUC was 0.8657, indicating a good significance for the pathogenesis of CRSwNP. The AUCs of hsa-circ-0031593 and hsa-miR-145-5p were 0.8353 [(0.7291–0.9415), p < 0.0001] and 0.8690 [(0.76–0.978), p < 0.0001]. Hsa-circ-0031593 and hsa-miR-145-5p had the strongest evidence supporting their involvement in the occurrence and development of CRSwNP since they had higher AUCs than others and had p values < 0.05.

DISCUSSION

CRSwNP is a significant public health problem with a considerable socioeconomic burden. Previous studies have reported that CRSwNP is a complex, multifactorial disease. There have been many studies on its etiology, but its pathogenesis remains unclear. Dysregulated expression of miRNAs has been shown in psoriasis, rheumatoid arthritis, pulmonary fibrosis, and allergic asthma. CircRNA is also involved in inflammatory diseases, but there has been no research on the role of circRNA in CRSwNP. Although numerous researchers have attempted to clarify the pathogenesis of CRSwNP, the detailed mechanisms remain unclear. Overall, little is known on the role of ncRNAs in the pathogenesis of CRSwNP. Further understanding of the genetic level of pathogenesis is essential for developing new techniques for effective prevention and therapy to improve prognosis. Researchers have found that ncRNAs, such as circRNAs and miRNAs, play essential roles in the occurrence and development of many diseases, which is contrary to the traditional view that genes are mainly regulated by protein coding (Wu et al., 2019; Yang et al., 2019). To explore the functions of circRNAs and miRNAs in CRSwNP, we performed gene microarray analysis of circRNAs and miRNAs in a CRSwNP group and a control group. Functional enrichment analysis and prediction of differentially expressed genes were carried out by using public databases. In addition, we performed qRT-PCR to validate the reliability of RNA-seq analysis, and we confirmed that hsa-circ-0031593 and

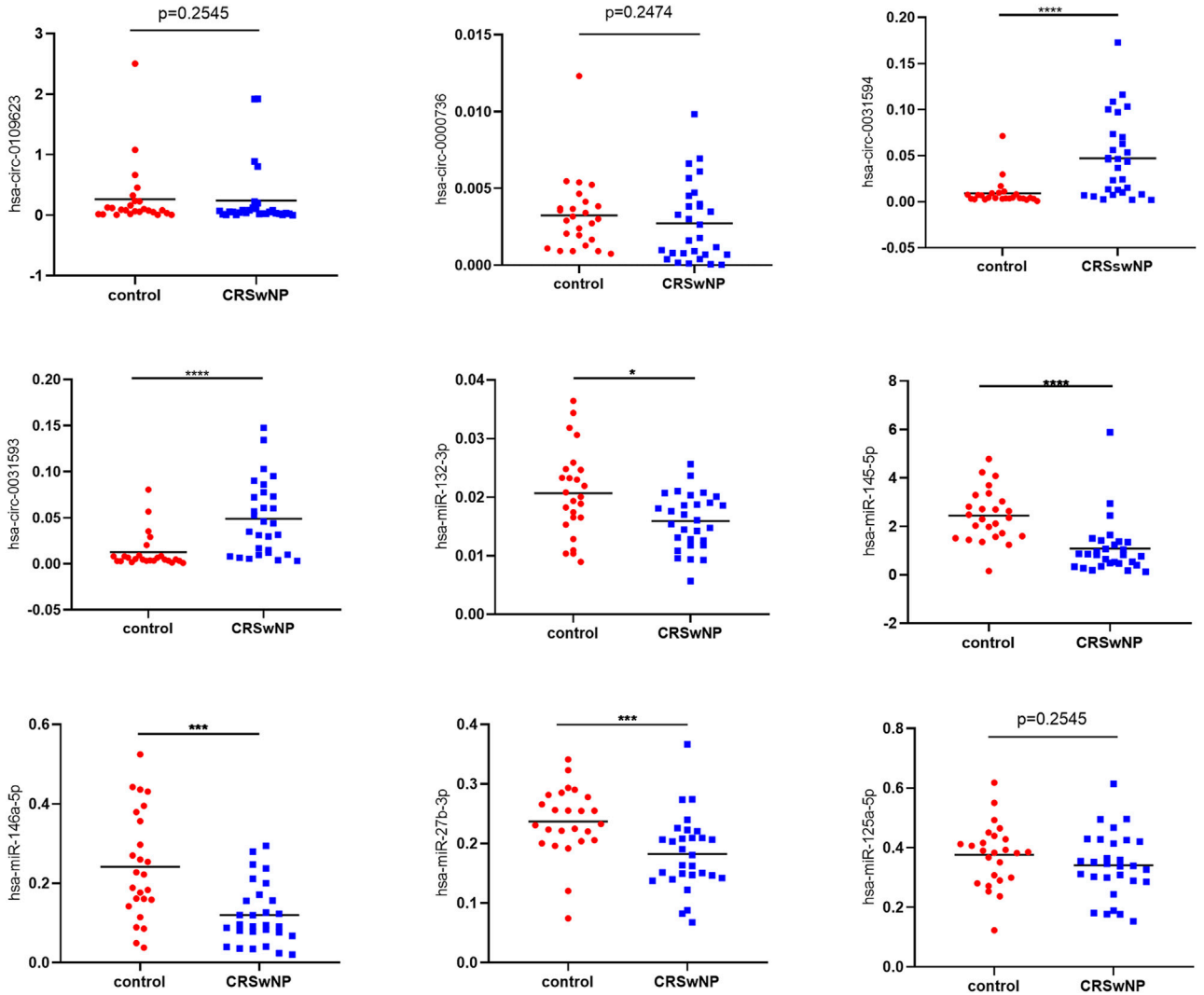


FIGURE 8 | Confirmation of the expression of circRNAs and miRNAs using quantitative real-time polymerase chain reaction (qRT-PCR). Sample numbers were expanded to perform qRT-PCR (CRSwNP (n = 29); control subjects (n = 25). GAPDH was used as the reference gene for circRNAs, and hsa-miR-16 was used as the reference gene for miRNAs. The heights of columns represent the fold changes of CRSwNP compared with the control group. The relative expression levels of hsa-circ-0031593, hsa-circ-0031594, hsa-circ-0109623, hsa-circ-0000736, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, hsa-miR-125a-5p and hsa-miR-27b-3p by qRT-PCR. *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001.

hsa-miR-145-5p had the strongest evidence supporting their involvement in the occurrence and development of CRSwNP by ROC curve analysis and PCA. In conclusion, our findings revealed a network potentially involved in CRSwNP pathogenesis, in which circRNAs and microRNAs play significant roles.

ECRSwNP differs greatly from non-ECRSwNP in many aspects (Shah et al., 2016; Lou et al., 2018), such as pathogenesis, development, prognosis, and CT scan images. Our original intention was to explore the functions of miRNA and circRNA in different CRSwNP subtypes, in order to help postoperative treatments like determining the eosinophil count. In addition, some researchers have found that ECRSwNP is difficult to treat and has a high recurrence rate, leading to poor clinical outcomes. However, in our study, the expressions of circRNA and miRNA between ECRSwNP and non-ECRSwNP had no statistical differences. It is possible that individual differences of the samples or the regulation of the next biological process has changed, resulting in different types of polyps. The development process of CRSwNP is complex and diverse (Schleimer, 2017). Meanwhile, the classification between ECRSwNP and non-ECRSwNP is limited. First, there is no unified view on the determination of ECRSwNP throughout the world. Second, the count of eosinophils is objective. Therefore, we combined ECRSwNP and non-ECRSwNP into the CRSwNP group.

After analyzing the different expressions of circRNAs and miRNAs between the CRSwNP and control groups, we found that 1794 circRNAs were significantly downregulated and 1,081 were significantly upregulated in the CRSwNP group. Additionally, 192 miRNAs showed significant downregulation in the CRSwNP

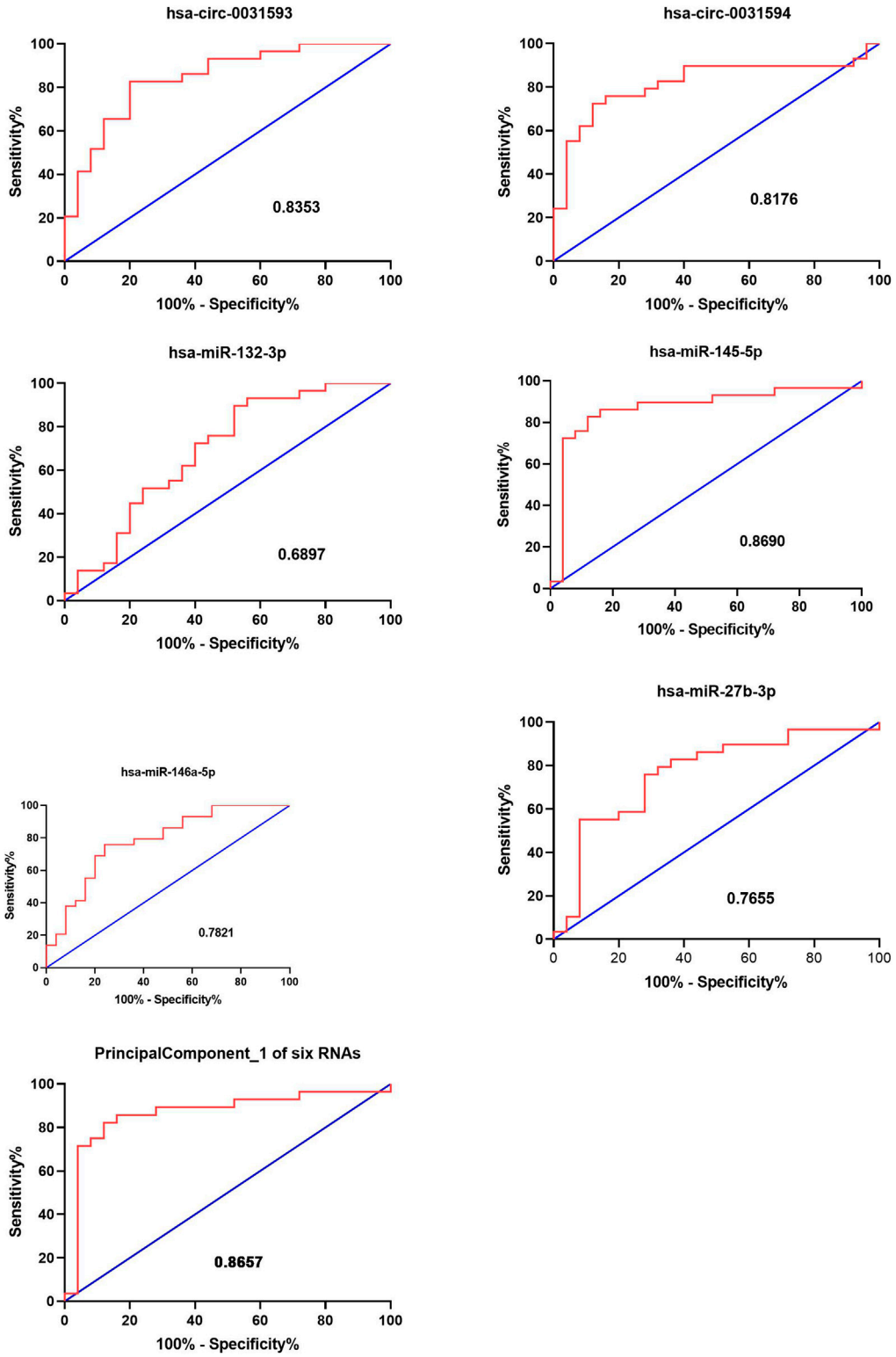


FIGURE 9 | Functional value of circRNAs and miRNAs. The receiver operating characteristic (ROC) curve analysis for the function of CRSwNP. The ROC curve analysis of hsa-circ-0031593, hsa-circ-0031594, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p and principal component 1 of these six ncRNAs for the significance in the occurrence and development of CRSwNP. The AUC (area under curves) values are given on the graphs.

TABLE 1 | Validation of the selected circRNAs and miRNAS by quantitative real-time polymerase chain reaction and the data of ROC curve analysis.

	AUC	95% CI	*p* value	Sensitivity	Specificity
Hsa-circ-0031593	0.8353	0.7291–0.9415	<0.0001	0.8235	0.80
Hsa-circ-0031594	0.8176	0.7047–0.9306	<0.0001	0.7059	0.88
Hsa-miR-132-3p	0.6897	0.5438–0.8355	0.0171	0.8966	0.48
Hsa-miR-145-5p	0.8690	0.76–0.978	<0.0001	0.8276	0.88
Hsa-miR-146a-5p	0.7821	0.6579–0.9063	0.0004	0.7586	0.76
Hsa-miR-27b-3p	0.7655	0.6332–0.8979	0.0008	0.7586	0.72
Principal component 1 of six RNAs	0.8657	0.7547–0.9768	<0.0001	0.8214	0.88

AUC, area under curves.

group and none showed upregulation. The different expressions of circRNAs and miRNAs between these two groups may be involved in the pathogenesis of CRSwNP.

Studies have shown that one of the important functions of circRNAs is to act as "miRNA sponges" and competitively bind miRNAs to regulate post-transcriptional activity (Ergun and Oztuzcu, 2015). Co-expression networks have been constructed to obtain the relationship between circRNAs and miRNAs. **Figure 3** shows that a single circRNA is associated with multiple miRNAs, and a single miRNA is associated with multiple circRNAs. Little is known about the relationship between circRNAs and miRNAs in CRSwNP. In fact, the present study is the first to detect circRNA in CRSwNP.

Without functional analysis, the huge quantity of data on gene expression is unintelligible. Hence, functional enrichment analysis and prediction of differentially expressed genes were carried out using public databases, such as GO enrichment analysis and KEGG pathway analysis, and we found that differential expressions of circRNAs between the CRNwNP and control groups were related to "amoebiasis," "salivary secretion," "cell adhesion molecules (CAMs)," "cAMP signaling pathway," "focal adhesion", "adherens junction," "TNF signaling pathway," and others. The differential expressions of miRNAs between the CRSwNP and control groups were enriched in "pathways in cancer," "endocytosis," "thyroid hormone signaling pathway," "salivary secretion," "regulation of actin cytoskeleton," "insulin secretion," and so on. Based on the functional analysis, the most significantly enriched pathway was CAMs, which was consistent with previous research (Milonski et al., 2015; Xu et al., 2017). The pathological characteristics of CRSwNP include inflammatory cells migrating to and infiltrating the nasal mucosa. During different stages of the progression of CRSwNP, the expressions of CAMs are various, which stimulate eosinophil and mast cell aggregation and contribute to Th2 skewing (Oyer et al., 2013). Compared to normal nasal mucosa, CRSwNP was shown to be more sensitive to IL-32 through lipopolysaccharides acting at the cAMP signaling pathway (Cho et al., 2016). Further, a study found that thromboxane A2 is involved in platelet aggregation and tissue inflammation in CRS, and cAMP regulates the expression of the thromboxane-prostanoid receptor and cxcl1/8, which participates in the pathogenesis of CRS (Elion et al., 2018). TNF, a complex and important inflammatory factor, induces local production of IgA and stimulates eosinophils, and it plays an important role in the pathogenesis of CRSwNP (Kato et al., 2008; Cho et al., 2015; Shimizu et al., 2016). The integrity of the airway epithelium is a prerequisite for its good barrier function, which depends on the intercellular junctions, including tight junctions and adhesion junctions (Suzuki et al., 2016; Jiao et al., 2018). Studies have shown that the breakdown of tight junctions and adhesion junctions of CRS and the decrease of protein component expression are major factors leading to the occurrence of CRS (Kim et al., 2018; Tian et al., 2018). Studies on the key genes and pathways in CRSwNP showed that salivary secretion was the most significantly enriched pathway for downregulated genes, which was consistent with our findings (Yao et al., 2019). Above all, our findings confirmed the validity of previous research and showed high reliability. Little is known about the regulation of the actin cytoskeleton and insulin secretion in CRSwNP. Studies have shown that the actin cytoskeleton is associated with several inflammatory diseases and is involved in leukocyte transendothelial migration (Schnoor, 2015; Ao et al., 2016; Lechuga and Ivanov, 2017). Besides, numerous studies have shown that the actin cytoskeleton plays an important role in regulating insulin secretion (Martínez-García et al., 2015; Sorrenson et al., 2016; Deyev et al., 2017). However, the functions of the actin cytoskeleton and insulin secretion in CRSwNP need to be further explored.

The heterogeneity of CRSwNP is gradually being recognized, prompting the discovery of novel biomarkers to describe specific endotypes and determine optimized treatment (Dennis et al., 2016; Kuhar et al., 2017). Recent studies on potential biomarkers in CRSwNP mainly focused on eosinophils, exhaled gas components, and inflammatory cells in nasal secretions, nasal tissues, and peripheral blood (Drake et al., 2016; Tsybikov et al., 2016; Asano et al., 2017; Chen et al., 2017). Yan indicated that miR-145-5p negatively regulates the proliferation and chemokine secretion of NHEKs by targeting MLK3, and the downregulation of miR-145-5p contributes to skin inflammation in psoriasis lesions (Yan et al., 2019). Dihydroquercetin attenuates lipopolysaccharide-induced acute lung injury by modulating FOXO3-mediated NF-κB signaling via miR-132–3p (Liu J.-H. et al., 2020). Human neutrophil elastase induces MUC5AC overexpression in chronic rhinosinusitis through miR-146a (Yan et al., 2020). Furthermore, miR-27b-3p, miR-181a-1-3p, and miR-326-5p are involved in the inhibition of macrophage activation in chronic liver injury (Li et al., 2017). Circ_0134111 knockdown relieves IL-1β-induced apoptosis, inflammation, and extracellular matrix degradation in human chondrocytes through the circ_0134111-miR-515-5p-SOCS1 network (Wu et al., 2021). An inducible circular RNA circKcnt2 inhibits ILC3 activation to facilitate colitis resolution (Liu B. et al., 2020). These studies show that circRNA and miRNA play vital functions in the process of inflammation. In the present study, ROC curve analysis and PCA indicated that aberrantly expressed circRNAs and miRNAs may be related to biological dysfunction

TABLE 2 | The data of principal component 1 by PCA with combinations of hsa-circ-0031593, hsa-circ-0031594, hsa-miR-132-3p, hsa-miR-145-5p, hsa-miR-146a-5p, and hsa-miR-27b-3p.

Principal component 1	Type of sample
−0.207152043	Control
2.500343003	Control
0.987185596	Control
1.304084045	Control
−0.361852501	Control
0.755023033	Control
−1.570053914	Control
−0.481308455	Control
0.906674823	Control
0.30217685	Control
2.353877347	Control
1.965290304	Control
−0.28100719	Control
1.564228751	Control
3.054599405	Control
−0.126055013	Control
0.567948512	Control
1.087283413	Control
−0.009538338	Control
0.62720042	Control
0.255341755	Control
−0.162273621	Control
0.38535951	Control
1.632552725	Control
0.966021961	Control
−1.33220639	CRSwNP
−0.361203299	CRSwNP
−0.386665708	CRSwNP
−1.597233545	CRSwNP
−1.396044339	CRSwNP
−1.205543467	CRSwNP
−1.376653623	CRSwNP
4.157249255	CRSwNP
−0.902849991	CRSwNP
−0.860962834	CRSwNP
−0.22755494	CRSwNP
−1.464884451	CRSwNP
−0.876447159	CRSwNP
0.712907312	CRSwNP
−0.962998826	CRSwNP
1.21901289	CRSwNP
−0.685225616	CRSwNP
−1.19975492	CRSwNP
−1.560321	CRSwNP
−0.314747233	CRSwNP
−0.088432428	CRSwNP
−1.263311269	CRSwNP
−0.777328458	CRSwNP
−1.238227185	CRSwNP
−0.490926392	CRSwNP
−1.537602896	CRSwNP
−0.910574069	CRSwNP
−1.087419795	CRSwNP

and play important roles in the pathogenesis of CRSwNP. Furthermore, hsa-circ-0031593 and hsa-miR-145-5p were more likely to correlate with the occurrence and development of CRSwNP.

There are still some limitations in our study. First, individual differences of the samples may have led to the lack of statistical differences between the ECRSwNP and non-ECRSwNP groups, although our findings were consistent with previous research (Cho et al., 2016; Suzuki et al., 2016; Xu et al., 2017; Kim et al., 2018; Yao et al., 2019). Second, all patients' data were from The First Affiliated Hospital of Nanchang University, and all patient were from Jiangxi Province. Although most of the Chinese population is Han, given that ethnic and regional variations may be involved in the development of CRSwNP, we will consider these variables in future studies. Third, our findings were only based on gene chip analysis, database comparison and prediction, and tissue experiment verification. Experiments *in vivo* and *in vitro* should be carried out to further explore the function of these aberrant genes in CRSwNP.

CONCLUSION

In our study, the expression profiles of ECRSwNP and non-ECRSwNP had no statistical differences. The differentially expressed circRNAs and miRNAs between the CRSwNP and control groups may play important roles in the pathogenesis of CRSwNP. Altered expression of hsa-circ-0031593 and hsa-miR-145-5p had the strongest evidence for involvement in the occurrence and development of CRSwNP.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The name of the repository and accession numbers can be found below: National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), https://www.ncbi.nlm.nih.gov/geo/, GSE169375 and GSE169376'

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Department of Otorhinolaryngology and Head Neck Surgery, the First Affiliated Hospital of Nanchang University, Nanchang, China in 2017. The study was approved by the Medical Research Ethics Committee of the hospital (2017080). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JYU, XK, and JYE conceived of or designed the work; JYU, XK, and JYE drafted the work; data acquisition was undertaken by YX and QL; data analysis was completed by DD; supervision or mentorship was done by JYU, YX, and XK. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by National Natural Science Foundation of China (No. 81860182), Jiangxi Natural Science Foundation (No.20181BAB205036) and Jiangxi Natural Science Foundation (No. 20192BBGL70025).

ACKNOWLEDGMENTS

We thank LetPub (www.letpub.com) for providing linguistic assistance during the preparation of this manuscript. Thank Yanqing Yu for her assistance in counting the tissue eosinophil number.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.643504/full#supplementary-material

REFERENCES

Ai-Sayed, A. A., Agu, R. U., and Massoud, E. (2017). Models for the Study of Nasal and Sinus Physiology in Health and Disease: A Review of the Literature. *Laryngoscope Invest. Otolaryngol.* 2 (6), 398–409. doi:10.1002/lio2.117

Ao, M., Wu, L., Zhou, X., and Chen, Y. (2016). Methyl-β-Cyclodextrin Impairs the Monocyte-Adhering Ability of Endothelial Cells by Down-Regulating Adhesion Molecules and Caveolae and Reorganizing the Actin Cytoskeleton. *Biol. Pharm. Bull.* 39 (6), 1029–1034. doi:10.1248/bpb.b16-00047

Asano, T., Kanemitsu, Y., Takemura, M., Yokota, M., Fukumitsu, K., Takeda, N., et al. (2017). Serum Periostin as a Biomarker for Comorbid Chronic Rhinosinusitis in Patients with Asthma. *Ann. ATS* 14 (5), 667–675. doi:10.1513/annalsats.201609-720oc

Beermann, J., Piccoli, M.-T., Viereck, J., and Thum, T. (2016). Non-coding Rnas in Development and Disease: Background, Mechanisms, and Therapeutic Approaches. *Physiol. Rev.* 96 (4), 1297–1325. doi:10.1152/physrev.00041.2015

Cao, P.-P., Li, H.-B., Wang, B.-F., Wang, S.-B., You, X.-J., Cui, Y.-H., et al. (2009). Distinct Immunopathologic Characteristics of Various Types of Chronic Rhinosinusitis in Adult Chinese. *J. Allergy Clin. Immunol.* 124 (3), 478–484. doi:10.1016/j.jaci.2009.05.017

Chen, C., Yin, P., Hu, S., Sun, X., and Li, B. (2020). Circular RNA-9119 Protects IL-1β-treated Chondrocytes from Apoptosis in an Osteoarthritis Cell Model by Intercepting the microRNA-26a/PTEN axis. *Life Sci.* 256, 117924. doi:10.1016/j.lfs.2020.117924

Chen, F., Hong, H., Sun, Y., Hu, X., Zhang, J., Xu, G., et al. (2017). Nasal Interleukin 25 as a Novel Biomarker for Patients with Chronic Rhinosinusitis with Nasal Polyps and Airway Hypersensitiveness. *Ann. Allergy Asthma Immunol.* 119 (4), 310–316. doi:10.1016/j.anai.2017.07.012

Cho, J.-S., Kim, J.-A., Park, J.-H., Park, I.-H., Han, I.-H., and Lee, H.-M. (2016). Toll-like Receptor 4-mediated Expression of Interleukin-32 via the C-Jun N-Terminal Kinase/protein Kinase B/cyclic Adenosine Monophosphate Response Element Binding Protein Pathway in Chronic Rhinosinusitis with Nasal Polyps. *Int. Forum Allergy Rhinol.* 6 (10), 1020–1028. doi:10.1002/alr.21792

Cho, S.-W., Kim, D. W., Kim, J.-W., Lee, C. H., and Rhee, C.-S. (2017). Classification of Chronic Rhinosinusitis According to a Nasal Polyp and Tissue Eosinophilia: Limitation of Current Classification System for Asian Population. *Asia Pac. Allergy* 7 (3), 121–130. doi:10.5415/apallergy.2017.7.3.121

Cho, S. H., Kim, D. W., Lee, S. H., Kolliputi, N., Hong, S. J., Suh, L., et al. (2015). Age-related Increased Prevalence of Asthma and Nasal Polyps in Chronic Rhinosinusitis and its Association with Altered IL-6 Trans-signaling. *Am. J. Respir. Cel Mol Biol* 53 (5), 601–606. doi:10.1165/rcmb.2015-0207rc

Dennis, S. K., Lam, K., and Luong, A. (2016). A Review of Classification Schemes for Chronic Rhinosinusitis with Nasal Polyposis Endotypes. *Laryngoscope Invest. Otolaryngol.* 1 (5), 130–134. doi:10.1002/lio2.32

Deyev, I. E., Popova, N. V., Serova, O. V., Zhenilo, S. V., Regoli, M., Bertelli, E., et al. (2017). Alkaline pH Induces IRR-Mediated Phosphorylation of IRS-1 and Actin Cytoskeleton Remodeling in a Pancreatic Beta Cell Line. *Biochimie* 138, 62–69. doi:10.1016/j.biochi.2017.04.002

Drake, V. E., Rafaels, N., and Kim, J. (2016). Peripheral Blood Eosinophilia Correlates with Hyperplastic Nasal Polyp Growth. *Int. Forum Allergy Rhinol.* 6 (9), 926–934. doi:10.1002/alr.21793

Elion, R. A., Althoff, K. N., Zhang, J., Moore, R. D., Gange, S. J., Kitahata, M. M., et al. (2018). Recent Abacavir Use Increases Risk of Type 1 and Type 2 Myocardial Infarctions Among Adults with HIV. *Jaids J. Acquired Immune Deficiency Syndromes* 78 (1), 62–72. doi:10.1097/qai.0000000000001642

Ergun, S., and Oztuzcu, S. (2015). Oncocers: Cerna-Mediated Cross-Talk by Sponging Mirnas in Oncogenic Pathways. *Tumor Biol.* 36 (5), 3129–3136. doi:10.1007/s13277-015-3346-x

Ferreira, A. F., Calin, G. A., Picanço-Castro, V., Kashima, S., Covas, D. T., and de Castro, F. A. (2018). Hematopoietic Stem Cells from Induced Pluripotent Stem Cells - Considering the Role of Microrna as a Cell Differentiation Regulator. *J. Cel Sci* 131 (4), jcs203018. doi:10.1242/jcs.203018

Fokkens, W. J., Lund, V. J., Mullol, J., Bachert, C., Alobid, I., Baroody, F., et al. (2012). EPOS 2012: European Position Paper on Rhinosinusitis and Nasal Polyps 2012. A Summary for Otorhinolaryngologists. 2012. A Summary for Otorhinolaryngologists. *Rhinology* 50 (1), 1–12. doi:10.4193/rhino50e2

He, Q., Liu, N., Hu, F., Shi, Q., Pi, X., Chen, H., et al. (2021). Circ_0061012 Contributes to IL-22-induced Proliferation, Migration and Invasion in Keratinocytes through miR-194-5p/GAB1 axis in Psoriasis. *Biosci. Rep.* 41 (1), BSR20203130. doi:10.1042/bsr20203130

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., et al. (2013). Circular Rnas Are Abundant, Conserved, and Associated with Alu Repeats. *RNA* 19 (2), 141–157. doi:10.1261/rna.035667.112

Jiao, J., Wang, M., Duan, S., Meng, Y., Meng, N., Li, Y., et al. (2018). Transforming Growth Factor-B1 Decreases Epithelial Tight junction Integrity in Chronic Rhinosinusitis with Nasal Polyps. *J. Allergy Clin. Immunol.* 141 (3), 1160–1163. doi:10.1016/j.jaci.2017.08.045

Kato, A., Peters, A., Suh, L., Carter, R., Harris, K. E., Chandra, R., et al. (2008). Evidence of a Role for B Cell-Activating Factor of the TNF Family in the Pathogenesis of Chronic Rhinosinusitis with Nasal Polyps. *J. Allergy Clin. Immunol.* 121 (6), 1385–1392. doi:10.1016/j.jaci.2008.03.002

Kim, B., Lee, H.-J., Im, N.-R., Lee, D. Y., Kang, C. Y., Park, I.-H., et al. (2018). Effect of Matrix Metalloproteinase Inhibitor on Disrupted E-Cadherin after Acid Exposure in the Human Nasal Epithelium. *The Laryngoscope* 128 (1), E1–E7. doi:10.1002/lary.26932

Korde, A., Ahangari, F., Haslip, M., Zhang, X., Liu, Q., Cohn, L., et al. (2020). An Endothelial microRNA-1-Regulated Network Controls Eosinophil Trafficking in Asthma and Chronic Rhinosinusitis. *J. Allergy Clin. Immunol.* 145 (2), 550–562. doi:10.1016/j.jaci.2019.10.031

Kuhar, H. N., Tajudeen, B. A., Mahdavinia, M., Gattuso, P., Ghai, R., and Batra, P. S. (2017). Inflammatory Infiltrate and Mucosal Remodeling in Chronic Rhinosinusitis with and without Polyps: Structured Histopathologic Analysis. *Int. Forum Allergy Rhinol.* 7 (7), 679–689. doi:10.1002/alr.21943

Kulcheski, F. R., Christoff, A. P., and Margis, R. (2016). Circular Rnas Are Mirna Sponges and Can Be Used as a New Class of Biomarker. *J. Biotechnol.* 238 (20), 42–51. doi:10.1016/j.jbiotec.2016.09.011

Lechuga, S., and Ivanov, A. I. (2017). Disruption of the Epithelial Barrier during Intestinal Inflammation: Quest for New Molecules and Mechanisms. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1864 (7), 1183–1194. doi:10.1016/j.bbamcr.2017.03.007

Li, G., Qin, Y., Qin, S., Zhou, X., Zhao, W., and Zhang, D. (2020). Circ_WBSCR17 Aggravates Inflammatory Responses and Fibrosis by Targeting miR-185-5p/SOX6 Regulatory axis in High Glucose-Induced Human Kidney Tubular Cells. *Life Sci.* 259, 118269. doi:10.1016/j.lfs.2020.118269

Li, W., Chang, N., Tian, L., Yang, J., Ji, X., Xie, J., et al. (2017). miR-27b-3p, miR-181a-1-3p, and miR-326-5p Are Involved in the Inhibition of Macrophage Activation in Chronic Liver Injury. *J. Mol. Med.* 95 (10), 1091–1105. doi:10.1007/s00109-017-1570-0

Liu, B., Ye, B., Zhu, X., Yang, L., Li, H., Liu, N., et al. (2020). An Inducible Circular RNA circKcnt2 Inhibits ILC3 Activation to Facilitate Colitis Resolution. *Nat. Commun.* 11 (1), 4076. doi:10.1038/s41467-020-17944-5

Liu, J.-H., Cao, L., Zhang, C.-H., Li, C., Zhang, Z.-H., and Wu, Q. (2020). Dihydroquercetin Attenuates Lipopolysaccharide-Induced Acute Lung Injury through Modulating FOXO3-Mediated NF-Kb Signaling via miR-132-3p. *Pulm. Pharmacol. Ther.* 64, 101934. doi:10.1016/j.pupt.2020.101934

Lou, H., Meng, Y., Piao, Y., Wang, C., Zhang, L., and Bachert, C. (2015). Predictive Significance of Tissue Eosinophilia for Nasal Polyp Recurrence in the Chinese Population. *Am. J. Rhinol Allergy* 29 (5), 350–356. doi:10.2500/ajra.2015.29.4231

Lou, H., Zhang, N., Bachert, C., and Zhang, L. (2018). Highlights of Eosinophilic Chronic Rhinosinusitis with Nasal Polyps in Definition, Prognosis, and Advancement. *Int. Forum Allergy Rhinol.* 8 (11), 1218–1225. doi:10.1002/alr.22214

Martínez-García, C., Izquierdo-Lahuerta, A., Vivas, Y., Velasco, I., Yeo, T. K., and Chen, S. (2015). Renal Lipotoxicity-Associated Inflammation and Insulin Resistance Affects Actin Cytoskeleton Organization in Podocytes. *PLoS One* 10 (11), e0142291. doi:10.1371/journal.pone.0142291

Martínez-Rivera, V., Negrete-García, M., Ávila-Moreno, F., and Ortiz-Quintero, B. (2018). Secreted and Tissue Mirnas as Diagnosis Biomarkers of Malignant Pleural Mesothelioma. *Int. J. Mol. Sci.* 19 (2), 595–622. doi:10.3390/ijms19020595

Milonski, J., Zielinska-Blizniewska, H., Majsterek, I., Przybyłowska-Sygut, K., Sitarek, P., Korzycka-Zaborowska, B., et al. (2015). Expression of POSTN, IL-4, and IL-13 in Chronic Rhinosinusitis with Nasal Polyps. *DNA Cel Biol.* 34 (5), 342–349. doi:10.1089/dna.2014.2712

Oude Voshaar, M. A. H., Das Gupta, Z., Bijlsma, J. W. J., Boonen, A., Chau, J., Courvoisier, D. S., et al. (2019). International Consortium for Health Outcome Measurement Set of Outcomes that Matter to People Living with Inflammatory Arthritis: Consensus from an International Working Group. *Arthritis Care Res.* 71 (12), 1556–1565. doi:10.1002/acr.23799

Oyer, S. L., Nagel, W., and Mulligan, J. K. (2013). Differential Expression of Adhesion Molecules by Sinonasal Fibroblasts Among Control and Chronic Rhinosinusitis Patients. *Am. J. Rhinol Allergy* 27 (5), 381–386. doi:10.2500/ajra.2013.27.3934

Peng, K., Jiang, P., Du, Y., Zeng, D., Zhao, J., Li, M., et al. (2021). Oxidized Low-density Lipoprotein Accelerates the Injury of Endothelial Cells via circ-USP36/miR -98-5p/VCAM1 axis. *IUBMB Life* 73 (1), 177–187. doi:10.1002/iub.2419

Plager, D. A., Kahl, J. C., Asmann, Y. W., Nilson, A. E., Pallanch, J. F., Friedman, O., et al. (2010). Gene Transcription Changes in Asthmatic Chronic Rhinosinusitis with Nasal Polyps and Comparison to Those in Atopic Dermatitis. *PLoS One* 5 (7), e11450. doi:10.1371/journal.pone.0011450

Schleimer, R. P. (2017). Immunopathogenesis of Chronic Rhinosinusitis and Nasal Polyposis. *Annu. Rev. Pathol. Mech. Dis.* 12, 331–357. doi:10.1146/annurev-pathol-052016-100401

Schnoor, M. (2015). Endothelial Actin-Binding Proteins and Actin Dynamics in Leukocyte Transendothelial Migration. *J. Immunol.* 194 (8), 3535–3541. doi:10.4049/jimmunol.1403250

Shah, S. A., Ishinaga, H., and Takeuchi, K. (2016). Pathogenesis of Eosinophilic Chronic Rhinosinusitis. *J. Inflamm. (Lond)* 13 (1), 11. doi:10.1186/s12950-016-0121-8

Shi, L.-L., Xiong, P., Zhang, L., Cao, P.-P., Liao, B., Lu, X., et al. (2013). Features of Airway Remodeling in Different Types of Chinese Chronic Rhinosinusitis Are Associated with Inflammation Patterns. *Allergy* 68 (1), 101–109. doi:10.1111/all.12064

Shimizu, S., Kouzaki, H., Kato, T., Tojima, I., and Shimizu, T. (2016). HMGB1-TLR4 Signaling Contributes to the Secretion of Interleukin 6 and Interleukin 8 by Nasal Epithelial Cells. *Am. J. Rhinol Allergy* 30 (3), 167–172. doi:10.2500/ajra.2016.30.4300

Sorrenson, B., Cognard, E., Lee, K. L., Dissanayake, W. C., Fu, Y., Han, W., et al. (2016). A Critical Role for β-Catenin in Modulating Levels of Insulin Secretion from β-Cells by Regulating Actin Cytoskeleton and Insulin Vesicle Localization. *J. Biol. Chem.* 291 (50), 25888–25900. doi:10.1074/jbc.m116.758516

Suzuki, H., Koizumi, H., Ikezaki, S., Tabata, T., Ohkubo, J.-i., Kitamura, T., et al. (2016). Electrical Impedance and Expression of Tight junction Components of the Nasal Turbinate and Polyp. *ORL J. Otorhinolaryngol. Relat. Spec.* 78 (1), 16–25. doi:10.1159/000442024

Tian, T., Zi, X., Peng, Y., Wang, Z., Hong, H., Yan, Y., et al. (2018). H3N2 Influenza Virus Infection Enhances Oncostatin M Expression in Human Nasal Epithelium. *Exp. Cel Res.* 371 (2), 322–329. doi:10.1016/j.yexcr.2018.08.022

Tsybikov, N. N., Egorova, E. V., Kuznik, B. I., Fefelova, E. V., and Magen, E. (2016). Neuron-specific Enolase in Nasal Secretions as a Novel Biomarker of Olfactory Dysfunction in Chronic Rhinosinusitis. *Am. J. Rhinol Allergy* 30 (1), 65–69. doi:10.2500/ajra.2016.30.4264

Wu, G., Sun, Y., Xiang, Z., Wang, K., Liu, B., Xiao, G., et al. (2019). Preclinical Study Using Circular Rna 17 and Micro Rna 181c-5p to Suppress the Enzalutamide-Resistant Prostate Cancer Progression. *Cell Death Dis* 10 (2), 37. doi:10.1038/s41419-018-1048-1

Wu, R., Zhang, F., Cai, Y., Long, Z., Duan, Z., Wu, D., et al. (2021). Circ_0134111 Knockdown Relieves IL-1β-induced Apoptosis, Inflammation and Extracellular Matrix Degradation in Human Chondrocytes through the Circ_0134111-miR-515-5p-SOCS1 Network. *Int. Immunopharmacology* 95, 107495. doi:10.1016/j.intimp.2021.107495

Xia, S., Feng, J., Lei, L., Hu, J., Xia, L., Wang, J., et al. (2017). Comprehensive Characterization of Tissue-specific Circular Rnas in the Human and Mouse Genomes. *Brief Bioinform* 18 (6), 984–992. doi:10.1093/bib/bbw081

Xu, M., Chen, D., Zhou, H., Zhang, W., Xu, J., and Chen, L. (2017). The Role of Periostin in the Occurrence and Progression of Eosinophilic Chronic Sinusitis with Nasal Polyps. *Sci. Rep.* 7 (1), 9479. doi:10.1038/s41598-017-08375-2

Yan, D., Ye, Y., Zhang, J., Zhao, J., Yu, J., and Luo, Q. (2020). Human Neutrophil Elastase Induces MUC5AC Overexpression in Chronic Rhinosinusitis through miR-146a. *Am. J. Rhinol Allergy* 34 (1), 59–69. doi:10.1177/1945892419871798

Yan, J. J., Qiao, M., Li, R. H., Zhao, X. T., Wang, X. Y., and Sun, Q. (2019). Downregulation of miR-145-5p Contributes to Hyperproliferation of Keratinocytes and Skin Inflammation in Psoriasis. *Br. J. Dermatol.* 180 (2), 365–372. doi:10.1111/bjd.17256

Yang, J., Cheng, M., Gu, B., Wang, J., Yan, S., and Xu, D. (2020). CircRNA_09505 Aggravates Inflammation and Joint Damage in Collagen-Induced Arthritis Mice via miR-6089/AKT1/NF-Kb axis. *Cel Death Dis* 11 (10), 833. doi:10.1038/s41419-020-03038-z

Yang, R., Xing, L., Zheng, X., Sun, Y., Wang, X., and Chen, J. (2019). The circRNA circAGFG1 Acts as a Sponge of miR-195-5p to Promote Triple-Negative Breast Cancer Progression through Regulating CCNE1 Expression. *Mol. Cancer* 18 (1), 4. doi:10.1186/s12943-018-0933-7

Yao, Y., Xie, S., and Wang, F. (2019). Identification of Key Genes and Pathways in Chronic Rhinosinusitis with Nasal Polyps Using Bioinformatics Analysis. *Am. J. Otolaryngol.* 40 (2), 191–196. doi:10.1016/j.amjoto.2018.12.002

Zhang, X.-H., Zhang, Y.-N., Li, H.-B., Hu, C.-Y., Wang, N., Cao, P.-P., et al. (2012). Overexpression of Mir-125b, a Novel Regulator of Innate Immunity, in Eosinophilic Chronic Rhinosinusitis with Nasal Polyps. *Am. J. Respir. Crit. Care Med.* 185 (2), 140–151. doi:10.1164/rccm.201103-0456oc

Zhang, Y.-N., Cao, P.-P., Zhang, X.-H., Lu, X., and Liu, Z. (2012). Expression of microRNA Machinery Proteins in Different Types of Chronic Rhinosinusitis. *The Laryngoscope* 122 (12), 2621–2627. doi:10.1002/lary.23517

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

frontiers in Genetics

METHODS
published: 31 May 2021
doi: 10.3389/fgene.2021.664654

Integrative Ranking of Enhancer Networks Facilitates the Discovery of Epigenetic Markers in Cancer

Qi Wang[1,2], *Yonghe Wu*[3], *Tim Vorberg*[2], *Roland Eils*[1,4] *and Carl Herrmann*[1]*

[1] Health Data Science Unit, Medical Faculty Heidelberg and BioQuant, Heidelberg, Germany, [2] Faculty of Biosciences, Heidelberg University, Heidelberg, Germany, [3] Division of Molecular Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany, [4] Digital Health Center, Berlin Institute of Health (BIH) and Charité, Berlin, Germany

Regulation of gene expression through multiple epigenetic components is a highly combinatorial process. Alterations in any of these layers, as is commonly found in cancer diseases, can lead to a cascade of downstream effects on tumor suppressor or oncogenes. Hence, deciphering the effects of epigenetic alterations on regulatory elements requires innovative computational approaches that can benefit from the huge amounts of epigenomic datasets that are available from multiple consortia, such as Roadmap or BluePrint. We developed a software tool named IRENE (Integrative Ranking of Epigenetic Network of Enhancers), which performs quantitative analyses on differential epigenetic modifications through an integrated, network-based approach. The method takes into account the additive effect of alterations on multiple regulatory elements of a gene. Applying this tool to well-characterized test cases, it successfully found many known cancer genes from publicly available cancer epigenome datasets.

Keywords: enhancer, epigenetics, histone modification, chromatin interaction, network analysis

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Fuhai Li,
Washington University in St. Louis, United States
Loredana Martignetti,
INSERM U900 Cancer Et Génome Bioinformatique, Biostatistiques Et Épidémiologie, France

***Correspondence:**
Carl Herrmann
carl.herrmann@bioquant.uni-heidelberg.de

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

Received: *05 February 2021*
Accepted: *29 March 2021*
Published: *31 May 2021*

Citation:
Wang Q, Wu Y, Vorberg T, Eils R and Herrmann C (2021) Integrative Ranking of Enhancer Networks Facilitates the Discovery of Epigenetic Markers in Cancer. Front. Genet. 12:664654. doi: 10.3389/fgene.2021.664654

INTRODUCTION

Epigenetic alterations are frequent in many cancers. In particular, DNA methylation and histone modifications are two main mechanisms that allow cancer cells to alter transcription without changing the DNA sequences, and lead to many abnormalities such as persistent activation of cell cycle control genes or deactivation of DNA repair genes. For example, promoter DNA hypo-methylation accompanied by histone hyper-acetylation is frequently observed in the activation of oncogenes in cancer. Besides, aberrant activation of distal regulatory elements is often associated with the up-regulation of cancer-promoting genes. Interestingly, epigenetic modifications at proximal and distal regulatory elements often appear to be earlier events than the gene expression (Hartley et al., 2013; Ziller et al., 2014), and can hence serve as potential early markers in cancer diagnosis.

Various histone modifications on promoters have been categorized into either activation or repression effects on gene expression. Such effects can be measured by comparing histone alteration levels between tumor and their corresponding normal tissues using ChIP-Seq (Karlic et al., 2010). A number of tools, such as ChIPComp (Chen et al., 2015), ChIPDiff (Xu et al., 2008), ChIPnorm (Nair et al., 2012), csaw (Lun and Smyth, 2015), DBChIP (Liang and Keles, 2012), DiffBind (Stark and Brown, 2011), MAnorm (Shao et al., 2012), RSEG (Song and Smith, 2011) have demonstrated their usefulness in cancer studies by comparing the histone intensities between two conditions (see Steinhauser et al., 2016 for a review of these tools). However, they are limited to the comparison

of a single histone mark. Furthermore, many tools such as jMOSAiCS (Zeng et al., 2013), IDEAS (Zhang et al., 2016), and ChromHMM (Ernst and Kellis, 2012) are able to perform integrative analyses across multiple epigenetic marks. However, while these tools provide an integrated description of the epigenetic characteristics at individual genome loci, they do not take into account the combined effects of these changes at multiple regulatory elements controlling a gene.

As previously mentioned, many histone modifications that potentially regulate gene expression also occur in other genomic regions besides promoters. Enhancers are distal regulatory elements that interact with gene promoters through chromosomal loops to regulate gene transcription. Most of the enhancers are located within ±1 Mb of the transcription start site (TSS) of their target genes (Maston et al., 2006). Enhancer activity is regulated through epigenetic modifications (Zentner et al., 2011), including positive regulation from histone marks, such as H3K27ac (Creyghton et al., 2010; Stasevich et al., 2014) and H3K4me1 (Heintzman et al., 2007; Calo and Wysocka, 2013), and negative regulation by H3K27me3 (Charlet et al., 2016) and H3K9me3 (Zhu et al., 2012).

Given the complexity of epigenetic regulation, novel tools are required to combine this information, and create a comprehensive overview of the differential epigenetic landscape, integrating multiple data layers. The method we developed, named IRENE (Integrative ranking with an epigenetic network of enhancers), combines a quantitative analysis on multiple differential epigenetic modifications with an integrated, network-based approach, in which we integrated two levels of epigenetic information: the signal intensity of each epigenetic mark, and the relationships between promoters and distal regulatory elements known as enhancers (**Figure 1**). In this paper, we describe the method and present the test cases. In our benchmarking tests on cancer datasets, the IRENE ranked lists have higher relevance to cancer marker genes (CMGs) than the other approaches. Being implemented as an R package, IRENE is an easy to use method allowing gene ranking between two conditions and highlighting potential cancer biomarkers.

RESULTS

IRENE: Epigenetic Ranking With an Epigenetic Network of Enhancers

IRENE analyzes epigenetic changes between two biological conditions (e.g., ChIP-seq data for histone modifications or whole-genome bisulfite sequencing for DNA methylation), and translates the differential signals at multiple regulatory elements into a unique score (**Figure 1**). Hence, IRENE performs a double integration, both across multiple epigenetic datasets and across different regulatory regions linked to a gene. To integrate multiple datasets, we use dPCA, which captures the directions of the greatest differential variance comparing two conditions, at each regulatory element (see section Materials and Methods) (Ji et al., 2013). As the goal of our method is to capture the differential signal at proximal and distal regulatory elements, we performed a dPCA analysis both at gene promoters and distal regulatory elements, which we call promoter interacting regions (PIRs) extracted from the 4DGenome database (Teng et al., 2015). Similar to standard PCA, differential PCA captures the directions of the greatest differential variance along several differential principal components (dPCs). We selected the first two dPCs, which appear to capture the differential signal both from activating and repressive epigenetic marks. The sum of the absolute values of dPC1 and dPC2 at each regulatory element was used as a score for this element. These scores are summarized as a weighted network relating regulatory elements to their target genes. The network consists of promoters and connected PIRs. Oriented edges from PIRs to promoters indicate a 3D interaction between these regulatory elements. Despite being in principle a bipartite graph (with nodes being either PIRs or promoters), we do not make a distinction between these two types of regulatory elements. A random walk based method then assigns a score to the corresponding gene. The output of the method is a ranked list of genes from the most to the least affected one, which incorporates both promoter and enhancer alterations. As a comparison, we also generated ranked lists based only on the promoter score (named promoter ranked lists in the following), discarding the contributions from distal PIR elements. This approach can be applied whenever two conditions are to be compared, for example, normal/tumor tissue, various tumor subtypes, or different developmental stages. More details are given in the Materials and Methods section. In order to benchmark our method, we used seven test cases consisting of tumor samples for seven different tumor types and normal matching samples. For each of these test cases, we compiled a list of CMGs (**Supplementary Table 2**) from the literature, and considered tissue-specific genes (TSGs) obtained from the ArchS4 database (Lachmann et al., 2018) as controls.

Cancer Marker Genes Are Scored Higher by Incorporating Enhancer in the Ranking

In our analysis, we determined that taking into account the first two dPCs is able to capture most of the differential variance for both activating and repressive epigenetic modifications (**Figures 2A,B**). After comparing the dPC1+dPC2 values between the CMGs and TSGs in each test case, we found that the scores from CMGs are generally higher than the scores of the TSGs for the enhancers, whereas the situation is less clear at promoters. This might indicate that most of the differential signal between tumor and normal occurs at distal regulatory regions. (**Figure 2C**).

Using the ranked gene lists generated by IRENE, we further computed the area under the curve (AUC) for the empirical cumulative density function (ECDF) of the high-confidence CMG ranks as a benchmarking approach, as described in the methods. First, we examined the IRENE ranks computed using the dPC1+dPC2 on gene promoters and their targeting enhancers, and found that the marker genes are ranked higher than TSGs in every test case, indicating that our approach captures the specific differential epigenetic signals at CMGs (**Figure 3A**). Moreover, both for CMGs and TSGs, the IRENE AUC values are higher than the AUC values computed using the

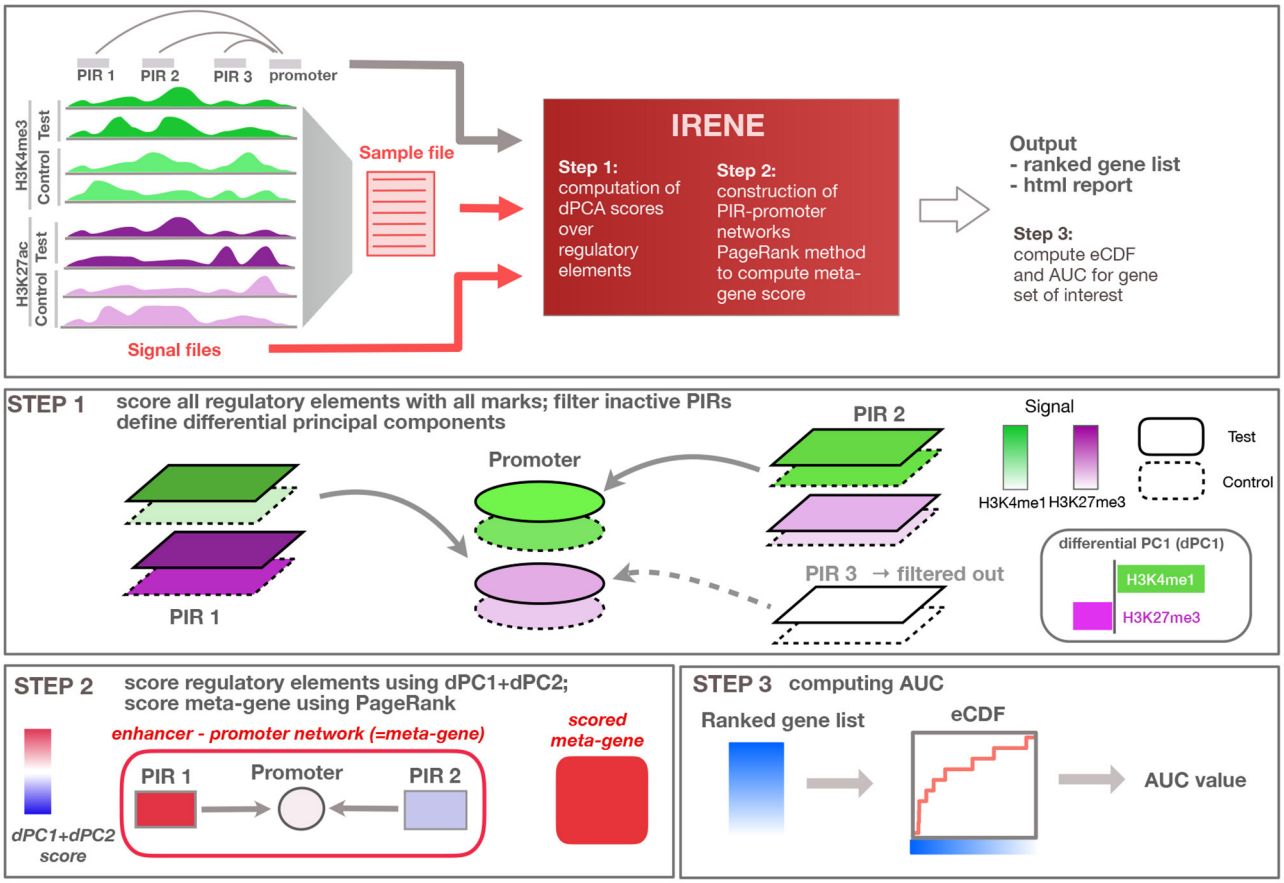


FIGURE 1 | Overview of the method. General overview of the input/output of the IRENE method; user defined input is indicated in red, while input provided by the method (e.g., regulatory loci and interactions) are displayed in gray. Step 1: Regulatory elements are scored using all epigenetic modifications, and related to the target gene. Step 2: Epigenetic alterations are scored using the first dPC and combined using PageRank into an integrated meta-gene score. Step 3: Ranked gene lists based on the score are converted to eCDF curves showing the enrichment of a given gene set within the top-ranked genes, and corresponding area under the curve (AUC) values are computed.

dPC1+dPC2 of gene promoters only (**Figure 3A**). The fact that the genes ranked higher in IRENE suggests that a significant part of the altered epigenetic alteration arises from distal enhancer regions. We then validated these findings on the larger CMG and TSG gene sets, and we found the AUCs of CMGs are all significantly higher (one-tailed *t*-test **p**-value<0.01) than the AUCs of TSGs (**Figure 3B**).

Some genes have a much high number of linked enhancers than others. To test whether this might bias the ranks of these genes, we performed 1,000 degree-preserving random perturbations, which completely rewired the enhancer–promoter graph but maintaining the degree distribution. We used the high-confidence CMGs in the benchmarking, and the AUCs with randomly assigned enhancers dropped 5–10% on average, indicating that the higher ranks of CMGs are not explained by their higher connectivity (**Figure 4**).

We compared the target gene assignment provided by the 4DGenome database, which is based on experimental evidence, with the simpler nearest-gene assignment. As can be observed in **Figure 4**, both approaches lead to comparable results, in line with recent reports indicating that the nearest gene assignment is reasonably effective in linking enhancers with target genes (Moore et al., 2020).

As mentioned in the Introduction, several other methods have been developed to integrate multiple epigenetic marks over genomic regions. Most of these methods provide qualitative analysis in the form of discrete chromatin states. To our knowledge, none of these methods apply a network-based integration as in IRENE to summarize regulatory elements related to the same gene. In order to provide a comparison, we focused on one of the mostly used such method, ChromHMM, which integrates various histone marks into discrete chromatin states (Ernst and Kellis, 2012). We combined ChromHMM with the Chromswitch method (Jessa and Kleinman, 2018), which computes a differential score between two groups of samples over specific regions. Applying this scoring approach to promoter regions, we compared the ranked lists obtained by IRENE at promoter regions with the ChromHMM-based ranks for the

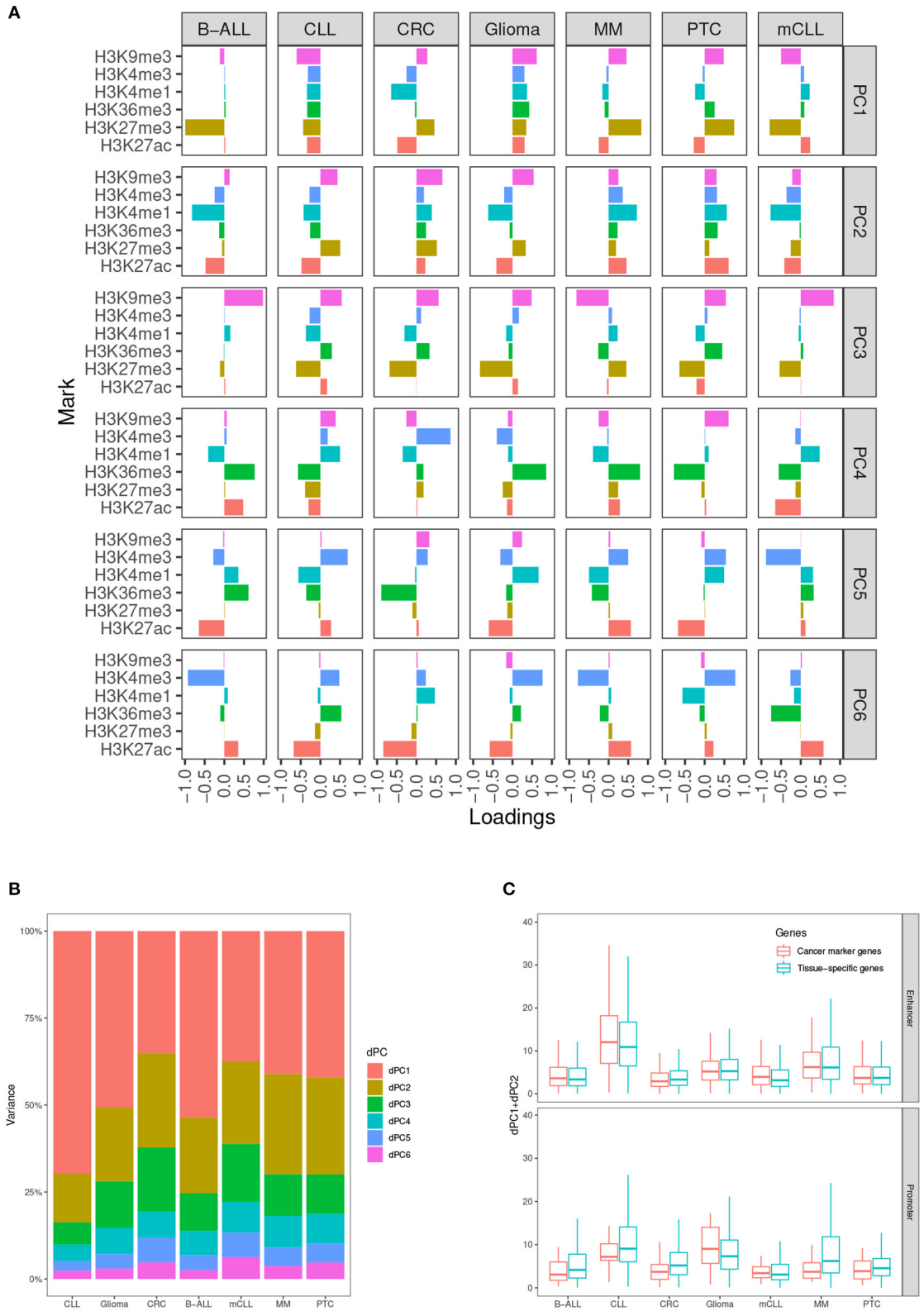


FIGURE 2 | Differential principal components. **(A)** Contributions of the six histone marks to each differential principal component (dPC). **(B)** Variances accounted for each dPC in the seven test cases. **(C)** Values of dPC1+dPC2 in the seven test cases, comparing cancer marker genes (CMGs) with TSGs, both for enhancers (top), and promoters (bottom).

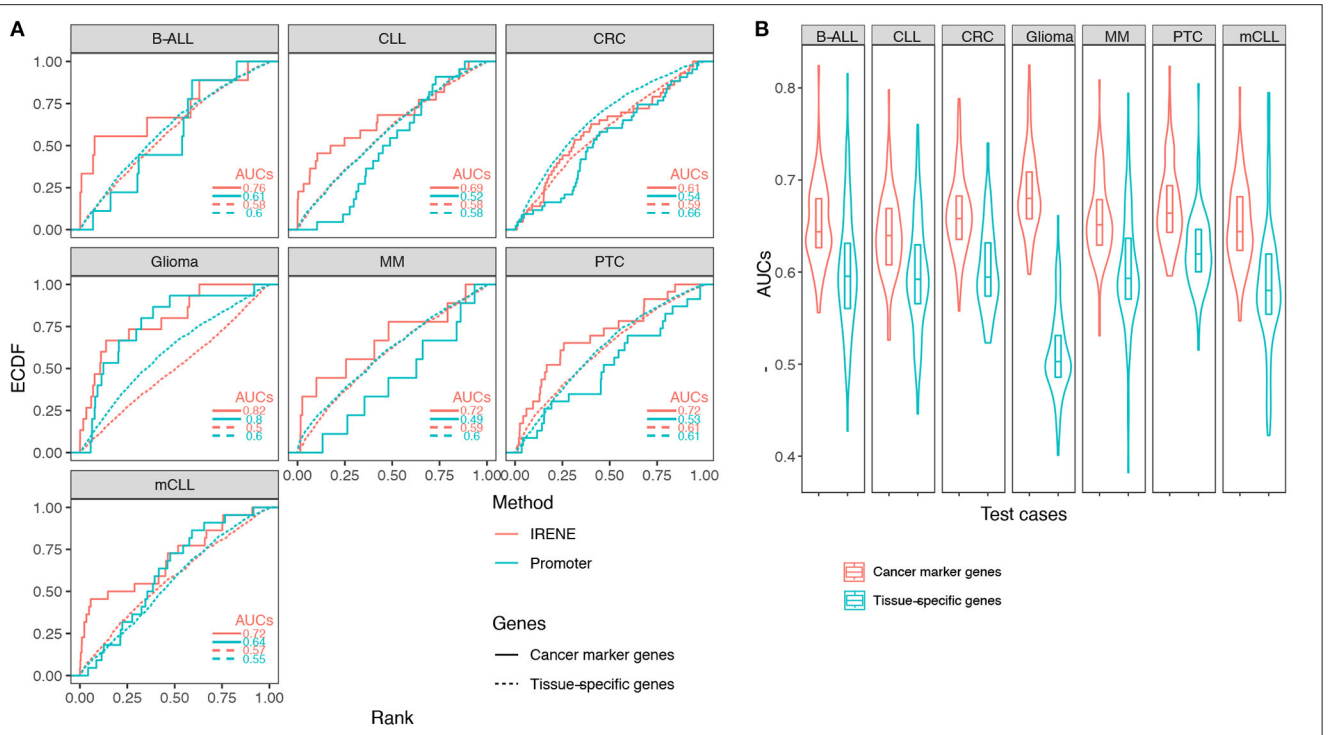


FIGURE 3 | (A) Empirical cumulative density function (ECDF) curves regarding the cancer marker genes and tissue-specific genes in seven test cases. The marker gene ranks using IRENE scores (red) are compared against their ranks using the promoter scores (cyan). **(B)** Distribution of the area under the curve (AUC) values using the cancer marker gene (CMG) sets from CancerMine in the seven test cases, compared to randomly picked tissue-specific genes to define equal size sets.

Glioma/normal brain test case, and found that the AUC values of the CMGs related to Glioma are significantly higher for the IRENE method (**Supplementary Figure 2**).

Network Analyses Characterized the Highly Ranked Genes in the IRENE and Promoter List

We downloaded 184 KEGG pathways in KGML format and loaded them as directed graphs using KEGGgraph (Zhang and Wiemann, 2009). Then we took the top 15% genes from the IRENE and promoter rank lists in each one of the seven test cases, and mapped the genes to the KEGG cancer signaling pathway (hsa05200). In total, the reference pathway contains 531 genes and 1989 interactions, and on average 208 of the 531 genes are found in the IRENE rank lists, while only 152 genes are found in the promoter rank lists. In addition, the IRENE-ranked genes differ from promoter-ranked genes in both in-degrees and out-degrees of the nodes (**Table 1**). As the IRENE nodes generally have higher in-degrees than out-degrees in the graph presentation of the reference pathway, implying the IRENE genes are more often targeted by the other regulatory genes on their enhancers as they harbor more differential enhancers. We further examined the glioma signaling pathway (hsa05214) and found 19 genes from the IRENE rank list and 10 genes from the promoter rank list in the glioma test case (**Figure 5**). One common gene, *EGFR*, is in both lists and has been reported to undergo tight control through epigenetic regulation on both promoters and enhancers (McInerney et al., 2000; Liu et al., 2015; Jameson et al., 2019). Moreover, nine genes are present only in the IRENE rank list, such as *CCND1*, which has been reported to be regulated by an estrogen-mediated enhancer (Eeckhoute et al., 2006). In conclusion, this analysis shows that the IRENE methods provide a ranked gene list, which is enriched for high-ranking, cancer-relevant genes.

DISCUSSION

From the above benchmarking on seven cancer test case studies, we showed that IRENE is a more comprehensive approach comparing to the current frequently used approaches such as separate ranking gene promoters and enhancers. This highlights the importance of epigenetic regulation through distant enhancer regions. Using IRENE, users cannot only discover the genes which show significantly epigenetic alterations on their promoters, but also the ones that are connected with strong epigenetic modifications on distal interacting enhancers, which facilitates the discovery of potential epigenetic marker genes. On the other hand, by interpreting the higher ranked genes mapped to the existing pathways, the user may also find the enhancers of interests from their differential epigenetic modifications. For example, we found the *PAX5* gene to have a significantly higher rank in the IRENE list compared to the promoter-only list in the two CLL case studies, which implies that *PAX5* is extensively regulated by enhancers. *PAX5* is a key

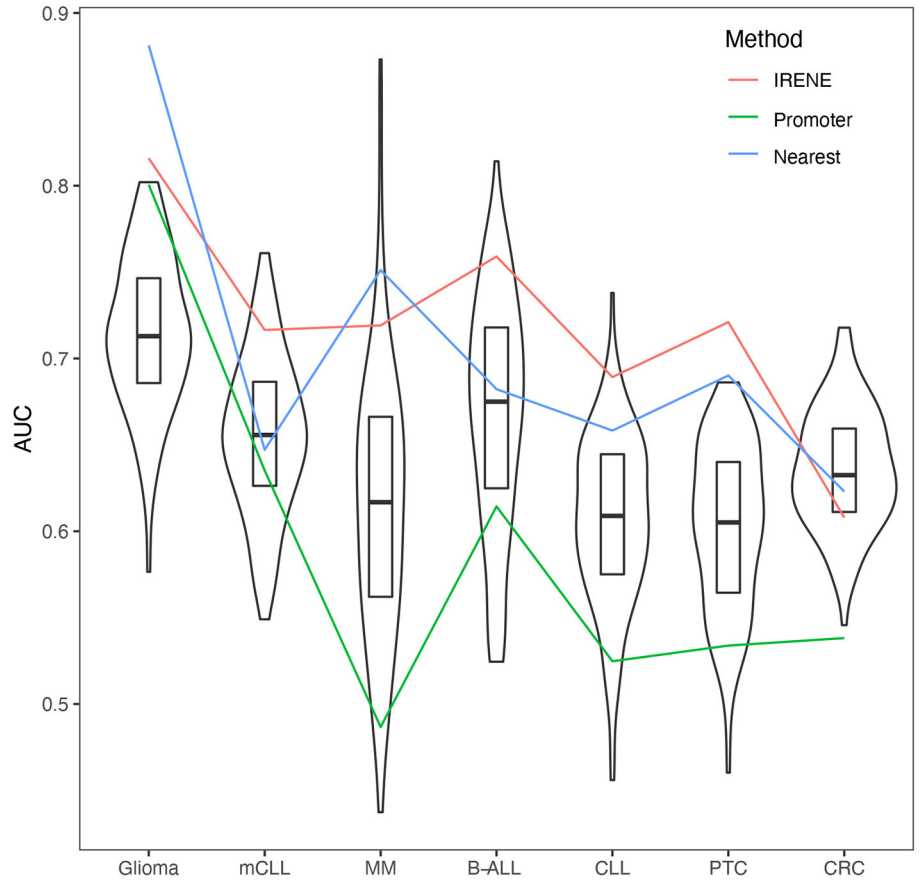


FIGURE 4 | Area under the curves (AUCs) of empirical cumulative density function (ECDF) curves of dPC1+dPC2 ranks from randomized promoter–enhancer interactions. The boxplots indicate the 25–75% quantile ranges from benchmarking each cancer marker gene set with 1,000 different rewired promoter–enhancer networks, whereas the red lines show the AUCs with the original promoter–enhancer interactions from IRENE using experimentally detected interactions (red), and interactions assigned by the nearest promoters (blue), and only promoters (green) rank lists.

TABLE 1 | Graph properties in respect of the nodes from the IRENE and promoter rank lists.

	Node number		Median in-degree		Median out-degree	
	IRENE	**Promoter**	**IRENE**	**Promoter**	**IRENE**	**Promoter**
CLL	214	167	2	2	1	3
Glioma	193	133	2	1	1	1
CRC	219	168	2	0	1	3
B-ALL	180	124	1	1	1	0
mCLL	211	168	2	0	1	3
MM	219	165	2	1	1	1
PTC	219	137	2	1	1	3

transcription factor in B-cell development, and its promoters have no significant epigenetic alterations in the CLL case studies. However, this gene is associated with several hyperacetylated and hypomethylated distal enhancers, one of which is located at 330 kilobases (kb) upstream of the *PAX5* TSS, and has been also found as extensively mutated in CLL (Puente et al., 2015) (**Figure 6**). The deletion of this enhancer resulted in a 40% reduction in the expression of *PAX5* expression and chromatin interaction of this enhancer and *PAX5* has been proven from chromosome conformation capture sequencing (4C-Seq) analysis (Puente et al., 2015). The main difficulty of this study is obtaining cell type specific enhancer–promoter interactions, as the high-resolution chromatin interaction map for the cancer cells is currently not available. We have tested two alternative approaches in this study, using either the experimentally validated chromatin interaction or distance-based interactions. The performance of the above two approaches are similar (**Figure 4**). We believe better performance can be achieved when cell type specific enhancer–promoter

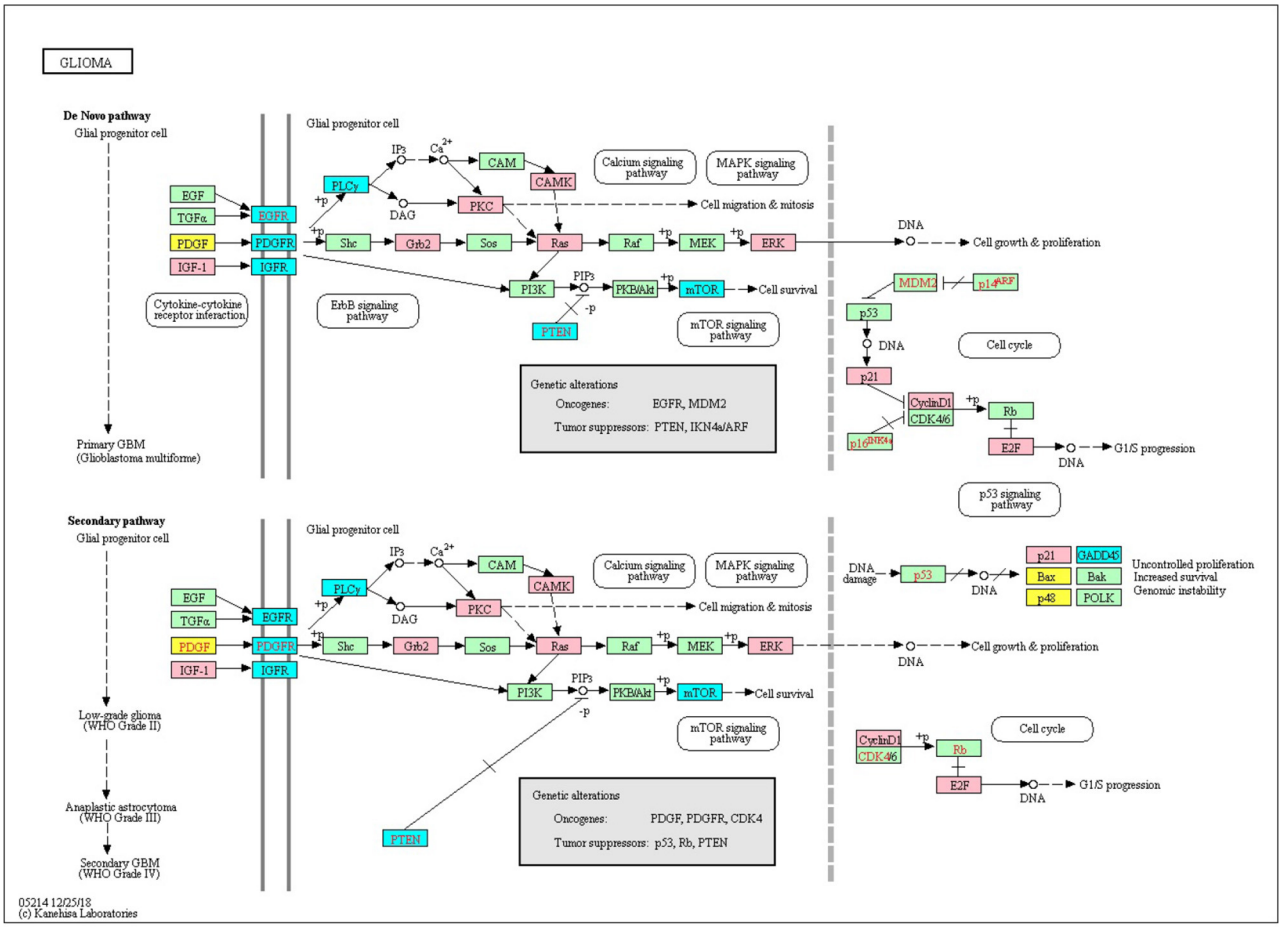


FIGURE 5 | The top 25% genes from the IRENE and promoter rank list are highlighted on the KEGG glioma signaling pathway. Pink, genes from the IRENE list; yellow, genes from the promoter list; cyan, genes from both lists.

interactions are available in the future, and using IRENE, user can replace the interaction map with a more specific one when applicable. Being a differential approach comparing two conditions, it might be affected by the possible heterogeneity of the groups being compared. If the heterogeneity is due to biological reasons (for example, different subtypes in the disease group), the comparison will be affected by the greater variance within one group. However, if the heterogeneity is of technical nature, then this noise will likely be buffered by the fact that our method integrates multiple regions to score the genes.

CONCLUSIONS

Genome-wide integrative epigenetic analysis is challenging and essential in many comparative studies. As far as we know, IRENE is the first tool that integrates quantitative and genome context information in the differential epigenetic analysis. Applying this tool to well-characterized test cases, it detects a number of candidate genes with significant epigenetic alterations, and comprehensive benchmarking validated these findings in cancer studies. As epigenomic datasets accumulate, the computational approaches employed in this study would be highly relevant in both comparative and integrative analysis of the epigenetic landscape. The discovery of novel epigenetic targets in cancers not only unfolds the fundamental mechanisms in tumorigenesis and development but also serves as an emerging resource for molecular diagnosis and treatment.

MATERIALS AND METHODS

Data Preparation

Retrieving Epigenetic Modification and Chromatin Interaction Datasets

Genome-wide ChIP-seq data are downloaded in BigWig format from NIH Roadmap Epigenomics (Bernstein et al., 2010), Blueprint (Adams et al., 2012), and the International Human Epigenome Consortium (IHEC) (Stunnenberg et al., 2016). We selected the six most frequently studied histone marks: H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3. These resources allow us to investigate the histone modification differences between tumor and normal tissues (**Supplementary Table 1**). For restricting the comparisons to

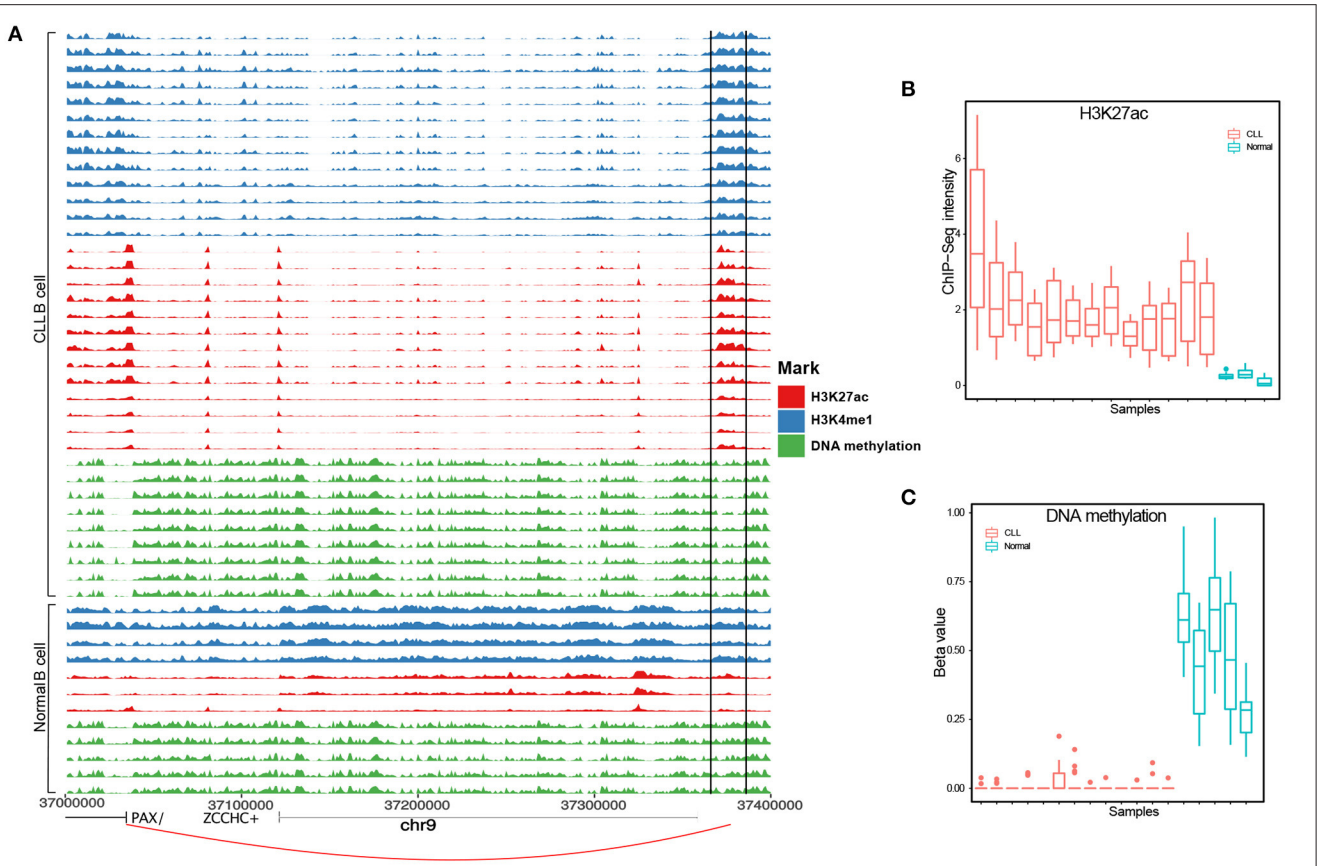


FIGURE 6 | A known PAX5 enhancer (chr9:37,370,916-37,371, and 635) in CLL exhibits hyperacetylation and hypomethylation [analysis of variance (ANOVA) *p*-value <0.05]. **(A)** PAX5 enhancer positions in each track are surrounded by two black solid lines. **(B)** H3K27ac and **(C)** DNA methylation levels between CLL and healthy samples.

the genomic loci of interests (promoters and enhancers), we downloaded the GRCh37 and GRCh38 coordinates of promoters from the eukaryotic promoter database (EPD) (Dreos et al., 2013), and the promoter interacting regions (PIRs) from the 4DGenome database (Teng et al., 2015). We treated the PIRs as potential enhancer regions, and filtered for tissue-specific enhancers by requiring the presence of H3K4me1 or H3K27ac peaks (peak calls provided in the **Supplementary Table 1**) in at least two samples from either tumor or normal tissues. By doing this, we enrich for cell type specific PIRs, which show a tissue-driven clustering (**Supplementary Figure 1**). The promoter coordinates were extended to ±1000 base pairs around the original coordinates. The sum of the numeric values from the BigWig blocks which overlap with the promoter and interacting regions are available from our project homepage. To build the relationships between and enhancers and promoters, we also download all the experimentally validated chromatin interaction datasets in various human tissues from 4DGenome.

Defining Disease and Control Datasets

We used histone modification datasets from seven cancer types in this study, i.e., B-ALL, CRC, glioma, MM, PTC, CLL, and mCLL from the Blueprint and IHEC consortia. For each cancer dataset, we paired it with the available dataset from the healthy tissue from which the cancer is most likely originated from. For example, the B-ALL, CLL, and MM were all compared against the healthy B cells in our design (see **Supplementary Table 1** for the pairs of normal/tumors used).

Definition of Cancer Marker Genes and Tissue-Specific Genes

We evaluated our algorithm on a small set of high-confidence CMGs, which is based on the tier-1 genes of the corresponding tissues from the Cancer Gene Consensus (CGC-t1) (Sondka et al., 2018) (**Supplementary Table 2**). As a negative control, we compiled a list of tissue-specific genes (TSGs) related to the tissues of interest for the tumor cases from ARCHS4_Tissues (https://maayanlab.cloud/archs4/). There are 2,318 genes for every tissue in the list. To validate our findings on independent, larger datasets of CMGs and TSGs, we compiled additional CMG lists containing 4,212 CMGs from 90 different cancer types from CancerMine (Lever et al., 2019), which incorporates the manual curated lists including the Cancer Gene Consensus (Sondka et al., 2018) and IntOGen (Gonzalez-Perez et al., 2013).

Data Processing Procedures

Combining Histone Marks

The epigenetic intensities on regulatory elements were summarized on a 1 kb scale, then power-transformed and

quantile normalized. We use the dPCA (Ji et al., 2013) to decompose the matrix D representing the difference between M epigenetic datasets at G genomic loci comparing two groups of samples, into matrices B and V (1)

$$D_{G\times M} = B_{G\times R} V_{R\times M} + E \quad (1)$$

where E is the random sampling noise.

We use the first k dPCs to represent the major changes between two conditions. We implemented an R wrapper function for dPCA in our tool, which takes the mean differences of the normalized ChIP-Seq signals in each genomic locus between two biological conditions as input, and returns the dPCs from dPCA. The definition of dPCs varies between the test cases (**Figure 2A**). The largest variances of the positive and negative histone mark components are captured by dPC1 and dPC2 in our test case studies (**Figure 2B**). Therefore, we selected the sum of the absolute values of the first two dPCs for representing the overall differences of these epigenetic marks.

Promoter–Enhancer Interaction Analyses

In our approach, the enhancer–promoter relationships are described as a weighted bipartite graph, in which both enhancers and promoters are represented as vertices, and edges are directed from enhancers to their target promoters (**Figure 1** Step 1). The weights of the vertices are defined as the sum of the absolute values of the first two dPCs when combining multiple epigenetic marks, or the absolute value of the difference if a single epigenetic mark is considered. We adopt an algorithm called "PageRank," which is originally designed for evaluating the importance of web pages (Brin and Page, 1998), for ranking the magnitude of epigenetic alterations in each gene. We use the "personalized" PageRank implemented in igraph (Rye et al., 2011) to summarize the weights of one promoter and its connected enhancers into a unique meta-gene score (**Figure 1** Step 2). Since our enhancer–promoter network is a directed graph, all the enhancer weights will eventually be attributed to their target promoter using PageRank, yielding a unified score for each gene, which can be used to rank the genes. Overall, there are $\sim$ 251,000 promoter interacting fragments in the promoter–enhancer interaction networks in our case studies, which is 8.5 times the number of promoters in the networks. The number of the interacting fragments targeting a gene varies from none to 227, and on average, 21 interacting fragments are targeting a promoter in the networks.

Scoring Ranked Lists

Using the gene ranks computed as described in the previous section, we can now evaluate the enrichment of a specific gene set $\mathcal{G}$ in the ranked list by computing the empirical cumulative distribution function (ECDF) obtained ranking the genes in decreasing order based on the previously described rank, and summing the indicator function

$$eCDF_{\mathcal{G}}(k) = \sum_{i=1}^{k} \delta_i \quad with \quad \delta_i = \begin{cases} 1 & if\ g_i \in \mathcal{G} \\ 0 & if\ g_i \notin \mathcal{G} \end{cases} \quad (2)$$

We use the area under the curve (AUC) as a measure of the enrichment of the gene set $\mathcal{G}$, with AUC = 0.5 corresponding to a random distribution of the genes in $\mathcal{G}$ inside the ranked list.

Comparison With ChromHMM

We applied the ChromHMM method (version v1.22) to the Glioma and the healthy brain control samples (see **Supplementary Table 1**). The 6 histone marks were integrated into 10 chromatin states, of which 2 correspond to active promoter regions and one to active enhancer regions (**Supplementary Figure 2B**). The chromswitch package (Jessa and Kleinman, 2018) (v. 1.12.0) from Bioconductor was applied to the promoter and PIR regions linked to promoters for specific chromatin states. The chromswitch method determines a consensus score between changes occurring in chromatin state within a group of sample, and the labels of these samples. Hence, a maximal consensus score for a region of interest would correspond to changes in a chromatin state within the region of interest occurring only in the samples of one of the two groups. A minimal consensus score would on the opposite correspond to changes in chromatin states in the region of interest occurring in samples, which are randomly distributed over the two groups. For each gene, we compute a score by averaging the consensus score of all regulatory elements related to this gene, and use this score to rank the genes, as a comparison to the IRENE ranking.

DATA AVAILABILITY STATEMENT

The R package is available at https://github.com/hdsu-bioquant/irene. The datasets generated for this study can be found in the https://github.com/hdsu-bioquant/ irene-data. We also designed a web interface that allows users to trace back the epigenetic alterations of every enhancer and promoter, as well as every sample which is used for computing the score. We use Rmarkdown to generate static HTML pages and created a web site for presenting the results from our test case studies, which can also be found under the project home page. Users may also take advantage of this function to create a report that highlights a few genes of their interests and share the studies with the audience.

AUTHOR CONTRIBUTIONS

CH designed and supervised this project. QW and CH drafted the manuscript. QW wrote and tested the software. YW and TV participated in software testing. YW, TV, and RE revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the BMBF-funded PRECISe project (#031L0076A).

ACKNOWLEDGMENTS

The manuscript is deposited in the bioRxiv preprint server as Wang et al. (2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.664654/full#supplementary-material

REFERENCES

Adams, D., Altucci, L., Antonarakis, S. E., Ballesteros, J., Beck, S., Bird, A., et al. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechol.* 30, 224–226. doi: 10.1038/nbt.2153

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., et al. (2010). The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.* 28, 1045–1048. doi: 10.1038/nbt1010-1045

Brin, S., and Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 107–117. doi: 10.1016/S0169-7552(98)00110-X

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: What, How, and Why? *Mol. Cell* 49, 825–837. doi: 10.1016/j.molcel.2013.01.038

Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Dalsgaard Sørensen, K., Liang, G., et al. (2016). Bivalent regions of cytosine methylation and H3K27 acetylation suggest an active role for DNA methylation at enhancers. *Mol. Cell* 62, 422–431.doi: 10.1016/j.molcel.2016.03.033

Chen, L., Wang, C., Qin, Z. S., and Wu, H. (2015). A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics* 31, 1889–1896. doi: 10.1093/bioinformatics/btv094

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proce. Natl. Acad. Sci. U.S.A.* 107, 21931–21936. doi: 10.1073/pnas.1016071107

Dreos, R., Ambrosini, G., Périer, R. C., and Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* D157–D164. doi: 10.1093/nar/gks1233

Eeckhoute, J., Carroll, J. S., Geistlinger, T. R., Torres-Arzayus, M. I., and Brown, M. (2006). A cell-type-specific transcriptional network required for estrogen regulation of cyclin D1 and cell cycle progression in breast cancer. *Genes Dev.* 20, 2513–2526. doi: 10.1101/gad.1446006

Ernst, J., and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215–216. doi: 10.1038/nmeth.1906

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., et al. (2013). IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10, 1081–1084. doi: 10.1038/nmeth.2642

Hartley, I., Elkhoury, F. F., Heon Shin, J., Xie, B., Gu, X., Gao, Y., et al. (2013). Long-lasting changes in DNA methylation following short-term hypoxic exposure in primary hippocampal neuronal cultures. *PLoS ONE* 8:e77859. doi: 10.1371/journal.pone.0077859

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39, 311–318. doi: 10.1038/ng1966

Jameson, N. M., Ma, J., Benitez, J., Izurieta, A., Han, J. Y., Mendez, R., et al. (2019). Intron 1-mediated regulation of EGFR expression in EGFR-dependent malignancies is mediated by AP-1 and BET proteins. *Mol. Cancer Res.* 17, 2208–2220. doi: 10.1158/1541-7786.MCR-19-0747

Jessa, S., and Kleinman, C. L. (2018). Chromswitch: a flexible method to detect chromatin state switches. *Bioinformatics* 34, 2286–2288. doi: 10.1093/bioinformatics/bty075

Ji, H., Li, X., Wang, Q.-f., and Ning, Y. (2013). Differential principal component analysis of ChIP-seq. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6789–6794. doi: 10.1073/pnas.1204398110

Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K., and Vingron, M. (2010). Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2926–2931. doi: 10.1073/pnas.0909344107

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., et al. (2018). Massive mining of publicly available rna-seq data from human and mouse. *Nat. Commun.* 9, 1–10. doi: 10.1038/s41467-018-03751-6

Lever, J., Zhao, E. Y., Grewal, J., Jones, M. R., and Jones, S. J. M. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat. Methods* 16, 505–507. doi: 10.1038/s41592-019-0422-y

Liang, K., and Keles, S. (2012). Detecting differential binding of transcription factors with ChIP-seq. *Bioinformatics* 28, 121–122. doi: 10.1093/bioinformatics/btr605

Liu, F., Hon, G. C., Villa, G. R., Turner, K. M., Ikegami, S., Yang, H., Ye, Z., et al. (2015). EGFR mutation promotes glioblastoma through epigenome and transcription factor network remodeling. *Mol. Cell* 60, 307–18. doi: 10.1016/j.molcel.2015.09.002

Lun, A. T. L., and Smyth, G. K. (2015). Csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.* 44, 1–10. doi: 10.1093/nar/gkv1191

Maston, G. A., Evans, S. K., and Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Ann. Rev. Genomics Hum. Genet.* 7, 29–59. doi: 10.1146/annurev.genom.7.080505.115623

McInerney, J. M., Wilson, M. A., Strand, K. J., and Chrysogelos, S. A. (2000). A strong intronic enhancer element of the EGFR gene is preferentially active in high EGFR expressing breast cancer cells. *J. Cell. Biochem.* 80, 538–549. doi: 10.1002/1097-4644(20010315)80:4<538::AID-JCB1008>3.0.CO;2-2

Moore, J. E., Pratt, H. E., Purcaro, M. J., and Weng, Z. (2020). A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* 21, 17. doi: 10.1186/s13059-019-1924-8

Nair, N. U., Das Sahu, A., Bucher, P., and Moret, B. M. (2012). Chipnorm: a statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PLoS ONE* 7:e39573. doi: 10.1371/journal.pone.0039573

Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., et al. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524. doi: 10.1038/nature1466

Rye, M. B., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res.* 39:e25. doi: 10.1093/nar/gkq1187

Shao, Z., Zhang, Y., Yuan, G.-C., Orkin, S. H., and Waxman, D. J. (2012). MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.* 13:R16. doi: 10.1186/gb-2012-13-3-r16

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., and Forbes, S. A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* 18, 696-705. doi: 10.1038/s41568-018-0060-1

Song, Q., and Smith, A. D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 27, 870–871. doi: 10.1093/bioinformatics/btr030

Stark, R., and Brown, G. (2011). *DiffBind: Differential Binding Analysis of ChIP-Seq Peak Data.* Bioconductor. Available online at: http://bioconductor.org/packages/release/bioc/html/DiffBind.html

Stasevich, T. J., Hayashi-Takanaka, Y., Sato, Y., Maehara, K., Ohkawa, Y., Sakata-Sogawa, K., et al. (2014). Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* 516, 272–275. doi: 10.1038/nature13714

Steinhauser, S., Kurzawa, N., Eils, R., and Herrmann, C. (2016). A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinforma.* 17, 953–966. doi: 10.1093/bib/bbv110

Stunnenberg, H. G., Consortium, T. I. H. E., Hirst, M., International Human Epigenome Consortium, and Hirst, M. (2016). The International Human Epigenome Consortium: a Blueprint for Scientific Collaboration and Discovery. *Cell* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007

Teng, L., He, B., Wang, J., and Tan, K. (2015). 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* 31, 2560–2564. doi: 10.1093/bioinformatics/btv158

Wang, Q., Wu, Y., Vorberg, T., Eils, R., and Herrmann, C. (2020). Integrative ranking of enhancer networks facilitates the discovery of epigenetic markers in cancer. *bioRxiv*. doi: 10.1101/2020.11.25.397844

Xu, H., Wei, C. L., Lin, F., and Sung, W. K. (2008). An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 24, 2344–2349. doi: 10.1093/bioinformatics/btn402

Zeng, X., Sanalkumar, R., Bresnick, E. H., Li, H., Chang, Q., and Keleş, S. (2013). jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.* 14:R38. doi: 10.1186/gb-2013-14-4-r38

Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* 21, 1273–1283. doi: 10.1101/gr.122382.111

Zhang, J. D., and Wiemann, S. (2009). KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics* 25, 1470–1471. doi: 10.1093/bioinformatics/btp167

Zhang, Y., An, L., Yue, F., and Hardison, R. C. (2016). Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.* 44, 6721–6731. doi: 10.1093/nar/gkw278

Zhu, Y., van Essen, D., and Saccani, S. (2012). Cell-type-specific control of enhancer activity by H3K9 trimethylation. *Mol. Cell* 46, 408–423. doi: 10.1016/j.molcel.2012.05.011

Ziller, M. J., Edri, R., Yaffe, Y., Donaghey, J., Pop, R., Mallard, W., et al. (2014). Dissecting neural differentiation regulatory networks through epigenetic footprinting. *Nature* 518, 355–359. doi: 10.1038/nature13990

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

CASE REPORT
published: 17 June 2021
doi: 10.3389/fgene.2021.638636

Case Report: Review of CT Findings and Histopathological Characteristics of Primary Liver Carcinosarcoma

Lu Huang[1] *and Lijian Lu*[2]*

[1] Department of Infectious Diseases, The First Affiliated Hospital of Guangxi Medical University, Nanning, China, [2] Department of Radiology, The Wuming Affiliated Hospital of Guangxi Medical University, Nanning, China

Objectives: The aim of the present study was to describe the computed tomography (CT) characteristics of primary liver carcinosarcoma (PLCS) and to explore the pathological basis for the diagnosis of primary hepatocellular carcinoma sarcoma.

Methods: Three male patients with PLCS were included in the present retrospective research, and the age was ranged from 52 to 63 years. The plain CT scan and third-stage enhancement scan were performed on patients. The pathological characteristics were analyzed. Stomachache was the main clinical symptoms of the three patients. Cirrhosis background was confirmed in one patients, and chronic Hepatitis B background was confirmed in other two patients.

Results: According to the results of CT, the inner diameter of the tumors ranged from 8.6 to 27.0 cm. The fibrous pseudocapsule around the tumor tissues was observed in two patients. Tumor tissues from all three patients were composed of sarcomatous and carcinomatous components. For carcinomatous components, hepatocellular carcinoma was observed in one patient and cholangiocarcinoma was observed in the other two patients. For sarcomatous components, angiosarcoma was observed in two patients and malignant fibrous histiocytoma was observed in another one patient. The tumor tissues were visualized as heterogeneous low density with large sheets of necrotic cystic lesions or thick-walled areas of multilocular cystic lesions using the plain CT scan. Edge-to-center filling and strengthening lesions, mild to moderate enhanced parenchyma at the arterial phase, and isodensity between the tumor parenchyma and the surrounding liver parenchyma at the portal vein phase or delayed phase were observed using the third-stage enhancement scan.

Conclusions: CT characteristics observed in the present study were of great benefit for the diagnosis of PLCS.

Keywords: liver, tumora, tomography, x-ray computed, diagnosis

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Sasi Arunachalam,
St. Jude Children's Research Hospital, United States
Jie Li,
Harbin Institute of Technology, China

***Correspondence:**
Lijian Lu
lulijiannet@163.com

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

Received: *07 December 2020*
Accepted: *29 April 2021*
Published: *17 June 2021*

Citation:
Huang L and Lu L (2021) Case Report: Review of CT Findings and Histopathological Characteristics of Primary Liver Carcinosarcoma. Front. Genet. 12:638636. doi: 10.3389/fgene.2021.638636

INTRODUCTION

Primary liver carcinosarcoma (PLCS) is defined as a malignant tumor concomitantly composed of a mixture of sarcomatous and carcinomatous by the World Health Organization (WHO), which is either hepatocyte-derived or cholangiocyte-derived or mixed. Currently, the pathological mechanism underlying PLCS is unclear (Xiang et al., 2015). PLCS is a type of rare and complex hepatic malignant tumor with aggressive growth characteristics, propensity for recurrence, and a poor prognosis (Li et al., 2016). Preoperative diagnosis of PLCS is typically challenging, which relies on the postoperative pathological examination (Shu et al., 2010). Both epithelial and mesenchymal sarcoma components can be observed on PLCS tumor tissues using a microscope, and immunohistochemical assay plays an important role for the further diagnosis of PLCS (Lao et al., 2007). In the present case-series report, three patients diagnosed with PLCS using surgical pathology in our hospital were included. The purpose of the present study was to describe the clinical, histopathological, and imaging characteristics of PLCS and to document the associated imaging presentations and results.

MATERIALS AND METHODS

The present retrospective research was authorized by the institutional research ethics committee of The First Affiliated Hospital of Guangxi Medical University. Informed consent was not applicable. The image data from all the three patients diagnosed with PLCS from January 2011 and February 2018 were analyzed. Two pathologists confirmed the pathological diagnosis of the cases. The medical records were consulted to determine the clinical manifestation, treatment, and outcome of the cases.

Three patients underwent the plain CT scan and third-stage enhancement scan (64 MDCT TK LIGHT SPEED GE Medical System). The scanning parameters were shown as the following: slice thickness: 5 mm; pitch: 1.375; bed speed: 5.5 mm/s; tube voltage: 120 kV; and tube current: 100 mA. Multi-planar recombination (MPR) was used for post-processing of images. Enhanced scanning was performed using a high-pressure syringe. The contrast agent administered was iopromide (includes 300 mg/mL of iodine) for a total of 70–85 mL with a flow rate of 3 mL/s.

Imaging results were reviewed independently by two abdominal imaging radiologists with 15 and 16 years of working experience, which were cross-checked by another radiologist to obtain the consistent conclusion. In the present study, the characteristics of the results of CT scans on tumors were evaluated, including position, size, relationship with hepatic envelope, edge, uniformity of density, and presence of adipose tissue, hemorrhage, cystic components, calcification, and vascular tumor.

RESULTS

Clinical Characteristics

Three male patients (52–63 years old) with PLCS who were treated at our hospital between January 2011 and February 2018 were included in the present study. All three patients were admitted to the hospital due to abdominal pain and a space-occupying lesion in the liver tissues. Two patients had a history of chronic hepatitis B, and one patient had a history of cirrhosis. A significant elevated level of carcinoembryonic antigen 199 (CA199) was observed in two patients, and an elevated level of alpha-fetoprotein (AFP) was observed in another one patients. All three patients had normal levels of carcinoembryonic antigen (CEA) (**Table 1**).

Pathological Characteristics

Two experienced abdominal pathologists individually analyzed the pathological data, which were cross-checked by another experienced abdominal pathologist. The maximum diameter of the lesion was ranged from 8.6 to 27.0 cm. The tangent plane of the lesions from all three patients was grayish white. A pseudo-envelope of fibrous tissue around the tumor was observed in two of the patients.

PLCS Consisted of Cancerous and Sarcoma Components

Tumor tissues composed of both cancerous and sarcomatous components interspersed with each other were observed in all three patients. For carcinomatous components, hepatocellular carcinoma was observed in one patient and cholangiocarcinoma was observed in the other two patients. For sarcomatous components, angiosarcoma was observed in two patients and malignant fibrous histiocytoma was observed in another one patient **Figures 1F**, **2D**, **3D**. Immunohistochemical results were shown as follows: Hep-1 (+) (one patient), AFP (+) (one patient), CK (+) (one patient), CK19 (+) (two patients), Vim (+) (two patients), CD34 (+) (two patients), and CD68 (+) (two patient), which were consistent with the diagnosis of PLCS **Table 2**.

CT Imaging Findings

Two experienced abdominal radiologists independently analyzed the imaging data, which were cross-checked by another experienced abdominal pathologist. All three patients had a

TABLE 1 | Clinical features of three patients with PLCS.

Case	age (years)	Sex	Main clinical symptoms	Liver disease	CA125	CA199	CEA	AFP	Prothrombin
1	52	male	Upper abdominal pain	Chronic hepatitis B	-	-	-	-	-
2	42	male	Right upper quadrant pain	Cirrhosis	-	+	-	+	+
3	63	male	Right upper quadrant pain	Chronic hepatitis B	-	+	-	-	-

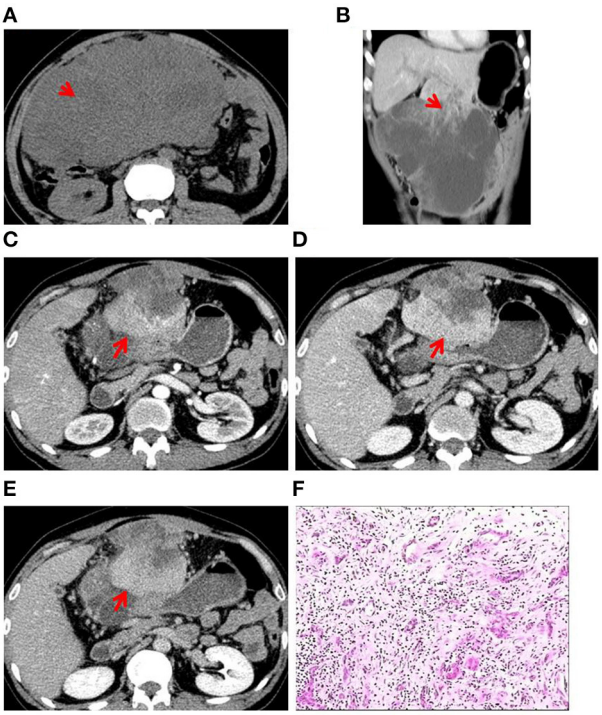


FIGURE 1 | Case 1: A 52-year-old man presented with upper abdominal pain since 3 months. **(A)** CT scan showed heterogeneously hypo-dense lesions (red arrow). **(B)** A large exogenous mass in the S3 segment of the liver was observed with irregular thick-walled areas inside the lesion (red arrow). **(C)** Contrast-enhanced CT in the arterial phase showed uneven mild to moderate enhancement of the lesion with intratumoral vessels and that the tumor has invaded the anterior abdominal wall (red arrow). **(D)** CT scan in the portal-vein phase showed gradual filling of the lesion from the edge to the center (red arrow). **(E)** CT scan in the prolongation phase showed gradual filling of the lesion from the edge to the center (red arrow). **(F)** Histopathological examination of surgical specimen: the carcinomatous component (cholangiocarcinoma) is interspersed with the sarcomatous component (angiosarcoma).

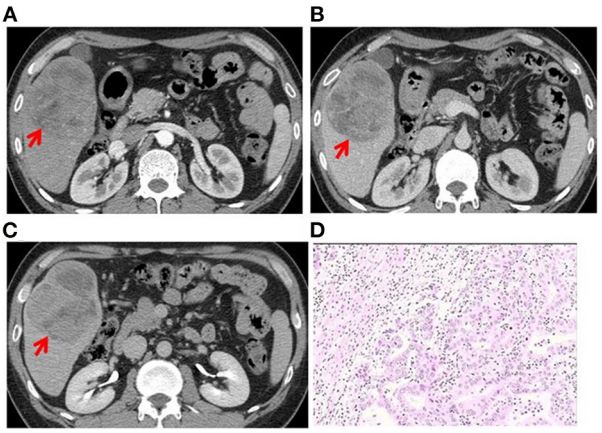


FIGURE 2 | Case 2: A 42-year-old man had abdominal pain in the right upper quadrant since 1 month. **(A)** A round mass in the S5 segment of the liver was observed. CT scan in the arterial phase showed uneven, mild to moderate enhancement (red arrow). **(B)** CT scan in the portal-vein phase showed gradual filling of the lesion from the edge to the center, with irregular thick-walled sac-variable regions (red arrow). **(C)** CT scan in the prolongation phase showed gradual filling of the lesion from the edge to the center, with irregular thick-walled sac-variable regions (red arrow). The tumor was surrounded by a pseudocapsule. **(D)** Histopathological examination of the surgical specimen shows PLCS. The carcinomatous component (cholangiocarcinoma) was interspersed with the sarcomatous component (angiosarcoma).

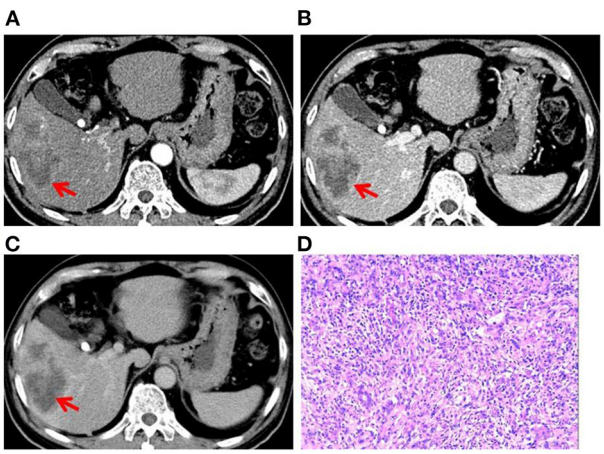


FIGURE 3 | Case 3: A 63-year-old man presented with pain in the right upper quadrant since 1 month. **(A)** A round mass in the S5 segment of the liver. CT scan in the arterial phase showed uneven, mild to moderate enhancement of the lesion. Tumor vessels were visible in the tumor. Irregular thick-walled sac-variable regions (red arrows) were seen in the lesion. **(B)** CT scan in the portal-vein phase showed gradual filling of the lesion from the edge to the center, with irregular thick-walled sac-variable regions (red arrow). **(C)** CT scan in the prolongation phase showed gradual filling of the lesion from the edge to the center, with irregular thick-walled sac-variable regions (red arrow). The tumor was surrounded by a pseudocapsule. **(D)** Histopathological examination of surgical specimen showing PLCS. Cancerous tissue (cholangiocarcinoma) was interspersed with sarcomatous tissue (malignant fibrous tissue).

single lesion in the liver, and CT scan showed an uneven and low-density zone (**Figure 1A**). An irregular and exogenous shaped tumor lesion was found to be located in the left lobe of the liver of case 1 (**Figure 1B**), while a pseudo-envelope with a clear boundary was formed around the tumor lesion in case 2 and case 2 (**Figures 2A,B**, **3A,B**). The tumor boundary was blurred, and there was no pseudo-envelope formation in case 1. Moreover, all three lesions showed mixed density and irregular thick-walled separation changes in the cystic zone (**Figures 1B**, **2A–C**, **3A–C**). No sign of calcification or intratumoral bleeding was observed in any of the patients.

According to the results of enhanced CT scan, the tumor margins in lesions from all patients were gradually filled and intensified toward the center (**Figures 1C–E**, **2A–C**, **3A–C**). Uneven and mildly enhanced tumor parenchyma and enriched tortuous tumor vessels were observed in the arterial phase (**Figures 1C**, **2A**, **3A**). Isodensity with hepatic parenchyma was observed in the portal vein and lag phase (**Figures 1D,E**, **2B,C**, **3B,C**). Invasion into the left branch of the portal vein and established tumor thrombus were observed in case 1, in which one lesion broke through the hepatic liver capsule into adjacent tissues. Tumor recurrence and distant metastasis

TABLE 2 | Pathological features of three cases with PLCS.

Item	Case 1	Case 2	Case 3
Pathology	One large liver mass, inner diameter: 27 cm, grayish cut surface, large necrotic area, and old bleeding	One large liver mass, inner diameter: 15 cm, grayish cut surface, some areas accompanied by hemorrhage and necrosis, and fibrous tissue wrapping around the tumor	One large liver mass, inner diameter: 8.6 cm, grayish cut surface, large necrotic area and old bleeding, and fibrous tissue wrapping around the tumor
Microscopy	The cancer tissue and the sarcoma tissue arranged in a mixed manner; cholangiocarcinoma in the cancer tissue, and angiosarcoma in the sarcoma. Immunohistochemistry CK19, Vim, CD34 (+)	The cancer tissue and the sarcoma tissue arranged in a mixed manner; hepatocellular carcinoma in the cancer tissue and angiosarcoma in the sarcoma. The sarcoma is an angiosarcoma, and the fibrous tissue is surrounded around the tumor. Immunohistochemistry AFP, Hep-, CK, CD34(+)	The cancer tissue and the sarcoma tissue arranged in a mixed manner; cholangiocarcinoma in the cancer tissue; malignant fibrous tissue tumor in the sarcoma; Immunohistochemistry CK19,CD68/34, Vim(+)

were observed in case 2 and case 2 within 3 months after operation (**Table 3**).

DISCUSSION

In 1989, Craig et al. proposed the definition of PLCS, which refers to primary liver malignant tumor containing both hepatocellular carcinoma and sarcoma. Subsequently, PLCS is further defined by the World Health Organization as a complex malignant liver tumor composed of a mixture of hepatocellular carcinoma or cholangiocarcinoma components and sarcoma components (Seifert et al., 1990). PLCS is a rare malignant tumor with rare reports (Celikbilek et al., 2011; Liu et al., 2012; Yamamoto et al., 2014; Xiang et al., 2015; Yu, 2015; Li et al., 2018a), and the specific clinical symptoms of PLCS are uncertain. Abdominal pain and abdominal distension are regarded as the main complaints of PLCS. Approximately 80% of PLCS patients possess a history of chronic liver disease, and a significantly elevated serum alpha fetoprotein (AFP) level is observed in about 27.6% PLCS patients (Li et al., 2018a,b). In the present study, all three patients were middle-aged men with a history of chronic liver disease, which suggests that middle-aged men and chronic liver disease might be risk factors for PLCS. Among the three patients, the serum CA199 level was increased in two patients, while the serum AFP level was increased in one patient. All three patients were CEA-negative, and abnormal prothrombin level was found in one patient. These observations might be associated with the number and type of tumor cell components, which was similar to those previously reported (Li et al., 2018b). Lung and lymph nodes, peritoneum, gallbladder, omentum, stomach, diaphragm, and adrenal gland are common metastatic positions. These clinical features indicate that PLCS has high levels of aggression and is metastatic (Celikbilek et al., 2011; Yasutake et al., 2014; Gu et al., 2015; Xiang et al., 2015).

The pathogenesis of PLCS is unclear. Current evidence (Lao et al., 2007; Celikbilek et al., 2011; Yasutake et al., 2014; Gu et al., 2015) supports the theory that carcinosarcoma is monoclonal in origin. In previous studies, most PLCSs were developed in normal livers with no cirrhosis background, which indicated that tumors develop from pluripotent liver progenitor cells or stem cells. The imaging characteristics of PLCS are currently unclear due to its low incidence, which makes it difficult for radiologists to make accurate preoperative imaging diagnosis. In the present study, all three patients were misdiagnosed preoperatively as hepatocellular carcinoma. The PLCS tumor was huge, irregularly shaped, and with unclear boundaries, which was consistent with the reports described previously (Lin et al., 2013; Gu et al., 2015; Xiang et al., 2015).

Computed tomography (CT) is the most commonly used imaging method for PLCS. However, currently few reports have described the CT findings of PLCS. Previous reports have described liver cancer sarcoma as generally large and irregular low-density masses, which tends to grow across the liver segment. The boundary of tumor is blurred, and the tumor directly invades into the surrounding tissues. Necrotic cystic degeneration is commonly observed in the central part of PLCS tumor tissues. Mild to moderate intensity is reported on PLCS using enhanced CT scan (Celikbilek et al., 2011; Liu et al., 2012; Xiang et al., 2015). In the present study, the size of PLCS tumor in all three patients was relatively large, which was irregular in one patient and nearly round in the other two patients. In one patient, the tumor had broken through the liver capsule and invaded into the surrounding tissues, which were supposed to be related to the high degree of malignancy and rapid growth of liver cancer sarcoma. These observations were consistent with previous reports, in which the pseudocapsule was rarely formed in hepatocarcinoma sarcoma (Celikbilek et al., 2011; Liu et al., 2012; Xiang et al., 2015; Li et al., 2018a,b). However, in the present study, the fibrous pseudocapsule was found in two patients, which might be related to massive proliferation of liver parenchymal fibrous tissue around the tumor induced by chronic liver diseases. In all three cases, irregular thick-walled multi-segmental cystic changes were observed, which might be related to the degree of necrosis in the lesion. Moreover, in all three cases, the tumors were gradually filled and enhanced from the margin to the center in the third-stage enhancement scan. Unevenness and mild-to-moderate enhancement were observed in the arterial phase, with several distorted tumor vessels. The parenchyma density of PLCS tumor was slightly higher than that of the adjacent liver parenchyma. The parenchymal enhancement in the portal vein or delayed phase showed an equal density change. These CT imaging characteristics have not been reported in previous literature (Celikbilek et al., 2011; Liu et al., 2012; Lin et al., 2013; Yamamoto et al., 2014; Xiang et al., 2015; Li et al., 2018a,b), which indicated that the isodense area in the portal vein or delayed phase of the tumor might be related to

TABLE 3 | CT features of three cases with PLCS.

Item	Location	Number of lesions	Morphology	Edge	Density	Calcification	Hemorrhage	Thick-walled sac change	Enhanced features	Pseudo capsule	Vascular invasion	Invasion of adjacent tissue	Bile duct invasion	Ascites	Metastasis	Recurrence
Case 1	S3/external	Single	Irregular	Blurry	Uneven	-	-	+	The edge is gradually filled and enhanced toward the center, and iso-density area is visible in the portal-vein or delayed phase.	-	+	+	-	-	Left abdominal mass/2 months	2 months
Case 2	S5	Single	Round	Clear	Uneven	-	-	+	The edge is gradually filled and enhanced toward the center, and iso-density area is visible in the portal-vein or delayed phase.	+	-	-	-	-	Pancreatic head mass/11 months	-
Case 3	S5	Single	Irregular	Clear	Uneven	-	-	+	The edge is gradually filled and enhanced toward the center, and iso-density area is visible in the portal-vein or delayed phase.	+	-	-	-	-	-	2 months

the abundant fibrous components or vascular components in the tumor parenchyma.

As described previously, calcification and bone tissue are observed in some tumors that contain the components of chondrosarcoma and osteosarcoma, which are suggested to be important CT signs for the diagnosis of PLCS (Lai et al., 2011). However, no signs of calcification or bone tissue were observed in any of the three patients in the present study, as chondrosarcoma and osteosarcoma were not included in the sarcomatous components of the tumors. In addition, tumor recurrence and distant metastasis were observed in two patients, indicating a poor prognosis of patients with PLCS.

Currently, the diagnosis of PLCS mainly depends on pathological results. As it is difficult to distinguish PLCS with other liver malignancies, such as hepatocellular carcinoma and cholangiocarcinoma, the imaging diagnosis for PLCS is difficult. Hepatocellular carcinoma is the most common primary malignancy of the liver, which is generally derived from chronic liver disease and commonly diagnosed in the elderly population (McEvoy et al., 2013). In the CT images of hepatocellular carcinoma, a low-density mass, varying in size, and significant enhancement are regularly presented, accompanied by satellite lesions and portal vein thrombosis. Capsules on the margin were commonly observed in well-differentiated hepatocellular carcinoma. Cholangiocarcinoma occurs in the bile duct epithelium and is usually located in the left hepatic lobe. Cholangiocarcinoma is found mostly in older men with a cirrhosis background. Typical imaging features of cholangiocarcinoma include more homogeneous low-density lesions, irregular appearance, gradual centripetal enhancement, contraction of adjacent hepatic envelope, and peripheral bile duct dilatation (Lewis et al., 2010). Compared to PLCS, less extensive necrosis, cystic degeneration, or isodensity changes were observed in hepatocellular carcinoma and cholangiocarcinoma.

Shortcomings of the Present Study

The number of cases included in the present retrospective analysis is small. The results obtained in the present study need to be further verified by more cases. In the present study, based on data collected from 2011 to 2018, the conditions of tumor recurrence and distant metastasis of patients were recorded. However, how tumor CT characteristics evolved over time was not explored yet, which will be explored in more cases in our future work.

CONCLUSIONS

Specific CT characteristics, such as huge tumor size, large-scale cystic and necrotizing degeneration, edge-to-center filling enhancement in the enhanced CT scan, and isodensity between the tumor parenchyma and the surrounding liver parenchyma at the portal vein phase or delayed phase, may help to distinguish PLCS from other malignancies. PLCS needs to be treated by surgical resection and careful CT follow-up due to their invasiveness and poor prognosis.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Wuming Hospital Affiliated to Guangxi Medical University of china. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article. All data published here are under the consent for publication.

AUTHOR CONTRIBUTIONS

LL designed/performed most of the investigation and data analysis. LH wrote the manuscript and provided clinical assistance. Both of the authors have read and approved the manuscript.

REFERENCES

Celikbilek, M., Deniz, K., Torun, E., Artis, T., Ozaslan, E., Karahan, O. I., et al. (2011). Primary hepatic carcinosarcoma. *Hepatobiliary Pancreat. Dis. Int.* 10, 101–103. doi: 10.1016/S1499-3872(11)60015-5

Gu, Y. J., Zhu, Y. Y., Lu, X. Y., Zhao, Q., and Cong, W. M. (2015). Hepatic carcinosarcoma: evidence of polyclonal origin based on microsatellite analysis. *Pathol. Res. Pract.* 211, 905–910. doi: 10.1016/j.prp.2015.09.007

Lai, Q., Levi Sandri, G. B., Melandro, F., Di Laudo, M., Garofalo, M., Guglielmo, N., et al. (2011). An unusual case of hepatic carcinosarcoma. *G. Chir.* 32, 372–373.

Lao, X. M., Chen, D. Y., Zhang, Y. Q., Xiang, J., Guo, R. P., Lin, X. J., et al. (2007). Primary carcinosarcoma of the liver: clinicopathologic features of 5 cases and a review of the literature. *Am. J. Surg. Pathol.* 31, 817–826. doi: 10.1097/01.pas.0000213431.07116.e0

Lewis, R. B., Lattin, G. E. Jr., Makhlouf, H. R., and Levy, A. D. (2010). Tumors of the liver and intrahepatic bile ducts: radiologic-pathologic correlation. *Magn. Reson. Imaging Clin. N. A.* 18, 587–609. doi: 10.1016/j.mric.2010.08.010

Li, B., Zhang, Y., Hou, J., Yu, H., and Shi, H. (2016). Primary liver carcinosarcoma and 18F-FDG PET/CT. *Clin. Nuclear Med.* 41, e383–385. doi: 10.1097/RLU.0000000000001232

Li, J., Liang, P., Zhang, D., Liu, J., Zhang, H., Qu, J., et al. (2018a). Primary carcinosarcoma of the liver: imaging features and clinical findings in six cases and a review of the literature. *Cancer Imaging Off. Publ. Int. Cancer Imaging Soci.* 18:17. doi: 10.1186/s40644-018-0141-0

Li, Z., Wu, X., Bi, X., Zhang, Y., Huang, Z., Lu, H., et al. (2018b). Clinicopathological features and surgical outcomes of four rare subtypes of primary liver carcinoma. *Chinese J. Cancer Res.* 30, 364–372. doi: 10.21147/j.issn.1000-9604.2018.03.08

Lin, Y. S., Wang, T. Y., Lin, J. C., Wang, H. Y., Chou, K. F., Shih, S. C., et al. (2013). Hepatic carcinosarcoma: clinicopathologic features and a review

of the literature. *Ann. Hepatol.* 12, 495–500. doi: 10.1016/S1665-2681(19)31015-4

Liu, Q. Y., Lin, X. F., Li, H. G., Gao, M., and Zhang, W. D. (2012). Tumors with macroscopic bile duct thrombi in non-HCC patients: dynamic multi-phase MSCT findings. *World J. Gastroenterol.* 18, 1273–1278. doi: 10.3748/wjg.v18.i11.1273

McEvoy, S. H., McCarthy, C. J., Lavelle, L. P., Moran, D. E., Cantwell, C. P., Skehan, S. J., et al. (2013). Hepatocellular carcinoma: illustrated guide to systematic radiologic diagnosis and staging according to guidelines of the American Association for the Study of Liver Diseases. *Radiographics* 33, 1653–1668. doi: 10.1148/rg.336125104

Seifert, G., Brocheriou, C., Cardesa, A., and Eveson, J. W. (1990). WHO international histological classification of tumours. tentative histological classification of salivary gland tumours. *Pathol. Res. Pract.* 186, 555–581. doi: 10.1016/S0344-0338(11)80220-7

Shu, R. Y., Ye, M., and Yu, W. Y. (2010). A case of primary liver carcinosarcoma: CT findings. *Chin. J. Cancer.* 29, 346–348. doi: 10.5732/cjc.009.10473

Xiang, S., Chen, Y. F., Guan, Y., and Chen, X. P. (2015). Primary combined hepatocellular-cholangiocellular sarcoma: an unusual case. *World J. Gastroenterol.* 21, 7335–7342. doi: 10.3748/wjg.v21.i23.7335

Yamamoto, T., Kurashima, Y., Ohata, K., Hashiba, R., Tanaka, S., Uenishi, T., et al. (2014). Carcinosarcoma of the liver: report of a case. *Surgery Today* 44, 1161–1170. doi: 10.1007/s00595-013-0612-7

Yasutake, T., Kiryu, S., Akai, H., Watadani, T., Akahane, M., Tomizawa, N., et al. (2014). MR imaging of carcinosarcoma of the liver using Gd-EOB-DTPA. *Magn. Reson. Med.* 13, 117–121. doi: 10.2463/mrms.2013-0011

Yu, G. Y. Y. (2015). Primary liver cancer sarcoma: a case report. *J. Clin. Hepatobiliary Dis.* 31, 1899–1901.

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 Molecular Biosciences

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fmolb.2021.675193

A Novel Ferroptosis-Related Biomarker Signature to Predict Overall Survival of Esophageal Squamous Cell Carcinoma

Jiahang Song[1,2†], *Yanhu Liu*[1†], *Xiang Guan*[1†], *Xun Zhang*[1], *Wenda Yu*[1*] *and Qingguo Li*[1,3*]

[1]Cardiovascular Center, The Second Affiliated Hospital of Nanjing Medical University, Nanjing, China, [2]Department of Radiation Oncology, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China, [3]Department of Cardiovascular Surgery, The Affiliated Hospital of Qinghai University, Xining, China

Esophageal squamous cell carcinoma (ESCC) accounts for the main esophageal cancer (ESCA) type, which is also associated with the greatest malignant grade and low survival rates worldwide. Ferroptosis is recently discovered as a kind of programmed cell death, which is indicated in various reports to be involved in the regulation of tumor biological behaviors. This work focused on the comprehensive evaluation of the association between ferroptosis-related gene (FRG) expression profiles and prognosis in ESCC patients based on The *Cancer* Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). ALOX12, ALOX12B, ANGPTL7, DRD4, MAPK9, SLC38A1, and ZNF419 were selected to develop a novel ferroptosis-related gene signature for GEO and TCGA cohorts. The prognostic risk model exactly classified patients who had diverse survival outcomes. In addition, this study identified the ferroptosis-related signature as a factor to independently predict the risk of ESCC. Thereafter, we also constructed the prognosis nomogram by incorporating clinical factors and risk score, and the calibration plots illustrated good prognostic performance. Moreover, the association of the risk score with immune checkpoints was observed. Collectively, the proposed ferroptosis-related gene signature in our study is effective and has a potential clinical application to predict the prognosis of ESCC.

Keywords: esophageal squamous cell carcinoma, ferroptosis, prognosis, gene signature, TCGA, GEO

OPEN ACCESS

Edited by:
Xiangqian Guo,
Henan University, China

Reviewed by:
Imtaiyaz Hassan,
Jamia Millia Islamia, India
Yinan Jiang,
University of Pittsburgh, United States

***Correspondence:**
Qingguo Li
Liqg@njmu.edu.cn
Wenda Yu
clover0610@hotmail.com

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

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *02 March 2021*
Accepted: *11 June 2021*
Published: *05 July 2021*

Citation:
Song J, Liu Y, Guan X, Zhang X, Yu W and Li Q (2021) A Novel Ferroptosis-Related Biomarker Signature to Predict Overall Survival of Esophageal Squamous Cell Carcinoma. Front. Mol. Biosci. 8:675193. doi: 10.3389/fmolb.2021.675193

INTRODUCTION

Esophageal cancer (ESCA), a global malignancy, ranks sixth and eighth in terms of tumor-related mortality and morbidity of all tumors, respectively. ESCA is associated with a dismal prognostic outcome, and its five-year survival rate has been reported to be 15–25%. Esophageal squamous cell carcinoma (ESCC) accounts for a major ESCA subtype, which is predominant in eastern Asia (Matsushima et al., 2010). The poor outcome of ESCC is associated with its insidious initial symptoms, susceptibility to metastasis, resistance to radiotherapy, and tumor recurrence (Pennathur et al., 2013). Over the past few years, multidisciplinary and surgical treatments have been developed, but the median survival of ESCC cases is only 10 months (Wang et al., 2020a). Moreover, considering the limited prediction of prognosis for ESCC patients, there is an urgent need for the exploration of novel biomarkers.

Ferroptosis, the novel regulated cell death (RCD) type that is different from necrosis, apoptosis, and autophagy, is featured by lipid hydroperoxide accumulation till the lethal dose (Dixon et al.,

2012). As revealed by more and more studies, ferroptosis exerts an important part in tumor progression and treatment (Stockwell et al., 2017; Shen et al., 2018; Gan, 2019). Besides, various tumor types such as adrenocortical carcinoma, hepatocellular carcinoma, and ovarian cancer have been demonstrated to be sensitive to ferroptosis (Yang et al., 2014; Belavgeni et al., 2019; Carbone and Melino, 2019). Numerous reports have indicated that ferroptosis-related genes (FRGs) are involved in the regulation of tumor initiation and progression (Junttila and Evan, 2009; Arrigo and Gibert, 2012; Liu et al., 2018; Enz et al., 2019). ALOX12 exhibits a context-dependent role in mediating lipid peroxidation, resulting in PUFA oxidation which promotes cell ferroptosis. An outstanding report was performed by Chu et al., who uncovered that ALOX12 is essential for p53-mediated tumor ferroptosis through the ACSL4-independent pathway (Chu et al., 2019). Recent studies confirmed that ANGPTL7 and DRD4 were inhibited by ferroptotic erastin, indicating the potential role of being ferroptosis markers (Yang et al., 2014; Wang et al., 2016). Gao et al. proved that repression of glutamine metabolism could reduce cell ferroptosis, which revealed a novel function of SLC38A1 in regulated cell death (Gao et al., 2015). However, the relationship between these FRGs and prognostic outcomes for ESCC cases remains to be further examined.

This study downloaded ESCC patient samples and corresponding clinical information from GEO and TCGA public databases. Afterward, we successfully established the prognosis risk signature that incorporated seven FRGs based on the GEO training set and validated it in the GEO test set, entire GEO set, and TCGA dataset. Ultimately, we initially explored the oncogenic effect of SLC38A1 through *in vitro* studies. This work develops a novel FRG prognostic signature to improve the prediction of the clinical outcomes of ESCC patients.

MATERIALS AND METHODS

Data Collection

Expression RNA-seq data together with associated clinical data from ESCC cases were acquired from The *Cancer* Genome Atlas (TCGA) data portal (https://tcga-data.nci.nih.gov/tcga/) and the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) database, which defined the entire GSE53625 set (n = 179) and TCGA set (n = 81), respectively. A total of 255 FRGs were extracted from the FerrDb website (http://www.zhounan.org/ferrdb).

Identification of Ferroptosis-Related Gene Prognostic Signature

Firstly, the entire GSE53625 set (n = 179) was randomized as the training set together with the internal test set in the 1:1 ratio. Then, we performed univariate Cox regression analysis for identifying prognostic FRGs ($p < 0.05$) in the training cohort. To remove the potential overfitting genes, the "glmnet" package was adopted for least absolute shrinkage and selection operator (LASSO) regression. At last, the optimal prognosis model based on FRGs was constructed by multivariate Cox regression. To be specific, we determined the risk score for ESCC cases by the following formula: risk score = (Gene 1 expression × coefficient) + (Gene 2 expression × coefficient) + . . . + (Gene n expression × coefficient). Meanwhile, the cases were separated into high- or low-risk groups based on the median score. In addition, the test set, entire set, and TCGA set were used to validate our signature.

Nomogram Establishment and Validation

For predicting the clinical outcomes of ESCC patients, we utilized the R package "rms" to construct the nomogram which incorporated clinical factors and risk signature. Additionally, the nomogram performance and prediction accuracy were determined to plot the calibration curves.

Gene Set Enrichment Analysis

GSEA was employed to detect biological functions as well as related signaling pathways in the high-risk group. The expression of genes in both the high- and low-risk groups, together with the collection of Hallmark and KEGG gene sets in Molecular Signatures Database v7.1, was analyzed by GSEA software. Gene sets conforming to | NES |> 1 and NOM $p < 0.05$ were deemed significant based on the GSEA User Guide.

Validation of Protein Expressions of Signature Genes by the HPA Database

Immunohistochemistry (IHC) helps to uncover relative protein distribution and expression according to particular binding of antigens with antibodies. IHC was conducted to determine the prognostic FRG expression in ESCC and non-carcinoma samples from the Human Protein Atlas (HPA, https://www.proteinatlas.org/) database at the protein level.

Cell Culture and Cell Transfection

ESCC cell lines (Eca109 and KYSE-150), together with the normal human esophageal epithelial cells (HEECs), were cultivated within the RPMI-1640 medium containing 10% fetal bovine serum (FBS, Gibco Company) and 10% streptomycin–penicillin (Sigma-Aldrich) and incubated in an incubator under 37°C and 5% CO_2 conditions. In addition, si-SLC38A1 and siRNA negative control (si-NC) were prepared via Ribobio (Guangzhou, China). The sense sequence of si-SLC38A1 was 5′-GUUACCUUCAAUUCAAAGATT-3′. Later, Lipofectamine 3000 reagent (Invitrogen) was employed to transfect siRNAs to specific cells in line with specific protocols. After transfection for 48 h, we harvested cells to conduct later experiments.

Quantitative Reverse Transcription Polymerase Chain Reaction

We utilized Trizol reagent (Vazyme Biotech, Nanjing, China) to isolate the total cellular RNA from ESCC cells. All extraction steps were performed in line with specific protocols. The BioSpec-nano spectrophotometer (Shimadzu, Japan) was used to measure the extracted RNA content. We deemed RNA samples that had the A260/A280 ratio of 1.8–2 as suitable samples. We then reverse transcribed the RNA using Prime Script RT Master Mix reagent

(Takara Bio, Dalian, China) for obtaining cDNA. The PCR system was prepared to utilize TB Green® Premix Ex Taq™ (Takara Bio, Dalian, China). We performed qRT-PCR on the Applied Biosystems StepOnePlus real-time PCR system (Thermo Fisher Scientific). In addition, the $2^{-\Delta\Delta CT}$ approach was applied in calculating the relative gene level. The SLC38A1 level was analyzed by the following primers: 5′-GATGGGTGATGGTGA TAGGG-3′ (forward) and 5′-TACTGGTCTAGGGGCCACAC-3′ (reverse). GAPDH was used as a reference gene.

Western Blot Analysis

Western blot analysis was conducted for determining SLC38A1 and GAPDH levels. The SLC38A1 (#36057, 1:1,000) and GAPDH (#5174, 1:1,000) antibodies were provided by Cell Signaling Technology (CST, Danvers, MA, United States).

Cell Counting Kit-8 Assay

We used the CCK-8 kit (Beyotime, Shanghai, China) for determining cell proliferation following specific protocols. The cells (2000/well) were inoculated into the 96-well plates and cultured within RPMI-1640 that contained 10% FBS. At a fixed time of day, we added CCK-8 solution into each well to incubate cells under 37°C for additional 2 h. The absorbance (OD) value was detected at 450 nm.

Colony Formation Assay

The cells (250/well) after transfection were inoculated to six-well plates in the colony formation assay and cultured within the RPMI-1640 medium that contained 10% FBS for a period of 10 days. Later, 1% formaldehyde was used to fix the growing colonies, whereas 1% crystal violet was utilized to stain the colonies. After taking images, we counted the colony number.

Transwell Assay

The Transwell chamber (pore size, 8 μm; Corning Costar Corp, United States) was utilized to examine cell migration. In brief, after suspending the stably transfected ESCC cells into the serum-free RPMI-1640 medium (200 μL), the upper chamber was loaded with cell suspension. Afterward, the RPMI-1640 medium (500 μL) that contained 10% FBS was placed into the lower chamber, followed by 24 h of cell incubation under 37°C. Later, 1% crystal violet was used to stain cells for 20 min, and then cotton swabs were used to remove cells on the upper membrane surface. A microscope (Olympus) was used to take photographs of cells on the bottom membrane surface, and four random fields were utilized to count the migration cells.

Statistical Analysis

R software (3.6.3) and GraphPad (8.0) were employed for all statistical data analyses. The log-rank test and Kaplan–Meier analysis were adopted for evaluating different OS between high- and low-risk groups. Besides, univariate and multivariate Cox regression was applied in identifying those independent factors for predicting prognosis. Time-dependent receiver operating characteristic (ROC) curves were used to evaluate our risk model for its prediction performance. A difference of $p < 0.05$ indicated statistical significance.

RESULTS

Construction and Verification of the Ferroptosis-Related Gene Prognostic Signature

A total of 179 ESCC patients from GSE53625 were randomized in a 1: 1 ratio into a training cohort (90 samples) and an internal validation cohort (89 samples). LASSO regression and multivariate Cox regression were performed in the training set to identify seven ferroptosis-related genes (ALOX12, ALOX12B, ANGPTL7, DRD4, MAPK9, SLC38A1, and ZNF419) for constructing a novel prognostic signature (**Figure 1**). The formula is shown as follows: Risk score = [ALOX12 expression× (−0.097)] + [ALOX12B expression× (−0.147)] + [ANGPTL7 expression× (0.326)] + [DRD4 expression × (−0.254)] + [MAPK9 expression × (0.288)] + [SLC38A1 expression× (−0.904)] + [ZNF419 expression × (0.782)]. We classified the ESCC cases into low- and high-risk groups based on the median risk score. The predictive performance of our seven-FRG–based risk model to predict patient OS can be observed in **Figure 2A**. As suggested through the Kaplan–Meier curve plotted according to the log-rank test, high-risk patients had poor OS compared with low-risk patients ($p < 0.05$, **Figure 2B**). For evaluating the credibility of our constructed model in predicting prognosis, we conducted ROC curve analysis. According to **Figure 2C**, area under the curve (AUC) values for the one-, three-, and five-year survival were determined to be 0.656, 0.765, and 0.788, respectively, for the GEO training set. The same analysis was conducted in the GEO validation cohort, and the AUC values for one-, three-, and five-year survival were 0.609, 0.697, and 0.647, respectively (**Figure 2C**). Moreover, we observed similar results in TCGA and the entire GEO sets, which proved the strong predictive potential of our risk model (**Figure 2**).

Subgroup Analysis for the Ferroptosis-Related Gene Prognostic Signature

This study determined the predictive performance of the prognostic signature for OS of patients who had diverse clinical parameters. As a result, subgroups were categorized according to age (≤65 vs. >65 years), gender (male vs. female), clinical stage (I-II vs. III), T stage (T1 + T2 vs. T3 + T4), and N stage (N0 vs. N1–N3). Based on age, gender, clinical stage, T stage, and N stage, high-risk patients had markedly poor five-year OS rates compared with low-risk patients (**Figure 3**).

Prognostic Nomogram Establishment and Validation

For investigating the possibility of using the as-constructed prognosis nomogram as the factor to independently predict the prognosis for ESCC cases, univariate together with multivariate Cox regression was carried out. As revealed by univariate analysis, age ($p = 0.009$), risk score ($p < 0.001$), N stage ($p < 0.001$), and clinical stage ($p < 0.001$) predicted the dismal OS (**Figure 4A**). In addition, according to multivariate

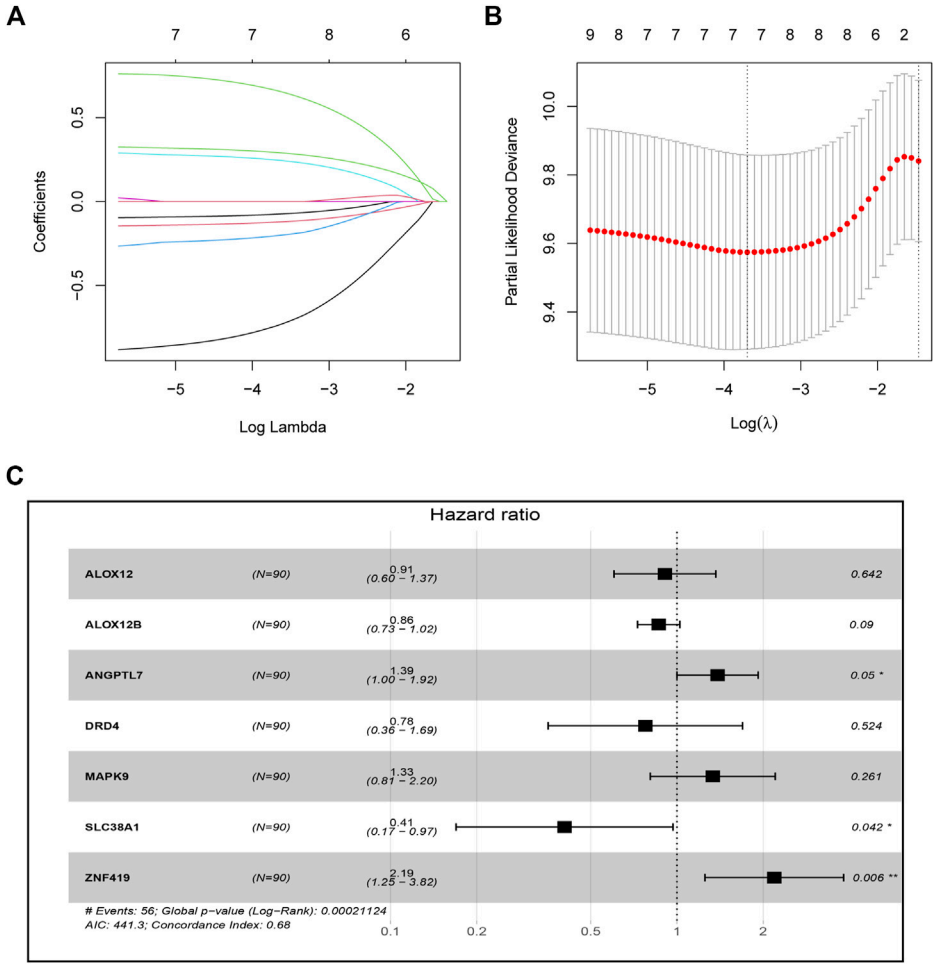


FIGURE 1 | Construction of the seven-ferroptosis-gene signature. **(A)** Cross-validation for tuning parameter screening upon LASSO regression analysis. **(B)** LASSO coefficient profiles for those intersected genes. **(C)** Forest plot of hazard ratios exhibiting the prognostic worth of seven FRGs.

Cox regression, the risk score (HR = 2.009, 95% CI = 1.559–2.589, $p < 0.001$) and age (HR = 1.034, 95% CI = 1.010–1.059, $p = 0.005$) were identified as the independent prognostic factors that predicted the poor OS for ESCC cases (**Figure 4B**). Subsequently, we incorporated the risk score and other clinicopathologic characteristics to establish a novel nomogram to predict the one-, three-, and five-year OS rates of ESCC patients (**Figure 4C**). Every individual patient would acquire a corresponding score, and a higher total point demonstrates a poorer outcome for the patient. Moreover, the one-, three-, and five-year survival calibration curves well fitted our constructed nomogram in the GEO entire cohort (**Figures 4D–F**).

Gene Set Enrichment Analysis With the Ferroptosis-Related Gene Prognostic Signature

We also conducted GSEA for clarifying the possible biological functions and signal transduction pathways among high-risk patients. As shown in **Figure 5**, a higher risk score was correlated with adhesion molecules, chemokine signaling pathway, KRAS signaling, and IL-2/STAT5 signaling, indicating that the patients with these pathways might be more prone to a worse clinical outcome.

Difference of Immune Checkpoints Between the High-Risk and Low-Risk Groups

To further explore the relationship between the immune checkpoints and two risk groups, we performed differentiation analysis for the expression of 22 immune checkpoints, including the TNF superfamily (BTLA, CD27, CD40LG, CD40, CD70, TNFRSF18, TNFRSF9, and TNFSF9) and B7-CD28 family (CD274, CD276, CTLA4, HHLA2, ICOS, ICOSLG, PDCD1, PDCD1LG2, and VTCN1), along with additional immune checkpoints (IDO1, HAVCR2, VSIR, LAG3, and NCR3). As shown in **Figure 6**, BTLA, CD40, CD40LG, CTLA4, and HAVCR2 were significantly

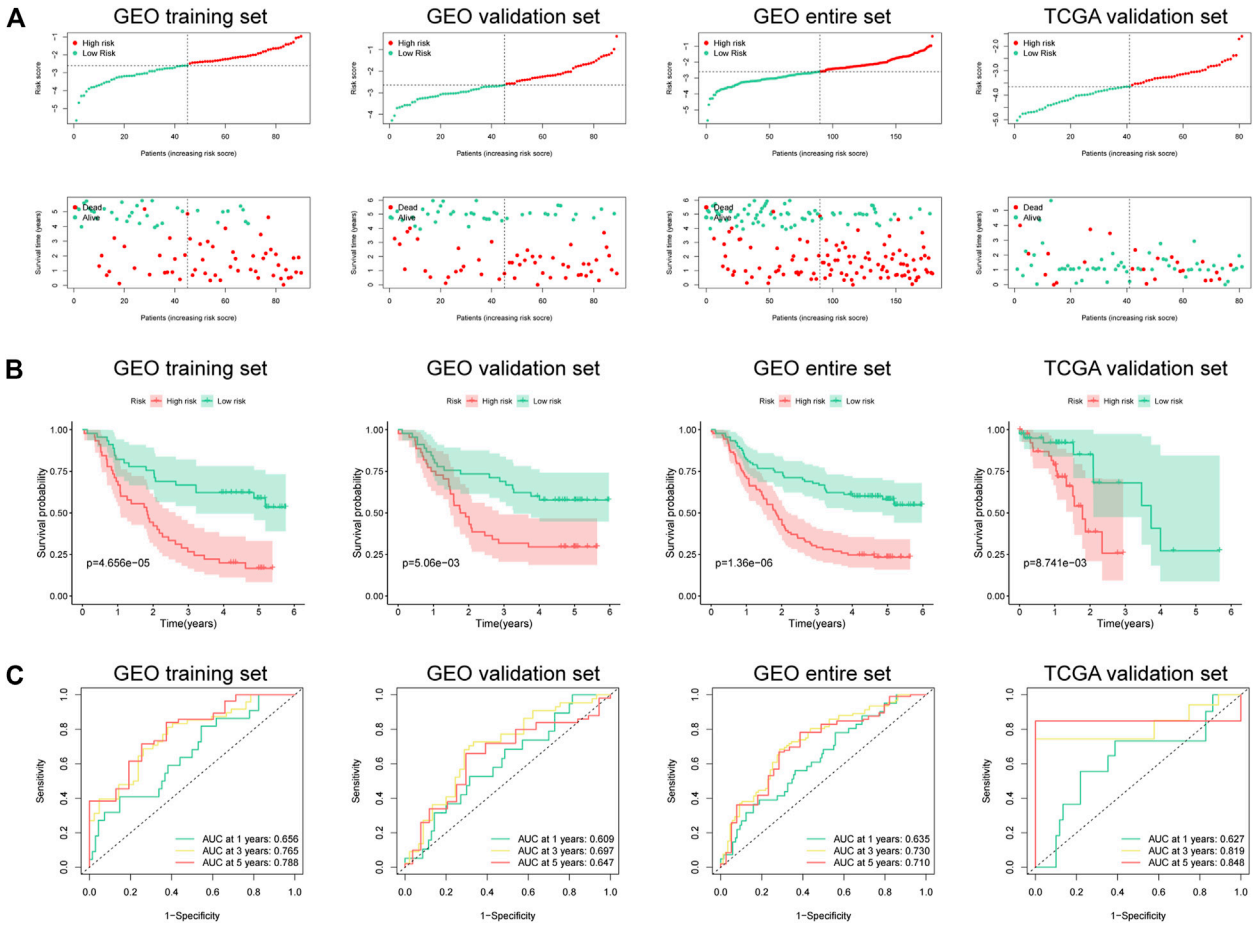


FIGURE 2 | Risk score of the prognostic signature that comprises seven FRGs for OS in four cohorts. **(A)** Risk score distribution and survival status of high- and low-risk patients. **(B)** Kaplan–Meier analysis on high- and low-risk patients. **(C)** Time-dependent ROC curve analyses on the GEO training set, GEO validation set, entire GEO set, and TCGA validation set.

upregulated in the high-risk group, while HHLA2 was enriched in the low-risk group.

Validation of the Expression Patterns and Protein Expression of Prognostic Signature Genes

We confirmed the expression levels of the seven signature genes (ALOX12, ALOX12B, ANGPTL7, DRD4, MAPK9, SLC38A1, and ZNF419) among the patients from GSE53625. The results showed that ALOX12, ANGPTL7, DRD4, and MAPK9 remarkably decreased within ESCC samples relative to non-carcinoma samples, whereas SLC38A1 and ZNF419 were highly expressed. Only ALOX12B expression showed no significant difference in tumor samples compared with normal samples (**Figure 7**). Consistent with the above results, the HPA database showed that ALOX12 and MAPK9 in ESCC tissues were lowly expressed, while SLC38A1 and ZNF419 were upregulated relative to normal samples. But DRD4 and ANGPTL7 protein expressions were not measured in the database (**Figure 8**).

Inhibition of SLC38A1 Decreased Esophageal Squamous Cell Carcinoma Cell Proliferation and Migration

Finally, we used the SLC38A1 gene to further explore the underlying role of our model in ESCC. First, the qRT-PCR assay and western blot analysis were performed to verify the differential expression between normal esophageal epithelial cells and ESCC cells (**Figure 9A**). As a result, SLC38A1 expression increased within ESCC cells relative to normal esophageal epithelial cells. Next, the siRNAs were applied to knock down the SLC38A1 levels within Eca109 and KYSE-150 cells, and both the qRT-PCR assay and western blot analysis confirmed the efficacy (**Figure 9B**). The CCK-8 proliferation assay and colony formation assay showed that downregulation of SLC38A1 can markedly reduce Eca109 and KYSE-150 cell proliferation (**Figures 9C–E**). Moreover, migration of Eca109 and KYSE-150 cells transfected with siRNA was inhibited (**Figure 9F**). These results suggest that SLC38A1 possibly promotes tumorigenesis of ESCC, yet the possible mechanism should be further explored.

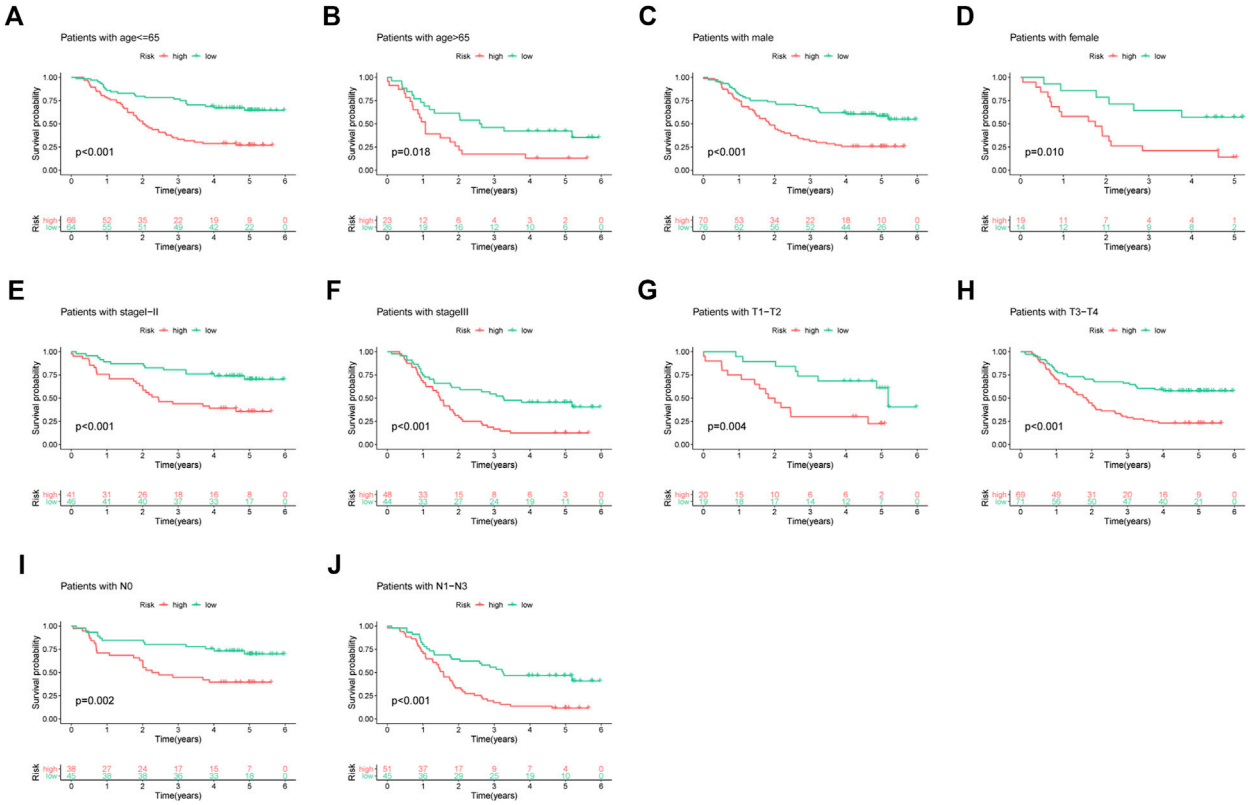


FIGURE 3 | Subgroup analyses based on different clinical features of ESCC cases: **(A)** age ≤65, **(B)** age >65, **(C)** male, **(D)** female, **(E)** stage I-II, **(F)** stage III, **(G)** T1-T2, **(H)** T3-T4, **(I)** N0, and **(J)** N1–N3.

DISCUSSION

ESCC is still a progressive and challenging disease with high morbidity and poor prognosis. Presently, the TNM staging system is still a crucial prognostic factor for assessing the prognosis of cancer patients, but it has limitation in elucidation of genetic variations, and those at the identical stage present a powerful heterogeneity for prognostic outcome. Selectively inducing the death of cancer cells may account for an efficient way to treat cancer. Mounting evidence indicated that ferroptosis plays a significant role in tumorigenesis and treatment for cancer (Yang et al., 2014; Stockwell et al., 2017; Liu et al., 2018; Carbone and Melino, 2019; Hassannia et al., 2019; Liang et al., 2019; Tesfay et al., 2019). However, there has not yet been much systematic analysis in the context of ferroptosis in ESCC, and the underlying mechanism of ESCC remains poorly illustrated.

This work concentrated on ferroptosis-related gene signatures with the prognosis value of ESCC patients. In the GEO training set, we first identified prognostic ferroptosis-related genes and then built the predictive model comprising seven FRGs through integration of LASSO regression and Cox regression analysis. According to Kaplan–Meier curve analysis, high-risk cases were associated with dismal OS compared with low-risk counterparts. Meanwhile, the ROC curve illustrated good performance of our model. The AUCs of ROC plots for five-year OS in the GEO cohort and TCGA cohort were 0.788 and 0.848, respectively. Furthermore, ROC curves were utilized to compare the prediction capability of our proposed model with that of other signatures. As a result, our risk signature achieved consistently excellent predictive value, compared with other published risk prognostic signatures in ESCC (Wang et al., 2020b; Gao et al., 2021; Zhao et al., 2021). The constructed prognostic signature was also verified in the GEO test set, entire GEO set, and TCGA set. Next, the seven ferroptosis-related genes' signature predicted the dismal OS for ESCC cases after subgroup analysis according to age, gender, clinical stage, T stage, and N stage. The results of Cox regression analysis showed that the as-constructed risk model might serve as an independent risk factor for ESCC. Moreover, the nomogram was established and the calibration plots were used to examine whether our nomogram was accurate in the prediction of one-, three-, and five-year OS. All these results revealed that the ferroptosis-related signature could be a superior predictor compared with the traditional clinical indicator.

Our proposed ferroptotic signature was composed of seven ferroptosis-related genes (ALOX12, ALOX12B, ANGPTL7, DRD4, MAPK9, SLC38A1, and ZNF419). Among the seven genes, ANGPTL7, MAPK9, and ZNF419 are latent hazardous genes and ALOX12, ALOX12B, DRD4, and SLC38A1 are potential protective genes. All these genes were shown to participate in the initiation and development of various cancers. ALOX12 belongs to a family of lipoxygenases (LOXs)

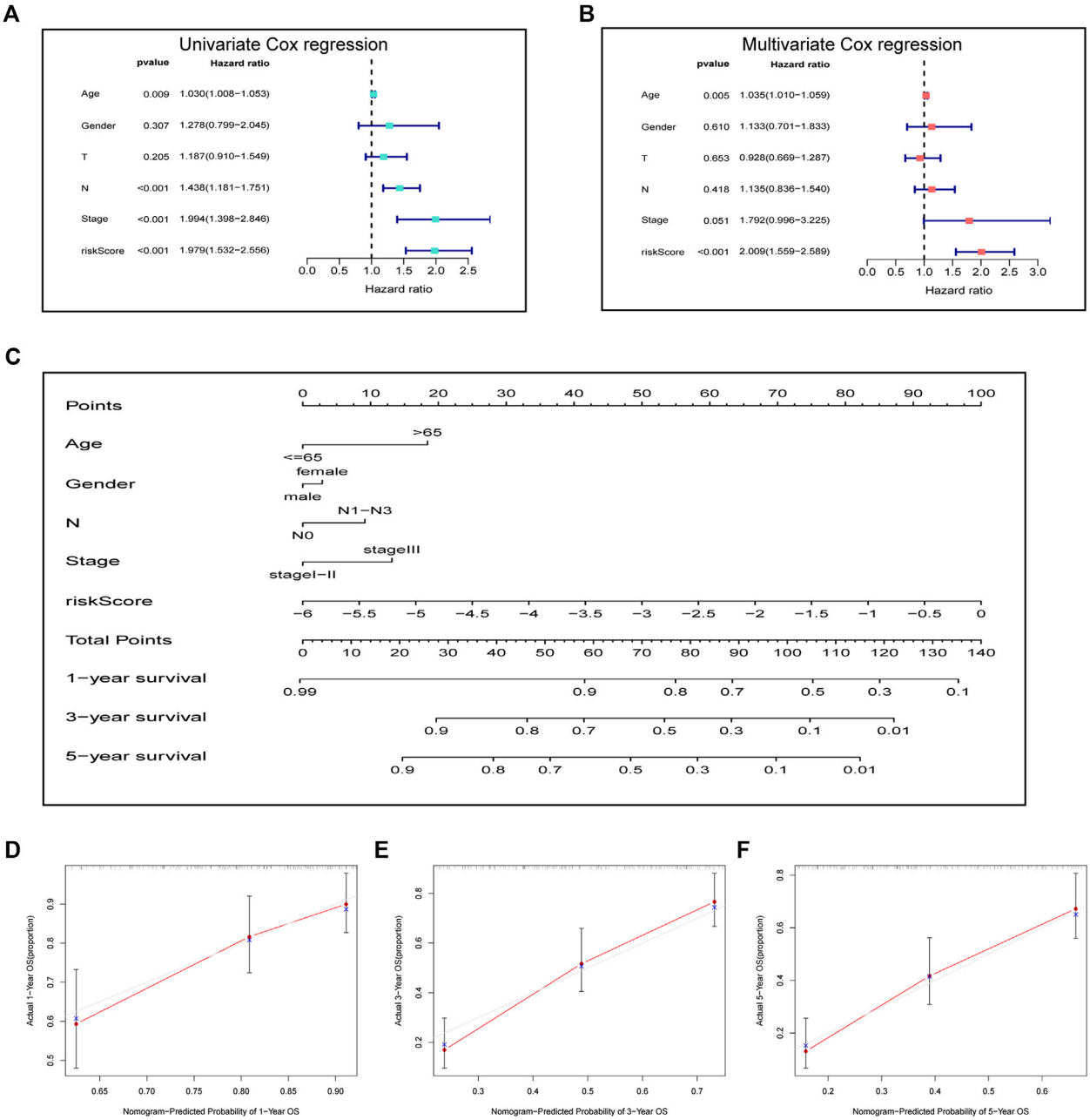


FIGURE 4 | Prognostic signature in combination with clinical parameters for predicting prognostic outcomes for ESCC cases. **(A)** Univariate analysis and **(B)** multivariate analysis containing the risk score and clinical factors. **(C)** Nomogram for predicting one-, three-, and five-year OS. **(D–F)** Calibration curves of nomogram on consistency between predicted and observed one-, three-, and five-year survival.

with a reported role in the promotion of the oxidation activity of polyunsaturated fatty acids (Yoshimoto et al., 1992). The ALOX12 protein could foster the biosynthesis of 12-hydroxyeicosatetraenoic acid by specifically metabolizing arachidonic acid (Honn et al., 1994). It has been confirmed that ALOX12 has the capability of mediating inflammation, cell migration, apoptosis, and tumor cell proliferation (Zheng et al., 2020). Yang et al. found that ALOX12 was downregulated in recurrence of hepatocellular carcinoma and regulated the ALOX12–12HETE–GPR31 signaling pathway (Yang et al., 2019). In lung cancer, overexpression of ALOX12 facilitated cell growth and migration by promoting RhoA and NF-κB activity (Chen et al., 2020). ALOX12B protein, another isoform of arachidonic acid 12-lipoxygenase, mainly catalyzes arachidonic acid to 12R-hydroxyeicosatetraenoic acid (Zheng et al., 2011). Jiang et al. revealed that the inhibition of ALOX12B could restrain cervical cancer cell proliferation and growth through suppressing the PI3K/ERK1 pathway, suggesting

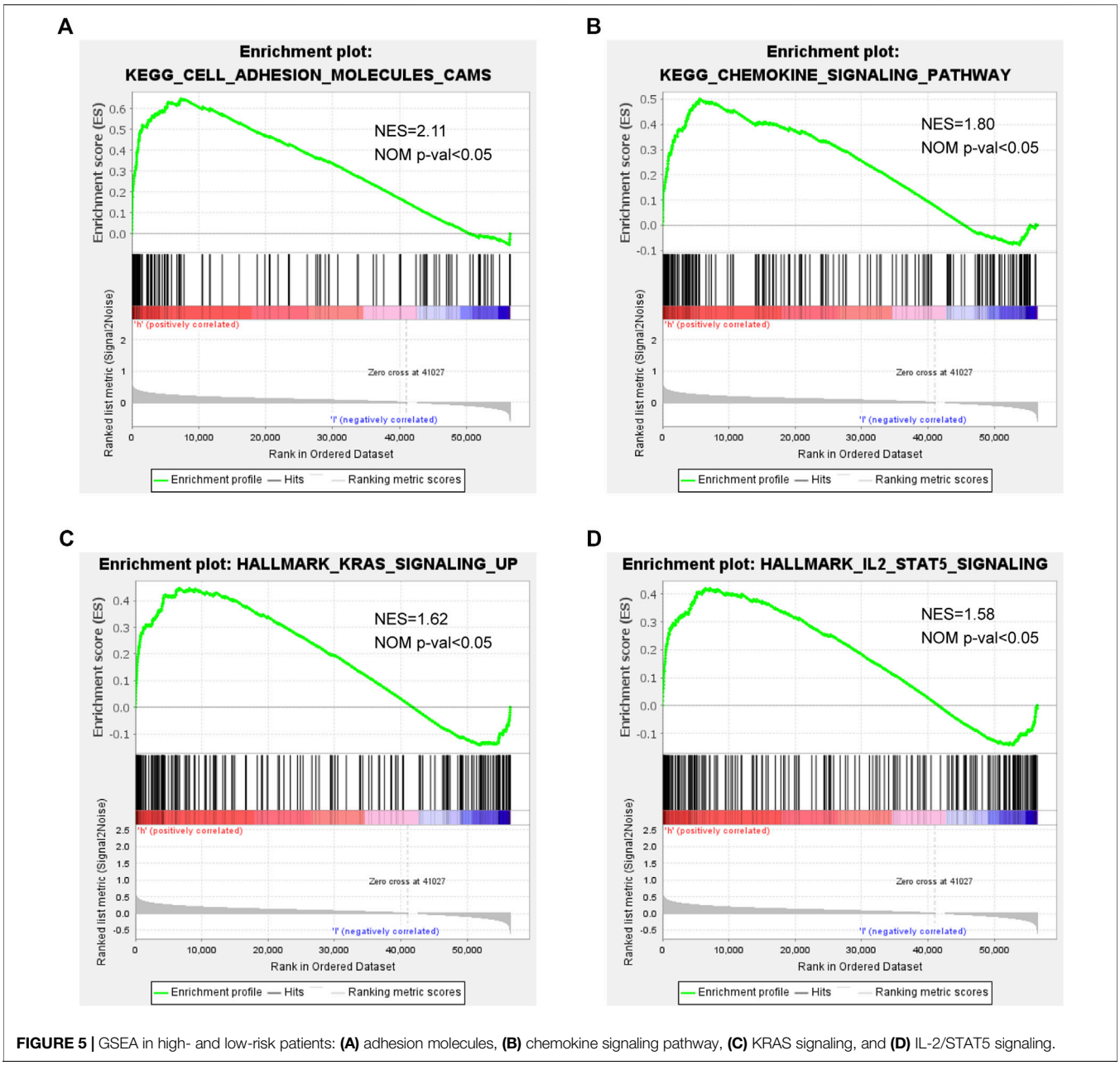


FIGURE 5 | GSEA in high- and low-risk patients: **(A)** adhesion molecules, **(B)** chemokine signaling pathway, **(C)** KRAS signaling, and **(D)** IL-2/STAT5 signaling.

that it can be taken as a good biomarker to provide new therapeutic strategies for cervical cancer patients (Jiang et al., 2020). In addition, Chu et al. reported that ALOX12 could oxygenate polyunsaturated fatty acids, which in turn induce p53-mediated tumor cell ferroptosis (Chu et al., 2019). Consistent with previous studies, our results indicate a negative correlation between ALOX12 and the poor prognosis of patients.

ANGPTL7, a member of the angiopoietin-like protein (ANGPTL) family, consists of an N-terminal coiled-coil domain and a C-terminal fibrinogen-like domain. The same structural domain as angiopoietin ensures ANGPTL7 to promote angiogenesis (Carbone et al., 2018). For instance, Parri et al. gave us a hint that hypoxia induced ANGPTL7 expression in tumor cells, which exert a vital part in pro-angiogenetic development (Parri et al., 2014). It was reported that ferroptosis induced by erastin or RSL3 could downregulate ANGPTL7, which might be involved in the onset of ferroptosis in cancer cells (Yang et al., 2014). The higher expression level of ANGPTL7 was also observed in colorectal cancer based on the gene profile analysis (Liu and Zhang, 2017). Our results are in line with these research studies, pointing out that ANGPTL7 is a risky gene (HR > 1) in ESCC. The DRD4 gene encodes the G-protein–coupled receptor which could suppress the activity of adenylyl cyclase. In glioblastoma, DRD4 could promote proliferation and autophagic flux and enhance survival of

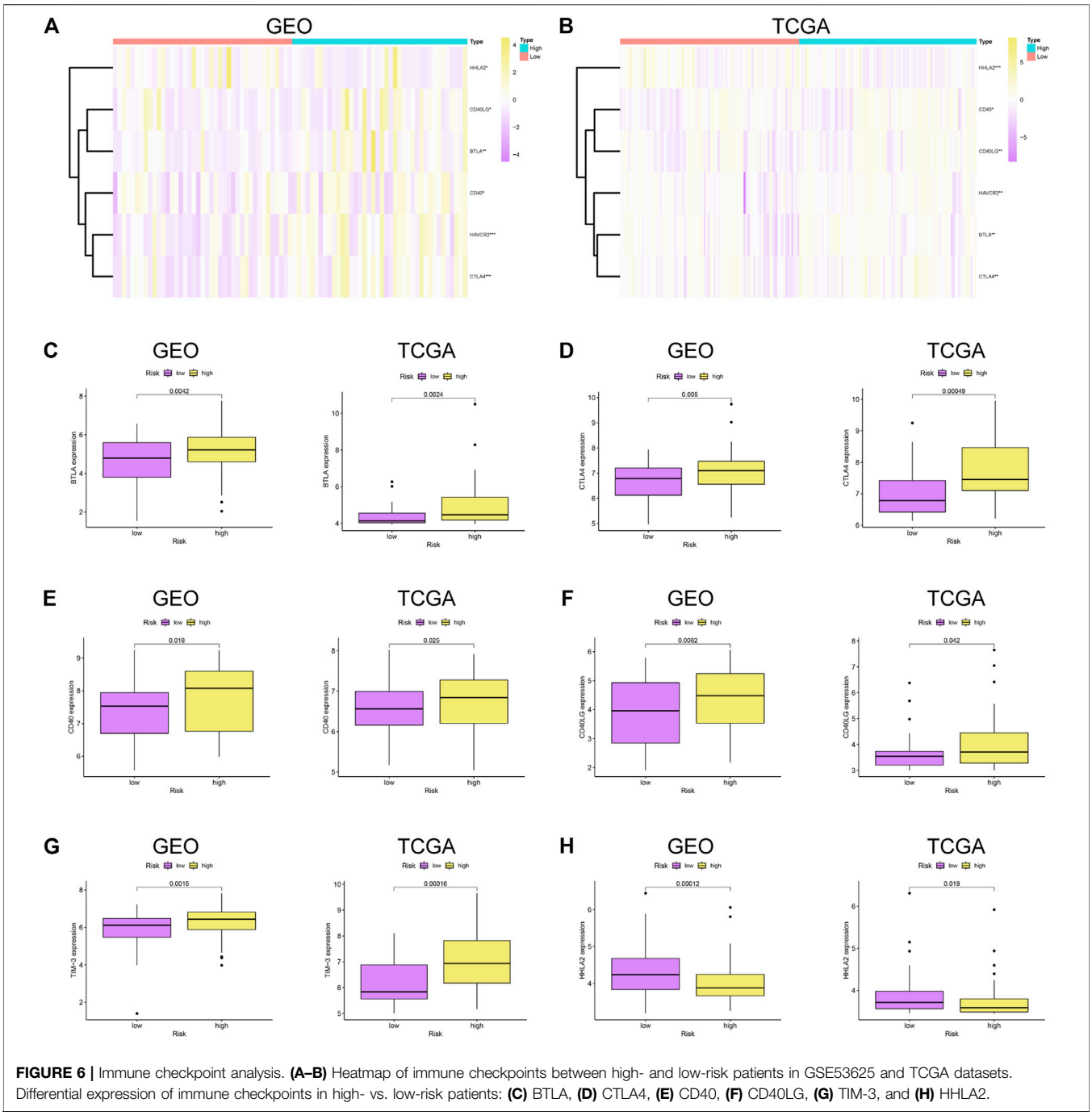


FIGURE 6 | Immune checkpoint analysis. **(A–B)** Heatmap of immune checkpoints between high- and low-risk patients in GSE53625 and TCGA datasets. Differential expression of immune checkpoints in high- vs. low-risk patients: **(C)** BTLA, **(D)** CTLA4, **(E)** CD40, **(F)** CD40LG, **(G)** TIM-3, and **(H)** HHLA2.

glioblastoma stem cells (Dolma et al., 2016). Wang et al. demonstrated that ferroptotic erastin contributed to degradation of DRD4 protein and anti-ferroptotic dopamine impeded DRD4 protein declination (Wang et al., 2016). MAPK9 could phosphorylate a series of transcription factors, which subsequently regulates cell proliferation, migration, and programmed cell death. Li et al. discovered that MEG3 and MIAT may foster the progression of lung adenocarcinoma through interacting with miR-106, thus regulating the involvement of MAPK9 in the MAPK signal transduction pathways (Li et al., 2016). SLC38A1, also known as amino acid transporter system A1, was initially identified as a crucial transporter of glutamine (Gu et al., 2001). SLC38A1 has been proved to be a potential oncogene in colorectal cancer and gastric cancer (Xie et al., 2014; Zhou et al., 2017). As a transcriptional regulator, ZNF419 polymorphism at the splice donor site might result in novel minor histocompatibility antigen ZAPHIR related to renal cell carcinoma (Broen et al., 2011).

Immune checkpoints could exert tumor immunosuppressive effects, which in turn prevent tumors from immune attack. BTLA was a member of the TNF superfamily, and its expression was

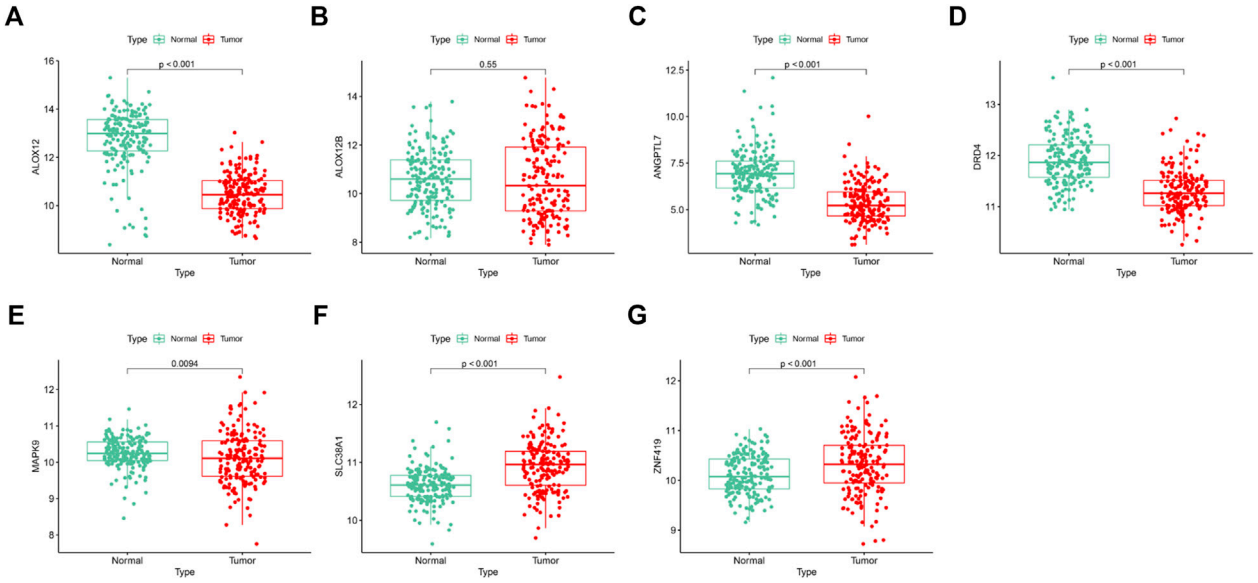


FIGURE 7 | Differentiation analysis of the signature genes among ESCCs and normal tissues based on GSE53625 dataset.

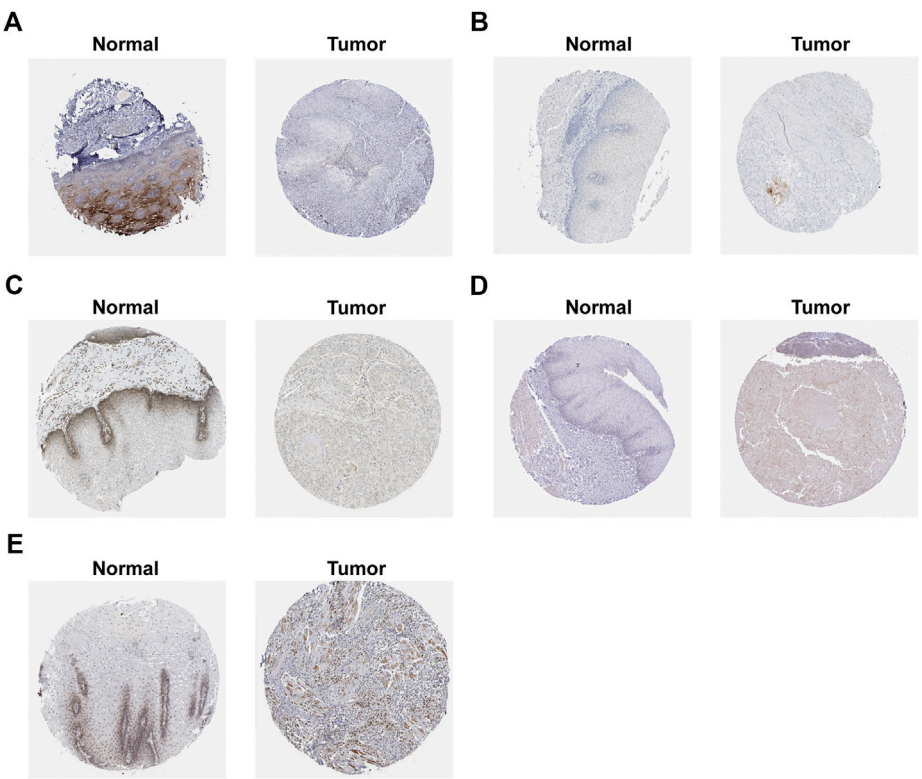


FIGURE 8 | Protein expression of signature genes in HPA database. **(A)** protein expression of ALOX12. **(B)** protein expression of ALOX12B. **(C)** protein expression of MAPK9. **(D)** protein expression of SLC38A1. **(E)** protein expression of ZNF419.

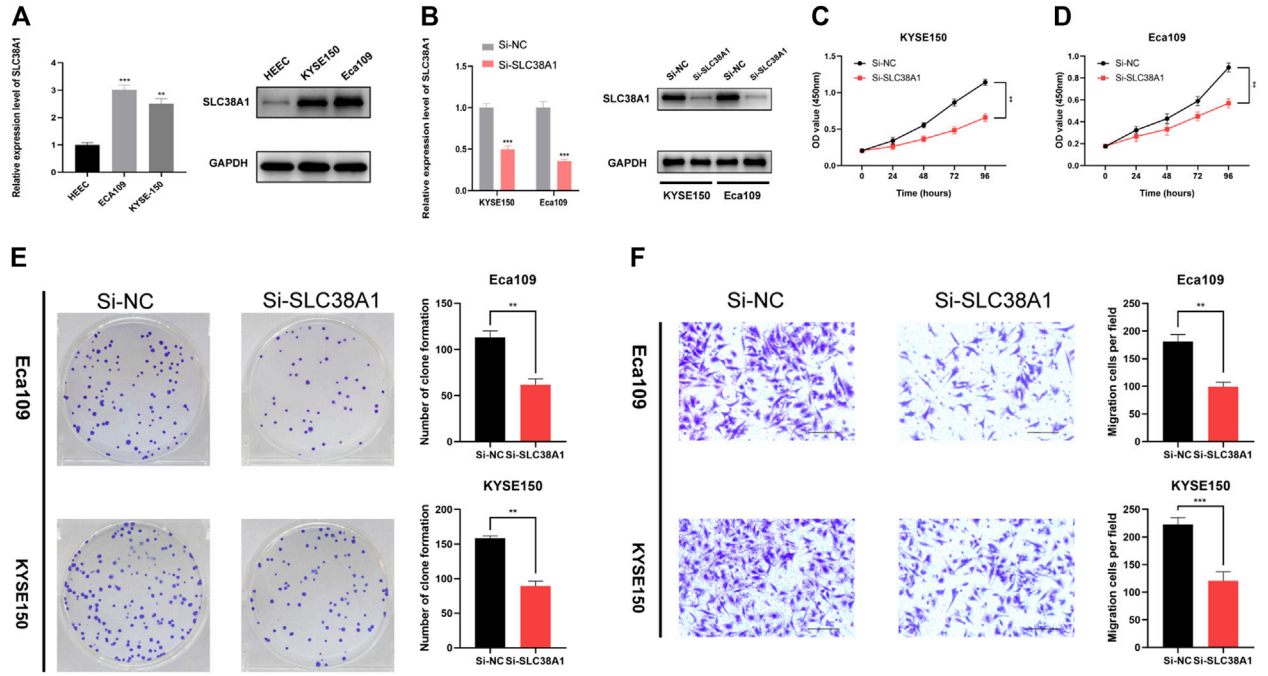


FIGURE 9 | Effects of inhibiting the expression of SLC38A1 on ESCC cell proliferation and migration. **(A)** SLC38A1 was upregulated within KYSE-150 and Eca109 cells relative to HEECs by the qRT-PCR assay and western blot analysis. **(B)** The expression of SLC38A1 was downregulated in KYSE-150 and Eca109 cells by siRNAs. **(C–E)** KYSE-150 and Eca109 cell proliferation after anti-SLC38A1 siRNA transfection was measured using CCK-8 assays **(C, D)** and colony formation assays **(E)**. **(F)** Transwell assays for ESCC cell migration (scale bars: 100 μm).

associated with cancer aggressiveness (Wang et al., 2020c). TIM-3, also known as HAVCR2, was predominantly located on NK cells and macrophages, inhibiting the activation of anti-tumor immunity (Datar et al., 2019). Matsumura et al. indicated that CD40 expression in ESCC is closely correlated with tumorigenesis and lymph node metastasis (Matsumura et al., 2016). In our results, most of immune checkpoints were related to the high-risk group, which verify the reliability of the signature in evaluating the prognosis of patients. Notably, some of the signature genes also have intricate connection with immune checkpoints. For example, MAPK9, also known as JNK2, was confirmed to be involved in the regulation of B7.1 (CD80) which could interact with CTLA-4 to mediate the development of immune responses. Lim et al. found that the expression of B7.1 induced by LPS was significantly suppressed by siJNK2 RNAs (Lim et al., 2005). It is reasonable to speculate that the downregulation of MAPK9 in ESCC might facilitate carcinogenesis through inhibiting B7.1-mediated activation of immune responses. In addition, restriction of glutamine utilization could enhance anti-programmed death ligand-1 (PD-L1) levels in tumor, which promote the effectiveness of PD-L1 antibody (Byun et al., 2020). Therefore, we hypothesized that SLC38A1, a key transporter of glutamine, might block the effectiveness of PD-L1 antibody by stimulating glutamine metabolism in ESCC.

Finally, we sought to detect the relationship between SLC38A1 and ESCC progression. The results showed that inhibiting SLC38A1 suppressed the cell viability and migration of Eca109 and KYSE-150 cells, which further proved the carcinogenic role of SLC38A1 in digestive-system neoplasms.

There are several limitations of this study. First, the data analyzed in the present work might be acquired from the public database. The clinical effectiveness and credibility of the as-constructed signature should be further verified by more practical data. Second, the functional mechanisms of signature need to be explicated through more profound *in vivo* and *in vitro* experiments.

To sum up, this work first identifies a new FRG-based prognostic signature, which predicts the OS of ESCC and mirrors the immune status. This constructed signature will provide new options for individualized treatment.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: TCGA (https://portal.gdc.cancer.gov/), GEO database (https://www.ncbi.nlm.nih.gov/geo/), FerrDb website (http://www.zhounan.org/ferrdb) and HPA database (https://www.proteinatlas.org/).

AUTHOR CONTRIBUTIONS

QL, JS, and WY were responsible for study conception and design. JS, XZ, and YL were in charge of data collection and analysis. JS and XG contributed to data interpretation. JS, YL, and

XG were in charge of manuscript drafting. QL and WY were responsible for manuscript revision. All authors read and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant numbers 81670421 and 81800078); Six Talent Peaks Project of Jiangsu Province, China (TD-SWYY-005); and Jiangsu Provincial Social Development Project (BE2016798).

ACKNOWLEDGMENTS

The authors would like to thank the Core Facility of the First Affiliated Hospital of Nanjing Medical University for their help in the detection of experimental samples.

REFERENCES

Arrigo, A.-P., and Gibert, B. (2012). HspB1 Dynamic Phospho-Oligomeric Structure Dependent Interactome as Cancer Therapeutic Target. *Cmm* 12 (9), 1151–1163. Epub 2012/07/19.PubMed PMID: 22804238. doi:10.2174/156652412803306693

Belavgeni, A., Bornstein, S. R., von Mässenhausen, A., Tonnus, W., Stumpf, J., Meyer, C., et al. (2019). Exquisite Sensitivity of Adrenocortical Carcinomas to Induction of Ferroptosis. *Proc. Natl. Acad. Sci. USA* 116 (44), 22269–22274. Epub 2019/10/16PubMed PMID: 31611400; PubMed Central PMCID: PMCPMC6825277. doi:10.1073/pnas.1912700116

Broen, K., Levenga, H., Vos, J., van Bergen, K., Fredrix, H., Greupink-Draaisma, A., et al. (2011). A Polymorphism in the Splice Donor Site of ZNF419 Results in the Novel Renal Cell Carcinoma-Associated Minor Histocompatibility Antigen ZAPHIR. *PLoS One* 6 (6), e21699, 2011 . Epub 2011/07/09PubMed PMID:21738768; PubMed Central PMCID: PMCPMC3125305. doi:10.1371/journal.pone.0021699

Byun, J.-K., Park, M., Lee, S., Yun, J. W., Lee, J., Kim, J. S., et al. (2020). Inhibition of Glutamine Utilization Synergizes with Immune Checkpoint Inhibitor to Promote Antitumor Immunity. *Mol. Cel* 80 (4), 592–606. e8Epub 2020/11/08PubMed PMID: 33159855. doi:10.1016/j.molcel.2020.10.015

Carbone, C., Piro, G., Merz, V., Simionato, F., Santoro, R., Zecchetto, C., et al. (2018). Angiopoietin-Like Proteins in Angiogenesis, Inflammation and Cancer. *Ijms* 19 (2), 431, 2018 . Epub 2018/02/02PubMed PMID: 29389861; PubMed Central PMCID:PMCPMC5855653. doi:10.3390/ijms19020431

Carbone, M., and Melino, G. (2019). Stearoyl CoA Desaturase Regulates Ferroptosis in Ovarian Cancer Offering New Therapeutic Perspectives. *Cancer Res.* 79 (20), 5149–5150. Epub 2019/10/17PubMed PMID: 31615810. doi:10.1158/0008-5472.CAN-19-2453

Chen, J., Tong, W., Liao, M., and Chen, D. (2020). Inhibition of Arachidonate Lipoxygenase12 Targets Lung Cancer through Inhibiting EMT and Suppressing RhoA and NF-Kb Activity. *Biochem. Biophysical Res. Commun.* 524 (4), 803–809. Epub 2020/02/11PubMed PMID: 32037090. doi:10.1016/j.bbrc.2020.01.166

Chu, B., Kon, N., Chen, D., Li, T., Liu, T., Jiang, L., et al. (2019). ALOX12 Is Required for P53-Mediated Tumour Suppression through a Distinct Ferroptosis Pathway. *Nat. Cel Biol* 21 (5), 579–591. Epub 2019/04/10PubMed PMID: 30962574;PubMed Central PMCID: PMCPMC6624840. doi:10.1038/s41556-019-0305-6

Datar, I., Sanmamed, M. F., Wang, J., Henick, B. S., Choi, J., Badri, T., et al. (2019). Expression Analysis and Significance of PD-1, LAG-3, and TIM-3 in Human Non-small Cell Lung Cancer Using Spatially Resolved and Multiparametric Single-Cell Analysis. *Clin. Cancer Res.* 25 (15), 4663–4673. Epub 2019/05/06PubMed PMID: 31053602; PubMed Central PMCID: PMCPMC7444693. doi:10.1158/1078-0432.CCR-18-4142

Dixon, S. J., Lemberg, K. M., Lamprecht, M. R., Skouta, R., Zaitsev, E. M., Gleason, C. E., et al. (2012). Ferroptosis: an Iron-dependent Form of Nonapoptotic Cell Death. *Cell* 149 (5), 1060–1072. Epub 2012/05/29PubMed PMID: 22632970; PubMed Central PMCID: PMCPMC3367386. doi:10.1016/j.cell.2012.03.042

Dolma, S., Selvadurai, H. J., Lan, X., Lee, L., Kushida, M., Voisin, V., et al. (2016). Inhibition of Dopamine Receptor D4 Impedes Autophagic Flux, Proliferation, and Survival of Glioblastoma Stem Cells. *Cancer Cell* 29 (6), 859–873. Epub 2016/06/15PubMed PMID: 27300435; PubMed Central PMCID: PMCPMC5968455. doi:10.1016/j.ccell.2016.05.002

Enz, N., Vliegen, G., De Meester, I., and Jungraithmayr, W. (2019). CD26/DPP4 - a Potential Biomarker and Target for Cancer Therapy. *Pharmacol. Ther.* 198, 135–159. Epub 2019/03/02PubMed PMID: 30822465. doi:10.1016/j.pharmthera.2019.02.015

Gan, B. (2019). DUBbing Ferroptosis in Cancer Cells. *Cancer Res.* 79 (8), 1749–1750. Epub 2019/04/17PubMed PMID: 30987975; PubMed Central PMCID: PMCPMC7193871. doi:10.1158/0008-5472.CAN-19-0487

Gao, J., Tang, T., Zhang, B., and Li, G. (2021). A Prognostic Signature Based on Immunogenomic Profiling Offers Guidance for Esophageal Squamous Cell Cancer Treatment. *Front. Oncol.* 11, 603634, 2021 . Epub 2021/03/16PubMed PMID: 33718151; PubMed Central PMCID:PMCPMC7943886. doi:10.3389/fonc.2021.603634

Gao, M., Monian, P., Quadri, N., Ramasamy, R., and Jiang, X. (2015). Glutaminolysis and Transferrin Regulate Ferroptosis. *Mol. Cel* 59 (2), 298–308. Epub 2015/07/15PubMed PMID: 26166707; PubMed Central PMCID: PMCPMC4506736. doi:10.1016/j.molcel.2015.06.011

Gu, S., Roderick, H. L., Camacho, P., and Jiang, J. X. (2001). Characterization of an N-System Amino Acid Transporter Expressed in Retina and its Involvement in Glutamine Transport. *J. Biol. Chem.* 276 (26), 24137–24144. Epub 2001/04/28ubMed PMID: 11325958. doi:10.1074/jbc.M009003200

Hassannia, B., Vandenabeele, P., and Vanden Berghe, T. (2019). Targeting Ferroptosis to Iron Out Cancer. *Cancer Cell* 35 (6), 830–849. Epub 2019/05/21PubMed PMID: 31105042. doi:10.1016/j.ccell.2019.04.002

Honn, K. V., Tang, D. G., Gao, X., Butovich, I. A., Liu, B., Timar, J., et al. (1994). 12-lipoxygenases and 12(S)-HETE: Role in Cancer Metastasis. *Cancer Metast Rev.* 13 (3-4), 365–396. Epub 1994/12/01PubMed PMID: 7712597. doi:10.1007/BF00666105

Jiang, T., Zhou, B., Li, Y., Yang, Q., Tu, K., and Li, L. (2020). ALOX12B Promotes Carcinogenesis in Cervical Cancer by Regulating the PI3K/ERK1 Signaling Pathway. *Oncol. Lett.* 20 (2), 1360–1368. Epub 2020/07/30PubMed PMID: 32724378; PubMed Central PMCID: PMCPMC7377187. doi:10.3892/ol.2020.11641

Junttila, M. R., and Evan, G. I. (2009). p53 - a Jack of All Trades but Master of None. *Nat. Rev. Cancer* 9 (11), 821–829. Epub 2009/09/25PubMed PMID: 19776747. doi:10.1038/nrc2728

Li, D. S., Ainiwaer, J. L., Sheyhiding, I., Zhang, Z., and Zhang, L. W. (2016). Identification of Key Long Non-coding RNAs as Competing Endogenous RNAs for miRNA-mRNA in Lung Adenocarcinoma. *Eur. Rev. Med. Pharmacol. Sci.* 20 (11), 2285–2295. Epub 2016/06/25. PubMed PMID: 27338053.

Liang, C., Zhang, X., Yang, M., and Dong, X. (2019). Recent Progress in Ferroptosis Inducers for Cancer Therapy. *Adv. Mater.* 31 (51), 1904197, 2019 . Epub 2019/10/09PubMed PMID: 31595562. doi:10.1002/adma.201904197

Lim, W., Gee, K., Mishra, S., and Kumar, A. (2005). Regulation of B7.1 Costimulatory Molecule Is Mediated by the IFN Regulatory Factor-7 through the Activation of JNK in Lipopolysaccharide-Stimulated Human Monocytic Cells. *J. Immunol.* 175 (9), 5690–5700. Epub 2005/10/21PubMed PMID: 16237059. doi:10.4049/jimmunol.175.9.5690

Liu, H., Schreiber, S. L., and Stockwell, B. R. (2018). Targeting Dependency on the GPX4 Lipid Peroxide Repair Pathway for Cancer Therapy. *Biochemistry* 57 (14), 2059–2060. Epub 2018/03/28PubMed PMID: 29584411; PubMed Central PMCID:PMCPMC5962875. doi:10.1021/acs.biochem.8b00307

Liu, H. Y., and Zhang, C. J. (2017). Identification of Differentially Expressed Genes and Their Upstream Regulators in Colorectal Cancer. *Cancer Gene Ther.* 24 (6), 244–250. Epub 2017/04/15PubMed PMID: 28409560. doi:10.1038/cgt.2017.8

Matsumura, Y., Hiraoka, K., Ishikawa, K., Shoji, Y., Noji, T., Hontani, K., et al. (2016). CD40 Expression in Human Esophageal Squamous Cell Carcinoma Is

Associated with Tumor Progression and Lymph Node Metastasis. *Ar* 36 (9), 4467–4476. Epub 2016/09/16PubMed PMID: 27630283. doi:10.21873/anticanres.10991

Matsushima, K., Isomoto, H., Kohno, S., and Nakao, K. (2010). MicroRNAs and Esophageal Squamous Cell Carcinoma. *Digestion* 82 (3), 138–144. Epub 2010/07/01PubMed PMID: 20588024. doi:10.1159/000310918

Parri, M., Pietrovito, L., Grandi, A., Campagnoli, S., De Camilli, E., Bianchini, F., et al. (2014). Angiopoietin-like 7, a Novel Pro-angiogenetic Factor Over-expressed in Cancer. *Angiogenesis* 17 (4), 881–896. Epub 2014/06/07PubMed PMID: 24903490. doi:10.1007/s10456-014-9435-4

Pennathur, A., Gibson, M. K., Jobe, B. A., and Luketich, J. D. (2013). Oesophageal Carcinoma. *The Lancet* 381 (9864), 400–412. Epub 2013/02/05PubMed PMID: 23374478. doi:10.1016/S0140-6736(12)60643-6

Shen, Z., Song, J., Yung, B. C., Zhou, Z., Wu, A., and Chen, X. (2018). Emerging Strategies of Cancer Therapy Based on Ferroptosis. *Adv. Mater.* 30 (12), 1704007, 2018 , Epub 2018/01/23PubMed PMID: 29356212; PubMed Central PMCID: PMCPMC6377162. doi:10.1002/adma.201704007

Stockwell, B. R., Friedmann Angeli, J. P., Bayir, H., Bush, A. I., Conrad, M., Dixon, S. J., et al. (2017). Ferroptosis: A Regulated Cell Death Nexus Linking Metabolism, Redox Biology, and Disease. *Cell* 171 (2), 273–285. Epub 2017/10/07PubMed PMID: 28985560; PubMed Central PMCID: PMCPMC5685180. doi:10.1016/j.cell.2017.09.021

Tesfay, L., Paul, B. T., Konstorum, A., Deng, Z., Cox, A. O., Lee, J., et al. (2019). Stearoyl-CoA Desaturase 1 Protects Ovarian Cancer Cells from Ferroptotic Cell Death. *Cancer Res.* 79 (20), 5355–5366. Epub 2019/07/05PubMed PMID: 31270077; PubMed Central PMCID: PMCPMC6801059. doi:10.1158/0008-5472.CAN-19-0369

Wang, D., Peng, Y., Xie, Y., Zhou, B., Sun, X., Kang, R., et al. (2016). Antiferroptotic Activity of Non-oxidative Dopamine. *Biochem. Biophysical Res. Commun.* 480 (4), 602–607. Epub 2016/10/30PubMed PMID: 27793671. doi:10.1016/j.bbrc.2016.10.099

Wang, L., Wei, Q., Zhang, M., Chen, L., Li, Z., Zhou, C., et al. (2020). Identification of the Prognostic Value of Immune Gene Signature and Infiltrating Immune Cells for Esophageal Cancer Patients. *Int. Immunopharmacol.* 87, 106795, 2020 . Epub 2020/07/25PubMed PMID: 32707495. doi:10.1016/j.intimp.2020.106795

Wang, Q., Yang, L., Fan, Y., Tang, W., Sun, H., Xu, Z., et al. (2020). Circ-ZDHHC5 Accelerates Esophageal Squamous Cell Carcinoma Progression *In Vitro* via miR-217/ZEB1 Axis. *Front. Cel Dev. Biol.* 8, 570305, 2020 . Epub 2021/01/05PubMed PMID: 33392180; PubMed Central PMCID: PMCPMC7773775. doi:10.3389/fcell.2020.570305

Wang, Q., Ye, Y., Yu, H., Lin, S.-H., Tu, H., Liang, D., et al. (2020). Immune Checkpoint-Related Serum Proteins and Genetic Variants Predict Outcomes of Localized Prostate Cancer, a Cohort Study. *Cancer Immunol. Immunother.* 70, 701–712. Epub 2020/09/11PubMed PMID: 32909077. doi:10.1007/s00262-020-02718-1

Xie, J., Li, P., Gao, H.-F., Qian, J.-X., Yuan, L.-Y., and Wang, J.-J. (2014). Overexpression of SLC38A1 Is Associated with Poorer Prognosis in Chinese Patients with Gastric Cancer. *BMC Gastroenterol.* 14, 70, 2014 . Epub 2014/04/10PubMed PMID: 24712400; PubMed CentralPMCID: PMCPMC3984425. doi:10.1186/1471-230X-14-70

Yang, F., Zhang, Y., Ren, H., Wang, J., Shang, L., Liu, Y., et al. (2019). Ischemia Reperfusion Injury Promotes Recurrence of Hepatocellular Carcinoma in Fatty Liver via ALOX12-12hete-GPR31 Signaling axis. *J. Exp. Clin. Cancer Res.* 38 (1), 489, 2019 . Epub 2019/12/14PubMed PMID: 31831037; PubMed Central PMCID: PMCPMC6909624. doi:10.1186/s13046-019-1480-9

Yang, W. S., SriRamaratnam, R., Welsch, M. E., Shimada, K., Skouta, R., Viswanathan, V. S., et al. (2014). Regulation of Ferroptotic Cancer Cell Death by GPX4. *Cell* 156 (1-2), 317–331. Epub 2014/01/21PubMed PMID: 24439385; PubMed Central PMCID: PMCPMC4076414. doi:10.1016/j.cell.2013.12.010

Yoshimoto, T., Arakawa, T., Hada, T., Yamamoto, S., and Takahashi, E. (1992). Structure and Chromosomal Localization of Human Arachidonate 12-lipoxygenase Gene. *J. Biol. Chem.* 267 (34), 24805–24809. Epub 1992/12/05. PubMed PMID: 1447217. doi:10.1016/s0021-9258(18)35835-6

Zhao, Y., Xu, L., Wang, X., Niu, S., Chen, H., and Li, C. (2021). A Novel Prognostic mRNA/miRNA Signature for Esophageal Cancer and its Immune Landscape in Cancer Progression. *Mol. Oncol.* 15 (4), 1088–1109. Epub 2021/01/20PubMed PMID: 33463006; PubMed Central PMCID: PMCPMC8024720. doi:10.1002/1878-0261.12902

Zheng, Y., Yin, H., Boeglin, W. E., Elias, P. M., Crumrine, D., Beier, D. R., et al. (2011). Lipoxygenases Mediate the Effect of Essential Fatty Acid in Skin Barrier Formation. *J. Biol. Chem.* 286 (27), 24046–24056. Epub 2011/05/12PubMed PMID: 21558561; PubMed Central PMCID: PMCPMC3129186. doi:10.1074/jbc.M111.251496

Zheng, Z., Li, Y., Jin, G., Huang, T., Zou, M., and Duan, S. (2020). The Biological Role of Arachidonic Acid 12-lipoxygenase (ALOX12) in Various Human Diseases. *Biomed. Pharmacother.* 129, 110354, 2020 . Epub 2020/06/17PubMed PMID: 32540644. doi:10.1016/j.biopha.2020.110354

Zhou, F.-F., Xie, W., Chen, S.-Q., Wang, X.-K., Liu, Q., Pan, X.-K., et al. (2017). SLC38A1 Promotes Proliferation and Migration of Human Colorectal Cancer Cells. *J. Huazhong Univ. Sci. Technol. [Med. Sci.* 37 (1), 30–36. Epub 2017/02/23PubMed PMID: 28224429. doi:10.1007/s11596-017-1690-3

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 Molecular Biosciences

TECHNOLOGY AND CODE
published: 19 July 2021
doi: 10.3389/fmolb.2021.698334

BGM-Net: Boundary-Guided Multiscale Network for Breast Lesion Segmentation in Ultrasound

Yunzhu Wu[1†], Ruoxin Zhang[2†], Lei Zhu[3], Weiming Wang[4], Shengwen Wang[5,6], Haoran Xie[7], Gary Cheng[8], Fu Lee Wang[4], Xingxiang He[2] and Hai Zhang[1,9*]*

[1]Department of Ultrasound, Shenzhen People's Hospital, The Second Clinical College of Jinan University, Shenzhen, China, [2]Department of Gastroenterology, The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, China, [3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom, [4]School of Science and Technology, The Open University of Hong Kong, Hong Kong, China, [5]Department of Neurosurgery, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China, [6]Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China, [7]Department of Computing and Decision Sciences, Lingnan University, Hong Kong, China, [8]Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong, China, [9]The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen, China

OPEN ACCESS

Edited by:
William C. Cho,
Queen Elizabeth Hospital, China

Reviewed by:
Nawab Ali,
University of Arkansas at Little Rock, United States
Kenneth S. Hettie,
Stanford University, United States

***Correspondence:**
Xingxiang He
hexingxiang@gdpu.edu.cn
Hai Zhang
szzhhans.scc.jnu@foxmail.com

ORCID:
Hai Zhang
orcid.org/0000-0002-9018-1858

[†]*These authors share first authorship*

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *21 April 2021*
Accepted: *14 June 2021*
Published: *19 July 2021*

Citation:
Wu Y, Zhang R, Zhu L, Wang W, Wang S, Xie H, Cheng G, Wang FL, He X and Zhang H (2021) BGM-Net: Boundary-Guided Multiscale Network for Breast Lesion Segmentation in Ultrasound. Front. Mol. Biosci. 8:698334. doi: 10.3389/fmolb.2021.698334

Automatic and accurate segmentation of breast lesion regions from ultrasonography is an essential step for ultrasound-guided diagnosis and treatment. However, developing a desirable segmentation method is very difficult due to strong imaging artifacts e.g., speckle noise, low contrast and intensity inhomogeneity, in breast ultrasound images. To solve this problem, this paper proposes a novel boundary-guided multiscale network (BGM-Net) to boost the performance of breast lesion segmentation from ultrasound images based on the feature pyramid network (FPN). First, we develop a boundary-guided feature enhancement (BGFE) module to enhance the feature map for each FPN layer by learning a boundary map of breast lesion regions. The BGFE module improves the boundary detection capability of the FPN framework so that weak boundaries in ambiguous regions can be correctly identified. Second, we design a multiscale scheme to leverage the information from different image scales in order to tackle ultrasound artifacts. Specifically, we downsample each testing image into a coarse counterpart, and both the testing image and its coarse counterpart are input into BGM-Net to predict a fine and a coarse segmentation maps, respectively. The segmentation result is then produced by fusing the fine and the coarse segmentation maps so that breast lesion regions are accurately segmented from ultrasound images and false detections are effectively removed attributing to boundary feature enhancement and multiscale image information. We validate the performance of the proposed approach on two challenging breast ultrasound datasets, and experimental results demonstrate that our approach outperforms state-of-the-art methods.

Keywords: breast lesion segmentation, boundary-guided feature enhancement, multiscale image analysis, ultrasound image segmentation, deep learning

1 INTRODUCTION

Breast cancer is the most commonly occurring cancer in women and is also the second leading cause of cancer death Siegel et al. (2017). Ultrasonography has been an attractive imaging modality for the detection and analysis of breast lesions because of its various advantages, e.g., safety, flexibility and versatility Stavros et al. (1995). However, clinical diagnosis of breast lesions based on ultrasound imaging generally requires well-trained and experienced radiologists as ultrasound images are hard to interpret and quantitative measurements of breast lesion regions are tedious and difficult tasks. Thus, automatic localization of breast lesion regions will facilitate the process of clinical detection and analysis, making the diagnosis more efficient, as well as achieving higher sensitivity and specificity Yap et al. (2018). Unfortunately, accurate breast lesion segmentation from ultrasound images is very challenging due to strong imaging artifacts, e.g., speckle noise, low contrast and intensity inhomogeneity. Please refer to **Figure 1** for some ultrasound samples.

To solve this problem, we propose a boundary-guided multiscale network (BGM-Net) to boost the performance of breast lesion segmentation from ultrasound images based on the feature pyramid network (FPN) Lin et al. (2017). Specifically, we first develop a boundary-guided feature enhancement (BGFE) module to enhance the feature map for each FPN layer by learning a boundary map of breast lesion regions. This step is particularly important for the performance of the proposed network because it improves the capability of the FPN framework to detect the boundaries of breast lesion regions in low contrast ultrasound images, eliminating boundary leakages in ambiguous regions. Then, we design a multiscale scheme to leverage the information from different image scales in order to tackle ultrsound artifacts, where the segmentation result is produced by fusing a fine and a coarse segmentation maps predicted from the testing image and its coarse counterpart, respectively. The multiscale scheme can effectively remove false detections that result from strong imaging artifacts. We demonstrate the superiority of the proposed network over state-of-the-art methods on two challenging breast ultrasound datasets.

2 RELATED WORK

In the literature, algorithms for breast lesion segmentation from ultrasound images have been extensively studied. Early methods Boukerroui et al. (1998), Madabhushi and Metaxas (2002), Madabhushi and Metaxas (2003), Shan et al. (2008), Shan et al. (2012), Xian et al. (2015), Gómez-Flores and Ruiz-Ortega (2016) mainly exploit hand-crafted features to construct segmentation models to infer the boundaries of breast lesion regions, and can be divided into three categories according to Xian et al. (2018), including region growing methods Kwak et al. (2005), Shan et al. (2008), Shan et al. (2012) deformable models Yezzi et al. (1997), Chen et al. (2002), Chang et al. (2003), Madabhushi and Metaxas (2003), Gao et al. (2012), and graph models Ashton and Parker (1995), Chiang et al. (2010), Xian et al. (2015).

Region growing methods start the segmentation from a set of manual or automatic selected seeds, which gradually expand to capture the boundaries of target regions according to the predefined growing criteria. Shan et al. Shan et al. (2012) developed an efficient mehtod to automatically generate region-of-interest (ROI) for breast lesion segmentation, while Kwak et al. Kwak et al. (2005) utilized common contour smoothness and region similarity (mean intensity and size) to define the growing criteria.

Deformable models first construct an initial model and then deform the model to reach object boundaries according to internal and external energies. Madabhushi et al. Madabhushi and Metaxas (2003) initialized the deformable model using boundary points and employed balloon forces to define the extern energy, while Chang et al. Chang et al. (2003) applied the stick filter to reduce speckle noise in ultrasound images before deforming the model to segment breast lesion regions.

Graph models perform breast lesion segmentation with efficient energy optimization by using Markov random field or graph cut framework. Chiang et al. Chiang et al. (2010) employed a pre-trained Probabilistic Boosting Tree (PBT) classifier to determine the data term of the graph cut energy, while Xian et al. Xian et al. (2015) formulated the energy function by modeling the information from both frequency and space domains. Although many a priori models haved been designed to assist breast lesion segmentation, these methods have limited capability to capture high-level semantic features in order to identify weak boundaries in ambiguous regions, leading to boundary leakages in low contrast ultrasound images.

In contrast, Learning-based methods utilize a set of manually designed features to train the classifier for segmentation tasks Huang et al. (2008), Lo et al. (2014), Moon et al. (2014), Othman and Tizhoosh (2011). Liu et al. Liu et al. (2010) extracted 18 local image features to train a SVM classifier to segment breast lesion regions, and Jiang et al. Jiang et al. (2012) utilized 24 Harr-like features and trained Adaboost classifier for breast tumor segmentation. Recently, convolution neural networks (CNNs) have been demonstrated to achieve excellent performance in a lot of medical applications by building a series of deep convolutional layers to learn high-level semantic features from labeled data. Inspired from this, several CNN frameworks Yap et al. (2018), Xu et al. (2019) have been developed to segment breast lesion regions from ultrasound images. For example, Yap et al. Yap et al. (2017) investigated the performance of three networks: a Patch-based LeNet, a U-Net, and a transfer learning approach with a pretrained FCN-AlexNet, for breast lesion detection. Lei et al. Lei et al. (2018) proposed a deep convolutional encoder-decoder network equipped with deep boundary supervision and adaptive domain transfer for the segmentation of breast anatomical layers. Hu et al. Hu et al. (2019) combined a dilated fully convolutional network with an active contour model to segment breast tumors. Although CNN-based methods improve the performance of breast lesion segmentation in low contrast ultrasound images, they still suffer from strong artifacts of speckle noise and intensity

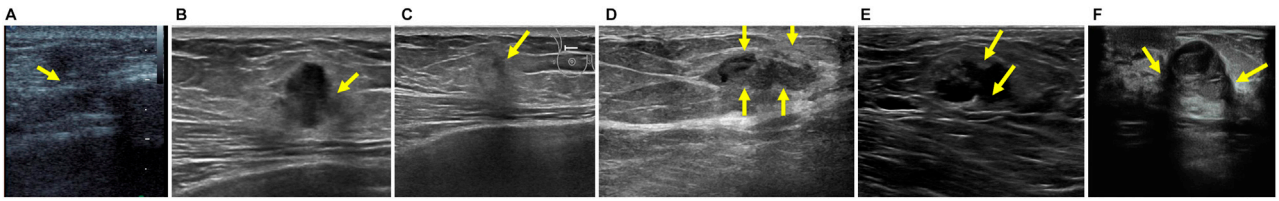


FIGURE 1 | Examples of breast ultrasound images. **(A–C)** Ambiguous boundaries due to similar appearance between lesion and non-lesion regions. **(D–F)** Intensity inhomogeneity inside lesion regions. Note that the green arrows are marked by radiologists.

inhomogeneity, which typically occur in clinical scenarios, and tend to generate inaccurate segmentation results.

3 OUR APPROACH

3.1 Overview

Figure 2 illustrates the architecture of the proposed approach. Given a testing breast ultrasound image I, we first downsample I into a coarse counterpart J, and then input both I and J into the feature pyramid network to obtain a set of feature maps with different spatial resolutions. After that, a boundary-guided feature enhancement module is developed to enhance the feature map for each FPN layer by learning a boundary map of breast lesion regions. All of the refined feature maps are then upsampled and concatenated to predict a fine S_I and a coarse S_J segmentation maps for I and J, respectively. Finally, the segmentation result S_f is produced by fusing S_I and S_J so as to leverage the information from different image scales. By combining enhanced boundary features and multiscale image information into a unified framework, our approach precisely segments the breast lesion regions from ultrasound images and effectively removes false detections resulting from various imaging artifacts.

3.2 Boundary-Guided Feature Enhancement

The FPN framework first uses a convolutional neural network to extract a set of feature maps with different spatial resolutions and then iteratively merges two adjacent layers from the last layer to the first layer. Although FPN improves the performance of breast lesion segmentation, it still suffers from the inaccuracy of boundary detection because of strong ultrasound artifacts. To solve this problem, we develop a boundary-guided feature enhancement module to improve the boundary detection capability of the feature map for each FPN layer by learning a boundary map of breast lesion regions.

Figure 3 shows the flowchart of the BGFE module. Given a feature map F, we first apply a 3×3 convolutional layer on F to obtain the first intermediate image X, followed by a 1×1 convolutional layer to obtain the second intermediate image Y, which will be used to learn a boundary map B of breast lesion regions. Then, we apply a 3×3 convolutional layer on Y to obtain the third intermediate image Z, and multiply each channel of Z with B in an element-wise manner. Finally, we concatenate X and Z, followed by a 1×1 convolutional layer, to obtain the enhanced feature map $\widehat{\boldsymbol{F}}$. Mathematically, the cth channel of $\widehat{\boldsymbol{F}}$ is computed as:

$$\widehat{\boldsymbol{F}}_c = f_{conv}\left(concate\left(\left(Z_c \times B\right), X\right)\right), \tag{1}$$

where f_{conv} is the 1×1 convolutional parameter; Z_c is the cth channel of Z; and *concate* is the concatenation operation on the feature map.

3.3 Multiscale Scheme

After the BGFE module, all of the refined feature maps will be upsampled and concatenated to predict the segmentation map of the input image. To account for various ultrasound artifacts, we design a multiscale scheme to produce the final segmentation result by fusing the information from different image scales. Specifically, for each testing breast ultrasound image, we first downsample it into a coarse counterpart with the resolution of 320×320. In our experiment, the training images are all resized to the resolution of 416×416 according to previous experience, and thus the testing image is also resized to the same resolution. Then, both the testing image and its coarse counterpart are input into the proposed network to predict a fine and a coarse segmentation maps, respectively. Finally, the segmentation result is produced by fusing the fine and the coarse segmentation maps so that false detections from the fine scale can be counteracted by the information from the coarse scale, leading to an accurate segmentation of breast lesion regions.

3.4 Loss Fuction

In our study, there is an annotated mask of breast lesion regions for each training image, which will serve as the ground true for breast lesion segmentation. In addition, we employ a canny detector Canny (1986) on the annotated mask to obtain a boundary map of breast lesion regions, which will serve as the ground true for boundary detection. Based on the two ground truths, we combine a segmentation loss and a boundary detection loss to compute the total loss function $\mathcal{L}$ as following:

$$\mathcal{L} = \mathcal{D}_{seg} + \alpha\mathcal{D}_{edge}, \tag{2}$$

where $\mathcal{D}_{seg}$ and $\mathcal{D}_{edge}$ are the segmentation loss and the boundary detection loss, respectively. α is used to balance $\mathcal{D}_{seg}$ and $\mathcal{D}_{edge}$, and is empirically set to 0.1. The definitions of $\mathcal{D}_{seg}$ and $\mathcal{D}_{edge}$ are given by:

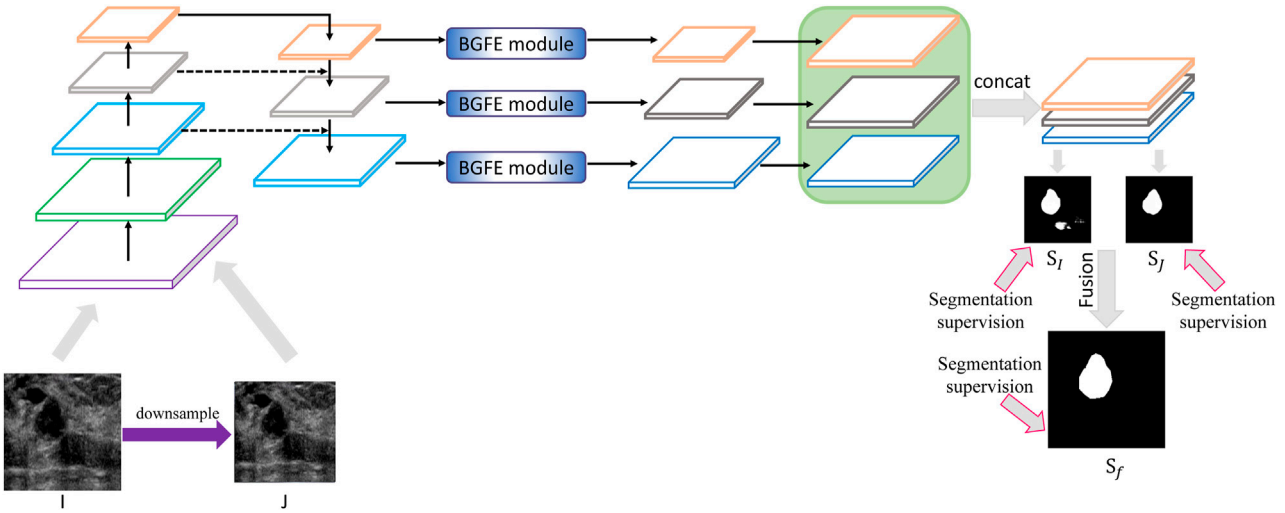


FIGURE 2 | Schematic illustration of the proposed approach for breast lesion segmentation from ultrasound images. Please refer to **Figure 3** for BGFE module. Best viewed in color.

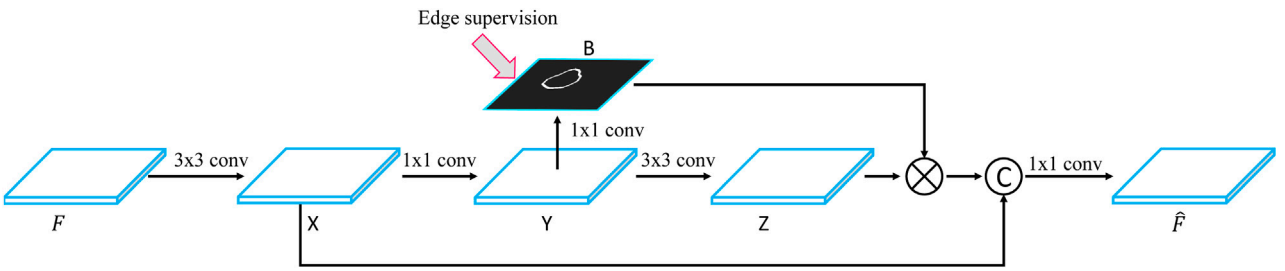


FIGURE 3 | Flowchart of the BGFE module. **F** and $\hat{\boldsymbol{F}}$ are the feature map and the refined feature map, respectively. Best viewed in color.

$$\mathcal{D}_{seg} = \hat{\Phi}\left(S_I, G_s\right) + \hat{\Phi}\left(S_J, G_s\right) + \hat{\Phi}\left(S_f, G_s\right), \quad and$$
$$\mathcal{D}_{edge} = \sum_{k=1}^{3} \hat{\Phi}\left(B_k, G_e\right), \tag{3}$$

where G_s and G_e are the ground truths for breast lesion segmentation and boundary detection, respectively. S_I and S_J are the segmentation maps of I and J, respectively, and S_f is the final segmentation result. B_k is the predicted boundary map of breast lesion regions at the kth BGFE module. The function $\hat{\Phi}$ includes a dice loss and a cross entropy loss, and is defined as:

$$\hat{\Phi} = \Phi_{CE} + \beta\Phi_{dice}, \tag{4}$$

where Φ_{CE} and Φ_{dice} are the functions of the cross entropy loss and the dice loss, respectively. β is used to balance Φ_{CE} and Φ_{dice}, and is empirically set to 0.5.

3.5 Training and Testing Strategies

Training Parameters

We initialize the parameters of the basic convolutional neural network by a pre-trained DenseNet-121 Huang et al. (2017) on ImageNet while the others are trained from scratch noise. The breast ultrasound images in our training dataset are randomly rotated, cropped, and horizontally flipped for data augmentation. We use Adam optimizer to train the whole framework by 10, 000 iterations. The learning rate is initialized as 0.0001 and reduced to 0.00001 after 5, 000 iterations. We implement our BGM-Net on Keras and run it on a single GPU with a mini-batch size of 8.

Inference

We take S_f as the final segmentation result for each testing image.

4 EXPERIMENTS

This section conducts extensive experiments, as well as an ablation study, to evaluate the performance of the proposed approach for breast lesion segmentation from ultrasound images.

4.1 Dataset

Two challenging breast ultrasound datasets are utilized for the evaluation. The first dataset (i.e., Al-Dhabyani et al., 2020) is from

TABLE 1 | Measurement results of different segmentation methods on the BUSZPH dataset. Our results are highlighted in bold.

Method	Dice	ADB	Jaccard	Precision	Recall
U-Net Ronneberger et al. (2015)	0.7819	15.6556	0.6990	0.8055	0.8429
U-Net++ Zhou et al. (2018)	0.7895	11.3389	0.7092	0.8408	0.8029
FPN Lin et al. (2017)	0.8597	5.6913	0.7829	0.9001	0.8518
DeeplabV3+ Chen et al. (2018)	0.8418	6.6364	0.7583	0.8870	0.8289
ConvEDNet Lei et al. (2018)	0.8368	5.7943	0.7540	0.8987	0.8249
Our approach	**0.8688**	**4.7966**	**0.7961**	**0.9080**	**0.8603**

TABLE 2 | Measurement results of different segmentation methods on the BUSI dataset. Our results are highlighted in bold.

Method	Dice	ADB	Jaccard	Precision	Recall
U-Net Ronneberger et al. (2015)	0.7696	33.4737	0.6777	0.8451	0.7833
U-Net++ Zhou et al. (2018)	0.7622	30.6443	0.6685	0.8222	0.7861
FPN Lin et al. (2017)	0.8267	16.6268	0.7409	0.8479	0.8539
DeeplabV3+ Chen et al. (2018)	0.8268	16.2611	0.7348	0.8720	0.8337
ConvEDNet Lei et al. (2018)	0.8270	17.3333	0.7357	0.8490	0.8551
Our approach	**0.8397**	**12.5637**	**0.7597**	**0.8931**	**0.8345**

the Baheya Hospital for Early Detection and Treatment of Womenś Cancer (Cairo, Egypt). BUSI includes 780 tumor images from 600 patients. We randomly select 661 images as the training dataset and the remaining 119 images serve as the testing dataset. The second dataset includes 632 breast ultrasound images (denoted as BUSZPH), collected from Shenzhen People's Hospital where informed consent is obtained from all patients. We randomly select 500 images as the training dataset and the remaining 132 images serve as the testing dataset. The breast lesion regions in all the images are manually segmented by experienced radiologies, and each annotation result is confirmed by three clinicians.

4.2 Evaluation Metric

We adopt five widely used metrics for quantitative comparison, including Dice Similarity Coefficient (Dice), Average Distance between Boundaries (ADB, in pixel), Jaccard, Precision, and Recall. Please refer to Chang et al. (2009), Wang et al. (2018) for more details about these metrics. Dice and Jaccard measure the similarity between the segmentation result and the ground truth. ADB measures the pixel distance between the boundaries of the segmentation result and the ground truth. Precision and Recall compute pixel-wise classification accuracy to evaluate the segmentation result. Overall, a good segmentation result shall have a low ADB value, but high values for the other four metrics.

4.3 Segmentation Performance

Comparison Methods

We validate the proposed approach by comparing it with five state-of-the-art methods, including U-Net Ronneberger et al. (2015), U-Net++ Zhou et al. (2018), feature pyramid network (FPN) Lin et al. (2017), DeeplabV3+ Chen et al. (2018) and ConvEDNet Lei et al. (2018). For consistent comparison, we obtain the segmentation results of the five methods by the public code (if available) or by our implementation, which is tuned for the best result.

Quantitative Comparison

Tables 1, **2** present the measurement results of different segmentation methods on the two datasets, respectively. Apparently, our approach achieves higher values on Dice, Jaccard, Precision and Recall measurements, and lower value on ADB measurement, demenstrating the high accuracy of the proposed approach for breast lesion segmentation from ultrasound images.

Visual Comparison

Figure 4 visually compares the segmentation results obtained by our approach and the other five segmentation methods. As shown in the figure, our approach precisely segments the breast lesion regions from ultrasound images despite of sevious artifacts, while the other methods tend to generate over or under-segmentation results as they wrongly classify some non-lesion regions or miss parts of lesion regions. In the first and second rows where high speckle noise is presented, our result shows the highest similarity against the ground true. This is because the boundary detection loss in our loss function explicitly regularizes the boundary shape of the detected regions using the boundary information in the ground true. In addition, non-lesion regions are greatly removed even though there are ambiguous regions with weak boundaries, see the third and fourth rows, since the multiscale shceme in our approach effectively fuses the information from different image scales. Moreover, our approach accurately locate the boundaries of breast lesion regions in inhomogeneous ultrasound images attributing to the boundary feature enhancement of the BGFE module, see the fifth and sixth rows. In contrast, segmentation results from the other methods are inferior as these methods have limited capability to cope with strong ultrasound artifacts.

4.4 Ablation Study

Network Design

We conduct an ablation study to evaluate the key components of the proposed approach. Specifically, three baseline networks are considered and their quantitative results on the two datasets are

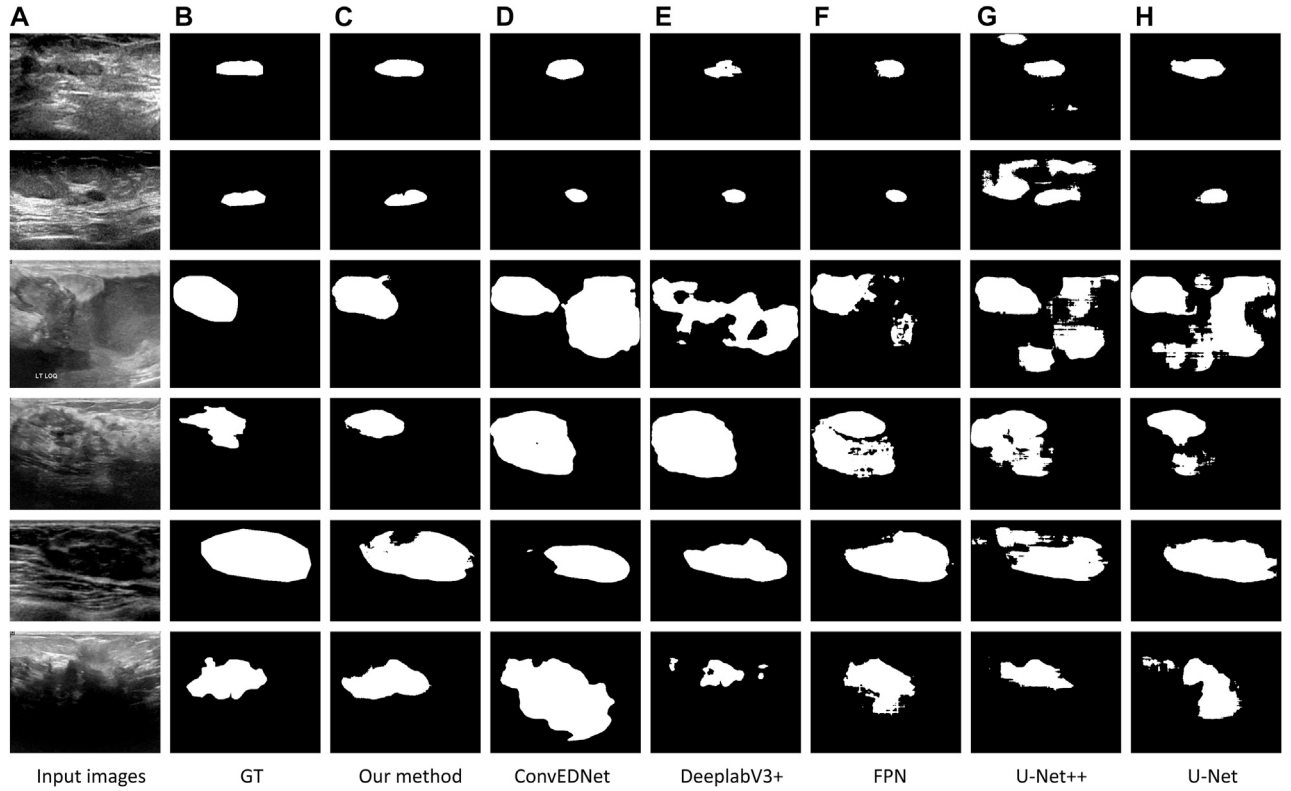


FIGURE 4 | Comparison of breast lesion segmentation among different methods. **(A)** Testing images. **(B)** Ground truth (denoted as GT). **(C–H)**: Segmentation results obtained by our approach (BGM-Net), ConvEDNet Lei et al. (2018), DeeplabV3+ Chen et al. (2018), FPN Lin et al. (2017), U-Net++ Zhou et al. (2018), and U-Net Ronneberger et al. (2015), respectively. Note that the images in first three rows are from BUSZPH, while the images in last three rows are from BUSI.

TABLE 3 | Measurement results of different baseline networks on the BUSZPH dataset. Our results are highlighted in bold.

Method	Dice	ADB	Jaccard	Precision	Recall
Basic	0.8496	6.9231	0.7665	0.8840	0.8553
Basic + Multiscale	0.8578	6.3899	0.7816	0.8853	0.8600
Basic + BGFE	0.8619	6.1084	0.7855	0.9006	0.8602
Our approach	**0.8688**	**4.7966**	**0.7961**	**0.9080**	**0.8603**

TABLE 4 | Measurement results of different baseline networks on the BUSI dataset. Our results are highlighted in bold.

Method	Dice	ADB	Jaccard	Precision	Recall
Basic	0.8158	13.9902	0.7325	0.8641	0.8253
Basic + Multiscale	0.8246	16.6773	0.7385	0.8831	0.8117
Basic + BGFE	0.8300	12.4873	0.7503	0.8669	0.8329
Our approach	**0.8397**	**12.5637**	**0.7597**	**0.8931**	**0.8345**

reported in comparison with our approach. The first baseline network (denoted as "Basic") removes both the BGFE modules and multiscale scheme from our approach, meaning that both boundary feature enhancement and multiscale fusing are disabled and the proposed approach degrades to the FPN framework. The second baseline network (denoted as "Basic + Multiscale") removes the BGFE modules from our approach, meaning that boundary feature enhancement is disabled while multiscale fusing is enabled. The third baseline network (denoted as "Basic + BGFE") removes the multiscale scheme from our approach, meaning that multiscale fusing is disabled while boundary feature enhancement is enabled.

Quantitative Comparison

Tables 3, **4** present the measurement results of different baseline networks on the two datasets, respectively. As shown in the table, both "Basic + BGFE" and "Basic + Multiscale" perform better than "Basic" by showing higher values on Dice, Jaccard, Precision and Recall measurements, but a lower value on ADB measurement. This clearly demonstrates the benifits from the FPN module and the multiscale scheme. In addition, our approach achieves the best result compared with the three baseline networks, which validates the superiority of the proposed approach by combining boundary feature enhancement and multiscale fusing into a unified framework.

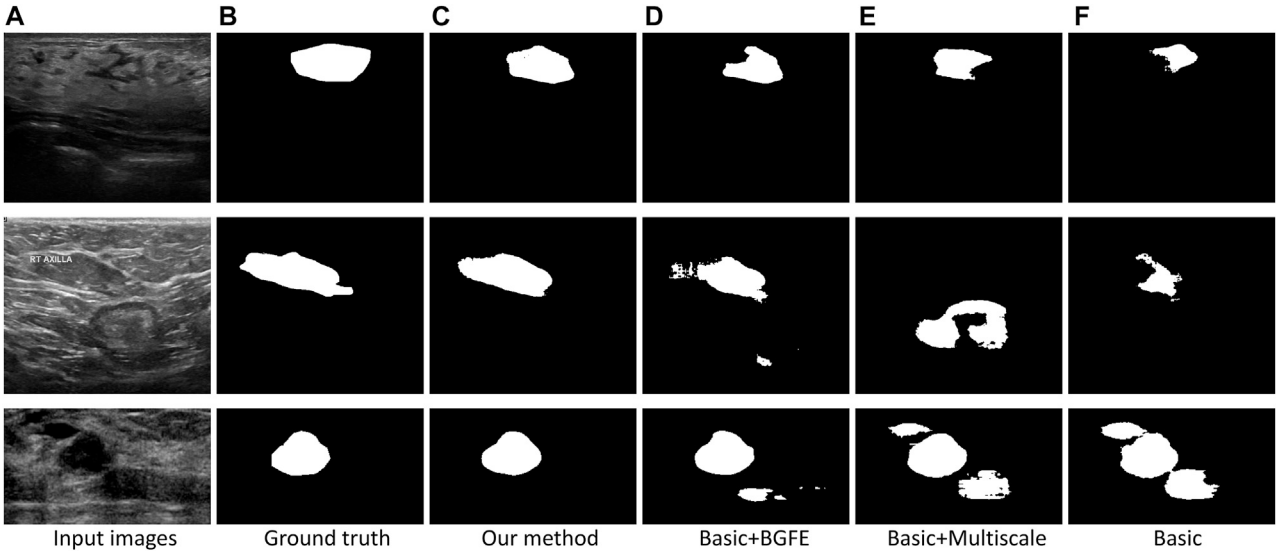


FIGURE 5 | Comparison of breast lesion segmentation between our approach **(C)** and the three baseline networks **(D–F)** against the ground truth **(B)**.

Visual Comparison

Figure 5 visually compares the segmentation results obtained by our approach and the three baseline networks. Apparently, our approach better segments breast lesion regions than the three baseline networks. False detections resulted from speckle noise are observed in the result of "Basic + BGFE", while "Basic + Multiscale" wrongly classifies a large part of non-lesion regions due to unclear boundaries in ambiguous regions. In contrast, our approach accurately locates the boundaries of breast lesion regions by learning an enhanced boundary map using the BGFE module. Moreover, false detections are effectively removed attributing to the multiscale scheme. Thus, our result achieves the highest similarity against the ground true.

5 CONCLUSION

This paper proposes a novel boundary-guided multiscale network to boost the performance of breast lesion segmentation from ultrasound images based on the FPN framework. By combining boundary feature enhancement and multiscale image information into a unified framework, the boundary detection capability of the FPN framework is greatly improved so that weak boundaries in ambiguous regions can be correctly identified. In addition, the segmentation accuracy is notably increased as false detections resulted from strong ultrasound artifacts are effectively removed attributing to the multiscale scheme. Experimental results on two challenging breast ultrasound datasets demonstrate the superiority of our approach compared with state-of-the-art methods. However, similar to previous work, our approach also relies on labeled data to train the network, which limits its applications in scenarios where unlabeled data is presented. Thus, the future work will consider the adaptation from labeled data to unlabeled data in order to improve the generalization of the proposed approach.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation. The data presented in the study are deposited in https://sites.google.com/site/indexlzhu/datasets.

AUTHOR CONTRIBUTIONS

YW, RZ, LZ, and WW designed the convolutional neural network and draft the paper. SW, XH, and HZ collected the data and pre-processed the original data. HX, GC, and FL prepared the quantitative and qualitative comparisons and revised the paper.

ACKNOWLEDGMENTS

This paper was supported by Natural Science Foundation of Shenzhen city (No. JCYJ20190806150001764), Natural Science Foundation of Guangdong province (No. 2020A1515010978), The Sanming Project of Medicine in Shenzhen training project (No. SYJY201802), National Natural Science Foundation of China (No. 61802072), General Research Fund (No. 18601118) of Research Grants Council of Hong Kong SAR, One-off Special Fund from Central and Faculty Fund in Support of Research from 2019/20 to 2021/22 (MIT02/19-20), Research Cluster Fund (RG 78/2019-2020R), Dean's Research Fund 2019/20 (FLASS/DRF/IDS-2) of The Education University of Hong Kong, and the Faculty Research Grant (DB21B6) of Lingnan University, Hong Kong.

REFERENCES

Al-Dhabyani, W., Gomaa, M., Khaled, H., and Fahmy, A. (2020). Dataset of Breast Ultrasound Images. *Data in Brief* 28, 104863.

Ashton, E. A., and Parker, K. J. (1995). Multiple Resolution Bayesian Segmentation of Ultrasound Images. *Ultrason. Imaging* 17, 291–304. doi:10.1177/016173469501700403

Boukerroui, D., Basset, O., Guérin, N., and Baskurt, A. (1998). Multiresolution Texture Based Adaptive Clustering Algorithm for Breast Lesion Segmentation. *Eur. J. Ultrasound* 8, 135–144. doi:10.1016/s0929-8266(98)00062-7

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-8, 679–698. doi:10.1109/tpami.1986.4767851

Chang, H.-H., Zhuang, A. H., Valentino, D. J., and Chu, W.-C. (2009). Performance Measure Characterization for Evaluating Neuroimage Segmentation Algorithms. *Neuroimage* 47, 122–135. doi:10.1016/j.neuroimage.2009.03.068

Chang, R.-F., Wu, W.-J., Moon, W. K., Chen, W.-M., Lee, W., and Chen, D.-R. (2003). Segmentation of Breast Tumor in Three-Dimensional Ultrasound Images Using Three-Dimensional Discrete Active Contour Model. *Ultrasound Med. Biol.* 29, 1571–1581. doi:10.1016/s0301-5629(03)00992-x

Chen, C.-M., Lu, H. H.-S., and Huang, Y.-S. (2002). Cell-based Dual Snake Model: a New Approach to Extracting Highly Winding Boundaries in the Ultrasound Images. *Ultrasound Med. Biol.* 28, 1061–1073. doi:10.1016/s0301-5629(02)00531-8

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *ECCV* (Springer), 801–818.

Chiang, H.-H., Cheng, J.-Z., Hung, P.-K., Liu, C.-Y., Chung, C.-H., and Chen, C.-M. (2010). "Cell-based Graph Cut for Segmentation of 2d/3d Sonographic Breast Images," in *ISBI* (IEEE), 177–180.

Gao, L., Liu, X., and Chen, W. (2012). Phase-and Gvf-Based Level Set Segmentation of Ultrasonic Breast Tumors. *J. Appl. Math.* 2012. doi:10.1155/2012/810805

Gómez-Flores, W., and Ruiz-Ortega, B. A. (2016). New Fully Automated Method for Segmentation of Breast Lesions on Ultrasound Based on Texture Analysis. *Ultrasound Med. Biol.* 42, 1637–1650. doi:10.1016/j.ultrasmedbio.2016.02.016

Hu, Y., Guo, Y., Wang, Y., Yu, J., Li, J., Zhou, S., et al. (2019). Automatic Tumor Segmentation in Breast Ultrasound Images Using a Dilated Fully Convolutional Network Combined with an Active Contour Model. *Med. Phys.* 46, 215–228. doi:10.1002/mp.13268

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely Connected Convolutional Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4700–4708.

Huang, S.-F., Chen, Y.-C., and Moon, W. K. (2008). "Neural Network Analysis Applied to Tumor Segmentation on 3d Breast Ultrasound Images," in IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 1303–1306.

Jiang, P., Peng, J., Zhang, G., Cheng, E., Megalooikonomou, V., and Ling, H. (2012). "Learning-based Automatic Breast Tumor Detection and Segmentation in Ultrasound Images," in *ISBI* (IEEE), 1587–1590.

Kwak, J. I., Kim, S. H., and Kim, N. C. (2005). "Rd-based Seeded Region Growing for Extraction of Breast Tumor in an Ultrasound Volume," in International Conference on Computational and Information Science (Springer), 799–808. doi:10.1007/11596448_118

Lei, B., Huang, S., Li, R., Bian, C., Li, H., Chou, Y.-H., et al. (2018). Segmentation of Breast Anatomy for Automated Whole Breast Ultrasound Images with Boundary Regularized Convolutional Encoder-Decoder Network. *Neurocomputing* 321, 178–186. doi:10.1016/j.neucom.2018.09.043

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature Pyramid Networks for Object Detection," in *CVPR*, 2117–2125.

Liu, B., Cheng, H. D., Huang, J., Tian, J., Tang, X., and Liu, J. (2010). Fully Automatic and Segmentation-Robust Classification of Breast Tumors Based on Local Texture Analysis of Ultrasound Images. *Pattern Recognition* 43, 280–298. doi:10.1016/j.patcog.2009.06.002

Lo, C., Shen, Y.-W., Huang, C.-S., and Chang, R.-F. (2014). Computer-aided Multiview Tumor Detection for Automated Whole Breast Ultrasound. *Ultrason. Imaging* 36, 3–17. doi:10.1177/0161734613507240

Madabhushi, A., and Metaxas, D. (2002). "Automatic Boundary Extraction of Ultrasonic Breast Lesions," in Proceedings IEEE International Symposium on Biomedical Imaging, 601–604.

Madabhushi, A., and Metaxas, D. N. (2003). Combining Low-, High-Level and Empirical Domain Knowledge for Automated Segmentation of Ultrasonic Breast Lesions. *IEEE Trans. Med. Imaging* 22, 155–169. doi:10.1109/tmi.2002.808364

Moon, W. K., Lo, C.-M., Chen, R.-T., Shen, Y.-W., Chang, J. M., Huang, C.-S., et al. (2014). Tumor Detection in Automated Breast Ultrasound Images Using Quantitative Tissue Clustering. *Med. Phys.* 41, 042901. doi:10.1118/1.4869264

Othman, A. A., and Tizhoosh, H. R. (2011). "Segmentation of Breast Ultrasound Images Using Neural Networks," in *Engineering Applications of Neural Networks* (Springer), 260–269. doi:10.1007/978-3-642-23957-1_30

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: Convolutional Networks for Biomedical Image Segmentation," in International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) (Springer), 234–241. doi:10.1007/978-3-319-24574-4_28

Shan, J., Cheng, H.-D., and Wang, Y. (2008). "A Novel Automatic Seed point Selection Algorithm for Breast Ultrasound Images," in International Conference on Pattern Recognition, 1–4.

Shan, J., Cheng, H. D., and Wang, Y. (2012). Completely Automated Segmentation Approach for Breast Ultrasound Images Using Multiple-Domain Features. *Ultrasound Med. Biol.* 38, 262–275. doi:10.1016/j.ultrasmedbio.2011.10.022

Siegel, R. L., Miller, K. D., and Jemal, A. (2017). Cancer Statistics, 2017. *CA: A Cancer J. Clinicians* 67, 7–30. doi:10.3322/caac.21387

Stavros, A. T., Thickman, D., Rapp, C. L., Dennis, M. A., Parker, S. H., and Sisney, G. A. (1995). Solid Breast Nodules: Use of Sonography to Distinguish between Benign and Malignant Lesions. *Radiology* 196, 123–134. doi:10.1148/radiology.196.1.7784555

Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., et al. (2018). "Deep Attentional Features for Prostate Segmentation in Ultrasound," in *MICCAI* (Springer), 523–530. doi:10.1007/978-3-030-00937-3_60

Xian, M., Zhang, Y., Cheng, H.-D., Xu, F., Huang, K., Zhang, B., et al. (2018). *A Benchmark for Breast Ultrasound Image Segmentation (BUSIS)*. arXiv: 1801.03182.

Xian, M., Zhang, Y., and Cheng, H. D. (2015). Fully Automatic Segmentation of Breast Ultrasound Images Based on Breast Characteristics in Space and Frequency Domains. *Pattern Recognition* 48, 485–497. doi:10.1016/j.patcog.2014.07.026

Xu, Y., Wang, Y., Yuan, J., Cheng, Q., Wang, X., and Carson, P. L. (2019). Medical Breast Ultrasound Image Segmentation by Machine Learning. *Ultrasonics* 91, 1–9. doi:10.1016/j.ultras.2018.07.006

Yap, M. H., Pons, G., Martí, J., Ganau, S., Sentís, M., Zwiggelaar, R., et al. (2018). Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J. Biomed. Health Inform.* 22, 1218–1226. doi:10.1109/JBHI.2017.2731873

Yap, M. H., Pons, G., Martí, J., Ganau, S., Sentís, M., Zwiggelaar, R., et al. (2018). Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks. *IEEE J. Biomed. Health Inform.* 22, 1218–1226. doi:10.1109/jbhi.2017.2731873

Yezzi, A., Kichenassamy, S., Kumar, A., Olver, P., and Tannenbaum, A. (1997). A Geometric Snake Model for Segmentation of Medical Imagery. *IEEE Trans. Med. Imaging* 16, 199–209. doi:10.1109/42.563665

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). "UNet++: A Nested U-Net Architecture for Medical Image Segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer), 3–11. doi:10.1007/978-3-030-00889-5_1

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

frontiers in Molecular Biosciences

REVIEW
published: 30 July 2021
doi: 10.3389/fmolb.2021.699906

Single-Cell RNA Sequencing of Retina: New Looks for Gene Marker and Old Diseases

Peixi Ying[1†], Chang Huang[2,3,4,5†], Yan Wang[6†], Xi Guo[7†], Yuchen Cao[1], Yuxi Zhang[1], Sheng Fu[8], Lin Chen[9], Guoguo Yi[10] and Min Fu[11]**

[1]The Second Clinical School, Southern Medical University, Guangzhou, China, [2]Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University, Shanghai, China, [3]NHC Key Laboratory of Myopia, Fudan University, Shanghai, China, [4]Key Laboratory of Myopia, Chinese Academy of Medical Sciences, Shanghai, China, [5]Shanghai Key Laboratory of Visual Impairment and Restoration, Shanghai, China, [6]Department of Ophthalmology, South China Hospital, Health Science Center, Shenzhen University, Shenzhen, China, [7]Medical College of Rehabiliation, Southern Medical University, Guangzhou, China, [8]The University of South China, Hengyang, China, [9]Department of Anesthesiology, Shenzhen Hospital, Southern Medical University, Shenzhen, China, [10]Department of Ophthalmology, the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [11]Department of Ophthalmology, Zhujiang Hospital, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:
Fengfeng Zhou, Jilin University, China

Reviewed by:
Siddharth Shukla, Howard Hughes Medical Institute (HHMI), United States
Sumit Mukherjee, Microsoft, United States

***Correspondence:**
Guoguo Yi
yigg@mail.sysu.edu.cn
Min Fu
min_fu1212@163.com

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

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *24 April 2021*
Accepted: *01 July 2021*
Published: *30 July 2021*

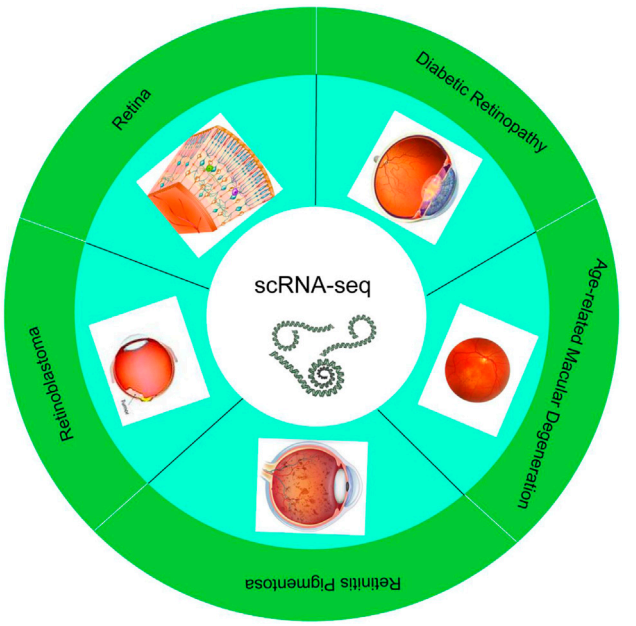
Citation:
Ying P, Huang C, Wang Y, Guo X, Cao Y, Zhang Y, Fu S, Chen L, Yi G and Fu M (2021) Single-Cell RNA Sequencing of Retina:New Looks for Gene Marker and Old Diseases. Front. Mol. Biosci. 8:699906. doi: 10.3389/fmolb.2021.699906

The retina is composed of 11 types of cells, including neurons, glial cells and vascular bed cells. It contains five types of neurons, each with specific physiological, morphological, and molecular definitions. Currently, single-cell RNA sequencing (sRNA-seq) is emerging as one of the most powerful tools to reveal the complexity of the retina. The continuous discovery of retina-related gene targets plays an important role in helping us understand the nature of diseases. The revelation of new cell subpopulations can focus the occurrence and development of diseases on specific biological activities of specific cells. In addition, sRNA-seq performs high-throughput sequencing analysis of epigenetics, transcriptome and genome at the single-cell level, with the advantages of high-throughput and high-resolution. In this paper, we systematically review the development history of sRNA-seq technology, and summarize the new subtypes of retinal cells and some specific gene markers discovered by this technology. The progress in the diagnosis of retinal related diseases is also discussed.

Keywords: single-cell RNA sequencing, ScRNA-seq, retina, gene, retinal disease

INTRODUCTION

With the development of high-throughput sequencing technology, humans can already analyze genomes and their products on a large scale, including DNA sequences, chromatin structure, RNA transcripts, proteins and metabolites (Botond, 2018). Traditional high-throughput sequencing requires sufficient DNA samples to be obtained from a large number of cells. However, the accuracy of high-throughput sequencing is quite low, and the result of sequencing should be corrected. Single-cell RNA sequencing (scRNA-seq) refers to the technology of high-throughput sequencing analysis of the genome, transcriptome and epigenetic genome at the single cell level. Currently, scRNA-seq technology is commonly used in the fields such as development of stem cell, embryo and tumor. For example, in the study of tumor tissues, researchers classify subgroups based on single-cell transcription maps (Masland, 2012; Wang et al., 2014), and based on the gene expression profiles, they can study the mechanism of cancer cell metastasis (Zheng et al., 2017) and discover new targets for immunotherapy (Pauly et al., 2019).



GRAPHICAL ABSTRACT |

In the field of ophthalmology, single-cell RNA sequencing research has been mostly applied to retina, from cell subtypes to targeted treatments for related diseases. Both humans and monkeys have fovea and macula, but mice are nocturnal dichromats and humans are diurnal trichromats. Therefore, studies on subtypes of retinal cells in humans and primates should ideally be published separately (Pauly et al., 2019). This review summarizes and discusses the latest progress and applications of scRNA-seq technology in the field of retina. So far, scRNA-Seq has been used in mouse, primate, human embryo and adult retinal tissue cell subtype research, as well as the pathogenic gene pathway research of various retinal-related diseases. In this review, we systematically reviewed the rapid progress of single-cell technology (**Figure 1**) and summarized the current challenges and unanswered questions in the field of retinal development and disease.

DEVELOPMENT OF SINGLE-CELL RNA SEQUENCING TECHNOLOGY

Single-cell transcriptome sequencing technology (scRNA-seq) is to analyze the expression profile of the cell transcriptome from the single cell level to identify cell-specific markers, discover rare cell types, cell subtypes, and reveal differences between cells expression (Zerti et al., 2020). The basic technical principles of scRNA-seq technology include: 1) separation technology, such as micromanipulation, laser capture microdissection, fluorescence activated cell sorting, 2) single-cell transcriptome amplification and sequencing library construction. Cells are the basic structural and functional units of organisms (Zerti et al., 2020). During their growth and development, due to different cell states and environmental stimuli, changes in transcriptome information show diversified manifestations. scRNA-seq can study the differential expression of RNA from a single cell level. Since the Tang team first applied scRNA-seq technology in 2009, scRNA-seq technology has received more research and development (**Table 1**). Besides, single-cell sequencing technology has been used to study stem cell differentiation, embryonic organ development, tumor tissue, immune tissue, nervous tissue, and other fields in recent years (Picelli, 2017). In the field of ophthalmology, it is mainly used to study the gene expression of normal retinal tissues and common retinal diseases, such as age-related macular degeneration and diabetic retinopathy.

APPLICATION OF SCRNA-SEQ IN NORMAL RETINAL TISSUE

In the field of ophthalmology, the single-cell RNA sequencing research in the past 5 years has mainly focused on the retina, and most of the research focuses on the exploration of cell subtypes, related genes and pathway. In particular, many researchers choose amacrine cells, bipolar cells and microglia cells for study. Among them, amacrine cells are the most diverse neurons, and most of them lack obvious molecular markers (Grunert and Martin, 1991), which has stimulated curiosity of various researchers in recent years. The retina is a highly heterogeneous tissue, and it is estimated that there are more than 100 nerve cell subtypes.

Primates (including humans) have a fovea on the retina, which is a small central area responsible for high vision and most color vision. However, the retina of mice does not have a fovea. This difference also limits some experimental studies. The distribution and number of primate and mouse retinal cells have a certain difference between the fovea and the periphery. Yi-Rong Peng et al. (Peng et al., 2019) used 165,000 single-cell RNA sequence maps to perform a comprehensive cell classification of the central fovea and peripheral retina of rhesus monkeys, of which 64 fovea (3 PRs, 2 HC, 12 BC, 27 AC, 16 RGC and four non-neurons) and 71 peripherals (2 PR, 2 HC, 11 BC, 34 AC, 18 RGC and four non-neurons) clusters. Comparison with the mouse retina type shows that the middle neuron type is tightly conserved, but the type and procedure of the projection neuron are different.

Based on the previous study, Wen-jun Yan et al. (Yan et al., 2020) compared the gene expression characteristics of human and cynomolgus monkey and fascicular monkey cell types. Besides, they identified five types of neurons (9,070 photoreceptors, 2,868 horizontal cells, 25,908 bipolar cells, 13,607 amacrine cells and 11,404 RGCs) and four types of non-neuronal cells. By comparing the retinal cell types of human and rhesus monkeys, the differentially expressed genes are summarized: the genes that are highly expressed in rhesus monkey retinal tissues are EPHX2, DB1 and DB6. GPATCH1 and CRHBP genes are highly expressed in human retinal tissues. In addition, by comparing retinal foveal cells with surrounding cells, they found that: 1) EPB41L2 and VTN are expressed by the fovea instead of the peripheral cone. 2) expression level of TTR in the fovea is higher than that of the surrounding bipolar types of DB3b and DB4. 3)TULP1 is expressed by peripheral but not foveal bipolar FMB and DB2. Besides, the transcriptome of

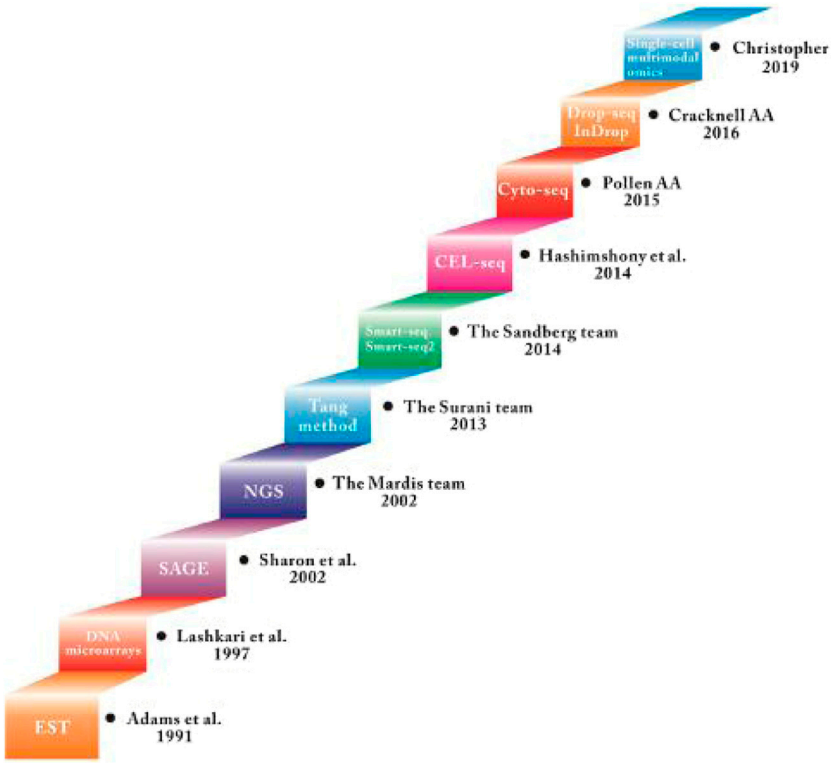


FIGURE 1 | Summary of the development of Single-cell RNA sequecing technology.

TABLE 1 | Principal characteristics of the most widely used scRNA-seq methods.

Name	Transcript coverage	Year	First discovery	Positional bias	Strand specificity	References
Tang method	Nearly full length	2013	Surani et al.	Strongly 3′	No	Gao, (2018)
STRT-seq	5′ only	2013	Sten linnarssons et al.	5′ only	No	Islam et al. (2012)
Smart-seq	Full length	2014	Sandberg et al.	Medium 3′	No	Valdes-Mora et al. (2018)
Smart-seq2	Full length	2014	Sandberg et al.	Weakly 3′	No	Valdes-Mora et al. (2018)
CEL-seq	3′ only	2014	Hashimshony et al.	3′ only	Yes	Hashimshony et al. (2012)
CEL-seq2	3′ only	2014	Hashimshony et al.	3′ only	Yes	Hashimshony et al. (2012)
MARS-seq	3′ only	2014	Hashimshony et al.	3′ only	Yes	Jaitin et al. (2014)
CytoSeq	Predefined genes only	2015	Pollen AA et al.	3′ only	Yes	Islam et al. (2014)
Drop-seq/InDrop	3′ only	2016	Cracknell JA et al.	3′ only	Yes	Pollen et al. (2014)
DroNC-seq	3′ only	2017	Habib et al.	3′ only	Yes	Habib et al. (2017)
Sci-RNA-seq	3′ only	2017	Cao et al.	3′ only	Yes	Cao et al. (2017)
Seq-well	3′ only	2017	Gierahn et al.	3′ only	Yes	Gierahn et al. (2017)
SPLiT-seq	3′ only	2018	Rosenberg et al.	3′ only	Yes	Rosenberg et al. (2018)
Quartz-Seq2	3′ only	2018	Sasagawa et al.	3′ only	Yes	Sasagawa Y et al. (2018)
Single-cell multimodal omics	Full length	2019	Christopher et al.	3′ only	Yes	Author Anonymous (2020); Moncada et al., 2020)

approximately 85,000 cells from the fovea and surrounding retinas of seven adult donors was analyzed by Wen-jun Yan et al. (Yan W et al., 2020) using single-cell RNA sequencing (sRNA-seq) in 2020. The results showed that FOXP2+, FOXP1-, FOXP2+, FOXP1+, and F-RGCs were highly expressed in RGCS cell clusters. In addition, the comparison showed that more than 90% of the human types were transcriptionally consistent with those previously identified in macaques, and that the expression of disease-related genes was highly conserved between the two species. These results confirm the usefulness of macaques in simulating blinding diseases and provide a basis for investigating the molecular mechanisms of visual processing.

TABLE 2 | New discoveries of genes and cell subtypes related to retina.

Study name	Methodology	Sample source	Number of cells sequenced	Year of publication	Molecules/pathways identified	References
Evan Z macosko et al.	Drop-seq	Mouse retinal cells	44,808	2015	Found 39 different cell populations	Macosko et al. (2015)
Shekhar et al.	Drop-seq	Mouse retinal bipolar cells	4 clusters	2016	Divided four types of BC5 (BC5A-BC5D)	Shekhar et al. (2016)
Yi-rong peng et al.	scRNA-seq	Macaque fovea and peripheral retina	165,000	2019	Fovea and peripheral retina contain more than 65 cell types	Peng et al. (2019)
Emily G O'Koren et al.	scRNA-seq	Mouse retina	4 clusters	2019	Found two types of microglia	O'Koren et al. (2019)
Sharon et al.	scRNA-seq	Mouse retina	6 clusters	2019	Reported reported 20 highly expressed genes in the macular region and 23 highly expressed genes in the peripheral region	Banin et al. (2015)
Li et al.	RNA-seq	10 non-proliferative DR patients and 11 non-DR T2DM patients	2051	2019	Found 1,239 areas the macular area is highly expressed and the 812 peripheral areas	Li et al. (2004)
Xiying mao et al.	scRNA-seq	Human embryonic stem cell (hESC)-derived 3D retinal organoids	16,348	2019	Found the RPC specific markers VSX2 and PAX6	Mao et al. (2019)
Yuqiong hu et al.	scRNA-seq	Human fetal NR and RPE	13,000	2019	Identified the main cell types of human fetal retina	Trimarchi et al. (2009)
The mariona esquerdo-barragán team	ChiP-seq	Mouse retina	87	2019	Found TOPORS, KLHL7, PRPF8, USP45, and Usp-20 were expressed at low levels in the retina	Esquerdo-Barragán et al. (2019)
Wen-Junyan et al. (2020)	scRNA-seq	Adult human donors	62,857	2020	Identified 5 types of neurons and 4 types of non-neuronal cells	Yan et al. (2020)
Wen-Junyan et al. (2020)	scRNA-seq	Adult human donors	85,000	2020	FOXP2+, FOXP1-, FOXP2+, FOXP1+, and F-RGCs were highly expressed in RGCS cell clusters	Yan W et al. (2020)
Masahito yamagata et al.	scRNA-seq	Chicken retinas	4,000	2021	VSX2 (CHX10) in the basal cells, TFAP2A in the central retinal cells, and RBPMS2 in retinal ganglion cells	Yamagata M and Sanes, (2021)

Regarding the study of fovea and peripheral cells, Sharon et al. (Banin et al., 2015) have reported 20 highly expressed genes in the macular region (such as SLC17A6, SNCG, NEFL, NET1, STMN2, YWHAH, UCHL1, DPYSL2, APP, NDRG4, TUBA1B,MDH1, EEF2) and 23 highly expressed genes in the peripheral region of the retina (SAG, RCVRN, UNC119, GPX3, PDE6G, ROM1, ABCA4, DDC, PDE6B, GNB1, NRL); based on bulk RNA Seq, Li et al. (Li et al., 2004) reported 1,239 The macular area is highly expressed and the 812 peripheral areas are highly expressed. These related studies provide a basic framework for single-cell analysis of species and across tissue regions.

In addition, scRNA-seq can be used to further study genetic markers and typing of specific optic nerve tissues and retinal cells. Macosko et al. (Macosko et al., 2015) analyzed the transcripts of 44,808 mouse retinal cells and identified 39 different transcribed cell populations, establishing a molecular map of gene expression for known retinal cell types and new candidate cell subtypes. Among them, 21 clusters of amacrine cells were mainly studied. 12 were identified as GABAergic (Gad1 and/or Gad2 positive), and the other five were glycine (Glcine transporter Slc6a9 positive). Ebf3 is a transcription factor found in SEG-glycine and nGnG-amacrine proteins and is specific for clusters 17 and 20.

To further study the gene expression of bipolar cells, Shekhar et al. (Shekhar et al., 2016) used mouse retinal bipolar cells (BCs) as the research object through DROP-SEQ and classified them by two different criteria. Firstly, according to whether the RBC is marked or not, a rod-shaped or cone-shaped BC is divided; secondly, according to the bipolar mark Isl1 and/or Grm6, the cone BC cluster can be further divided into on (3–6, 13, 15) and off (7–10, 12, 14) BC type. It is also worth mentioning that on the basis of the predecessors, the team further divided four types of BC5 (BC5A-BC5D), specifically BC5A (Sox6+) and BC5B (Chrm2+), BC5C (Slitrk5+), BC5D (Lrrtm1+).

On the basis of previous studies, O'Koren team (O'Koren et al., 2019) used single-cell sequencing to reveal the unique transcriptome-related genes of microglia in photoreceptor degeneration, such as Lsp1, asApoe, Ppiaf4, and Alox5ap, which were temporarily induced in the middle of the trajectory; Fabp5, Lgals3, Cd63, Lpl, Cybb, Mmp12, and Spp1 are adjusted up late in the trajectory.

The developmental pathways of mouse neural retina (NR) and retinal pigment epithelium (RPE) have been extensively revealed. However, the molecular mechanism of human NR and RPE formation and the interaction between these two tissues have not been well elucidated (Dulken et al., 2017). In recent years, some studies have used scRNA-seq technology to conduct experimental design with retinal multifunctional stem cells (RPCs) as the research object. RPCs are located in the inner layer of the optic cup (Oppikofer et al., 2017). They produced six types of neurons in retinal cells. The processes that retinal

development needs to go through: RPC proliferation, cell fate determination, and specific neuronal differentiation (Gordon et al., 2013).

Yuqiong Hu et al. (Trimarchi et al., 2009) identified the main cell types of human fetal retina, which are RGC expressing γ-synuclein (SNCG), NEFL, ATOH7 and EBF3; HCs express ONECUT1/2/3; ACs express MEIS2, GAD1 and GAD2; BCs express VSX1 and VSX2; PCs express PDC, PDE6G, SAG, CRX and NRL; microglia express CX3CR1, C1QA, C1QB and C1QC; fibroblasts express COL3A1 and COL1A1.

According to the report, Xiying Mao et al. (Mao et al., 2019) found that the RPC specific markers VSX2 and PAX6 were co-expressed 28 days ago; after 28 days, the expression of VSX2 began to disappear on the central basal side of the retina, expressing the retinal ganglion cell (RGC) marker ELAVL3/ The cells of four began to appear simultaneously, and the number of ELAVL3/4 positive cells gradually increased thereafter. HES1 and HES5 are briefly activated in RPC (Lukaszewicz and Anderson, 2011) and then suppressed in terminally differentiated neurons, and HES6 continues to be up-regulated after the lineage bifurcation point.

Based on the previous evidence, Brian S. Clark et al. (Clark et al., 2019) used single-cell RNA sequencing to describe ten developmental stages covering the entire process of retinal neurogenesis, our results indicate that NFI transcription factors (NFIA, NFIB, and NFIX) are selectively expressed in late RPCs and indicate that they regulate the fate of bipolar interneurons and Miller glial cells and promote proliferation and imactivation. Besides, Mariona Esquerdo-Barragán team (Esquerdo-Barragán et al., 2019) found that TOPORS, KLHL7, PRPF8, USP45, and Usp-20 were expressed at low levels in the retina through scRNA-seq. Josd1, Pan2, Usp11, Usp14, Usp15, Usp10, Usp22, Usp39 and cone cells Compared to the expression of rod differentiation, the expression of three genes (Otud7b, Usp46, and Usp48) increased in late cone cells; the expression of Usp45, Usp53, and Usp54 was limited to the photosensitive layer; Usp28, Usp37, or Otub1 is highly expressed in the embryonic period, but expression is stopped after birth; Usp12, Zranb1 or Usp32, its expression is extremely low in the embryonic period, but significantly increased before and after birth (Hojo et al., 2000). These genes are related to the ubiquitin proteasome system (UPS), which has important research significance for retinal precursor cell differentiation.

In order to further explain the tissue structure and cell subtypes of the chicken retina, based on previous studies, this year Masahito Yamagata et al. (Yamagata M and Sanes, 2021)used single-cell RNA sequencing (sRNA-seq) to generate a cellular atlas of chicken retinas (40,000 single-cell transcriptome), 136 cell types plus 14 sites or developmental intermediates were identified. The team mapped genes expressed in the majority of three types of retinal cells, namely VSX2 (CHX10) in the basal cells, TFAP2A in the central retinal cells, and RBPMS2 in retinal ganglion cells. The results provide new insights into the structure and evolution of the retina and lay the foundation for the study of the anatomy, physiology and development of the retina in birds.

For the past few years, the continuous application and development of scRNA-seq technology has been improved. The study of normal retinal cells in animals and human eyes can redefine the cluster of cells based on the marker gene (**Table 2**). It also enables a deeper understanding of tissue cells and subsequently the cluster of cells. Carrying out a more in-depth classification of cells helps to understand the heterogeneity of cells well, and also brings along a new perspective for our subsequent diagnosis and treatment of disease.

APPLICATION OF SCRNA-SEQ IN RETINAL DISEASES

scRNA-Seq in the Research of Targeted Therapy of Ocular Tumors Application

Different cells change differently at seperate stages of the disease. The transcriptome of many cell subtypes in the retina, especially rare cells, is usually obscured by a large number of RNA sequences. Therefore, understanding the transcriptome at the cell type or single cell level will expand research related to disease.

Single-cell RNA-seq can be used for targeted therapy of eye tumors (Kawaguchi et al., 2008). It is well established that the molecular and cellular characteristics of tumors can indicate the origin of tumor cells and provide a basis for targeted therapy. Retinoblastoma is a malignant tumor in infants and young children. In recent years, researchers have used single-cell sequencing technology to study the pathogenetic gene pathway and treatment of it.

Mcevoy et al. (McEvoy et al., 2011) performed single-cell gene expression array analysis on tumor cells of retinoblastoma patients and mouse models, showing that multiple cell types are specifically expressed in a single retinoblastoma cell. The results showed that human retinoblastoma expressed high levels of MDMX gene and MDMX protein. Some monoamine/ catecholamine receptors in mice include serotonin receptors (HTR3A, HTR1E), dopamine receptors (DRD5) and histamine receptors (HRH3) Expression levels in retinoblastoma It is equal to or higher than the normal human retina.

Based on the previous study, Joseph Collin et al. (Collin et al., 2021) used nine human embryonic and fetal retinal tissues by sRNA-seq and ATAC sequence method. The results showed that Glu137Ter and Tyr655Ter were highly expressed in 4 month old embryonic tumor tissues. However, the Rb1c.763C and Arg255TER genes were overexpressed in embryonic tumor tissue at 34 months. In addition, CCNE1, CCNE2, CCNB2, CCNA2, and CDK1 genes were highly expressed in fetal tumor tissues. In addition, this study provides evidence of the heterogeneity of RB tumors and defines molecular pathways and new targeted therapeutic strategies.

scRNA-Seq on the Pathogenesis of Age-Related Macular Degeneration and Treatment Research

In addition, single-cell sequencing technology also aids the study and treatment of retinal degenerative and vision loss diseases by

TABLE 3 | Studies of gene expression in retina diseases.

Study name	Methodology	Sample source	Diseases	Number of cells sequenced	Year of publication	Molecules/pathways identified	References
Justina McEvoy et al.	Single-cell gene expression array analysis	Human and mouse retina	Retinoblastoma	120	2011	Showed that there are multiple cell type-specific expressions in a single retinoblastoma cell	McEvoy et al. (2011)
Melissa K jones et al.	scRNA-seq	Rat retina	AMD	11,215	2016	Used human brain-derived neural precursor cells to treat retinal degenerative lesions	Jones et al. (2016)
Jacob S heng et al.	scRNA-seq	Rat retina	Autoimmune uveitis retinitis	64,196	2019	Defined the main immune effector cell types	Heng et al. (2019)
Nicholas M. Tran et al.	scRNA-seq	Mouse retinal ganglion cells	Optic nerve crush	46	2019	Generate a comprehensive molecular map of the 46RGC type in the adult retina	Tran et al. (2019)
Radeke et al.	scRNA-seq	Mouse retina	AMD	118	2019	Discovered new age-related macular degeneration (AMD) biomarkers and gene expression characteristics of AMD pathogenesis	Newman et al. (2012)
Xian Zhang et al.	RNA-seq	60 diabetic retinopathy patients	DR	383	2019	Found that overexpression of AK077216 in DR patients resulted in downregulation of miR-383	Zhang et al. (2019)
Madhvi menon et al.	scRNA-seq	Human retina	AMD	23,339	2019	CFI, TIMP3, VEGFA and COL4A3 genes were highly expressed in AMD retinal cells	Menon M et al. (2019)
Wen-Junyan et al. (2020)	scRNA-seq	Human retina	Retinitis pigmentosa	1756	2020	Used cell atlas to evaluate the retinal expression of 1756 disease-related genes	Yan W. et al. (2020)
Joseph collin et al.	scRNA-seq	Human retina	Retinoblastoma	655	2021	CCNE1, CCNE2, CCNB2, CCNA2, and CDK1 genes were highly expressed in fetal tumor tissues	Collin et al. (2021)

analyzing the pathogenesis of related diseases and discovering new biological targets and markers.

Radeke et al. (Newman et al., 2012) discovered new age-related macular degeneration (AMD) biomarkers and gene expression characteristics of AMD pathogenesis. These findings indicate that the cell-based inflammatory response in the RPE choroid is a core feature of AMD. All AMD phenotypes in the RPE choroid are associated with high expression of all or a subset of the following chemokines, namely CXCL1, CXCL2, CXCL9, CXCL10, CXCL11, CCL2 and CCL8. AMD expression in retinal pigment epithelium Related bases and chemokines are C10orf18, ARL9, CXCL10, FZD10, CTSL2, CXCL. AMD may be a single disease with a common immune response process. The genes that regulate these immune activities, as well as many other genes found, represent promising new targets for the treatment and diagnosis of AMD.

On the basis of the previous research, Madhvi Menon et al. (Menon M et al., 2019) retinal cells were isolated and sequenced from six postmortem human retinal macular and surrounding panretinal suspension using droplet based microfluidic (20,091 cells) and nanopore based Seq Well (3,248 cells) to investigate cell types associated with age-related macular degeneration. The results showed that CFI, TIMP3, VEGFA and COL4A3 genes were highly expressed in AMD retinal cells.

Besides, Jones et al. (Jones et al., 2016) used human brain-derived neural precursor cells (hNPCs) to treat retinal degenerative lesions. The results showed that the top five genes with the greatest changes included Mir671, Lcn2, Cd74, Gfap, and Cebpd; Lcn2, Cd74, Gfap, and Cebpd (Hughes et al., 2003). All show that as retinal degeneration increases, Mir671, Lcn2, Cd74, and Cebpd play a role in the immune response of macrophages and/or microglia, suggesting that the activity of macrophages/microglia increases as the retina degenerates (Lawson et al., 1990).

scRNA-Seq on the Pathogenesis of Diabetic Retinopathy and Treatment Research

In recent years, a series of studies on Diabetic Retinopathy (DR) have suggested that vision loss in DR patients is no longer considered to be a simple microvascular complication, also known as neurodegenerative disease (Kamalden et al., 2017). Different retinal cells, trophic factors, neurotransmitters, and inflammatory factors play an important role in the pathogenesis of diabetic retinopathy (Kamalden et al., 2017). Moreover, there are not many studies on diabetic retinopathy by single cell histology (Pastukh et al., 2019).

Xian Zhang et al. (Zhang et al., 2019) found that overexpression of AK077216 in DR patients resulted in downregulation of miR-383, but overexpression of miR-383 had no significant effect on the expression of AK077216; overexpression of AK077216 inhibited apoptosis of ARPE-19 cells (Ru et al., 2014), miR Overexpression of -383 plays the opposite role and attenuates the overexpression of AK077216; therefore it is concluded that AK077216 is down-

regulated in diabetic retinopathy, and inhibits ARPE-19 cell apoptosis by down-regulating miR-383. Zimeng Li et al. found that miR-4448, miR-338-3p, miR-190a-5p, miR-485-5p and miR-9-5p are highly expressed in the serum of DR patients (Shaker et al., 2019).

Application of scRNA-Seq Studies in Other Retinal Diseases

The types of neurons in the central nervous system are significantly different in terms of resilience to injuries or other injuries (Della Santina et al., 2013). In recent years, some researchers have provided a systematic framework to analyze the specific types of injuries through single-cell sequencing technology. Sexual response, and demonstrate that differential gene expression can be used to reveal the molecular targets of intervention.

Nicholas Tran et al. (Tran et al., 2019) first used single-cell RNA-seq (scRNA-seq) to generate a comprehensive molecular map of the 46RGC type in the adult retina. By tracking their survival after ONC (Optic Nerve Crush), the transcription and morphological changes before degradation were described, and each type of selectively expressed genes was determined. Among them, Igf1 (7/7 resRGCs), Opn4 (5/7) and Spp1 (3/7)/OE-Ucn, Ucn protein, OE-Timp2, KO-Crhbp, and KOMmp9 all promoted significant overall regeneration of the optic nerve.

Another experimental study, using single-cell sequencing technology, determined whether multiple explosion exposures caused greater damage to RGC than single explosion exposures (Hong et al., 2015). The results show that Cd40, Mrpl34, Kmo, Lmcd1, BC030870, I830077J02Rik, and Ms4a14 (Kim et al., 2008) are related genes that mediate neuroprotection.

scRNA-seq has been used as a comprehensive and fair method to study cell types and gene expression patterns in the retina of spontaneous, chronic and progressive autoimmune uveitis. Jacob et al. (Heng et al., 2019) used Aire−/− mice to establish a model of autoimmune uveitis retinitis. Mouse models offer a unique opportunity to study the mechanisms of autoimmune uveiretinitis, which is an important cause of vision loss. The team characterized 64,196 isolated retinal cells from eight samples using a droplet based sRNA-seq platform (10×genomics). The results showed that experimental uveiretinitis is a T-cell-driven disease, and the highly expressed genes in the following types of cells were: Th1 cells (T-bet+, IFNG+, CXCR6+, CD4+, CD8a, KLRA1), CD8a + T cells ($CD8a^+$, $CD4^-$, KLRA1), T follicular helper cells (BCL6+, CXCR5+, $CD4^+$, CD8a) and regulatory T cells (Foxp3+, $CD4^+$, IL10+). In addition, TGFb2 is the main TGF-β family member expressed in Aire mouse retina, mainly in the inner layer (INL). In conclusion, this study supports a similar central role of Th1 cells in Aire/uveoretitis, which has important implications for clinical treatment.

Besides, Wen-junYan et al. (Yu-Wai-Man et al., 2010) used cell atlas to evaluate the retinal expression of 1756 disease-related genes. Studies have shown that among the genes associated with retinitis pigmentosa (RP), RPGR and TOPORS, SLC25A46, SLC7A14 and RP9 are highly expressed in RGC (Delettre et al., 2002). In addition, RGR and RLBP1 are highly expressed in Müller glial cells. CRX, RAX2, GNAT2, PDE6H genes are highly expressed in rods and cones. RHO, NRL and NR2E3, all show the enrichment of the fovea (Miller et al., 2019). Lebers congenital amaurosis (LCA) is a group of severe hereditary retinal dystrophy, which is characterized by nystagmus, delayed or missing pupil light reflection, and blindness. Experimental results show that CEP290, GUCY2D and CRB1 genes are highly expressed in RGC (Anguita et al., 2021). In studies related to congenital quiescent night blindness (CSNB), it was found that GNAT1 and SLC24A1 were highly expressed in rod cells, while GRM6 and TRPM1 were highly expressed in bipolar cells (Clemons et al., 2013).

It can be seen that sc RNA-seq research can provide differentiated gene expression of cells for retinal diseases, transcription factor prediction, and the network communication interaction of each cell in the process of disease progression, which can provide new targets for the diagnosis and treatment of disease prediction. Using this technology, we can discover new cell subtypes and identify genetic markers of individual retinal subtype cells to help study and locate targets related to specific visual functions, thereby gaining a deeper understanding of cell function and cell heterogeneity explore the establishment of genetic networks that maintain cell diversity (**Table 3**).

CONCLUSION

Single-cell sequencing has opened up a new field to study different cell subtypes and genetic markers, and reveal the development mechanism and therapeutic targets of retinal-related diseases, and established itself as a valuable and unique tool to further study retinal tissue at the cellular level.

Single-cell sequencing can be used to study the classification of cell types and subtypes in the retina at the transcriptome level, and can help solve the heterogeneity and molecular complexity of the retina. An ideal scRNA-seq method can be used to analyze all coding and efficient non-coding cell transcripts, and even reveal subtle changes in gene expression. The past decade has witnessed significant technological development ever since the first scRNA-seq protocol was published in 2009 (Baden et al., 2016). With the steady decline in sequencing costs and the introduction of methods to significantly increase production every year, the genome, transcriptome, epigenome, and proteome of millions of cells can be sequenced simultaneously in the near future (Benowitz et al., 2017).

The main remaining problem is the challenge of efficiently separating individual cells from biological samples and analyzing large amounts of sequencing data. The close combination of scRNA-seq and bioinformatics technology can provide a powerful detection method to reveal the gene regulatory networks during cell development and differentiation. At present, the application of scrNA-seq in ophthalmic research is still limited. With the continuous progress of the technology, it may be rapidly expanded to the research of ocular diseases in the next few years.

AUTHOR CONTRIBUTIONS

PY, CH and YW have contributed equally to this work and share co-first authorship. Other authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Anguita, R., Tasiopoulou, A., Shahid, S., Roth, J., Sim, S. Y., and Patel, P. J. (2021). A Review of Aflibercept Treatment for Macular Disease[J]. *Ophthalmol. Ther.* doi:10.1007/s40123-021-00354-1

Author Anonymous (2020). Method of the Year 2019: Single-Cell Multimodal Omics. *Nat. Methods* 17 (1), 1. doi:10.1038/s41592-019-0703-5

Baden, T., Berens, P., Franke, K., Román Rosón, M., Bethge, M., and Euler, T. (2016). The Functional Diversity of Retinal Ganglion Cells in the Mouse. *Nature* 529 (7586), 345–350. doi:10.1038/nature16468

Banin, E., Gootwine, E., Obolensky, A., Ezra-Elia, R., Ejzenberg, A., Zelinger, L., et al. (2015). Gene Augmentation Therapy Restores Retinal Function and Visual Behavior in a Sheep Model of CNGA3 Achromatopsia. *Mol. Ther.* 23 (9), 1423–1433. doi:10.1038/mt.2015.114

Benowitz, L. I., He, Z., and Goldberg, J. L. (2017). Reaching the Brain: Advances in Optic Nerve Regeneration. *Exp. Neurol.* 287 (3), 365–373. doi:10.1016/j.expneurol.2015.12.015

Botond, R. (2018). The First Steps in Vision: Cell Types, Circuits, and Repair[J]. *EMBO Mol. Med.* 11 (3), e10218. doi:10.15252/emmm.201810218

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., et al. (2017). Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism. *Science* 357 (6352), 661–667. doi:10.1126/science.aam8940

Clark, B. S., Stein-O'Brien, G. L., Shiau, F., Cannon, G. H., Davis-Marcisak, E., Sherman, T., et al. (2019). Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification. *Neuron* 102 (6), 1111–1126. doi:10.1016/j.neuron.2019.04.010

Clemons, T. E., Gillies, M. C., Chew, E. Y., Bird, A. C., Peto, T., Wang, J. J., et al. (2013). Medical Characteristics of Patients with Macular Telangiectasia Type 2 (MacTel Type 2) MacTel Project Report No. 3. *Ophthalmic Epidemiol.* 20 (2), 109–113. doi:10.3109/09286586.2013.766757

Collin, J., Queen, R., Zerti, D., Steel, D. H., Bowen, C., Parulekar, M., et al. (2021). Dissecting the Transcriptional and Chromatin Accessibility Heterogeneity of Proliferating Cone Precursors in Human Retinoblastoma Tumors by Single Cell Sequencing-Opening Pathways to New Therapeutic Strategies? *Invest. Ophthalmol. Vis. Sci.* 62 (6), 18. doi:10.1167/iovs.62.6.18

Delettre, C., Lenaers, G., Pelloquin, L., Belenguer, P., and Hamel, C. P. (2002). OPA1 (Kjer Type) Dominant Optic Atrophy: A Novel Mitochondrial Disease. *Mol. Genet. Metab.* 75 (2), 97–107. doi:10.1006/mgme.2001.3278

Della Santina, L., Inman, D. M., Lupien, C. B., Horner, P. J., and Wong, R. O. L. (2013). Differential Progression of Structural and Functional Alterations in Distinct Retinal Ganglion Cell Types in a Mouse Model of Glaucoma. *J. Neurosci.* 33 (44), 17444–17457. doi:10.1523/jneurosci.5461-12.2013

Dulken, B. W., Leeman, D. S., Boutet, S. C., Hebestreit, K., and Brunet, A. (2017). Single-Cell Transcriptomic Analysis Defines Heterogeneity and Transcriptional Dynamics in the Adult Neural Stem Cell Lineage. *Cel Rep.* 18 (3), 777–790. doi:10.1016/j.celrep.2016.12.060

Esquerdo-Barragán, B., Toulis, V., Swaroop, A., and Marfany, G. (2019). Expression of Deubiquitinating Enzyme Genes in the Developing Mammal Retina[J]. *Mol. Vis.* 25, 800–813.

Gao, S. (2018). Data Analysis in Single-Cell Transcriptome Sequencing[J]. *Methods Mol. Biol.* 1754, 311–326. doi:10.1007/978-1-4939-7717-8_18

Gierahn, T. M., Wadsworth, M. H., Hughes, T. K., Bryson, B. D., Butler, A., Satija, R., et al. (2017). Seq-Well: Portable, Low-Cost RNA Sequencing of Single Cells at High Throughput. *Nat. Methods* 14 (4), 395–398. doi:10.1038/nmeth.4179

Gordon, P. J., Yun, S., Clark, A. M., Monuki, E. S., Murtaugh, L. C., and Levine, E. M. (2013). Lhx2 Balances Progenitor Maintenance with Neurogenic Output and Promotes Competence State Progression in the Developing Retina. *J. Neurosci.* 33 (30), 12197–12207. doi:10.1523/jneurosci.1494-13.2013

Grunert, U., and Martin, P. (1991). Rod Bipolar Cells in the Macaque Monkey Retina: Immunoreactivity and Connectivity. *J. Neurosci.* 11 (9), 2742–2758. doi:10.1523/jneurosci.11-09-02742.1991

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., et al. (2017). Massively Parallel Single-Nucleus RNA-Seq with DroNc-Seq. *Nat. Methods* 14 (10), 955–958. doi:10.1038/nmeth.4407

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cel Rep.* 2 (3), 666–673. doi:10.1016/j.celrep.2012.08.003

Heng, J. S., Hackett, S. F., Stein-O'Brien, G. L., Winer, B. L., Williams, J., Goff, L. A., et al. (2019). Comprehensive Analysis of a Mouse Model of Spontaneous Uveoretinitis Using Single-Cell RNA Sequencing. *Proc. Natl. Acad. Sci. USA* 116 (52), 26734–26744. doi:10.1073/pnas.1915571116

Hojo, M., Ohtsuka, T., Hashimoto, N., Gradwohl, G., Guillemot, F., and Kageyama, R. (2000). Glial Cell Fate Specification Modulated by the bHLH Gene Hes5 in Mouse Retina. *Development* 127 (12), 2515–2522. doi:10.1242/dev.127.12.2515

Hong, G., Fu, T.-M., Zhou, T., Schuhmann, T. G., Huang, J., and Lieber, C. M. (2015). Syringe Injectable Electronics: Precise Targeted Delivery with Quantitative Input/Output Connectivity. *Nano Lett.* 15 (10), 6979–6984. doi:10.1021/acs.nanolett.5b02987

Hughes, E. H., Schlichtenbrede, F. C., Murphy, C. C., Sarra, G.-M., Luthert, P. J., Ali, R. R., et al. (2003). Generation of Activated Sialoadhesin-Positive Microglia during Retinal Degeneration. *Invest. Ophthalmol. Vis. Sci.* 44 (5), 2229–2234. doi:10.1167/iovs.02-0824

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., et al. (2012). Highly Multiplexed and Strand-specific Single-Cell RNA 5′ End Sequencing. *Nat. Protoc.* 7 (5), 813–828. doi:10.1038/nprot.2012.022

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., et al. (2014). Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers. *Nat. Methods* 11 (2), 163–166. doi:10.1038/nmeth.2772

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-free Decomposition of Tissues into Cell Types. *Science* 343 (6172), 776–779. doi:10.1126/science.1247651

Jones, M. K., Lu, B., Saghizadeh, M., and Wang, S. (2016). Gene Expression Changes in the Retina Following Subretinal Injection of Human Neural Progenitor Cells into a Rodent Model for Retinal Degeneration. *Mol. Vis.* 22, 472–490.

Kamalden, T. A., Macgregor-Das, A. M., Kannan, S. M., Dunkerly-Eyring, B., Khaliddin, N., Xu, Z., et al. (2017). Exosomal MicroRNA-15a Transfer from the Pancreas Augments Diabetic Complications by Inducing Oxidative Stress. *Antioxid. Redox Signaling* 27 (13), 913–930. doi:10.1089/ars.2016.6844

Kawaguchi, A., Ikawa, T., Kasukawa, T., Ueda, H. R., Kurimoto, K., Saitou, M., et al. (2008). Single-cell Gene Profiling Defines Differential Progenitor Subclasses in Mammalian Neurogenesis. *Development* 135 (18), 3113–3124. doi:10.1242/dev.022616

Kim, I.-J., Zhang, Y., Yamagata, M., Meister, M., and Sanes, J. R. (2008). Molecular Identification of a Retinal Cell Type that Responds to Upward Motion. *Nature* 452 (7186), 478–482. doi:10.1038/nature06739

Lawson, L. J., Perry, V. H., Dri, P., and Gordon, S. (1990). Heterogeneity in the Distribution and Morphology of Microglia in the normal Adult Mouse Brain. *Neuroscience* 39 (1), 151–170. doi:10.1016/0306-4522(90)90229-w

Li, S., Mo, Z., Yang, X., Price, S. M., Shen, M. M., and Xiang, M. (2004). Foxn4 Controls the Genesis of Amacrine and Horizontal Cells by Retinal Progenitors. *Neuron* 43 (6), 795–807. doi:10.1016/j.neuron.2004.08.041

Lukaszewicz, A. I., and Anderson, D. J. (2011). Cyclin D1 Promotes Neurogenesis in the Developing Spinal Cord in a Cell Cycle-independent Manner. *Proc. Natl. Acad. Sci.* 108 (28), 11632–11637. doi:10.1073/pnas.1106230108

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161 (5), 1202–1214. doi:10.1016/j.cell.2015.05.002

Mao, X., An, Q., Xi, H., Yang, X.-J., Zhang, X., Yuan, S., et al. (2019). Single-Cell RNA Sequencing of hESC-Derived 3D Retinal Organoids Reveals Novel Genes Regulating RPC Commitment in Early Human Retinogenesis. *Stem Cel Rep.* 13 (4), 747–760. doi:10.1016/j.stemcr.2019.08.012

Masland, R. H. (2012). The Neuronal Organization of the Retina[J]. *Neuron* 76 (2), 266–80. doi:10.1016/j.neuron.2012.10.002

McEvoy, J., Flores-Otero, J., Zhang, J., Nemeth, K., Brennan, R., Bradley, C., et al. (2011). Coexpression of Normally Incompatible Developmental Pathways in Retinoblastoma Genesis. *Cancer Cell* 20 (2), 260–275. doi:10.1016/j.ccr.2011.07.005

Menon M, M. S., Davila-Velderrain, J., Goods, B. A., Cadwell, T. D., Xing, Y., Stemmer-Rachamimov, A., et al. (2019). Single-cell Transcriptomic Atlas of the Human Retina Identifies Cell Types Associated with Age-Related Macular Degeneration[J]. *Nat. Commun.* 10 (1), 4902. doi:10.1038/s41467-019-12780-8

Miller, S. J., Philips, T., Kim, N., Dastgheyb, R., Chen, Z., Hsieh, Y.-C., et al. (2019). Molecularly Defined Cortical Astroglia Subpopulation Modulates Neurons via Secretion of Norrin. *Nat. Neurosci.* 22 (5), 741–752. doi:10.1038/s41593-019-0366-7

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., et al. (2020). Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas. *Nat. Biotechnol.* 38 (3), 333–342. doi:10.1038/s41587-019-0392-8

Newman, A. M., Gallo, N. B., Hancox, L. S., Miller, N. J., Radeke, C. M., Maloney, M. A., et al. (2012). Systems-level Analysis of Age-Related Macular Degeneration Reveals Global Biomarkers and Phenotype-specific Functional Networks. *Genome Med.* 4 (2), 16. doi:10.1186/gm315

O'Koren, E. G., Yu, C., Klingeborn, M., Wong, A. Y. W., Prigge, C. L., Mathew, R., et al. (2019). Microglial Function Is Distinct in Different Anatomical Locations during Retinal Homeostasis and Degeneration[J]. *Immunity* 50 (3), 723–737. doi:10.1016/j.immuni.2019.02.007

Oppikofer, M., Bai, T., Gan, Y., Haley, B., Liu, P., Sandoval, W., et al. (2017). Expansion of the ISWI Chromatin Remodeler Family with New Active Complexes. *EMBO Rep.* 18 (10), 1697–1706. doi:10.15252/embr.201744011

Pastukh, N., Meerson, A., Kalish, D., and Blum, A. (2019). Serum miR-122 Levels Correlate with Diabetic Retinopathy. *Clin. Exp. Med.* 19 (2), 255–260. doi:10.1007/s10238-019-00546-x

Pauly, D., Agarwal, D., Dana, N., Schäfer, N., Biber, J., Wunderlich, K. A., et al. (2019). Cell-Type-Specific Complement Expression in the Healthy and Diseased Retina. *Cel Rep.* 29 (9), 2835–2848. doi:10.1016/j.celrep.2019.10.084

Peng, Y.-R., Shekhar, K., Yan, W., Herrmann, D., Sappington, A., Bryman, G. S., et al. (2019). Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell* 176 (5), 1222–1237. doi:10.1016/j.cell.2019.01.004

Picelli, S. (2017). Single-cell RNA-Sequencing: The Future of Genome Biology Is Now[J]. *RNA Biol.* 14 (5), 637–650. doi:10.1080/15476286.2016.1201618

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage Single-Cell mRNA Sequencing Reveals Cellular Heterogeneity and Activated Signaling Pathways in Developing Cerebral Cortex. *Nat. Biotechnol.* 32 (10), 1053–1058. doi:10.1038/nbt.2967

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., et al. (2018). Single-cell Profiling of the Developing Mouse Brain and Spinal Cord with Split-Pool Barcoding. *Science* 360 (6385), 176–182. doi:10.1126/science.aam8999

Ru, Y., Kechris, K. J., Tabakoff, B., Hoffman, P., Radcliffe, R. A., Bowler, R., et al. (2014). The multiMiR R Package and Database: Integration of microRNA-Target Interactions along with Their Disease and Drug Associations. *Nucleic Acids Res.* 42 (17), e133. doi:10.1093/nar/gku631

Sasagawa Y, D. H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A., et al. (2018). Quartz-Seq2: a High-Throughput Single-Cell RNA-Sequencing Method that Effectively Uses Limited Sequence Reads[J]. *Genome Biol.* 19 (1), 29. doi:10.1186/s13059-018-1407-3

Shaker, O. G., Abdelaleem, O. O., Mahmoud, R. H., Abdelghaffar, N. K., Ahmed, T. I., Said, O. M., et al. (2019). Diagnostic and Prognostic Role of Serum miR-20b, miR-17-3p, HOTAIR, and MALAT1 in Diabetic Retinopathy. *IUBMB Life* 71 (3), 310–320. doi:10.1002/iub.1970

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., et al. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* 166 (5), 1308–1323. doi:10.1016/j.cell.2016.07.054

Tran, N. M., Shekhar, K., Whitney, I. E., Jacobi, A., Benhar, I., Hong, G., et al. (2019). Single-Cell Profiles of Retinal Ganglion Cells Differing in Resilience to Injury Reveal Neuroprotective Genes. *Neuron* 104 (6), 1039–1055. doi:10.1016/j.neuron.2019.11.006

Trimarchi, J. M., Cho, S.-H., and Cepko, C. L. (2009). Identification of Genes Expressed Preferentially in the Developing Peripheral Margin of the Optic Cup. *Dev. Dyn.* 238 (9), 2327–2329. doi:10.1002/dvdy.21973

Valdes-Mora, F., Handler, K., Law, A. M. K., Salomon, R., Oakes, S. R., Ormandy, C. J., et al. (2018). Single-Cell Transcriptomics in Cancer Immunobiology: The Future of Precision Oncology[J]. *Front. Immunol.* 12 (9), 2582. doi:10.3389/fimmu.2018.02582

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., et al. (2014). Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing. *Nature* 512 (7513), 155–160. doi:10.1038/nature13600

Yamagata M, Y. W., and Sanes, J. R. (2021). A Cell Atlas of the Chick Retina Based on Single-Cell Transcriptomics[J]. *Elife* 10, e63907. doi:10.7554/elife.63907

Yan W, P. Y., van Zyl, T., Regev, A., Shekhar, K., Juric, D., and Sanes, J. R. (2020). Cell Atlas of the Human Fovea and Peripheral Retina[J]. *Sci. Rep.* 10 (1), 9802. doi:10.1038/s41598-020-66092-9

Yan, W., Laboulaye, M. A., Tran, N. M., Whitney, I. E., Benhar, I., and Sanes, J. R. (2020). Mouse Retinal Cell Atlas: Molecular Identification of over Sixty Amacrine Cell Types. *J. Neurosci.* 40 (27), 5177–5195. doi:10.1523/jneurosci.0471-20.2020

Yu-Wai-Man, P., Sitarz, K. S., Samuels, D. C., Griffiths, P. G., Reeve, A. K., Bindoff, L. A., et al. (2010). OPA1 Mutations Cause Cytochrome C Oxidase Deficiency Due to Loss of Wild-type mtDNA Molecules. *Hum. Mol. Genet.* 19 (15), 3043–3052. doi:10.1093/hmg/ddq209

Zerti, D., Collin, J., Queen, R., Cockell, S. J., and Lako, M. (2020). Understanding the Complexity of Retina and Pluripotent Stem Cell Derived Retinal Organoids with Single Cell RNA Sequencing: Current Progress, Remaining Challenges and Future Prospective. *Curr. Eye Res.* 45 (3), 385–396. doi:10.1080/02713683.2019.1697453

Zhang, X., Shi, E., Yang, L., Fu, W., Hu, F., and Zhou, X. (2019). LncRNA AK077216 Is Downregulated in Diabetic Retinopathy and Inhibited the Apoptosis of Retinal Pigment Epithelial Cells by Downregulating miR-383. *Endocr. J.* 66 (11), 1011–1016. doi:10.1507/endocrj.ej19-0080

Zheng, C., Zheng, L., Yoo, J.-K., Guo, H., Zhang, Y., Guo, X., et al. (2017). Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. *Cell* 169 (7), 1342–1356. doi:10.1016/j.cell.2017.05.035

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fgene.2021.706986

Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression

Fengchun Ke†, Wei Kong† and Shuaiqun Wang*

College of Information Engineering, Shanghai Maritime University, Shanghai, China

Imaging genetics combines neuroimaging and genetics to assess the relationships between genetic variants and changes in brain structure and metabolism. Sparse canonical correlation analysis (SCCA) models are well-known tools for identifying meaningful biomarkers in imaging genetics. However, most SCCA models incorporate only diagnostic status information, which poses challenges for finding disease-specific biomarkers. In this study, we proposed a multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model to reveal disease-specific associations between single nucleotide polymorphisms and quantitative traits derived from multi-modal neuroimaging data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. MT-SCCAR uses complementary information carried by multiple-perspective cognitive scores and encourages group sparsity on genetic variants. In contrast with two other multi-modal SCCA models, MT-SCCAR embedded more accurate neuropsychological assessment information through linear regression and enhanced the correlation coefficients, leading to increased identification of high-risk brain regions. Furthermore, MT-SCCAR identified primary genetic risk factors for Alzheimer's disease (AD), including rs429358, and found some association patterns between genetic variants and brain regions. Thus, MT-SCCAR contributes to deciphering genetic risk factors of brain structural and metabolic changes by identifying potential risk biomarkers.

Keywords: imaging genetics, sparse canonical correlation analysis, magnetic resonance imaging, positron emission tomography, single nucleotide polymorphisms, multi-task learning

OPEN ACCESS

Edited by:
Jie Li,
Harbin Institute of Technology, China

Reviewed by:
Lei Du,
Northwestern Polytechnical University, China
Jin Li,
Harbin Medical University, China

***Correspondence:**
Wei Kong
weikong@shmtu.edu.cn

†These authors share first authorship

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

Received: *08 May 2021*
Accepted: *19 July 2021*
Published: *05 August 2021*

Citation:
Ke F, Kong W and Wang S (2021) Identifying Imaging Genetics Biomarkers of Alzheimer's Disease by Multi-Task Sparse Canonical Correlation Analysis and Regression. Front. Genet. 12:706986. doi: 10.3389/fgene.2021.706986

INTRODUCTION

Imaging genetics has recently emerged as a method for investigating imaging and genetic biomarkers related to diseases such as Alzheimer's disease (AD) (Bogdan et al., 2017). Identified neuroimaging and genetics biomarkers can provide a complementary understanding of the brain's structure and metabolism (Zhang et al., 2011). Moreover, the vast amounts of diagnostic and neuropsychological information from various perspectives enable the discovery of disease-specific biomarkers. Therefore, it is essential to simultaneously analyze multiple neuroimaging techniques, such as magnetic resonance imaging (MRI), fluorodeoxyglucose positron emission tomography (FDG-PET), genotyping, and clinical diagnostic data. In this study, we aimed to build a model to identify disease-specific biomarkers across multiple imaging modalities, which can be used as an effective clue for disease diagnosis and targeted therapy.

Numerous studies have attempted to identify the associations between genotypic data such as single nucleotide polymorphisms (SNPs) and neuroimaging quantitative traits (QTs) (Rasetti and Weinberger, 2011). Because genotypic data and imaging QTs are multivariate, several bi-multivariate methods have been proposed to better characterize their associations. Liu et al. explored parallel independent component analysis (PICA) to detect the associations between brain function and genetic variants. However, this method cannot restore meaningful SNPs and regions of interest (ROIs), which has led to a lack of reasonable biomarker interpretation (Liu et al., 2009). Sparse canonical correlation analysis (SCCA) has a strong capability for bi-multivariate association identification and interpretable variable selection. Accordingly, many efforts have attempted to apply SCCA to neuroimaging genetics. Boutte et al. introduced an SCCA model with least absolute shrinkage and selection operator (LASSO) constraints on neuroimaging genetics data fusion (Boutte and Liu, 2010). Hao et al. presented a multi-view SCCA model to establish associations between SNPs, QTs, and cognitive outcomes (Hao et al., 2017). However, these multi-view SCCA models are a simple extension to conventional SCCA models. The requirement that SNP canonical weight vectors associate with all modal data is too strict, and could result in not making full use of all modal information. To address this limitation, Du et al. developed a multi-task SCCA model that could be used to jointly analyze SNPs and multiple neuroimaging data by treating each association as an individual learning task (Du et al., 2021). However, this model's neglect of diagnostic information means that biomarkers identified by these multiple-data models may not be sufficiently disease-specific.

To detect more complex and meaningful associations, studies to date have applied diagnostic information into SCCA methods (Yan et al., 2018; Du et al., 2020). Yan et al. proposed an outcome-relevant SCCA model based on a subject similarity matrix (Yan et al., 2018). Du et al. integrated multi-task SCCA and logistic regression in a sophisticated model to identify robust disease-related imaging and genetic patterns by incorporating diagnostic status information (Du et al., 2020). Classified diagnostic information, such as AD, mild cognitive impairment (MCI), and healthy control (HC), facilitates the association between SNPs and QTs; however, roughly dividing the disease stages does not provide any more accurate information than do continuous neuropsychological assessments measured from different angles.

To address the above problems, we proposed a novel SCCA model with the capacity to extract disease-specific biomarkers across multiple neuroimaging modalities. The proposed multi-task sparse canonical correlation analysis and regression (MT-SCCAR) model integrates multi-task SCCA and multi-task linear regression in a fused model and uses multiple cognitive scores (CSs) as auxiliary information to induce associations between SNPs and QTs. Multi-task sparse canonical correlation analysis and regression considers the relationships within subjects from different disease courses and can find disease-specific biomarkers. We also considered underlying hierarchical information among SNPs by modeling structural relationships as divided by gene or by linkage disequilibrium (LD) in a group sparsity penalty. To evaluate MT-SCCAR's effectiveness, we performed extensive experiments to find associations between SNPs and two imaging QTs, including gray matter density and standard uptake value ratio (SUVR) extracted from MRI and positron emission tomography (PET), respectively. Compared with the other two multi-modal SCCA models that used real Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort data, MT-SCCAR not only outperformed these models in its ability to identify genetic AD risk factors, but also detected robust AD brain risk regions across multiple neuroimaging modalities. Thus, our proposed model has the potential to understand disease mechanisms from both structural and metabolic perspectives.

MATERIALS AND METHODS

Data Sources and Preprocessing

Real neuroimaging and genetic data used in this study were obtained from the ADNI1 database. A total of 305 non-Hispanic Caucasian subjects with genotype, neuroimaging, and cognitive assessment data at the ADNI1 baseline were downloaded from the LONI website,[1] including 83 HC, 148 MCI, and 74 AD subjects. The Mini-Mental State Examination (MMSE) is a numeric scale to test cognitive functions, including attention, calculation, and responsiveness to simple commands (Tombaugh and McIntyre, 1992). The Functional Activities Questionnaire (FAQ) evaluates instrumental activities of daily life, such as financial management and meal preparation (Teng et al., 2010). The Alzheimer's Disease Assessment Scale Cognitive Subscale (ADAS-Cog) mainly measures cognitive ability such as word recall, comprehension of spoken language, and orientation (Cano et al., 2010). **Table 1** shows the characteristics of the subjects.

Genotyping Data and Processing

Genotypes for 305 subjects were performed using the Illumina HumanHap610-Quad BeadChips from the ADNI1 database. The SNP data were lifted to hg19 build using lift over tool (Kent et al., 2002). To get pure SNP data, we used a genetic analysis tool PLINK (Purcell et al., 2007) to filter the SNPs using the following quality control criteria: gender check, sibling pair identification, call rate check (<90%) per subject and SNP marker, the Hardy-Weinberg Equilibrium (HWE $p < 10–6$), and marker removal by the minor allele frequency (MAF < 0.05), SNP data were further imputed using Michigan imputation server to estimate

[1] http://adni.loni.usc.edu/

TABLE 1 | Characteristics of the subjects.

Subjects	HC	MCI	AD
Number	83	148	74
Gender(M/F)	50/33	98/50	39/35
Age(mean ± std)	77.76 ± 4.59	76.62 ± 6.92	76.96 ± 6.91
Education(mean ± std)	15.68 ± 3.09	15.88 ± 2.77	14.27 ± 3.37
MMSE (mean ± std)	29.18 ± 1.11	26.09 ± 3.14	20.57 ± 3.20
FAQ (mean ± std)	0.54 ± 1.25	6.62 ± 8.96	19.75 ± 3.50
ADAS-Cog(mean ± std)	6.00 ± 2.89	13.72 ± 3.03	26.09 ± 11.64

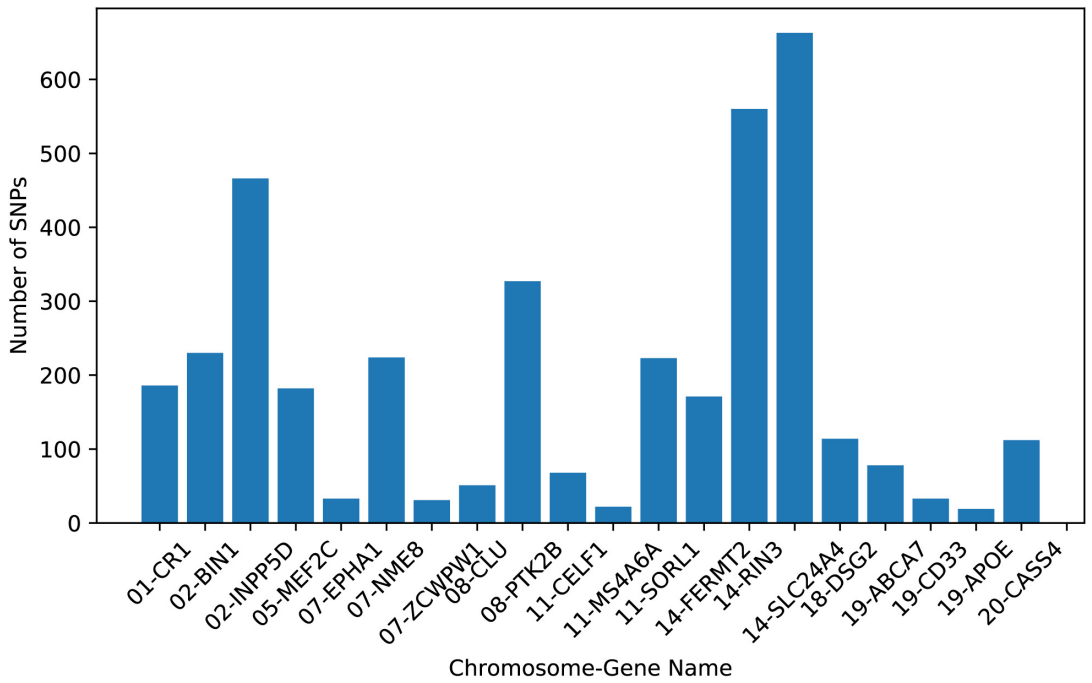


FIGURE 1 | The numbers of SNPs belonging to each AD risk gene used in this study.

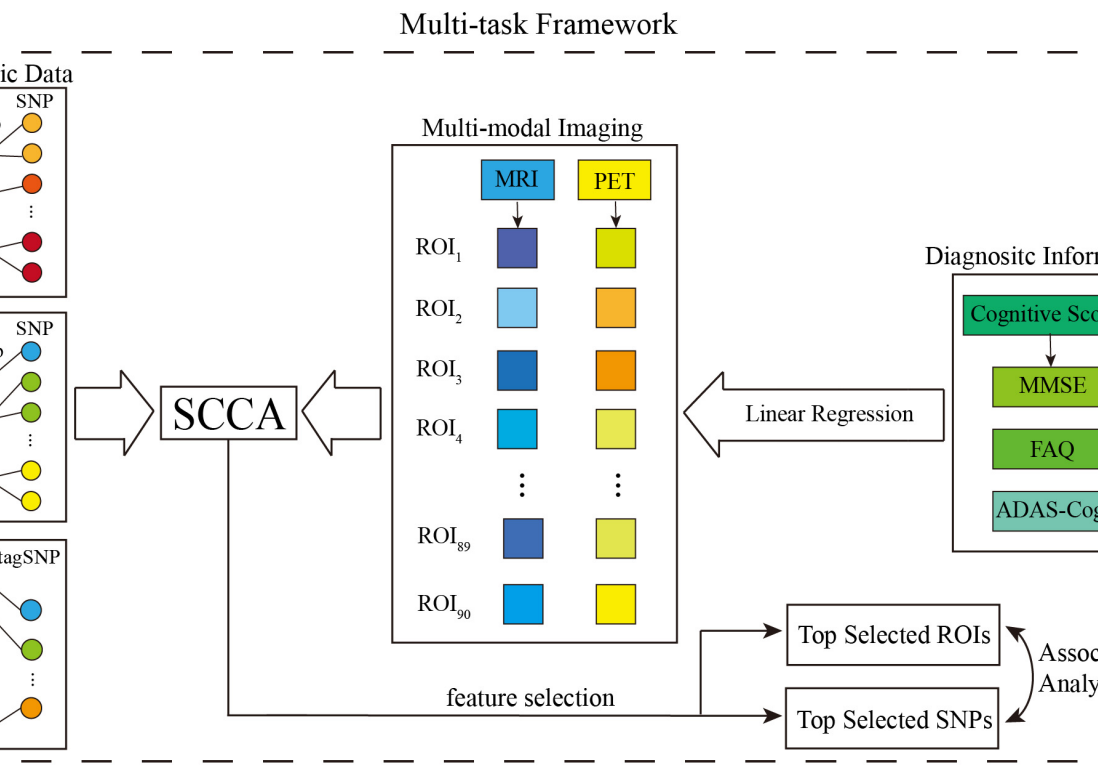


FIGURE 2 | Schematic illustration of MT-SCCAR.

the missing genotypes based on the HRC r1.1 2016 panel (Das et al., 2016). The post-imputation quality control used the rsq > 0.3 and MAF of 0.1 (Li et al., 2010).

Since our study focused on the top 20 AD risk genes listed on the AlzGene database[2] and references (Tanzi et al., 2007; Wang et al., 2012a). After imputation, we selected all the SNPs within ± 5k base pairs of the gene boundary using the ANNOVAR annotation (Wang et al., 2010). The above procedures yielded 3793 SNPs belonging to the top 20 risk genes. **Figure 1** presents the AD risk genes and the number of pre-selected SNPs. Moreover, considering the structural relationship among SNPs, we used Haploview (Barrett et al., 2004) to divide the LD block using the LD-Spline algorithm with $D^{'} > 0.8$, resulting in 209 blocks containing 3770 SNPs. A total of 894 tag SNPs were also assigned by Haploview in pairwise mode and an r^2 threshold was set to 0.8. These tagged SNPs represented the genetic variation across a particular region and could facilitate the association study (Montpetit et al., 2006). Furthermore, each SNP value was coded in an additive fashion to reflect the number of minor alleles.

Neuroimaging Data and Processing

The baseline 1.5T MRI scans were aligned to the standard Montreal Neurological Institute (MNI) space, resampled to $2 \times 2 \times 2$ mm^3 voxels, registered by SPM software package (Ashburner and Friston, 2007). Then, we extracted the gray matter tissue from the MRI scans and calculated mean gray matter densities of 116 ROIs based on MarsBar AAL atlas (Tzourio-Mazoyer et al., 2002). After removing 26 ROIs of the cerebellum, mean gray matter densities of 90 ROIs were used as QTs in our study.

The FDG-PET scans were co-registered to each subject's same visit MRI scans and normalized to MNI space by SPM tool. We further excluded white matter regions by masking the PET with gray matter masks obtained by the segmentation of the same subject's co-registered MRI. Then, the PET scans were normalized into the cerebellar gray matter reference region defined on the AAL atlas to generate SUVR images. After this, we used SUVR of 90 ROIs as QTs in our study by removing the 26 ROIs of cerebellum. Moreover, all the QTs were adjusted to exclude the influence of gender, age, and education.

Methods

In this paper, we denote lowercase letters as vectors, uppercase letters as matrices. $||\boldsymbol{x}||_2$ denotes the Euclidean norm, $||\boldsymbol{X}||_{2,1}$ denotes the sum of the Euclidean norms of the rows of $\boldsymbol{X}$, and $||\boldsymbol{X}||_{1,1}$ denotes the absolute sum of all elements of X.

The CS-Related Features Selection Model for Imaging Genetics

Assuming that there are n subjects with pSNPs, qROIs from M imaging modalities, and G different cognitive outcomes. We used $\boldsymbol{X} \in R^{n \times p}$,$\boldsymbol{Y}_m \in R^{n \times q}\,(m = 1, ..., M)$, and $\boldsymbol{z}_g \in R^{n \times 1}\,(g = 1, ..., G)$ to represent genetic data, multiple imaging data, and cognitive scores, respectively. The basic principle of MT-SCCAR is to find $\boldsymbol{U} \in R^{p \times M}$ and $\boldsymbol{V} \in R^{q \times M}$ to maximize the correlation between $\boldsymbol{X}\boldsymbol{u}_m$ and $\boldsymbol{Y}_m\boldsymbol{u}_m$, where u_{im} indicates the weight of the ith SNP for the mth modality, and v_{jm} indicates the weight of the jth ROI for the mth modality. To identify imaging genetic biomarkers that are relevant to CS and disease, the multi-task linear regression objective was combined with the multi-task SCCA (MTSCCA) objective, which can be formulated as:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \mathcal{L}_R(\boldsymbol{V}) + \mathcal{L}_{SCCA}(\boldsymbol{U}, \boldsymbol{V}) + \Omega(\boldsymbol{U}) + \Omega(\boldsymbol{V}). \quad (1)$$

The above model consists of four parts, $\mathcal{L}_R(\boldsymbol{V})$ detects disease-relevant imaging QTs. $\mathcal{L}_{SCCA}(\boldsymbol{U}, \boldsymbol{V})$ captures the bi-multivariate associations between SNPs and multiple imaging QTs. $\Omega(\boldsymbol{U})$ and $\Omega(\boldsymbol{V})$ are the regularization terms to enforce sparsity of $\boldsymbol{U}$ and $\boldsymbol{V}$, so only a small number of interpretable variables can be selected. This model integrates the advantages of MTSCCA and linear regression, which has a certain superiority in using complementary cognitive information. **Figure 2** provides a schematic overview of MT-SCCAR. SNPs were classified into the same group by either gene or LD. Accordingly, SNPs with gene or LD information and tagSNPs were input to the SCCA component separately, which was used to establish the relationships between genetic data and multiple imaging data. The linear regression component was used to introduce CSs into the SCCA part. The multi-task modeling method guaranteed the ability to process multiple imaging and CS data. Unlike conventional unsupervised SCCA models, MT-SCCAR is a supervised SCCA model, which considers the relationships within subjects from different disease courses.

The Linear Regression Model for CS-QT Associations

In the proposed model, the associations between CSs and multi-modal neuroimaging QTs were established by multi-task regression. For each task, we built a regression model for revealing CS-related neuroimaging QTs:

$$\mathcal{L}_R(\boldsymbol{V}) = \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2, \quad (2)$$

TABLE 2 | Specific procedure of MT-SCCAR algorithm.

Algorithm: MT-SCCAR algorithm

Input: The genetic data $\boldsymbol{X} \in R^{n \times p}$, the neuroimaging data $\boldsymbol{Y} \in R^{n \times q}$ of M modalities, and the CS data $\boldsymbol{Z} \in R^{n \times C}$.
λ_{u1}, λ_{u2}, λ_{u3}, λ_{v1}, λ_{v2}, γ_u, and γ_v.
Ensure: canonical weights $\boldsymbol{V}$ and $\boldsymbol{U}$
1: While not converged regarding to $\boldsymbol{V}$, $\boldsymbol{U}$ do
2:Update the diagonal matrix $\boldsymbol{D}_{v1}$ and $\boldsymbol{D}_{v2}$;
3:Solve $\boldsymbol{v}_m$ according to Equation (12);
4:Normalize $\boldsymbol{v}_m$ so that $||\boldsymbol{Y}\boldsymbol{v}_m||_2^2 = 1$;
5:Update the diagonal matrix $\boldsymbol{D}_{u1}$, $\boldsymbol{D}_{u2}$ and $\boldsymbol{D}_{u3}$;
6:Solve $\boldsymbol{U}$ according to Equation (15);
7:Normalize $\boldsymbol{u}_m$ so that $||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1$;
8: End while

[2]www.alzgene.org

where M is the number of neuroimaging modalities, C is the number of cognitive assessments, and n is the total amount of subjects. $\boldsymbol{v}_m$ is the canonical weight of QTs for the mth modalities,$\boldsymbol{y}_m^l$ is the neuroimaging data vector of the lth subjects for the mth modalities, and z_c^l is the score of the lth subjects for the cth cognitive assessments. This multi-task regression model can jointly utilize neuropsychological assessments from different complementary perspectives.

The MTSCCA Model for SNP-QT Associations

Unlike conventional multi-view SCCA models, MTSCCA learns multiple SCCA tasks together by treating each imaging modality association model as a task. This model was proposed by Du et al. (2021) and can be defined as:

$$\min_{\boldsymbol{u}_m,\boldsymbol{v}_m} \sum_{m=1}^{M} -\boldsymbol{u}_m^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m \; s.t. \; ||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1, \; ||\boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 = 1, \; \forall m. \tag{3}$$

For canonical weights $\boldsymbol{U}$ and $\boldsymbol{V}$, each column $\boldsymbol{u}_m$ and $\boldsymbol{v}_m$ represents an individual learning task for different modalities. The main advantage of this multi-task strategy is that SNP canonical weight vectors do not need to be associated with all imaging modalities simultaneously. Each task focuses on identifying SNPs that are associated with only one imaging modality.

The Regularization Terms

Multiple neuroimaging modalities can provide more comprehensive information in terms of both structural and functional perspectives. In our model, two principal tasks corresponded to two neuroimaging modalities. MT-SCCAR should be able to identify neuroimaging QTs shared among multiple modalities and to enforce individual level sparsity. Hence, $\boldsymbol{\Omega}(\boldsymbol{V})$ was composed of two parts, which can be defined as:

$$\boldsymbol{\Omega}(\boldsymbol{V}) = \lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1}, \tag{4}$$

where λ_{v1} and λ_{v2} are positive parameters and can be tuned via cross-validation.

The first penalty was defined as:

$$||\boldsymbol{V}||_{2,1} = \sum_{i=1}^{q}\sqrt{\sum_{m=1}^{M}\boldsymbol{V}_{i,j}^2} = \sum_{i=1}^{q}\left|\left|\boldsymbol{V}_{i,:}\right|\right|_2, \tag{5}$$

This term aims to enforce task-consistent (modality-consistent) sparsity on $\boldsymbol{V}$, which encourages multi-modal imaging QTs to share similar canonical weights.

The second penalty was defined as:

$$||\boldsymbol{V}||_{1,1} = \sum_{j=1}^{q}\sum_{m=1}^{M}\left|v_{jm}\right|, \tag{6}$$

This term indicates the absolute sum of all elements of $\boldsymbol{V}$, which helps to screen the entire ROIs to find the relevant ROIs.

Similarly, the regularization terms of $\boldsymbol{U}$ also include the above two penalties, which can help discover SNPs that may affect multiple brain regions. It is common knowledge that some SNPs located in the same gene or LD block often have similar functions and are jointly related to specific ROIs. It is essential to model underlying hierarchical information among SNPs by adding an extra penalty. Therefore, we defined $\boldsymbol{\Omega}(\boldsymbol{U})$ as follows:

$$\boldsymbol{\Omega}(\boldsymbol{U}) = \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \lambda_{u3}||\boldsymbol{U}||_G, \tag{7}$$

where λ_{u1}, λ_{u2}, and λ_{u3} are positive parameters, the third penalty (Wang et al., 2012a) can be formulated as:

$$||\boldsymbol{U}||_G = \sum_{k=1}^{K}\sqrt{\sum_{i\in g_k}\sum_{j=1}^{M}u_{ij}^2}, \tag{8}$$

where K denotes the number of groups divided by gene or LD. This penalty penalizes canonical weights as a whole for each task and thus can fully use the structural information.

The Optimization Algorithm

In order to address the problem defined in Equation (1), according to the method that has been well studied previously (Du et al., 2021), we can rewrite Equation (1):

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2 + \sum_{m=1}^{M}||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1} + \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \lambda_{u3}||\boldsymbol{U}||_G \; s.t. \; ||\boldsymbol{X}\boldsymbol{u}_m||_2^2 = 1, \; ||\boldsymbol{Y}\boldsymbol{v}_m||_2^2 = 1, \; \forall m. \tag{9}$$

We then use the Lagrange multiplier to solve this problem by taking the partial derivatives of Equation (9) regarding $\boldsymbol{u}_m$ and$\boldsymbol{v}_m$ separately, which can change the formula from non-convex to convex.

First, we treat $\boldsymbol{U}$ as constant, the Lagrange multiplier of Equation (9) can be simplified as:

$$\sum_{m=1}^{M}\sum_{c=1}^{C}\sum_{l=1}^{n}\left|\left|\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{y}_m^l - z_c^l\right|\right|_2^2 + \sum_{m=1}^{M}||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \lambda_{v1}||\boldsymbol{V}||_{2,1} + \lambda_{v2}||\boldsymbol{V}||_{1,1} + \gamma_v\sum_{m=1}^{M}||\boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 \tag{10}$$

by dropping the constant terms, and γ_v is a positive parameter. For each $\boldsymbol{v}_m$, We further take the partial derivatives of Equation (10) and let the result be zero:

$$\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m - \sum_{c=1}^{C}\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{z}_c - \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}_m + \lambda_{v1}\boldsymbol{D}_{v1}\boldsymbol{v}_m + \lambda_{v2}\boldsymbol{D}_{v2}\boldsymbol{v}_m + (\gamma_v + 1)\,\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\boldsymbol{v}_m = \boldsymbol{0}, \tag{11}$$

where $\boldsymbol{D}_{v1}$ is a diagonal matrix with the ith element as $\frac{1}{2||\boldsymbol{v}_{i,:}||_2}(i \in [1, q])$, and $\boldsymbol{D}_{v2}$ is a diagonal matrix with ith element as $\frac{1}{2||\boldsymbol{v}_{im}||_2}(i \in [1, q]$, and $m \in [1, M])$. Obviously, we can take an

iterative rule to solve this problem since both $\boldsymbol{D}_{v1}$ and $\boldsymbol{D}_{v2}$ are rely on canonical weights $\boldsymbol{V}$. This rule can be formulated as:

$$\boldsymbol{v}_m = \left(\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m + \lambda_{v1}\boldsymbol{D}_{v1} + \lambda_{v2}\boldsymbol{D}_{v2} + (\gamma_v + 1)\,\boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{Y}_m\right)^{-1} \left(\sum_{c=1}^{C} \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{z}_c + \boldsymbol{Y}_m^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}_m\right). \quad (12)$$

Then, we treat $\boldsymbol{V}$ as a constant, the Lagrange multiplier of Equation (9) can be simplified as:

$$\sum_{m=1}^{M} ||\boldsymbol{X}\boldsymbol{u}_m - \boldsymbol{Y}_m\boldsymbol{v}_m||_2^2 + \lambda_{u1}||\boldsymbol{U}||_{2,1} + \lambda_{u2}||\boldsymbol{U}||_{1,1} + \lambda_{u3}||\boldsymbol{U}||_G + \gamma_u\,||\boldsymbol{X}\boldsymbol{u}||_2^2 \quad (13)$$

by dropping the constant terms, and γ_u is also a positive parameter. Similar to $\boldsymbol{v}_m$, for $\boldsymbol{U}$, we let the partial derivatives of Equation (13) to be zero:

$$-\boldsymbol{X}^{\mathrm{T}}\mathbb{Y} + \lambda_{u1}\boldsymbol{D}_{u1}\boldsymbol{U} + \lambda_{u2}\boldsymbol{D}_{u2}\boldsymbol{U} + \lambda_{u3}\boldsymbol{D}_{u3}\boldsymbol{U} + \gamma_u\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}U = \boldsymbol{0}, \quad (14)$$

where $\boldsymbol{D}_{u1}$ is a diagonal matrix with the ith element as $\frac{1}{2||\boldsymbol{u}_{i,:}||_2}(i \in [1, p])$, $\boldsymbol{D}_{u2}$ is a diagonal matrix with ith element as $\frac{1}{2||\boldsymbol{u}_{im}||_2}(i \in [1, p]$, and $m \in [1, M])$, $\boldsymbol{D}_{u3}$ is a block diagonal matrix with element as $\frac{1}{2||U_{k,:}||_{\mathrm{F}}}\boldsymbol{I}_k(k \in [1, K])$, I_k is an identity matrix of the same size with kth SNP groups, and $\mathbb{Y} = [\boldsymbol{Y}_1\boldsymbol{v}_1\boldsymbol{Y}_2\boldsymbol{v}_2 \ldots \boldsymbol{Y}_m\boldsymbol{v}_m]$. Hence, the iterative rules can be formulated as:

$$\boldsymbol{U} = (\lambda_{u1}\boldsymbol{D}_{u1} + \lambda_{u2}\boldsymbol{D}_{u2} + \lambda_{u3}\boldsymbol{D}_{u3} + (\gamma_u + 1)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\mathbb{Y}. \quad (15)$$

Based on the above analysis, the optimization algorithm of the proposed method is shown in **Table 2**. We can update $\boldsymbol{V}$ and $\boldsymbol{U}$ alternatively in each iteration until the predefined convergence criterion is satisfied.

RESULTS AND DISCUSSION

Experimental Settings

To comprehensively evaluate the effectiveness of our proposed MT-SCCAR model, two similar models that can analyze multi-modal data were compared with MT-SCCAR. They are three-view SCCA (TSCCA) and MTSCCA. Three-view SCCA can process neuroimaging, genetics, and cognitive scores data by extending conventional two-view association to three data types. MTSCCA was used to evaluate the regression part of our proposed model performance.

There are seven parameters in our model. Tuning all these parameters will pay a high cost. In our experiment, we fixed γ_u and γ_v to 1 since they mainly control the amplitude of $\boldsymbol{V}$ and $\boldsymbol{U}$ (Chen and Liu, 2011). To tune these

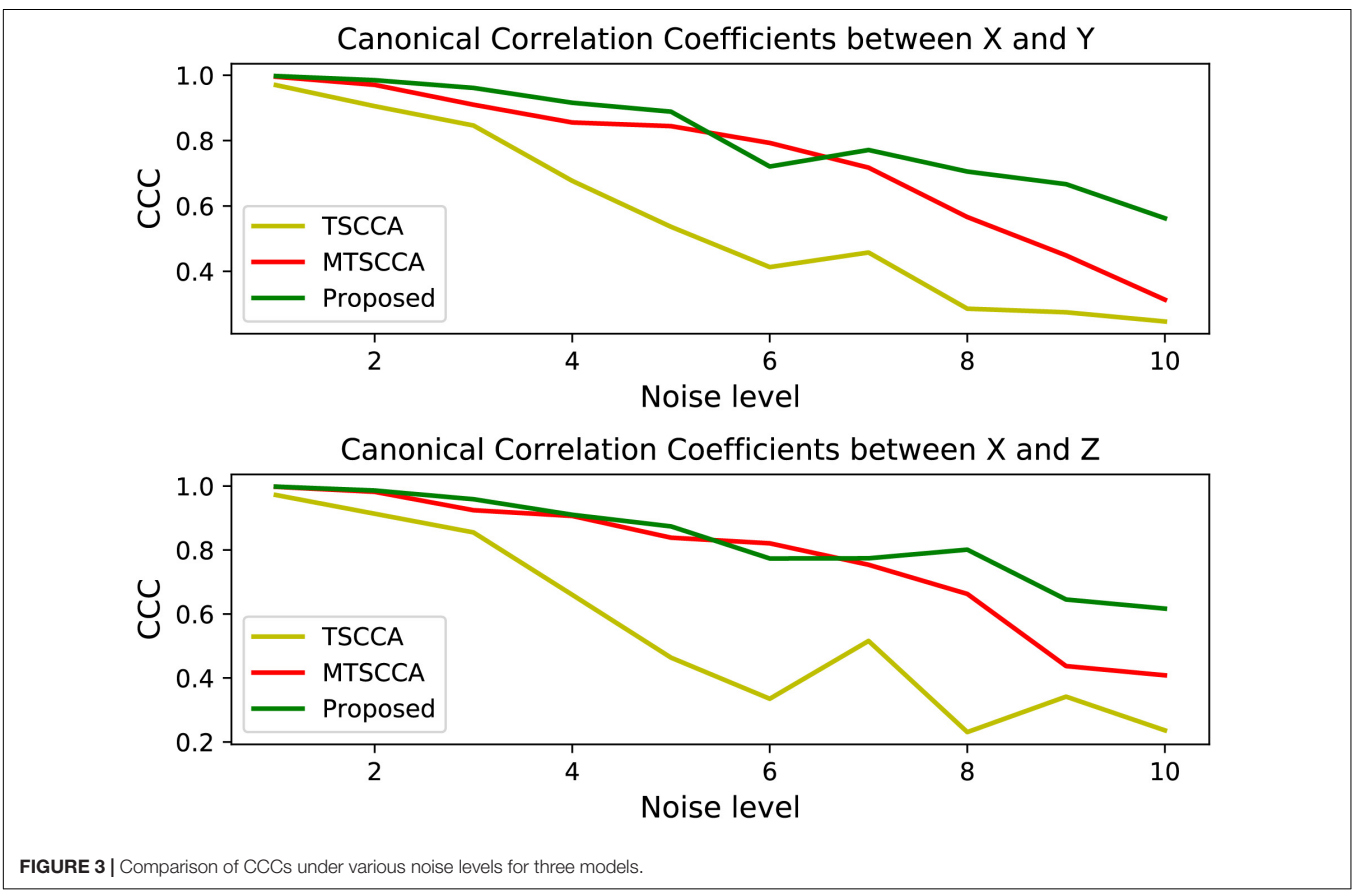


FIGURE 3 | Comparison of CCCs under various noise levels for three models.

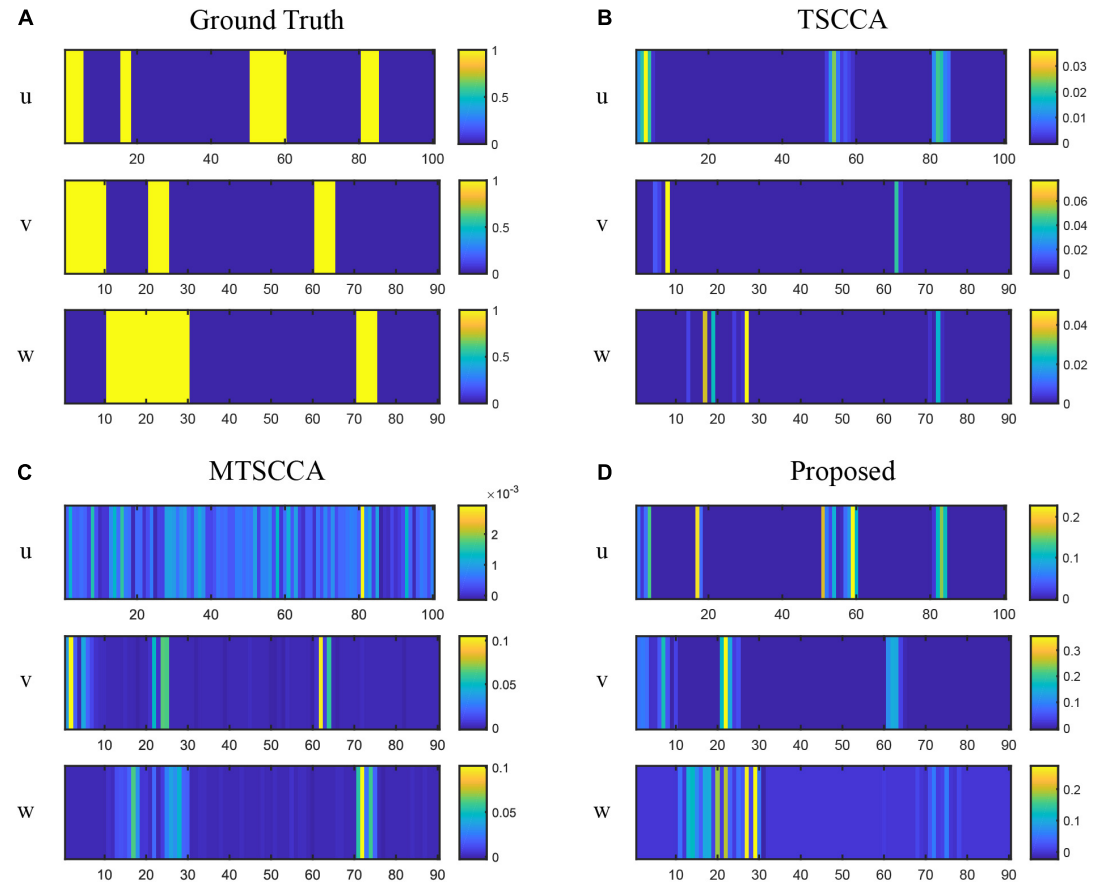


FIGURE 4 | Comparison of canonical weights on synthetic data with the high noise level. **(A)** The ground truth canonical weights. **(B)** The estimated canonical weights of TSCCA. **(C)** The estimated canonical weights of MTSCCA. **(D)** The estimated canonical weights of the proposed model.

parameters to appropriate values, we adopted a nested five-fold cross-validation strategy. Specifically, we tuned them in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}\}$ until the highest mean testing canonical correlation coefficients (CCCs) was generated in the inner loop. CCC was defined as the Pearson correlation coefficient between $\boldsymbol{X}u$ and $\boldsymbol{Y}v$, and can be used as a quantitative measure of SCCA model performance (Hao et al., 2017). For multi-task learning, CCC can be calculated by $corr\ (\boldsymbol{X}_m\boldsymbol{u}_m, \boldsymbol{Y}_m\boldsymbol{v}_m)$ for mth task. Also, we terminated the iteration when both $\max\left|u_i^{(t+1)} - u_i^t\right| \leq 10^{-5}$ and $\max\left|v_j^{(t+1)} - v_j^t\right| \leq 10^{-5}$ were satisfied. All models in our experiment have taken the same parameter adjustment steps.

Results on Synthetic Data

We generated ten synthetic datasets with the same ground truth of loading vectors but different noise levels. Assuming that $\boldsymbol{X} \in R^{n \times p}$,$\boldsymbol{Y} \in R^{n \times q}$, and $\boldsymbol{Z} \in R^{n \times q}$ denote SNP, MRI, and PET for all synthetic data sets, respectively. $\boldsymbol{X}$ was generated by $\boldsymbol{X} = \boldsymbol{ul} + \boldsymbol{e}$,$\boldsymbol{Y}$ was generated by $\boldsymbol{Y} = \boldsymbol{vl} + \boldsymbol{e}$, and $\boldsymbol{Z}$ was generated by $\boldsymbol{Z} = \boldsymbol{wl} + \boldsymbol{e}$, where $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$are known loading vectors, $\boldsymbol{l}$ is a latent vector with a 3-component Gaussian distribution to simulate the disease course (Yan et al., 2018), and $\boldsymbol{e}$ is derived from the Gaussian distribution $N\left(0,\ \sigma_e^2\right)$ with σ_e^2 as the noise variance. In our study, n, p, and q were set to 90, 100, and 90, respectively. All the 90 samples were classed into three groups with centers -5, 0, 5. For neuropsychological assessment data, $\boldsymbol{c}$ was generated by $\boldsymbol{c} = \boldsymbol{l} + \boldsymbol{e}$. To assess the model performance at various noise levels, we tested different noise variances ranging from 1 to 10, with a step size of 1. The five-fold cross-validation results are shown in **Figures 3**, **4**.

Figure 3 plots the testing CCC for three models with changing noise levels. Higher CCC indicates better performance in identifying underlying associations. As expected, the performance decreased with increased noise levels for all models. All three models performed similarly well at low noise levels. Models with the multi-task framework (MTSCCA, MT-SCCAR) performed better than TSCCA at medium noise levels. Then MT-SCCAR outperformed the other two models as the noise level was further increased, suggesting that MT-SCCAR had a strong ability to resist noise. **Figure 4** shows the true signal of canonical weights and canonical weights estimated by three models with a noise level of 10. Important features were highlighted in the heatmaps displaying ground truth. We could clearly observe that the weight $\boldsymbol{u}$ estimated by MTSCCA was ambiguous. It was

therefore difficult to recognize important features. TSCCA did not identify complete important features. MT-SCCAR estimated the best canonical weights that were consistent with the ground truths. These results implied that the proposed model had the potential to extract important features in real neuroimaging genetics studies.

Results on Neuroimaging and Genetics Data

In real neuroimaging genetics data application, all subjects with SNP, MRI, PET, and three different cognitive information data were inputted into MT-SCCAR.A total of 3793 SNPs with LD or gene group information and 894 tag SNPs were used separately. The group sparsity penalty treated each tagSNP as an individual group. We then averaged the CCCs based on five-fold cross-validation, representing the mean strength of identified associations between SNPs and two imaging QTs.

As illustrated in **Table 3**, TSCCA achieved the highest training CCCs but performed poorly in testing CCCs. These unreasonable results may be caused by overfitting (Du et al., 2021). Multi-task sparse canonical correlation analysis and regression achieved the highest testing CCCs on both MRI and PET. Specifically, MT-SCCAR (LD) and MT-SCCAR (gene) achieved the highest testing CCC on SNP-MRI association and SNP-PET association, respectively. Notably, MT-SCCAR (gene) achieved relatively small testing CCC on SNP-MRI association; MT-SCCAR (LD) achieved a more balanced result than those of MT-SCCAR (gene), which indicates that using LD group information is more beneficial than using gene group information. The training CCCs of MT-SCCAR with tagSNP were higher than those of MT-SCCAR with group information since the different numbers

TABLE 3 | Comparison of canonical correlation coefficients (mean ± std) in terms of each model.

	Training CCCs		Testing CCCs	
	SNP-MRI	**SNP-PET**	**SNP-MRI**	**SNP-PET**
TSCCA	**0.82 ± 0.01**	**0.82 ± 0.01**	0.21 ± 0.05	0.23 ± 0.03
MTSCCA	0.55 ± 0.05	0.46 ± 0.11	0.21 ± 0.03	0.30 ± 0.06
Proposed (LD)	0.55 ± 0.01	0.48 ± 0.01	**0.34 ± 0.04**	0.36 ± 0.05
Proposed(gene)	0.56 ± 0.02	0.47 ± 0.01	0.22 ± 0.02	**0.39 ± 0.03**
Proposed(tagSNP)	0.60 ± 0.03	0.52 ± 0.01	0.26 ± 0.05	0.27 ± 0.04

The best correlation coefficients are shown in boldface.

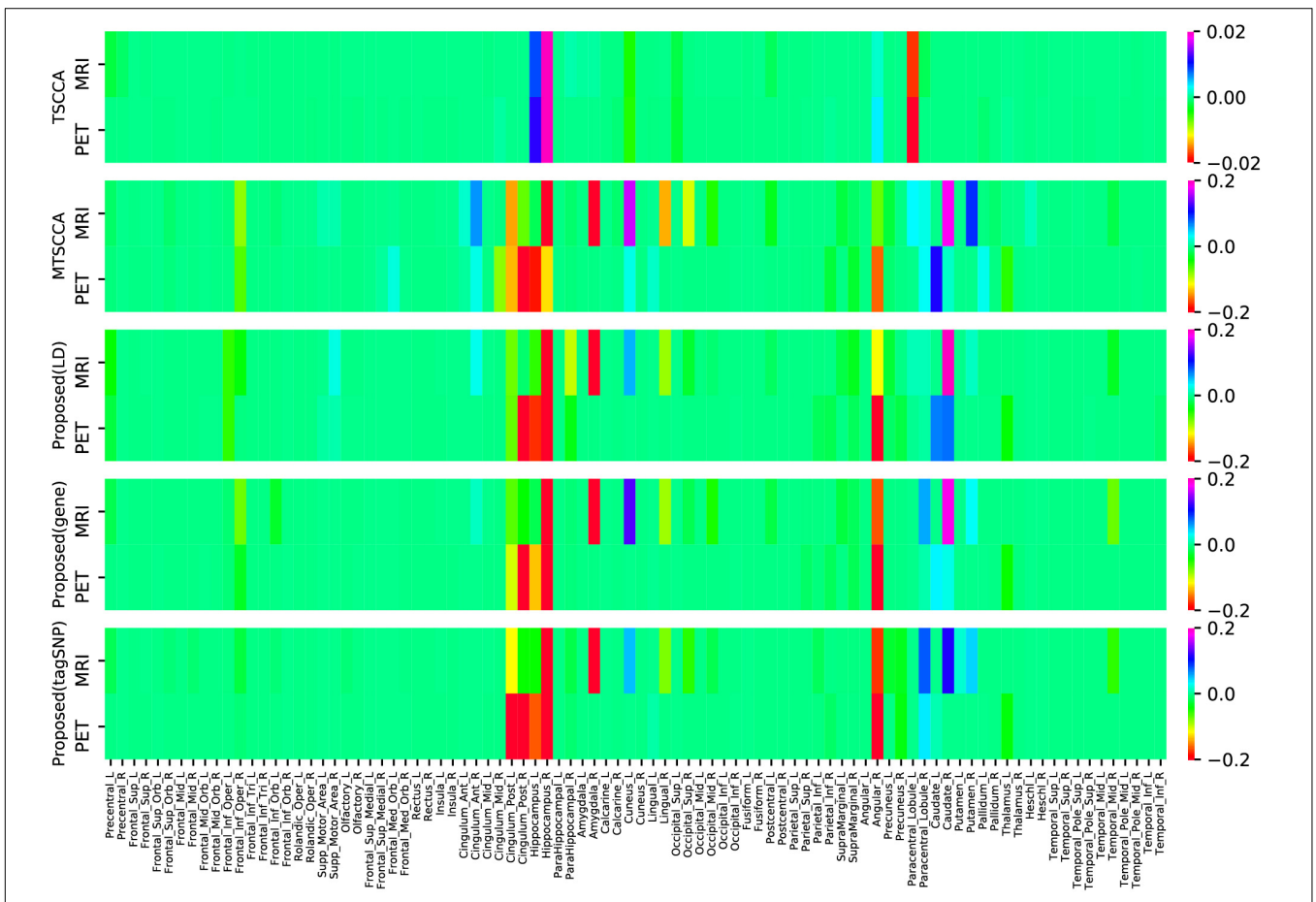


FIGURE 5 | Comparison of estimated canonical weights of imaging QTs. Each row represents: (1) TSCCA; (2) MTSCCA; (3) Proposed (LD); (4) Proposed (gene); (5) Proposed (tagSNP). Within each row, there are two parts represent two imaging modalities.

TABLE 4 | The top ten selected ROIs by the proposed model.

MRI	PET
Hippocampus_R	Cingulum_Post_R
Amygdala_R	**Angular_R**
Caudate_R	**Hippocampus_R**
Angular_R	**Hippocampus_L**
ParaHippocampal_R	**Caudate_R**
Lingual_R	Caudate_L
Cingulum_Post_L	**Cingulum_Post_L**
Cuneus_L	**Frontal_Inf_Oper_L**
Hippocampus_L	Thalamus_L
Frontal_Inf_Oper_L	**ParaHippocampal_R**

The jointly selected ROIs are shown in boldface.

of SNPs were used. Moreover, MTSCCA also performed better than TSCCA, which means the superiority of multi-task models when dealing with multiple imaging QTs and genetic data.

The Top Selected ROIs

In addition to the CCCs, the canonical weights were also one of the focuses of our study since they can help us find brain regions being highly related to AD. **Figure 5** shows the comparison of mean canonical weights of two imaging QTs based on five-fold cross-validation trials. Each row represents an SCCA model. The heatmap color represents the estimated weight of each model, so the selected QTs were highlighted in **Figure 5**. We can clearly observe that several brain regions were selected by both MRI and PET scans, such as the right hippocampal and the right angular gyrus, indicating that these regions may be modality-consistent. Additionally, TSCCA identified only modality-consistent QTs but failed to identify modality-specific QTs. This was due to the nature of its modeling strategy and may have resulted in crucial biomarkers being ignored. Multi-task models can identify modality-specific and modality-consistent QTs, which also implied the limitations of conventional multi-view SCCA models. In order to more accurately analyze the identified brain regions, using the proposed model with LD group information, the top ten ROIs of each modality were selected and sorted according to the absolute values of canonical weights.

As shown in **Table 4**, ROIs that were jointly selected by two imaging modalities are shown in boldface, all of which are known to be closely related to the pathogenesis of AD according to previous research. The hippocampus is essential for forming new memories and was reported as one of the earliest affected brain regions in AD and MCI (Moreno-Jimenez et al., 2019). Both left and right caudate nucleus have been reported that their volume is significantly different between AD and normal control (Cho et al., 2014; Botzung et al., 2019). The right angular gyrus is considered to be closely related to language ability, and patients with angular gyrus syndrome are often found to have damage in this brain

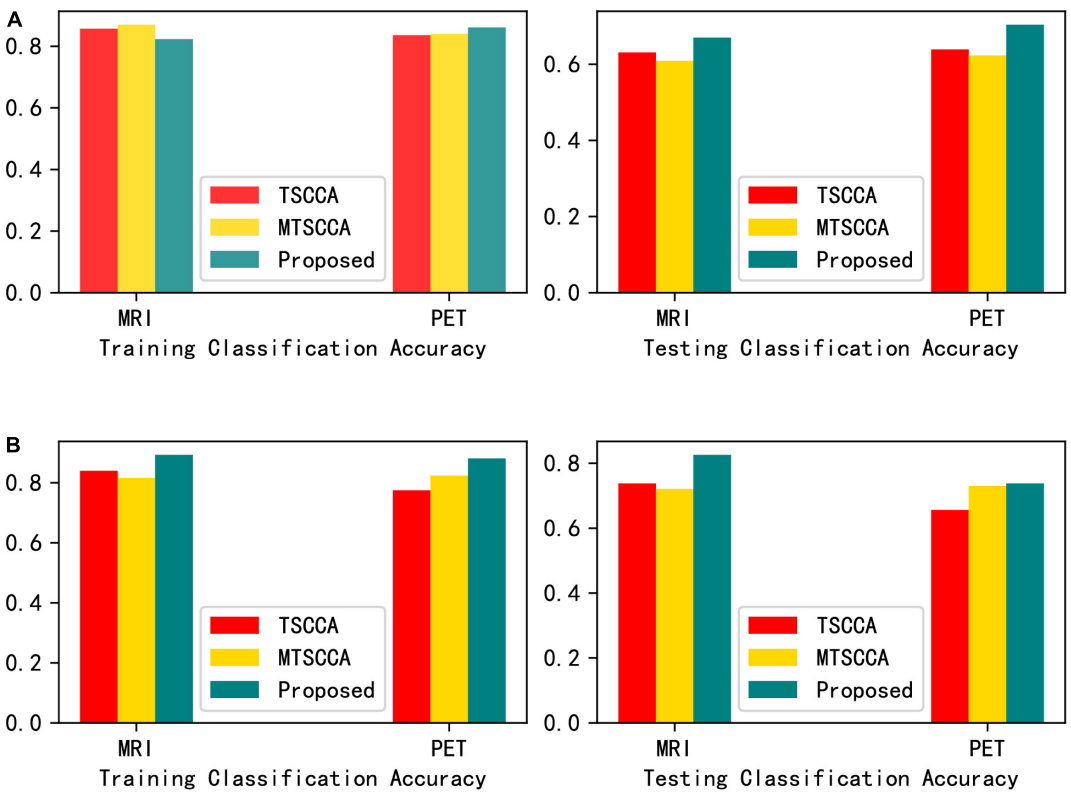


FIGURE 6 | (A) Comparison of the classification accuracy of the selected imaging QTs by support vector machine (SVM). **(B)** Comparison of the classification accuracy of the selected imaging QTs by random forest (RF).

TABLE 5 | The top ten selected SNPs.

TSCCA	MTSCCA	Proposed (LD)	Proposed (gene)	Proposed (tagSNP)
rs735780	rs769449	rs769449	rs7256200	rs117641527
rs405509	rs7256200	rs7256200	rs10414043	rs8012948
rs578506	rs10414043	rs10414043	rs769449	rs1884910
rs4904901	rs4904901	rs4901317	rs7157639	rs78015388
rs7157639	rs61975596	rs429358	rs405509	rs2598123
rs429358	rs7794735	rs4904901	rs4904901	rs4335936
rs4257390	rs55636820	rs7157639	rs429358	rs59325138
rs7412	rs77640937	rs449647	rs75773078	rs439401
rs7794735	rs34273097	rs11629428	rs11629428	rs112097633
rs10256195	rs9972149	rs3829947	rs4901317	rs429358

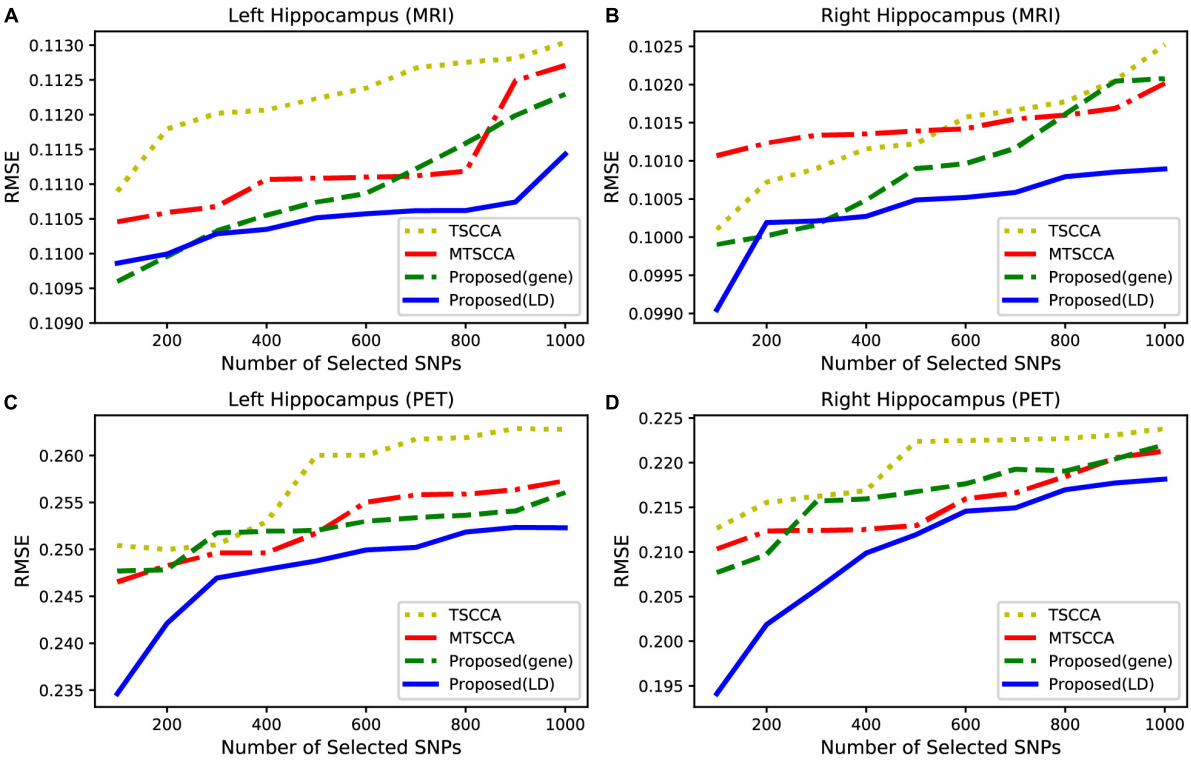


FIGURE 7 | Comparison of the RMSE with respect to different numbers of SNPs from 100 to 1000. **(A)** QT of left hippocampus based on MRI scan. **(B)** QT of right hippocampus based on MRI scan. **(C)** QT of left hippocampus based on PET scan. **(D)** QT of right hippocampus based on PET scan.

area (Horwitz et al., 1998). The right parahippocampal gyrus affects the encoding and maintenance of bound information related to working memory (Luck et al., 2010). The metabolic reduction in the posterior cingulate gyrus is a very early sign in AD (Minoshima et al., 1997). Notably, all the remaining brain regions have also been reported to be associated with AD in published literature. These satisfactory results were due to the inclusion of cognitive information into the linear regression to adjust weighting.

In order to further thoroughly verify that the neuroimaging biomarkers found by the proposed model are more disease-related than those found by the other two models. Selecting the top ten QTs as input features, support vector machine (SVM) with Gaussian radial basis function (RBF) kernel and random forest (RF) were adopted as classification methods. The parameters were tuned with five-fold cross-validation based on the training sets. **Figure 6** presents the classification accuracies of the two classifiers. The testing classification results showed that the classifier using the features selected by MT-SCCAR achieved the highest accuracies, thus indicating the superiority of MT-SCCAR in identifying disease-related biomarkers. Notably, the testing classification accuracies were relatively low for both SVM and RF, probably due to inevitable noise during the feature extraction process of brain imaging. These results were also consistent with previous studies (Wang et al., 2012b; Adeli et al., 2017).

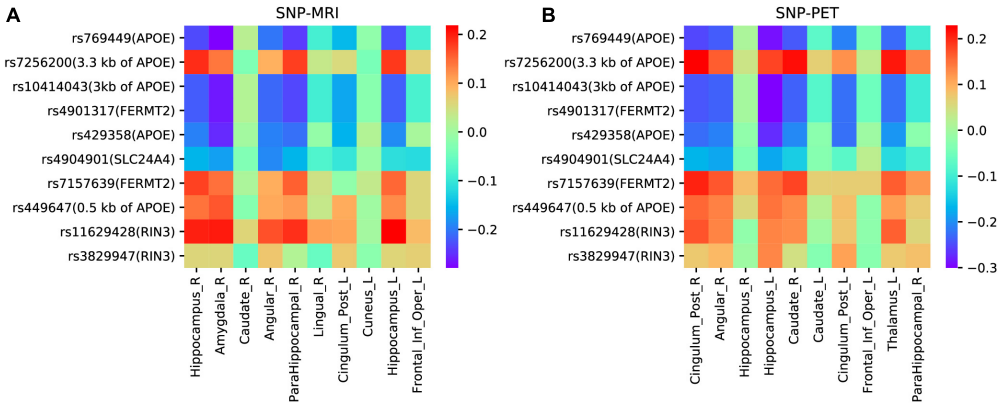


FIGURE 8 | (A) The heatmap of pairwise correlations between SNPs and ROIs of MRI scan. **(B)** The heatmap of pairwise correlations between SNPs and ROIs of PET scan.

TABLE 6 | The correlation coefficients and p-values of eight SNP-ROI pairs.

SNP-ROI pairs	Correlation coefficient	p-value
rs4904901-Angular_R(MRI)	−0.189	0.002
rs4904901- Angular_R(PET)	0.180	0.003
rs7157639-Hippocampus_R(MRI)	0.176	0.003
rs7157639-Cingulum_Post_R(PET)	0.204	0.001
rs11629428-Hippocampus_L(MRI)	0.218	0.0003
rs11629428- Cingulum_Post_R(PET)	0.171	0.004
rs3829947- Angular_R(MRI)	0.078	0.067
rs3829947- Hippocampus_L(PET)	0.135	0.025

The Top Selected SNPs

In addition to neuroimaging biomarkers, SCCA models can also identify genetic biomarkers. We averaged the SNP canonical weights into a single vector and selected the top ten SNPs. As illustrated in **Table 5**, the proposed model with LD or gene group information yielded meaningful results. For example, rs769449 (APOE) is located in promoter and enhancer areas for multiple brain tissues and is associated with AD (Liu et al., 2018). Moreover, the well-known AD risk biomarker rs429358 (APOE) was also identified by the proposed model, demonstrating its strong correlation ability. The remaining five SNPs of the proposed model, i.e., rs7256200 (3.3 kb of APOE), rs10414043 (3kb of APOE), rs4901317 (FERMT2), rs449647 (0.5 kb of APOE), and rs405509 (0.2 kb of APOE), have also been documented to increase the risk of AD in previous studies (Lin et al., 2017; Xiao et al., 2017). However, four selected SNPs have not yet been reported to be related to AD. They still need further research to confirm in the future. Next, we compared the top ten SNPs identified by MT-SCCAR (LD and gene) with the 894 tagSNPs. Interestingly, MT-SCCAR (LD) identified six tagSNPs (rs7256200, rs4901317, rs429358, rs7157639, rs449647, and rs3829947). Multi-task sparse canonical correlation analysis and regression (gene) identified five tagSNPs (rs7256200, rs7157639, rs405509, rs429358, and rs4901317). This implied that using tagSNP will reduce the number of SNPs that need to be analyzed and facilitate identifying significant SNPs. The proposed model with tagSNP also identified some significant SNPs. For example, rs59325138 (3.6 kb of APOE) has been reported to modify the cerebrospinal fluid apolipoprotein E protein levels (Cervantes et al., 2011). The Beta-Amyloid (1-42), an AD biomarker, is associated with rs439401 (1.8kb of APOE) (Xu et al., 2014). The TSCCA identified the rs4292358 and three other SNPs (rs405509, rs7412, and rs7794735) that have been reported previously (Arking et al., 2008; Ma et al., 2016; Zhen et al., 2017). The MTSCCA also identified four SNPs (rs769449, rs7256200, rs10414043, and rs7794735) but cannot identify rs429358. In summary, the proposed model was more accurate for identifying disease-specific genetic biomarkers than the other two models.

Alzheimer's disease (AD) usually first affects the hippocampus, resulting in cognitive decline and memory loss (Moreno-Jimenez et al., 2019). Therefore, when selecting the same number of features, the predictive effect of the QTs of the hippocampus can be used to evaluate model performance. Based on this analysis, we built a regression model to predict the QTs of the hippocampus from MRI and PET scans. Different numbers of SNPs were selected from 100 to 1000 with a step of 100. Using a support vector machine (SVR) with RBF kernel, we calculated the average root mean squared error (RMSE) for each model based on five-fold cross-validation. For a fair comparison, we only compared TSCCA, MTSCCA, MT-SCCAR (gene), and MT-SCCAR(LD) since MT-SCCAR (tagSNP) used only 894 tagSNPs. **Figure 7** shows the testing RMSE of the left and right hippocampus obtained by different imaging techniques. Smaller RMSE indicates that the selected SNPs are more related to AD. According to **Figure 7**, the prediction errors were lowest for the proposed model. These results suggested that the proposed model outperformed the other two models on four imaging QTs.

Pairwise Correlation Analyses

Based on the top ten selected ROIs and SNPs obtained by the proposed model with LD group information, we drew heatmaps of pairwise correlation coefficients between SNPs and two imaging QTs. As illustrated in **Figure 8**, it is clearly observed that

the selected SNPs were mainly located in and around the APOE region. APOE is the major genetic risk factor for AD (Munoz et al., 2019). Moreover, the association patterns of SNPs and ROIs selected by MRI and PET were very similar, which indicated the ability of our model to identify modality-consistent biomarkers.

To gain more insight, we further analyzed four undocumented SNPs (rs4904901, rs7157639, rs11629428, and rs3829947) identified by MT-SCCAR with LD group information. The imaging QTs which had the strongest association with these four SNPs were singled out. Consequently, a total of eight SNP-ROI pairs were generated to validate the proposed model. These associations can also allow us to explore relationships from the microscopic molecular level to the macroscopic brain level. **Table 6** shows the Pearson correlation coefficients and p-values of eight SNP-ROI pairs. The p-values of all eight pairs were small, indicating a significant correlation within each pair. For rs4904901,it was correlated strongest with the same brain region across both imaging modalities, which suggests it is a modality-consistent association pattern. For the rest of the SNPs, the heterogeneous association patterns may have great potential to help us understand how changes in molecular level influence brain structure and metabolic.

CONCLUSION

In this paper, we proposed the MT-SCCAR model to investigate potential neuroimaging and genetic biomarkers. Compared with TSCCA and MTSCCA, the proposed model integrated genotype, multiple neuroimaging, and neuropsychological assessments into a single model to analyze multi-modal information. We tested our model on synthetic and ADNI data sets and compared its association results with those of TSCCA and MTSCCA. We found that our model demonstrated higher CCCs of 0.34 ± 0.04 (LD) and 0.39 ± 0.03 (gene) compared with the CCCs of TSCCA (0.23 ± 0.03) and MTSCCA (0.30 ± 0.06). Moreover, MT-SCCAR identified a small number of SNPs from enormous SNPs that were related to AD, wherein all of the top ten selected ROIs were AD brain risk regions. These satisfactory results show that MT-SCCAR outperforms TSCCA and MT-SCCA in detecting disease-specific biomarkers on multi-modal data.

The proposed model incorporates SNPs, neuroimaging measurements, and cognitive scores. However, there are a number of biological pathways that correlate with structural changes in the brain. Therefore, future efforts should aim to integrate data across more levels (i.e., gene expression, cell, and DNA methylation) for a more sophisticated understanding of the biological pathways leading from gene to disease.

DATA AVAILABILITY STATEMENT

The datasets for this article are not publicly available but are available upon request at the following private repository: Alzheimer's Disease Neuroimaging Initiative, http://adni.loni.usc.edu/data-samples/access-data/, https://ida.loni.usc.edu/pages/access/studyData.jsp (The dataset contains the neuropsychological assessment data), and https://ida.loni.usc.edu/pages/access/geneticData.jsp (The dataset contains the genetics data). Requests to access the datasets should be directed to (Alzheimer's Disease Neuroimaging Initiative or catherine.conti@ucsf.edu). The code is available at https://github.com/ftorange/MT-SCCAR.

AUTHOR CONTRIBUTIONS

WK and FK designed the model and analyzed the results. FK prepared data and drafted the manuscript. SW and FK performed the pre-processing with imaging and genetics data. WK helped with data interpretation and manuscript drafting. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Natural Science Foundation of Shanghai (No. 18ZR1417200) and National Natural Science Foundation of China (No. 61803257).

ACKNOWLEDGMENTS

We appreciate the Alzheimer's Disease Neuroimaging Initiative (ADNI) for contributing data.

REFERENCES

Adeli, E., Wu, G., Saghafi, B., An, L., Shi, F., and Shen, D. (2017). Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson's disease. *Sci. Rep.* 7, 41069–41069. doi: 10.1038/srep41069

Arking, D. E., Cutler, D. J., Brune, C. W., Teslovich, T. M., West, K., Ikeda, M., et al. (2008). A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *Am. J. Hum. Genet.* 82, 160–164. doi: 10.1016/j.ajhg.2007.09.015

Ashburner, J., and Friston, K. (2007). "CHAPTER 7 – Voxel-based morphometry," in *Statistical Parametric Mapping*, eds K. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny (London: Academic Press), 92–98.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2004). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bogdan, R., Salmeron, B. J., Carey, C. E., Agrawal, A., Calhoun, V. D., Garavan, H., et al. (2017). Imaging genetics and genomics in psychiatry: a critical review of progress and potential. *Biol. Psychiatry* 82, 165–175. doi: 10.1016/j.biopsych.2016.12.030

Botzung, A., Philippi, N., Noblet, V., Loureiro de Sousa, P., and Blanc, F. (2019). Pay attention to the basal ganglia: a volumetric study in early dementia with Lewy bodies. *Alzheimers Res. Ther.* 11, 108. doi: 10.1186/s13195-019-0568-y

Boutte, D., and Liu, J. (2010). Sparse canonical correlation analysis applied to fMRI and genetic data fusion. *Proc. IEEE Int. Conf. Bioinform. Biomed.* 2010, 422–426. doi: 10.1109/BIBM.2010.5706603

Cano, S. J., Posner, H. B., Moline, M. L., Hurt, S. W., Swartz, J., Hsu, T., et al. (2010). The ADAS-cog in Alzheimer's disease clinical trials: psychometric evaluation of the sum and its parts. *J. Neurol. Neurosurg. Psychiatry* 81, 1363–1368. doi: 10.1136/jnnp.2009.204008

Cervantes, S., Samaranch, L., Vidal-Taboada, J. M., Lamet, I., Bullido, M. J., Frank-Garcia, A., et al. (2011). Genetic variation in APOE cluster region and Alzheimer's disease risk. *Neurobiol Aging* 32, 2107.e7–17. doi: 10.1016/j.neurobiolaging.2011.05.023

Chen, X., and Liu, H. (2011). An efficient optimization algorithm for structured sparse CCA, with applications to eQTL mapping. *Stat. Biosci.* 4, 3–26. doi: 10.1007/s12561-011-9048-z

Cho, H., Kim, J. H., Kim, C., Ye, B. S., Kim, H. J., Yoon, C. W., et al. (2014). Shape changes of the basal ganglia and thalamus in Alzheimer's disease: a three-year longitudinal study. *J. Alzheimers Dis.* 40, 285–295. doi: 10.3233/JAD-132072

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi: 10.1038/ng.3656

Du, L., Liu, F., Liu, K., Yao, X., Risacher, S. L., Han, J., et al. (2020). Identifying diagnosis-specific genotype–phenotype associations via joint multitask sparse canonical correlation analysis and classification. *Bioinformatics* 36(Suppl.1), i371–i379. doi: 10.1093/bioinformatics/btaa434

Du, L., Liu, K., Yao, X., Risacher, S. L., Han, J., Saykin, A. J., et al. (2021). Multi-task sparse canonical correlation analysis with application to multi-modal brain imaging genetics. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18, 227–239. doi: 10.1109/TCBB.2019.2947428

Hao, X., Li, C., Du, L., Yao, X., Yan, J., Risacher, S. L., et al. (2017). Mining Outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Sci. Rep.* 7:44272. doi: 10.1038/srep44272

Horwitz, B., Rumsey, J., and Donohue, B. (1998). Functional connectivity of the angular gyrus in normal reading and dyslexia. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8939–8944.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* 34, 816–834. doi: 10.1002/gepi.20533

Lin, E., Tsai, S.-J., Kuo, P.-H., Liu, Y.-L., Yang, A. C., and Kao, C.-F. (2017). Association and interaction effects of Alzheimer's disease-associated genes and lifestyle on cognitive aging in older adults in a Taiwanese population. *Oncotarget* 8, 24077–24087.

Liu, C., Chyr, J., Zhao, W., Xu, Y., Ji, Z., Tan, H., et al. (2018). Genome-wide association and mechanistic studies indicate that immune response contributes to Alzheimer's disease development. *Front. Genet.* 9:410. doi: 10.3389/fgene.2018.00410

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., and Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* 30, 241–255. doi: 10.1002/hbm.20508

Luck, D., Danion, J.-M., Marrer, C., Pham, B.-T., Gounot, D., and Foucher, J. (2010). The right parahippocampal gyrus contributes to the formation and maintenance of bound information in working memory. *Brain Cogn.* 72, 255–263. doi: 10.1016/j.bandc.2009.09.009

Ma, C., Zhang, Y., Li, X., Zhang, J., Chen, K., Liang, Y., et al. (2016). Is there a significant interaction effect between apolipoprotein E rs405509 T/T and epsilon4 genotypes on cognitive impairment and gray matter volume? *Eur. J. Neurol.* 23, 1415–1425. doi: 10.1111/ene.13052

Minoshima, S., Giordani, B., Berent, S., Frey, K. A., Foster, N. L., and Kuhl, D. E. (1997). Metabolic reduction in the posterior cingulate cortex in very early Alzheimer's disease. *Ann. Neurol.* 42, 85–94. doi: 10.1002/ana.410420114

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., et al. (2006). An evaluation of the performance of tag SNPs derived from HapMap in a Caucasian population. *PLoS Genet* 2:e27. doi: 10.1371/journal.pgen.0020027

Moreno-Jimenez, E. P., Flor-Garcia, M., Terreros-Roncal, J., Rabano, A., Cafini, F., Pallas-Bazarra, N., et al. (2019). Adult hippocampal neurogenesis is abundant in neurologically healthy subjects and drops sharply in patients with Alzheimer's disease. *Nat. Med.* 25, 554–560. doi: 10.1038/s41591-019-0375-9

Munoz, S. S., Garner, B., and Ooi, L. (2019). Understanding the role of ApoE fragments in Alzheimer's disease. *Neurochem. Res.* 44, 1297–1305. doi: 10.1007/s11064-018-2629-1

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rasetti, R., and Weinberger, D. R. (2011). Intermediate phenotypes in psychiatric disorders. *Curr. Opin. Genet. Dev.* 21, 340–348. doi: 10.1016/j.gde.2011.02.003

Tanzi, R. E., Blacker, D., Bertram, L., McQueen, M. B., and Mullin, K. (2007). Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat. Genet.* 39, 17–23. doi: 10.1038/ng1934

Teng, E., Becker, B. W., Woo, E., Knopman, D. S., Cummings, J. L., and Lu, P. H. (2010). Utility of the functional activities questionnaire for distinguishing mild cognitive impairment from very mild Alzheimer disease. *Alzheimer Dis. Assoc. Disord.* 24, 348–353. doi: 10.1097/WAD.0b013e3181e2fc84

Tombaugh, T. N., and McIntyre, N. J. (1992). The mini-mental state examination: a comprehensive review. *J. Am. Geriatr. Soc.* 40, 922–935. doi: 10.1111/j.1532-5415.1992.tb01992.x

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., et al. (2012a). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28, 229–237. doi: 10.1093/bioinformatics/btr649

Wang, H., Nie, F., Huang, H., Risacher, S. L., Saykin, A. J., Shen, L., et al. (2012b). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* 28, i127–i136. doi: 10.1093/bioinformatics/bts228

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164–e164. doi: 10.1093/nar/gkq603

Xiao, H., Gao, Y., Liu, L., and Li, Y. (2017). Association between polymorphisms in the promoter region of the apolipoprotein E (APOE) gene and Alzheimer's disease: a meta-analysis. *EXCLI J.* 16, 921–938. doi: 10.17179/excli2017-289

Xu, Z., Shen, X., Pan, W., and Alzheimer's Disease Neuroimaging I. (2014). Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes. *PLoS One* 9:e102312. doi: 10.1371/journal.pone.0102312

Yan, J., Liu, K., Lv, H., Amico, E., Risacher, S. L., Wu, Y. C., et al. (2018). "Joint exploration and mining of memory-relevant brain anatomic and connectomic patterns via a three-way association model," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018))*, (Washington, DC: IEEE), 6–9.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., and Alzheimer's Disease Neuroimaging I. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55, 856–867. doi: 10.1016/j.neuroimage.2011.01.008

Zhen, J., Huang, X., Van Halm-Lutterodt, N., Dong, S., Ma, W., Xiao, R., et al. (2017). ApoE rs429358 and rs7412 polymorphism and gender differences of serum lipid profile and cognition in aging chinese population. *Front. Aging Neurosci.* 9:248. doi: 10.3389/fnagi.2017.00248

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 August 2021
doi: 10.3389/fgene.2021.709660

A Novel Method for Identifying Essential Proteins Based on Non-negative Matrix Tri-Factorization

Zhihong Zhang[1,3†], Meiping Jiang[2], Dongjie Wu[4], Wang Zhang[5], Wei Yan[1] and Xilong Qu[3,6†]*

[1] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China, [2] Department of Ultrasound, Hunan Provincial Maternal and Child Health Care Hospital, Changsha, China, [3] School of Information Technology and Management, Hunan University of Finance and Economics, Changsha, China, [4] Department of Banking and Finance, Monash University, Clayton, VIC, Australia, [5] Department of Optoelectronic Engineering, Jinan University, Guangzhou, China, [6] Hunan Provincial Key Laboratory of Finance and Economics Big Data Science and Technology, Hunan University of Finance and Economics, Changsha, China

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Lei Yang,
Harbin Medical University, China
Cheng Liang,
Shandong Normal University, China

***Correspondence:**
Meiping Jiang
meipingjiang123@163.com

[†] These authors have contributed equally to this work

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

Received: *14 May 2021*
Accepted: *06 July 2021*
Published: *06 August 2021*

Citation:
Zhang Z, Jiang M, Wu D, Zhang W, Yan W and Qu X (2021) A Novel Method for Identifying Essential Proteins Based on Non-negative Matrix Tri-Factorization. Front. Genet. 12:709660. doi: 10.3389/fgene.2021.709660

Identification of essential proteins is very important for understanding the basic requirements to sustain a living organism. In recent years, there has been an increasing interest in using computational methods to predict essential proteins based on protein–protein interaction (PPI) networks or fusing multiple biological information. However, it has been observed that existing PPI data have false-negative and false-positive data. The fusion of multiple biological information can reduce the influence of false data in PPI, but inevitably more noise data will be produced at the same time. In this article, we proposed a novel non-negative matrix tri-factorization (NMTF)-based model (NTMEP) to predict essential proteins. Firstly, a weighted PPI network is established only using the topology features of the network, so as to avoid more noise. To reduce the influence of false data (existing in PPI network) on performance of identify essential proteins, the NMTF technique, as a widely used recommendation algorithm, is performed to reconstruct a most optimized PPI network with more potential protein–protein interactions. Then, we use the PageRank algorithm to compute the final ranking score of each protein, in which subcellular localization and homologous information of proteins were used to calculate the initial scores. In addition, extensive experiments are performed on the publicly available datasets and the results indicate that our NTMEP model has better performance in predicting essential proteins against the start-of-the-art method. In this investigation, we demonstrated that the introduction of non-negative matrix tri-factorization technology can effectively improve the condition of the protein–protein interaction network, so as to reduce the negative impact of noise on the prediction. At the same time, this finding provides a more novel angle of view for other applications based on protein–protein interaction networks.

Keywords: non-negative matrix factorization, protein-protein interaction, essential protein, PageRank, network

INTRODUCTION

Essential proteins play an indispensable role in the survival of organisms, and the criticality of proteins is mainly determined by their biological functions. Studies have shown that essential proteins have abundant functions such as translation, transcription, and replication (Glass et al., 2009). The prediction of essential proteins can apply the important reference information of biology and medicine, which has a wide application prospect in the fields of disease diagnosis and drug design. Currently, researchers have proposed a variety of biological methods to identify essential proteins, such as single-gene knockout (Kobayashi et al., 2003). However, these experimental methods have some limitations such as high cost and long time consumption. Therefore, it is urgent to improve the prediction performance of the computational method to identify essential proteins.

In recent years, researchers have proposed many computational methods to identify essential proteins relying on different ideas and technologies. Researchers have proposed many classic algorithms for predicting essential proteins based on PPI network topological characteristics, such as degree centrality (DC) (Hahn and Kern, 2005), information centrality (IC) (Björnsdottir, 2001), closeness centrality (CC) (Wuchty and Stadler, 2003), betweenness centrality (BC) (Joy et al., 2014), subgraph centrality (SC) (Estrada and Rodríguez-Velázquez, 2005), and sum of edge clustering coefficient centrality (NC) (Wang et al., 2012). Li et al. (2018) found that in the PPI network, the frequency of essential proteins in triangular structures is significantly higher than that of non-essential proteins. Based on this research discovery, they proposed a new measure of pure Centrality-Neighborhood Closeness Centrality (NCC). Although this type of approach allows direct identification of essential proteins in the absence of known essential proteins, there are limitations to these approaches. First, the existing PPI data are incomplete with a large number of false positives and false negatives, affecting the accuracy of predicting essential proteins. Second, most of these methods just use the topological properties of the network while ignoring other properties of essential proteins.

In order to make up for the limitations of incomplete protein interaction networks, many research groups have combined PPI networks with other biological information in recent years to improve the accuracy of essential protein identification. Tew et al. (2007) proposed a novel method called NFC, which defines the functional similarity between two proteins based on the GO term similarity and scores the protein based on the sum of the functional similarity between the protein and its neighboring proteins. Zhang et al. (2018) proposed an essential protein prediction method named TEO by combining the network topology characteristics, gene expression information, and GO annotation information. A weighted protein interaction network was established by calculating the Edge Clustering Coefficient (ECC), Pearson Correlation Coefficient (PCC), and functional similarity, so as to realize essential protein recognition. Lei et al. (2019) proposed an essential protein identification method called RWEP. Firstly, a weighted PPI network was established using network topology, gene expression, and GO annotations; then, each protein in the network was identified according to subcellular localization and protein complexes. Finally, the restart random walk algorithm is used to iteratively calculate the protein score in the weighted network. Due to the strong clustering of essential proteins, Ren et al. (2011) proposed a new centrality method that combines PPI network topology and protein complex information to identify essential proteins. By fusing the topological feature of PPI networks and gene expression information, Zhang et al. (2013) and Li et al. (2012) proposed two different models to predict essential proteins, called CoEWC and PeC, respectively. Based on the modular characteristics of essential proteins, Zhao et al. (2014) proposed an essential protein identification method called POEM. Based on the network topological characteristics and gene expression information, a highly reliable weighted network was established, and on this basis, overlapping functional modules with high cohesion and low coupling were dug. Finally, scores were calculated according to the weighted density of the modules to which the proteins belong, so as to realize the identification of essential proteins. Peng et al. (2012) considered that essential proteins were more conservative than non-essential proteins and often combined with each other. They proposed an iterative method ION that combines direct homology and PPI networks to predict essential proteins. The probability transfer matrix was established by using the edge clustering coefficient (ECC) and interaction network, and the initial score vector of protein was established by using homology information. According to the similarities of active PPI networks of each time, Peng et al. (Zhang et al., 2019) established a novel PPI network. Then, based on this network and orthologous information of protein, they developed a dynamic protein–protein interaction network-based model called FDP. Zhong et al. (2021) proposed a new measure method called JDC, which offers a dynamic threshold method to binarize gene expression data and combines Jaccard similarity index and degree centrality to predict essential proteins. However, the methods based on multisource data are relatively simple. It not only will conceal the complex relationship between the multisource data but also may introduce artificial noise.

In this article, we utilize non-negative matrix tri-factorization (NMTF) to deal with the challenges introduced above and propose a novel method named NTMEP for identifying essential proteins. NTMEP focuses on the following three important aspects. First, it is well known that the multiple kinds of biological data about proteins can be integrated to construct a weighted PPI network with similar functions. As a result, the more different types of data are used, the more artificial noise is produced inevitably. Considering this problem, NTMEP constructs the weighted PPI by using original protein–protein interaction information merely. Second, the NMTF algorithm is extensively used for many applications in pattern recognition, text mining, DNA gene expressions, and so on. This is also extended to community detection and the recommendation system. Hence, to mine more potential protein–protein associations, the NMTF algorithm is introduced in our progress. It takes the internal possibility of associations between proteins into account, which contributes to generation of a more reliable prediction model

that excludes the noisy candidates. Third, distinct from previous approaches, we employ homologous and subcellular localization information in the course of ranking proteins, which can improve the accuracy of predicting essential proteins effectively.

MATERIALS AND METHODS

Our purpose is to develop a novel method which can improve the accuracy of predicting essential proteins. We firstly constructed a weighted PPI network to represent the complex relationships between proteins. Moreover, a novel prediction method based on NMTF was proposed specifically for the network to find the potential associations between proteins. Finally, the PageRank algorithm was performed to identify the essential protein candidates by integrating subcellular localization and homologous information.

Let $G(V, E)$ be the PPI network that contains node set $V = (p_1, p_2, \ldots, p_n)$ (n is the number of proteins) and edge set $E = [(p_1, p_2, w_1), (p_2, p_3, w_2), \ldots, (p_i, p_j, w_m)]$ where (p_i, p_j, w_m) is the interaction between protein p_i and p_j with weighted value w_m which was set to 1 in original protein–protein interaction information.

Protein Association Measurement

In this subsection, a weighted PPI network was constructed in which the association value of two proteins would be calculated based on their topological characteristics. In analyzing the topological characteristics of PPI networks, researchers have found that the PPI networks are one kind of small-world and scale-free network. Therefore, the topological features of the PPI network can be used to predict essential proteins. In recent years, the item of common neighbors of two proteins in the PPI network has been used in many prediction algorithms to realize the task of predicting essential proteins. They demonstrate that the more common neighbors exist between two proteins, the more deeply is the association they have with other. In this article, if proteins p_i and p_j share at least one common neighbor, we assume that p_i and p_j are interacting. This kind of connection between proteins is called the co-neighbor (CoN) relationships and calculated as follows:

$$P_{CoN}(i,j) = \begin{cases} \dfrac{\left|S_{Nei}(i) \bigcap S_{Nei}(j)\right|^2}{(|S_{Nei}(i)| - 1) * (\left|S_{Nei}(j)\right| - 1)} & \text{if } |S_{Nei}(i)| > 1 \text{ and } \left|S_{Nei}(j)\right| > 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $S_{Nei}(i)$ and $S_{Nei}(j)$ present the neighborhood sets of p_i and p_j, respectively. As can be seen from the above equation, the value of the CoN relationships of the two-protein range is between 0 and 1.

Reconstruction of the Weighted PPI Network Based on NMTF

Non-negative matrix tri-factorization as a general technology takes or compresses a data matrix into a compact latent space. It has been used to model topics in text data (Hua et al., 2011), to predict cancer driver genes from clinical data (Xi et al., 2018), and to detect disease–disease associations (Žitnik et al., 2013). It is an efficient data representation technique, which has been widely used in recommender systems (Hernando et al., 2016; Luo et al., 2016). This new understanding should help to improve prediction accuracy of the essential proteins.

To take full advantage of NMTF, we perform it on the weighted PPI network (P_{CoN}) to mine the potential interactions of proteins. In contrast to classic non-negative matrix factorization (Lee and Seung, 1999) where the input matrix is separated into two parts, NMTF resolves the input matrix into three latent matrices. Here, we consider that the input adjacency matrix $P_{CoN} \in R^{n*n}$ has missing records, that is to say, the interactions between proteins have not been discovered. By using NMTF, a new matrix $Y \in R^{n*n}$ containing some new records would be constructed, as follows:

$$P_{CoN} \approx Y = FSG^T \quad (2)$$

Here, NMTF is designed to describe the matrix $P_{CoN} \in R^{n*n}$ with a product of three non-negative potential matrices$F \in R^{n*k}$, $S \in R^{k*k}$, and $G \in R^{n*k}$, while parameter k denotes factorization ranks and represents the number of potential vectors which form the column and row column space. For a given non-negative data matrix P_{CoN}, the issue can be solved as the following optimization problem:

$$D = \min J(F, S, G) = \left|\left|P_{CoN} - FSG^T\right|\right|_F^2 \quad (3)$$

where $||\cdot||_F$ is the Frobenius norm. Since the objective function in Eq. (3) is a joint non-convex problem, we employ the rule of multiplicative iteration to solve the objective function on the basis of using auxiliary functions. The squared Frobenius norm can be written as $||X||^2 = Tr(X^T X)$; therefore, Eq. (3) equals to:

$$D = Tr\left(P_{CoN}^T P_{CoN} - 2P_{CoN}^T FSG^T + GS^T F^T FSG^T\right) \quad (4)$$

Its partial derivative equations for factor F, S, and G are as follows, respectively:

$$\begin{aligned} \frac{\partial D}{\partial F} &= 2FSG^T GS^T - 2P_{CoN} GS^T \\ \frac{\partial D}{\partial S} &= 2F^T FSG^T G - 2F^T P_{CoN} G \\ \frac{\partial D}{\partial G} &= 2GS^T F^T FS - 2P_{CoN}^T FS \end{aligned} \quad (5)$$

It is well known that the static point can be detected using the Karush–Kuhn–Tucker (KKT) complementarity conditions. The KKT condition for factor F is as follows:

$$\frac{\partial D}{\partial F_{ik}} F_{ik} = 0 \quad (6)$$

In this connection, the conditions are assumed to be functional if the derivative is zero:

$$\left(FSG^TGS^T - P_{CoN}GS^T\right)_{iu} F_{iu} = 0$$

$$F_{iu} = F_{iu}\frac{\left(P_{CoN}GS^T\right)_{iu}}{\left(FSG^TGS^T\right)_{iu}} \tag{7}$$

Similarly, the updating rules for G and S can be derived as follows:

$$G_{iu} = G_{iu}\frac{\left(P_{CoN}^TFS\right)_{iu}}{\left(GS^TFFS^T\right)_{iu}}$$

$$S_{iu} = S_{iu}\frac{\left(F^TP_{CoN}G\right)_{iu}}{\left(F^TFSG^TG\right)_{iu}} \tag{8}$$

The multiplication iteration rules are shown as follows:

$$F_{iu} \leftarrow F_{iu}\frac{\left(P_{CoN}GS^T\right)_{iu}}{\left(FSG^TGS^T\right)_{iu}}$$

$$G_{iu} \leftarrow G_{iu}\frac{\left(P_{CoN}^TFS\right)_{iu}}{\left(GS^TFFS^T\right)_{iu}}$$

$$S_{iu} \leftarrow S_{iu}\frac{\left(F^TP_{CoN}G\right)_{iu}}{\left(F^TFSG^TG\right)_{iu}} \tag{9}$$

From the above Eq. (9), the optimal matrix Y, which is closest to P_{CoN}, can be computed. Finally, to recover the symmetry of the protein–protein interactions, we transformed the matrix Y to a symmetrical transition probability matrix P^*_{CoN}, as follows:

$$P^*_{CoN}(i,j) = \begin{cases} \dfrac{\max\left(Y_{ij}, Y_{ji}\right)}{\sum_{k=0}^{N} Y_{ik}}, & \sum_{k=0}^{N} Y_{ik} \neq 0 \\ 0, & \text{else} \end{cases} \tag{10}$$

The NMTF-Based Model for Identifying Essential Proteins

Through the description of the above algorithm, based on the information of the original PPI network, an optimized weighted PPI network can be established. Therefore, we can use an iterative method to rank protein scores. This method mainly includes two parts: the calculation of the initial score and the calculation of the ranking score, as detailed below.

Computation of Initial Scores

In this part, we will initially score each protein in the PPI network using homologous and subcellular localization information. Taking the *Saccharomyces cerevisiae* PPI network as an example, Tang et al. (2018) analyzed whether all the proteins in this network had direct homologous proteins in 99 reference species. They concluded that the more homologous a protein has in the reference species, the more likely it is to become a required protein. In order to obtain the given protein p_i in the PPI network $G = (V, E)$, we mainly use the homology information to calculate the homology score (S_H) of the protein. Among them, S_H (p_i) refers to the conservative score of p_i, and the calculation formula is as follows:

$$S_H(p_i) = \frac{H(p_i)}{\max\limits_{1\le j\le |V|}(H(p_j))} \tag{11}$$

Among them, $H(p_i)$ refers to the number of times that the protein p_i has direct homologous proteins in the reference species.

We know that an important feature of proteins is subcellular localization. By studying the characteristics of protein subcellular localization, researchers (Li et al., 2016; Zhao et al., 2016; Lei et al., 2018) found that essential proteins are more likely to appear in specific subcellular locations. Based on this, we calculated the subcellular localization score (S_L) of the protein based on the subcellular localization information. If the protein p_i exists in the final subcellular localization dataset R, then the frequency of each subcellular location r can be calculated by the following formula:

$$OF(r) = \frac{|SN(r)|}{\max\limits_{1\le k\le n}\left(\left|SN(k)\right|\right)} \tag{12}$$

where SN represents the relationship between the protein and the subcellular location data set, $SN(r)$ refers to the number of proteins corresponding to the subcellular location r, and n is the number of subcellular locations.

Based on a fixed protein p_i, the subcellular localization score S_L (p_i) refers to the highest score for all subcellular locations.

$$S_L(p_i) = \max_{r\in C(p_i)}(OF(r)) \tag{13}$$

where $C(p_i)$ represents the subcellular location corresponding to the protein p_i.

Finally, according to Eq. (11–13), the unique initial score $S_L(p_i)$ of protein p_i is expressed as follows:

$$S_I(p_i) = S_H(p_i) \times S_L(p_i) \tag{14}$$

Computation of Ranking Scores

The ranking of protein p_i is called $S_F(p_i)$, and $\sum_{p_j\in S_{CoN}(i)} P^*_{CoN}(p_i, p_j)S_F(p_j)$ refers to the neighbor induction score. Based on this, the ranking score of each protein in the PPI network can be calculated by Eq. (15), as shown below:

$$S_F(p_i) = \alpha \sum_{p_j\in S_{CoN}(i)} P^*_{CoN}(p_i, p_j)S_F(p_j) + (1-\alpha)S_I(p_i) \tag{15}$$

Among them, the function of the parameter α ($0 \le \alpha < 1$) is to adjust the weight of the two scores in the final ranking score. Based on the above analysis, the protein ranking score is a linear combination of its initial score and the neighborhood correlation score at the edge of the network. Therefore, formula (15) can be rewritten in matrix vector format as follows:

$$S_F = \alpha * P^*_{CoN} * S_F + (1-\alpha) * S_I \tag{16}$$

In our study, the Jacobi iterative method is used to solve Eq. (16), as shown below:

$$S^t_F = \alpha * P^*_{CoN} * S^{t-1}_F + (1-\alpha) * S_I \tag{17}$$

Algorithm 1 | NTMEP
Input: A PPI network G, subcellular localization information, homologous proteins information, stopping error ε, parameters k, α, and K
Output: Top K proteins sorted by S_F in descending order
Step 1: Calculate adjacency matrix P_{CoN} of the weighted PPI network according to Eq. (1)
Step 2: Reconstruct matrix P_{CoN} to P^*_{CoN} by Eq. (2)–(10)
Step 3: Initialize initial vector S_I with $S_F{}^0 = S_I$ and $t = 0$
Step 4: Compute $S_F{}^t$ according Eq. (17)
Step 5. If $||S_F{}^t - S_F{}^{t-1}|| < \varepsilon$, then $PR\ S_F = S_F{}^t$ and terminate the algorithm. Otherwise, let $t = t+1$ and repeat Step 4
Step 6. Sort proteins by the value of S_F in the descending order
Step 7. Output top K of sorted proteins

where S_F^t is the protein's scores obtained in the tth iteration.

Through the above analysis, we conclude that the overall framework of the NMTF-based model for the identification of essential protein (NTMEP) can be referred to as the following **Algorithm 1**.

RESULTS AND DISCUSSION

Experimental Data

In the experiments, we use four data sets including protein–protein interaction set, experimentally verified essential protein set, subcellular location set, and homologous protein information set. We downloaded the relationships among proteins from the DIP database (Xenarios et al., 2002), which includes 1,167 essential proteins and a total of 24,743 interactions between 5,093 proteins after removing self-interactions and duplicate interactions. Also, these data are adopted to construct the weighted protein network based on the topological structures. The experimentally verified essential protein dataset with 1,285 essential proteins are derived from MIPS (Mewes et al., 2006), SGD (Cherry et al., 1998), DEG (Zhang and Lin, 2009), and SGDP (Saccharomyces Genome Deletion Project, 2012). From the COMPARTMENTS (Binder et al., 2014) database, we obtained the subcellular location data, which cover 11 categories (Endoplasmic, Nucleus, Cytoskeleton, Golgi, Cytosol, Vacuole, Plasma, Mitochondrion, Endosome, Peroxisome, and Extracellular) (Peng et al., 2015). The homologous protein information is collected come from the seventh edition of the InParanoid database (Ostlund et al., 2010) including paired comparisons of 100 whole genomes (99 eukaryotes and one prokaryote).

Parameter α Sensitivity Analysis

In the NTMEP, the parameter α in Eq. (16), which used to weigh up the contribution of neighbor-induced score and initial score, was set to 0, 0.1, 0.2,..., and 1. While considering only the neighbor-induced score, α was set to 1. On the other hand, α was set to 0 when considering only the initial score. The impact of the parameter α to the performance of NTMEP is presented in **Table 1**. After the ranking scores of proteins were calculated with the different value of parameter α, we get the number of true essential proteins in the top 100, 200, 300, 400, 500, and 600 candidates, respectively. **Table 1** shows that the performance of the NTMEP is very poor when α was set to 0 or 1. It can be seen from the data in **Table 1** that the 0.1 and 0.2 groups have better prediction results. Especially, the best performance was achieved in the top 100 candidates when α was set to 0.1. Consequently, α was set to 0.2 in this article to make the NTMEP obtain good performance.

Comprehensive Comparison With Other Methods

To comparatively study the performance of NTMEP in predicting essential proteins, we also implement 10 types of representative essential proteins prediction methods, like DC (Joy et al., 2014), IC (Estrada and Rodríguez-Velázquez, 2005), CC (Wang et al., 2012), BC (Li et al., 2018), SC (Tew et al., 2007), NC (Zhang et al., 2018), PeC (Li et al., 2012), CoEWC (Zhang et al., 2013), POEM (Zhao et al., 2014), and JDC (Zhong et al., 2021), which are state-of-the-art prediction methods for the well essential protein prediction.

The higher number of essential proteins within the top k of the ranking list means the more real essential proteins are predicted successfully. Parameter k, which is set to 100, 200, 300, 400, 500, and 600, denotes the number of essential protein candidates selected. The number of real essential proteins within top k candidates is shown in **Figure 1**. NTMEP consistently outperformed the other competitive methods at various k cutoffs and ranked 92, 85.5, 78.7, 73.8, 69.4, and 65.2% of positive samples in top 100, 200, 300, 400, 500, and 600, respectively. Especially, as for the top 100 of essential protein candidates, NTMEP has higher predict accuracy 46, 48, 55, 48, 51, 37, 18, 19, 11, and 12% than that obtained from DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, and

TABLE 1 | The impact of parameter α to the performance of NTMEP.

	Top 100	Top 200	Top 300	Top 400	Top 500	Top 600
0	78	154	221	289	335	378
0.1	94	167	232	293	341	390
0.2	92	171	236	295	347	391
0.3	90	167	234	293	347	391
0.4	88	164	230	290	349	396
0.5	85	161	224	286	339	393
0.6	83	155	221	275	321	378
0.7	83	152	214	263	315	371
0.8	79	151	206	257	307	357
0.9	79	147	197	249	299	346
1	80	140	194	241	281	321

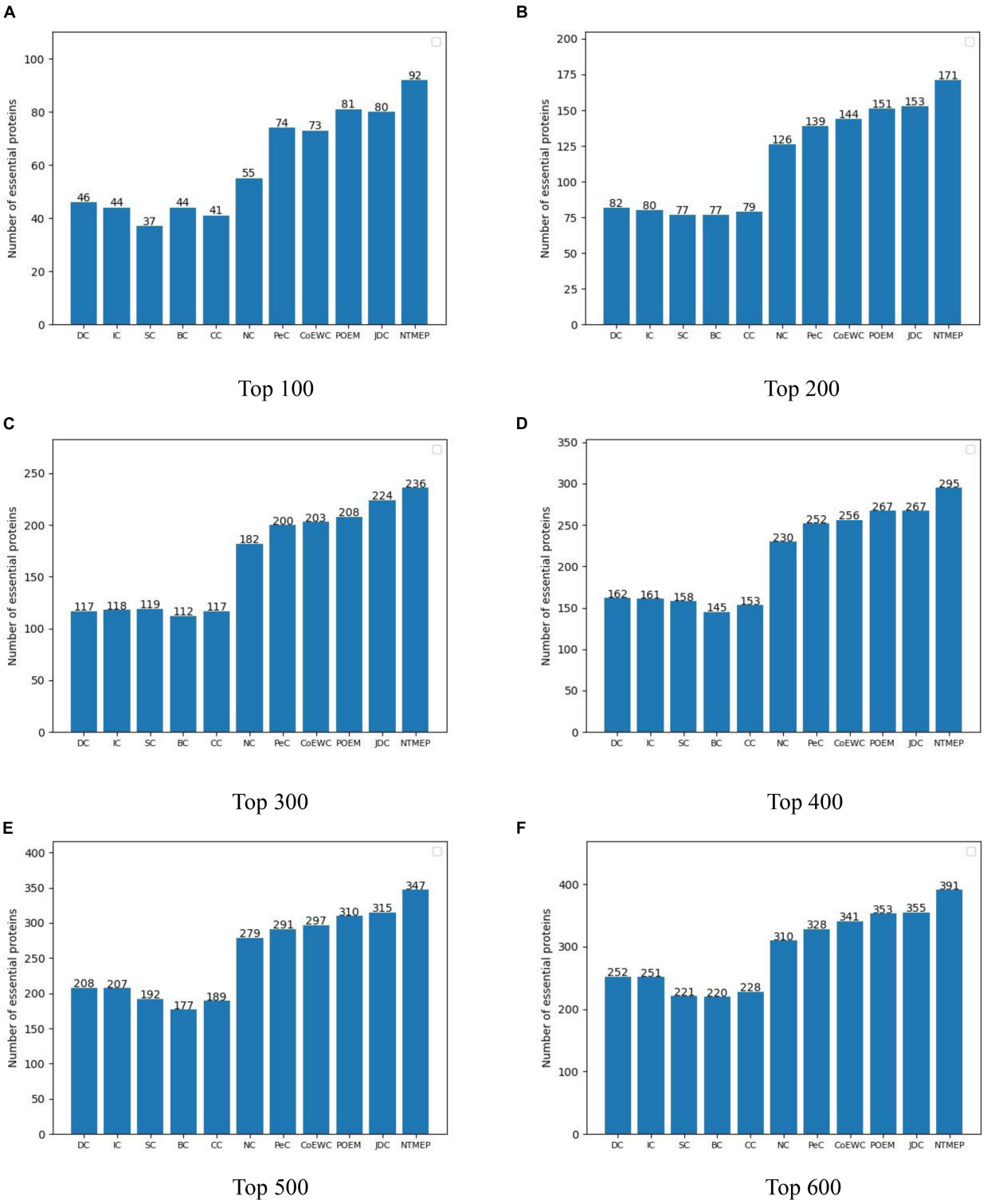


FIGURE 1 | Number of actual essential proteins identified by NTMEP and other ten previously competitive methods at various *k* values. **(A)** Top 100 ranked proteins; **(B)** Top 200 ranked proteins; **(C)** Top 300 ranked proteins; **(D)** Top 400 ranked proteins; **(E)** Top 500 ranked proteins; **(F)** Top 600 ranked proteins.

JDC, respectively. In those competitive methods, JDC had the best accuracy and ranked 80, 76.5, 74.7, 66.8, 63, and 59.2% in the top 100–600, respectively. Compared with JDC, NTMEP improved by 15% in top 100, 11.8% in top 200, 5.4% in top 300, 10.5% in top 400, 10.2% in top 500, and 10.1% in top 600.

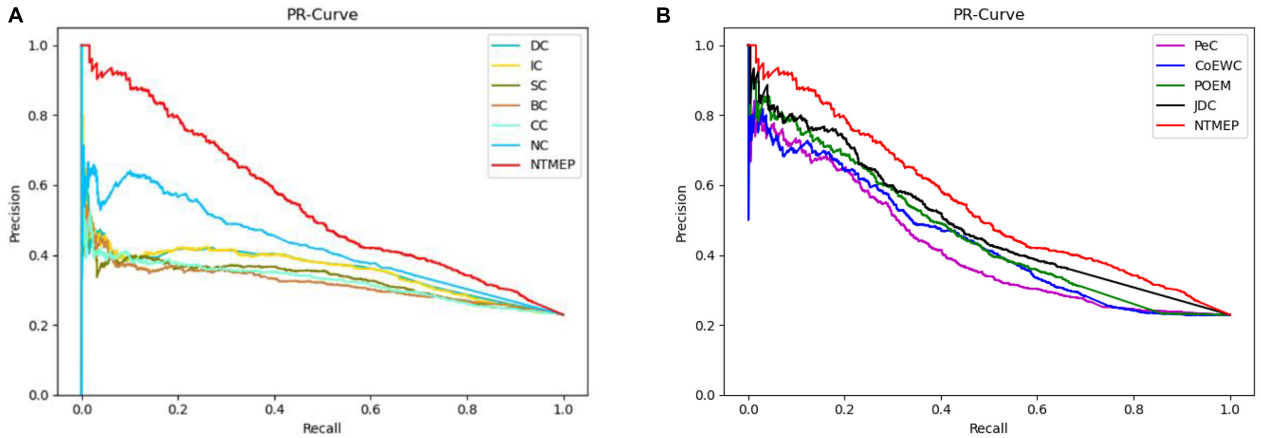


FIGURE 2 | Performance comparison between NTMEP and other ten methods in terms of PR curves. **(A)** PR curves of DC, IC, SC, BC, CC, NC, and NTMEP; **(B)** PR curves of PeC, CoEWC, POEM, JDC, and NTMEP.

Validated by Precision–Recall Curves

To obtain a fair and convincing comparison, the precision–recall (PR) curve is used to evaluate the prediction performance for essential proteins of our method and other state-of-the-art methods. The value of cutoffs, presented as *k*, is ranged from 1 to 5,093. We compute the scores of all proteins by using each algorithm and sorted it in descending order, respectively. The top *k* proteins are selected as a positive set, namely, essential protein candidates, and others as the negative set, namely, non-essential protein candidates. **Figure 2** compares the results obtained from the different methods. As shown in **Figure 2A**, compared with DC, IC, BC, CC, SC, and NC, the PR curves of NTMEP reported significantly higher capability for identifying essential proteins. The results obtained from our method and PeC, CoEWC, POEM, and JDC are presented in **Figure 2B**. Looking at **Figure 2B**, in the first part of the PR curve, it is apparent that the precision of our method has the best performance compared to those methods. In order to give quantitative comparison results, the area under the curve (AUC) values of the PR curve were computed, respectively, as shown in **Table 2**. As a whole, the NTMEP dramatically outperformed those competitive methods.

Validated by Jackknife Methodology

In this subsection, we employ the jackknife curves to assess the performance of our NTMEP method and other existing methods (DC, BC, CC, SC, IC, NC, PeC, CoEWC, POEM, and JDC), the various top number of ranked proteins as candidates. The jackknife curves of all the methods are displayed in **Figure 3**, where the horizontal axis denotes the number of proteins ranked at the top in descending order with each corresponding method, and the vertical axis is the accumulative quantity of the real essential proteins within the ranked proteins. **Figures 3A,B** illustrate the jackknife curves of all the competitive methods compared with NTMEP, respectively. As is seen from **Figure 3A**, the curve of NTMEP reported a higher number of real essential proteins than other existing centrality measure methods, such as DC, BC, CC, SC, IC, and NC. As shown in **Figure 3B**, NTMEP is also better than PeC, CoEWC, POEM, and JDC. To give quantitative comparison results, the AUC values of jackknife curve were computed, respectively, as shown in **Table 3**. From **Figure 3** and **Table 3**, it is clear that the NTMEP method outperforms the other 10 essential protein prediction methods.

In summary, these results demonstrated the powerful ability of NTMEP in identifying essential proteins. This finding is reasonable because our method adopts NMTF to find the potential interactions between proteins, which could provide additional interaction information and help to improve the prediction results by a large margin.

TABLE 2 | The AUC values of the PR curve obtained from NTMEP and other 10 competitive methods.

Method	NTMEP	DC	IC	SC	BC	CC
AUC value of PR curve	0.549	0.359	0.357	0.331	0.319	0.326
		NC	PeC	CoEWC	POEM	JDC
		0.425	0.492	0.463	0.439	0.417

Analysis of the Differences Between NTMEP and Other 10 Competitive Prediction Methods

This subsection will analyze the difference between NTMEP and other prediction methods through experimental results. Firstly, 11 protein sets were constructed by NTMEP and other 10 prediction methods (DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, and JDC), and each protein set contains the top 100 essential proteins predicted by each prediction method. The number of proteins that overlap between the NTMEP method and other methods and the number of proteins that differ are shown in **Table 4**.

In **Table 4**, Mi refers to one of the 10 prediction methods (DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, and JDC);

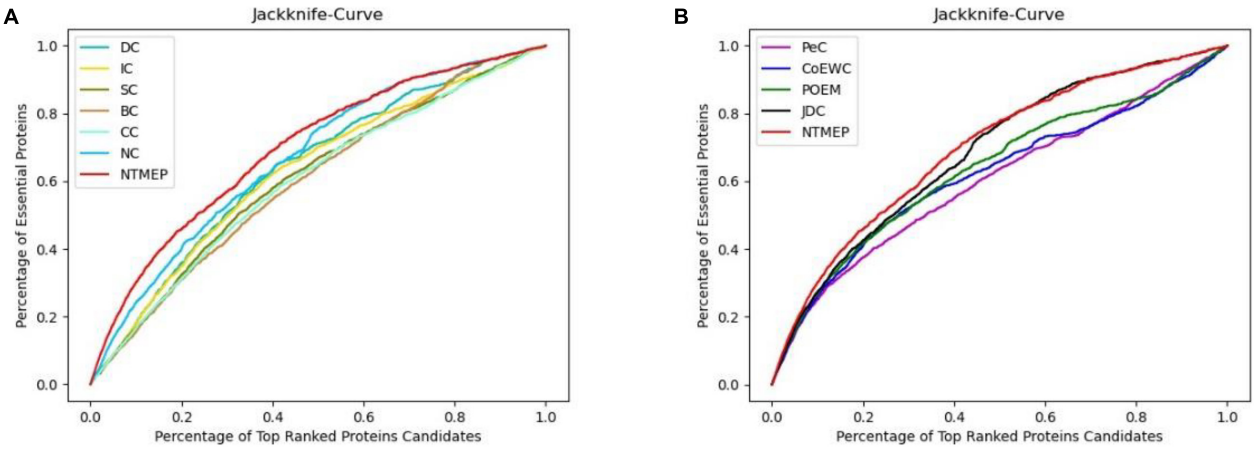


FIGURE 3 | Performance comparison between NTMEP and other ten methods in terms of Jackknife curves. **(A)** Jackknife curves of DC, IC, SC, BC, CC, NC, and NTMEP; **(B)** Jackknife curves of PeC, CoEWC, POEM, JDC, and NTMEP.

TABLE 3 | The AUC values of jackknife curve obtained from NTMEP and other 10 methods.

Method	NTMEP	DC	IC	SC	BC	CC
AUC value of jackknife curve	0.697	0.640	0.628	0.607	0.601	0.600
		NC	**PeC**	**CoEWC**	**POEM**	**JDC**
		0.670	0.603	0.618	0.635	0.684

|Mi∩NTMEP| represents the number of common proteins predicted by both Mi and NTMEP in the top 100 ranked proteins. {Mi-NTMEP} refers to the difference set in the top 100 ranked proteins, while proteins were selected as essential proteins by Mi but not by NTMEP. Moreover, | Mi-NTMEP| represents the number of proteins in the difference set. Similarly, {NTMEP-Mi} denotes the difference set constituted by the proteins belonging to NTMEP but not to Mi, and the number is denoted by |NTMEP-Mi|.

As shown in **Table 4**, the second row of the table shows that 85 essential protein candidates out of the top 100 essential protein candidates predicted by DC are different from those predicted by NTMEP, while 32 of these 85 predicted essential protein candidates are true essential proteins; thus, the percentage of essential proteins in the difference set is 37.6%. Among the top 100 essential protein candidates predicted by NTMEP, 85 essential protein candidates were different from those predicted by DC, but 78 of them were accurate; thus, the percentage of essential proteins in the difference set was 91.8%. From this line of data, it can be seen that most of the top 100 essential protein candidates predicted by NTMEP are different from those candidates predicted by DC. Moreover, NTMEP predicts far more true key proteins than DC. This indicates that NTMEP not only is a different method from DC but also shows that NTMEP is much better than DC in distinguishing essential proteins from common proteins. Similarly, it can be seen from the other rows of the table that NTMEP maintains this advantage over all other prediction methods.

CONCLUSION

In reviewing the literature, previous studies developed many computational methods to predict essential proteins effectively.

TABLE 4 | Comparison of the overlap and difference of the top 100 proteins identified by NTMEP and other 10 methods.

Methods (Mi)	| Mi∩NTMEP|	| NTMEP-Mi| and | Mi-NTMEP|	Number of essential proteins in {Mi-NTMEP}	Number of essential proteins in {NTMEP-Mi }	Percentage of essential proteins in {Mi-NTMEP}	Percentage of essential proteins in {NTMEP-Mi }
DC	15	85	32	78	37.6%	91.8%
IC	15	85	30	78	35.3%	91.8%
SC	12	88	25	80	28.4%	90.9%
BC	11	89	34	82	38.2%	92.1%
CC	13	87	29	80	33.3%	92.0%
NC	33	67	25	62	37.3%	92.5%
PeC	44	56	33	50	58.9%	91.1%
CoEWC	46	54	30	49	55.6%	90.7%
POEM	49	51	35	46	68.6%	90.2%
JDC	40	60	42	54	70.0%	90.0%

However, these methods do not take full account of the false-positive and -negative noise generated from high-throughput experimentation and the process of the weighted PPI network construction. To get the utmost out of the complex association between proteins, NMTF is introduced into our proposed method. Moreover, subcellular localization and homologous protein information are used in the final scoring stage instead of the stage of establishing the weighted network. Also, a comprehensive experiment is carried out and the results show that our new method can obtain a better performance compared with other methods. A possible explanation for these results might be that there are deep relationships between proteins which are not founded by high-throughput experimentation, and fusion of multiple data raises the cost and reduces the overall efficiency of the process. These results add to the rapidly expanding field of computational methods for predicting essential proteins. It is unfortunate that the study did not solve the problem of noise generated by multisource data fusion. This is an important issue for future research.

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

ZZ and MJ obtained the protein–protein interaction data, benchmark essential protein dataset, subcellular location data, and homologous protein information. ZZ, MJ, and XQ designed the new method, NTMEP, and analyzed the results. ZZ, DW, and WZ drafted and revised the manuscript together. All authors have read and approved the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61772089, 61873221, and 61672447; Natural Science Foundation of Hunan Province (2018JJ3566, 2018JJ3565, 2018JJ4058, and 2020JJ4648); Major Scientific and Technological Projects for Collaborative Prevention and Control of Birth Defects in Hunan Province (2019SK1010); Research Foundation of Education Bureau of Hunan Province (19A048, 18A441, and 18C0958); Educational Planning Key Project of Hunan Province (XJK18DJA1); and Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals (2018TP1027).

REFERENCES

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., et al. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* 2014:bau012. doi: 10.1093/database/bau012

Björnsdottir, K. (2001). Language, research and nursing practice. *J. Adv. Nurs.* 33, 159–166. doi: 10.1111/j.1365-2648.2001.01648.x

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., et al. (1998). SGD: *Saccharomyces* genome database. *Nucleic Acids Res.* 26, 73–79. doi: 10.1093/nar/26.1.73

Estrada, E., and Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys.* 71:056103. doi: 10.1103/PhysRevE.71.056103

Glass, J. I., Hutchison, C. A., Smith, H. O., and Venter, J. C. (2009). A systems biology tour de force for a near-minimal bacterium. *Mol. Syst. Biol.* 5:330. doi: 10.1038/msb.2009.89

Hahn, M. W., and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806. doi: 10.1093/molbev/msi072

Hernando, A., Bobadilla, J., and Ortega, F. (2016). A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowl. Based Syst.* 97, 188–202. doi: 10.1016/j.knosys.2015.12.018

Hua, W., Nie, F., Huang, H., and Makedon, F. (2011). "Fast nonnegative matrix Tri-factorization for large-scale data co-clustering," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence. DBLP*, 2011, (Barcelona, CA).

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2014). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 96–103. doi: 10.1155/JBB.2005.96

Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G., Andersen, K. K., Arnaud, M., et al. (2003). Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* 100, 4678–4683. doi: 10.1073/pnas.0730515100

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

Lei, X., Fang, M., Wu, F. X., and Chen, L. (2018). Improved flower pollination algorithm for identifying essential proteins. *BMC Syst. Biol.* 12:46. doi: 10.1186/s12918-018-0573-y

Lei, X., Yang, X., and Fujita, H. (2019). Random walk based method to identify essential proteins by integrating network topology and biological characteristics. *Knowl. Based Syst.* 167, 53–67. doi: 10.1016/j.knosys.2019.01.012

Li, G., Li, M., Wang, J. X., Li, Y., and Pan, Y. (2018). United neighborhood closeness centrality and orthology for predicting essential proteins. *IEEE ACM Trans. Comput. Biol. Bioinform.* 17, 1451–1458. doi: 10.1109/TCBB.2018.2889978

Li, G., Li, M., Wang, J. X., Wu, J., Wu, F. X., and Pan, Y. (2016). Predicting essential proteins based on subcellular localization, orthology and PPI networks. *BMC Bioinformatics* 17:279. doi: 10.1186/s12859-016-1115-5

Li, M., Zhang, H., Wang, J. X., and Pan, Y. (2012). A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* 6:15. doi: 10.1186/1752-0509-6-15

Luo, X., Zhou, M., Li, S., You, Z., Xia, Y., and Zhu, Q. (2016). A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 579–592. doi: 10.1109/TNNLS.2015.2415257

Mewes, H. W., Frishman, D., Mayer, K. F. X., Munsterkotter, M., Noubibou, O., Pagel, P., et al. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 34, D169–D172. doi: 10.1093/nar/gkj148

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., et al. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Peng, W., Wang, J. X., Wang, W., Liu, Q., Wu, F. X., and Pan, Y. (2012). Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks. *BMC Syst. Biol.* 6:87. doi: 10.1186/1752-0509-6-87

Peng, X., Wang, J., Zhong, J., Luo, J., and Pan, Y. (2015). "An efficient method to identify essential proteins for different species by integrating protein subcellular localization information," in *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, (Washington, DC), 277–280.

Ren, J., Wang, J. X., Li, M., Wang, H., and Liu, B. (2011). Prediction of essential proteins by integration of PPI network topology and protein complexes information. *Bioinform. Res. Appl.* 6674, 12–24. doi: 10.1007/978-3-642-21260-4_6

Saccharomyces Genome Deletion Project (2012). *Saccharomyces Genome Deletion Project*. Available online at: http://yeastdeletion.stanford.edu/ (accessed June 20, 2012).

Tang, X., Li, X., Hu, S., and Zhao, B. (2018). A framework for identifying functional modules in dynamic networks. *Int. J. Data Mining Bioinform.* 21, 1–17. doi: 10.1504/IJDMB.2018.095554

Tew, K. L., Li, X. L., and Tan, S. H. (2007). Functional centrality: detecting lethality of proteins in protein interaction networks. *Genome Inform.* 19, 166–177.

Wang, J. X., Li, M., Wang, H., and Pan, Y. (2012). Identification of essential proteins based on edge clustering coefficient. *IEEE ACM Trans. Comput. Biol. Bioinform.* 9, 1070–1080. doi: 10.1109/TCBB.2011.147

Wuchty, S., and Stadler, P. F. (2003). Centers of complex networks. *J. Theor. Biol.* 223, 45–53. doi: 10.1016/s0022-5193(03)00071-7

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Xi, J., Li, A., and Wang, M. (2018). A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. *Neurocomputing* 296, 64–73. doi: 10.1016/j.neucom.2018.03.026

Zhang, F., Peng, W., Yang, Y., Dai, W., and Song, J. (2019). A novel method for identifying essential genes by fusing dynamic protein-protein interactive networks. *Genes (Basel)* 10:31. doi: 10.3390/genes10010031

Zhang, R., and Lin, Y. (2009). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic Acids Res.* 37, D455–D458. doi: 10.1093/nar/gkn858

Zhang, W., Xu, J., Li, Y., and Zou, X. (2018). Detecting essential proteins based on network topology, gene expression data, and gene ontology information. *IEEE ACM Trans. Comput. Biol. Bioinform.* 15, 109–116. doi: 10.1109/TCBB.2016.2615931

Zhang, X., Xu, J., and Xiao, W. X. (2013). A new method for the discovery of essential proteins. *PLoS One* 8:e58763. doi: 10.1371/journal.pone.0058763

Zhao, B. H., Wang, J. X., Li, M., Wu, F. X., and Pan, Y. (2014). Prediction of essential proteins based on overlapping essential modules. *IEEE Trans. Nanobiosci.* 13, 415–424. doi: 10.1109/TNB.2014.2337912

Zhao, B. H., Wang, J. X., Li, X., and Wu, F. X. (2016). Essential protein discovery based on a combination of modularity and conservatism. *Methods* 110, 54–63. doi: 10.1016/j.ymeth.2016.07.005

Zhong, J., Tang, C., Peng, W., Xie, M., Sun, Y., Tang, Q., et al. (2021). A novel essential protein identification method based on PPI networks and gene expression data. *BMC Bioinformatics* 22:248. doi: 10.21203/rs.3.rs-55902/v2

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* 3:3202. doi: 10.1038/srep03202

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

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

frontiers in Genetics

METHODS
published: 10 August 2021
doi: 10.3389/fgene.2021.708981

deepMNN: Deep Learning-Based Single-Cell RNA Sequencing Data Batch Correction Using Mutual Nearest Neighbors

Bin Zou[1], Tongda Zhang[1], Ruilong Zhou[1,2], Xiaosen Jiang[1,2], Huanming Yang[1,3], Xin Jin[1,4,5] and Yong Bai[1]**

[1] BGI-Shenzhen, Shenzhen, China, [2] College of Life Science, University of Chinese Academy of Sciences, Beijing, China, [3] James D. Watson Institute of Genome Sciences, Hangzhou, China, [4] School of Medicine, South China University of Technology, Guangzhou, China, [5] Guangdong Provincial Key Laboratory of Human Disease Genomics, Shenzhen Key Laboratory of Genomics, BGI-Shenzhen, Shenzhen, China

It is well recognized that batch effect in single-cell RNA sequencing (scRNA-seq) data remains a big challenge when integrating different datasets. Here, we proposed deepMNN, a novel deep learning-based method to correct batch effect in scRNA-seq data. We first searched mutual nearest neighbor (MNN) pairs across different batches in a principal component analysis (PCA) subspace. Subsequently, a batch correction network was constructed by stacking two residual blocks and further applied for the removal of batch effects. The loss function of deepMNN was defined as the sum of a batch loss and a weighted regularization loss. The batch loss was used to compute the distance between cells in MNN pairs in the PCA subspace, while the regularization loss was to make the output of the network similar to the input. The experiment results showed that deepMNN can successfully remove batch effects across datasets with identical cell types, datasets with non-identical cell types, datasets with multiple batches, and large-scale datasets as well. We compared the performance of deepMNN with state-of-the-art batch correction methods, including the widely used methods of Harmony, Scanorama, and Seurat V4 as well as the recently developed deep learning-based methods of MMD-ResNet and scGen. The results demonstrated that deepMNN achieved a better or comparable performance in terms of both qualitative analysis using uniform manifold approximation and projection (UMAP) plots and quantitative metrics such as batch and cell entropies, ARI F1 score, and ASW F1 score under various scenarios. Additionally, deepMNN allowed for integrating scRNA-seq datasets with multiple batches in one step. Furthermore, deepMNN ran much faster than the other methods for large-scale datasets. These characteristics of deepMNN made it have the potential to be a new choice for large-scale single-cell gene expression data analysis.

Keywords: scRNA-seq data integration, batch effect correction, residual network, mutual nearest neighbor, deep learning

Abbreviations: MNN, mutual nearest neighbor; ARI, adjusted rand index; ASW, average silhouette width; PCA, principal component analysis; UMAP, uniform manifold approximation and projection; RAM, random access memory; GPU, graphics processing unit.

OPEN ACCESS

Edited by:
Feng Liu,
Wuhan University, China

Reviewed by:
Shixiong Zhang,
Xidian University, China
Jun Li,
University of Notre Dame, United States

****Correspondence:***
Xin Jin
jinxin@genomics.cn
Yong Bai
baiyong@genomics.cn

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

Received: *13 May 2021*
Accepted: *16 July 2021*
Published: *10 August 2021*

Citation:
Zou B, Zhang T, Zhou R, Jiang X, Yang H, Jin X and Bai Y (2021) deepMNN: Deep Learning-Based Single-Cell RNA Sequencing Data Batch Correction Using Mutual Nearest Neighbors. Front. Genet. 12:708981. doi: 10.3389/fgene.2021.708981

INTRODUCTION

High-throughput single-cell RNA sequencing (scRNA-seq) has enabled the gene expression profiling of a large number of individual cells at a single-cell resolution, offering unprecedented insights into the transcriptomic characterization of cell heterogeneity and dynamics (Stegle et al., 2015; Consortium, 2018; Han et al., 2018; Svensson et al., 2018). Considerable efforts have been made over the past decade to promote the rapid development of this technology, leading to massive single-cell gene expression data compiled from different experiments at different times and even with various sequencing platforms. However, like other sequencing technologies, these differences inevitably cause an unexpected batch effect due to the technical or biologically irrelevant variations across batches (Goh et al., 2017; Tran et al., 2020). The batch effect in the scRNA-seq data has been plaguing downstream analysis as it may interrupt the gene expression patterns. Consequently, the issue of batch effect may lead to a spurious conclusion when jointly investigating the comprehensive biological process of cells on the basis of integrating multiple datasets. Hence, batch effect correction is crucial for analyzing scRNA-seq data, allowing investigators to capture the intrinsically biological features across batches.

Currently, a myriad of batch effect correction algorithms has been proposed to tackle the problem (Tran et al., 2020). MNNCorrect (Haghverdi et al., 2018) assumed the orthogonality of batch effect to the biological manifold and corrected batch effect by calculating average difference in the high-dimensional gene expression space between similar cells across batch pairs (called mutual nearest neighbors, MNNs). Yet due to its high consumption of memory usage and CPU runtime, a number of methods were further developed to enhance the performance, for example, fastMNN (Haghverdi et al., 2018) and Seurat Integration (Seurat V3) (Stuart et al., 2019) followed the MNN scheme to carry out MNN search in a subspace by applying principal component analysis (PCA) and canonical correlation analysis (CCA), respectively. Scanorama (Hie et al., 2019) performed a faster approximate nearest neighbor search in the low-dimensional space computed by the randomized singular value decomposition. BBKNN (Polański et al., 2020) found MNNs in a low-dimensional, reduced space by computing k nearest neighbors and transformed the neighbor information into connectivity to construct a graph that linked all cells across batches. Harmony (Korsunsky et al., 2019) projected cells across different batches into a PCA space, followed by iteratively grouping similar cells into multiple clustering while simultaneously maximizing the diversity of batches within each cluster. LIGER (Welch et al., 2019) employed integrative non-negative matrix factorization to reduce the dimension and identified shared and batch-specific features across datasets. It then detected joint clusters and normalized the factor loading quantiles to perform batch correction. scMerge (Lin et al., 2019) constructed a graph connecting mutual nearest clusters between batches to remove batch effects.

Deep learning-based methods for single-cell analysis have experienced a tremendous progress in recent years and were already applied to remove batch effects in scRNA-seq data, for instance, MMD-ResNet (Shaham et al., 2017) has attempted to remove batch effect by minimizing the maximum mean discrepancy (MMD) using residual neural networks. BERMUDA (Wang et al., 2019) sought to remove batch effect locally based on MMD loss between similar cell clusters using an autoencoder structure. scGen (Lotfollahi et al., 2019) corrected batch effect based on the distributions of the cells that were inferenced from a reference dataset using a variational autoencoder model. However, scGen was a supervised method that required cell types in advance. scGAN (Bahrami et al., 2020) labeled multiple batches of the input cells that were represented in latent embedding space using a generative adversarial network model.

Although several batch correction methods are available, most of them struggle with excessive running time or resource requirements, which are likely to be further exacerbated as the cell numbers of scRNA-seq experiments continue growing. In this study, we propose deepMNN, a deep learning-based scRNA-seq batch correction model using MNN. We first identified MNN pairs among batches in a PCA subspace. A residual-based batch correction network was then constructed and employed to remove batch effects based on these MNN pairs. The overall loss of deepMNN was designed as the sum of a batch loss and a weighted regularization loss. The batch loss was used to compute the distance between cells in MNN pairs in the PCA subspace, while the regularization loss was to make the output of the network similar to the input. We compared the performance of deepMNN with state-of-the-art batch correction methods, including the widely used methods of Harmony, Scanorama, and Seurat V4, as well as the recently developed deep learning-based methods of MMD-ResNet and scGen. To comprehensively investigate the performance of these methods, we employed different scRNA-seq datasets under various scenarios, such as datasets with non-identical cell types, datasets with multiple batches, and large-scale datasets. In addition to qualitative analysis using uniform manifold approximation and projection (UMAP) plots, we calculated three metrics to quantitatively compare their performance on batch correction, including batch and cell type entropies, adjusted rand index (ARI) F1 score, and average silhouette width (ASW) F1 score. The experiment results showed that, in comparison to other correction methods, deepMNN not only reached a better or comparable performance in terms of the quantitative metrics and running time but also allowed for integrating scRNA-seq datasets with multiple batches in one step.

MATERIALS AND METHODS

Architecture of deepMNN

The deepMNN encompassed two main steps: pre-processing and batch correction (**Figure 1A**). The pre-processing step followed the standard workflow for scRNA-seq data analysis in Scanpy (Wolf et al., 2018), such as quality control (QC), filtering, normalization, identification of highly variable genes, scaling, and linear dimensional reduction using PCA. The dimensional-reduced data X^{pca} was used to find MNN pairs among the different batches. In the batch correction step, the scaled data

was fed into the batch correction network, and the output was further transformed into the PCA subspace. Here, the batch correction network was formed by the stack of two residual blocks. Each residual block received an input x and computed output $y = x + \delta(x)$, where $\delta(x)$ was the output of the residual block (**Figure 1B**). The batch loss measured the distance between cells in MNN pairs in the PCA subspace. We also used a regularization loss to make the output of batch correction network resemble the input.

Data Pre-processing

The steps of data pre-processing for scRNA-seq data included (1) QC and filtering, which was performed to remove the unwanted cells based on user-defined criteria, (2) normalization, the gene expression measurements for each cell were normalized by the total expression, followed by multiplication of a scale factor of 10,000, and (3) log-transformation, the normalized data was processed using log-transformation. Subsequently, 2,000 highly variable genes (HVGs, i.e., genes exhibiting high cell-to-cell variation in the dataset) were identified. We then scaled the data by calculating the z-score for each gene expression to have zero mean and unit variance. It should be noted that the z-score values exceeding the standard deviation of 10 were clipped. Next, we applied PCA on the scaled data and reduced the dimension using the first 50 principal components (PCs) empirically. The resulting matrix X^{pca} was further used to find MNN pairs across different batches. In addition, the first 50 PCs were also used to reduce the dimension of the outputs from the batch correction network as well (**Figure 1A**).

Searching for MNN Pairs Among Batches

To find MNN pairs across batches, deepMNN searched 20 nearest neighbors for every cell in one batch from the remaining other batches in the dimensional-reduced PCA subspace. After repeating this process for all batches, we identified MNN pairs where a cell in one batch is the nearest neighbor of a cell in another batch and *vice versa*. Since the computational load of nearest neighbor queries was exponential in the size of the dataset, we improved the efficiency of our method using an approximate nearest neighbor searching algorithm that was implemented in the Annoy package[1].

Batch Correction Network

Inspired by the well-known residual network, the batch correction network was formed by the stack of two residual blocks. A residual block received an input x (or the output of the previous block) and computed output $y = x + \delta(x)$, where $\delta(x)$ is a residual term resulting from two sequences of three consecutive layers: weight layer, batch normalization layer, and PReLU activation layer (**Figure 1B**). The first weight layer in a residual block had $2 \times d$ nodes, while the second weight layer had d nodes, where d is the input dimension of the residual block.

In our work, the initial input into the batch correction network was the scaled data with 2,000 selected HVGs. Consequently, the number of nodes in the first and the second weight layers of the first residual block was 4,000 and 2,000, respectively. The number of nodes in the two weight layers of the second residual block was correspondingly the same as that in the first residual block. Therefore, the number of nodes in the output layer of the batch correct network was 2,000.

Loss Function

There were two types of losses in this study: (1) the batch loss that was the sum of the Euclidean distances between cells in the MNN

[1]https://github.com/spotify/annoy

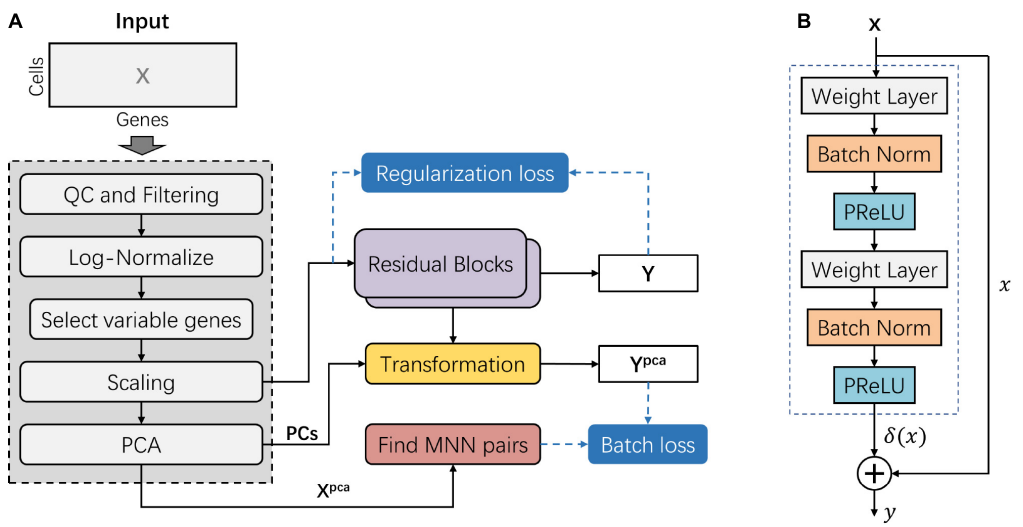


FIGURE 1 | Overview of the deepMNN framework. **(A)** Illustration of the deepMNN workflow that was comprised of data pre-processing, principal component analysis transformation, mutual nearest neighbor pair search, batch correction network with a stack of two residual blocks, and calculation of batch loss and regularization loss. **(B)** Residual block comprised of two sequences of three consecutive layers, including the weight layer, batch normalization layer, and PReLU activation layer.

pairs and (2) the regularization loss aimed to make the output of the network similar to the input.

To compute the batch loss, we first calculated the dimensional-reduced vector Y_i^{pca} for cell i as follows:

$$Y_i^{\text{pca}} = Y_i \cdot \text{PCs}$$

where Y_i is the output of the batch correction network for cell i; PCs are the first 50 principal components as described in section "Data Pre-processing." Suppose two cells i and j were in the MNN pair k and, thus, denoted as $Y_{i_k}^{\text{pca}}$ and $Y_{j_k}^{\text{pca}}$, respectively. Then, the batch loss L_{b} can be written as follows:

$$L_{\text{b}} = \sum_k \left|\left|Y_{i_k}^{\text{pca}} - Y_{j_k}^{\text{pca}}\right|\right|_2$$

where $\left|\left|Y_{i_k}^{\text{pca}} - Y_{j_k}^{\text{pca}}\right|\right|_2$ represents the Euclidian distance between cells i and j in the MNN pair k, $k = 1,2,3,\ldots, K$, and K is the total number of MNN pairs.

We hypothesized that the cells in an MNN pair had the same cell type, their distance should be small when no batch effect existed, and hence the batch loss was used to remove the batch effect between different batches. However, if the batch correction network had a zero vector output, the batch loss should have been zero, which was not our expectation. As such, we further utilized a regularization loss to make the output of the network not far away from the input.

The regularization loss L_{r} was defined as the sum of the Euclidian distances between the output and the input of the batch correction network.

$$L_{\text{r}} = \sum_i \left|\left|Y_i - X_i\right|\right|_2$$

where Y_i is the output of the batch correction network of cell i, and X_i is the cell i in the scaled data with 2,000 HVGs.

Finally, the overall loss of deepMNN was defined as the combination of a batch loss and a weighted regularization loss:

$$\text{L} = L_{\text{b}} + \alpha \cdot L_{\text{r}}$$

The value of α was set as 0.001 in our experiments.

Hyperparameters for Training deepMNN

We trained the deepMNN batch correction network *de novo* with default initialization of weights as provided by the PyTorch library (version 1.6.0). We employed the Adam optimizer (Kingma and Ba, 2014) with default parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and a batch size of 1,024 for all experiments. The maximum number of epochs was set as 200. The training procedure would stop early when the total loss did not decrease for 10 consecutive epochs. The learning rate (LR) was initialized as 0.1 and decayed by 0.8 every 20 epochs. In general, the hyperparameters of the network were manually optimized. We searched primarily over the residual block structure, empirically chose the number of the residual blocks, and manually tuned the LR to obtain optimal performance.

Batch Correction Through Other Methods

Three widely used methods of Harmony, Scanorama, and Seurat V4 and two deep learning-based methods of MMD-ResNet and scGen were used to compare the performance on batch correction with deepMNN.

We first applied the same data pre-processing as described in section "Data Pre-processing" for all these methods, including QC and filtering, normalization, and log-transformation. For Harmony, the first 50 PCs were determined by applying PCA on the pre-processed data, followed by utilization of the RunHarmony function in its R package (version 0.1.0) to conduct the batch correction experiments. The parameters of maximum clusters and maximum iterations were set as 50 and 100, respectively. For Scanorama, we first identified 2,000 HVGs after data pre-processing and then employed its Python implementation (version 1.7.1) to perform the experiments with default parameter settings. For Seurat V4, we followed the Seurat integration workflow recommended by the Seurat package (version 4.0.3). Briefly, we first selected 2,000 HVGs from the pre-processed data and then computed the anchors using the FindIntegrationAnchors function, followed by integration of the batches using the IntegrateData function to accomplish the experiments. For MMD-ResNet, the PyTorch implementation[2] was used to perform the experiments. After data pre-processing and dimension reduction using PCA, we selected the first 50 PCs to train the MMD-ResNet model with default hyperparameters but with a batch size of 256. The training stopped when the loss did not decrease for five consecutive epochs. For scGen, we used the PyTorch implementation (version 2.0.0) to carry out the experiments in our work. We selected the top 7,000 HVGs by default from the pre-processed data to train the scGen model with default hyperparameters except for epochs of 100 and a batch size of 32.

To assess the performance of each method including deepMNN, the top 50 PC vectors extracted from the batch-corrected expression matrix were used for the calculation of evaluation metrics and visualization.

Datasets

Human Peripheral Blood Mononuclear Cell

The data included two batches of human peripheral blood mononuclear cells (PBMCs) from two healthy donors, which were generated by the 3′ and 5′ Genomics protocols, respectively (Zheng et al., 2017). The data and the cell type annotated by Polański et al. (2020) were downloaded from ftp://ngs.sanger.ac.uk/production/teichmann/BBKNN/PBMC.merged.h5ad. We excluded cells without annotation and only retained common genes, resulting in nine different cell types for a total of 8,098 cells in the 3′ batch and 7,378 cells in the 5′ batch, each with 17,430 genes.

Human Pancreas

The data consisted of five published pancreas datasets: Baron (GSE84133) (Baron et al., 2016), Muraro (GSE85241)

[2]https://github.com/ushaham/batchEffectRemoval2020

(Muraro et al., 2016), Segerstolpe (E-MTAB-5061) (Segerstolpe et al., 2016), Wang (GSE83139) (Wang et al., 2016), and Xin (GSE81608) (Xin et al., 2016), generated using inDrop, CEL-Seq2, SMART-Seq2, SMARTer, and SMARTer protocols, respectively. The data batches and annotations were downloaded from https://hemberg-lab.github.io/scRNA.seq.datasets/human/pancreas/. We removed the cells annotated with unknown cell types and only retained the genes detected in all batches. As a result, there were 15 different cell types for a total of 14,767 cells, each with 15,558 genes.

Human Cell Atlas

The Human Cell Atlas (HCA) dataset was downloaded from https://github.com/immunogenomics/harmony2019/tree/master/data/figure3, processed by Korsunsky et al. (2019). This data had two batches, including 275,264 bone marrow cells and 253,024 cord blood cells, respectively (Li et al., 2018). 10× Genomics protocol was used to generate the data, and 24,823 genes were acquired for each cell. We removed the cell types whose number of cells was less than 200, resulting in 22 different cell types for a total of 528,014 cells.

Evaluation Metrics for Batch Correction

To assess the batch correction performance of deepMNN and other methods as described above, we calculated three types of metrics, batch and cell type entropies (Chazarra-Gil et al., 2021), ARI F1 score (Hubert and Arabie, 1985; Tran et al., 2020), and ASW F1 score (Rousseeuw, 1987; Tran et al., 2020).

Batch and Cell Type Entropies

The entropies of batch and cell type can be used to measure batch mixing and cell type separation. To compute the batch and cell type entropies, we first constructed a KNN graph where each cell was a node and connected to its 20 nearest neighbors. Then, the batch entropy E_i^{b} and cell type entropy E_i^{c} for cell i were calculated as follows:

$$P_{i\mathrm{b}} = \frac{\mathrm{N}_{i\mathrm{b}}}{\mathrm{N}_i}$$

$$E_i^{\mathrm{b}} = -\frac{1}{B}\sum_{\mathrm{b}} P_{i\mathrm{b}}\log\left(P_{i\mathrm{b}}\right)$$

$$P_{i\mathrm{c}} = \frac{\mathrm{N}_{i\mathrm{c}}}{\mathrm{N}_i}$$

$$E_i^{\mathrm{c}} = -\frac{1}{C}\sum_{\mathrm{c}} P_{i\mathrm{c}}\log\left(P_{i\mathrm{c}}\right)$$

where N_i is the number of neighbors of cell i ($\mathrm{N}_i = 20$ for each cell i), $\mathrm{N}_{i\mathrm{b}}$ is the number of neighbors of cell i with batch b, $\mathrm{N}_{i\mathrm{c}}$ is the number of neighbors of cell i with cell type c, and B and C are the number of batches and the number of cell types, respectively. A high batch entropy indicates a homogeneous mixture of different batches, while a low cell type entropy suggests that the cell types remain separate.

Adjusted Rand Index F1 Score

The rand index (RI) measures the similarity of results between two clustering methods. It is useful to compare the true label distribution with the clustering prediction and, therefore, can also be applied to measure batch mixing and cell type separation. The RI is defined as:

$$RI = \frac{a + b}{\binom{n}{2}}$$

where a is the number of pairs of cells with the same true label that belongs to the same cluster, b is the number of pairs of cells with a different true label that are assigned to different clusters, and $\binom{n}{2}$ is the number of unordered pairs in a set of n cells. To ensure a value close to 0 for random labeling, the RI score is "adjusted for chance," which gives the ARI:

$$ARI = \frac{RI - E(RI)}{max\left(RI\right) - E(RI)}$$

where $E(RI)$ and $max(RI)$ are the expectation and maximum of RI, respectively. The ARI score ranges from −1 to 1. A positive high ARI score suggests that the result of clustering prediction is much consistent with the true label distribution.

To obtain the ARI score, we first applied the k-means algorithm to generate cluster labels for comparison against batch labels and cell type labels. We then randomly selected 80% of cells and calculated the ARI scores for batch and cell type. This procedure was repeated 20 times to ensure stability. The batch ARI score and cell type ARI score were further normalized into an interval of [0, 1], which were denoted as $\mathrm{ARI}_{\mathrm{batch\_norm}}$ and $\mathrm{ARI}_{\mathrm{celltype\_norm}}$, respectively. Finally, the ARI F1 score was defined as:

$$F1_{ARI} = \frac{2(1 - ARI_{batch\_norm})(ARI_{celltype\_norm})}{1 - ARI_{batch\_norm} + ARI_{celltype\_norm}}$$

The ARI F1 score is the harmonic mean of the ARI batch score and the ARI cell type score. As a combined measurement of batch mixing and cell type separation, a higher ARI F1 score indicates a better performance of the batch correction method.

Average Silhouette Width F1 Score

The silhouette score measures how well a cell lies within its own cluster in comparison with other clusters. It is defined as:

$$s_i = \frac{(b_i - a_i)}{max\left(a_i, b_i\right)}$$

where a_i is the average distance between cell i and other cells in its cluster, and b_i is the average distance between cell i and the cells in its nearest cluster. The silhouette score is between −1 and 1. A positive high silhouette score suggests that the cell is close to its own cluster but discrepant to other clusters. The ASW score over the entire dataset is then given by:

$$\mathrm{ASW} = \frac{1}{n}\sum_i s_i$$

where n is the total number of cells in the dataset. The ASW score indicates whether the clusters are well separated and, hence, can also be used to evaluate the performance of the batch correction methods.

Like the calculation of the ARI score, we randomly selected 80% of cells to compute the ASW batch score and the ASW cell type score and repeated this procedure 20 times. We normalized the ASW batch score and the ASW cell type score into an interval [0, 1]. The ASW F1 score was then obtained by calculating the harmonic mean of the normalized ASW batch score and the normalized ASW cell type score as follows:

$$F1_{ASW} = \frac{2(1-ASW_{batch\_norm})(ASW_{celltype\_norm})}{1-ASW_{batch\_norm} + ASW_{celltype\_norm}}$$

The ASW F1 score is a combined metric to assess batch mixing and cell type separation. A higher ASW F1 score indicates better performance.

Statistical Test and Visualization

The Mann–Whitney *U*-test with the Benjamini–Hochberg correction was applied to the ARI F1 scores and the ASW F1 scores to compare the performance on batch correction between deepMNN and other methods.

We used UMAP (Becht et al., 2019) implemented in the Scanpy library (version 1.6.0) to visualize our batch correction results with default parameters.

RESULTS

We utilized the three datasets of PBMCs with two batches, pancreas cells with five batches, and HCA cells with two batches (**Table 1**) to evaluate all batch correction methods under four different scenarios: identical cell types, non-identical cell types, multiple batches, and large datasets.

The experiments were carried out on a workstation with four NVIDIA GeForce GTX 1080 Ti graphics cards, two Intel Xeon E5-2620 v4 CPUs, and 64G random access memory (RAM). We performed experiments for all methods in the CPU environment except the deep learning-based methods of deepMNN, scGen, and MMD-ResNet, for which a single GPU card was used.

Scenario 1: Identical Cell Types

We first used the PBMC dataset to evaluate the batch correction methods. This dataset was comprised of nine identical cell types and possessed a similar proportion of cells for each cell type between the two batches (**Figure 2A**). The UMAP plots depicted that all methods except MMD-ResNet successfully merged the common cells (**Figure 3A**). The deepMNN, Harmony, and Seurat V4 produced a distinct megakaryocyte cluster from other cell type clusters. By comparison, most megakaryocyte cells were mixed up with monocyte CD14 cells by Scanorama and scGen. Moreover, the CD8 cells located much closer within the compact clusters that resulted from deepMNN and Seurat V4. However, these cells scattered around the CD4 T cells in the clusters generated by Harmony, Scanorama, and scGen.

With regards to the batch and cell type entropies (**Figure 3B**), deepMNN achieved a comparable or a slightly lower batch entropy than Harmony, scGen, and Seurat v4, but higher than MMD-ResNet and Scanorama. A lower cell type entropy was reached by deepMNN compared to other methods except for Harmony and Seurat V4. As for the ASW F1 score (**Figure 3C**), deepMNN was significantly higher than the other methods ($p < 0.00001$). Furthermore, the results of the ARI F1 scores (**Figure 3D**) showed that the performance of deepMNN was comparable with that of Harmony and Seurat V4 and significantly better than all the other methods ($p < 0.00001$).

TABLE 1 | Single-cell RNA sequencing datasets used for evaluating deepMNN.

Dataset	Batch	Protocol	Number of cells
PBMC	10× 3′	10× Chromium Single Cell 3′ v2 chemistry	8,098
	10× 5′	10× Chromium Single Cell 5′ paired-end chemistry	7,378
Pancreas	Baron	inDrops	8,569
	Muraro	CelSeq2	2,122
	Segerstolpe	SMART-seq2	2,127
	Wang	SMARTer	457
	Xin	SMARTer	1,492
HCA	Bone Marrow	10×	275,264
	Cord blood	10×	253,024

Scenario 2: Non-identical Cell Types

To evaluate deepMNN under the scenario where batches had non-identical cell types, we downsampled the PBMC dataset using the following criteria: (1) the CD8 and B cells were removed from the 10× 3′ batch and (2) the monocyte CD14 and NK cells were removed from the 10× 5′ batch. As a result, the two batches had different cell types except for CD4, megakaryocyte, and monocyte FCGR3A cells (**Figure 2B**). Similar to the results from scenario 1, we observed that all the methods, except MMD-ResNet, merged the two batches (**Figure 4A**). The deepMNN, Harmony, scGen, and Seurat V4 produced well-separated clusters for megakaryocyte cells that, however, were mixed up with monocyte CD14 cells using Scanorama. Moreover, it was observed that the methods of Harmony, Scanorama, and Seurat V4 mixed up some CD8 T cells with CD4 T cells, some other CD8 T cells with NK cells, and some monocyte FCGR3A cells with monocyte CD14 cells. In contrast, all cell types were clearly distinguished by deepMNN except that only a few of CD8 T cells were mixed up with NK cells.

Regarding the batch and cell type entropies, deepMNN was one of the methods that obtained the lowest cell entropy (**Figure 4B**). It had a lower batch entropy than Harmony, scGen, and Seurat v4 did. The ASW F1 score of deepMNN was lower than scGen but significantly higher than all other methods ($p < 0.00001$) (**Figure 4C**). No significant difference in the ARI F1 scores was observed between deepMNN and the methods of Harmony, scGen, and Seurat V4. However, deepMNN reached a significantly higher ARI F1 score than MMD-ResNet and Scanorama ($p < 0.00001$) (**Figure 4D**).

Scenario 3: Multiple Batches

To assess the performance of deepMNN on a dataset with multiple batches, we employed the dataset of human pancreatic cells that consisted of five batches. The dataset had different

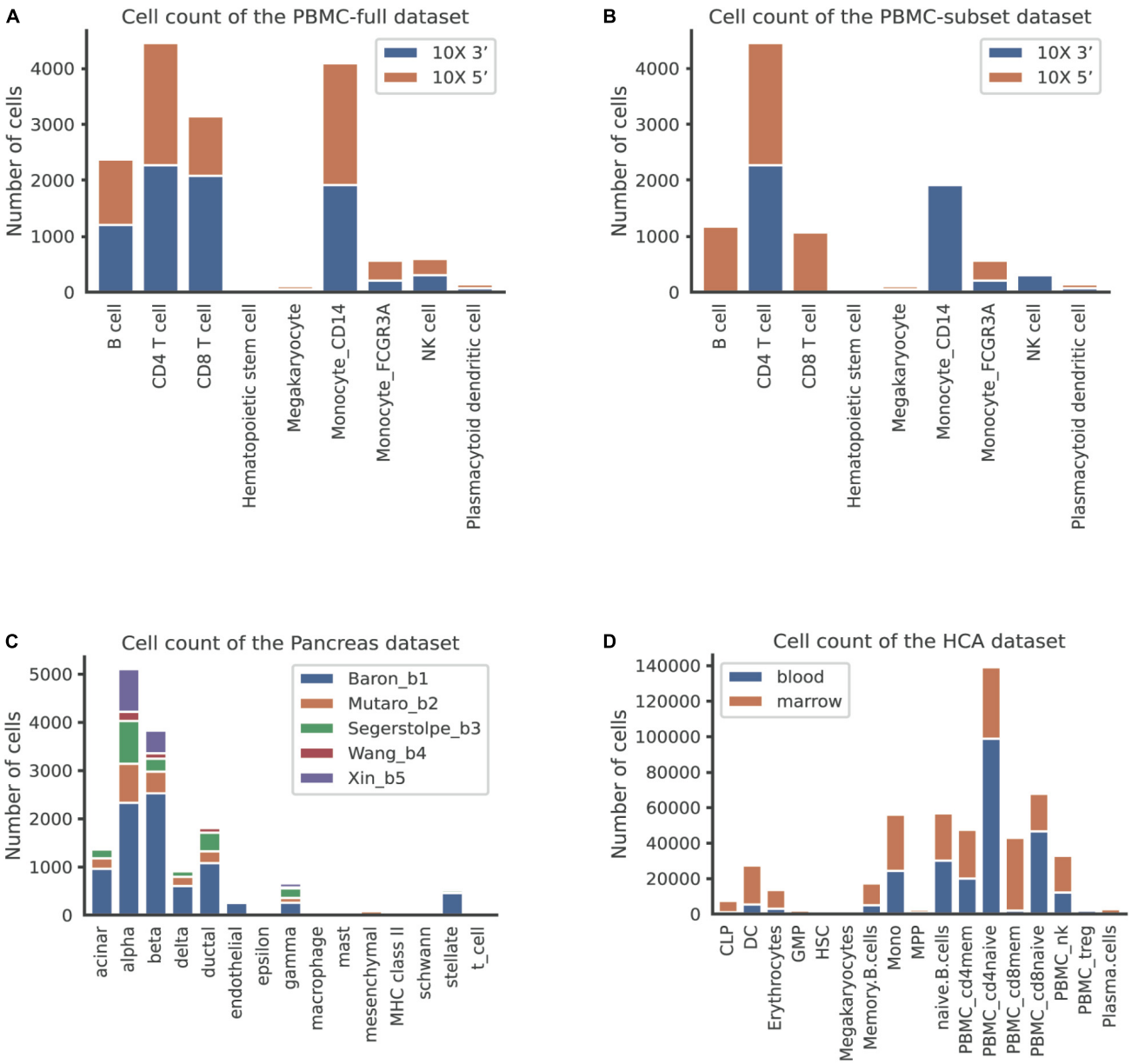


FIGURE 2 | The number of cells per cell type across batches in different datasets. **(A)** The human peripheral blood mononuclear cell (PBMC) full dataset, corresponding to the scenario of identical cell types. **(B)** The PBMC subset dataset, corresponding to the scenario of non-identical cell types. **(C)** The pancreas dataset, corresponding to the scenario of multiple batches. **(D)** The human cell atlas dataset, corresponding to the scenario of large-scale datasets.

numbers of cells and non-identical cell types between batches (**Figure 2C**). The UMAP plots demonstrated that Harmony, scGen, and Seurat v4 can merge all batches, while deepMNN and Scanorama were more likely to make cell-specific clusters close together (**Figure 5A**). Interestingly, all methods appeared to have maintained a relatively good cell type separation.

For the evaluation metrics, deepMNN obtained a lower batch entropy than Harmony, scGen, and Seurat V4 and was one of the methods that achieved the lowest cell entropy (**Figure 5B**). It reached a significantly higher ASW F1 score compared to the other methods ($p < 0.00001$) (**Figure 5C**). The ARI F1 score from deepMNN was also significantly higher than that from Harmony ($p < 0.05$), Scanorama ($p < 0.00001$), and scGen ($p < 0.001$) except for Seurat V4 ($p > 0.05$) (**Figure 5D**). Due to the bad performance of MMD-ResNet in the experiments using two-batch datasets as shown above, we did not evaluate the method of MMD-ResNet under this multiple-batch scenario.

Scenario 4: Large-Scale Dataset

We further evaluated the batch correction methods using the large-scale HCA dataset that was comprised of two batches, where one batch had 275,184 bone marrow cells, while another

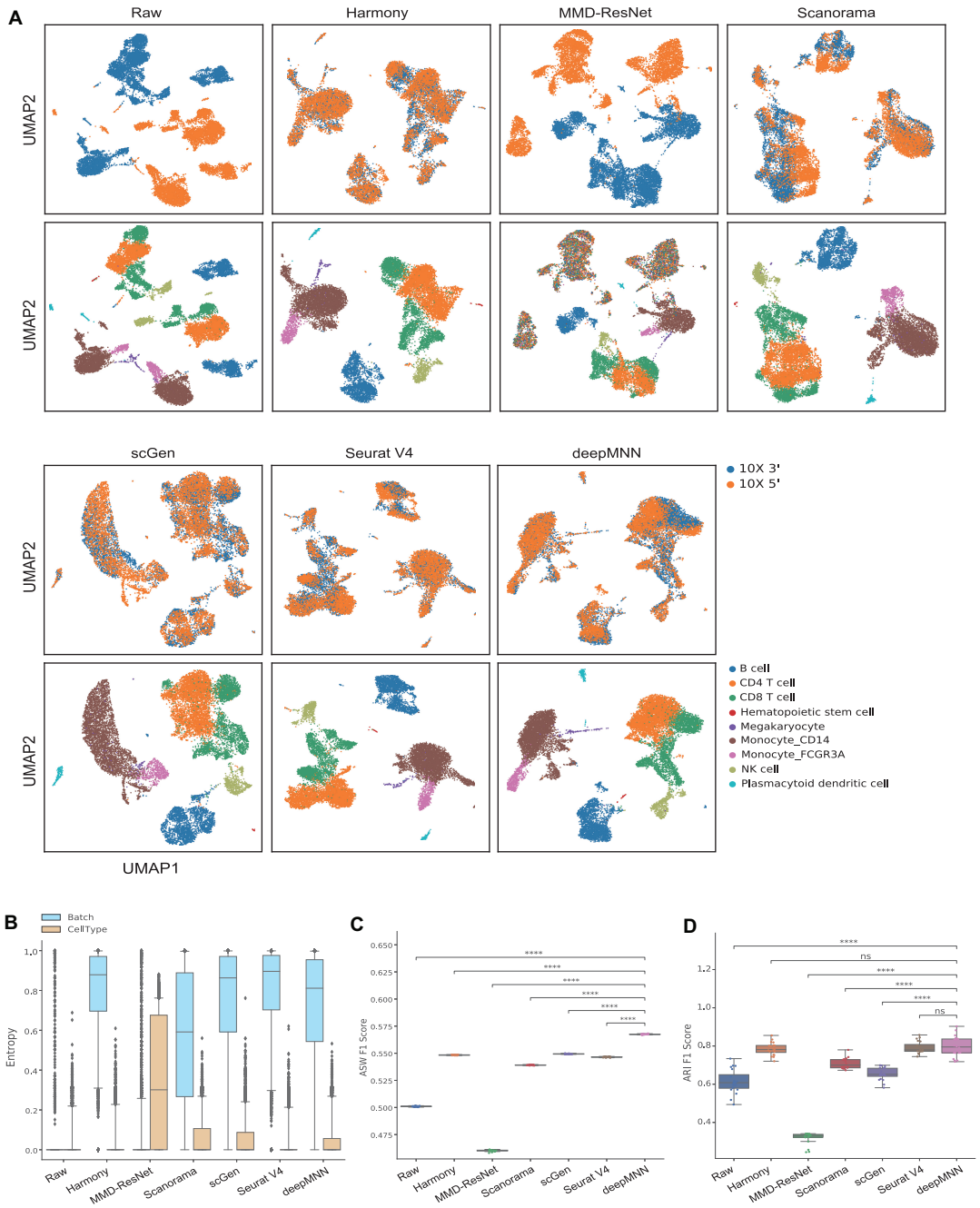


FIGURE 3 | Comparison of batch effect correction methods for the human peripheral blood mononuclear cell dataset of identical cell types with two batches. **(A)** Qualitative evaluation of the raw data, Harmony, MMD-ResNet, Scanorama, scGen, Seurat V4, and deepMNN using UMAP. The cells were colored by batches on the top row and colored by cell type on the bottom row. **(B)** The batch and cell type entropies resulting from the batch correction methods. The plots show the median (line within box), 25th and 75th percentiles (box), 5th and 95th percentiles (whiskers), and outliers (diamond points). **(C)** The ASW F1 score resulting from different batch correction methods. **(D)** The ARI F1 scores resulting from different batch correction methods. ****$p \leq 0.0001$.

had 252,830 cord blood cells (Li et al., 2018; **Figure 2D**). Seurat V4 and scGen were not capable of running successfully on our server with 64GB RAM due to the exceedingly huge size of the dataset. The deepMNN took approximately 17 min to complete the process of batch effect correction, which was significantly faster than Harmony (∼35 min) and Scanorama (∼77 min). Since the computation of batch and cell type entropies required more than 1 TB RAM and the calculation of the ASW F1 score was unable to be completed within 48 h on our server, we did not provide the results of the quantitative metrics. However, it was observed that deepMNN, Harmony, and Scanorama were able to bring cell-specific clusters close together (**Figure 6**).

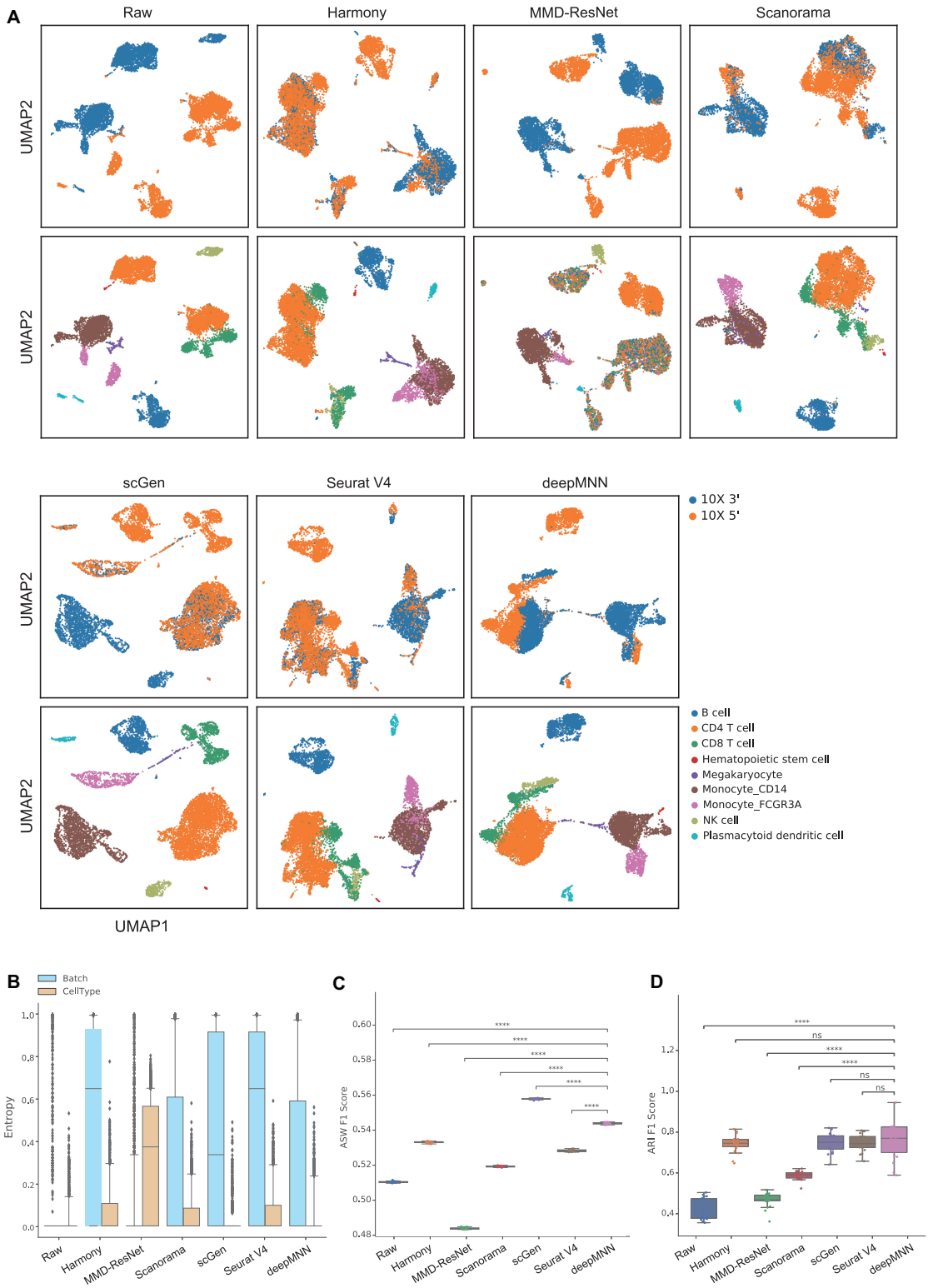


FIGURE 4 | Comparison of batch effect correction methods for the human peripheral blood mononuclear cell dataset of non-identical cell types with two batches. **(A)** Qualitative evaluation of the raw data, Harmony, MMD-ResNet, Scanorama, scGen, Seurat V4, and deepMNN using UMAP. The cells were colored by batches on the top row and colored by cell type on the bottom row. **(B)** The batch and cell type entropies resulting from the batch correction methods. The plots show the median (line within box), 25th and 75th percentiles (box), 5th and 95th percentiles (whiskers), and outliers (diamond points). **(C)** The ASW F1 score resulting from different batch correction methods. **(D)** The ARI F1 scores resulting from different batch correction methods. ****$p \leq 0.0001$.

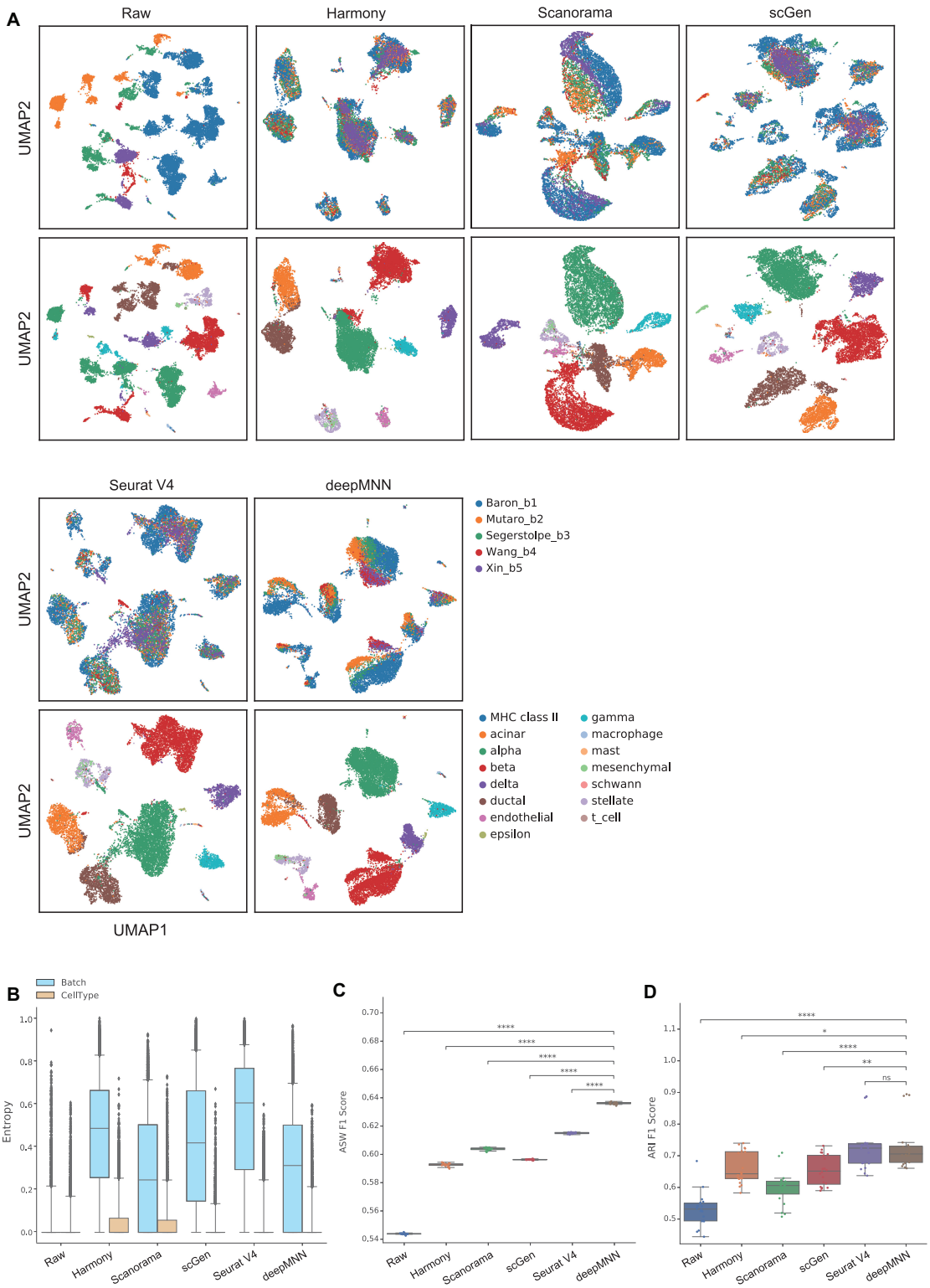


FIGURE 5 | Comparison of batch effect correction methods for the pancreas datasets with five batches. **(A)** Qualitative evaluation of the raw data, Harmony, Scanorama, scGen, Seurat V4, and deepMNN using UMAP. The cells were colored by batches on the top row and colored by cell type on the bottom row. **(B)** The batch and cell type entropies resulting from the batch correction methods. The plots show the median (line within box), 25th and 75th percentiles (box), 5th and 95th percentiles (whiskers), and outliers (diamond points). **(C)** The ASW F1 score resulting from different batch correction methods. **(D)** The ARI F1 scores resulting from different batch correction methods. *$0.01 < p \leq 0.05$, **$0.001 < p \leq 0.01$, ****$p \leq 0.0001$.

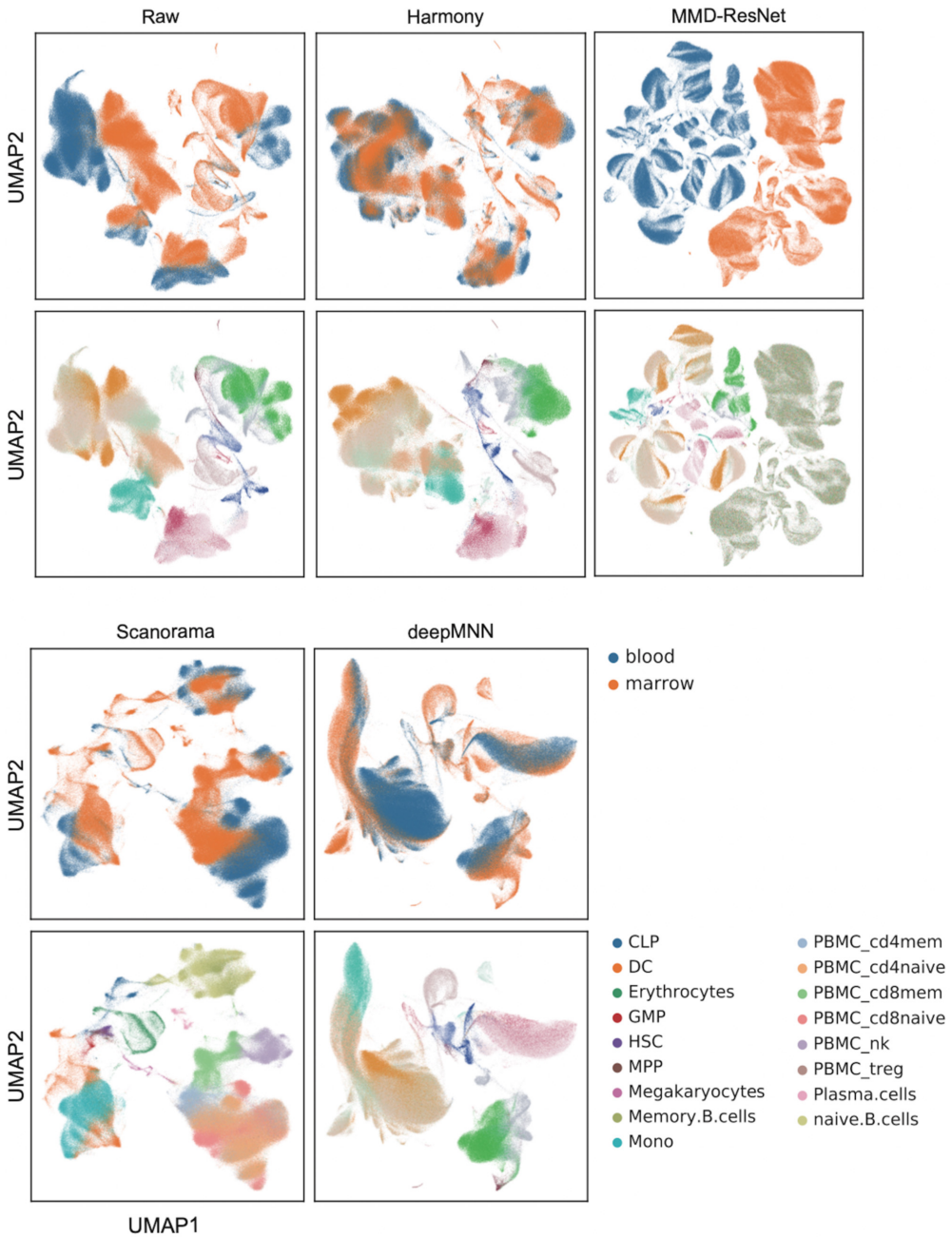


FIGURE 6 | Evaluation of raw data, Harmony, MMD-ResNet, Scanorama, scGen, and deepMNN on the large-scale HCA dataset with 528,014 cells. The cells were colored by batches on the top row and colored by cell type on the bottom row.

DISCUSSION

Batch effect poses a big challenge in scRNA-seq data analysis. In this study, we proposed deepMNN, a novel deep learning-based scRNA-seq batch correction method. The deepMNN was constructed by a residual-based batch correction network in conjunction with MNN pairs to remove batch effects in scRNA-seq data. The experiment results showed that deepMNN can successfully align different datasets under four scenarios such as identical cell types, non-identical cell types, multiple batches, and large-scale datasets. We compared the performance of deepMNN with state-of-the-art batch correction methods, including Harmony, Scanorama, and Seurat V4 as well as MMD-ResNet and scGen. The results demonstrated that deepMNN achieved a better or comparable performance in terms of both qualitative analysis using UMAP plots and quantitative metrics such as batch and cell entropies, ARI F1 score, and ASW F1 score as well as running time. Two review papers (Tran et al., 2020;

Chazarra-Gil et al., 2021) reported that Harmony and Seurat were the best batch correction methods in most scenarios, which, in turn, suggested the high efficiency of deepMNN to correct batch effect.

The cell types and their proportions may be considerably different across batches. For MNN-based batch correction methods, such as MNNCorrect, Scanorama, and deepMNN, the MNN pairs across batches need to be computed first. When two cells from two datasets were identified in an MNN pair, they were likely the same cell type. To remove the batch effect, traditional methods usually calculated reference vectors based on the identified MNN pairs and mapped one dataset to the space obtained from the reference dataset. By comparison, deepMNN applied a batch correction network that was formed by the stack of two residual blocks for batch removal. Since the residual block contained a residual term $\delta(x)$ and an identity term x, deepMNN can easily learn a representation similar to the identity term. In addition, the distributions of the same cell types from different batches were theoretically close to each other, and the discrepancy may be introduced by the batch effect. Thus, the residual structure of deepMNN attempted to learn a representation for the identity term, and the residual term can be regarded as the batch effect.

Methods like Scanorama and Seurat V4 merged only two datasets at once and iterated the same procedure to accomplish the integration of multiple datasets. To our best knowledge, deepMNN was the first method to integrate multiple batches of scRNA-seq data in one step. After identifying MNN pairs among batches, we minimized the batch loss that measured the distance between cells in the MNN pairs, which can promote the network removing the multiple-batch effect simultaneously. It should be noted that the batch loss was not directly based on the output of the batch correction network. We applied the PCA instead to reduce the dimension of the output first and then calculated the distance between cells in the MNN pairs.

Compared to the state-of-the-art batch correction methods, deepMNN achieved almost significantly high ARI F1 scores and ASW F1 scores under the scenarios of identical cell types, non-identical cell types, and multiple batches. The scGen reached a higher ASW F1 score than deepMNN under the scenario of non-identical cell types. This was partially due to the feature of scGen that was a supervised learning method and required cell type labels. As for computation time, deepMNN was comparable with other methods when the dataset was small. However, it was significantly fast when dealing with large-scale datasets – for example, deepMNN spent around 17 min on batch correction for the 528k HCA dataset, while Harmony and Scanorama needed about 35 and 77 min, respectively. Korsunsky et al. (2019) compared the runtimes for different batch correction methods and reported Harmony as one of the fastest batch correction methods, which took 68 min on 500,000 cells. One reason for the ability of quick batch correction by deepMNN was likely that it removed batch effect in one step. Another reason might probably be that deepMNN converged fast and can complete batch correction within tens of epochs. In our experiments, deepMNN only required 50 to 100 epochs to accomplish the removal of batch effect. The last reason was partially due to the deep learning-based method of deepMNN that used GPU to speed up the computation. Seurat V4 and scGen cannot run on our 64GB server for the 528k HCA dataset due to their high RAM requirement.

The overall loss of deepMNN was the sum of a batch loss and a weighted regularization loss that was controlled by the tradeoff parameter α. The use of regularization loss was to make the output of the network similar to the input and to prevent the output from being zero when no batches existed in a dataset. We investigated the effect of α on the batch correction performance of deepMNN in terms of the ARI F1 score and ASW F1 score under three different scenarios. Generally, the ASW F1 score tended to rise first and then declined with the decrease of α, and it reached

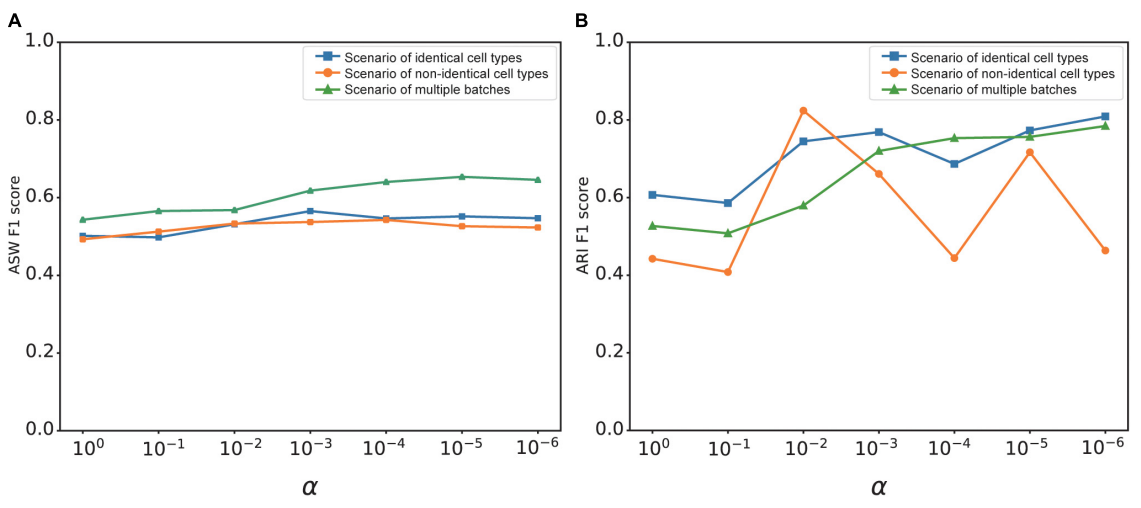


FIGURE 7 | The effect of value changes in α on the batch correction performance of deepMNN under three scenarios of identical cell types, non-identical cell types, and multiple batches. **(A)** The ASW F1 scores *versus* various α values under different scenarios. **(B)** The ARI F1 scores *versus* various α values under different scenarios.

almost the highest value when α was 0.001 under each of the three scenarios (**Figure 7A**). Although the ARI F1 score exhibited much fluctuation with the change of α, it can also have the highest value with α of 0.001 under the scenario of identical cell types (**Figure 7B**). Therefore, we chose 0.001 as the optimal value of parameter α .

One key limitation of our method was that deepMNN depended heavily on the identified MNN pairs. Only a small number of MNN pairs can be found when a handful of cells represented a shared biological state across batches, which was not sufficient to remove batch effects in the entire datasets effectively. On the other hand, even though a large number of MNN pairs have been identified but a low percentage of them have had the same cell types, deepMNN would result in a poor performance on batch correction. In our experiments, about 80–90% of MNN pairs had the same cell types. In the future, more reliable schemes of searching MNN pairs will be investigated. Another aspect of limitation in this study was related to the dimension reduction method. In this study, deepMNN used the PCA to project raw single-cell gene expression data into low-dimensional space. However, a previous study (Butler et al., 2018) demonstrated that PCA could intrinsically identify biologically irrelevant variations caused by technical effects. Other data embedding methods like CCA (Butler et al., 2018) and autoencoder (Li et al., 2020) would be further considered to improve the batch correction performance of deepMNN.

CODE AVAILABILITY

The source code of deepMNN, including the experimental results of the study, can be found at https://github.com/zoubin-ai/deepMNN.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These datasets can be found here: Human peripheral blood mononuclear cell (PBMC): ftp://ngs.sanger.ac.uk/production/teichmann/BBKNN/PBMC.merged.h5ad; Human pancreas: https://hemberg-lab.github.io/scRNA.seq.datasets/human/pancreas/; Human cell atlas (HCA): https://github.com/immunogenomics/harmony2019/tree/master/data/figure3.

AUTHOR CONTRIBUTIONS

BZ and YB conceived the algorithm and wrote the manuscript. BZ and RZ developed and performed the computational experiments. TZ performed the scRNA-seq experiments. BZ and XJa plotted the figures. YB, XJn, and HY supervised the study. All authors read and approved the final manuscript.

FUNDING

We gratefully acknowledge the support of the National Natural Science Foundation of China (32000398), Natural Science Foundation of Guangdong Province, China (2017A030306026), and Guangdong-Hong Kong Joint Laboratory on Immunological and Genetic Kidney Diseases (2019B121205005).

ACKNOWLEDGMENTS

We thank Shiping Liu and Yinqi Bai for the helpful discussion of the algorithm.

REFERENCES

Bahrami, M., Maitra, M., Nagy, C., Turecki, G., Rabiee, H. R., and Li, Y. (2020). Deep feature extraction of single-cell transcriptomes by generative adversarial network. *Bioinformatics* 37, 1345–1351. doi: 10.1093/bioinformatics/btaa976

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Syst.* 3, 346–360.e4.

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

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* 36, 411–420. doi: 10.1038/nbt.4096

Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y., and Hemberg, M. (2021). Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.* 49:e42. doi: 10.1093/nar/gkab004

Consortium, T. M. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* 562, 367–372. doi: 10.1038/s41586-018-0590-4

Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* 35, 498–507. doi: 10.1016/j.tibtech.2017.02.012

Haghverdi, L., Lun, A. T. L., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* 36, 421–427. doi: 10.1038/nbt.4091

Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., et al. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell* 172, 1091–1107.e17.

Hie, B., Bryson, B., and Berger, B. (2019). Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* 37, 685–691. doi: 10.1038/s41587-019-0113-3

Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi: 10.1007/bf01908075

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

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., et al. (2019). Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* 16, 1289–1296. doi: 10.1038/s41592-019-0619-0

Li, B., Kowalczyk, M. S., Dionne, K., Ashenberg, O., Tabaka, M., Tickle, T., et al. (2018). *HCA Data Portal-Census of Immune Cells*. Available online at: https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79. (accessed January 9, 2021).

Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., et al. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.* 11:2338. doi: 10.1038/s41467-020-15851-3

Lin, Y., Ghazanfar, S., Wang, K. Y. X., Gagnon-Bartsch, J. A., Lo, K. K., Su, X., et al. (2019). scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets. *Proc. Natl. Acad. Sci. U.S.A.* 116, 9775–9784. doi: 10.1073/pnas.1820006116

Lotfollahi, M., Wolf, F. A., and Theis, F. J. (2019). scGen predicts single-cell perturbation responses. *Nat. Methods* 16, 715–721. doi: 10.1038/s41592-019-0494-8

Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., et al. (2016). A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* 3, 385–394.e3.

Polański, K., Young, M. D., Miao, Z., Meyer, K. B., Teichmann, S. A., and Park, J.-E. (2020). BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* 36, 964–965.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Segerstolpe, Å, Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., et al. (2016). Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metabol.* 24, 593–607. doi: 10.1016/j.cmet.2016.08.020

Shaham, U., Stanton, K. P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., et al. (2017). Removal of batch effects using distribution-matching residual networks. *Bioinformatics* 33, 2539–2546. doi: 10.1093/bioinformatics/btx196

Stegle, O., Teichmann, S. A., and Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* 16, 133–145. doi: 10.1038/nrg3833

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M. III, et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21.

Svensson, V., Vento-Tormo, R., and Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* 13, 599–604.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., et al. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* 21:12. doi: 10.1186/s13059-019-1850-9

Wang, T., Johnson, T. S., Shao, W., Lu, Z., Helm, B. R., Zhang, J., et al. (2019). BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol.* 20:165. doi: 10.1186/s13059-019-1764-6

Wang, Y. J., Schug, J., Won, K.-J., Liu, C., Naji, A., Avrahami, D., et al. (2016). Single-cell transcriptomics of the human endocrine pancreas. *Diabetes* 65, 3028–3038. doi: 10.2337/db16-0405

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873–1887.e17.

Wolf, F. A., Angerer, P., and Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19:15.

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metabol.* 24, 608–615. doi: 10.1016/j.cmet.2016.08.018

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049.

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

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

frontiers in Genetics

METHODS
published: 10 September 2021
doi: 10.3389/fgene.2021.682841

MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis

Inuk Jung[1], Minsu Kim[2], Sungmin Rhee[3], Sangsoo Lim[4] and Sun Kim[2,3,4]**

[1] Department of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea, [2] Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3] Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea, [4] Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-Gu, Seoul, South Korea

Multi-omics data is frequently measured to enrich the comprehension of biological mechanisms underlying certain phenotypes. However, due to the complex relations and high dimension of multi-omics data, it is difficult to associate omics features to certain biological traits of interest. For example, the clinically valuable breast cancer subtypes are well-defined at the molecular level, but are poorly classified using gene expression data. Here, we propose a multi-omics analysis method called MONTI (Multi-Omics Non-negative Tensor decomposition for Integrative analysis), which goal is to select multi-omics features that are able to represent trait specific characteristics. Here, we demonstrate the strength of multi-omics integrated analysis in terms of cancer subtyping. The multi-omics data are first integrated in a biologically meaningful manner to form a three dimensional tensor, which is then decomposed using a non-negative tensor decomposition method. From the result, MONTI selects highly informative subtype specific multi-omics features. MONTI was applied to three case studies of 597 breast cancer, 314 colon cancer, and 305 stomach cancer cohorts. For all the case studies, we found that the subtype classification accuracy significantly improved when utilizing all available multi-omics data. MONTI was able to detect subtype specific gene sets that showed to be strongly regulated by certain omics, from which correlation between omics types could be inferred. Furthermore, various clinical attributes of nine cancer types were analyzed using MONTI, which showed that some clinical attributes could be well explained using multi-omics data. We demonstrated that integrating multi-omics data in a gene centric manner improves detecting cancer subtype specific features and other clinical features, which may be used to further understand the molecular characteristics of interest. The software and data used in this study are available at: https://github.com/inukj/MONTI.

Keywords: feature selection, tensor decomposition, cancer, multi-omics, integrative analysis

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Florian Buettner,
German Cancer Research Center (DKFZ), Germany
Fuhai Li,
Washington University in St. Louis, United States

****Correspondence:***
Inuk Jung
inukjung@knu.ac.kr
Sun Kim
sunkim.bioinfo@snu.ac.kr

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

Received: *19 March 2021*
Accepted: *12 August 2021*
Published: *10 September 2021*

Citation:
Jung I, Kim M, Rhee S, Lim S and Kim S (2021) MONTI: A Multi-Omics Non-negative Tensor Decomposition Framework for Gene-Level Integrative Analysis. Front. Genet. 12:682841. doi: 10.3389/fgene.2021.682841

1. INTRODUCTION

Genes are among the most important building blocks of all organisms. Their transcription and translation are essential for maintaining fundamental cellular mechanisms. Genes are continuously and precisely regulated by a wide variety of mechanisms, including transcription factors, miRNAs, methylation, and mutations, which are often cumulatively referred to as multi-omics. When

investigating a biological mechanism, each omics can only provide a single perspective. By matching multi-omics data sampled from a common subject, a multiple-perspective view can be generated for an enhanced understanding of the complex dynamics of biology in the subject. For each additionally integrated omics data type, a new relationship can be mined between a gene and the newly added, which increases the ability to represent complex relationships across multi-omics data types, as shown in **Figure 1**. However, due to their heterogeneous nature, it is difficult to integrate such different omics data types within a common data structure and even more difficult to analyze them in a combined manner due to their high dimension.

A number of initiative projects have made great effort to collect and publicly provide large amounts of multi-omics data, such as TCGA (Weinstein et al., 2013), GTEx (Carithers et al., 2015), ENCODE (The ENCODE Project Consortium, 2012), and HFGP (Li et al., 2016). These databases provide more than 10,000 high-throughput sequencing data sets generated using various platforms and collected from cancer patients, normal human tissues and model organisms. Compared to the availability of such large amounts of multi-omics data, the development of analytic methods that can encompass such large-scale heterogeneous data is just recently gaining interest (Hasin et al., 2017).

It is well understood that more data can improve the accuracy of data mining. However, this is true only if the data are precisely understood and, more importantly, correctly integrated. Omics data are generated on different platforms, which implies unique measurement scales, data formats, as well as different emphasis on molecular domains and relationships among molecular entities. Hence, normalization, pre-processing, as well as how to evaluate associations with genes or other entities must be carefully taken into account for each omics data set. Finally, the data must be analyzed in an integrative manner in order to data mine inter-relationships across the multi-omics domains.

While the aforementioned initiative projects are focused on providing large-scale multi-omics data, other databases have gathered and processed these large data sets to allow statistical queries. The LinkedOmics project (Vasaikar et al., 2017) collected multi-omics data from TCGA that includes 32 cancer types, surpassing 1 billion data points in total. Using simple correlation methods (i.e., Pearson, Spearman), a user may search for genes that are significantly correlated with the query gene. Here, the correlation is in the context of multi-omics. In addition to issues around data collection and analysis, methods for visualizing multi-omics data is important. With an increasing number of omics comes increased difficulty in visualizing the relationships between multiple omics. PaintOmics3 (Hernández-de Diego et al., 2018) is a web-based visualization tool that allows users to observe multi-omics relationships in a graphical manner. It supports nearly every sequencing technology platform, including proteomics and region-based omics data, such as ATAC (Buenrostro et al., 2015) or ChIP-seq (Park, 2009) data.

To date, studies sought to analyze high-throughput multi-omics sequencing data, with the majority reporting results using a single or a pair of omics (e.g, mRNA-miRNA, mRNA-methylation). In addition, the majority of such studies focus on identifying genes showing significant correlation with a certain omics type using statistical methods, such as Pearson's correlation or cosine similarity. Furthermore, such approaches tend to focus on finding a matching omics relation for a single gene with each iteration of the analysis rather than analyzing all genes and omics data in a combined manner. This is mainly due to the heavy computation load and requirements of multiple testing, which makes statistical analysis difficult.

A number of studies have reviewed multi-omics integration methods. A recent study (Huang et al., 2017) grouped multi-omics integration methods into four categories: (1) Matrix factorization methods, (2) Bayesian methods, (3) Network-based methods, and (4) Multiple step-analysis. In addition to those categories, the recently popular deep learning technique has been applied to predict genes that yield significant survival results in liver cancer (Chaudhary et al., 2017). Such multi-omics integration methods can also be categorized as supervised

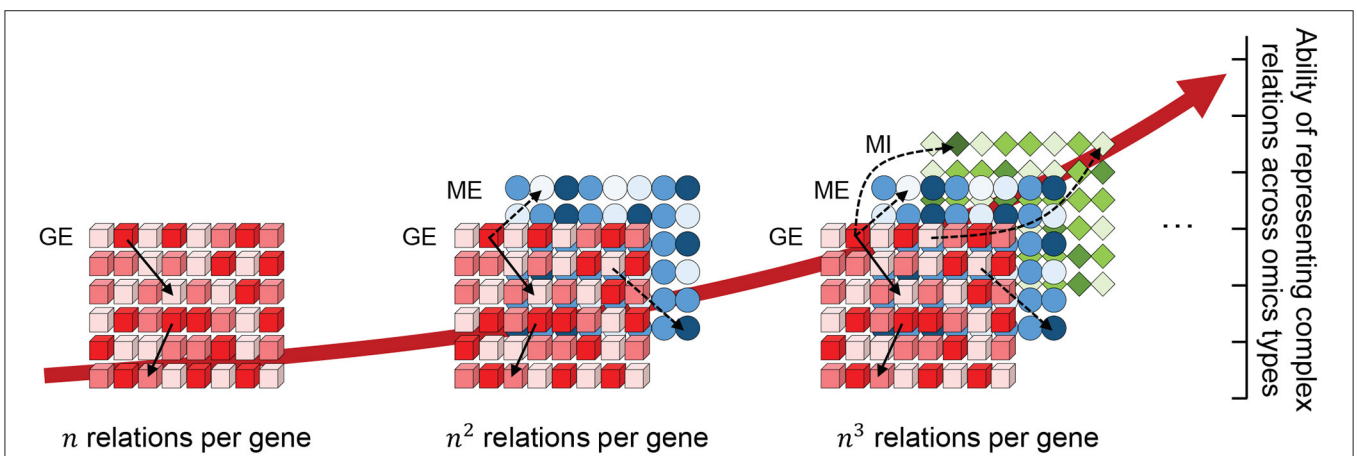


FIGURE 1 | The possible number of relations that a gene can have across omics layers (GE, gene expression; ME, methylation; MI, miRNA) increases exponentially with each omics data type added to the integration. Here, *n* indicates the number of genes within a single omics layer.

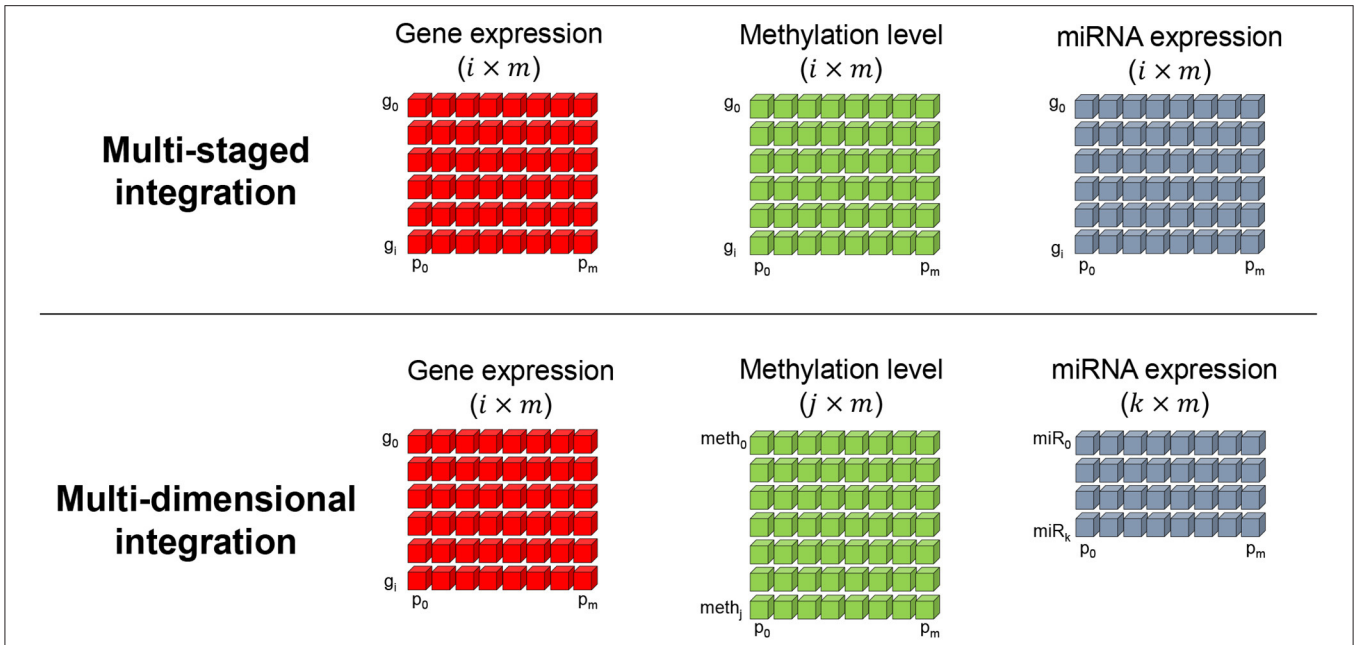


FIGURE 2 | Two prevalently used multi-omics integration methods. The multi-staged **(top)**, or gene-centric, method encodes all omics measurement values in a per gene basis. Hence, the number of genes (*g*) and samples (or patients *p*) in each omics matrix are required to have equal dimensions. The multi-dimensional **(bottom)** integration method is less restrictive in the dimensions and makes use of each omics data as is.

and unsupervised by making use of labels that represent the phenotype of the data, such as normal vs tumor sample. Tools such as jNMF (Zhang et al., 2012), MOFA (Argelaguet et al., 2018), and PARADIGM (Vaske et al., 2010) are unsupervised methods that mine gene clusters or modules associated with a phenotype of interest. Also, a network based multi-omics clustering method, SNF (Similarity Network Fusion) (Wang et al., 2014), was proposed that integrates multiple omics networks by weighted similarity of cluster samples.

More importantly, the aspect of the result greatly depends on how the multiple omics data are integrated. Two studies well-categorized and defined two important integration methods, which are the meta-dimensional and multi-staged integration approaches (Ritchie et al., 2015; Sathyanarayanan et al., 2020). The multi-staged integration method focuses on identifying omics factors that effect gene expression level, which is expected to find the causal relationship of a certain phenotype of interest. Hence, the omics data are integrated in a gene-centric manner and requires that each omics data have the same dimensions in sample and gene numbers as shown in **Figure 2** (top). Here, *g* and *p* refers to the gene and patient (or sample) indices *i* and *m*, respectively. Such gene-level multi-omics integration can be advantageous in assessing the flow of information from omics to genes. For example, gene-level analysis of mRNA, methylation, and miRNA omics data can discover strong relationships across the three omics layers in means to explain the dynamics of gene expression (Subramanian et al., 2020). However, with limited number of omics data, the landscape of gene expression modulation may not be fully explained. Also, the selection of omics data need to be focused on the assumption that they influence the gene expression regulation. In the other hand, the multi-dimensional integration method makes us of each omics data as is. Thus, the number of entities in each omics matrix may differ. The two integration methods both assume a matched multi-omics, that is, multi-omics data are retrieved from the same subject and therefore have the same number of samples. Such assumption is also referred to as multi-modal data. Such omics-level integration may capture the bigger dynamics underlying a phenotype since the entire data is analyzed as is (Sathyanarayanan et al., 2020). However, to analyze relationships across the omics layers, post-processing of the result is required, which can become very complex with larger number of omics data since the combinations of omics exponentially increase.

Utilizing multi-omics data, we can identify important biomarkers and also identify multi-omics features specific to a given sample or phenotype. In the context of cancer, multi-omics features specific to cancer subtypes can be identified, which can serve as valuable information for constructing highly accurate subtype classification models. This approach will eventually facilitate enhanced identification of subtype-specific genes. Delineation between cancer and normal tissues or across different cancer types have long been a popular problem (Furey et al., 2000; Ramaswamy et al., 2001; Sotiriou et al., 2003), with a classification accuracy reaching 85% (Gevaert et al., 2006). However, classifying cancer subtypes (Network et al., 2012; Shen et al., 2012; Paquet and Hallett, 2015) is more difficult than distinguishing tumor and normal samples. For example, classification accuracy for predicting breast cancer subtypes is low, ranging from 56.7 to 75% (Wu et al., 2017; Tao et al., 2019).

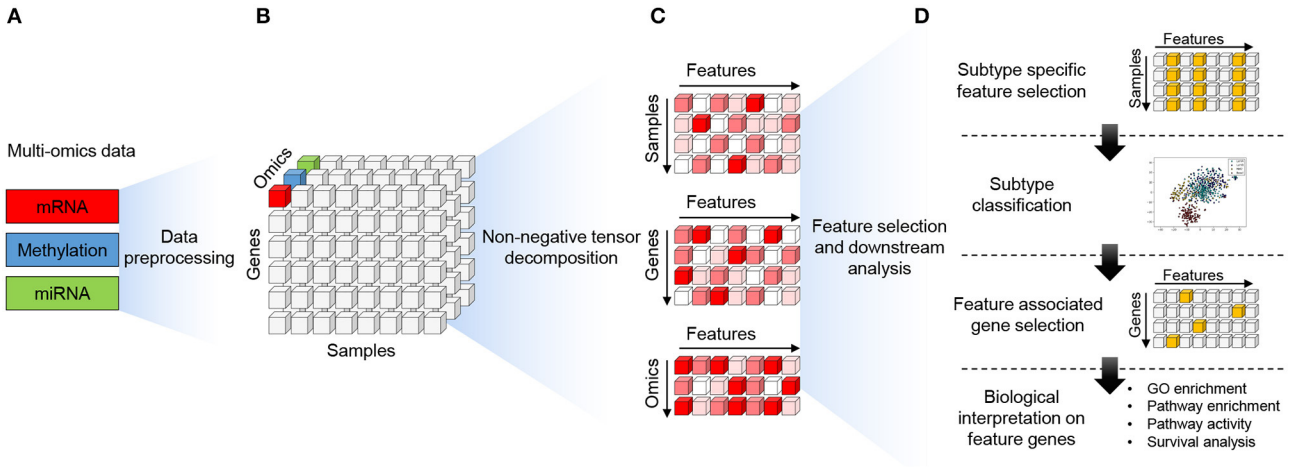


FIGURE 3 | The workflow of the MONTI framework. **(A)** Each omics data (gene expression, methylation, miRNA expression) is pre-processed as a two-dimensional gene-centric matrix comprised of genes and samples. **(B)** The omics matrices are then stacked to form a three-dimensional tensor structure (genes, samples, omics) all sharing the same genes and samples. **(C)** Using the PARAFAC approach, the tensor is decomposed into two-dimensional gene, patient and omics components. Here, the components share the rank features. **(D)** The patient component is used to select subtype-specific features using subtype-specific L1 classifiers. The selected subtype-specific features are used to build a subtype classifier model using MLP (Multi-layer perceptron). Genes associated to the subtype-specific features are then selected for biological function analysis.

In this study, we developed MONTI (Multi-Omics Non-negative Tensor Decomposition Integration) that learns hidden features through tensor decomposition for the integration of multi-omics data. MONTI is based on the gene-level integration method, which we find to be more helpful in understanding the results. The objective of MONTI is to extract feature genes that well explain some clinical attribute of interest in large multi-omics data. Being able to extract such a genes list with significant relation to clinical attributes can serve as a source that can naturally be used for simpler downstream analysis, such as, gene set enrichment of pathway analysis. Also, MONTI constraints the multi-omics data to be subject matched, where each omics data are collected from a common subject (i.e., patient). Such design may avoid omics variance within a same group, thus, amplifying the signals of hidden features.

In experiments with TCGA multi-omics data sets from breast, colon and stomach cancer samples, MONTI achieved significantly higher cancer subtype classification accuracy than existing multi-omics analysis methods. For the downstream analysis, genes associated with subtype-specific features were identified for biological interpretation.

2. MATERIALS AND METHODS

2.1. MONTI Framework Overview

The MONTI workflow operates in two phases. In the first phase, the multi-omics data are integrated and decomposed using non-negative tensor decomposition. In the second phase, subtype-specific features and genes associated with them are selected using L1 regularization, and these features are then used to generate a subtype classifier using the multi-layer perceptron (MLP) neural network. The overall workflow is depicted in **Figure 3**.

2.2. Data Preparation and Preprocessing of Multi-Omics Data

Samples with matched gene expression, methylation, and miRNA expression data sets were collected for three case studies from TCGA: (1) 597 breast cancer samples, (2) 314 colon cancer, and (3) 305 stomach cancer samples. Only primary tumor samples with all three matching omics data sets were selected for the analysis. The pre-quantified gene and miRNA expression values from TCGA were used as provided. For the methylation data, we used the HumanMethylation450 BeadChip-based data and further selected probes located within the gene promoter regions (i.e., 2 Kb upstream of a gene's transcription start site). Subtype information were acquired from the original studies. The partially missing subtype information of the breast cancer case study was taken from Lim et al. (2018), which were generated by the PAM50 classification method (Parker et al., 2009). Sample case IDs and annotated cancer subtypes of the samples used in this study are in **Supplementary Table 1**.

Because we aim to discover gene regulatory multi-omics features, each omics data is individually processed to form a *gene-centric* two-dimensional sample(patient)-gene matrix. The values in each omics matrix are computed and assigned with respect to each gene. The tensor structure requires all slices to be of the same size. Thus, while each omics matrix is independently processed, they share the same set of genes and samples.

The gene expression values were preprocessed according to the provided TCGA level 3 gene expression data, which were subject to log_2 quantile normalization across samples. For miRNA, they were first bundled per target gene, such that

the number of bundles matched the number of genes. The geometric mean of miRNA expression per bundle was assigned to each corresponding gene. The expression values were then log_2 quantile normalized. For methylation data, probes located within the transcription start site and 2 Kb upstream of gene promoter regions were grouped per gene. The average methylation level per gene was further quantile normalized.

Due to the nature of tensor decomposition, the omics value in each matrix need to be scaled within a common range. If not, an omics matrix with comparably large values, such as gene expression, would have a diminishing effect on other omics matrices with relatively lower values. Hence, normalized matrices are further scaled within the range of 0–1. Finally, the omics matrices were stacked on an orthogonal axis to form a three dimensional tensor structure.

2.3. Tensor Decomposition

There are several ways to decompose a tensor. PARAFAC (Carroll and Chang, 1970; Harshman, 1970) (a.k.a CANDECOMP-canonical decomposition) and TUCKER3 (Kroonenberg, 1983) are the most widely used methods. Both are multi- or bi-linear decomposition methods, which decompose the array into sets of scores and loadings. The decomposed scores and loadings describe the original data in a more compressed form. PARAFAC is based on factorization, whereas TUCKER3 utilizes principal component analysis. The resulting decomposition structure also differs between the two. PARAFAC decomposes a tensor into three two-dimensional components or matrices, while TUCKER3 generates three two-dimensional components along with an additional core matrix that is shared by the components. Due to the core matrix, interpreting data with the TUCKER3 model is more complicated (due to the increased number of parameters) than PARAFAC (Bro, 1997). Hence, here we used the PARAFAC method to decompose the multi-omics tensor.

A PARAFAC model of a three-way array T with elements x_{ijk} is given by three loading matrices, C_g, C_p, and C_o with elements g_{if}, p_{jf}, and o_{kf}. Here, we refer to C_g, C_p, and C_o as the gene, patient and omics components, respectively. The tensor T is decomposed using a predefined number of ranks R, which we will refer to as features $f = 1, \ldots, R$.

Due to the non-negative constraint, the interpretation of the feature values are much easier, since they are cumulative and do not negate themselves. Thus, a larger value will imply a strong signal of the feature. Furthermore, since omics data are most non-negative, the non-negative constraint can be naturally applied.

The trilinear model minimizes the sum of squares of the residuals, e_{ijk} in the model

$$x_{ijk} = \sum_{f=1}^{R} g_{if} p_{jf} o_{kf} + e_{ijk}, \tag{1}$$

which can also be written as

$$T = \sum_{f=1}^{R} g_f \otimes p_f \otimes o_f \tag{2}$$

An illustration of the PARAFAC model using gene expression, methylation level and miRNA expression data is in **Figure 4**. Here, $g_n(n = 0, ..., N)$ refers to the genes, $o_k(k = 0, ..., K)$ indicates the type of omics and $p_m(m = 0, ..., M)$ refers to patient samples. N, M and O indicate the number of genes, samples, and omics types, respectively. Three omics types are used in this illustration; thus, $K = 2$.

2.4. Feature Selection

Subtype-associated tensor features, a subset of features selected from the tensor decomposition result, significantly improved subtype classification accuracy. To select such subtype-specific features, L1 regularization was used for each subtype and applied to the (C_p) component (i.e., patient component) with the following equation,

$$min \sum_{i=1}^{M} (y_i - \sum_{f=1}^{R} z_{if} w_f)^2 + \alpha \sum_{f=1}^{R} |w_f| . \tag{3}$$

Here, M refers to the number of patient samples and R the number of features, or columns, in C_p. y_i refers to the target subtype value. Because an L1 model is built for each subtype, the target value is set to 1 for the corresponding subtype and 0 for the other subtype samples. For example, for the breast cancer case study, four L1 models were generated, one for each subtype of Luminal A, Luminal B, Her2, and Basal. z refers to the values of each feature in C_p. $w_f(f = 1, ..., R)$ refers to the weight of each feature to be inferred. The α value is the weight of the penalty term. Larger α values yields greater penalty, which will result in more features having zero weight and causing fewer features to be selected. We found that the L1 regularization achieved greater performance compared to the L2 regularization (**Figure 5**).

The feature selection performance using L1 and L2 were measured using the BRCA, COAD, and STAD data with varying ranks. As show in **Figure 5**, L1 showed better feature selection performance in terms of subtype classification accuracy in the three cancer types.

2.5. Selecting Feature Associated Genes

Based on the L1 selected features from C_p, feature genes were further selected from C_g. This procedure outputs a sparse set of genes, where each gene has a membership to a single feature. The association of a gene g to a feature is decided by $g_f = max(g_{0,R})$, where the weight is maximum at the corresponding feature index f.

2.6. Cancer Subtype Classification Analysis

The significance of the selected feature genes was measured by their power of subtype classification accuracy. The classification accuracy was measured using a multi-layer perceptron (MLP) classifier with 10-fold cross validation. Here, values of the feature genes from C_g were given as input to build the MLP classifier.

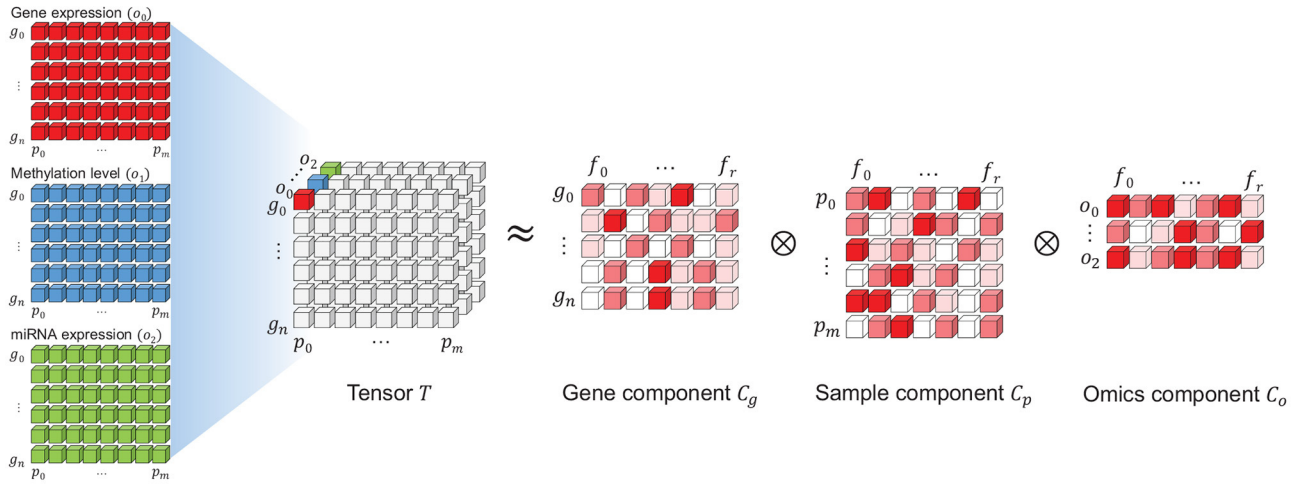


FIGURE 4 | An illustration of the non-negative tensor decomposition (PARAFAC) using three types of omics. The tensor T is decomposed into three components: gene, sample (patient), and omics. Each component corresponds to one axis in tensor T. Each component is a two-dimensional matrix where one axis embeds the rank features $f_i (i = 0, ..., r)$ for the entities in each component (i.e., genes, samples, and omics), similar to the traditional matrix factorization method.

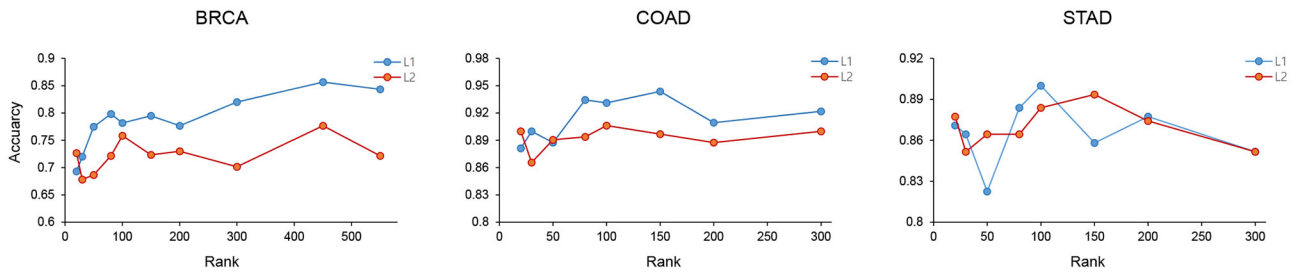


FIGURE 5 | The cancer subtype classification accuracy of BRCA, COAD, and STAD was measured using features selected by the L1 and L2 method with different ranks.

3. RESULTS

3.1. Three Case Studies

MONTI was applied to three cancer types: breast cancer (BRCA), colorectal cancer (COAD), and stomach cancer (STAD). The cancer types were chosen based on the number of samples that had matched multi-omics data from the same patient. There were 597, 314, and 305 matched omics data for BRCA, COAD, and STAD, respectively. To avoid an overly sparse tensor, genes that do not have any methylation probes located within their promoter and 2 Kb upstream of transcription start site (TSS) regions were discarded, which resulted in 14,513 genes with 60,707 methylation probes in total. The average methylation beta values were taken and assigned per gene. Similarly, miRNA expression values were grouped per target gene and the arithmetic mean of miRNA expression values in a group was assigned to its target gene. The multi-omics data items were used to produce gene centric omics matrices, which were then combined to form a three dimensional tensor of each cancer type, i.e., genes×multi-omics×patient samples.

3.2. Subtype Classification Results

Before deriving cancer subtype-specific features through tensor decomposition, a pre-defined rank R value for decomposing the tensor were needed to be chosen. In addition, a penalty strength, α value needed to be set for L1 regularization. Both were empirically chosen over a range of values by testing the subtype classification accuracy.

First, we evaluated the subtype classification accuracy using the feature in C_p over different ranks. The subtype classification accuracy for BRCA, COAD, and STAD was the highest with ranks 450, 150, and 100, respectively. The α value for L1 regularization determines the strength of the penalty for the features. The larger the α is the smaller number of features and genes be selected. Subtype classification performance was further investigated using α values ranging from 0 to 0.1. To further select informative features, the non-zero weight features were ranked by their absolute coefficient value from which top 20% features were chosen.

The subtype classification accuracy was the highest when $\alpha = 0.01$ (**Figure 6**). As a result, 26, 31, and 37 features from C_p were

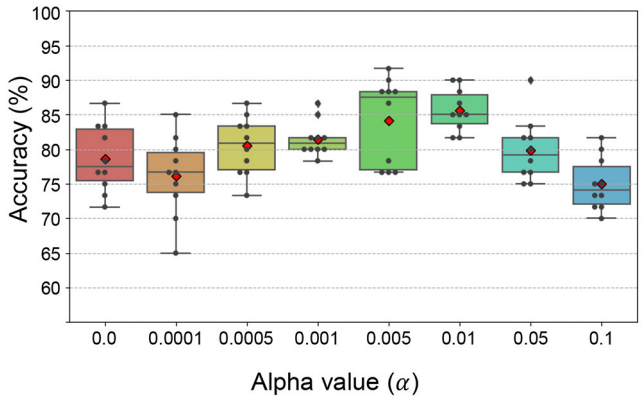


FIGURE 6 | The classification accuracy with varying α values. The classification accuracy was the highest with $\alpha = 0.01$.

selected for subtype classification from the BRCA, COAD, and STAD tensors, respectively.

The multi-omics tensors for the three cancer case studies were decomposed with the optimal rank numbers and α values that were chosen as explained above. We then investigated how much contributions feature genes (i.e., from C_g) made to the improvement in subtype classification accuracy.

Our primary interest in this study was whether the selected features would better represent the underlying biological mechanism when using multiple omics data compared to single or a smaller subset of omics data. As shown in **Figure 7A**, subtype classification the accuracy was the highest when all available multi-omics data were used and combined by the tensor features, which are labeled as GE, ME, and MI for gene expression, methylation, and miRNA expression respectively.

Here, we find that such accuracy reflects how much the subtypes are explainable by the selected features and their associated genes in multi-omics manner.

The number of features and their associated genes are shown in **Table 1**. Since a feature can be associated with multiple subtypes, the sum of features in the `St-Features` column may be larger than the number of selected features. Here, `Features` and `Genes` refer to the total number of genes and the number of features in each cancer case study and `St-Features` and `St-Genes` to the number of genes and the number of features in each subtype `St`, respectively. A total of 2,385 genes, 3,831 genes, and 5,461 genes were found to be associated with BRCA, COAD, and STAD subtypes, respectively. The majority of genes were exclusively assigned to a certain subtype in all three cancer data sets (**Figure 7B**). This was more intuitive in the tSNE plot in **Figure 7C**. While the number of features was the largest in BRCA, the total number of genes did not necessarily differ with the other cancer types.

The 10-fold cross validated F1 scores of MONTI were 0.844, 0.9, and 0.91 for BRCA, COAD, and STAD, respectively. As far as we are aware of, the classification accuracy are highest among classification results reported in the literature so far and, in our experiments, MONTI outperformed existing methods such as MOFA2, iCluster, and SNF. For BRCA and COAD, the classification accuracy increased significantly when at least two omics data were used involving gene expression omics (GE). Improvement in classification accuracy was dramatic for COAD where use of single omics resulted in poor performance. Interestingly, methylation showed to be more influential in STAD, where ME alone achieved high classification accuracy. The CpG island methylator phenotype (CIMP) information can be used to characterize distinct subtypes of gastric cancer well and it is known that specific methylation patterns and clinicopathological features are associated (Network et al., 2014; Tahara and Arisawa, 2015) with it. While the majority of feature genes were associated with a single subtype (**Figure 7B**), some had membership to multiple. For example, the Venn diagram of BRCA shows that Luminal A and Luminal B subtypes share 265 genes while Her2 and Basal shared 53, which is true in the biological concept. Luminal A and Luminal B are hormone-receptor positive subtypes whereas Her2 and Basal are hormone-receptor negative subtypes, which also reflects the aggressiveness of the cancer (i.e., hormone-receptor negative cancers grow faster). Such characteristics are well-observed in the tSNE plots in **Figure 7C**.

3.3. Performance Evaluation

While few tools are available for multi-omics analysis with the goal of classifying cancer subtypes, all such tools aim to discover genes that have a strong correlation with one or more omics. In other words, such relational information is expected to differ between the cancer subtypes, which information is used to build classifiers or to mine subtype-specific data on genes or features. We compared the BRCA, COAD, and STAD subtype classification accuracy of five methods, which are MONTI, SNF (Wang et al., 2014), MOFA2 (Multi-Omics Factor Analysis) (Argelaguet et al., 2020), iCluster (Shen et al., 2009), and PCA.

The three cancer data sets consist of four subtypes. In BRCA, the number of samples per subtype were 220, 152, 91, and 132 for Luminal A, Luminal B, Her2, and Basal, respectively. In COAD, the number of samples per subtype are 43, 125, 48, 99 for CMS1, CMS2, CMS3, and CMS4, respectively. In STAD, the number of samples per subtype are 188, 26, 42, and 49 for CIN, EBV, GS, and MSI, respectively.

The genes used for analysis were chosen by two criteria. First, only protein coding genes were selected. Second, genes where the methylation values in the TSS 2 k upstream region was missing in more than 80% of the samples were filtered out. The miRNA data was used as is and the target gene information was acquired from mirDB (Chen and Wang, 2020). As a result, 14,514 genes were selected based on the BRCA, COAD, and STAD data sets. Methylation probes with missing values in all samples were dropped, resulting in 62,070 probes. Similarly, miRNAs with zero expression in all samples were excluded, resulting in 1,882 miRNAs. Each omics data were normalized as described in section 2.

The optimal number of ranks for MONTI were selected using the `nmfEstimateRank` function in the R `preprocessCore` package. For each gene-level omics data the optimal number of ranks were investigated based on the dispersion metric, from

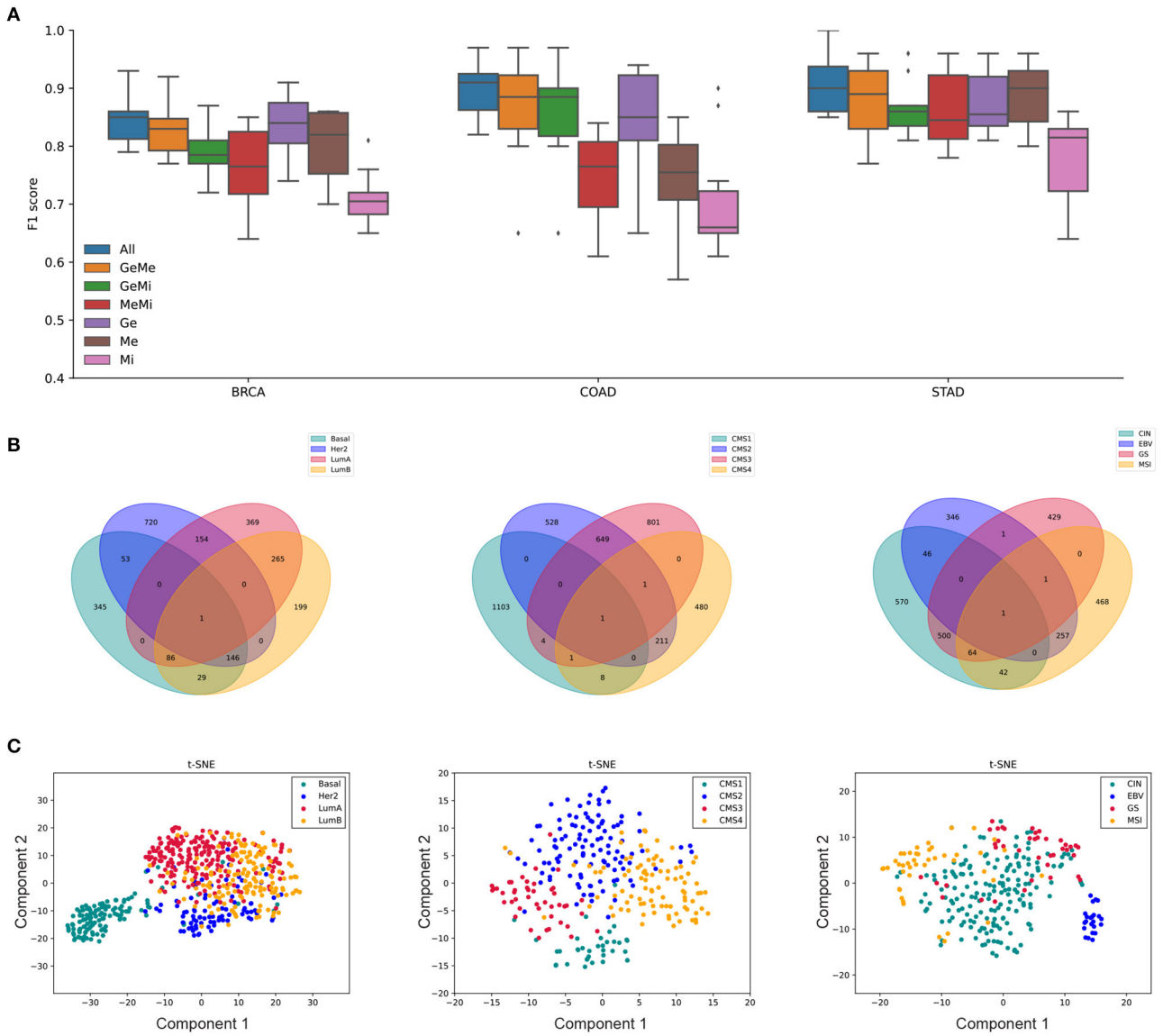


FIGURE 7 | The MONTI analysis results of BRCA, COAD, and STAD subtypes are shown. **(A)** The subtype classification accuracy was the highest when using all three omics data for all three cancer types. **(B)** The cancer subtype specific genes. Here, the genes are shared by at most two different subtypes. **(C)** tSNE plots that were drawn using the selected features from C_p of each cancer type.

which we chose an appropriate rank number based on the elbow method. As a result, 120 ranks were chosen for BRCA, COAD and STAD. As an example, the dispersion plot of BRCA omics data are shown in **Figure 8**. The feature genes omics values were used for measuring the F1 score.

SNF (Similarity Network Fusion) integrates multi-omics data by constructing networks for each omics data in terms of the sample similarity using the omics data and then fusing the networks iteratively using the message-passing method. The principle is to keep edges between samples that are consistent across the different omics networks and to remove that are inconsistent and of low similarity. The optimal hyper parameters K, the number of neighbors in K-nearest neighbor, and T, the number of iterations for the diffusion process, where determined via the parameter grid search. The (K, T) parameters were set as (10, 30), (10, 10), and (5, 20) for BRCA, COAD, and STAD data sets, respectively. The output of SNF is the sample clusters, which was used to measure the F1 score.

MOFA2 utilizes matrix decomposition with the purpose of identifying sources of heterogeneity in multi-omics data sets. It decomposes multiple two-dimensional matrices, where each matrix represents an omics data type comprised of genes and samples. The decomposition yields feature matrices, each associated to one of the input omics matrices, and an additional factor matrix, which represents the activation values of each feature per sample. Thus, if three omics data are given as input,

TABLE 1 | The number of selected features and genes in BRCA, COAD, and STAD.

Case study	Ranks	Features	Genes	Subtypes	St-Features	St-Genes
BRCA	120	26	2,385	Luminal A	10	879
				Luminal B	9	732
				Her2	11	1,080
				Basal	8	665
COAD	120	31	3,831	CMS1	7	1,129
				CMS2	9	1,403
				CMS3	11	1,473
				CMS4	10	704
STAD	120	37	5,461	CIN	9	1,234
				GS	9	1,007
				MSI	9	839
				EBV	8	652

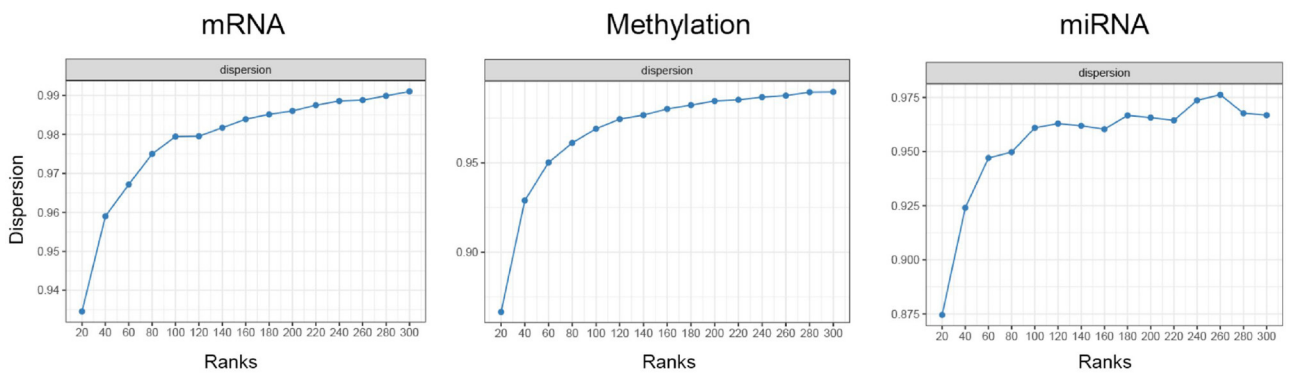


FIGURE 8 | The dispersion plot using different ranks using BRCA omics data for estimating optimal NMF ranks.

they will be decomposed into four matrices (i.e., three feature and one factor matrices). MOFA2 allows to chose the number of factors or features from the decomposed factor matrix, where we utilized as many as possible for each dataset. The maximum features that could be used was 10 for BRCA, COAD, and STAD, respectively. The output of MOFA was the Z sample factor matrix, which was used for measuring the F1 score.

iCluster adopts a joint latent variable model for integrative clustering of multi-omics data. iCluster aims to data mine significant associations between different omics data types through likelihood-inference using the Expectation-Maximization algorithm. iCluster supports a omics optimal weight estimation function, which we used for each data set for clustering. The output of iCluster is the sample clusters, which was used to measure the F1 score.

At last, sample PCA features were extracted and used for classifying the cancer subtypes. For each cancer and omics data, optimal number of PCA features were selected based on the classification accuracy via a parameter grid search. For BRCA, 10, 6, and 10 PCs were selected from gene, methylation, and miRNA data, respectively. Similarly, 8, 5, and 2 PCs for COAD and 20, 2, and 18 PCs for STAD were selected from gene, methylation, and miRNA data, respectively. The selected PCs were stacked and given as input to the random forest classifier to measure the F1 score.

The average F1 score was measured via 10-cross validation for each tool with configurations described above. The train and test data were split before any normalization or feature selection in each BRCA, COAD, and STAD data set. The same train and test data sets were used to measure the F1 score in each method. Furthermore, the input data were both prepared in gene-level (i.e., multi-staged) and omics-level (i.e., multi-dimension) format to observe the difference between the two integration methods. Thus, each method, except MONTI, was subject to two types of input data and were tested for classification accuracy accordingly. The tools measured with gene-level input data are labeled as SNF_g, MOFA2_g, iCluster_g, and PCA_g.

The comparison results are shown in **Figure 9**. The F1 score was the highest in MONTI for all cancer subtypes, followed by iCluster and SNF. We observed that the gene-level input data yielded lower F1 scores in MOFA2, while it remained relatively similar in SNF, iCluster, and PCA methods. The significant drop of F1 score in MOFA2_g may be due to its feature extraction method. While the omics-level input data matrix is very dense, the gene-level matrix is relatively sparse, especially for the miRNA data. Hence, the latent factors associated with the miRNA

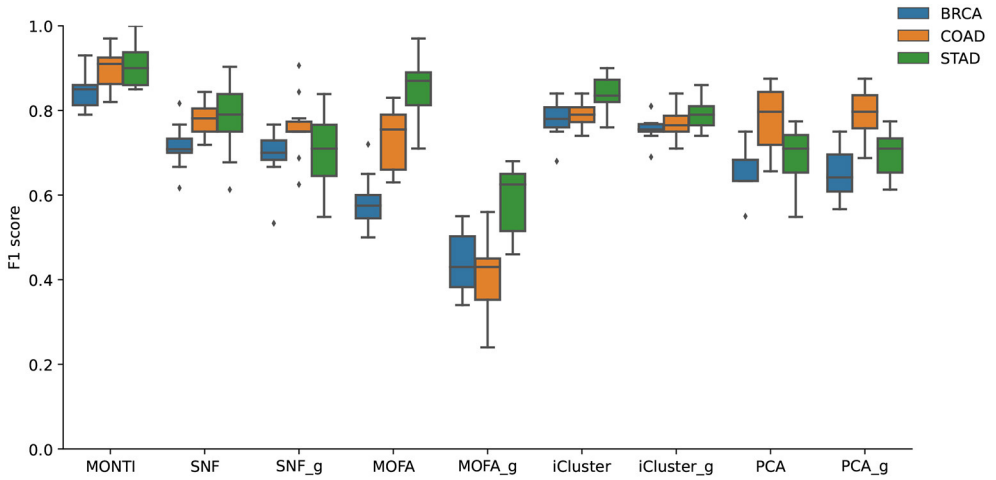


FIGURE 9 | The F1 scores of five tools using gene-level and omics-level data sets of BRCA, COAD, and STAD subtypes.

data will loose information. Furthermore, while MONTI utilizes larger number of rank features, MOFA2 utilized 10 features, which may have reduced the dimension too much, thus, loosing more information accordingly.

3.4. Analysis of Pan-Cancer Clinical Features

The relatively high classification accuracy of the cancer subtypes above implies that they may be explained using the feature extracted genes in terms of multi-omics. Thus, we further investigated whether clinical attributes, other than cancer subtypes, such as gender, mutation groups or metastasis can be explained using multi-omics data. Among the many clinical attributes, categorical attributes with <5 groups were used. Also, clinical attributes with high sample bias were excluded. As a result, a total of nine cancer types and 95 clinical attributes were analyzed using mRNA, methylation and miRNA data. For example, the "Pathologic M" feature of STAD, which is the TNM staging of metastasis, has three classes, which are M0, M1, and MX. If the cancer has spread, the sample is labeled as M0, and if not it is labeled as M1. If metastasis cannot be measured, it is labeled as MX. Thus, similar to the cancer subtype classification, we measured the classification accuracy of each of the categorical clinical attributes that were selected by the criteria described above. The details of the data set and clinical attributes are provided in **Supplementary Table 2**.

MONTI was executed on each cancer type and each clinical feature as described in section 2. The classification accuracy of the cancer clinical attributes are shown in **Figure 10**. Here, we observed that some clinical attributes were well classified while others showed poor classification.

All cancer subtypes showed relatively high accuracy in BRCA, COAD, STAD, and PRAD (Prostate adenocarcinoma), which hints that the multi-omics profile is highly correlated with cancer molecular subtypes. Also, while mutation data was not utilized, the BRAF and RAS mutation classes were well distinguished in THCA (Thyroid carcinoma). From such result, we may infer that at least mRNA, methylation and miRNA omics have causal relationship with BRAF and RAS mutations, which was also reported in Agrawal et al. (2014). In case of HNSC (Head and Neck squamous cell carcinoma), the gender attribute was classified with almost perfect accuracy, which was also reported in Yuan et al. (2016).

The Pan-cancer analysis results show that some clinical attributes are able to be explained using mRNA, methylation and miRNA data while others need further investigation using other omics or clinical data. Collectively, we find that such results may help selecting omics when performing research on clinical features in a cancer cohort.

4. DISCUSSION

While not shown in this study, the subtype classification accuracy decreased when involving certain omics types, particularly with the use of mutation profile data. For BRCA data, the accuracy dropped below 0.75 when SNP data were included in the tensor. The first short-coming of the SNP data was its extreme sparseness (i.e., 0.5% genes with SNP). We further attempted to impute the remaining missing values using the network-based stratification method for tumor mutations (Hofree et al., 2013). Unfortunately, the accuracy further decreased, which may be due to the introduction of additional uncertainty arising from large number of predictions. For sparse data, integration methods that are not gene-centric may be more advantageous, such as SNF. Such result implies that no single method may be universally applicable for incorporating all types of omics data, and that omics data must be well understood and integrated in a manner specific to the characteristics of each omics. Similar arguments have been discussed previously (Zhang et al., 2018).

Clustering of the selected sample features from the C_p component of the BRCA analysis result shows us that the Basal samples are well clustered together, whereas the Luminal A and

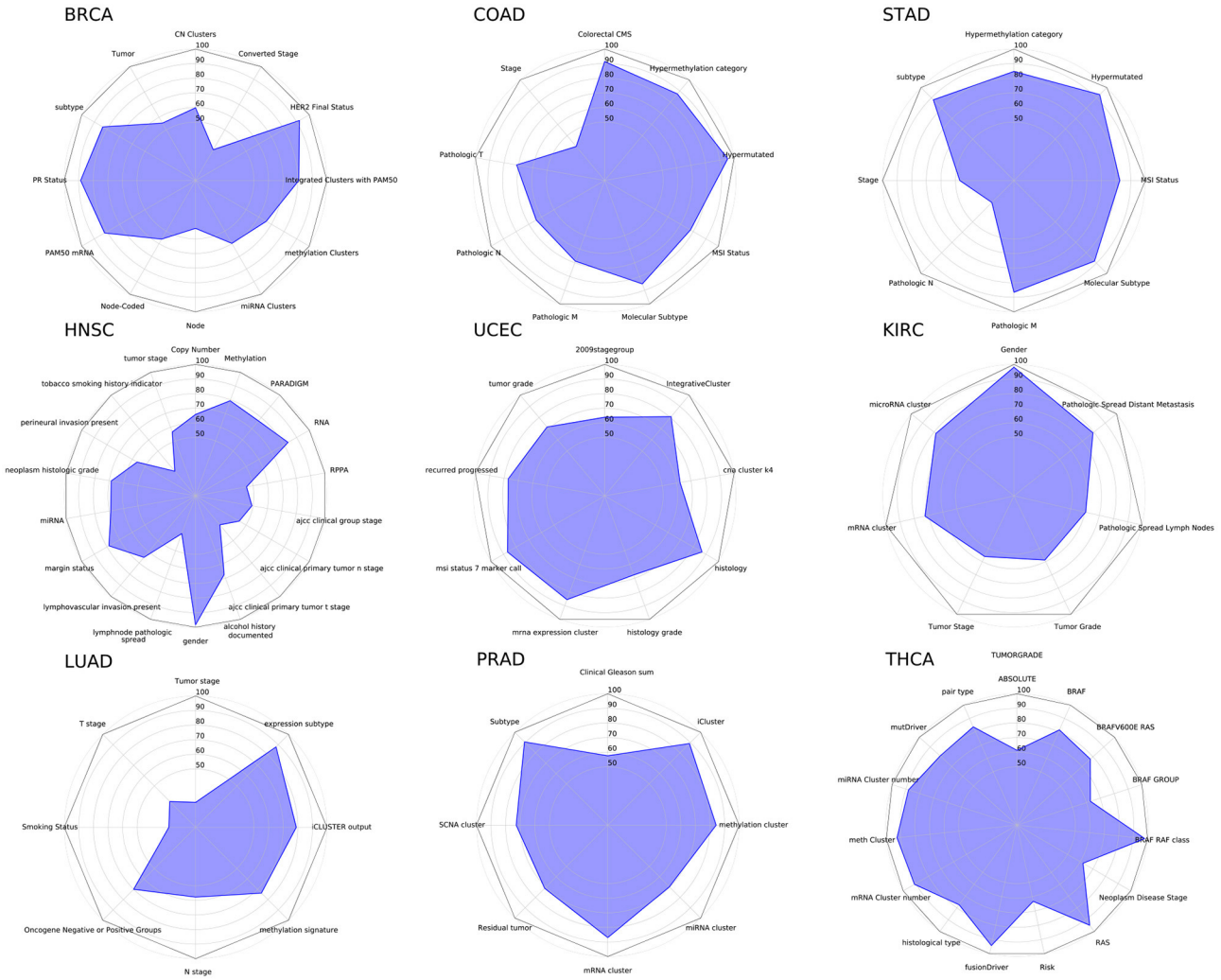


FIGURE 10 | The radar chart showing classification results of nine cancers and their clinical attributes.

Luminal B subtypes are relatively more mixed (**Figure 11A**). Similarly, the clustering of selected feature genes from the C_g component showed the feature activity of genes (**Figure 11B**). Here, the top color bars represent the maximum omics type of each feature. The feature four related genes had strong relation with methylation. Genes with high values in multiple features that are related with different omics types indicate that the gene has relationship across the two different omics types.

Furthermore, the selected features in all three case studies captured correlation among different omics data types. As shown in **Figure 12**, EXOC6 was most affected by DNA methylation in Basal subtype of BRCA. EXOC6 is reported to be an important respondent gene when the effects of a combination of the histone deacetylase inhibitor suberoylanilide hydroxamic acid (SAHA) and taxanes were tested for cytotoxicity using human breast cancer cell lines (Chang et al., 2011). Also, EXOC6 was found to be one out of five genes that was able to asses breast cancer risk with high accuracy (Winham et al., 2017). While EXOC6 was observed to have distinct methylation profiles in brain tissues (Farlik et al., 2016; Hira and Gillies, 2016), it was not actively investigated in breast cancer Basal subtype samples in terms of multi-omics correlation. The OLFML2B gene was found to be negatively correlated with miRNA in the CMS4 subtype in COAD. We found that the miRNA OLFML2B targeting miRNA, miR-30b, is a well-known oncogene suppressor miRNA in colorectal cancer (Liao et al., 2014), which may explain the omics relationship here. At last, the MAPK15 has been reported to be a regulator for redioresistance in nasopharyngeal carcinoma cells, which is tightly linked to the Epstein-Barr virus (EBV) infection (Li et al., 2018), which may relate to the EBV subtype of STAD. Collectively, we may induce that the MAPK15's expression is down-regulated by methylation, which was not the case in other STAD subtypes. Other than the selected genes, well known multi-omics correlated genes related to certain cancer subtypes were also detected. Although data not shown, the ESPL1, detected by MONTI, showed significant regulatory relationship between gene expression and methylation specific to Luminal A and Luminal B subtypes in BRCA, which

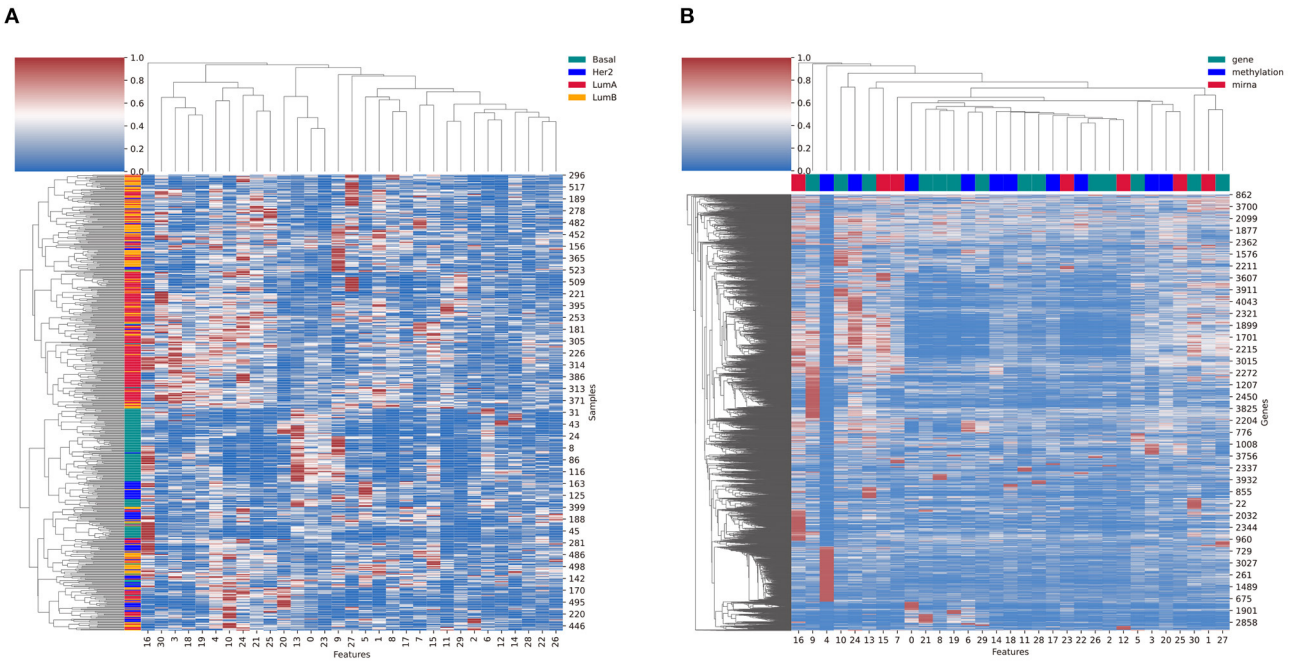


FIGURE 11 | (A) The cluster heatmap of sample features (C_p) and **(B)** the cluster heatmap of the feature genes (C_g) from the breast cancer result. The left color bars in **(A)** refer to the BRCA subtypes. The top color bars in **(B)** refer to the omics with the largest feature value in C_o.

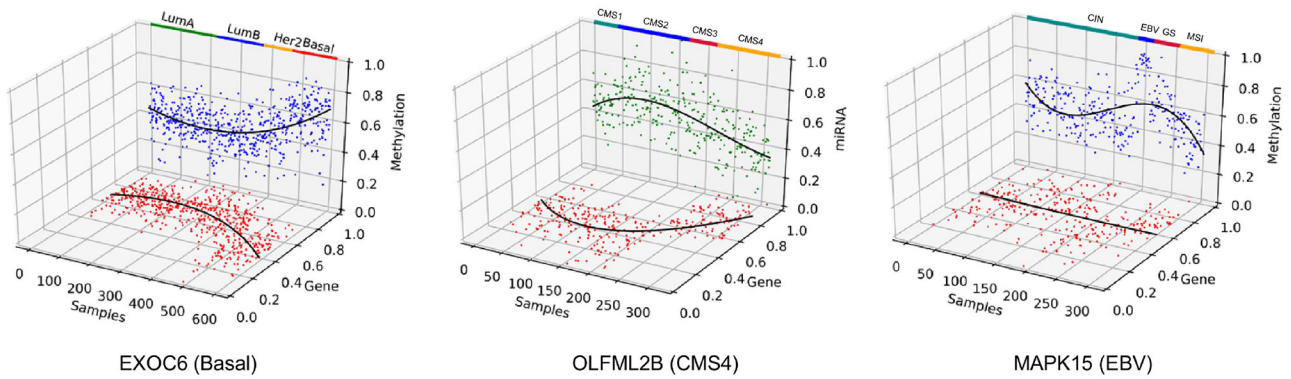


FIGURE 12 | Three genes were selected to show correlation between different types of omics data across patient samples. EXOC6 was associated with the Basal subtype of BRCA, OLFML2B was associated with CMS4 subtype of COAD and MAPK15 was associated with the EBV subtype of STAD.

was previously reported in Finetti et al. (2014) and Li and Li (2020).

OLFML2B was most affected by miRNA in CMS4 subtype of COAD. MAPK15 also showed strong gene expression regulation by methylation in EBV subtype of STAD. This kind of result by MONTI may suggest cancer subtype specific gene regulation mechanisms, which can help discover subtype-specific gene markers for further biological and clinical investigations.

The genes were further examined to see if they captured known signals of cancer subtype specific pathways by applying the Subsystem Activation Scoring (SAS) method (Lim et al., 2016). SAS is used to decompose molecular pathways into sub-pathways (named subsystems) and measure the activation levels of them in terms of gene expression. We expanded it to multi-omics levels to evaluate the association of each subsystem with each cancer subtype by constructing random forest classifiers using its SAS score. The detailed method and results are described in **Supplementary Table 3**. The detected pathway subsystems were highly specific to each cancer type. For example, the top 10 ranked pathways for the three case studies were all supported by previous studies. For example, the "Fanconi anemia" pathway was the top ranked pathway for the BRCA data, which is known to be a rare chromosomal instability disorder that is susceptible to cancer (Alan and D'Andrea, 2010). The "HIF-1 signaling" pathway was top ranked in STAD with association to miRNA. The study (He et al., 2017) suggests that miR-224 promotes cell growth migration and invasion by targeting the RASSF8 gene in STAD. Similarly, the top ranked

"Vascular smooth muscle contraction" pathway by SAS was also reported to be induced by colorectal cancer (Li et al., 2017).

The application of MONTI was demonstrated on cancer subtype multi-omics data. However, MONTI is not tailored to cancer subtype analysis but can be utilized to identify any categorical clinical features, such as gender, mutation groups, tumor grade, or age. Thus, the advantage of MONTI is that it is able to identify clinical feature associated genes in terms of multi-omics. Furthermore, the omics component C_o can be further used to investigate which omics are currently active and take part in gene expression regulation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: TCGA multi-omics data.

AUTHOR CONTRIBUTIONS

SK and IJ designed the project and MONTI algorithm framework. IJ and SR implemented multi-omics integration. SK, IJ, SL, and MK performed the biological analysis and interpretation.

FUNDING

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (2019M3E5D3073365), the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (No. NRF-2014M3C9A3063541), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020M3C9A5085604).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.682841/full#supplementary-material

REFERENCES

Agrawal, N., Akbani, R., Aksoy, B. A., Ally, A., Arachchi, H., Asa, S. L., et al. (2014). Integrated genomic characterization of papillary thyroid carcinoma. *Cell* 159, 676–690. doi: 10.1016/j.cell.2014.09.050

Alan, D., and D'Andrea, M. (2010). The fanconi anemia and breast cancer susceptibility pathways. *N. Engl. J. Med*. 362, 1909–1919. doi: 10.1056/NEJMra0809889

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., et al. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol*. 21, 1–17. doi: 10.1186/s13059-020-02015-1

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-omics factor analysis–a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol*. 14:e8124. doi: 10.15252/msb.20178124

Bro, R. (1997). Parafac. Tutorial and applications. *Chemometr. Intell. Lab. Syst*. 38, 149–171.

Buenrostro, J. D., Wu, B., Chang, H. Y., and Greenleaf, W. J. (2015). Atac-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol*. 109, 21–29. doi: 10.1002/0471142727.mb2129s109

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., et al. (2015). A novel approach to high-quality postmortem tissue procurement: the GTEx project. *Biopreserv. Biobank*. 13, 311–319. doi: 10.1089/bio.2015.0032

Carroll, J. D., and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart–young" decomposition. *Psychometrika* 35, 283–319. doi: 10.1007/BF02310791

Chang, H., Jeung, H.-C., Jung, J. J., Kim, T. S., Rha, S. Y., and Chung, H. C. (2011). Identification of genes associated with chemosensitivity to saha/taxane combination treatment in taxane-resistant breast cancer cells. *Breast Cancer Res. Treatm*. 125, 55–63. doi: 10.1007/s10549-010-0825-z

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2017). Deep learning based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res*. 24, 1248–1259. doi: 10.1101/114892

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

Farlik, M., Halbritter, F., Müller, F., Choudry, F. A., Ebert, P., Klughammer, J., et al. (2016). DNA methylation dynamics of human hematopoietic stem cell differentiation. *Cell Stem Cell* 19, 808–822. doi: 10.1016/j.stem.2016.10.019

Finetti, P., Guille, A., Adelaide, J., Birnbaum, D., Chaffanet, M., and Bertucci, F. (2014). ESPL1 is a candidate oncogene of luminal b breast cancers. *Breast Cancer Res. Treatm*. 147, 51–59. doi: 10.1007/s10549-014-3070-z

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16, 906–914. doi: 10.1093/bioinformatics/16.10.906

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with BAYESIAN networks. *Bioinformatics* 22, e184–e190. doi: 10.1093/bioinformatics/btl230

Harshman, R. A. (1970). "Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis," in *UCLA Working Papers in Phonetics* (Los Angeles, CA).

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol*. 18:83. doi: 10.1186/s13059-017-1215-1

He, C., Wang, L., Zhang, J., and Xu, H. (2017). Hypoxia-inducible microrna-224 promotes the cell growth, migration and invasion by directly targeting rassf8 in gastric cancer. *Mol. Cancer* 16:35. doi: 10.1186/s12943-017-0603-1

Hernández-de Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furio-Tari, P., Pappas, G. J., et al. (2018). PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucl. Acids Res*. 46, W503–W509. doi: 10.1093/nar/gky466

Hira, Z. M., and Gillies, D. F. (2016). Identifying significant features in cancer methylation data using gene pathway segmentation. *Cancer Inform*. 15, 189–198. doi: 10.4137/CIN.S39859

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10:1108. doi: 10.1038/nmeth.2651

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet*. 8:84. doi: 10.3389/fgene.2017.00084

Kroonenberg, P. M. (1983). *Three-Mode Principal Component Analysis: Theory and Applications, Vol. 2*. Los Angeles, CA: DSWO Press.

Li, J., and Li, X. (2020). Comprehensive analysis of prognosis-related methylated sites in breast carcinoma. *Mol. Genet. Genom. Med.* 8:e1161. doi: 10.1002/mgg3.1161

Li, W.-W., Wang, H.-Y., Nie, X., Liu, Y.-B., Han, M., and Li, B.-H. (2017). Human colorectal cancer cells induce vascular smooth muscle cell apoptosis in an exocrine manner. *Oncotarget* 8:62049. doi: 10.18632/oncotarget.18893

Li, Y., Oosting, M., Smeekens, S. P., Jaeger, M., Aguirre-Gamboa, R., Le, K. T., et al. (2016). A functional genomics approach to understand variation in cytokine production in humans. *Cell* 167, 1099–1110. doi: 10.1016/j.cell.2016.10.017

Li, Z., Li, N., Shen, L., and Fu, J. (2018). Quantitative proteomic analysis identifies MAPK15 as a potential regulator of radioresistance in nasopharyngeal carcinoma cells. *Front. Oncol.* 8:548. doi: 10.3389/fonc.2018.00548

Liao, W.-T., Ye, Y.-P., Zhang, N.-J., Li, T.-T., Wang, S.-Y., Cui, Y.-M., et al. (2014). MicroRNA-30b functions as a tumour suppressor in human colorectal cancer by targeting KRAS, PIK3CD and BCL2. *J. Pathol.* 232, 415–427. doi: 10.1002/path.4309

Lim, S., Lee, S., Jung, I., Rhee, S., and Kim, S. (2018). Comprehensive and critical evaluation of individualized pathway activity measurement tools on pan-cancer data. *Brief. Bioinform.* 21, 36–46. doi: 10.1093/bib/bby097

Lim, S., Park, Y., Hur, B., Kim, M., Han, W., and Kim, S. (2016). Protein interaction network (PIN)-based breast cancer subsystem identification and activation measurement for prognostic modeling. *Methods* 110, 81–89. doi: 10.1016/j.ymeth.2016.06.015

Paquet, E. R., and Hallett, M. T. (2015). Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl. Cancer Instit.* 10:357. doi: 10.1093/jnci/dju357

Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10:669. doi: 10.1038/nrg2641

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27:1160. doi: 10.1200/JCO.2008.18.1370

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. U.S.A.* 98, 15149–15154. doi: 10.1073/pnas.211566398

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* 16, 85–97. doi: 10.1038/nrg3868

Sathyanarayanan, A., Gupta, R., Thompson, E. W., Nyholt, D. R., Bauer, D. C., and Nagaraj, S. H. (2020). A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Brief. Bioinformatics* 21, 1920–1936. doi: 10.1093/bib/bbz121

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., et al. (2012). Integrative subtype discovery in glioblastoma using icluster. *PLoS ONE* 7:e35236. doi: 10.1371/journal.pone.0035236

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 2906–2912. doi: 10.1093/bioinformatics/btp543

Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., et al. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U.S.A.* 100, 10393–10398. doi: 10.1073/pnas.1732912100

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14:1177932219899051. doi: 10.1177/1177932219899051

Tahara, T., and Arisawa, T. (2015). Dna methylation as a molecular biomarker in gastric cancer. *Epigenomics* 7, 475–486. doi: 10.2217/epi.15.4

Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., et al. (2019). Classifying breast cancer subtypes using multiple kernel learning based on omics data. *Genes* 10:200. doi: 10.3390/genes10030200

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57. doi: 10.1038/nmeth.2238

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2017). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963. doi: 10.1093/nar/gkx1090

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics* 26, i237–i245. doi: 10.1093/bioinformatics/btq182

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11:333. doi: 10.1038/nmeth.2810

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

Winham, S. J., Mehner, C., Heinzen, E. P., Broderick, B. T., Stallings-Mann, M., Nassar, A., et al. (2017). Nanostring-based breast cancer risk prediction for women with sclerosing adenosis. *Breast Cancer Res. Treat.* 166, 641–650. doi: 10.1007/s10549-017-4441-z

Wu, T., Wang, Y., Jiang, R., Lu, X., and Tian, J. (2017). A pathways-based prediction model for classifying breast cancer subtypes. *Oncotarget* 8:58809. doi: 10.18632/oncotarget.18544

Yuan, Y., Liu, L., Chen, H., Wang, Y., Xu, Y., Mao, H., et al. (2016). Comprehensive characterization of molecular differences in cancer between male and female patients. *Cancer Cell* 29, 711–722. doi: 10.1016/j.ccell.2016.04.001

Zhang, S., Liu, C.-C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucl. Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Zhang, W., Ma, J., and Ideker, T. (2018). Classifying tumors by supervised network propagation. *Bioinformatics* 34, i484–i493. doi: 10.1093/bioinformatics/bty247

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

CORRECTION
published: 25 October 2021
doi: 10.3389/fgene.2021.778490

Corrigendum: MONTI: A Multi-Omics Non-Negative Tensor Decomposition Framework for Gene-Level Integrative Analysis

Inuk Jung[1]*, *Minsu Kim*[2], *Sungmin Rhee*[3], *Sangsoo Lim*[4] and *Sun Kim*[2,3,4]*

[1]Department of Computer Science and Engineering, Kyungpook National University, Daegu, South Korea, [2]Computing and Computational Sciences Directorate, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [3]Department of Computer Science and Engineering, Seoul National University, Seoul, South Korea, [4]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea

Keywords: feature selection, tensor decomposition, cancer, multi-omics, integrative analysis

A corrigendum on

MONTI: A Multi-Omics Non-Negative Tensor Decomposition Framework for Gene-Level Integrative Analysis

by Jung, I., Kim, M., Rhee, S., Lim, S., and Kim, S. (2021). Front. Genet. 12:682841. doi: 10.3389/fgene.2021.682841

There is an error in the **Funding statement**. The correct number for "the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education" is "2020M3C9A5085604." Corrected statement is given below:

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (2019M3E5D3073365), the Collaborative Genome Program for Fostering New Post-Genome Industry of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (MSIT) (No. NRF-2014M3C9A3063541), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020M3C9A5085604).

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

***Correspondence:**
Inuk Jung
inukjung@knu.ac.kr
Sun Kim
sunkim.bioinfo@snu.ac.kr

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

Received: *17 September 2021*
Accepted: *30 September 2021*
Published: *25 October 2021*

Citation:
Jung I, Kim M, Rhee S, Lim S and Kim S (2021) Corrigendum: MONTI: A Multi-Omics Non-Negative Tensor Decomposition Framework for Gene-Level Integrative Analysis. Front. Genet. 12:778490. doi: 10.3389/fgene.2021.778490

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 September 2021
doi: 10.3389/fgene.2021.700086

In vivo Estimation of Breast Cancer Tissue Volume in Subcutaneous Xenotransplantation Mouse Models by Using a High-Sensitivity Fiber-Based Terahertz Scanning Imaging System

Hua Chen[1], Juan Han[1], Dan Wang[1], Yu Zhang[1], Xiao Li[2] and Xiaofeng Chen[2]*

[1] School of Physics, Southeast University, Nanjing, China, [2] The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:
Fengfeng Zhou, Jilin University, China

Reviewed by:
Pekka Ruusuvuori, University of Turku, Finland
Vikram Dalal, Washington University in St. Louis, United States

***Correspondence:**
Hua Chen
chenhua@seu.edu.cn

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

Received: *25 April 2021*
Accepted: *20 August 2021*
Published: *27 September 2021*

Citation:
Chen H, Han J, Wang D, Zhang Y, Li X and Chen X (2021) In vivo Estimation of Breast Cancer Tissue Volume in Subcutaneous Xenotransplantation Mouse Models by Using a High-Sensitivity Fiber-Based Terahertz Scanning Imaging System. Front. Genet. 12:700086. doi: 10.3389/fgene.2021.700086

Absorption contrast between the terahertz (THz) frequency range of fatty and cancer tissues allows cancer diagnosis by THz imaging. We successfully demonstrated the ability of THz imaging to measure small breast cancer volume in the subcutaneous xenotransplantation mouse models even without external comparison. We estimated the volume detection limitation of the fiber-based THz scanning imaging system using a highly sensitive cryogenic-temperature-operated Schottky diode detector to be smaller than 1 mm^3, thus showing the potential application of this technique in preliminary early cancer diagnosis.

Keywords: THz, imaging, mouse model, breast cancer, caner volume

INTRODUCTION

Terahertz (THz) wavelength is from 0.003 to 3.0 mm, which is longer than far-infrared and light wave, so the scattering in a biological tissue is greatly reduced and no harmful photoionization occurs for the low photon energy (Kindt and Schmuttenmaer, 1996). Meanwhile, THz waves are very sensitive to polar substances (Pedersen and Keiding, 1992; Wang et al., 2010; Yamada et al., 2014) and can provide better contrast for the biological tissue than x-ray. So far, researchers have detected various human cancers by using THz wave. For example, skin cancer has been the focus of THz imaging research in recent years. It has been confirmed by *in vivo* and *in vitro* models that THz has a high diagnostic rate for the boundary and depth of invasion of skin cancer (Woodward et al., 2002; Rahman et al., 2016). Pickwell et al. (2005) measured the THz refractive index and absorptivity of normal tissues and cancer tissues of 10 patients with basal cell carcinoma and showed that the absorption characteristics of cancer tissues were significantly different from those of healthy tissues; this contrast between the two tissues proved that THz imaging can be used as a non-invasive

diagnostic tool for skin cancer. Fitzgerald et al. (2006) analyzed the THz images of isolated breast cancer tissues and compared the imaging edge with pathological examination results. Reese (Reid et al., 2011) and other researchers have studied the THz images of freshly resected colorectal cancer tissues and found that normal tissues have a good contrast with cancer tissues and it is possible to detect cancer in esophagus, colon, bladder, prostate, and other deep tissues by THz endoscopic imaging equipment (Wang and Mittleman, 2004).

Breast cancer is the second most common cancer affecting women and accounts for 23% of all cancer cases. Moreover, it is also the main cause of cancer death for females, and the mortality rate is 14% of the all cancer deaths (Jemal et al., 2011). Recently, several preliminary clinical studies have reported that the THz absorption contrast method could be used to diagnose breast tumors from normal tissues (Fitzgerald et al., 2006; Ashworth et al., 2009; Chen et al., 2011a,b; Bowman et al., 2017a,b, 2018; Chavez et al., 2018), and the contrast is induced by water content and cancer-induced structure change (Ashworth et al., 2009; Chen et al., 2011a,b). In our previous study, we not only demonstrated that THz wave can clearly identify breast cancer tissue without any other H&E staining (Bowman et al., 2017b), but also realize early detection breast cancer in the nude mice (Chen et al., 2011a). However, the detection capability is limited to tissues thinner than 5 mm (Chen et al., 2011a), which is too thin compared to the thickness of an actual female breast under magnetic resonance imaging or x-ray (>5 cm), thus limiting further clinical applications. In this study, the capability was improved to 8 cm by applying a high-sensitive cryogenic-temperature-operated Schottky diode detector to the fiber-based THz scanning imaging system. Using this technique, we realized *in vivo* early breast cancer detection in a subcutaneous xenotransplantation mouse model without any external comparison, and even estimated the detection limit of the THz imaging system to be smaller than 1 mm^3, which is a great advantage compared to the current detection limit of x-ray mammography (2 mm diameter).

EXPERIMENTAL

Setup of the Terahertz Imaging System

The results of *ex vivo* THz spectroscopy of thin breast tissue sections (Fitzgerald et al., 2006; Bowman et al., 2017b) revealed that high tissue absorption leads to low penetration depth, which makes transmission imaging difficult. However, the THz absorption of the breast tissue decreases at lower frequency, so we use 108 GHz frequency for *in vivo* imaging. A schematic picture of the fiber-based THz imaging system used in this study is shown in **Figure 1**. The parameters of polyethylene (PE) fibers (Chen et al., 2006, 2007; Lu et al., 2008) and the working principle of the system remains unchanged from those described in our previous system (Chen et al., 2011a). Briefly, the THz wave is radiated from a YIG oscillator module, and then the THz wave is collected by a pair of off-axis parabolic mirrors and focused into the PE sub-wavelength fiber with a diameter of 600 μm and a length of 45 cm (Chen et al., 2011a). Finally, the THz wave coupling by TE fiber is focused by a PE lens onto the sample and then the transmitted power is detected by the detector. To improve detection sensitivity, we introduced a cryogenic-temperature-operated Schottky diode detector with a working temperature of approximately 4 K. Cooling the Schottky diode detector reduces noise significantly, thus enhancing the sensitivity to 10^{-13} W/Hz with the same dynamic range and response time. Finally, a lock-in amplifier will analyze the collected signals. The image is obtained by two-dimensional (X–Y) direct scanning of the output end of the fiber with an imaging time of less than 1 min. The results show the signal-to-noise ratio of the imaging system to be about 10^8:1, which is improved about 10^3 times compared to our previous imaging system (Chen et al., 2011a).

Mouse Treatment

This work is approved by the Institutional Animal Care and Use Committee of Southeast University and Nanjing Medical University (No. 3207027381). We purchased 4- to 6-month-old female BALB/cAnN.Cg-Foxnlnu/CrlNarl mice, an immune inhibited laboratory mouse strain unable to reject breast cancer cell injection and fatty tissue xenograft from another species, from Slac Laboratory Animal, Shanghai, China.

To induce breast cancer, we directly implanted 0.3 ml of MDA MB 231 breast cancer cells into the dermis layer of the mouse skin. The cancer cells were cultivated in L-15 with 10% fetal bovine serum and 1% antibiotics to a cell concentration of 5×10^7 per milliliter of culture media. After injecting the cancer cells, we immediately marked the injection area, kept the mice warm around 36°C, and restored them to health. On the seventh day, we implanted mouse fatty tissue to embed the breast cancer cells. The implanted or *ex vivo* measured fatty tissue was aspirated from 12-week-old female B6.V-Lepob/J mice and rinsed thrice in the transport medium [NaCl 0.9% (w/v), glucose 56 mM, HEPES 25 mM, and PSA 10 ml (pH 7.4)]. The cancer cells and fatty tissue implantations as well as *in vivo* THz imaging were conducted after anesthetizing the mice by injecting ketamine-xylazine (50 + 15 mg/kg) intraperitoneally. THz imaging was conducted 7 days after fatty tissue implantation.

THz Absorption Spectra of Mouse Tissues

We first *in vivo* measured the mouse skin, fatty tissue, and breast cancer tissue by THz absorption spectroscopy at 108–143 GHz. To extract the properties of the constituent tissue types, THz absorbance (α) was averaged linearly by assuming that any reflections and scattering caused by heterogeneities within samples were negligible. The absorbance was calculated according to the Beer-Lambert law $\alpha = \ln(I_s/I_b)/d$, where I_s is the transmitted power of the THz wave through samples, I_b is the background (transmission power of THz wave through the cover glass), and d is the thickness of tissues. As shown in **Figure 2A**, after anesthetizing the mouse, we sandwiched the embedded dorsal area with two cover glasses. Then, the THz absorption spectra were measured by the YIG oscillator module and Schottky diode detector mentioned in section "Setup of THz Imaging System." The corresponding

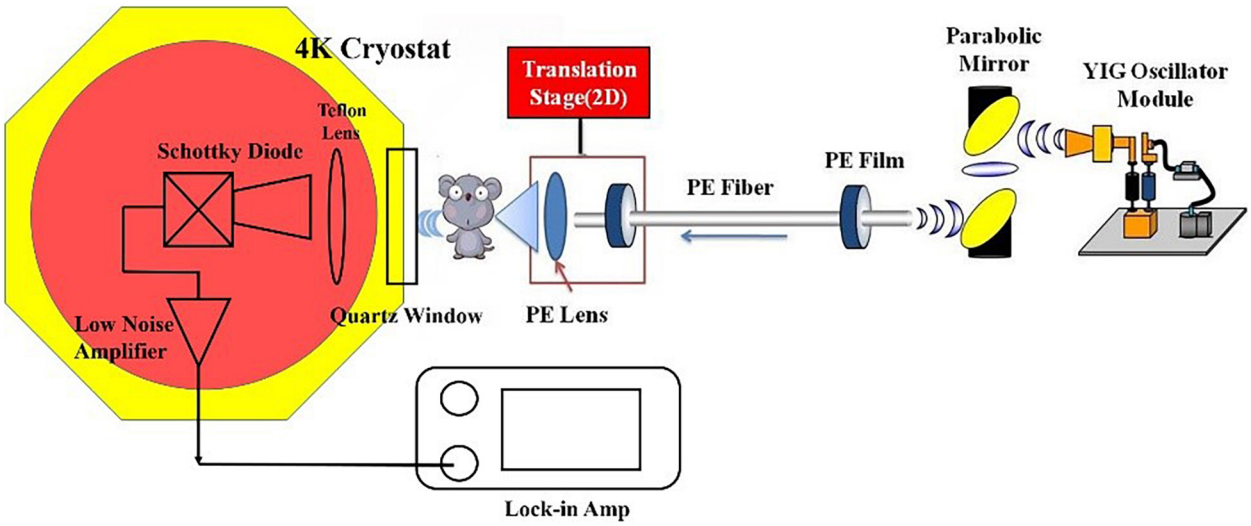


FIGURE 1 | Schematic picture of the imaging system. Fiber parameters: diameter: 600 μm; length: 40 cm; attenuation coefficient: 10^{-3} cm^{-1}.

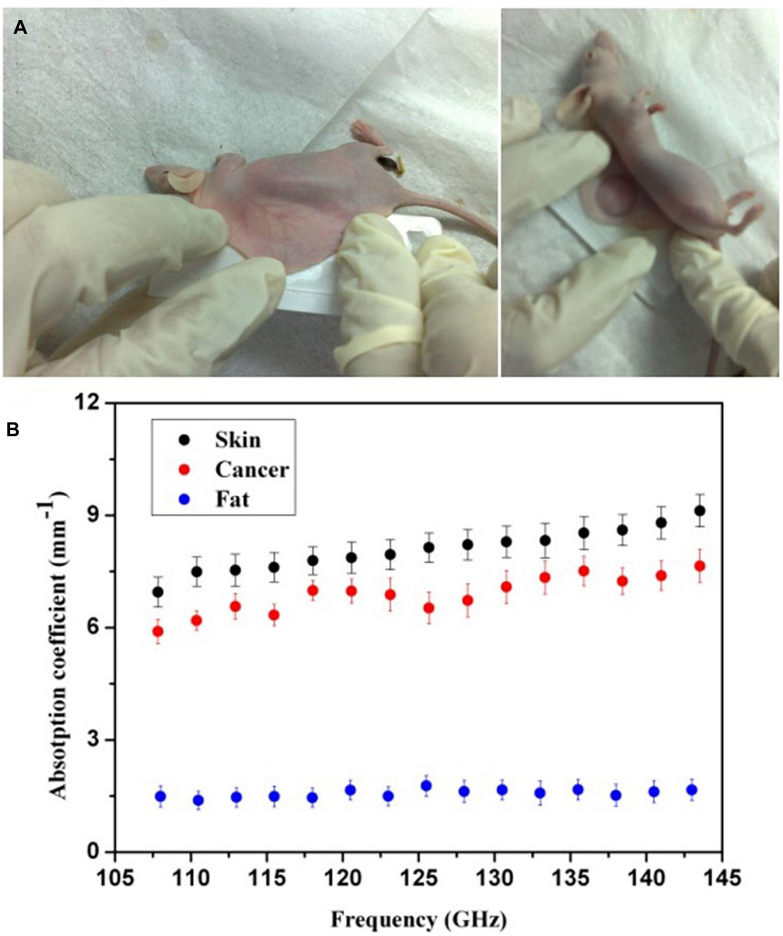


FIGURE 2 | (A) Photos of a mouse processing. **(B)** Absorption spectra of mouse skin and fatty and breast cancer tissues (black, blue, and red solid circles, respectively). The error bars represent the standard deviation of the mean; $n = 20$.

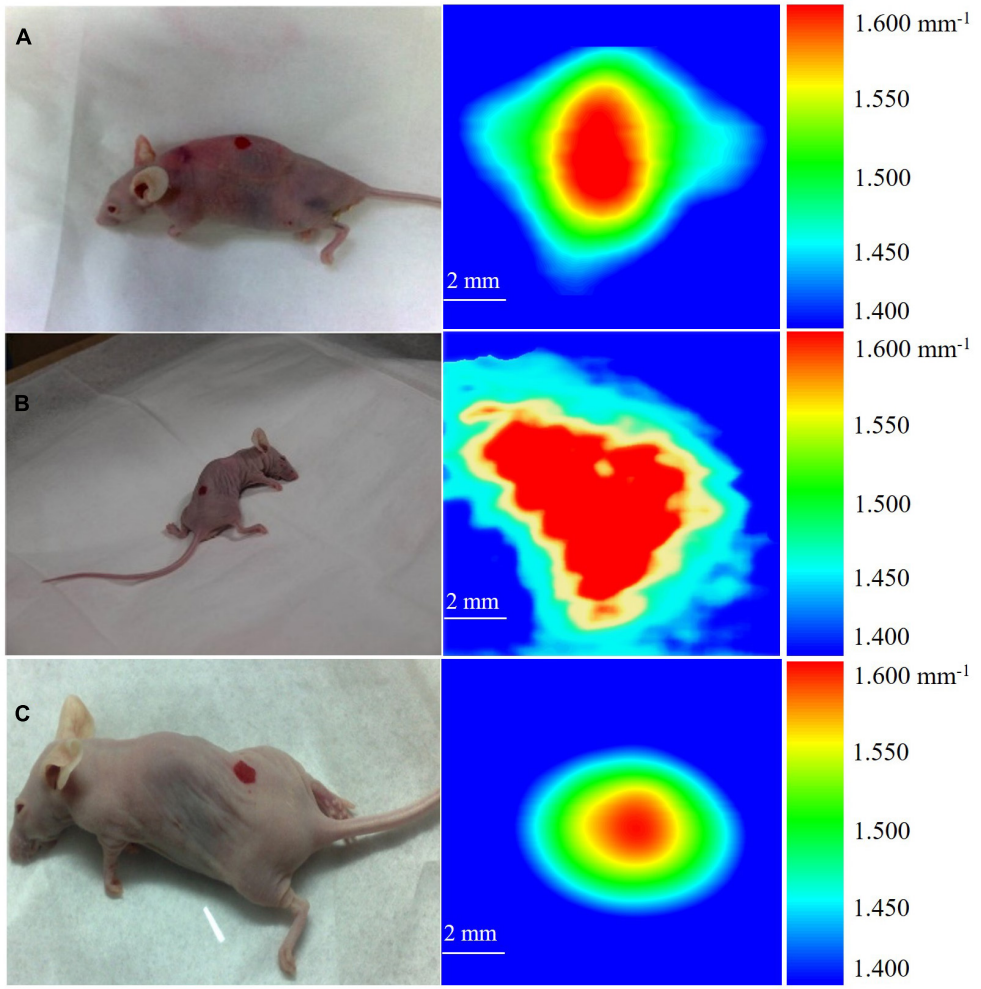


FIGURE 3 | THz images of breast cancer in three mice. **(A)** The cancer in mouse is about 0.480 mm^3. **(B)** The cancer in mouse is about 0.853 mm^3. **(C)** The cancer in mouse is about 0.704 mm^3.

absorption coefficients were calculated from the measurements in 20 mice, which is shown in **Figure 2B**. It has been clearly found that THz absorption spectra can differentiate between fatty and cancer tissues and the absorption coefficients of cancer tissues are much higher than those of the fatty tissues. As the water content of breast cancer tissues is higher than normal tissues, we believe that the water content in tissues may be the most related and dominant factor for the absorption contrast (Chen et al., 2015). Meanwhile, we found that the absorption coefficients of skin and cancer tissue were similar. Considering that skin thickness is relatively uniform and will not vary with time significantly, we calibrated the attenuation due to skin as a uniform and position-independent attenuation background. Moreover, considering the sensitivity of the cryogenic-temperature-operated detection system and absorption coefficients of mouse skin and fatty and breast cancer tissues, we estimated that the penetration capability of our system can be improved to 8 cm, which is similar to the average breast thickness in Asian females.

RESULTS

After the cancer cell injection, on the 7th day, we anesthetized the mouse and implanted mouse fatty tissue to embed the cancer cells. Starting from the 14th day, we measured the cancer implanted area (marked as red area in the picture of **Figure 3**) by THz imaging daily. The mouse was anesthetized and the dorsal cancer area was sandwiched by two cover glasses. Finally, once the scanning completed, mice were monitored, kept warm to 36°C, and allowed to wake up naturally. For further studies on estimating the breast cancer size in the mouse model, we first tested the sensitivity of the THz imaging system with 10 mice, and the limitation was investigated in three mice.

Figure 3 shows the 10 × 10 mm^2 THz images of three mice acquired on the 14th day after cancer cell implantation. During the imaging process, each mouse was scanned three times and the images were presented in the form of the mean absorption coefficient (α). We calibrated the attenuation due to skin as a uniform and position-independent

attenuation background. The scanned images show that the high absorption of breast cancer tissue provides endogenous contrast under THz imaging, making it easy to distinguish from the background absorption. We defined the color bar by absorption coefficient α from 1.400 mm^{-1} to 1.600 mm^{-1}. The background of the image, shown as blue color, is defined as 1.400 $mm^{-1} < \alpha < 1.450\ mm^{-1}$, corresponding to the absorption coefficient of fatty tissue (according to **Figure 2**). The absorption coefficients of the sandwiched tissues induced with early cancer development is 1.450 $mm^{-1} < \alpha < 1.600\ mm^{-1}$, while 1.600 mm^{-1} is the maximum absorption coefficient and is shown as red color. Since early cancer development differs individually, the absorption change Δα will be different for each individual, and the absorption change Δα for these three tested mice was 0.090, 0.160, and 0.132 mm^{-1}, respectively.

According to the concept of cell absorption cross (σ), we estimated tumor volume in these three mice. σ is defined as: $\sigma = \alpha/N = \alpha \times V_{cell}$, where N is the number of absorbing cells per unit volume and V_{cell} is the volume of a single cancer cell. The development of cancer cells embedded in fat then induced Δα and the corresponding cancer cell density N' was described as $N' = \Delta\alpha/\sigma = \Delta\alpha/(\alpha \times V_{cell})$. Finally, the volume of the total cancer tissue V was evaluated. Through the THz absorption spectra shown in **Figure 3**, we calibrated the value of σ. As shown in **Figure 3**, the measured absorption changes 0.090, 0.160, and 0.132 mm^{-1} in the three mice correspond to V = 0.480, 0.853, and 0.704 mm^3, respectively, while the sensitivity of x-ray mammography depends on breast density (Nass et al., 2001) and the detection limitation is as small as 2 mm in diameter (Onuigbo et al., 2001) currently.

DISCUSSION

According to our previous study on human breast cancer, we proved that THz imaging can clearly diagnose breast cancer tissues (Chen et al., 2011b) and detect cancer volume (Chen et al., 2011a). However, the detection capability of the imaging system is far from clinical application, for the reason that the detection thickness of the former system is smaller than 5 mm (Chen et al., 2011a). In this study, we successfully improved the capability to 8 cm and clinical application would become possible compared to the thickness of an actual female breast. In order to further demonstrate the potential clinical application of THz imaging in the detection of small breast cancer tissue volume, we conducted this study in mouse models. The results show that THz imaging has high sensitivity and potential for non-invasive early cancer detection without exogenous contrast. In this work, we did not consider human breast fibrous tissue because the available subcutaneous xenotransplantation animal models prevent us from implanting fibrous tissue to simulate real females breast conditions. However, the THz absorption spectra can distinguish breast cancer tissue from fibrous tissue very well (Fitzgerald et al., 2006; Ashworth et al., 2009; Bowman et al., 2017b). The future potential, specificity, and penetration ability for *in vivo* imaging in humans needs to be studied.

CONCLUSION

The fiber-based THz scanning imaging system based on cryogenic detection system was used to study human breast cancer tissue volume in the mouse model. Results show that THz imaging can not only monitor cancer development in real time but also identify small cancer tissue volume, and all the measurements are conducted without the need of exogenous contrast. Through calculation, we found that this method may be used to detect cancer tissue volume smaller than 1 mm^3, which is highly advantageous compared to the current detection limit (2 mm) of x-ray mammography. This non-invasive and non-ionizing imaging method has a potential application to breast cancer volume detection.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Care and Use Committee of Southeast University and Nanjing Medical University (No. 3207027381).

AUTHOR CONTRIBUTIONS

HC, JH, and DW conducted this study and clinical trials. DW, YZ, and JH conducted the experiments. XC and XL analyzed the data. HC wrote the main manuscript text. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by a Joint Research Project by Southeast University and Nanjing Medical University No. 3207027381.

REFERENCES

Ashworth, P. C., Pickwell-MacPherson, E., Provenzano, E., Pinder, S. E., Purushotham, A. D., Pepper, M., et al. (2009). Terahertz pulsed spectroscopy of freshly excised human breast cancer. *Opt. Express* 17, 12444–12454. doi: 10.1364/oe.17.012444

Bowman, T., Chavez, T., Khan, K., Wu, J., Chakraborty, A., Rajaram, N., et al. (2018). Pulsed terahertz imaging of breast cancer in freshly excised murine tumors. *J. Biomed. Opt.* 23:026004.

Bowman, T., Walter, A., Shenderova, O., Nunn, N., Guire, G. M., and El-Shenawee, M. (2017a). A phantom study of terahertz spectroscopy and imaging of micro- and nano-diamonds and nano-onions as contrast agents for

breast cancer. *Biomed. Phys. Eng. Express* 3:055001. doi: 10.1088/2057-1976/aa87c2

Bowman, T., Wu, Y., Gauch, J., Campbell, L. K., and El-Shenawee, M. (2017b). Terahertz imaging of three-dimensional dehydrated breast cancer tumors. *J. Infrared Millim. Terahertz Waves* 38, 766–786. doi: 10.1007/s10762-017-0377-y

Chavez, T., Bowman, T., Wu, J., Bailey, K., and El-Shenawee, M. (2018). Assessment of terahertz imaging for excised breast cancer tumors with image morphing. *J. Infrared Millim. Terahertz Waves* 39, 1283–1302. doi: 10.1007/s10762-018-0529-8

Chen, H., Chen, T. H., Tseng, T. F., Lu, J. T., Kuo, C. C., Fu, S. C., et al. (2011a). High-sensitivity in vivo THz transmission imaging of early human breast cancer in a subcutaneous xenograft mouse model. *Opt. Express* 19, 21552–21562. doi: 10.1364/oe.19.021552

Chen, H., Lee, W. J., Huang, H. Y., Chiu, C. M., Tsai, Y. F., Tseng, T. F., et al. (2011b). Performance of THz fiber-scanning near-field microscopy to diagnose breast tumors. *Opt. Express* 19, 19523–19531. doi: 10.1364/oe.19.019523

Chen, H., Ma, S., Wu, X., Yang, W., and Zhao, T. (2015). Diagnose human colonic tissues by terahertz near-field imaging. *J. Biomed. Opt.* 20:036017. doi: 10.1117/1.jbo.20.3.036017

Chen, H. W., Li, Y. T., Pan, C. L., Kuo, J. L., Lu, J. Y., Chen, L. J., et al. (2007). Investigation on spectral loss characteristics of subwavelength terahertz fibers. *Opt. Lett.* 32, 1017–1019. doi: 10.1364/ol.32.001017

Chen, L. J., Chen, H. W., Kao, T. F., Lu, J. Y., and Sun, C. K. (2006). Low-loss subwavelength plastic fiber for terahertz waveguiding. *Opt. Lett.* 31, 308–310. doi: 10.1364/ol.31.000308

Fitzgerald, A. J., Wallace, V. P., Jimenez-Linan, M., Bobrow, L., Pye, R. J., Purushotham, A. D., et al. (2006). Terahertz pulsed imaging of human breast tumors. *Radiology* 239, 533–540. doi: 10.1148/radiol.2392041315

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.

Kindt, J. T., and Schmuttenmaer, C. A. (1996). Far-infrared dielectric properties of polar liquids probed by femtosecond terahertz pulse spectroscopy. *J. Phys. Chem.* 100, 10373–10379. doi: 10.1021/jp960141g

Lu, J. Y., Kuo, C. C., Chiu, C. M., Chenc, H. W., Pan, C. L., and Sun, C. K. (2008). "THz interferometric imaging using subwavelength plastic fiber based THz endoscopes," in *Proceedings of the Optics InfoBase* (San Jose, CA: Institute of Electrical and Electronics Engineers), 2494–2501. doi: 10.1364/oe.16.002494

Nass, S. J., Henderson, I. C., and Lashof, J. C. (2001). *Mammography and Beyond: Developing Techniques for the Early Detection of Breast Cancer*. Washington DC: National Academy Press.

Onuigbo, C. M., Cuffy-Hallam, M. E., Dunsmore, N. A., and Zinreich, E. S. (2001). Mammography reveals a 2-mm intraductal breast carcinoma. *Hosp. Phys.* 37, 61–64.

Pedersen, J. E., and Keiding, S. R. (1992). Thz time-domain spectroscopy of nonpolar liquids. *IEEE J. Quantum Electron.* 28, 2518–2522. doi: 10.1109/3.159558

Pickwell, E., Fitzgerald, A. J., Cole, B. E., Taday, P. F., Pye, R. J., Ha, T., et al. (2005). Simulating the response of terahertz radiation to basal cell carcinoma using ex vivo spectroscopy measurements. *J. Biomed. Opt.* 10, 21–25.

Rahman, A., Rahman, A. K., and Rao, B. (2016). Early detection of skin cancer via terahertz spectral profiling and 3D imaging. *Biosens. Bioelectron.* 82, 64–70. doi: 10.1016/j.bios.2016.03.051

Reid, C. B., Fitzgerald, A., Reese, G., Goldin, R., Tekkis, P., O'Kelly, P. S., et al. (2011). Terahertz pulsed imaging of freshly excised human colonic tissues. *Phys. Med. Biol.* 56, 4333–4353. doi: 10.1088/0031-9155/56/14/008

Wang, C., Gong, J., Xing, Q., Li, Y., Liu, F., Zhao, X., et al. (2010). Application of terahertz time-domain spectroscopy in intracellular metabolite detection. *J. Biophotonics* 3, 641–645. doi: 10.1002/jbio.201000043

Wang, K., and Mittleman, D. M. (2004). Metal wires for terahertz wave guiding. *Nature* 432, 376–379. doi: 10.1038/nature03040

Woodward, R. M., Cole, B. E., Wallace, V. P., Pye, R. J., Arnone, D. D., Linfield, E. H., et al. (2002). Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue. *Phys. Med. Biol.* 47, 3853–3863. doi: 10.1088/0031-9155/47/21/325

Yamada, T., Tominari, Y., Tanaka, S., Mizuno, M., and Fukunaga, K. (2014). Vibration modes at terahertz and infrared frequencies of ionic liquids consisting of an imidazolium cation and a halogen anion. *Materials* 7, 7409–7422. doi: 10.3390/ma7117409

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

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 05 November 2021
doi: 10.3389/fmolb.2021.683032

Analysis and Construction of a Molecular Diagnosis Model of Drug-Resistant Epilepsy Based on Bioinformatics

Tenghui Han[1†], Zhenyu Wu[2†], Jun Zhu[3,4†], Yao Kou[5], Jipeng Li[3] and Yanchun Deng[1]**

[1]Department of Neurology, Xijing Hospital, Airforce Medical University, Xi'an, China, [2]Department of Anatomy, Histology and Embryology and K.K. Leung Brain Research Centre, School of Basic Medicine, Airforce Medical University, Xi'an, China, [3]State Key Laboratory of Cancer Biology, Institute of Digestive Diseases, Xijing Hospital, Airforce Medical University, Xi'an, China, [4]Department of General Surgery, The Southern Theater Air Force Hospital, Guangzhou, China, [5]Basic Medical College, Yan'an University, Yan'an, China

OPEN ACCESS

Edited by:
Feng Liu,
Wuhan University, China

Reviewed by:
Zhaowei Teng,
People's Hospital of Yuxi City, China
Qing Long,
Kunming Medical University, China

***Correspondence:**
Jipeng Li
jipengli1974@aliyun.com
Yanchun Deng
yqnchund@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: 19 March 2021
Accepted: 29 September 2021
Published: 05 November 2021

Citation:
Han T, Wu Z, Zhu J, Kou Y, Li J and Deng Y (2021) Analysis and Construction of a Molecular Diagnosis Model of Drug-Resistant Epilepsy Based on Bioinformatics. Front. Mol. Biosci. 8:683032. doi: 10.3389/fmolb.2021.683032

Background: Epilepsy is a complex chronic disease of the nervous system which influences the health of approximately 70 million patients worldwide. In the past few decades, despite the development of novel antiepileptic drugs, around one-third of patients with epilepsy have developed drug-resistant epilepsy. We performed a bioinformatic analysis to explore the underlying diagnostic markers and mechanisms of drug-resistant epilepsy.

Methods: Weighted correlation network analysis (WGCNA) was applied to genes in epilepsy samples downloaded from the Gene Expression Omnibus database to determine key modules. The least absolute shrinkage and selection operator (LASSO) regression and support vector machine-recursive feature elimination (SVM-RFE) algorithms were used to screen the genes resistant to carbamazepine, phenytoin, and valproate, and sensitivity of the three-class classification SVM model was verified through the receiver operator characteristic (ROC) curve. A protein–protein interaction (PPI) network was utilized to analyze the protein interaction relationship. Finally, ingenuity pathway analysis (IPA) was adopted to conduct disease and function pathway and network analysis.

Results: Through WGCNA, 72 genes stood out from the key modules related to drug resistance and were identified as candidate resistance genes. Intersection analysis of the results of the LASSO and SVM-RFE algorithms selected 11, 4, and 5 drug-resistant genes for carbamazepine, phenytoin, and valproate, respectively. Subsequent union analysis obtained 17 hub resistance genes to construct a three-class classification SVM model. ROC showed that the model could accurately predict patient resistance. Expression of 17 hub resistance genes in healthy subjects and patients was significantly different. The PPI showed that there are six resistance genes (*CD247*, *CTSW*, *IL2RB*, *MATK*, *NKG7*, and *PRF1*) that may play a central role in the resistance of epilepsy patients. Finally, IPA revealed that resistance genes (*PRKCH* and *S1PR5*) were involved in "CREB signaling in Neurons."

Conclusion: We obtained a three-class SVM model that can accurately predict the drug resistance of patients with epilepsy, which provides a new theoretical basis for research and treatment in the field of drug-resistant epilepsy. Moreover, resistance genes *PRKCH* and *S1PR5* may cooperate with other resistance genes to exhibit resistance effects by regulation of the cAMP-response element-binding protein (CREB) signaling pathway.

Keywords: epilepsy, drug-resistant epilepsy, bioinformatics analysis, CREB signaling pathway, resistance gene

1 INTRODUCTION

Epilepsy is a complex chronic neurological disease characterized by the recurrence of unprovoked seizures and has numerous neurobiological, cognitive, and psychosocial consequences (Fisher et al., 2014). It affects the health of over 70 million people worldwide (Thijs et al., 2019). Epilepsy has complex etiologies, diverse clinical symptoms and phenotypes, and high heterogeneity, which interfere with its diagnosis as well as treatment (Rawat et al., 2020). Moreover, approximately a third of patients with epilepsy are refractory to antiepileptic drugs (AEDs) when they are employed singly or even in various combinations (Lerche, 2020). There is thus an urgent need to find new diagnostic markers of refractory epilepsy to ameliorate the current situation of epilepsy diagnosis and treatment.

There are multitypes of AEDs for epilepsy treatment, among which carbamazepine (CBZ), phenytoin (PHT), and valproate (VPA) are the most widely used first-line drugs (Schmidt and Schachter, 2014). CBZ is a first-line treatment for partial and generalized convulsive seizures, trigeminal pain, and bipolar disorder, which functions as a Na^+ channel blocker (Harper and Topol, 2012). CBZ remains the most efficacious drug for focal and generalized seizures with focal onset (Baulac et al., 2012; Baulac et al., 2017). PHT is also speculated to work as a Na^+ channel blocker; it exhibits similar efficacy to CBZ and is the first-line drug for focal seizures and generalized seizures with focal onset. Unusually, PHT is mainly administered intravenously (Mattson et al., 1985). As the first-line and most effective intravenous drug for focal and generalized seizures in current clinical treatment, VPA performs multiple functions, including GABA potentiation, glutamate inhibition, and sodium channel and T-type calcium channel blockade (Tomson et al., 2016).

In 2009, the International League Against Epilepsy (ILAE) defined drug-resistant epilepsy as "failure of adequate trials of two tolerated, appropriately chosen and used AED schedules" (Kwan et al., 2010). Patients with drug-resistant epilepsy have a significantly increased risk of psychiatric and somatic comorbidities and adverse effects from AEDs. Furthermore, their seizures are not well controlled and recurrent, especially generally tonic–clonic seizures, which is the best-recognized risk factor for sudden unexplained death in epilepsy (Ryvlin et al., 2019). Recent research has demonstrated that after the failure of two well-tolerated AED schedules appropriately chosen for the seizure types, patients under long-term treatment for epilepsy have a progressively less likely chance of success with further drug treatment (Chen et al., 2018). Therefore, early-stage identification of AED resistance is crucial to patient treatment outcomes.

In our study, we used weighted correlation network analysis (WGCNA), the least absolute shrinkage and selection operator (LASSO) algorithm, and the support vector machine-recursive feature elimination (SVM-RFE) algorithm to analyze and select resistance genes. All genes in epilepsy patient samples were downloaded from the Gene Expression Omnibus (GEO) database. We constructed a novel three-class classification SVM model to accurately predict patient resistance, which may provide a new strategy for the treatment and research of drug-resistant epilepsy and also revealed that the resistance genes *PRKCH* and *S1PR5* may cooperate with other resistance genes through regulation of the cAMP-response element-binding protein (CREB) signaling pathway. The workflow is shown in **Figure 1**.

2 MATERIALS AND METHODS

2.1 Data Source

The original dataset of the whole gene expression profiles was downloaded from the GEO database. The accession number was GSE143272, which was based on GPL10558 (Illumina HumanHT-12 V4.0 expression beadchip). Gene sequences of a total of 34 drug-naïve patients with epilepsy and 57 followed-up patients showing differential response to AED monotherapy, along with 50 healthy subjects as a control group, were included in the study. The AED-treatment group included the CBZ-drug-treatment group (tolerance: 9; intolerance: 10), the PHT-drug-treatment group (tolerance: 6; intolerance: 7), and the VPA-drug-treatment group (tolerance: 9; intolerance: 16).

2.2 Definitions of Candidate Resistance Genes by WGCNA

In this study, we used the "WGCNA" R software package to construct modules related to clinical features in the epilepsy sample dataset (GSE143272) and identify candidate genes (Langfelder and Horvath, 2008). The clinical features were divided into eight categories: normal (health), unmedicated epilepsy (case), CBZ tolerance, CBZ intolerance, PHT tolerance, PHT intolerance, VPA tolerance, and VPA intolerance. The overall clustering of the GSE143272 dataset was found to be of relatively high quality, so no sample removal processing was performed (**Supplementary Figure S1A**). The traits of the samples are shown in **Supplementary**

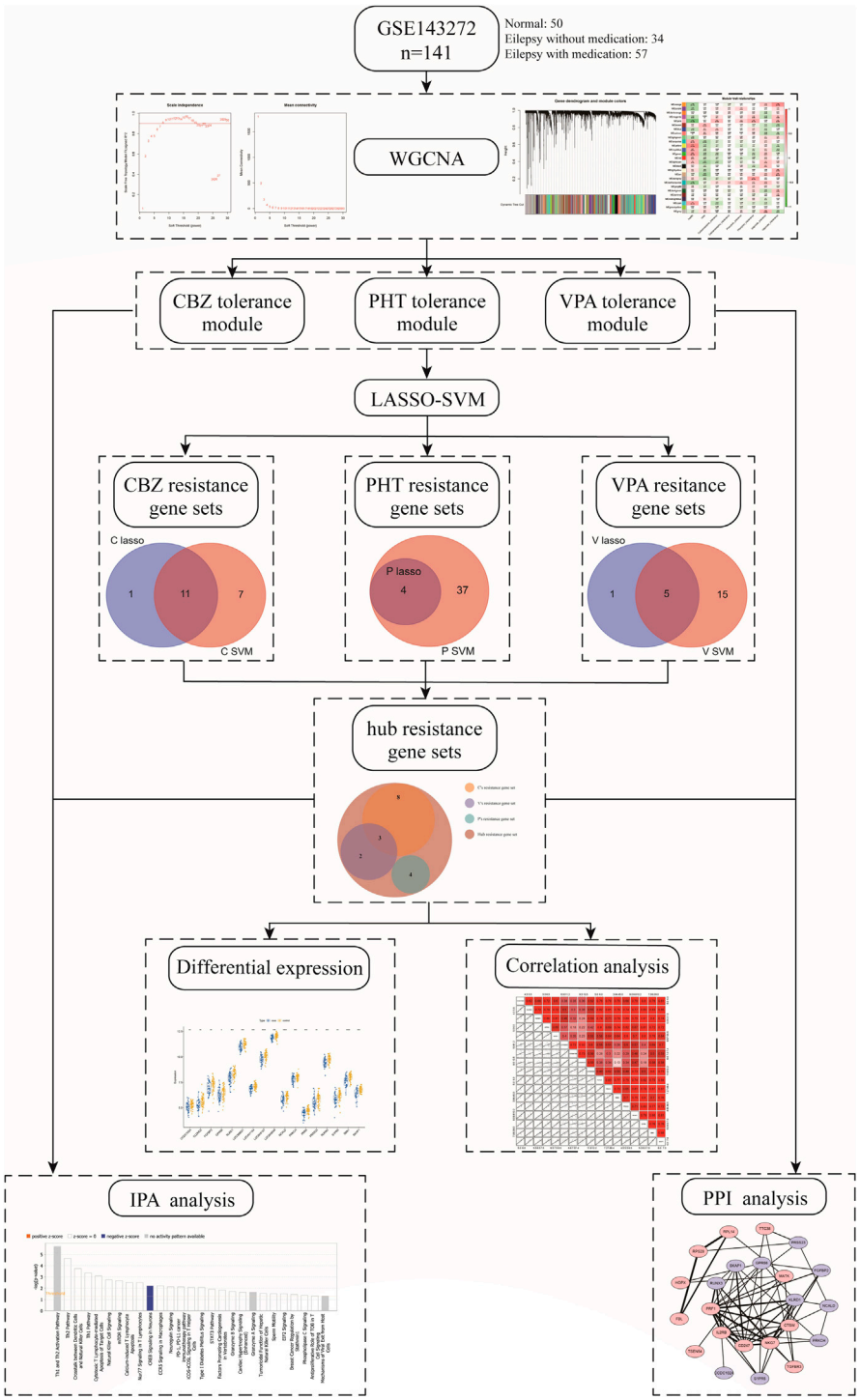


FIGURE 1 | Workflow of the study.

Figure S1B. The adjacency matrix was converted to a topological overlap matrix (TOM) (Li et al., 2019). According to the degree of TOM similarity, genes were divided into multiple gene modules (**Supplementary Figures S1C,D**). In this analysis, the soft threshold was set to 7 (scale-free R^2 = 0.85), and the minimum module size was 30. The correlations between the characteristic gene of each module and clinical characteristics were calculated. The screening of key modules was achieved by calculating the correlation between the module genes and clinical features. Moreover, a gene with |gene significance (GS)| >0.2 and |

module membership (MM)| > 0.8 in the key modules was considered as a candidate resistance gene.

2.3 Feature Selections by LASSO and SVM-RFE Algorithms

LASSO logistic regression and SVM-RFE were performed on the candidate resistance genes obtained in WGCNA to screen characteristic genes. LASSO is a regression analysis algorithm that uses regularization to improve the prediction accuracy. The penalty parameter (λ) of the LASSO regression model was determined by following a 10-fold cross-validation of the minimum criterion (i.e., the value of λ corresponding to the lowest partial likelihood deviation). The LASSO regression algorithm using the "glmnet" package (Friedman et al., 2010) in R was performed to identify genes significantly associated with the distinctions between CBZ-resistant and PHT + VPA-resistant samples, PHT-resistant and CBZ + VPA-resistant samples, and VPA-resistant and CBZ + PHT-resistant samples. Furthermore, SVM-RFE is an effective feature selection technique that finds the best variables by deleting the feature vector generated by SVM (Wang and Liu, 2015). In this study, the SVM-RFE algorithm screened the best variables based on a minimum 10 × CV error value. The performances of CBZ/PHT/VPA resistance LASSO and SVM models are shown in **Supplementary Table S1**. For each drug, resistance genes were defined as the common genes identified by the LASSO and SVM-RFE algorithms. Ultimately, we combined the resistance genes of CBZ, PHT, and VPA as hub resistance genes for further analysis. A three-class classification SVM module was established using the "e1071" software package in R (**Supplementary Figure S2**) (Cinelli et al., 2017), and the receiver operating characteristic (ROC) curve was used to further determine the diagnostic value of the hub resistance genes in epilepsy.

2.4 Construction of the Protein–Protein Interaction Network

To interpret the molecular mechanisms of hub resistance genes in epilepsy, the online tool, the Search Tool for the Retrieval of Interacting Genes (STRING) database, was used to construct the protein–protein interaction (PPI) network of 72 modular genes (Szklarczyk et al., 2015). The PPI was visualized with a confidence score >0.15 (Assenov et al., 2008).

2.5 Ingenuity Pathway Analysis for the Identification of Diseases and Function Pathways Involved

Ingenuity pathway analysis (IPA) is a web-based bioinformatic application for functional analysis, aggregation, and further understanding of data analysis results (Khan et al., 2016). Briefly, IPA was performed to identify diseases and functions and gene networks that were most significant to hub resistance genes. The Z-scores of significantly involved diseases and function pathways were also determined.

2.6 Statistical Analysis

All statistical analyses were performed using R version 3.4.1. The Wilcox test was used to analyze the relationship between drug resistance and clinicopathological characteristics. Pearson correlation analysis was adopted to understand the relevance of the 17 hub resistance genes. The area under the curve (AUC) was calculated to evaluate the property of the models. $p < 0.05$ was envisaged to indicate a statistically significant difference.

3 RESULTS

3.1 Determination of the Most Relevant Module Genes for Drug Tolerance in Epilepsy Treatment

We first clustered all the samples in the GSE143272 dataset to ensure the accuracy of the analysis (**Supplementary Figure S1A**). The coexpression network was constructed through coexpression analysis. A total of 27 modules (including gray modules) were identified via the average linkage hierarchical clustering. To ensure that the interaction between genes in the coexpression network could conform to the scale-free distribution to the greatest extent, the power of $\beta = 7$ was selected; to merge the highly similar modules, we chose a cutoff <0.25 and a minimum module size of 30 using the dynamic hybrid tree cut method. In this study, we focused on the drug-resistant traits of disease samples. Therefore, we included the two traits of the case and drug tolerance as reference factors to screen key modules. It was found that the MElightcyan module had the highest correlation with CBZ-tolerance traits (module-trait relationships = −0.27 and −0.12, respectively) and VPA-tolerance traits (module-trait relationships = −0.27 and 0.2, respectively) of cases. The MEyellow module (module-trait relationships = −0.19 and −0.12, respectively) was found to have the highest association with the PHT-tolerance status of the case (**Figure 2**). Hence, 1,016 genes in the two modules (MElightcyan: 206 and MEyellow: 810) were considered to be significant module genes for further intramodular analysis. Based on the candidate gene screening criteria in the key module (|GS| > 0.2 and |MM| > 0.8), a total of 72 candidate genes from the MElightcyan (25 genes) and MEyellow (47 genes) modules were chosen for further analysis (**Figures 2E,F**; **Supplementary Tables S2, S3**).

3.2 Identification of Hub Resistance Genes in Patients With Epilepsy

In this study, two distinct algorithms, LASSO and SVM-RFE, were utilized for screening potential resistance genes against CBZ, PHT, and VPA. For each drug, resistance genes were defined by the common signature genes identified by LASSO and SVM-RFE. Ultimately, the resistance genes of all three drugs were collectively termed as hub resistance genes in our research.

For the identification of potential resistance genes to CBZ, we built classifiers capable of distinguishing between CBZ-resistant

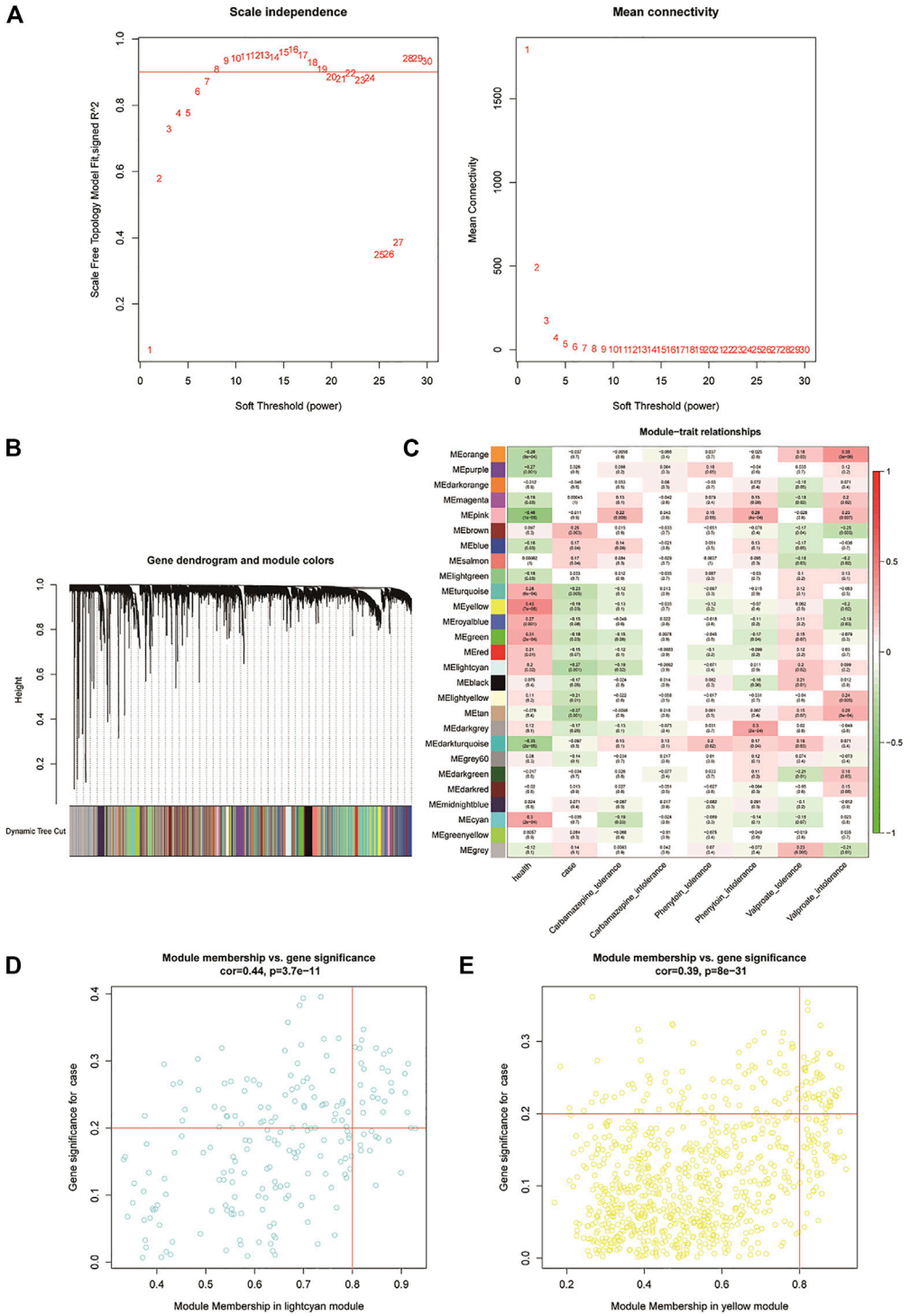


FIGURE 2 | Weighted gene co-expression network analysis of the potential resistance genes. **(A)** Weighted value β of scale-free networks. The relationship between the soft threshold and scale-free R2 is exhibited on the left. On the right, the relationship between the soft threshold and mean connectivity is shown. **(B)** Cluster dendrogram. Each branch in the figure represents the genes, which are divided into module colors based on the cluster analysis results. The oligogenics are assigned in the gray module. **(C)** Heatmap of the correlation analysis between modules and clinical characteristics. The vertical axis represents the different modules; the horizontal axis represents the different traits. The number in each cell represents the correlation coefficient and significance (*p*-value) between a module and a trait. **(D,E)** Scatter diagrams of MElightcyan and MEyellow modules. Using the criteria |GS| > 0.2 and |MM| > 0.8, we selected the key genes of each module in the upper right corner of the figure. Twenty-five key genes were screened from the MElightcyan module, a resistance module common to both CBZ and VPA drugs.

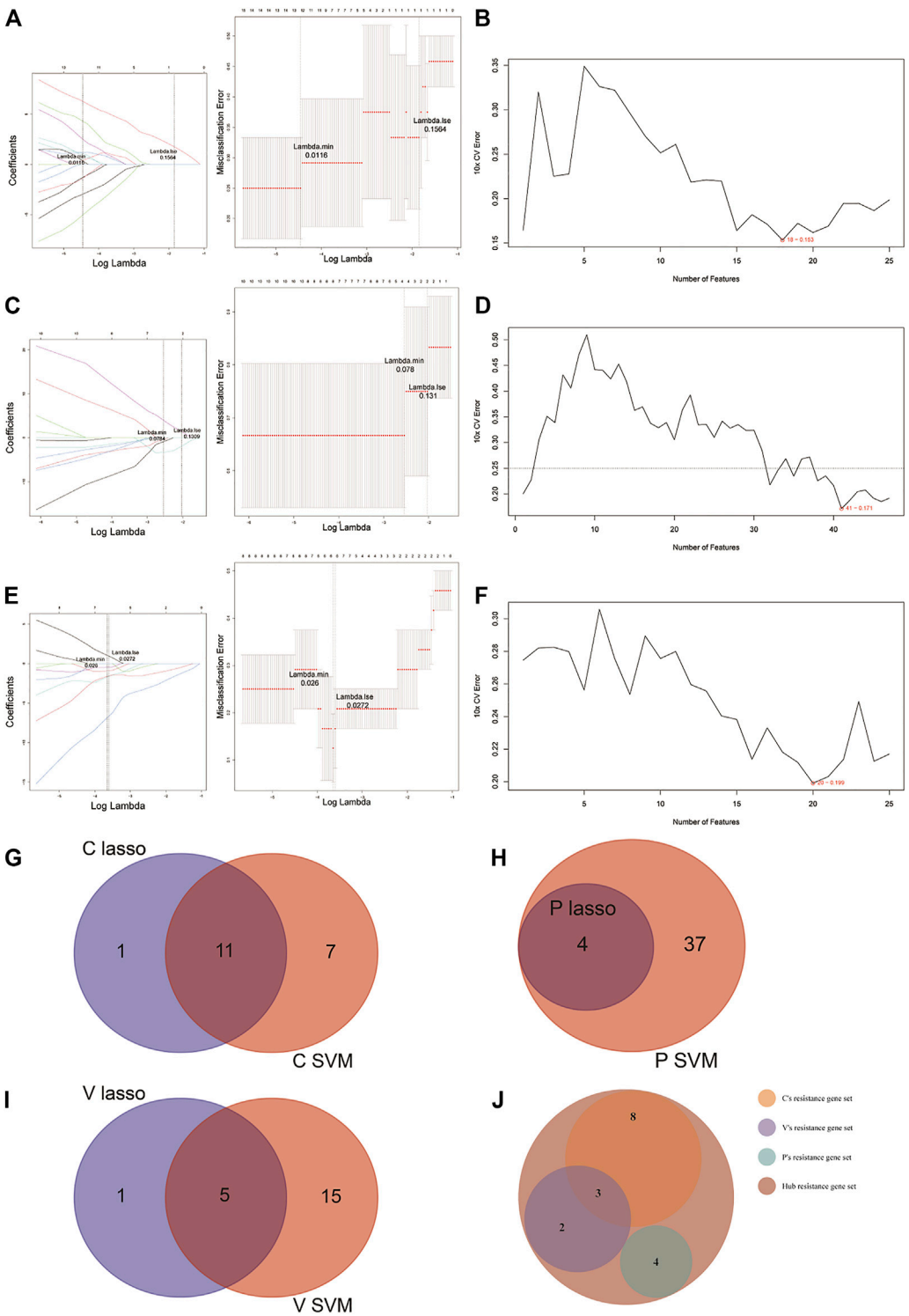
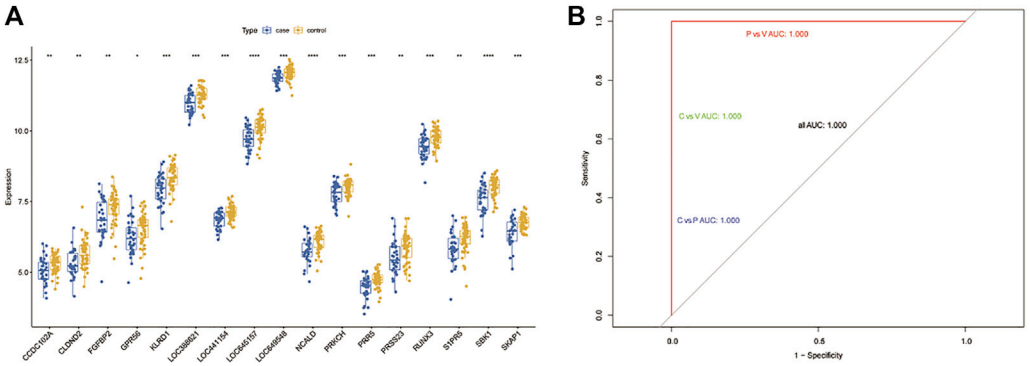
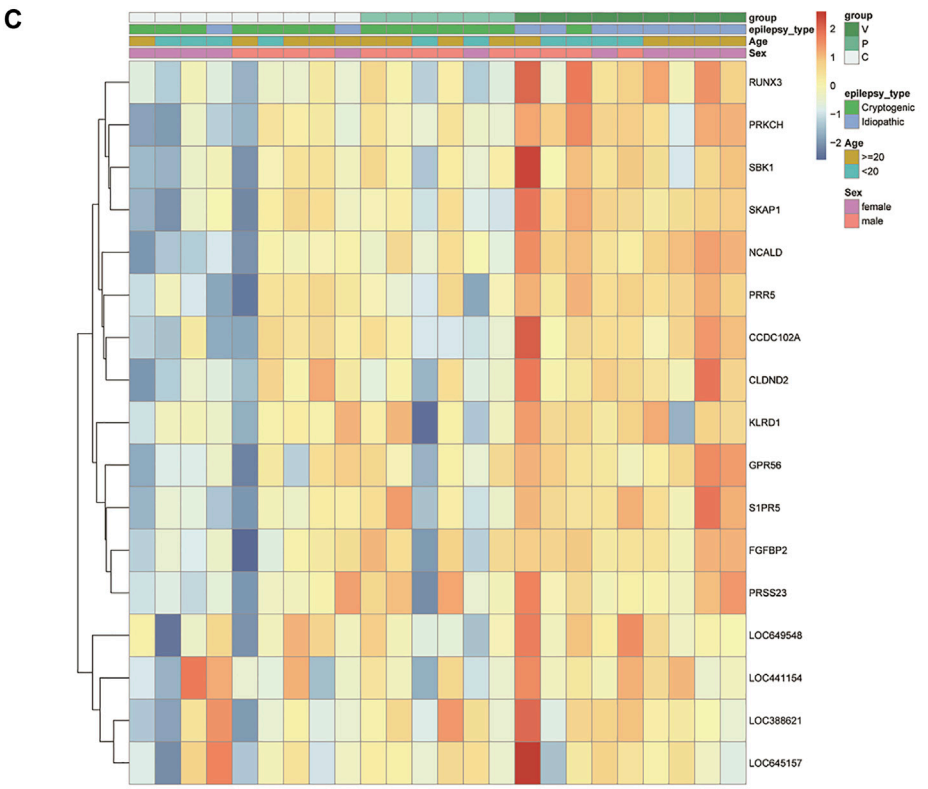


FIGURE 3 | LASSO and SVM-RFE algorithms were used for characteristic gene selection. **(A, C, E)** LASSO algorithm. Using the LASSO algorithm, we identified 12 potential resistance genes in the CBZ-resistance gene set, 4 in the PHT-resistance gene set, and 6 in the VPA-resistance gene set. **(B, D, F)** SVM-RFE algorithm. SVM-RFE algorithm separately indicated the resistance genes most closely corresponding with the lowest error rates in patients treated with CBZ **(B)**, PHT **(D)**, and VPA **(F)**. **(G–I)** Venn diagram of the characteristic genes for CBZ **(G)**, PHT **(H)**, and VPA **(I)**, which were selected from the LASSO or SVM-RFE algorithms. **(J)** We unified the LASSO + SVM characteristic resistance genes of CBZ, PHT, and VPA and obtained 17 characteristic genes.





D

	Total (N=24)	Expression			P-value
		Carbamazepine_tolerance (N=9)	Phenytoin_tolerance (N=6)	Valproate_tolerance (N=9)	
Gender					
female	11 (45.8%)	5 (55.6%)	1 (16.7%)	5 (55.6%)	0.254
male	13 (54.2%)	4 (44.4%)	5 (83.3%)	4 (44.4%)	
Age (years)					
>=20	14 (58.3%)	5 (55.6%)	4 (66.7%)	5 (55.6%)	0.892
<20	10 (41.7%)	4 (44.4%)	2 (33.3%)	4 (44.4%)	
epilepsy_type					
Cryptogenic	14 (58.3%)	7 (77.8%)	6 (100%)	1 (11.1%)	<0.001
Idiopathic	10 (41.7%)	2 (22.2%)	0 (0%)	8 (88.9%)	

FIGURE 4 | Assessment of the predictive value of the three-class classification SVM model. **(A)** Boxplot shows the expression patterns of 17 drug resistance genes in case and control samples from the GSE143272 dataset. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$. **(B)** ROC curve based on every two drugs in the model. Blue represents CBZ vs. PHT, green represents CBZ vs. VPA, and red represents PHT vs. VPA. Since all AUCs are 1.000, only one color is shown in the figure (other ROC curves are covered). **(C,D)** By using the Wilcox test, we analyzed the correlation between three clinical traits (age, gender, and pathological classification) and drug resistance. A heatmap of resistance genes and clinical traits is plotted **(C)**. Clinical traits and drug resistance were significantly correlated ($p < 0.05$) **(D)**.

samples ($n = 9$) and PHT + VPA-resistant samples ($n = 15$) using the LASSO and SVM-RFE algorithms. Specifically, the LASSO regression was performed to remove candidate genes that were related to each other to prevent overfitting of the model (**Figure 3A**). A total of 12 LASSO signature genes were obtained at λ min = 0.0116; they were *CCDC102A*, *CEP78, CLDND2*, *FGFBP2*, *GPR56*, *KLRD1*, *NCALD*, *PRKCH*, *RUNX3*, *S1PR5*, *SBK1*, and *SKAP1*. Meanwhile, based on the SVM-RFE algorithm (**Figure 3B**), 18 SVM-RFE signature genes were identified at a minimum 10-fold CV error (0.153), namely, *NCALD*, *FGFBP2*, *CCDC102A*, *KLRD1*, *S1PR5*, *SKAP1*, *TTC38*, *CLDND2*, *PRKCH*, *SBK1*, *CD247*, *RUNX3*, *ENPP4*, *TSEN54*, *NKG7*, *PRR5*, *GPR56*, and *HOPX*. Subsequently, a total of 11 genes (*CCDC102A*, *CLDND2*, *FGFBP2*, *GPR56*, *KLRD1*, *NCALD*, *PRKCH*, *RUNX3*, *S1PR5*, *SBK1*, and *SKAP1*) were identified by overlap analysis as common to both the LASSO signature gene set and the SVM-RFE signature gene set; these genes were defined as resistance genes for CBZ (**Figure 3G**).

Before identifying potential resistance genes to PHT, we divided all drug-resistant samples into PHT-resistant ($n = 6$) and CBZ + VPA-resistant ($n = 18$) groups. The 72 candidate genes previously identified were narrowed down using the LASSO regression algorithm, resulting in the identification of four variables (*LOC388621*, *LOC441154*, *LOC645157*, and *LOC649548*) as potential resistance genes for PHT at λ min = 0.0784 (**Figure 3C**). Based on the best point (10 × CV error = 0.171), the SVM-RFE algorithm obtained 41 eigenvalues (**Figure 3D**; **Supplementary Table S4**). By overlapping the genes from the two algorithms, we identified the four genes (*LOC388621*, *LOC441154*, *LOC645157*, and *LOC649548*) as resistance genes in patients treated with PHT (**Figure 3H**).

Based on 9 VPA-resistant samples and 15 CBZ + PHT-resistant samples, the LASSO regression algorithm identified *IL2RB*, *NCALD, PRKCH*, *PRR5*, *PRSS23*, and *RUNX3* as potential resistance genes to VPA based on λ min = 0.0272 from 72 candidate genes (**Figure 3E**). A subset of 16 features among the candidate genes was determined using the SVM-RFE algorithm (10 × CV error = 0.199; **Figure 3F**). The five overlapping features (*NCALD*, *PRKCH*, *PRR5*, *PRSS23*, and *RUNX3*) between these two algorithms were ultimately selected as the resistance genes in patients treated with VPA (**Figure 3I**).

Collectively, we obtained a total of 11 CBZ-resistant genes, 4 PHT-resistant genes, and 5 VPA-resistant genes (**Supplementary Table S5**). Overlap analysis revealed that *NCALD*, *RUNX3*, and *PRKCH* were the common resistance genes for CBZ and VPA (**Figure 3J**). Thus, a total of 17 hub resistance genes were obtained and included for further analysis.

3.3 Evaluation of the Three-Class Classification SVM Model

The 17 resistance genes were significantly different in control and case samples; i.e., compared with the control group, their expression in case samples was generally lower (**Figure 4A**). Then, the library ("e1071") package was used in the R software to construct a three-class classification SVM model for the 17 hub resistance genes obtained from the above analysis, and its prediction performance was evaluated in the GSE143272 dataset. The ROC curve was drawn based on the true and predicted values of each two drugs in the model. The results demonstrated that the three-class classification SVM model could distinguish the patient's tolerance to the three drugs (all AUC = 1.000), indicating that the resistance genes may be clinically useful (**Figure 4B**). We then compared the clinical characteristics of the three subgroups, namely, CBZ tolerance, PHT tolerance, and VPA tolerance. Subgroup analysis of clinical characteristics showed that the cryptogenic epilepsy type was characterized by significant differences (**Figures 4C,D**). Other clinical characteristics like gender, age, and idiopathic epilepsy type had no statistical significance.

3.4 Correlation Analysis of Resistance Genes

Pearson analysis was used to explore the correlation between 17 resistance genes. Studies have shown that all resistance genes have a strong positive correlation; as shown in **Figure 5A**, *SKAP1* has the highest correlation with *SBK1* and *NCALD* ($r = 0.88$). The relationship between some other resistance genes does not seem to be as close. For example, the correlation between *LOC645157* and *PRR5*/*S1PR5* ($r = 0.13$ and $r = 0.18$, respectively) and the correlation between *GPR56* and *LOC441154* were not considerable ($r = 0.18$).

Next, we used the STRING online tool to construct a PPI network for 72 modular genes. This was to show the maximum possible additional modular genes that interacted with the 17 resistance genes. We set the confidence level to 0.15. After removing discrete proteins, we obtained a PPI network with 23 proteins. The PPI network is illustrated in **Figure 5B**. The results showed that 6 of the 17 resistance genes were at the center of the network, indicating that they were associated with a higher number of genes. Therefore, we speculated that these genes played a major role in the corresponding drug-tolerance modules. Judging from the analysis of the degree of binding (combined score), we found that *CD247-IL3RB-PRF1-NKG7*/*KLRD1*/*CD274* may form a complete closed loop of tolerance and promote the patient's body to develop resistance. Also, although *RPL14*, *RPS28*, and *FBL* were out of the core of the PPI, these three resistance genes could form a complete closed chain of action and exert a powerful resistance effect (**Supplementary Table S6**). Regardless of the fact that only 10 resistance genes were displayed in the network, the remaining 7 resistance genes seem to have a unique relationship network that was not yet known to play their corresponding roles.

3.5 IPA of the Hub Resistance Genes

The complete list of enriched disease and function pathway analysis is included in **Supplementary Table S7**. A total of 27 enriched disease and function pathways were identified by applying the –log (p-value) > 1.3 threshold. All the 27 representative pathways that were found to associate tightly

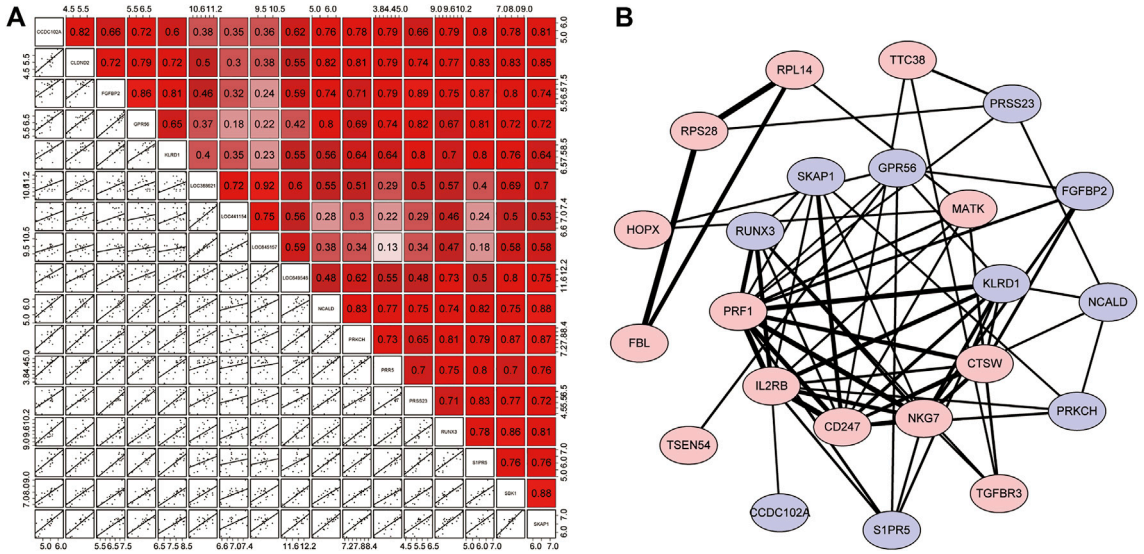


FIGURE 5 | Correlation analysis of resistance genes and the PPI network. **(A)** Correlation matrix of 17 resistance genes. The upper right half is the correlation score, where red represents positive correlation and purple indicates negative correlation. The lower left section shows scattered plots of correlations between each of the two genes. **(B)** PPI network. The red proteins are resistance genes, while the blue proteins are the remaining module genes. Based on the combined score, the lines represented the interaction between them, and the degree of thickness represented the degree of their combination.

with the tolerance module genes and resistance genes are shown in **Figure 6A**, ranked according to their −log (*p*-value). The "Th1 and Th2 activation pathway" was the highest-ranking signaling pathway with a −log (*p*-value) of 5.71. Although none of the detected signaling pathways had a Z-score > 2 (significant activation), one of the enriched signaling pathways, "CREB signaling in neurons," had a Z-score = −2. Of note, the involvement of CREB in the occurrence and development of epilepsy is well recognized (Sharma et al., 2019). These results suggest that these resistance genes (*PRKCH* and *S1PR5*) may induce resistance in patients with drug-treated epilepsy by regulation of the CREB pathway. Moreover, **Figure 6B** shows the interaction network between 72 modular genes. Among them, we found that *RUNX3* could directly interact with *S1PR5* and *PPR5* by acting on *Akt*. However, *CCD102A*, *FGFBP2*, *NCALD*, *PRSS23*, and *SRSS23* were intertwined into an intricate network through their direct or indirect interaction with beta-estradiol.

4 DISCUSSION

Epilepsy is one of the most common chronic diseases of the nervous system and extensively affects people of all ages, genders, and races worldwide (Fiest et al., 2017; Devinsky et al., 2018). Pharmacological treatment is widely recognized as the mainstay of the therapy approach for people with epilepsy. However, previous studies have indicated that more than one-third of the patients are likely to develop refractory epilepsy in the process of AED treatment (Kwan and Brodie, 2000; Löscher et al., 2020). The complex resistance mechanisms of AEDs are still not entirely clear. Recent studies have demonstrated that the application of bioinformatic analysis could provide a chance to explore the underlying mechanisms of drug resistance (Zhang et al., 2019; Zhu et al., 2019). Therefore, we utilized bioinformatic analysis techniques to construct a three-class SVM model to precisely predict the drug resistance of patients with epilepsy and explored the potential mechanisms of drug-resistant epilepsy.

In this study, we included 50 healthy patients, 34 patients with epilepsy untreated by medication, and 57 patients with epilepsy with three different AED treatments (CBZ, PHT, or VPA) from the GEO database (GSE143272 dataset). Then, 72 candidate resistance genes were identified by WGCNA. IPA revealed a total of 27 disease and functional associations of candidate resistance genes. The highest-ranked signaling pathway was the Th1 and Th2 activation pathway, indicating that candidate resistance genes were potentially involved in the regulation of immune response in patients. Subsequently, by employing the LASSO + SVM-RFE algorithm, we constructed a three-class classification SVM model based on 17 hub resistance genes (*CCDC102A*, *CLDND2*, *FGFBP2*, *GPR56*, *KLRD1*, *NCALD*, *PRKCH*, *PRR5*, *PRSS23*, *RUNX3*, *S1PR5*, *SBK1*, *SKAP1*, *LOC388621*, *LOC441154*, *LOC645157*, and *LOC649548*) from CBZ-resistant, PHT-resistant, and VPA-resistant gene sets. The model possessed a strong ability to predict drug tolerance in patients (AUC = 1.000). Furthermore, these genes displayed a significant Pearson correlation with each other. The PPI network analysis revealed that *CD247*, *CTSW*, *IL2RB*, *MATK*, *NKG7*, and *PRF1* were at the center of the network and may play essential roles in the development of drug resistance.

Our study screened 17 novel resistance genes and built a highly effective model to accurately predict the drug resistance of patients with epilepsy. Among the 17 hub resistance genes, we found that *NCALD* and *GPR56* were verified to be directly relevant to epilepsy in previous studies. Recent studies have

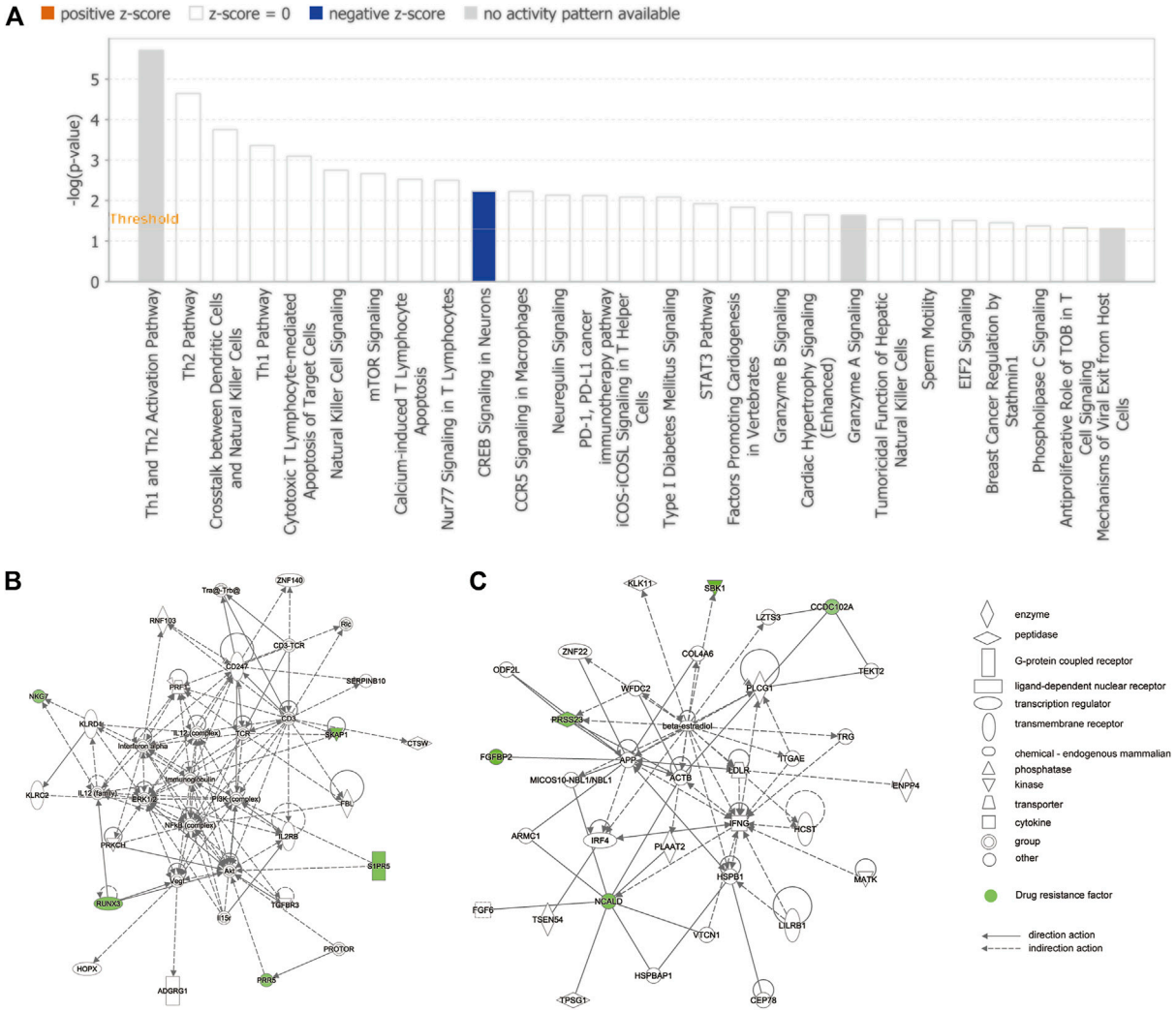


FIGURE 6 | IPA of module genes. **(A)** Diseases and functional pathway analysis of module genes. While the Y-axis is the pathway terms, the X-axis is the log (*p*-value). **(B)** Interaction network was constructed between modules of genes and the chemical/drug and other substances by using IPA. Green represents resistance genes, including *KLRD1*, *PRR5*, *RUNX3*, *S1PR5*, *SKAP1*, *CCDC102A*, *FGFBP2*, *NCALD*, *PRSS23*, and *SBK1*. Solid lines of the arrows indicate direct interactions between genes, while dotted lines indicate indirect interactions.

reported that intellectual disability and epilepsy were detected in patients with *NCALD* deletion, indicating that *NCALD* could be a crucial gene in epilepsy (Kuechler et al., 2011; Kuroda et al., 2014). Additionally, studies have demonstrated that *GPR56* mutations may cause malformations of cortical development, which could further result in epileptogenesis (Guerrini and Dobyns, 2014; Kuzniecky, 2015). However, the underlying mechanisms of *NCALD* and *GPR56* in AED resistance have not yet been reported and left a wide scope for further research.

Considering the above result of the IPA, we found that the CREB signaling pathway in neurons appeared to be closely associated with the tolerance module and resistance genes. Recent research has demonstrated that the CREB signaling pathway plays an essential role in mossy fiber sprouting, which is generally known to be a pathological result of recurrent epilepsy. CREB upregulation boosts the transcription of its target genes, which results in the enhancement of mossy fiber sprouting and an increase in the number of dysfunctional synapses in neural circuits, resulting in poor AED treatment outcomes for patients with epilepsy and ultimately developing into refractory epilepsy (Redmond et al., 2002; Finsterwald et al., 2010). Additionally, according to our results, two hub resistance genes (*PRKCH* and *S1PR5*) were closely involved in the CREB pathway, which is consistent with previous research. *PRKCH* encodes a protein kinase subtype, which is widely involved in brain functions (Boehm et al., 2006; Schwenk et al., 2013). Through pathway analysis on the identified single-nucleotide polymorphism component, researchers have found that *PRKCH* is strongly associated with the CREB signaling pathway (Chen et al., 2015). *S1PR5* encodes a G-protein-coupled receptor which is reported to be highly relevant to CREB activation (Rivera et al., 2008; Wang et al., 2020). Moreover, *PRKCH* was proved to be the joint gene

among CBZ-resistant and VPA-resistant gene sets in our findings. Integrating this evidence, we speculate that *PRKCH* and *S1PR5* may induce resistance in patients with drug-treated epilepsy through the CREB pathway.

Intriguingly, emerging evidence has demonstrated that *PRKCH* and *PPR5* are associated with the mTOR signaling pathway. The mTOR pathway regulates a variety of neuronal functions, including cell proliferation, survival, growth, metabolism, and plasticity. Compelling evidence has indicated that abnormal activity of the mTOR pathway plays an irreplaceable role in epileptogenesis (Lim et al., 2015; Curatolo et al., 2018). Moreover, recent studies have further confirmed the substantial therapeutic potential of targeting the mTOR signaling pathway in drug-resistant epilepsy (Hodges and Lugo, 2020). This implies that *PRKCH* and *PPR5* could be potential targets for the treatment of refractory epilepsy.

Additionally, other than the 5 hub genes mentioned above, we also identified 12 novel drug resistance genes, herein first reported to be related to refractory epilepsy. According to the correlation analysis, all 17 resistance genes have a strong positive relation, and *SKAP1* has the highest correlation, with *SBK1* and *NCALD*. Moreover, among the 12 novel resistance genes, *CCDC102A*, *FGFBP2*, *RUNX3*, *SKAP1*, *KLRD1*, and *PRSS23* were intertwined into a complex PPI network. *LOC388621*, *LOC441154*, *LOC645157*, and *LOC649548* were first screened out to be PHT-resistant genes, and their structure and function deserve to be further studied. Integrating the results above, we inferred that the 17 hub genes have intricate direct or indirect interactions in drug-resistant epilepsy.

Nevertheless, there were several limitations in this study. First, our research is based on a publicly available dataset. Prospective real-world data should be incorporated to validate the clinical utility of our model. Subsequently, further *in vitro* and *in vivo* experiments should be performed to confirm the mechanisms of the 17 hub genes in drug-resistant epilepsy.

5 CONCLUSION

Through this study, we have offered novel insights into the research and treatment of drug-resistant epilepsy and created a novel three-class SVM model with high prediction values. This is also the first study that has elucidated that the resistance genes *PRKCH* and *S1PR5* may work in coordination with other resistance genes to exhibit their resistance effects through regulation of the CREB signaling pathway.

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

YD designed the study; TH, ZW, and JZ contributed to the conception of the study and completed the manuscript together; YK contributed significantly to statistical analysis and manuscript preparation; JL helped perform the analysis with constructive discussions. All authors read and approved the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at https://www.frontiersin.org/articles/10.3389/fmolb.2021.683032/full#supplementary-material

REFERENCES

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., and Albrecht, M. (2008). Computing Topological Parameters of Biological Networks. *Bioinformatics* 24 (2), 282–284. doi:10.1093/bioinformatics/btm554

Baulac, M., Brodie, M. J., Patten, A., Segieth, J., and Giorgi, L. (2012). Efficacy and Tolerability of Zonisamide versus Controlled-Release Carbamazepine for Newly Diagnosed Partial Epilepsy: a Phase 3, Randomised, Double-Blind, Non-inferiority Trial. *Lancet Neurol.* 11 (7), 579–588. doi:10.1016/s1474-4422(12)70105-9

Baulac, M., Rosenow, F., Toledo, M., Terada, K., Li, T., De Backer, M., et al. (2017). Efficacy, Safety, and Tolerability of Lacosamide Monotherapy versus Controlled-Release Carbamazepine in Patients with Newly Diagnosed Epilepsy: a Phase 3, Randomised, Double-Blind, Non-inferiority Trial. *Lancet Neurol.* 16 (1), 43–54. doi:10.1016/s1474-4422(16)30292-7

Boehm, J., Kang, M.-G., Johnson, R. C., Esteban, J., Huganir, R. L., and Malinow, R. (2006). Synaptic Incorporation of AMPA Receptors during LTP Is Controlled by a PKC Phosphorylation Site on GluR1. *Neuron* 51 (2), 213–225. doi:10.1016/j.neuron.2006.06.013

Chen, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., van Hulzen, K., Fernández, G., et al. (2015). G-protein Genomic Association with normal Variation in gray Matter Density. *Hum. Brain Mapp.* 36 (11), 4272–4286. doi:10.1002/hbm.22916

Chen, Z., Brodie, M. J., Liew, D., and Kwan, P. (2018). Treatment Outcomes in Patients with Newly Diagnosed Epilepsy Treated with Established and New Antiepileptic Drugs. *JAMA Neurol.* 75 (3), 279–286. doi:10.1001/jamaneurol.2017.3949

Cinelli, M., Sun, Y., Best, K., Heather, J. M., Reich-Zeliger, S., Shifrut, E., et al. (2017). Feature Selection Using a One Dimensional Naïve Bayes' Classifier Increases the Accuracy of Support Vector Machine Classification of CDR3 Repertoires. *Bioinformatics* 33 (7), btw771–955. doi:10.1093/bioinformatics/btw771

Curatolo, P., Moavero, R., van Scheppingen, J., and Aronica, E. (2018). mTOR Dysregulation and Tuberous Sclerosis-Related Epilepsy. *Expert Rev. Neurotherapeutics* 18 (3), 185–201. doi:10.1080/14737175.2018.1428562

Devinsky, O., Vezzani, A., O'Brien, T. J., Jette, N., Scheffer, I. E., de Curtis, M., et al. (2018). Epilepsy. *Nat. Rev. Dis. Primers* 4, 18024. doi:10.1038/nrdp.2018.24

Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C.-S., Dykeman, J., et al. (2017). Prevalence and Incidence of Epilepsy. *Neurology* 88 (3), 296–303. doi:10.1212/wnl.0000000000003509

Finsterwald, C., Fiumelli, H., Cardinaux, J.-R., and Martin, J.-L. (2010). Regulation of Dendritic Development by BDNF Requires Activation of CRTC1 by Glutamate. *J. Biol. Chem.* 285 (37), 28587–28595. doi:10.1074/jbc.M110.125740

Fisher, R. S., Acevedo, C., Arzimanoglou, A., Bogacz, A., Cross, J. H., Elger, C. E., et al. (2014). ILAE Official Report: a Practical Clinical Definition of Epilepsy. *Epilepsia* 55 (4), 475–482. doi:10.1111/epi.12550

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33 (1), 1–22. doi:10.18637/jss.v033.i01

Guerrini, R., and Dobyns, W. B. (2014). Malformations of Cortical Development: Clinical Features and Genetic Causes. *Lancet Neurol.* 13 (7), 710–726. doi:10.1016/s1474-4422(14)70040-7

Harper, A. R., and Topol, E. J. (2012). Pharmacogenomics in Clinical Practice and Drug Development. *Nat. Biotechnol.* 30 (11), 1117–1124. doi:10.1038/nbt.2424

Hodges, S. L., and Lugo, J. N. (2020). Therapeutic Role of Targeting mTOR Signaling and Neuroinflammation in Epilepsy. *Epilepsy Res.* 161, 106282. doi:10.1016/j.eplepsyres.2020.106282

Khan, M. I., Dębski, K. J., Dabrowski, M., Czarnecka, A. M., and Szczylik, C. (2016). Gene Set Enrichment Analysis and Ingenuity Pathway Analysis of Metastatic clear Cell Renal Cell Carcinoma Cell Line. *Am. J. Physiology-Renal Physiol.* 311 (2), F424–F436. doi:10.1152/ajprenal.00138.2016

Kuechler, A., Buysse, K., Clayton-Smith, J., Le Caignec, C., David, A., Engels, H., et al. (2011). Five Patients with Novel Overlapping Interstitial Deletions in 8q22.2q22.3. *Am. J. Med. Genet.* 155 (8), 1857–1864. doi:10.1002/ajmg.a.34072

Kuroda, Y., Ohashi, I., Saito, T., Nagai, J.-i., Ida, K., Naruto, T., et al. (2014). Refinement of the Deletion in 8q22.2-q22.3: the Minimum Deletion Size at 8q22.3 Related to Intellectual Disability and Epilepsy. *Am. J. Med. Genet.* 164 (8), 2104–2108. doi:10.1002/ajmg.a.36604

Kuzniecky, R. (2015). Epilepsy and Malformations of Cortical Development. *Curr. Opin. Neurol.* 28 (2), 151–157. doi:10.1097/wco.0000000000000175

Kwan, P., Arzimanoglou, A., Berg, A. T., Brodie, M. J., Allen Hauser, W., Mathern, G., et al. (2010). Definition of Drug Resistant Epilepsy: Consensus Proposal by the Ad Hoc Task Force of the ILAE Commission on Therapeutic Strategies. *Epilepsia* 51 (6), 1069–1077. doi:10.1111/j.1528-1167.2009.02397.x

Kwan, P., and Brodie, M. J. (2000). Early Identification of Refractory Epilepsy. *N. Engl. J. Med.* 342 (5), 314–319. doi:10.1056/nejm200002033420503

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

Lerche, H. (2020). Drug-resistant Epilepsy - Time to Target Mechanisms. *Nat. Rev. Neurol.* 16 (11), 595–596. doi:10.1038/s41582-020-00419-y

Li, J., Liu, C., Chen, Y., Gao, C., Wang, M., Ma, X., et al. (2019). Tumor Characterization in Breast Cancer Identifies Immune-Relevant Gene Signatures Associated with Prognosis. *Front. Genet.* 10, 1119. doi:10.3389/fgene.2019.01119

Lim, J. S., Kim, W.-i., Kang, H.-C., Kim, S. H., Park, A. H., Park, E. K., et al. (2015). Brain Somatic Mutations in MTOR Cause Focal Cortical Dysplasia Type II Leading to Intractable Epilepsy. *Nat. Med.* 21 (4), 395–400. doi:10.1038/nm.3824

Liu, X., and Wang, Q. (2015). Screening of Feature Genes in Distinguishing Different Types of Breast Cancer Using Support Vector Machine. *Ott* 8, 2311–2317. doi:10.2147/ott.S85271

Löscher, W., Potschka, H., Sisodiya, S. M., and Vezzani, A. (2020). Drug Resistance in Epilepsy: Clinical Impact, Potential Mechanisms, and New Innovative Treatment Options. *Pharmacol. Rev.* 72 (3), 606–638. doi:10.1124/pr.120.019539

Mattson, R. H., Cramer, J. A., Collins, J. F., Smith, D. B., Delgado-Escueta, A. V., Browne, T. R., et al. (1985). Comparison of Carbamazepine, Phenobarbital, Phenytoin, and Primidone in Partial and Secondarily Generalized Tonic-Clonic Seizures. *N. Engl. J. Med.* 313 (3), 145–151. doi:10.1056/nejm198507183130303

Rawat, C., Kushwaha, S., Srivastava, A. K., and Kukreti, R. (2020). Peripheral Blood Gene Expression Signatures Associated with Epilepsy and its Etiologic Classification. *Genomics* 112 (1), 218–224. doi:10.1016/j.ygeno.2019.01.017

Redmond, L., Kashani, A. H., and Ghosh, A. (2002). Calcium Regulation of Dendritic Growth via CaM Kinase IV and CREB-Mediated Transcription. *Neuron* 34 (6), 999–1010. doi:10.1016/s0896-6273(02)00737-7

Rivera, J., Proia, R. L., and Olivera, A. (2008). The alliance of Sphingosine-1-Phosphate and its Receptors in Immunity. *Nat. Rev. Immunol.* 8 (10), 753–763. doi:10.1038/nri2400

Ryvlin, P., Rheims, S., and Lhatoo, S. D. (2019). Risks and Predictive Biomarkers of Sudden Unexpected Death in Epilepsy Patient. *Curr. Opin. Neurol.* 32 (2), 205–212. doi:10.1097/wco.0000000000000668

Schmidt, D., and Schachter, S. C. (2014). Drug Treatment of Epilepsy in Adults. *Bmj* 348, g254. doi:10.1136/bmj.g254

Schwenk, R. W., Vogel, H., and Schürmann, A. (2013). Genetic and Epigenetic Control of Metabolic Health. *Mol. Metab.* 2 (4), 337–347. doi:10.1016/j.molmet.2013.09.002

Sharma, P., Kumar, A., and Singh, D. (2019). Dietary Flavonoids Interaction with CREB-BDNF Pathway: An Unconventional Approach for Comprehensive Management of Epilepsy. *Cn* 17 (12), 1158–1175. doi:10.2174/1570159x17666190809165549

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res.* 43 (Database issue), D447–D452. doi:10.1093/nar/gku1003

Thijs, R. D., Surges, R., O'Brien, T. J., and Sander, J. W. (2019). Epilepsy in Adults. *The Lancet* 393 (10172), 689–701. doi:10.1016/s0140-6736(18)32596-0

Tomson, T., Battino, D., and Perucca, E. (2016). Valproic Acid after Five Decades of Use in Epilepsy: Time to Reconsider the Indications of a Time-Honoured Drug. *Lancet Neurol.* 15 (2), 210–218. doi:10.1016/s1474-4422(15)00314-2

Wang, G., Zhu, Z., Xu, D., and Sun, L. (2020). Advances in Understanding CREB Signaling-Mediated Regulation of the Pathogenesis and Progression of Epilepsy. *Clin. Neurol. Neurosurg.* 196, 106018. doi:10.1016/j.clineuro.2020.106018

Zhang, Y., Dong, H., Duan, L., Yuan, G., Liang, W., Li, Q., et al. (2019). SLC1A2 Mediates Refractory Temporal Lobe Epilepsy with an Initial Precipitating Injury by Targeting the Glutamatergic Synapse Pathway. *IUBMB Life* 71 (2), 213–222. doi:10.1002/iub.1956

Zhu, Y., Elemento, O., Pathak, J., and Wang, F. (2019). Drug Knowledge Bases and Their Applications in Biomedical Informatics Research. *Brief Bioinform* 20 (4), 1308–1321. doi:10.1093/bib/bbx169

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

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 03 January 2022
doi: 10.3389/fmolb.2021.781427

The Pyroptosis-Related Gene Signature Predicts the Prognosis of Hepatocellular Carcinoma

Shuqiao Zhang[1†], Xinyu Li[2†], Xiang Zhang[3], Shijun Zhang[4], Chunzhi Tang[2] and Weihong Kuang[5*]*

[1]First Affiliated Hospital of Chinese Medicine, Guangzhou University of Chinese Medicine, Guangzhou, China, [2]Medical College of Acupuncture-Moxibustion and Rehabilitation, Guangzhou University of Chinese Medicine, Guangzhou, China, [3]The Second Clinical Medical College, Zhejiang Chinese Medical University, Hangzhou, China, [4]Department of Traditional Chinese Medicine, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China, [5]Guangdong Key Laboratory for Research and Development of Natural Drugs, School of Pharmacy, Guangdong Medical University, Dongguan, China

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Vikram Dalal,
Washington University in St. Louis, United States
Weijia Liao,
Affiliated Hospital of Guilin Medical University, China

***Correspondence:**
Chunzhi Tang
jordan664@163.com
Weihong Kuang
Kuangwh@gdmu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: 22 September 2021
Accepted: 13 December 2021
Published: 03 January 2022

Citation:
Zhang S, Li X, Zhang X, Zhang S, Tang C and Kuang W (2022) The Pyroptosis-Related Gene Signature Predicts the Prognosis of Hepatocellular Carcinoma. Front. Mol. Biosci. 8:781427. doi: 10.3389/fmolb.2021.781427

Objective: Hepatocellular carcinoma (HCC) is a genetically and phenotypically heterogeneous tumor, and the prediction of its prognosis remains a challenge. In the past decade, studies elucidating the mechanisms that induce tumor cell pyroptosis has rapidly increased. The elucidation of their mechanisms is essential for the clinical development optimal application of anti-hepatocellular carcinoma therapeutics.

Methods: Based on the different expression profiles of pyroptosis-related genes in HCC, we constructed a LASSO Cox regression pyroptosis-related genes signature that could more accurately predict the prognosis of HCC patients.

Results: We identified seven pyroptosis-related genes signature (*BAK1*, *CHMP4B*, *GSDMC*, *NLRP6*, *NOD2*, *PLCG1*, *SCAF11*) in predicting the prognosis of HCC patients. Kaplan Meier survival analysis showed that the pyroptosis-related high-risk gene signature was associated with poor prognosis HCC patients. Moreover, the pyroptosis-related genes signature performed well in the survival analysis and ICGC validation group. The hybrid nomogram and calibration curve further demonstrated their feasibility and accuracy for predicting the prognosis of HCC patients. Meanwhile, the evaluation revealed that our novel signature predicted the prognosis of HCC patients more accurately than traditional clinicopathological features. GSEA analysis further revealed the novel signature associated mechanisms of immunity response in high-risk groups. Moreover, analysis of immune cell subsets with relevant functions revealed significant differences in aDCs, APC co-stimulation, CCR, check-point, iDCs, Macrophages, MHC class-I, Treg, and type II INF response between high- and low-risk groups. Finally, the expression of Immune checkpoints was enhanced in high-risk group, and m6A-related modifications were expressed differently between low- and high-risk groups.

Conclusion: The novel pyroptosis-related genes signature can predict the prognosis of patients with HCC and insight into new cell death targeted therapies.

Keywords: hepatocellular carcinoma, pyroptosis, prognosis, immune infiltration, signature

INTRODUCTION

Hepatocellular carcinoma (HCC) is the third most aggressive and lethal disease, accounting for approximately 75% of liver cancer cases, and is a highly genetically and phenotypically heterogeneous malignancy with 830,000 deaths in 2020 (Petrick et al., 2016; Moon and Ro, 2021). Alcohol abuse, obesity, diabetes, and metabolic syndromes are significant risk factors for HCC progression, and inflammation caused by these risk factors promotes liver fibrosis, leading to cirrhosis and ultimately HCC(Mittal and El-Serag, 2013; Kim and Viatour, 2020). Patients with HCC are asymptomatic at the early stage, which seriously delays timely diagnosis. Patients diagnosed at the late stage of HCC are not suitable for radical surgery, resulting in minimal availability and effectiveness of therapeutic options (Llovet et al., 2008). Thus, novel biomarkers that can discriminate patients at high risk for HCC are urgently needed to improve personalized HCC prognostic prediction accuracy and treatment.

In the past decade, studies elucidating the mechanisms that induce tumor cell pyroptosis has rapidly increased (Derangere et al., 2014; Jiang et al., 2017; Wang Y et al., 2018). Pyroptosis is an inflammatory caspase-dependent cell death type characterized by pore formation, cell swelling and rupture of the plasma membrane, and release of intracellular contents (Ruan et al., 2020). Pyroptosis therapies are increasing as opportunities to inhibit cancer development. Meanwhile, pyroptosis promotes inflammatory cell death and inhibits cancer cell proliferation and migration, and decreased expression of some pyroptotic inflammasomes has been found in cancer cells (Fang et al., 2020). Apoptosis is widely studied as a major form of regulated cell death underlying tumor pathogenesis and therapy. Still, cancer-associated defects in apoptosis induction and execution contribute to a significant proportion of treatment failures (Ng et al., 2012; Holohan et al., 2013; Hata et al., 2014). The clear molecular pathways mediating necrotic types of cell death have recently been uncovered, the long-standing view of apoptosis as a standard regulating mechanism of death programs has changed (Vanden Berghe et al., 2014; Conrad et al., 2016; Wallach et al., 2016). The previously unknown mechanism of pyroptosis as a molecularly targeted pathway to eradicate oncogene addicted tumor cells may have important implications for the clinical development and optimal application of anticancer therapeutics (Lu et al., 2018).

However, studies on the functions and mechanisms of pyroptosis-related genes in HCC progression remain scarce. A systematic evaluation of pyroptosis-related gene prognostic signatures and their correlation with HCC patients may further our understanding of HCC mechanisms and provide new applications for a rapid, effective, and specific diagnosis and effective therapy.

A novel pyroptosis-related prognostic signature of differentially expressed genes in HCC was established in our study. Then we studied their role in the prognosis of HCC patients and the associated immune response and the effect of N6- methylation on adenosine (m6A) modification.

METHODS

Data Collection

We extracted RNA sequencing (50 normal and 374 tumors) data of 377 patients from the TCGA-LIHC (https://portal.gdc.cancer.gov/repository) dataset, and RNA sequencing (273 tumors) data of 261 patients from the ICGC-LIRI-JP (https://dcc.icgc.org/releases/current/Projects/LIRI-JP) dataset. Clinical characteristics of HCC patients in the TCGA and ICGC dataset was shown in **Supplementary Table S1**. The corresponding pyroptosis-related genes in **Supplementary Table S2** were identified from the previous studies of multiple regulatory mechanisms of pyroptosis in the tumor microenvironment (Xia et al., 2019; Shao et al., 2021; Ye et al., 2021) and Molecular Signatures database (http://www.gsea-msigdb.org/gsea/login.jsp) (Liberzon et al., 2015). Before comparison, normalization of the expression data in both datasets values was performed using fragment per kilobase million (FPKM) values. The association between pyroptosis-related genes and HCC was assessed using the "limma" R package, and the correlation was considered significant if the p-value was <0.05. The protein-protein interaction (PPI) network of the pyroptosis-related differentially expressed genes (DEGs) was developed by STRING (Szklarczyk et al., 2021), version 11.5 (https://string-db.org/).

Functional Enrichment Analysis

First, the biological process (BP), cellular component (CC), and molecular function (MF) of the pyroptosis-related DEGs were investigated using Gene Ontology (GO). Then the biological pathway functions of DEGs were further analyzed by Kyoto Encyclopedia of Genes and Genomes (KEGG) based data in R software version 4.0.5.

Development of the Pyroptosis-Related Genes Prognostic Signature

To construct an accurate and reliable prognostic prediction signature for HCC patients, we first screened the resulting pyroptosis-related DEGs for those with predictive value using univariate Cox regression analysis and then further processed using LASSO regression analysis prevent the fitting of risk models. Finally, the pyroptosis-related genes signature was constructed and stratified according to the risk score ($e^{\text{sum (each genes' expression} \times \text{corresponing cofficient)}}$). Finally, HCC patients were divided into high-risk (≥median) and low-risk (<median) groups according to the median value of the risk score of the established prognostic model.

The Predictive Nomogram and Calibration Curves

To create a clinically practical approach in predicting the 1, 3, and 5-year overall survival rate of HCC patients, we developed a hybrid nomogram model incorporating independent prognostic factor including risk score signature, gender, age, TMN, stage,

and grade. We then validated the accuracy of the nomogram model for judging the prognosis situation of HCC patients using the degree of fit of the calibration curve to the actual observed values.

Immune Profile Analysis

Meanwhile, immune cell infiltration levels of the seven pyroptosis-related genes signatutre in individual samples in two risk groups were quantified by single-sample gene set enrichment analysis (ssGSEA) (Rooney et al., 2015). The cellular immune responses of the pyroptosis-related genes signature between subgroups were then evaluated by comparing the results of CIBERSORT(Newman et al., 2015; Charoentong et al., 2017), CIBERSORT−ABS (Wang L et al., 2020), QUANTISEQ (Plattner et al., 2020), MCPCOUNTER (Shi et al., 2020), XCELL (Aran et al., 2017), EPIC(Racle et al., 2017), and TIMER (Li et al., 2017) algorithms. In addition, we evaluated differences in immune function expression by tumor-infiltrating immune cell subsets in the two risk groups. Finally, we analyzed the status of m6A methylation modification in high and low-risk groups to explore the possible impact of the seven pyroptosis-related genes on the activities of methyltransferases, demethylases, and methylated reader proteins in HCC.

Independent Prognostic Validation of the Prognostic Signature

Information on clinical characteristics, including gender, age and staging data, of HCC patients in the TCGA dataset and HCC patients in the ICGC dataset was extracted. These clinical variables in combination with our risk score prognostic signature was analyzed by univariate and multivariate Cox regression.

Statistical Analysis

We used Bioconductor packages including "limma," "survival," "survminer" in Rstudio software (Version 1.4.1106) for analyzing data. Wilcoxon test and unpaired Student's t-test were used to comparing non-normal and normal distribution expression variables. Based on the false discovery rate, the different expression of genes was corrected by the Benjamin Hochberg method to control the elevated false-positive rate. Kaplan Meier (KM) survival analysis was performed to evaluate the feasibility of pyroptosis-related genes signature for predicting the overall survival of HCC patients. Time-dependent receiver operator characteristic curve (ROC) and decision curve analysis (DCA) (Vickers et al., 2008) was used to validate the reliability of the predictive model and to compare the accuracy of the novel pyroptosis related gene signature with traditional clinicopathological features in predicting the prognosis of HCC patients. Furthermore, Fisher's exact test was used to analyze pyroptosis-related gene expression profiles among the clinicopathological features. To analyze the pyroptosis-related DEGs associated immune status in each sample in the TCGA-LIHC cohort, the relative infiltration of 20 immune cell types in the tumor microenvironment was calculated via ssGSEA with the application of the "GSVA" package in R. $p < 0.05$ in the results of all analyses was considered statistically significant. The flow-process diagram of this study is shown in **Figure 1**.

RESULTS

Identification of Pyroptosis-Related DEGs

42 pyroptosis-related DEGs among HCC and normal liver tissues in the TCGA-LIHC dataset were identified using the limma R package (**Supplementary Table S3**). The expression level of these genes was presented as a heatmap in **Figure 2A**. Further by PPI analysis, we explored the interactions among these DEGs (**Figure 2B**). With the minimum required interaction score of 0.9 (the highest confidence) in the PPI analysis, we determined *NLRP3, CHMP4A, CASP8, CASP3, TP53, PYCARD, CHMP2A,* and *IL1B* were hub genes. The correlation network of the pyroptosis-related DEGs is shown in **Figure 2C**.

Pyroptosis-Related DEGs-Based HCC Classification Pattern

To explore the connections between the expression of the 42 pyroptosis-related DEGs and HCC subtypes, we performed a consensus clustering analysis with all 377 HCC patients in the TCGA-LIHC cohort. By increasing the clustering variable (k) from 2 to 9, we found that when $k = 2$, the intragroup correlations were the highest and the intergroup correlations were low, indicating that the 377 HCC patients could be well divided into two clusters based on the 42 DEGs (**Figure 3A**). The DEGs expression profile and the clinicopathological characteristics were presented in the heatmap (**Figure 3B**). We also compared the survival advantage between the two clusters, and the KM overall survival curves showed that the survival probability of cluster 1 was higher than cluster 2 (**Figure 3C**).

Enrichment Analysis of Pyroptosis-Related DEGs

Gene Ontology (GO) function and KEGG pathways enrichment analyses of the DEGs were performed. Enriched biological process (BP), including regulation of interleukin−1 production, midbody abscission, and mitotic cytokinetic process. Meanwhile, phospholipid binding, cytokine receptor binding, and cysteine−type endopeptidase activity were the regular molecular function (MF). Cellular component (CC) mainly comprised the ESCRT complex, multivesicular body, late endosome, and inflammasome complex (**Figure 4A**). Moreover, KEGG pathways analysis demonstrated that necroptosis, NOD−like receptor signaling pathway, apoptosis, hepatitis, P53 signaling pathway, MAPK signaling pathway, and MicroRNAs in cancer were markedly enriched (**Figure 4B**).

Development of Pyroptosis-Related Gene Prognostic Signature

First, ten HCC prognosis related pyroptosis genes were screened out from the DEGs by univariate Cox analysis (**Figure 5A**). Next,

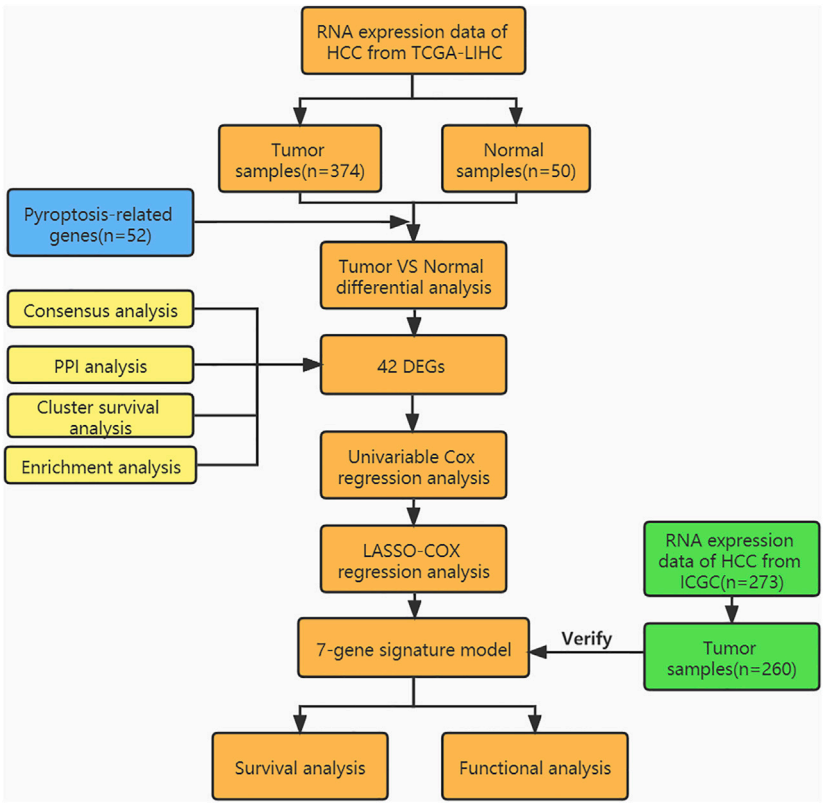


FIGURE 1 | Workflow diagram.

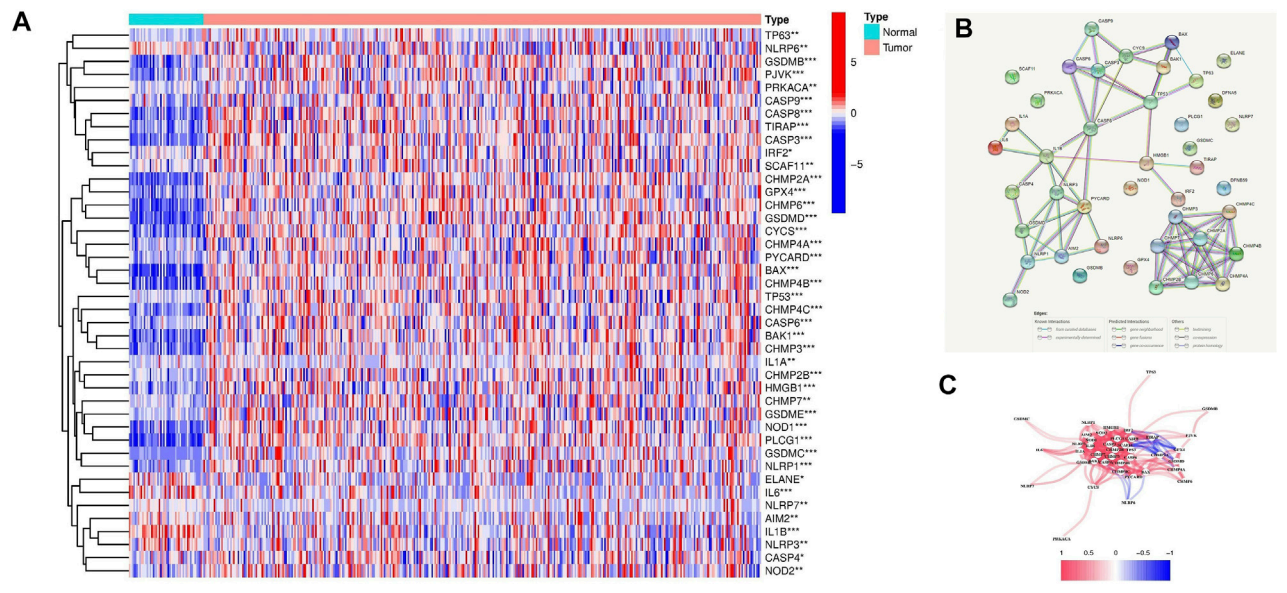


FIGURE 2 | Expression of the 42 pyroptosis-related DEGs and their interactions. **(A)** Heatmap of the pyroptosis-related DEGs between the normal and the tumor samples (blue: low expression level; red: high expression level). *p* values were presented as: **p* < 0.05; ***p* < 0.01; ****p* < 0.001. **(B)** The PPI network showed the interactions among the pyroptosis-related DEGs. **(C)** The correlation network of the pyroptosis-related DEGs (blue lines: negative correlations; red lines: positive correlations. The color depth reflected the strength of their relevance).

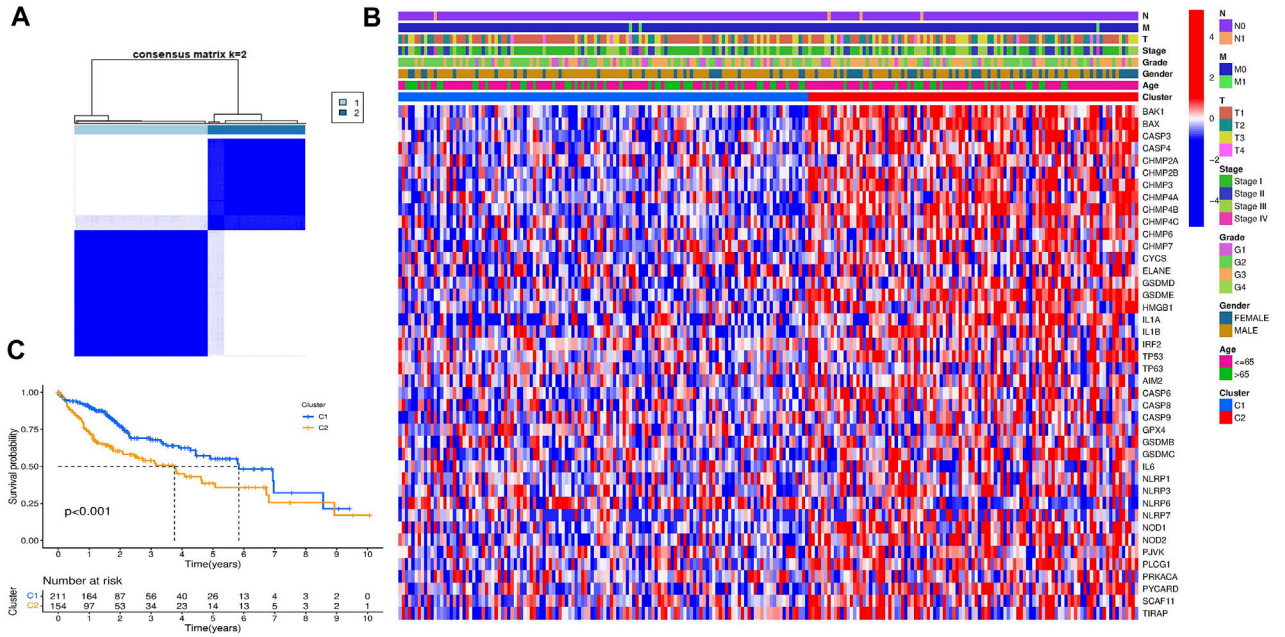


FIGURE 3 | HCC classification pattern based on the pyroptosis-related DEGs. **(A)** 377 HCC patients were divided into two groups when $k = 2$ in the TCGA cohort. **(B)** Heatmap of the clinicopathological characteristics between the two clusters classified by the DEGs. **(C)** KM overall survival curves of the two clusters.

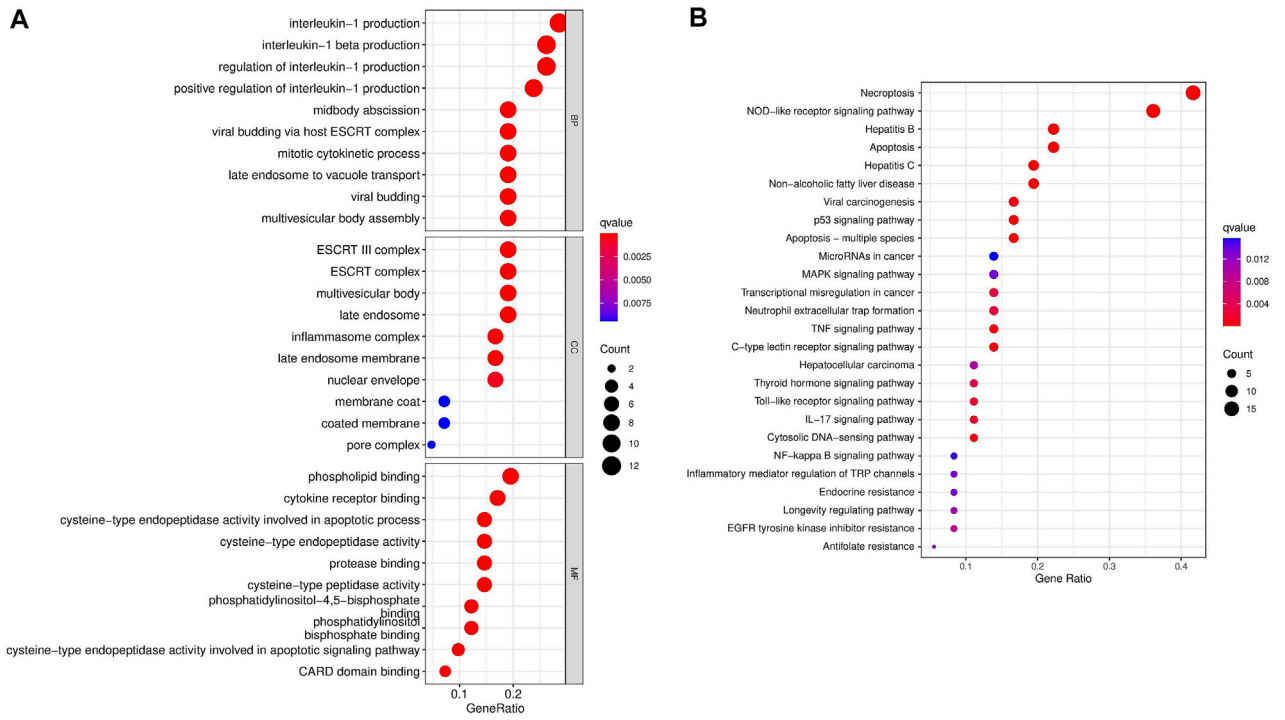


FIGURE 4 | Gene Ontology and KEGG enrichment analysis of the pyroptosis-related DEGs. **(A)** GO analysis. (**B**) KEGG analysis.

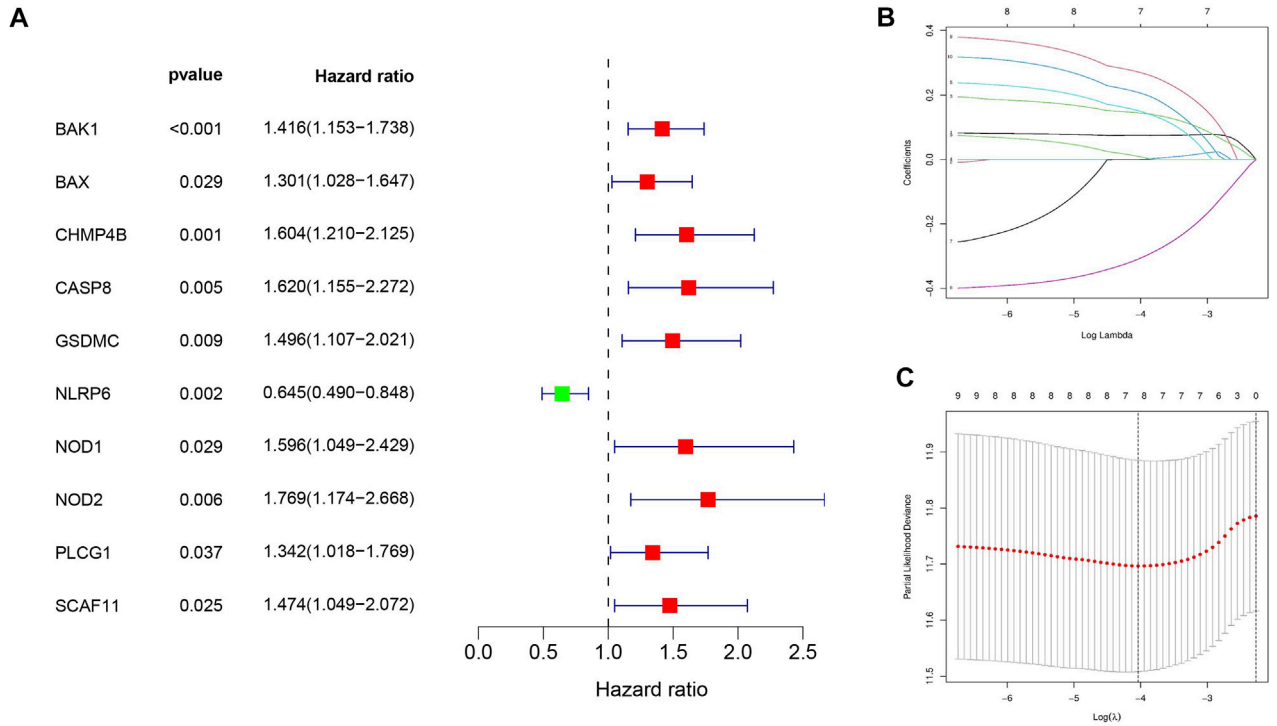


FIGURE 5 | Development of seven pyroptosis-related genes prognostic signature. **(A)** Univariate Cox regression revealed 10 pyroptosis-related genes associated with prognosis. **(B)** 10 pyroptosis-related genes were penalized by LASSO Cox regression analysis. **(C)** 10-fold cross-validation for the optimal parameter selection in the LASSO Cox regression.

the ten pyroptosis-related genes were penalized by LASSO Cox regression (**Figures 5B,C**). Finally, the pyroptosis-related genes signature was constructed based on the risk score= (0.07486**BAK1* exp.) + (0.14487 **CHMP4B* exp.) + (0.15165**GSDMC* exp.) + (−0.309234**NLRP6* exp.) + (0.27176 **NOD2* exp.) + (0.00979 **PLCG1* exp.) + (0.20830 **SCAF11* exp.).

Survival Results and Multivariate External Examination

KM analysis confirmed that the TCGA and ICGC cohorts HCC patients in the high-risk group were associated with worse OS (**Figures 6A,D**). At the same time, we could see from the hazard survival status plots of the high-risk groups that high expression of the novel predictive model is correlated with poor survival of HCC patients (**Figures 6B–F**). Besides, PCA analysis and t-SNE analysis presented that HCC patients in different risk groups were distributed in two directions (**Figures 6G–J**). Then, we performed ROC analysis using the timeROC package in R. The prognostic prediction power (AUC) of the seven pyroptosis-related genes signature in the TCGA-LIHC cohort was 0.753(1 year), 0.616(3 years), and 0.639 (5 years) (**Figure 7A**). Furthermore, the AUC of the seven pyroptosis-related genes signature in the IGCG validation cohort was 0.663(1 year), 0.643(3 years), and 0.638 (5 years) (**Figure 7C**). The clinical characteristics of ROC analysis revealed that compared with the traditional pathological characteristics, the risk score model could more accurately predict the prognosis of HCC patients in the TCGA cohort (AUC = 0.743, **Figure 7B**) and ICGC cohort (AUC = 0.772, **Figure 7D**).

Independent Prognostic Value Validation of the Risk Signature

Univariate and multivariate cox analyses were conducted to verify whether the novel pyroptosis-related genes risk score signature was an independent prognostic factor for overall survival of HCC patients. The risk score model in the TCGA and ICGC cohorts were significantly associated with overall survival of HCC patients in the univariate Cox analysis (TCGA cohort: HR = 4.385, 95% CI = 2.303–8.350, $p < 0.001$; ICGC cohort: HR = 3.468, 95% CI = 1.363–8.821, $p = 0.009$) (**Figures 8A,C**). After correcting for other confounders, the multivariate Cox analysis confirmed that the risk score signature remained an independent predictor of overall survival for HCC patients. (TCGA cohort: HR = 3.837, 95% CI = 2.008–7.329, $p < 0.001$; ICGC cohort: HR = 2.674, 95% CI = 1.114–6.418, $p = 0.028$) (**Figures 8B,D**). The clinical heatmap presented the relationship between the novel signature and traditional clinicopathological manifestations in **Figure 8E**. The fitting degree of calibration curve verified the accuracy of the nomogram model in predicting the prognosis of patients with HCC. (**Figures 9A,B**). Meanwhile, the net benefit of the risk score signature in the DCA was superior to traditional clinical and pathological characteristics in

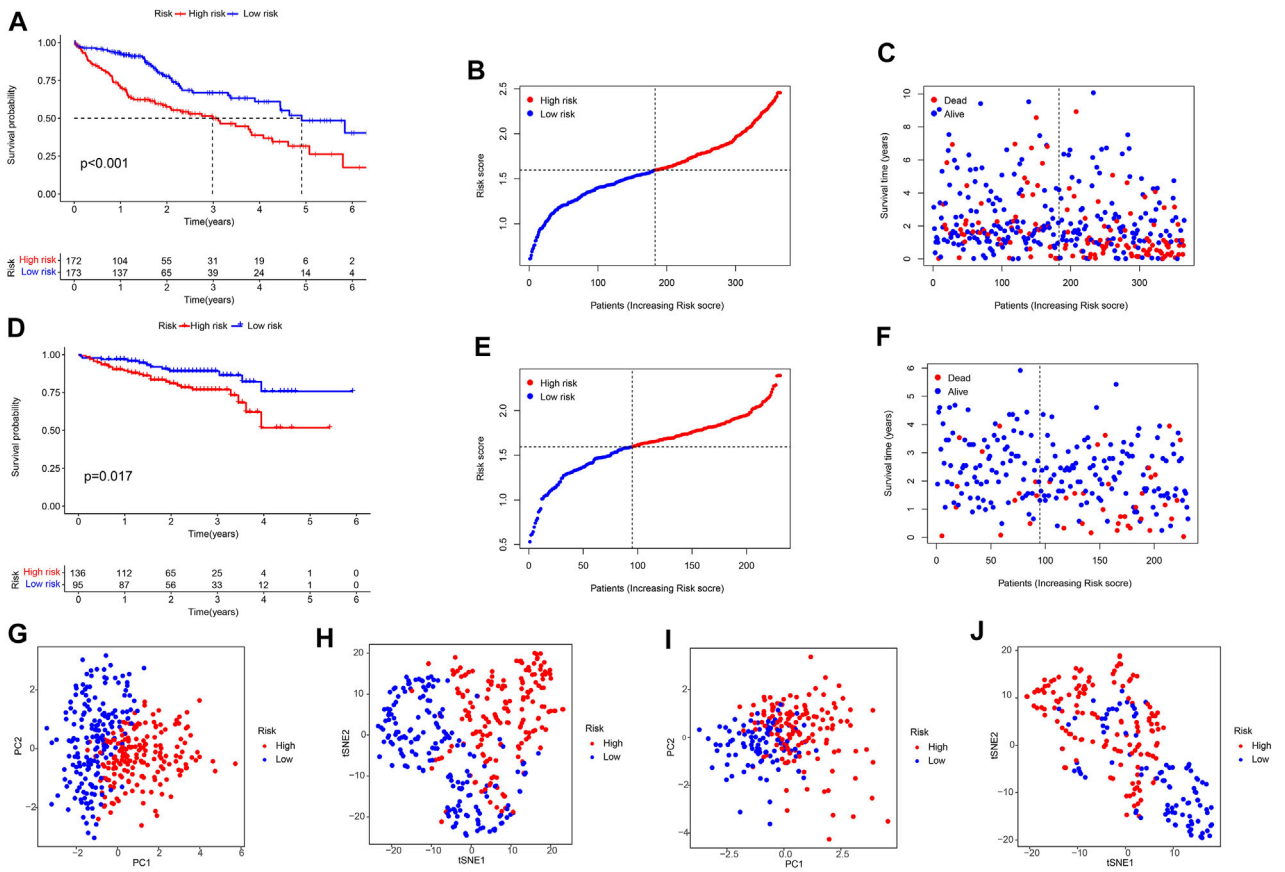


FIGURE 6 | Survival analysis of the seven pyroptosis-related genes signature in the TCGA-LIHC cohort and ICGC-LIRI-JP cohort. TCGA-LIHC cohort **(A–C,G,H)**, ICGC-LIRI-JP cohort **(D–F, I,J)**. **(A,D)** KM survival analysis result. **(B,C,E,F)** Survival status and the risk score distribution of HCC patients. **(G,I)** PCA plot. **(H,J)** t-SNE analysis.

predicting the prognosis of HCC patients (**Figure 9C**). Therefore, this nomogram could be used in predicting the prognostic of HCC patients.

Gene Set Enrichment Analysis

The potential pathways, mechanisms, and bioprocess of the pyroptosis-related genes signature were analyzed based on GSEA, which revealed those genes regulated both the tumor development and immune response, centrally including NOD-like receptor signaling pathway, T-cell receptor signaling pathway, WNT signaling pathway, regulation of autophagy, MAPK signaling pathway, spliceosome, VEGF signaling pathway and pathways in cancer (**Figure 10**; **Supplementary Table S4**).

Immunological Reaction and Immune Checkpoints Expression

The Heatmap showed that the expression of the immune cell infiltration responses of the novel pyroptosis-related genes signature was significantly upregulated in HCC under the QUANTISEQ, CIBERSORT, CIBERSORT-ABS, MCPCOUNTER, XCELL, TIMER, and EPIC algorithms (**Figure 11A**; **Supplementary Table S5**). Single-sample gene set enrichment analysis based on TCGA-LIHC data showed expression of immune cell subsets and relevant functions, significantly different between the two risk groups. p values were presented as: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. The high-risk group's most prominent up-regulated immune functions were aDCs, APC co-stimulation, CCR, check-point, iDCs, Macrophages, MHC class-I, Treg. In contrast, type II INF response was down-regulated in the high-risk group, implying one of the main causes that suppression of the production and release of IFNs leads to loss of control over HCC growth (**Figure 11B**). Given the importance of immunotherapy based on checkpoint inhibitors for HCC, we further investigated the expressions of immune checkpoints in the two risk groups. The results showed that most immunological checkpoints were more active in high-risk groups in **Figure 11C**. The analysis of the effect of the pyroptosis-related genes signature on m6A-related modification showed the methylation expression level of *YTHDF1, YTHDF2, WTAP, YTHDC1, YTHDF2, FTO, HNRNPC, ALKBH5, RBM15, YTHDC2,* and *METTL3* in the high-risk group was higher. (**Figure 11D**).

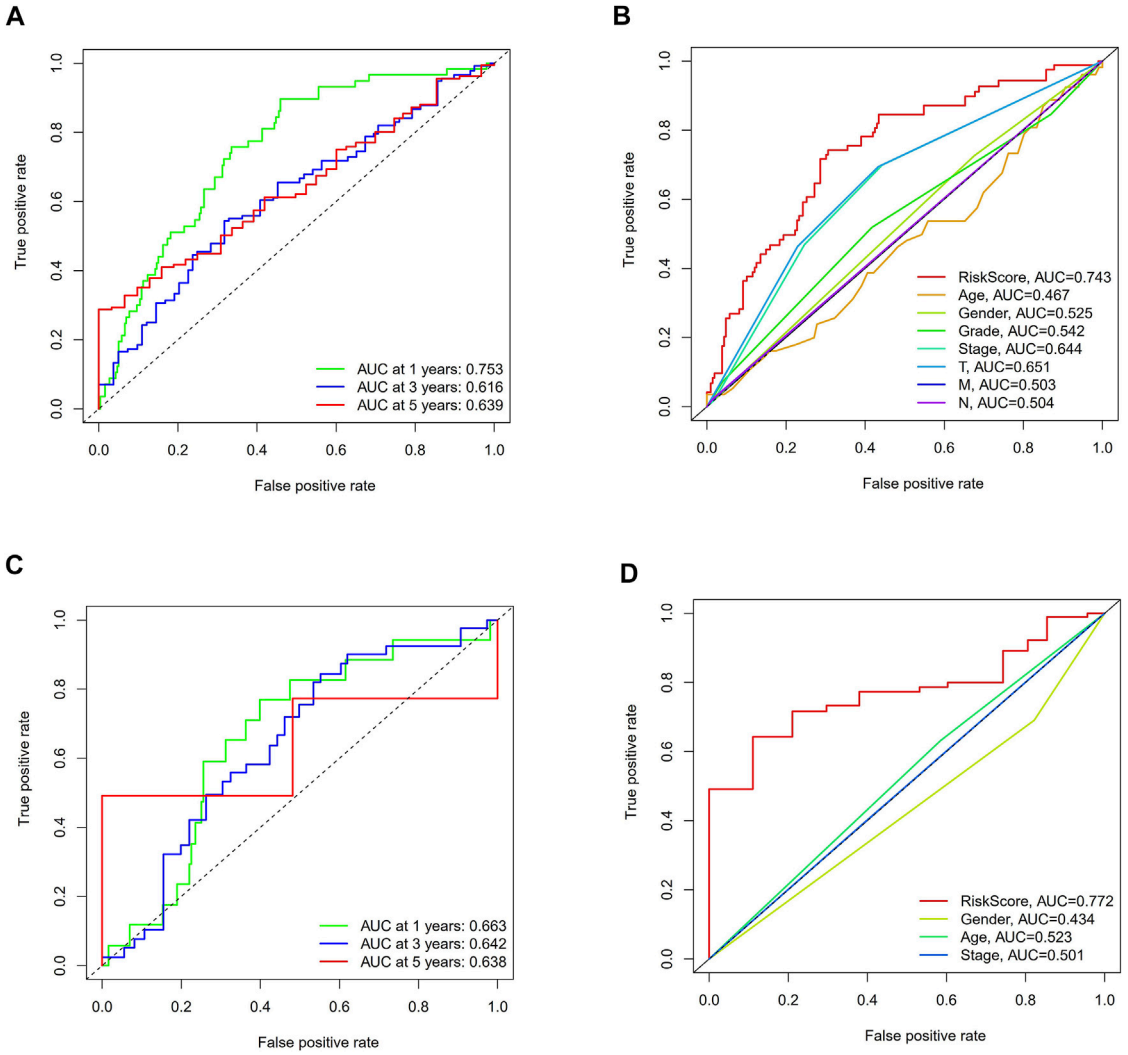


FIGURE 7 | The ROC curve analysis of the seven pyroptosis-related genes signature in the two cohorts. **(A,C)** Time-dependent ROC analysis for HCC patients. **(B, D)** The ROC analysis for clinical features and risk score signature.

DISCUSSION

Cell death is one of the most fundamental problems of life and plays a crucial role in organismal development, homeostasis, and cancer pathogenesis (Hanahan and Weinberg, 2011). As a model of programmed cell death, pyroptosis, although capable of suppressing tumor cell proliferation, can also create a microenvironment suitable for tumor cell growth and promotion (Minton, 2020; Yu et al., 2021), and thus has received increasing attention. Meanwhile, many recent studies have demonstrated that pyroptosis is closely related to developing liver diseases such as liver damage (Lebeaupin et al., 2015), fatty lesions (Miura et al., 2010), inflammation (Wei et al., 2019), and fibrosis (Wree et al., 2014). However, little is currently known about the role of pyroptosis in liver cancer development, and our study was undertaken to elucidate this role. In this study, we first analyzed 42 pyroptosis DEGs in HCC. Based on the pyroptosis -related DEGs, we determined two molecular subtypes using the consensus clustering algorithm. It was found that the survival probability of C2 was much worse than C1 in overall survival. Functional and KEGG pathways analysis further discovered that these DEGs in subtypes primarily participated in necroptosis, NOD–like receptor signaling pathway, apoptosis, hepatitis, P53 signaling pathway, and MAPK signaling pathway. Some recent studies showed that Caspase/granzyme-induced apoptosis could be switched to pyroptosis by the expression of GSDMs, appears to contribute to the killing of tumor cells by cytotoxic lymphocytes, and reprogram the tumor microenvironment to an immunostimulatory state (Van Opdenbosch and Lamkanfi, 2019; Tsuchiya, 2020; Tsuchiya, 2021). Zhang et al. (2019) reported that overexpression of p53 in human lung cancer alveolar basal epithelial cells significantly reduced tumor growth and mortality by increasing pyroptotic levels in an *in vivo* assay. Therefore, appropriate guiding the pyroptosis of

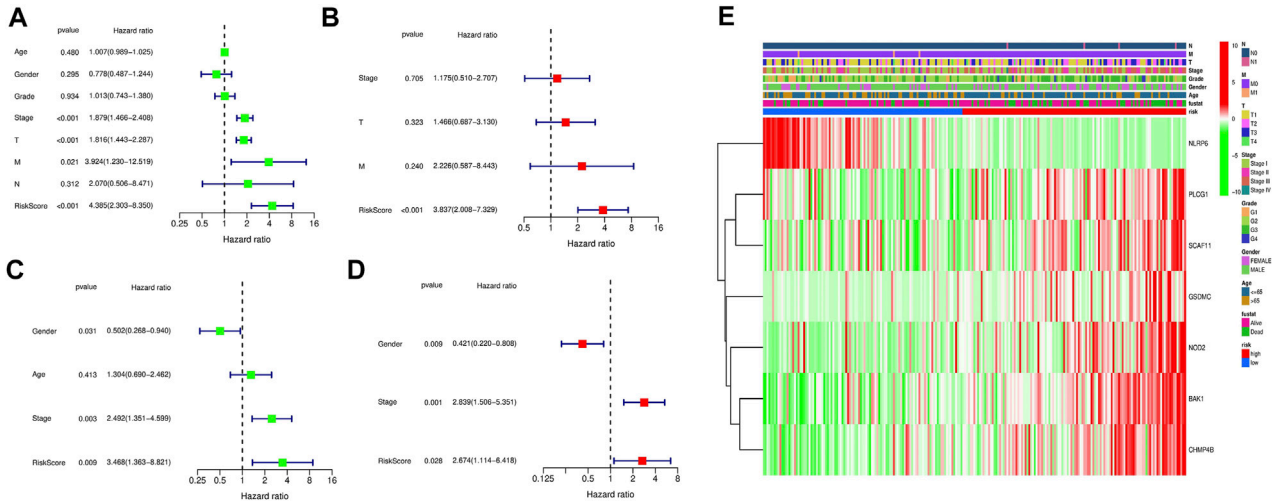


FIGURE 8 | Assessment of the clinical prognostic value of the risk score model in HCC patients by univariate and multivariate COX analyses. **(A)** Univariate independent Cox analysis for TCGA cohort. **(B)** Multivariate independent Cox analysis for TCGA cohort. **(C)** Univariate independent Cox analysis for ICGC cohort. **(D)** Multivariate independent Cox analysis for ICGC cohort. **(E)** Heatmap of the pyroptosis-related genes prognosis signature and clinicopathological manifestations.

hepatocellular carcinoma cells may inspire an advanced therapy strategy of HCC patients.

Next, our study identified seven differently expressed pyroptosis-related gene markers from DEGs as independent prognostic factors for HCC. Among the seven pyroptosis-related genes signature, *BAK1* is a vital cell death regulator that can initiate mitochondria-mediated apoptosis by interacting with proteins (Wang et al., 2013). The protection of *BAK1* by exosomal circ-0051443 through sponging mir-331-3p can inhibit the malignant biological behaviors of HCC(Chen et al., 2020). And silencing CHMP4B can promote epithelial-mesenchymal transition in HCC(Han et al., 2019). *GSDMC* is the only one of the human gasdermin family members whose biological function has not been determined (Kovacs and Miao, 2017). *GSDMC* was significantly associated with poorer prognosis liver cancer patients in our study, indicating that it acts as a tumor-promoting gene. Interestingly, the current study revealed that TNF α - activated caspase-8 switched apoptosis to pyroptosis in the presence of hypoxia-activated *GSDMC* and nPD-L1, leading to tumor necrosis in hypoxic regions (Hou et al., 2020; Du et al., 2021). Therefore, the effect of activating *GSDMC* in different environments on liver cancer is worthy of further exploration. Wang Q et al. (2018) reported that *NLRP6* inhibits gastric cancer cell proliferation, migration, and invasion by regulating the STAT3 signaling pathway, and its down-regulation is closely associated with poor patient prognosis. Similarly, down-regulation of *NLRP6* was associated with poorer prognosis in HCC patients in our study, suggesting that *NLRP6* may play a tumor suppressor role in HCC development. Meanwhile, hepatic *NOD2* promotes hepatocarcinogenesis through a *RIP2* mediated proinflammatory response and novel nuclear autophagy-mediated DNA damage mechanism, and its high expression is closely associated with poor prognosis in HCC patients (Zhou et al., 2021). Furthermore, increased *PLCG1* expression in tumor tissues was significantly associated with adverse clinical features of HCC, which may be a role played by *PLCG1* through activation of mitogen-activated protein kinase and NF-kB signaling pathways (Tang et al., 2019). To date, there are few studies on the regulation of pyroptosis by *SCAF11* in cancer (Xu et al., 2021; Ye et al., 2021). In our study, high expression of *SCAF11* was associated with poor prognosis in liver cancer, reflecting that it may be a liver cancer-promoting factor associated with positively regulating the pyroptosis pathway and inhibition of *SCAF11* should be considered as a target for the treatment of HCC. Based on the median value of the risk score of pyroptosis-related genes signature, HCC patients were divided into high-risk and low-risk group. The survival analyses indicated that the pyroptosis-related high-risk genes were positively related with worse prognosis HCC patients. Moreover, the pyroptosis-related genes signature performed well in the ROC and DCA validation. Finally, their reliability and applicability in predicting HCC prognosis were demonstrated in the nomogram and calibration curve and indicated that our novel risk signature outperformed traditional clinicopathological characteristics.

Pyroptosis serves as a bridge between the immune system and the tumor (Li et al., 2021). Its activation in immune cells and cancer cells will cause the release of inflammatory chemokines and subsequent immune cell infiltration, activating the tumor microenvironment and improving the tumor's efficiency of immunotherapy (Xia et al., 2019; Vietri et al., 2020). On the other hand, the chronic inflammatory response resulting from pyroptosis triggered inflammasomes, and produced cytokines can help tumor cells escape from immune system surveillance and promote the development of tumors (Cookson and Brennan, 2001; Wang Q et al., 2020). In GSEA analysis, the significant

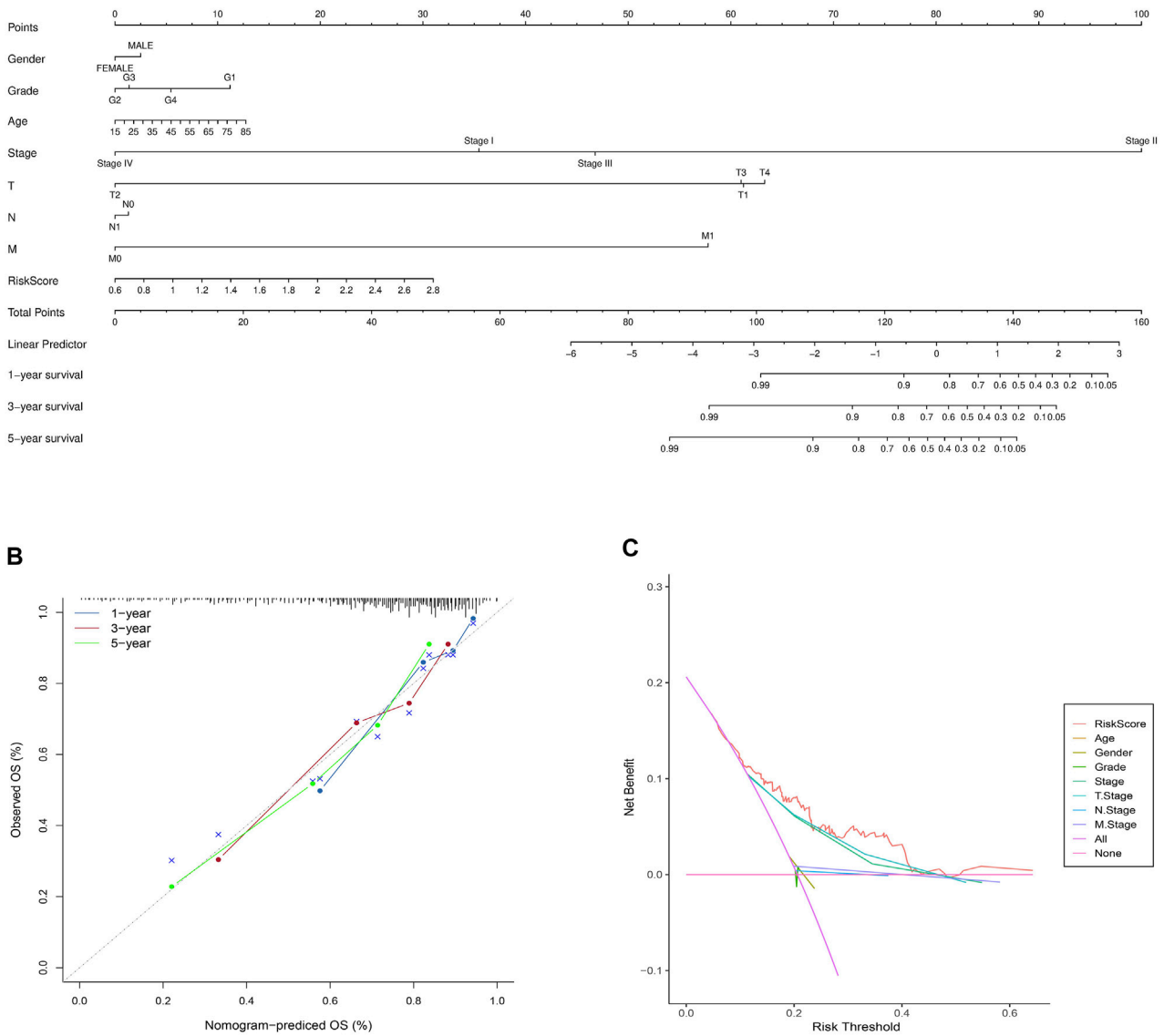


FIGURE 9 | The nomogram model and calibration curves developed based on the risk score signature and prognosis-related clinicopathological indicators. **(A)** The predictive nomogram. **(B)** The calibration curves of the nomogram. **(C)** The decision curve analyses plot.

enrichment of immune and tumor-related pathways among individuals in the high-risk group indicated two sides of the effect of pyroptosis on tumor cell survival, progression, and apoptosis. Furthermore, relevant functional analysis of immune cell subsets revealed that aDCs, APC co-stimulation, CCR, check-point, iDCs, Macrophages, MHC class-I, and Treg of pyroptosis-related genes signature were significantly attenuated in HCC high-risk group, suggesting that reduced levels of antitumor immunity may lead to poor prognosis. Therefore, promoting antitumor immune response is essential to prevent HCC at early stage from further development and generate effective clinical treatments. Moreover, the expression of Immune checkpoints such as *PDCD1*, *PDCDLG2*, *TIGIT*, *LAG3*, and *TNFRSF4* was enhanced in the high-risk group. The PD-1 pathway is a central pathway of immunosuppression in the human tumor microenvironment. Inhibition of PD-1 and PD-L1 can generate endogenous antitumor immunity to inhibit cancer development (Garg and Agostinis, 2017). However, the response rate may be low since inflammation in the cancer-immune microenvironment is ineffective for efficient infiltration and activation of immune cells. The efficiency of anti-PD-1 or PD-L1 therapy can be improved under pyroptosis-induced inflammation in the tumor microenvironment by chemotherapy, radiotherapy, and other therapeutic regimens (Bergsbaken et al., 2009; Reck et al., 2019). Published clinical trials have shown that antibiotic

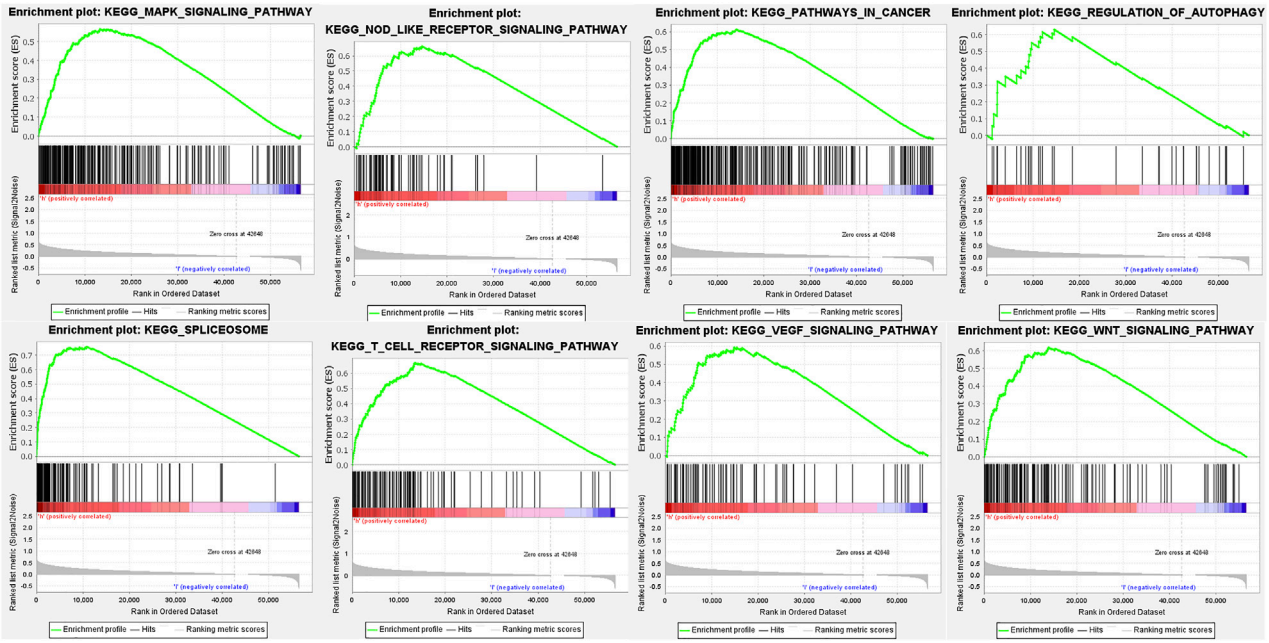


FIGURE 10 | GSEA for the seven pyroptosis-related genes signature.

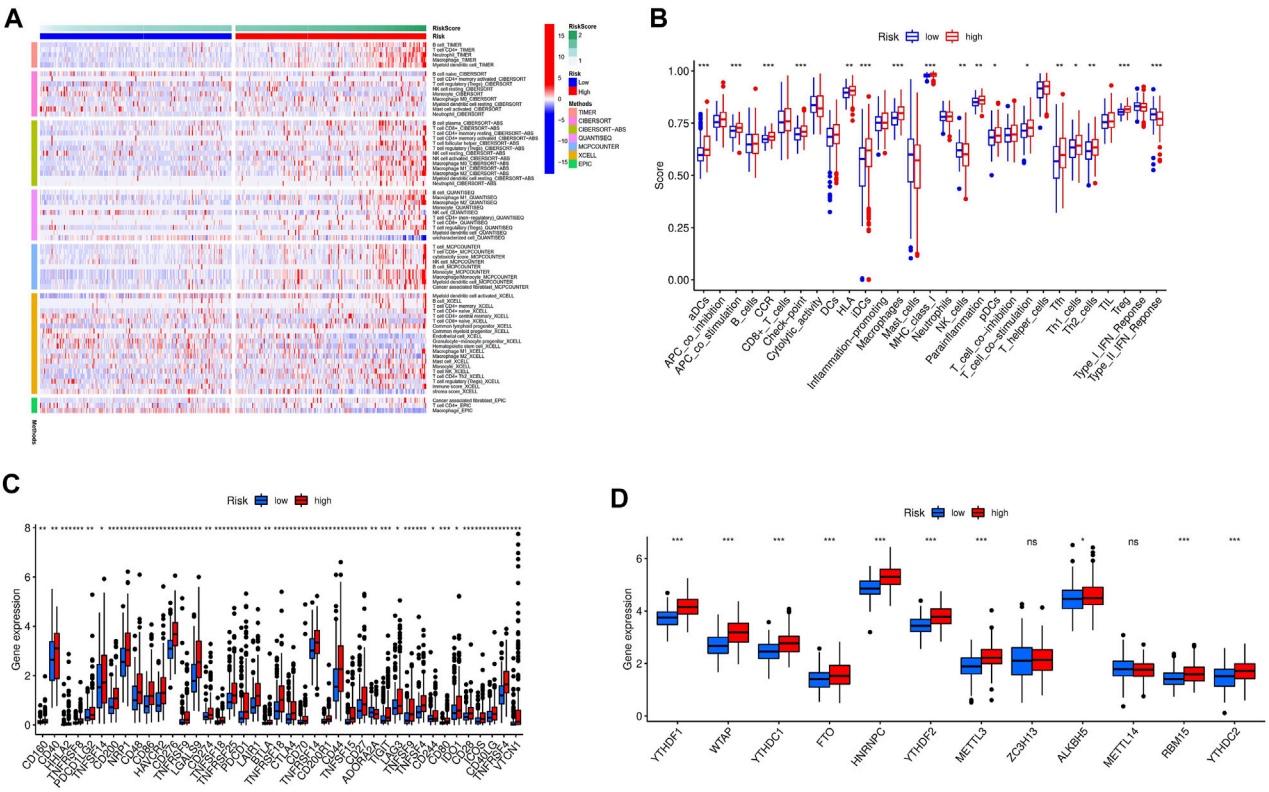


FIGURE 11 | The relationship between prognostic signature and immune response and m6A modification. **(A)** The immune cell infiltration profile of the novel pyroptosis-related genes signature. **(B)** Relevant functional analysis of immune cell subsets. **(C)** Analyses of immune checkpoints between the two HCC risk groups. **(D)** Analyses of m6A modification expression between low and high HCC risk groups.

chemotherapeutics can promote the combination of STAT3 and PD-L1 to upregulate *GSDMC* mediated pyroptosis under hypoxia (Blasco and Gomis, 2020), which may improve HCC patient survival compared to patients received only a single type of treatment to improve the efficiency of PD-L1 inhibitors. TIGIT, similar to LAG3, belongs to the immunoglobulin superfamily and is exclusively expressed on lymphocytes, including CD8 + T cells, memory, and regulatory CD4 + T cells, follicular CD4 + T cells, and NK cells (Stanietsky et al., 2009; Ge et al., 2021). In HCC tumor-bearing mice treated with anti-PD-1, concurrent anti-TIGIT treatment resulted in a combined blockade effect that expanded the effector memory CD8 + T cell population and increased the cytotoxic T cell to Treg ratio in the tumor, thereby suppressing tumor growth and prolonging survival (Li et al., 2018; Chiu et al., 2020; Lepletier et al., 2020), indicating that TIGIT can be used as a rational target to further improve the efficacy of anti-PD-1 therapy in HCC. Unlike standard immune checkpoint blockers that block surface receptors in tumors and T cells responsible for inhibiting antitumor immune responses, drugs that target *TNFRSF4* work by directly activating and modulating the immune response (Alves Costa Silva et al., 2020). Upon treatment of tumor models with an anti-*TNFRSF4* monoclonal antibody, IL-10 production by tumor-infiltrating Treg cells is reduced, allowing the maturation of dendritic cells (Burocchi et al., 2011; Zhang et al., 2018), creating a permissive immune state that allows for the maturation of dendritic accumulation of myeloid cells and development of innate and adaptive immunity (Piconese et al., 2008; Bulliard et al., 2014), opening an additional avenue for cancer therapy.

Although we verified two subtypes of HCC and validated the reliability of the novel predictive risk score model of seven pyroptosis genes and analyzed their functions in HCC progression, our study has serval limitations. This bioinformatic study needs to be tested further by experimental validation. Therefore, further laboratory experiments are required, including larger sample multicenter studies, especially studying the relationship between pyroptosis-related genes signature and immune activity. Compared with other traditional clinical characteristics, our risk score model is a better independent prognostic indicator. Thus, this novel risk model could serve as the prognostic predictor and provide clues for personalized immunotherapy for HCC patients.

CONCLUSION

The novel pyroptosis-related genes signature can predict the prognosis of patients with HCC and insight into new cell death targeted therapies.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SQZ designed and analyzed the research study; SQZ, XL, and SJZ wrote and revised the manuscript, XL and XZ collected and analyzed the data, and all authors have read and approved the manuscript.

FUNDING

This study was supported by Discipline construction project of Guangdong Medical University (No. 4SG21009G).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.781427/full#supplementary-material

REFERENCES

Alves Costa Silva, C., Facchinetti, F., Routy, B., and Derosa, L. (2020). New Pathways in Immune Stimulation: Targeting OX40. *ESMO Open* 5, e000573. doi:10.1136/esmoopen-2019-000573

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18, 220. doi:10.1186/s13059-017-1349-1

Berghe, T. V., Linkermann, A., Jouan-Lanhouet, S., Walczak, H., and Vandenabeele, P. (2014). Regulated Necrosis: the Expanding Network of Non-apoptotic Cell Death Pathways. *Nat. Rev. Mol. Cel Biol* 15, 135–147. doi:10.1038/nrm3737

Bergsbaken, T., Fink, S. L., and Cookson, B. T. (2009). Pyroptosis: Host Cell Death and Inflammation. *Nat. Rev. Microbiol.* 7, 99–109. doi:10.1038/nrmicro2070

Blasco, M. T., and Gomis, R. R. (2020). PD-L1 Controls Cancer Pyroptosis. *Nat. Cel Biol* 22, 1157–1159. doi:10.1038/s41556-020-00582-w

Bulliard, Y., Jolicoeur, R., Zhang, J., Dranoff, G., Wilson, N. S., and Brogdon, J. L. (2014). OX40 Engagement Depletes Intratumoral Tregs via Activating FcγRs, Leading to Antitumor Efficacy. *Immunol. Cel Biol* 92, 475–480. doi:10.1038/icb.2014.26

Burocchi, A., Pittoni, P., Gorzanelli, A., Colombo, M. P., and Piconese, S. (2011). Intratumor OX40 Stimulation Inhibits IRF1 Expression and IL-10 Production by Treg Cells while Enhancing CD40L Expression by Effector Memory T Cells. *Eur. J. Immunol.* 41, 3615–3626. doi:10.1002/eji.201141700

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cel Rep.* 18, 248–262. doi:10.1016/j.celrep.2016.12.019

Chen, W., Quan, Y., Fan, S., Wang, H., Liang, J., Huang, L., et al. (2020). Exosome-transmitted Circular RNA Hsa_circ_0051443 Suppresses Hepatocellular Carcinoma Progression. *Cancer Lett.* 475, 119–128. doi:10.1016/j.canlet.2020.01.022

Chiu, D. K.-C., Yuen, V. W.-H., Cheu, J. W.-S., Wei, L. L., Ting, V., Fehlings, M., et al. (2020). Hepatocellular Carcinoma Cells Up-Regulate PVRL1, Stabilizing PVR and Inhibiting the Cytotoxic T-Cell Response via TIGIT to Mediate Tumor Resistance to PD1 Inhibitors in Mice. *Gastroenterology* 159, 609–623. doi:10.1053/j.gastro.2020.03.074

Conrad, M., Angeli, J. P. F., Vandenabeele, P., and Stockwell, B. R. (2016). Regulated Necrosis: Disease Relevance and Therapeutic Opportunities. *Nat. Rev. Drug Discov.* 15, 348–366. doi:10.1038/nrd.2015.6

Cookson, B. T., and Brennan, M. A. (2001). Pro-inflammatory Programmed Cell Death. *Trends Microbiol.* 9, 113–114. doi:10.1016/s0966-842x(00)01936-3

Derangère, V., Chevriaux, A., Courtaut, F., Bruchard, M., Berger, H., Chalmin, F., et al. (2014). Liver X Receptor β Activation Induces Pyroptosis of Human and Murine colon Cancer Cells. *Cell Death Differ* 21, 1914–1924. doi:10.1038/cdd.2014.117

Du, T., Gao, J., Li, P., Wang, Y., Qi, Q., Liu, X., et al. (2021). Pyroptosis, Metabolism, and Tumor Immune Microenvironment. *Clin. Translational Med.* 11, e492. doi:10.1002/ctm2.492

Fang, Y., Tian, S., Pan, Y., Li, W., Wang, Q., Tang, Y., et al. (2020). Pyroptosis: A New Frontier in Cancer. *Biomed. Pharmacother.* 121, 109595. doi:10.1016/j.biopha.2019.109595

Garg, A. D., and Agostinis, P. (2017). Cell Death and Immunity in Cancer: From Danger Signals to Mimicry of Pathogen Defense Responses. *Immunol. Rev.* 280, 126–148. doi:10.1111/imr.12574

Ge, Z., Peppelenbosch, M. P., Sprengers, D., and Kwekkeboom, J. (2021). TIGIT, the Next Step towards Successful Combination Immune Checkpoint Therapy in Cancer. *Front. Immunol.* 12, 699895. doi:10.3389/fimmu.2021.699895

Han, Q., Lv, L., Wei, J., Lei, X., Lin, H., Li, G., et al. (2019). Vps4A Mediates the Localization and Exosome Release of β-catenin to Inhibit Epithelial-Mesenchymal Transition in Hepatocellular Carcinoma. *Cancer Lett.* 457, 47–59. doi:10.1016/j.canlet.2019.04.035

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144, 646–674. doi:10.1016/j.cell.2011.02.013

Hata, A. N., Yeo, A., Faber, A. C., Lifshits, E., Chen, Z., Cheng, K. A., et al. (2014). Failure to Induce Apoptosis via BCL-2 Family Proteins Underlies Lack of Efficacy of Combined MEK and PI3K Inhibitors for KRAS-Mutant Lung Cancers. *Cancer Res.* 74, 3146–3156. doi:10.1158/0008-5472.CAN-13-3728

Holohan, C., Van Schaeybroeck, S., Longley, D. B., and Johnston, P. G. (2013). Cancer Drug Resistance: an Evolving Paradigm. *Nat. Rev. Cancer* 13, 714–726. doi:10.1038/nrc3599

Hou, J., Zhao, R., Xia, W., Chang, C.-W., You, Y., Hsu, J.-M., et al. (2020). PD-L1-mediated Gasdermin C Expression Switches Apoptosis to Pyroptosis in Cancer Cells and Facilitates Tumour Necrosis. *Nat. Cel Biol* 22, 1264–1275. doi:10.1038/s41556-020-0575-z

Jiang, Z., Yao, L., Ma, H., Xu, P., Li, Z., Guo, M., et al. (2017). miRNA-214 Inhibits Cellular Proliferation and Migration in Glioma Cells Targeting Caspase 1 Involved in Pyroptosis. *Oncol. Res.* 25, 1009–1019. doi:10.3727/096504016X14813859905646

Kim, E., and Viatour, P. (2020). Hepatocellular Carcinoma: Old Friends and New Tricks. *Exp. Mol. Med.* 52, 1898–1907. doi:10.1038/s12276-020-00527-1

Kovacs, S. B., and Miao, E. A. (2017). Gasdermins: Effectors of Pyroptosis. *Trends Cel Biol.* 27, 673–684. doi:10.1016/j.tcb.2017.05.005

Lebeaupin, C., Proics, E., De Bieville, C. H. D., Rousseau, D., Bonnafous, S., Patouraux, S., et al. (2015). ER Stress Induces NLRP3 Inflammasome Activation and Hepatocyte Death. *Cell Death Dis* 6, e1879. doi:10.1038/cddis.2015.248

Lepletier, A., Madore, J., O'donnell, J. S., Johnston, R. L., Li, X.-Y., Mcdonald, E., et al. (2020). Tumor CD155 Expression Is Associated with Resistance to Anti-PD1 Immunotherapy in Metastatic Melanoma. *Clin. Cancer Res.* 26, 3671–3681. doi:10.1158/1078-0432.CCR-19-3925

Li, L., Jiang, M., Qi, L., Wu, Y., Song, D., Gan, J., et al. (2021). Pyroptosis, a New Bridge to Tumor Immunity. *Cancer Sci.* 112, 3979–3994. doi:10.1111/cas.15059

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77, e108–e110. doi:10.1158/0008-5472.CAN-17-0307

Li, X.-Y., Das, I., Lepletier, A., Addala, V., Bald, T., Stannard, K., et al. (2018). CD155 Loss Enhances Tumor Suppression via Combined Host and Tumor-Intrinsic Mechanisms. *J. Clin. Invest.* 128, 2613–2625. doi:10.1172/JCI98769

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cel Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004

Llovet, J. M., Ricci, S., Mazzaferro, V., Hilgard, P., Gane, E., Blanc, J.-F., et al. (2008). Sorafenib in Advanced Hepatocellular Carcinoma. *N. Engl. J. Med.* 359, 378–390. doi:10.1056/NEJMoa0708857

Lu, H., Zhang, S., Wu, J., Chen, M., Cai, M.-C., Fu, Y., et al. (2018). Molecular Targeted Therapies Elicit Concurrent Apoptotic and GSDME-dependent Pyroptotic Tumor Cell Death. *Clin. Cancer Res.* 24, 6066–6077. doi:10.1158/1078-0432.CCR-18-1478

Minton, K. (2020). Pyroptosis Heats Tumour Immunity. *Nat. Rev. Immunol.* 20, 274–275. doi:10.1038/s41577-020-0297-2

Mittal, S., and El-Serag, H. B. (2013). Epidemiology of Hepatocellular Carcinoma. *J. Clin. Gastroenterol.* 47 (Suppl. l), S2–S6. doi:10.1097/MCG.0b013e3182872f29

Miura, K., Kodama, Y., Inokuchi, S., Schnabl, B., Aoyama, T., Ohnishi, H., et al. (2010). Toll-Like Receptor 9 Promotes Steatohepatitis by Induction of Interleukin-1β in Mice. *Gastroenterology* 139, 323–334. doi:10.1053/j.gastro.2010.03.052

Moon, H., and Ro, S. W. (2021). MAPK/ERK Signaling Pathway in Hepatocellular Carcinoma. *Cancers* 13, 3026. doi:10.3390/cancers13123026

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

Ng, K. P., Hillmer, A. M., Chuah, C. T. H., Juan, W. C., Ko, T. K., Teo, A. S. M., et al. (2012). A Common BIM Deletion Polymorphism Mediates Intrinsic Resistance and Inferior Responses to Tyrosine Kinase Inhibitors in Cancer. *Nat. Med.* 18, 521–528. doi:10.1038/nm.2713

Petrick, J. L., Kelly, S. P., Altekruse, S. F., Mcglynn, K. A., and Rosenberg, P. S. (2016). Future of Hepatocellular Carcinoma Incidence in the United States Forecast through 2030. *Jco* 34, 1787–1794. doi:10.1200/JCO.2015.64.7412

Piconese, S., Valzasina, B., and Colombo, M. P. (2008). OX40 Triggering Blocks Suppression by Regulatory T Cells and Facilitates Tumor Rejection. *J. Exp. Med.* 205, 825–839. doi:10.1084/jem.20071341

Plattner, C., Finotello, F., and Rieder, D. (2020). Deconvoluting Tumor-Infiltrating Immune Cells from RNA-Seq Data Using quanTIseq. *Methods Enzymol.* 636, 261–285. doi:10.1016/bs.mie.2019.05.056

Racle, J., De Jonge, K., Baumgaertner, P., Speiser, D. E., and Gfeller, D. (2017). Simultaneous Enumeration of Cancer and Immune Cell Types from Bulk Tumor Gene Expression Data. *Elife* 6, e26476. doi:10.7554/eLife.26476

Reck, M., Schenker, M., Lee, K. H., Provencio, M., Nishio, M., Lesniewski-Kmak, K., et al. (2019). Nivolumab Plus Ipilimumab versus Chemotherapy as First-Line Treatment in Advanced Non-small-cell Lung Cancer with High Tumour Mutational burden: Patient-Reported Outcomes Results from the Randomised, Open-Label, Phase III CheckMate 227 Trial. *Eur. J. Cancer* 116, 137–147. doi:10.1016/j.ejca.2019.05.008

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell* 160, 48–61. doi:10.1016/j.cell.2014.12.033

Ruan, J., Wang, S., and Wang, J. (2020). Mechanism and Regulation of Pyroptosis-Mediated in Cancer Cell Death. *Chemico-Biological Interactions* 323, 109052. doi:10.1016/j.cbi.2020.109052

Shao, W., Yang, Z., Fu, Y., Zheng, L., Liu, F., Chai, L., et al. (2021). The Pyroptosis-Related Signature Predicts Prognosis and Indicates Immune Microenvironment Infiltration in Gastric Cancer. *Front. Cel Dev. Biol.* 9, 676485. doi:10.3389/fcell.2021.676485

Shi, J., Jiang, D., Yang, S., Zhang, X., Wang, J., Liu, Y., et al. (2020). LPAR1, Correlated with Immune Infiltrates, Is a Potential Prognostic Biomarker in Prostate Cancer. *Front. Oncol.* 10, 846. doi:10.3389/fonc.2020.00846

Stanietsky, N., Simic, H., Arapovic, J., Toporik, A., Levy, O., Novik, A., et al. (2009). The Interaction of TIGIT with PVR and PVRL2 Inhibits Human NK Cell Cytotoxicity. *Proc. Natl. Acad. Sci.* 106, 17858–17863. doi:10.1073/pnas.0903474106

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

Tang, W., Zhou, Y., Sun, D., Dong, L., Xia, J., and Yang, B. (2019). Oncogenic Role of Phospholipase C-γ1 in Progression of Hepatocellular Carcinoma. *Hepatol. Res.* 49, 559–569. doi:10.1111/hepr.13309

Tsuchiya, K. (2020). Inflammasome-associated Cell Death: Pyroptosis, Apoptosis, and Physiological Implications. *Microbiol. Immunol.* 64, 252–269. doi:10.1111/1348-0421.12771

Tsuchiya, K. (2021). Switching from Apoptosis to Pyroptosis: Gasdermin-Elicited Inflammation and Antitumor Immunity. *Ijms* 22, 426. doi:10.3390/ijms22010426

Van Opdenbosch, N., and Lamkanfi, M. (2019). Caspases in Cell Death, Inflammation, and Disease. *Immunity* 50, 1352–1364. doi:10.1016/j.immuni.2019.05.020

Vickers, A. J., Cronin, A. M., Elkin, E. B., and Gonen, M. (2008). Extensions to Decision Curve Analysis, a Novel Method for Evaluating Diagnostic Tests, Prediction Models and Molecular Markers. *BMC Med. Inform. Decis. Mak* 8, 53. doi:10.1186/1472-6947-8-53

Vietri, M., Radulovic, M., and Stenmark, H. (2020). The many Functions of ESCRTs. *Nat. Rev. Mol. Cel Biol* 21, 25–42. doi:10.1038/s41580-019-0177-4

Wallach, D., Kang, T.-B., Dillon, C. P., and Green, D. R. (2016). Programmed Necrosis in Inflammation: Toward Identification of the Effector Molecules. *Science* 352, aaf2154. doi:10.1126/science.aaf2154

Wang, L., Sebra, R. P., Sfakianos, J. P., Allette, K., Wang, W., Yoo, S., et al. (2020). A Reference Profile-free Deconvolution Method to Infer Cancer Cell-Intrinsic Subtypes and Tumor-type-specific Stromal Profiles. *Genome Med.* 12, 24. doi:10.1186/s13073-020-0720-0

Wang, Q., Wang, C., and Chen, J. (2018). NLRP6, Decreased in Gastric Cancer, Suppresses Tumorigenicity of Gastric Cancer Cells. *Cmar* 10,6431–6444. doi:10.2147/CMAR.S182980

Wang, Q., Wang, Y., Ding, J., Wang, C., Zhou, X., Gao, W., et al. (2020). A Bioorthogonal System Reveals Antitumour Immune Function of Pyroptosis. *Nature* 579, 421–426. doi:10.1038/s41586-020-2079-1

Wang, Y.-D., Cai, N., Wu, X.-L., Cao, H.-Z., Xie, L.-L., and Zheng, P.-S. (2013). OCT4 Promotes Tumorigenesis and Inhibits Apoptosis of Cervical Cancer Cells by miR-125b/BAK1 Pathway. *Cel Death Dis* 4, e760. doi:10.1038/cddis.2013.272

Wang, Y., Yin, B., Li, D., Wang, G., Han, X., and Sun, X. (2018). GSDME Mediates Caspase-3-dependent Pyroptosis in Gastric Cancer. *Biochem. Biophysical Res. Commun.* 495, 1418–1425. doi:10.1016/j.bbrc.2017.11.156

Wei, Q., Zhu, R., Zhu, J., Zhao, R., and Li, M. (2019). E2-Induced Activation of the NLRP3 Inflammasome Triggers Pyroptosis and Inhibits Autophagy in HCC Cells. *Oncol. Res.* 27, 827–834. doi:10.3727/096504018X15462920753012

Wree, A., Eguchi, A., Mcgeough, M. D., Pena, C. A., Johnson, C. D., Canbay, A., et al. (2014). NLRP3 Inflammasome Activation Results in Hepatocyte Pyroptosis, Liver Inflammation, and Fibrosis in Mice. *Hepatology* 59, 898–910. doi:10.1002/hep.26592

Xia, X., Wang, X., Cheng, Z., Qin, W., Lei, L., Jiang, J., et al. (2019). The Role of Pyroptosis in Cancer: Pro-cancer or Pro-"host"? *Cel Death Dis* 10, 650. doi:10.1038/s41419-019-1883-8

Xu, D., Ji, Z., and Qiang, L. (2021). Molecular Characteristics, Clinical Implication, and Cancer Immunity Interactions of Pyroptosis-Related Genes in Breast Cancer. *Front. Med.* 8, 702638. doi:10.3389/fmed.2021.702638

Ye, Y., Dai, Q., and Qi, H. (2021). A Novel Defined Pyroptosis-Related Gene Signature for Predicting the Prognosis of Ovarian Cancer. *Cell Death Discov.* 7, 71. doi:10.1038/s41420-021-00451-x

Yu, P., Zhang, X., Liu, N., Tang, L., Peng, C., and Chen, X. (2021). Pyroptosis: Mechanisms and Diseases. *Sig Transduct Target. Ther.* 6, 128. doi:10.1038/s41392-021-00507-5

Zhang, T., Li, Y., Zhu, R., Song, P., Wei, Y., Liang, T., et al. (2019). Transcription Factor P53 Suppresses Tumor Growth by Prompting Pyroptosis in Non-small-cell Lung Cancer. *Oxidative Med. Cell Longevity* 2019, 1–9. doi:10.1155/2019/8746895

Zhang, X., Xiao, X., Lan, P., Li, J., Dou, Y., Chen, W., et al. (2018). OX40 Costimulation Inhibits Foxp3 Expression and Treg Induction via BATF3-dependent and Independent Mechanisms. *Cel Rep.* 24, 607–618. doi:10.1016/j.celrep.2018.06.052

Zhou, Y., Hu, L., Tang, W., Li, D., Ma, L., Liu, H., et al. (2021). Hepatic NOD2 Promotes Hepatocarcinogenesis via a RIP2-Mediated Proinflammatory Response and a Novel Nuclear Autophagy-Mediated DNA Damage Mechanism. *J. Hematol. Oncol.* 14, 9. doi:10.1186/s13045-020-01028-4

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 06 January 2022
doi: 10.3389/fgene.2021.783391

Invasive Prediction of Ground Glass Nodule Based on Clinical Characteristics and Radiomics Feature

Hui Zheng, Hanfei Zhang, Shan Wang, Feng Xiao and Meiyan Liao**

Zhongnan Hospital, Wuhan University, Wuhan, China

Objective: To explore the diagnostic value of CT radiographic images and radiomics features for invasive classification of lung adenocarcinoma manifesting as ground-glass nodules (GGNs) in computer tomography (CT).

Methods: A total of 312 GGNs were enrolled in this retrospective study. All GGNs were randomly divided into training set (n = 219) and test set (n = 93). Univariate and multivariate logistic regressions were used to establish a clinical model, while the minimum redundancy maximum relevance (mRMR) and least absolute shrinkage and selection operator (LASSO) algorithm were used to select the radiomics features and construct the radiomics model. A combined model was finally built by combining these two models. The performance of these models was assessed in both training and test set. A combined nomogram was developed based on the combined model and evaluated with its calibration curves and C-index.

Results: Diameter [odds ratio (OR), 1.159; $p < 0.001$], lobulation (OR, 2.953; $p = 0.002$), and vascular changes (OR, 3.431; $p < 0.001$) were retained as independent predictors of the invasive adenocarcinoma (IAC) group. Eleven radiomics features were selected by mRMR and LASSO method to established radiomics model. The clinical model and radiomics mode showed good predictive ability in both training set and test set. When two models were combined, the diagnostic area under the curve (AUC) value was higher than the single clinical or radiomics model (training set: 0.86 vs. 0.83 vs. 0.82; test set: 0.80 vs. 0.78 vs. 0.79). The constructed combined nomogram could effectively quantify the risk degree of 3 image features and Rad score with a C-index of 0.855 (95%: 0.805~0.905).

Conclusion: Radiographic and radiomics features show high accuracy in the invasive diagnosis of GGNs, and their combined analysis can improve the diagnostic efficacy of IAC manifesting as GGNs. The nomogram, serving as a noninvasive and accurate predictive tool, can help judge the invasiveness of GGNs prior to surgery and assist clinicians in creating personalized treatment strategies.

Keywords: ground glass nodules (GGNs), Radiomics, Adenocarcinoma, computed tomography, Diagnostic model

OPEN ACCESS

Edited by:
Lin Hua,
ICAR Indian Institute of Soybean Research, India

Reviewed by:
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Yuqing Mao,
National Institutes of Health (NIH), United States

****Correspondence:***
Meiyan Liao
Liaomy@whu.edu.cn
Feng Xiao
seiya_0731@163.com

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

Received: *26 September 2021*
Accepted: *01 December 2021*
Published: *06 January 2022*

Citation:
Zheng H, Zhang H, Wang S, Xiao F and Liao M (2022) Invasive Prediction of Ground Glass Nodule Based on Clinical Characteristics and Radiomics Feature. Front. Genet. 12:783391. doi: 10.3389/fgene.2021.783391

INTRODUCTION

Lung cancer is the major cancer leading in cancer-related deaths, and imaging played an important role in diagnosis and treatment. With the popularity of computed tomography (CT) and artificial intelligence (AI), the discovery of lung cancer manifesting as ground-glass nodules (GGNs) increased sequentially during the process of CT screening. Early detection, follow-up, and timely intervention were of positive significance for GGNs. No doubt, these findings deserved the attention of society, medical professionals, and the general public.

GGNs could be divided into pure ground-glass nodules (pGGNs) and mixed ground-glass nodules (mGGNs) according to the presence of the solid composition. At present, the development mechanism of GGNs was not clear. GGNs may exist in various pathological entities, including tumor, inflammation, focal hemorrhage, and focal interstitial fibrosis (Park et al., 2007). Although GGN was in nonspecific radiologic findings, persistent GGN was more likely to be malignant. Studies had shown that 20% of pGGNs and 40% of mGGNs increase gradually or show a trend of increasing solid composition (Kobayashi et al., 2018). However, the GGN growth was slow and the process of deterioration may take several years, which was why multiple current guidelines recommend longer follow-up times.

Surgical resection was the most effective method for GGN treatment. Preinvasive lesions and minimally invasive adenocarcinoma (MIA) could also be well treated by lobectomy (wedge resection or segmental resection), with a 5-year disease-free survival rate of 100%. It was necessary to analyze the imaging characteristics of each pathological subtype before operation and to judge the infiltrability of the GGN.

Earlier studies had paid more attention to GGN imaging features, such as size, consolidation, and morphological characteristics. Medical imaging technology had been developing in recent years, and its use in clinical oncology had expanded from the initial diagnostic tools to personalized treatment and management tools. Artificial intelligence and radiomics diagnosis were widely concerned. Radiomics referred to the automatic extraction of a large number of quantitative features from medical images by computer software and the use of statistical methods to screen and establish diagnosis related to the results. The radiomics model showed good sensitivity and specificity in tumor pathological type discrimination and invasive judgment.

The aim of this study was to explore the diagnostic value of imaging features and radiomics features in the invasive diagnosis of lung adenocarcinoma manifested as GGN, so as to assist clinical diagnosis and treatment.

MATERIALS AND METHODS

Patients

This retrospective study was approved by the corresponding institutional review board (grant: 2021057), and the patients' informed consent was waived. Clinical data and chest CT image of resected GGN between July 2017 and December 2020 at Zhongnan Hospital of Wuhan University were retrospectively collected. A total of 291 patients with 312 GGNs were enrolled in this retrospective study. The inclusion criteria were as follows: (1) the nodules showed as GGN at lung window setting (width 1500 HU; level is −700 HU), image thickness ≤1.25 mm; 2) maximum diameter of nodules measured on lung windows <30 mm; 3) accurate surgical and pathological results must be obtained. Exclusion criteria were as follows: 1) incomplete chest CT image, heavy artifacts or poor quality; 2) GGN who have no pathological results or perform only a biopsy without surgery.

Data Flowchart

As seen in **Figure 1**, the data processing of this study could be divided into three parts. The first (**Figure 1A**) is the clinical

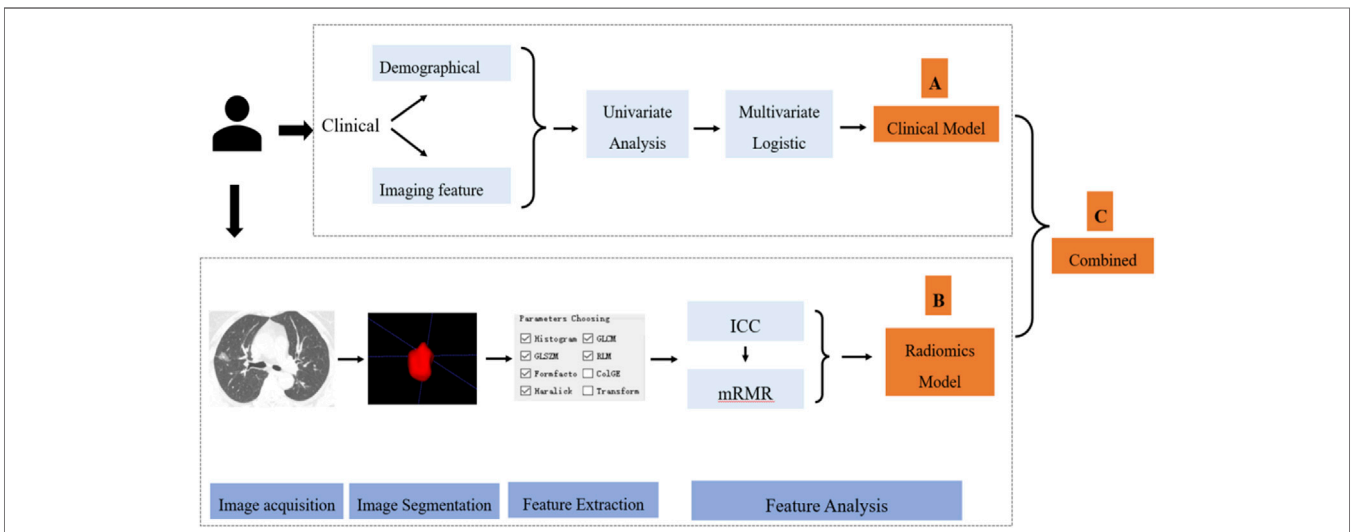


FIGURE 1 | Flowchart of clinical and radiomics feature analysis.

TABLE 1 | Summary of radiomics features used in this study.

Feature classes	No. of features	3 representative features
Histogram	42	FrequencySize, MaxIntensity, MeanValue,...
GLCM	144	ClusterProminence, ClusterShade, Correlation,...
GLSZM	11	SizeZoneVariability, HighIntensityEmphasis, IntensityVariability,...
RLM	180	GreyLevelNonuniformity, HighGreyLevelRunEmphasis, LongRunEmphasis,...
Formfactor	15	Compactness1, Maximum3DDiameter, Sphericity,...
Haralick	10	HaraEntroy, contrast, differenceEntropy,...
Total	402	

GLCM, gray-level co-occurrence matrix; GLSZM, gray-level size zone matrix; RLM, gray-level run-length matrix.

TABLE 2 | Clinical characteristics of GGNs.

Characteristics	Number
Sex	
Male	103 (33.0%)
Female	209 (67.0%)
Age, year	58 (50–65)
Pathological subtype	
Benign	25 (8.0%)
AAH	12 (3.8%)
AIS	20 (6.4%)
MIA	74 (23.7%)
IAC	181 (58.0%)
EGFR mutation (*n* = 30)	
Mutation in exon 21	12 (40.0%)
Mutation in exon 19	10 (33.3%)
Wild type	8 (26.7%)
Preoperative position (*n* = 75)	
Pneumothorax	29 (38.6%)
Hemorrhage	32 (42.7%)
Without complications	14 (18.7%)
Interoperative biopsy (*n* = 197)	
Misdiagnosis	7 (3.6%)
Underestimate the infiltration	20 (10.1%)

AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma in situ*; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma.*

characteristic analysis and modeling. Clinical characteristic analysis contained univariate logistic regression and multivariate logistic regression step by step; two types of characteristics (demographics and traditional imaging features) were considered in this part. The second (**Figure 1B**) is the image analysis and radiomics modeling, which contained image acquisition, image segmentation, radiomics feature extraction, and modeling step by step. In this study, several most used machine learning models and a deep learning (DL) method were tried and compared, then the most suitable model was selected for radiomics modeling. After the analysis of these two parts, the screened clinical risk factors and constructed radiomics model were combined to construct the combined model and radiomics + clinical nomogram (**Figure 1C**).

All data sets were divided into a training set and a test set according to a 7:3 ratio using the stratified random sampling method, in which the samples were stratified according to different groups of IAC, and then randomly sampled; feature analysis and modeling were performed based on the training set, and the performance of constructed models was validated based on both training and test set.

Clinical Characteristics Analysis and Clinical Modeling

Clinical characteristics contained two types: three demographics (patient sex, age, and operation mode) and 14 traditional imaging features, which were extracted from CT images, including diameter, volume, ratio of consolidation, mean CT value, mass, location, margin, shape, pleural indentation sign, bubble-like lucency, air bronchus sign, vascular change, speculation, and lobulation. A large number of studies (Yang et al., 2018) have confirmed that traditional imaging features play crucial roles in the diagnosis and pathological classification of GGN. The selection of these traditional imaging features was referred to these studies (Yang et al., 2018).

Diameter, mean CT value, volume, and ratio of consolidation were obtained by automatic cutting and calculation according to the Intelligent 4D Imaging System for Chest CT 6.8 (Hangzhou YITU Healthcare Technology Co., Ltd., Hangzhou, China). Mass was an important sign of tumor growth, which could reflect the change of tumor volume and the difference of cell density (Qi et al., 2020). Calculation formula $\text{Mass} = \frac{\text{volume}\times 1000 + (\text{mean CT value})}{1000}$.

Count data were defined as follows. Location: divided into left upper lobe (LUL), left lower lobe (LLL), right upper lobe (RUL), right middle lobe (RML), and right lower lobe (RLL). Margin: a clear demarcation between the lesions and the surrounding lung parenchyma range, more than 75% of the perimeter was defined as clear, otherwise defined as blurred. Pleural indentation sign: linear or small patch between the nodules and the local pleural. Bubble-like lucency: boundary-clear air density or cavity within the nodules. Air bronchus sign: the bronchial shadow was seen in the increased density area. Vascular change: morphological changes of the vessels when passing through the GGN, such as dilatation, stiffness, correction, distortion. Spiculation: fine lines around the nodules point to the lung. Lobulation: the outline of the nodules was raised in multiple arc due to different growth speed.

Two experienced chest radiologists blinded evaluated these CT traditional imaging features independently and resolve the differences through discussion.

Image Processing and Radiomics Modeling

Chest CT scans were performed using a GE Discovery 750HD scanner (GE Medical Systems, Milwaukee, WI, USA) and/or a

TABLE 3 | Univariate analysis of clinical and imaging features in the training and test sets.

Characteristic	Training set (219)		*p*	Test set (93)		*p*
	Non-IAC group (n = 86)	IAC group (*n* = 133)		Non-IAC group (*n* = 43)	IAC group (*n* = 50)	
Male	63 (71.6%)	86 (65.6%)	0.355	26 (60.5%)	34 (68.0%)	0.449
Age, year	57 (49–62)	61 (52–66)	0.009	55 (46–60)	60 (54–65)	0.027
Diameter, mm	11 (8–14)	17 (13–20)	< 0.001	11 (8–15)	17 (14–21)	< 0.001
Volume, mm^3	509 (238–1,047)	1,351 (796–2,639)	< 0.001	552 (248–1,184)	1,517 (816–3,104)	< 0.001
Ratio of consolidation	0.04 (0–0.22)	0.24 (0.10–0.45)	< 0.001	0.04 (0–0.14)	0.28 (0.13–0.54)	< 0.001
Mean CT value, HU	−588 (−660–489)	−442 (−566–361)	< 0.001	−593 (−675–530)	−445 (−553–322)	< 0.001
Mass, mg	199 (104–393)	775 (322–1,352)	< 0.001	256 (101–520)	755 (420–1725)	< 0.001
Location			0.201			0.411
RUL	30 (34.1%)	51 (38.9%)		12 (27.9%)	22 (44.0%)	
RML	2 (2.3%)	10 (7.6%)		4 (9.3%)	3 (6.0%)	
RLL	20 (22.7%)	20 (15.3%)		5 (11.6%)	8 (16.0%)	
LUL	26 (29.5%)	41 (31.3%)		15 (34.9%)	11 (22.0%)	
LLL	10 (11.4%)	9 (6.9%)		7 (16.3%)	6 (12.0%)	
pGGN	47 (53.4%)	32 (24.4%)	< 0.001	24 (55.8%)	9 (18.0%)	< 0.001
Margin			0.106			0.377
Clear	30 (34.1%)	59 (45.0%)		22 (51.2%)	21 (42.0%)	
Unclear	58 (65.9%)	72 (55.0%)		21 (48.8%)	29 (58.0%)	
Shape			< 0.001			0.008
Round or oval	60 (68.2%)	54 (41.2%)		29 (67.4%)	20 (40.0%)	
Irregular	28 (31.8%)	81 (58.8%)		14 (32.6%)	30 (60.0%)	
Pleural indentation sign	27 (30.7%)	74 (56.5%)	< 0.001	15 (34.9%)	22 (44.0%)	0.371
Bubble-like lucency	21 (23.9%)	34 (26.0%)	0.727	6 (14.0%)	14 (28.0%)	0.1
Air bronchus sign	15 (17.0%)	66 (50.4%)	< 0.001	14 (32.6%)	31 (62.0%)	0.005
Spiculation	33 (37.5%)	72 (55.0%)	0.011	13 (30.2%)	28 (56.0%)	0.013
Lobulation	24 (27.3%)	89 (67.9%)	< 0.001	14 (32.6%)	27 (54.0%)	0.038
Vascular change	31 (35.2%)	97 (74.0%)	< 0.001	17 (39.5%)	37 (74.0%)	0.001

LLL, left lower lobe; LUL, left upper lobe; RLL, right lower lobe; RML, right middle lobe; RUL, right upper lobe.

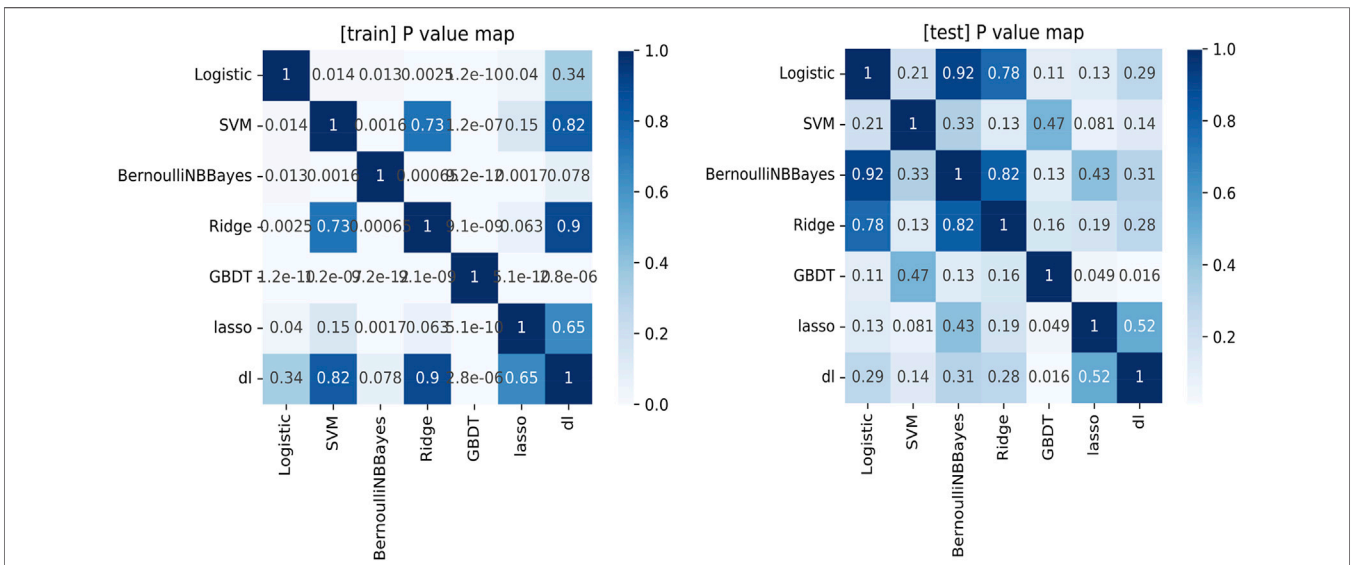


FIGURE 2 | The comparison of the ROC analysis among the machine learning models in the training set and test set.

SOMATOM Definition scanner (Siemens Healthineers, Forchheim, Germany), with a reconstruction slice thickness = 1.25 mm, slice interval = 1.25 mm, matrix size = 512 × 512, tube voltage = 120 kV, and tube current 100–350 mA. All images were then transmitted to the workstation and PACS for post-processing.

Before image analysis, all images were first resampled into the same sampling size (1 mm × 1 mm×1 mm) using the linear

TABLE 4 | Results of the ROC analysis for different machine learning methods.

	Training set			Test set		
	AUC	[0.025	0.975]	AUC	[0.025	0.975]
Logistic	0.806	0.748	0.864	0.776	0.683	0.869
SVM	0.836	0.781	0.891	0.750	0.651	0.849
Bernoulli naive Bayes	0.780	0.718	0.843	0.778	0.685	0.870
Ridge	0.833	0.780	0.887	0.773	0.678	0.867
GBDT	1.000	NaN	NaN	0.702	0.596	0.808
LASSO	0.819	0.763	0.874	0.793	0.702	0.885
DL	0.830	0.776	0.884	0.819	0.732	0.905

SVM, support vector machine; GBDT, gradient boosting decision tree; LASSO, least absolute shrinkage and selection operator; DL, deep learning.

interpolation method. Then, the open-source image analysis software ITK-SNAP (Version 3.6; http://www.itksnap.org) was used for manual segmentation and radiomics analysis was applied to the CT images using in-house software (Artificial Intelligence Kit; GE Healthcare, Chicago, IL, USA). A total of 402 imaging texture features from the category of histogram, the gray-level co-occurrence matrix (GLCM), the gray-level size zone matrix (GLSZM), the gray-level run-length matrix (RLM), and shape- and size-based features were finally extracted from one single image (**Table 1**). The details of each radiomics features are shown in the Appendix.

Another physician repeated the above segmentation and feature extraction steps for the test of feature reliability and reproducibility. The differences between the features generated by reader one and those by reader two (interobserver reliability), as well as the differences between the twice-generated features by reader 1 (intraobserver reproducibility), were all evaluated. Inter- and intraclass correlation coefficients (ICCs) were used to evaluate the agreement of feature extraction. A good agreement was reached when the ICC was greater than 0.8 in this study.

Minimum redundancy maximum relevance (mRMR) was used for feature reduction. Then, several machine learning models and a DL method (detailed in supplemental methods) were tried and compared in the radiomics modeling. The most suitable model was selected as the mathematical model of the radiomics model.

The combined model was constructed using multivariate logistic regression by combining the clinical risk factors with the radiomics model, which was used as an independent risk factor in the combined model. The radiomics + clinical nomogram transformed the combined model into a simple and visual graph, making the results of the prediction model more prominent and of higher clinical use value.

Model Validation

The receiver operating characteristic (ROC) curve-related metrics were employed for the evaluation of model diagnostic abilities. The area under the curve (AUC) and Delong's test were used to evaluate and compare the diagnosis abilities among different machine learning models and the DL method. Six ROC-related metrics, AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (PPV) were used to assess the constructed radiomics and combined models. The relationship between nomogram-predicted probability and actual probability was evaluated by the calibration curve and C-index.

Statistical Analysis

All statistical analyses were performed with SPSS (version 23.0, IBM) and R software (version 4.0.1, Vienna, Austria). Continuous variables with normal distribution were presented as mean ± SD and test by Student's t test. Continuous variables with non-normal distribution were presented as median (interquartile range, IQR) and tested by Mann–Whitney *U* test. The differences of count data between two groups were analyzed by the chi-square test.

RESULTS

Patients Characteristics

A total of 297 patients with 312 GGNs were included in the study; of these, 103 (33%) were male and 209 (67%) were female, and the median age was 58 (IQR: 50–65) years. There were 181 nodules in the IAC group and 131 nodules in the non-IAC group (25 benign lesions, 12 AAH, 20 AIS, 74 MIA). Detailed clinical information of patients is summarized in **Table 2**.

Clinical Analysis and Modeling

In the training set, the univariate analysis showed that multiple clinical parameters were larger in IAC groups (**Table 3**), including diameter (17 vs. 11 mm, $p < 0.001$), volume (1,351 vs. 509 mm^3, $p < 0.001$), ratio of consolidation (0.24 vs. 0.04, $p < 0.001$), mean CT value (−442 vs. −588 HU, $p < 0.001$), and mass (775 vs. 199 mg, $p < 0.001$). The IAC group had less pGGN and was easier to exhibit an irregular shape, pleural indentation sign, air bronchus sign, spiculation, lobulation, and vascular changes ($p < 0.05$).

The clinical model was built using multivariable logistic regression, where diameter [odds ratio (OR), 1.159; $p < 0.001$], lobulation (OR, 2.953; $p = 0.002$), and vascular changes (OR, 3.431; $p < 0.001$) were identified as independent risk factors. The AUC of the clinical model in the training set and the test set was 0.83 and 0.78, respectively.

Comparison of Diagnosis Efficacy for Different Methods

As shown in **Figure 2** and **Table 4**, we found that for both training and test sets, DL models showed the best diagnostic performance. However, the difference between it and other models was not significant, except for the GBDT model (obvious overfitting). The diagnostic ability of LASSO was the second highest in the test set, but similarly their difference was not significant. The LASSO model was a linear regression method using L1 regularization, which could make the learned weights of some features 0, so as to achieve the purpose of feature sparseness

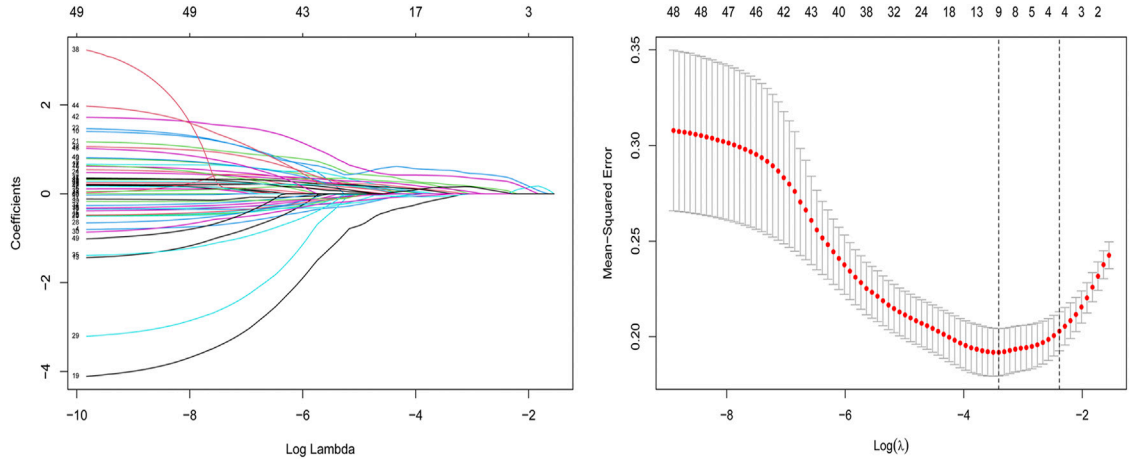


FIGURE 3 | Feature selection for the LASSO logistic regression. The selection of the tuning parameter (λ) using a 10-fold cross-validation. At minimal value of the mean square error of the classification, the dotted vertical line (λ = 0.033) was drawn, including 11 optimal features with nonzero coefficients. The histogram of 11 radiomics features was presented.

and selection. Considering that its model structure is simple and not easy to overfit with a strong clinical interpretability, we choose LASSO as the mathematical model of the radiomics model for this study.

Radiomics Analysis and Modeling

After ICC analysis, 217 variables were retained and included in mRMR and LASSO analysis. Finally, 11 optimal features with nonzero coefficients were selected to establish a radiomics model (**Figure 3** and **Table 5**). The radiomics model had AUC values of 0.82 and 0.79 in the training set and the test set, respectively.

Nomogram and Calibration Curve of IAC Manifested as GGNs

A logistic regression analysis identified the diameter, lobulation, vascular change, and Rad score as independent predictors, which were incorporated to develop an individualized prediction nomogram (**Figure 4**). The calibration curve showed a high consistency between predicted probability and observed probability, and a c-index of 0.855 (95%: 0.805–0.905).

TABLE 5 | 11 features selected by the LASSO method.

Index	Coefficients
InverseDifferenceMoment_AllDirection_offset1_SD	0.145
ShortRunEmphasis_angle135_offset1	0.119
GLCMEnergy_angle0_offset4	−0.102
MinorAxisLength	0.218
ShortRunHighGreyLevelEmphasis_angle45_offset7	0.533
RunLengthNonuniformity_AllDirection_offset1_SD	−0.016
kurtosis	−0.053
GLCMEntropy_angle45_offset7	0.406
Percentile35	0.028
HighIntensityEmphasis	0.061
HaralickCorrelation_angle45_offset1	0.149

Clinical Use of the Nomogram

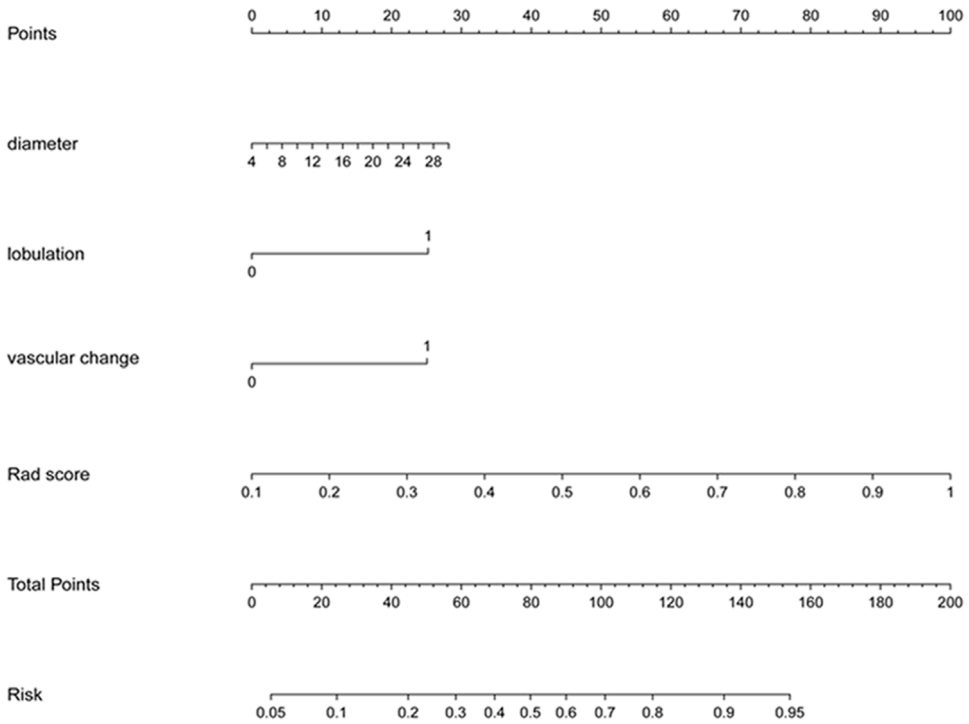
Figure 5 and **Figure 6** showed the important value of the nomogram for GGN diagnosis. The total score was calculated based on Rad score and the imaging performance of the lesion including diameter, presence of lobulation, and vascular change. Finally, the corresponding total score indicated the probability of IAC. In **Figure 5**, the nodule showed a low IAC risk probability of 0.249, and the final pathological was confirmed as AAH. **Figure 6** showed a GGN with high IAC risk probability of 0.943, and the final pathology result was consistent with the prediction of the nomogram.

Comparison of Diagnosis Efficiency Between Clinical Model and Radiomics Model

Delong's tests showed that the performance of the combined model was significantly better than that of a single clinical or radiomics model in the training set (clinical vs. combined, 0.83 vs. 0.86, p = 0.032; radiomics vs. combined, 0.82 vs. 0.86, p = 0.031). In the test set, there were no significant differences in ROC analysis for the three models. The diagnostic performances of the clinical model, radiomics model, and combined model are shown in **Table 6** and **Figure 7**.

DISCUSSION

In this study, we established a clinical model and radiomics models by analyzing the imaging and radiomics characteristics of GGN and compared the diagnostic values of different models to provide a highly effective GGN diagnostic tool for the clinical diagnosis. The results showed that the diagnostic accuracy of the clinical model and the radiomics model was similar to the combined model, but the AUC value increased when the clinical and radiomics models were combined. This suggested that radiomics analysis could also be a tool for clinical diagnosis.



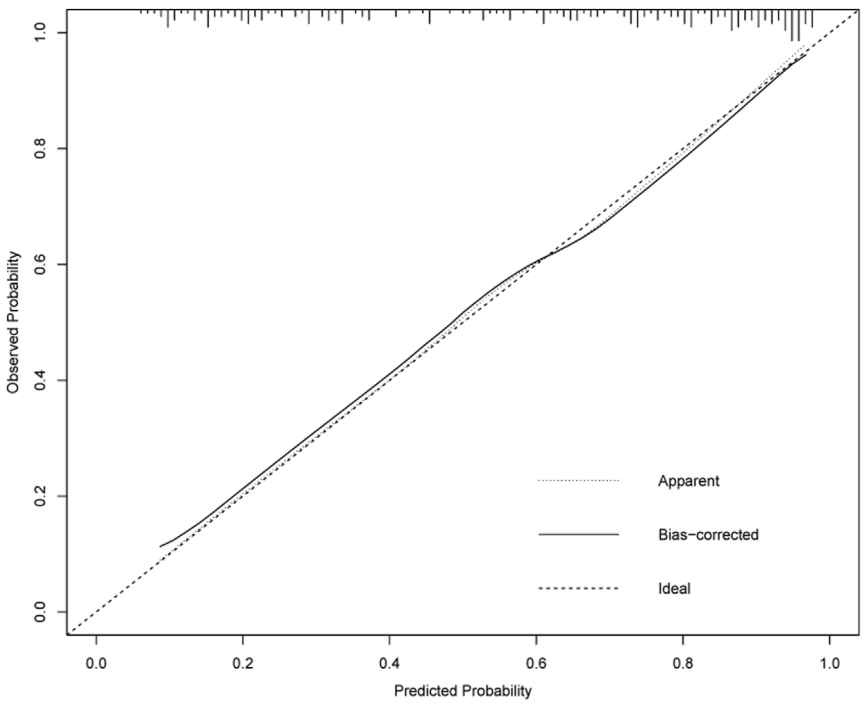


FIGURE 4 | A radiomics-based nomogram was developed in the training set. The radiomics-based nomogram was developed in the training set, and the diameter, lobulation, vascular change, and Rad score were incorporated. The total score was calculated by adding the score for each risk factor, and then the probability of IAC was predicted on the risk axis.

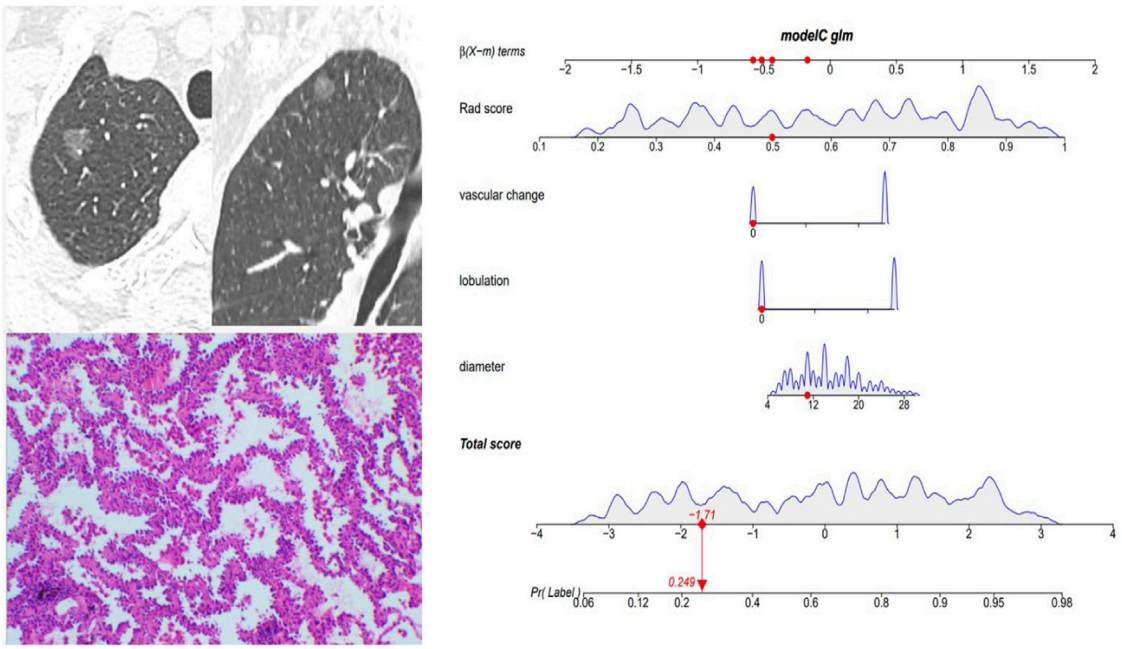


FIGURE 5 | Female, 57 years old; CT showed a pGGN of 11 mm in the right upper lobe, with no significant lobulation and vascular change, and the Rad score of pGGN was 0.491. Interactive nomogram showing that the IAC risk probability of this nodule was 0.249. The case was confirmed as AAH.

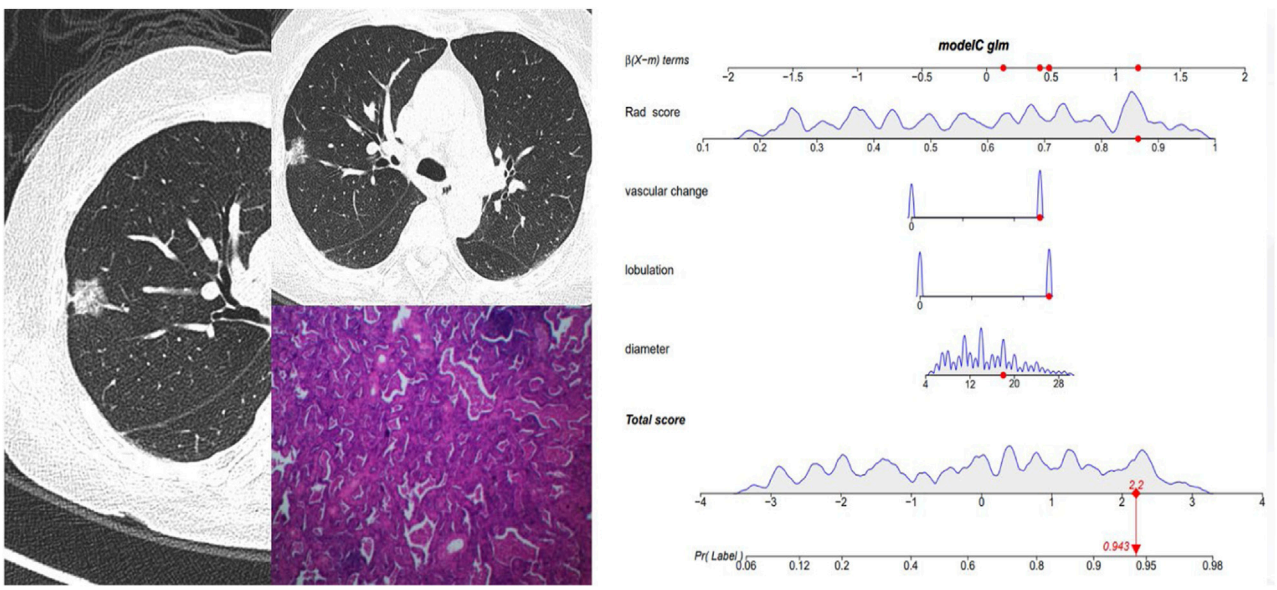


FIGURE 6 | Female, 62 years old. CT showed a mGGN of 18 mm in the right upper lobe, with significant lobulation and vascular change, and the Rad score of pGGN was 0.864. The interactive nomogram showing the IAC risk probability of this nodule was 0.943. The case was confirmed as IAC.

Not surprisingly in the selection process of different machine learning and DL models, the DL method obtained the highest diagnostic efficacy, which was due to its deep excavation of information of many high-dimensional and complex image features. Highly intelligent and automated processing data using the DL network were the mainstream direction of artificial intelligence in the future, and medical image analysis is its important application field. However, how to combine them organically is still a problem. For example, in this study, to conduct a personalized evaluation with strong clinical interpretability and high availability, we hope that the model is simple with easily understood image features. At the same time,

TABLE 6 | Comparison of diagnosis efficiency between clinical model and radiomics model.

	AUC	Sensitivity	Specificity	PPV	NPV	Accuracy
Training set						
Clinical	0.83	0.76	0.83	0.87	0.70	0.79
Radiomics	0.82	0.77	0.75	0.82	0.69	0.76
Combined	0.86	0.81	0.77	0.84	0.73	0.79
Test set						
Clinical	0.78	0.72	0.79	0.80	0.71	0.75
Radiomics	0.79	0.64	0.88	0.86	0.68	0.75
Combined	0.80	0.68	0.84	0.83	0.69	0.75

AUC, area under the curve; PPV, positive predictive value; NPV, negative predictive value.

we hope that the diagnostic efficacy of the model can be as high as possible. This is a contradiction in the modeling process, which is why we finally chose LASSO as the mathematical model.

Traditional imaging feature analysis found that diameter, lobulation, and vascular changes were independent risk factors for predicting IAC. In previous studies, several imaging characteristics were related to GGN. A meta-analysis (Dai et al., 2018) showed the limited diagnostic efficacy of single-image features of GGN, with a sensitivity range of 0.41–0.52, specificity range of 0.56–0.63, and AUC range of 0.60–0.67. Zhang et al. (Zhan et al., 2019) analyzed for GGN of 5–10 mm and found that GGNs larger than 8.12 mm and with attenuation greater than −449.52 HU were more likely to be IAC. Lobulation was another important independent risk predictor (Lee et al., 2013). Morphological changes such as lobulation justified the possibility of high invasiveness of small GGNs. Vascular changes were of important significance for the invasive judgment of GGN less than 10 mm. The IAC group was more likely to show vascular stiffness, distortion, expansion, or correction (Gao et al., 2019).

Size was a vital parameter for assessing the invasiveness of GGNs. Previous studies showed that the cutoff value of 10 mm was an optimal predictor for invasive lesions in pGGNs and 14 mm was an optimal predictor for invasive lesions in mGGNs (Lee et al., 2013). Another study showed a size difference between noninvasive and invasive group pGGN (0.74 vs. 0.90 cm, $p < 0.001$) (Sun et al., 2020).

The consolidation had the potential to identify the infiltration of the GGN. The consolidation/tumor ratio (CTR) was commonly used to assess the proportion of consolidation (Kobayashi et al., 2018). However, the ratio of consolidation in this study was not an independent risk factor of IAC, which may be related to different measurement methods. In 2013, Fleischner Society proposed that the consolidation should be evaluated in the mediastinal window and its size should be evaluated based on the average of the measured long and short diameters (Naidich et al., 2013). One study noted that the average diameter of consolidation in the mediastinum may not be the most suitable to assess mGGN progress (Kakinuma et al., 2015). Now most researchers observed and measured the consolidation of nodules on the lung window (Lee et al., 2014; Zhang et al., 2014). In addition, the size of the consolidation in the mediastinal window does not equal to the size of the infiltration focal point in the pathological specimen. Since part of the alveolar collapse, inflammatory, and fibrosis changes also appear as high density, the size of consolidation on the CT image may be larger than the actual range of pathological invasiveness.

Radiomics analysis provides a method to quantify and monitor changes in the treatment process (Aerts et al., 2014). Latest developments in image acquisition, standardization, and analysis promote an objective and accurate quantitative analysis that can

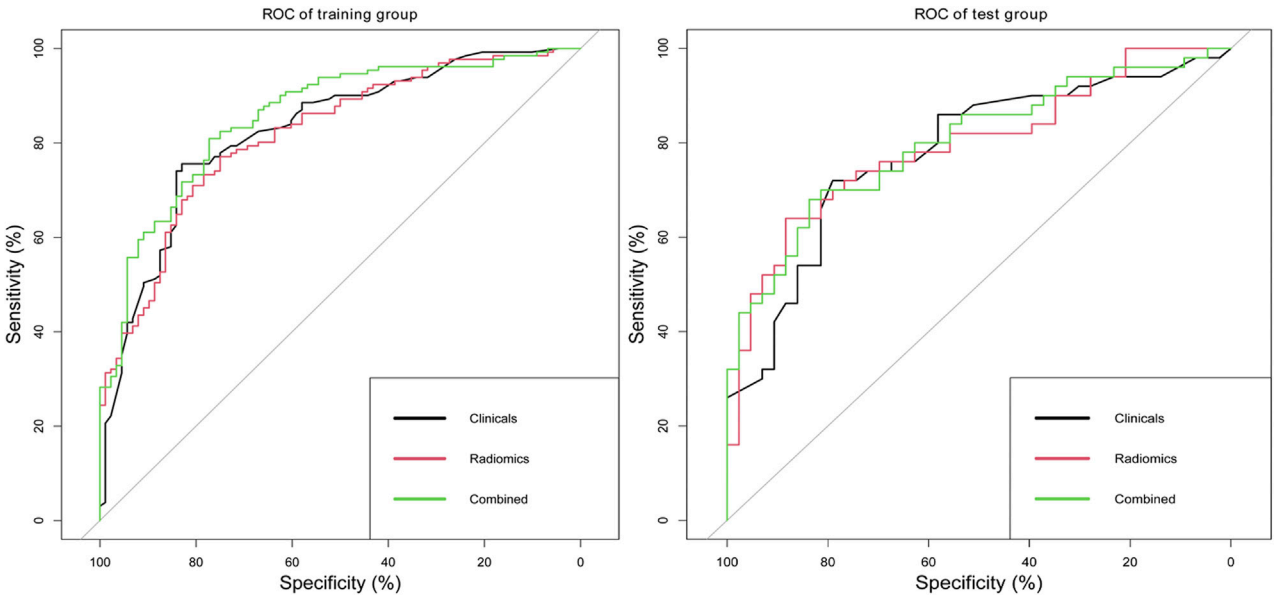


FIGURE 7 | ROC analysis of clinical model, radiomics model, and combined model in the training set and test set.

be used as a non-invasive diagnostic prediction method. Zhang et al. (2019) used histogram information and morphological features to construct invasive diagnostic models, with a sensitivity and specificity of 79.4% and 91.4%, respectively. Sun et al. (2020) found that the AUC of the combined model was higher than that of a single clinical model or radiomics model (training group: 0.8 vs. 0.75 vs. 0.73; validation group: 0.77 vs. 0.71 vs. 0.72). In addition to studying the tumor's own characteristics, radiomics can also further analyze the lung changes around the tumor by obtaining ROI in the peripheral region of nodules (Huang et al., 2018).

In terms of treatment, Ginsberg and Rubinstein (1995) had suggested that the long-term effect of lobectomy was better than sublobar resection. Recent studies have proposed sublobar resection rather than traditional lobectomy for AIS or pGGN manifesting as pGGN less than 20 mm (Watanabe et al., 2002; Yoshida et al., 2005). Surgical indications of GGN have not been uniform, and surgery is usually recommended for GGN with increased diameter or increased solid composition (Gould et al., 2013). Intraoperative freezing biopsy of early lung adenocarcinoma plays an important role in determining the surgical strategy. In this study, the diagnostic accuracy of frozen biopsy was high (benign/malignant diagnosis accuracy of 96.5%; pathological subtype diagnosis accuracy of 83.2%) and could help in diagnosis and classification and guide surgical treatment. When intraoperative frozen biopsy could not provide a timely diagnosis, radiomics may serve as a reliable reference for predicting pathological classification (Wang et al., 2020). In this study, the diagnostic accuracy of the clinical and radiomics models was lower than that of intraoperative freezing biopsy. The models still need further optimization in order to be more suitable for clinical diagnosis.

However, there are several limitations in the present study. First, this study is a retrospective research, conducted in a single center with a relatively smaller sample size. Larger sample size increases the statistical power of the diagnostic analysis which is necessary in the future, and thus a prospective cohort study should be conducted to validate these findings. More prospective data at different institutions should be analyzed to validate the clinical utility of the study results. Second, the repeatability of manual or semiautomatic tumor segmentation is an unsolved problem. Parts of GGN are close to the pleural or attached to blood vessels, which are more difficult to accurately segment and showed low repeatability (Kumar et al., 2012). Researchers propose new approaches to solve the segmentation problems of GGNs such as boundary leakage and small volume over-segmentation (Li et al., 2016). A review analysis shows that machine learning-based methods are useful for detecting and quantifying GGN (Mansoor et al., 2015). However, lung segmentation methods have not been amalgamated into single approaches or unified platforms using a single-user interface. Currently, the lung GGN segmentation is finished manually by experienced radiologists. This study will attempt to explore automated segmentation techniques to improve the efficiency of segmentation in future work.

CONCLUSION

Clinical and radiomics features have high accuracy in the invasive diagnosis of GGNs. Combined analysis can improve the diagnostic efficacy of IAC manifesting as GGNs. The nomogram serves as a noninvasive and accurate predictive tool to determine the invasiveness of GGNs prior to surgery and assist clinicians in creating personalized treatment strategies.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Medical Ethics Committee, Zhongnan Hospital of Wuhan University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements. 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

ML and HZ conceived ideas and designed the study. HZ and FX were responsible for the data collection, drafting of the manuscript, data analysis, and interpretation of the data. HZ and SW contributed to discussion. ML and FX have contributed equally to this work and share correspondence authorship.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.783391/full#supplementary-material

REFERENCES

Aerts, H. J. W. L., Velazquez, E. R., Leijenaar, R. T. H., Parmar, C., Grossmann, P., Carvalho, S., et al. (2014). Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. *Nat. Commun.* 5, 4006. doi:10.1038/ncomms5006

Dai, J., Yu, G., and Yu, J. (2018). Can CT Imaging Features of Ground-Glass Opacity Predict Invasiveness? A Meta-Analysis. *Thorac. Cancer* 9 (4), 452–458. doi:10.1111/1759-7714.12604

Gao, F., Sun, Y., Zhang, G., Zheng, X., Li, M., and Hua, Y. (2019). CT Characterization of Different Pathological Types of Subcentimeter Pulmonary Ground-Glass Nodular Lesions. *Bjr* 92 (1094), 20180204. doi:10.1259/bjr.20180204

Ginsberg, R. J., and Rubinstein, L. V. (1995). Randomized Trial of Lobectomy versus Limited Resection for T1 N0 Non-small Cell Lung Cancer. *Ann. Thorac. Surg.* 60 (3), 615–622. doi:10.1016/0003-4975(95)00537-u

Gould, M. K., Donington, J., Lynch, W. R., Mazzone, P. J., Midthun, D. E., Naidich, D. P., et al. (2013). Evaluation of Individuals with Pulmonary Nodules: when Is it Lung Cancer? Diagnosis and Management of Lung Cancer, 3rd Ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 143 (5 Suppl. l), e93S–e120S. doi:10.1378/chest.12-2351

Huang, P., Park, S., Yan, R., Lee, J., Chu, L. C., Lin, C. T., et al. (2018). Added Value of Computer-Aided CT Image Features for Early Lung Cancer Diagnosis with Small Pulmonary Nodules: A Matched Case-Control Study. *Radiology* 286 (1), 286–295. doi:10.1148/radiol.2017162725

Kakinuma, R., Muramatsu, Y., Kusumoto, M., Tsuchida, T., Tsuta, K., Maeshima, A. M., et al. (2015). Solitary Pure Ground-Glass Nodules 5 Mm or Smaller: Frequency of Growth. *Radiology* 276 (3), 873–882. doi:10.1148/radiol.2015141071

Kobayashi, Y., Ambrogio, C., and Mitsudomi, T. (2018). Ground-glass Nodules of the Lung in Never-Smokers and Smokers: Clinical and Genetic Insights. *Transl. Lung Cancer Res.* 7 (4), 487–497. doi:10.21037/tlcr.2018.07.04

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S. A., Schabath, M. B., et al. (2012). Radiomics: the Process and the Challenges. *Magn. Reson. Imaging* 30 (9), 1234–1248. doi:10.1016/j.mri.2012.06.010

Lee, K. H., Goo, J. M., Park, S. J., Wi, J. Y., Chung, D. H., Go, H., et al. (2014). Correlation between the Size of the Solid Component on Thin-Section CT and the Invasive Component on Pathology in Small Lung Adenocarcinomas Manifesting as Ground-Glass Nodules. *J. Thorac. Oncol.* 9 (1), 74–82. doi:10.1097/jto.0000000000000019

Lee, S. M., Park, C. M., Goo, J. M., Lee, H.-J., Wi, J. Y., and Kang, C. H. (2013). Invasive Pulmonary Adenocarcinomas versus Preinvasive Lesions Appearing as Ground-Glass Nodules: Differentiation by Using CT Features. *Radiology* 268 (1), 265–273. doi:10.1148/radiol.13120949

Li, B., Chen, Q., Peng, G., Guo, Y., Chen, K., Tian, L., et al. (2016). Segmentation of Pulmonary Nodules Using Adaptive Local Region Energy with Probability Density Function-Based Similarity Distance and Multi-Features Clustering. *Biomed. Eng. Online* 15 (1), 49. doi:10.1186/s12938-016-0164-3

Mansoor, A., Bagci, U., Foster, B., Xu, Z., Papadakis, G. Z., Folio, L. R., et al. (2015). Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends. *RadioGraphics* 35 (4), 1056–1076. doi:10.1148/rg.2015140232

Naidich, D. P., Bankier, A. A., MacMahon, H., Schaefer-Prokop, C. M., Pistolesi, M., Goo, J. M., et al. (2013). Recommendations for the Management of Subsolid Pulmonary Nodules Detected at CT: A Statement from the Fleischner Society. *Radiology* 266 (1), 304–317. doi:10.1148/radiol.12120628

Park, C. M., Goo, J. M., Lee, H. J., Lee, C. H., Chun, E. J., and Im, J.-G. (2007). Nodular Ground-Glass Opacity at Thin-Section CT: Histologic Correlation and Evaluation of Change at Follow-Up. *Radiographics* 27 (2), 391–408. doi:10.1148/rg.272065061

Qi, L.-L., Wu, B.-T., Tang, W., Zhou, L.-N., Huang, Y., Zhao, S.-J., et al. (2020). Long-term Follow-Up of Persistent Pulmonary Pure Ground-Glass Nodules with Deep Learning-Assisted Nodule Segmentation. *Eur. Radiol.* 30 (2), 744–755. doi:10.1007/s00330-019-06344-z

Sun, Y., Li, C., Jin, L., Gao, P., Zhao, W., Ma, W., et al. (2020). Radiomics for Lung Adenocarcinoma Manifesting as Pure Ground-Glass Nodules: Invasive Prediction. *Eur. Radiol.* 30 (7), 3650–3659. doi:10.1007/s00330-020-06776-y

Wang, B., Tang, Y., Chen, Y., Hamal, P., Zhu, Y., Wang, T., et al. (2020). Joint Use of the Radiomics Method and Frozen Sections Should Be Considered in the Prediction of the Final Classification of Peripheral Lung Adenocarcinoma Manifesting as Ground-Glass Nodules. *Lung Cancer* 139, 103–110. doi:10.1016/j.lungcan.2019.10.031

Watanabe, S.-i., Watanabe, T., Arai, K., Kasai, T., Haratake, J., and Urayama, H. (2002). Results of Wedge Resection for Focal Bronchioloalveolar Carcinoma Showing Pure Ground-Glass Attenuation on Computed Tomography. *Ann. Thorac. Surg.* 73 (4), 1071–1075. doi:10.1016/s0003-4975(01)03623-2

Yang, J., Wang, H., Geng, C., Dai, Y., and Ji, J. (2018). Advances in Intelligent Diagnosis Methods for Pulmonary Ground-Glass Opacity Nodules. *Biomed. Eng. Online* 17 (1), 20. doi:10.1186/s12938-018-0435-2

Yoshida, J., Nagai, K., Yokose, T., Nishimura, M., Kakinuma, R., Ohmatsu, H., et al. (2005). Limited Resection Trial for Pulmonary Ground-Glass Opacity Nodules: Fifty-Case Experience. *J. Thorac. Cardiovasc. Surg.* 129 (5), 991–996. doi:10.1016/j.jtcvs.2004.07.038

Zhan, Y., Peng, X., Shan, F., Feng, M., Shi, Y., Liu, L., et al. (2019). Attenuation and Morphologic Characteristics Distinguishing a Ground-Glass Nodule Measuring 5-10 Mm in Diameter as Invasive Lung Adenocarcinoma on Thin-Slice CT. *Am. J. Roentgenology* 213 (4), W162–w170. doi:10.2214/ajr.18.21008

Zhang, T., Pu, X.-H., Yuan, M., Zhong, Y., Li, H., Wu, J.-F., et al. (2019). Histogram Analysis Combined with Morphological Characteristics to Discriminate Adenocarcinoma *In Situ* or Minimally Invasive Adenocarcinoma from Invasive Adenocarcinoma Appearing as Pure Ground-Glass Nodule. *Eur. J. Radiol.* 113, 238–244. doi:10.1016/j.ejrad.2019.02.034

Zhang, Y., Qiang, J. W., Ye, J. D., Ye, X. D., and Zhang, J. (2014). High Resolution CT in Differentiating Minimally Invasive Component in Early Lung Adenocarcinoma. *Lung Cancer* 84 (3), 236–241. doi:10.1016/j.lungcan.2014.02.008

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 January 2022
doi: 10.3389/fgene.2021.791349

Construction of the Classification Model Using Key Genes Identified Between Benign and Malignant Thyroid Nodules From Comprehensive Transcriptomic Data

Qingxia Yang[1] *and Yaguo Gong*[2]*

[1]*Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, Department of Bioinformatics, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing, China,* [2]*School of Pharmacy, Macau University of Science and Technology, Macau, China*

OPEN ACCESS

Edited by:
Fengfeng Zhou,
Jilin University, China

Reviewed by:
Ravi Pandey,
Jackson Laboratory for Genomic Medicine, United States
Sumeet Gulati,
International Clinical Research Center (FNUSA-ICRC), Czechia

***Correspondence:**
Yaguo Gong
gongyglab@gmail.com

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

Received: *08 October 2021*
Accepted: *06 December 2021*
Published: *14 January 2022*

Citation:
Yang Q and Gong Y (2022) Construction of the Classification Model Using Key Genes Identified Between Benign and Malignant Thyroid Nodules From Comprehensive Transcriptomic Data. Front. Genet. 12:791349. doi: 10.3389/fgene.2021.791349

Thyroid nodules are present in upto 50% of the population worldwide, and thyroid malignancy occurs in only 5–15% of nodules. Until now, fine-needle biopsy with cytologic evaluation remains the diagnostic choice to determine the risk of malignancy, yet it fails to discriminate as benign or malignant in one-third of cases. In order to improve the diagnostic accuracy and reliability, molecular testing based on transcriptomic data has developed rapidly. However, gene signatures of thyroid nodules identified in a plenty of transcriptomic studies are highly inconsistent and extremely difficult to be applied in clinical application. Therefore, it is highly necessary to identify consistent signatures to discriminate benign or malignant thyroid nodules. In this study, five independent transcriptomic studies were combined to discover the gene signature between benign and malignant thyroid nodules. This combined dataset comprises 150 malignant and 93 benign thyroid samples. Then, there were 279 differentially expressed genes (DEGs) discovered by the feature selection method (Student's *t* test and fold change). And the weighted gene co-expression network analysis (WGCNA) was performed to identify the modules of highly co-expressed genes, and 454 genes in the gray module were discovered as the hub genes. The intersection between DEGs by the feature selection method and hub genes in the WGCNA model was identified as the key genes for thyroid nodules. Finally, four key genes (ST3GAL5, NRCAM, MT1F, and PROS1) participated in the pathogenesis of malignant thyroid nodules were validated using an independent dataset. Moreover, a high-performance classification model for discriminating thyroid nodules was constructed using these key genes. All in all, this study might provide a new insight into the key differentiation of benign and malignant thyroid nodules.

Keywords: classification model, key genes, transcriptomics, combined analysis, thyroid nodules

INTRODUCTION

Thyroid nodules are regarded as common clinical problems worldwide, and nearly 50% of the population harbor thyroid nodules (Burman and Wartofsky, 2015; Jasim et al., 2020). For benign thyroid nodules, there is no need to perform any medical treatment if it does not keep growing or cause other problems (Durante et al., 2015). Indeed, less than 10% of patients' thyroid nodules demonstrate disease progression after a median follow-up of 6 years (Ito et al., 2014). But the thyroid malignancy occurring in only 5–15% of thyroid nodules needed to be treated surgically (Wong et al., 2018). Therefore, to improve treatment efficiency, the main challenge is on how to differentiate the malignant nodules from the majority of benign ones reliably using the diagnostic methods (Cho et al., 2020; Singh Ospina et al., 2020).

Until now, to determine the risk of malignancy, fine-needle aspiration (FNA) with cytologic evaluation remains the diagnostic choice for ≥1.0 cm nodules (Heider et al., 2020). But one-third of thyroid nodules could not be discriminated as benign or malignant correctly (Cibas and Ali, 2009). Over the past decade, molecular testing has developed rapidly to improve the diagnostic accuracy as well as minimize cost and unnecessary testing for indeterminate cases (Roth et al., 2018). Moreover, transcript profiling is a widely used technique to discover the molecular changes. Transcriptomics could obtain information simultaneously based on the abundance of multiple mRNA transcripts for the biological sample (Knyazeva et al., 2020; Moncada et al., 2020). So, the gene signatures based on transcriptomic data could be used to distinguish benign from malignant thyroid nodules efficiently.

Recently, there have been a lot of transcriptomic studies to identify the gene signatures associated with thyroid nodules. For example, Giordano et al. found the three genes (PPARG, AQP7, and ENO3) implicated for the neoplastic mechanism of thyroid follicular carcinomas (Giordano et al., 2006). Wojtas et al. confirmed differential expression of seven genes (CPQ, PLVAP, TFF3, ACVRL1, ZFYVE21, FAM189A2, and CLEC3B) between malignant and benign follicular thyroid tumors (Wojtas et al., 2017). Schulten et al. revealed 55 transcripts (GABBR2, NRCAM, ECM1, HS6ST2, RXRG, *etc.*) differentially expressed between follicular variant of papillary thyroid carcinomas and follicular adenomas of the thyroid (Schulten et al., 2015). Hinsch et al. detected that QPRT was a potential marker for the immunohistochemical screening of follicular thyroid nodules (Hinsch et al., 2009). Although there were various signatures identified in different studies, it was reported that they were difficult to be applied in clinical diagnosis because of the inconsistency and unreliability (Singh Ospina et al., 2020).

The inconsistency among gene signatures from different studies might result from many sources, such as limited number of samples (Schwalbe et al., 2017; Osborn et al., 2018). It is understood that these transcriptomic studies were performed using dozens of samples of thyroid nodules. If the multiple independent studies could be combined as one comprehensive dataset, the sample size could be enlarged and the stability of the gene signatures could be enhanced significantly (Mistry et al., 2013). Moreover, weighted gene co-expression network analysis (WGCNA) could be used to identify the modules of co-expressed genes highly associated with the biological mechanism (He et al., 2019). WGCNA has been widely used to explore biomarkers and therapeutic targets of various diseases (Niemira et al., 2019; Chen et al., 2020). Therefore, it was highly needed to identify key genes between malignant and benign thyroid nodules by WGCNA from a comprehensive dataset.

In this work, five independent transcriptomic studies comprising 150 malignant and 93 benign thyroid nodule samples were combined to discover the gene signatures of thyroid nodules. First, 279 differentially expressed genes (DEGs) were identified by the feature selection method (Student's *t* test and fold change) after data preprocessing and batch effect removal. And various biological process terms (such as hormone metabolic process, platelet degranulation, and thyroid hormone generation) were enriched using these DEGs. Second, the WGCNA model was constructed to identify significant modules of highly co-expressed genes, and 454 hub genes in the gray module were identified. Third, the intersection between DEGs identified by the feature selection method and the hub genes using the WGCNA model was discovered as the key genes. In order to perform the systematic validation, four key genes participated in the pathogenesis of malignant thyroid nodules were validated by an independent dataset. Finally, a high-performance classification model for discriminating benign and malignant thyroid nodules was constructed using these key genes. All in all, this study might provide a useful classification model for discriminating benign and malignant thyroid nodules.

MATERIALS AND METHODS

Collection of Transcriptomic Data From Multiple Studies

A variety of microarray studies based on thyroid tissue were collected by searching the key word "thyroid nodules" in the Gene Expression Omnibus (GEO) database (Barrett et al., 2013). These collected datasets should meet the following criteria (Yang et al., 2020b): 1) the gene expression profiling was conducted using *cDNA* microarray for "*Homo Sapiens*"; 2) the tissues analyzed were thyroid nodules; 3) raw data could be available for further analysis; and 4) the collected datasets should consist of one group of malignant samples and another group of benign ones. As a result, five independent transcriptomic datasets were collected, and each comprised both benign and malignant thyroid nodules. The detailed information of these five collected datasets is provided in **Table 1**, including dataset ID, number of samples, microarray platform, and tissue indicated in the original publication and references.

Data Preprocessing and Batch Effect Removal

To enhance the consistency and classification capacity, all datasets in this study (**Table 1**) were combined to discover the

TABLE 1 | Datasets collected from five independent microarray studies of thyroid nodules (sorted by sample size). Each dataset contained one cohort of malignant and another cohort of another group of benign samples.

Id	No. of samples (malignant: benign)	Platform	Tissue	References
GSE27155	95 (78:17)	HG-U133A	Thyroid tissue	*Clin Cancer Res* 12 (7): 1983–93, 2006
GSE29315	71 (31:40)	HG-U95Av2	Thyroid tissue	Tomas G, *et al.* unpublished, 2012
GSE82208	52 (27:25)	HG-U133 Plus 2	Thyroid tissue	*Int J Mol Sci* 18 (6): 1,184, 2017
GSE54958	13 (6:7)	HuGene-1.0 ST	Thyroid tissue	*BMC Genomics* 16 (S1): S7, 2015
GSE15045	12 (8:4)	ABI Human Genome Survey Microarray v.2	Thyroid tissue	*BMC Cancer* 9: 93, 2009

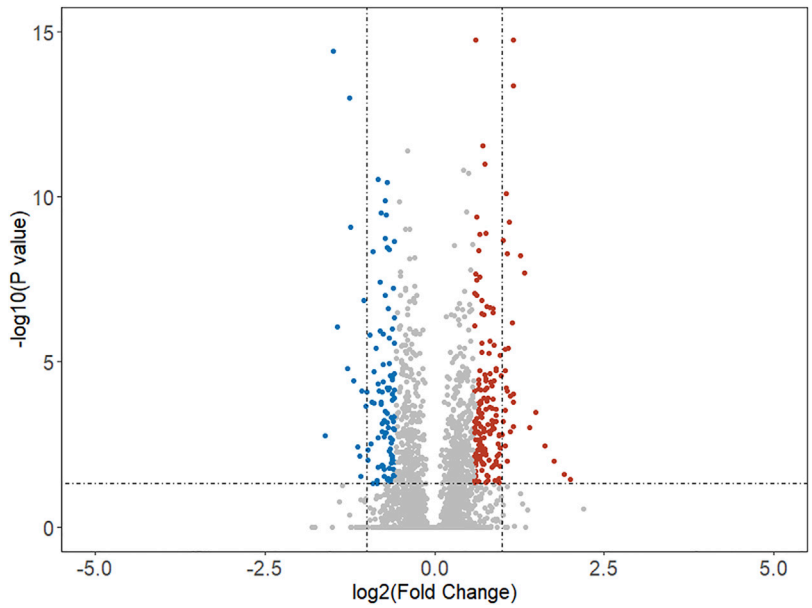


FIGURE 1 | Volcano map of differentially expressed genes in malignant samples compared with benign samples. The horizon line was the cutoff (adjusted *p*-value < 0.05) of Student's *t* test. The vertical line was the cutoff (logFC >0.58 or logFC < -0.58) of the fold change method. The blue and red dots indicated the downregulated and upregulated genes, respectively.

key genes of thyroid nodules. The combination of multiple datasets was carried out in *R* environment (v3.4.3, http://www.r-project.org) (Sepulveda, 2020). The raw data (*CEL* file) of all datasets were read, log-transformed, and normalized using the corresponding *R* package, and all parameters were set as default. All probe sets were then mapped to their corresponding gene names using *Bioconductor* (Tippmann, 2015). The average expression value was retained if one gene was mapped to multiple probes (Yang et al., 2020c). To remove batch effects among five independent datasets, *Z*-score transformation was used to adjust the gene expression levels in each dataset (Yang Q et al., 2019b; Yang et al., 2020a). *Z*-score transformation for each gene could be computed by subtracting the mean of all genes and dividing the difference by the standard deviation of all genes in one experiment. After data transformation, the mean value for each experiment became zero with standard deviation equaling one.

Differentially Expressed Genes Discovered Between Benign and Malignant Thyroid Nodules

In this study, there were five collected datasets integrated as a comprehensive dataset for discovering signatures. This comprehensive dataset consisted of 150 malignant and 93 benign samples of thyroid nodules. To the best of one's knowledge, this integrated dataset was the largest transcriptomic dataset in the analysis of thyroid nodules. Based on this comprehensive dataset, the DEGs were discovered using feature selection methods including Student's *t* test and fold change (FC). For Student's *t* test, *multtest*

TABLE 2 | Top 25 up- and downregulated DEGs identified by Student's *t* test and fold change method (logFC >0.58 or logFC < -0.58 and adjusted *p*-value < 0.05) combining all five datasets in **Table 1**.

ID	Entrez ID	Gene symbol	Adjusted *p*-value	logFC
Table A. The top 25 upregulated genes				
1	9,324	HMGN3	0.035423	1.999879
2	515	ATP5F1	0.02562	1.907751
3	5,800	PTPRO	0.010352	1.767712
4	23576	DDAH1	0.003481	1.626399
5	9,782	MATR3	0.000342	1.498593
6	11167	FSTL1	0.000987	1.408146
7	4,435	CITED1	2.04E-08	1.328755
8	301	ANXA1	5.86E-09	1.273075
9	1803	DPP4	1.81E-15	1.166173
10	55885	LMO3	9.26E-05	1.162304
11	10944	C11orf58	0.00016	1.162246
12	1,001	CDH3	4.16E-14	1.155315
13	722	C4BPA	0.000938	1.154525
14	10178	TENM1	6.51E-07	1.15377
15	439,921	MXRA7	0.001287	1.117048
16	159	ADSS	0.000106	1.113014
17	5,627	PROS1	5.72E-10	1.104001
18	6,447	SCG5	3.80E-06	1.081727
19	7,360	UGP2	7.51E-05	1.076941
20	25797	QPCT	5.05E-09	1.068464
21	1,622	DBI	0.009991	1.065552
22	5,906	RAP1A	6.06E-05	1.055333
23	7,991	TUSC3	7.96E-11	1.05345
24	7,498	XDH	1.86E-05	1.04801
25	10981	RAB32	0.000299	1.046273
Table B. The top 25 downregulated genes				
26	4,703	NEB	3.90E-06	−0.8582
27	432	ASGR1	2.01E-05	−0.89599
28	1805	DPT	0.00018	−0.8994
29	4,494	MT1F	4.58E-09	−0.91087
30	219,333	USP12	0.047108	−0.9167
31	2,117	ETV3	0.000167	-0.93059
32	6,722	SRF	0.003049	−0.94275
33	1,381	CRABP1	1.48E-06	−0.95542
34	6,921	TCEB1	0.004592	-0.98698
35	2,323	FLT3LG	0.009582	−0.98782
36	1,299	COL9A3	8.03E-05	-1.00485
37	4,713	NDUFB7	0.000215	−1.00738
38	4,495	MT1G	1.39E-07	-1.05177
39	9,265	CYTH3	7.71E-05	−1.07064
40	8,458	TTF2	0.030282	−1.09564
41	968	CD68	0.007163	−1.11098
42	6,624	FSCN1	0.003741	−1.12761
43	4,920	ROR2	3.74E-05	−1.19808
44	2,167	FABP4	8.24E-10	−1.24181
45	744	MPPED2	1.02E-13	−1.25312
46	3,292	HSD17B1	1.63E-05	-1.28357
47	1,014	CDH16	3.65E-16	−1.33575
48	1,733	DIO1	8.64E-07	−1.42927
49	7,173	TPO	3.90E-15	−1.49917
50	9,351	SLC9A3R2	0.00174	−1.61953

package of R language was applied, and the adjusted *p*-value < 0.05 was selected as the cutoff (Yan et al., 2019). The fold change was used to compare the mean expression of each gene between malignant and benign thyroid nodules (Yu et al., 2020). The cutoff level of FC was set to logFC >0.58 (FC > 1.5) or logFC < -0.58 (FC < 0.67). The equation of FC was shown below (as shown in **Eq. (1)**).

$$\log FC = mean(log\,2(Malignant\ Group)) - mean(\log 2(Benign\ Group)). \quad \text{(Eq.1)}$$

The volcano plot was applied to visualize and demonstrate the DEGs using *ggplot2* package. Then the analysis of gene ontology (GO) enrichment was performed to identify the key biological processes for thyroid nodules (Yang et al., 2019a). Moreover, *GOplot* and *clusterProfiler* packages were used for visualizing the biological processes (BP) of GO enrichment (Yu et al., 2012; Yang et al., 2021). The raw *p*-value < 0.05 of GO terms was considered statistically significant.

Hub Genes Identified Using Weighted Gene Co-Expression Network Analysis

The *WGCNA* package was applied to establish the scale-free weight gene co-expression networks for thyroid nodules (Langfelder and Horvath, 2008). The unqualified genes were screened out, and the matrix of genes' similarity by Pearson's correlation analysis was created. Appropriate soft threshold power (β) was applied to strengthen this matrix to a scale-free co-expression network (Yang et al., 2020b). The lowest power was chosen, so the scale-free topology fit index curve flattened out upon reaching a high value. The highly correlated genes were assigned into the same module. As a result, the intersection was obtained between DEGs identified by the feature selection method and hub genes in a key module using the WGCNA model. These genes in the intersection were regarded as the key genes for further validation.

Validation of the Key Genes Based on the Independent Dataset

A systematic validation was conducted by evaluating the upregulated and downregulated genes based on the independent dataset (GSE34289) (Alexander et al., 2012). This validation dataset consisted of two independent datasets from two different platforms. The first independent dataset was detected based on GPL5175 platform (Affymetrix Human Exon 1.0 ST Array). In this dataset, there were 23 malignant and 26 benign thyroid nodules. The second independent dataset was detected based on GPL14961 platform (Afirma-T Human Custom Array). There were 120 malignant and 198 benign samples in this second independent dataset. In this study, the boxplot was used to demonstrate the differential expression of these key genes between malignant and benign thyroid nodules.

Construction of the High-Performance Classification Model Using the Key Genes

To construct a classification model for thyroid nodules, four powerful classifiers, namely, support vector machine, linear discriminate analysis, partial least squares, and random forest algorithm, were applied in this study (Orru et al., 2012). The key genes between malignant and benign thyroid nodules were used to discriminate different samples. In the first step, the five-fold cross validation of the comprehensive dataset (**Table 1**) was performed to validate the performance of this classification

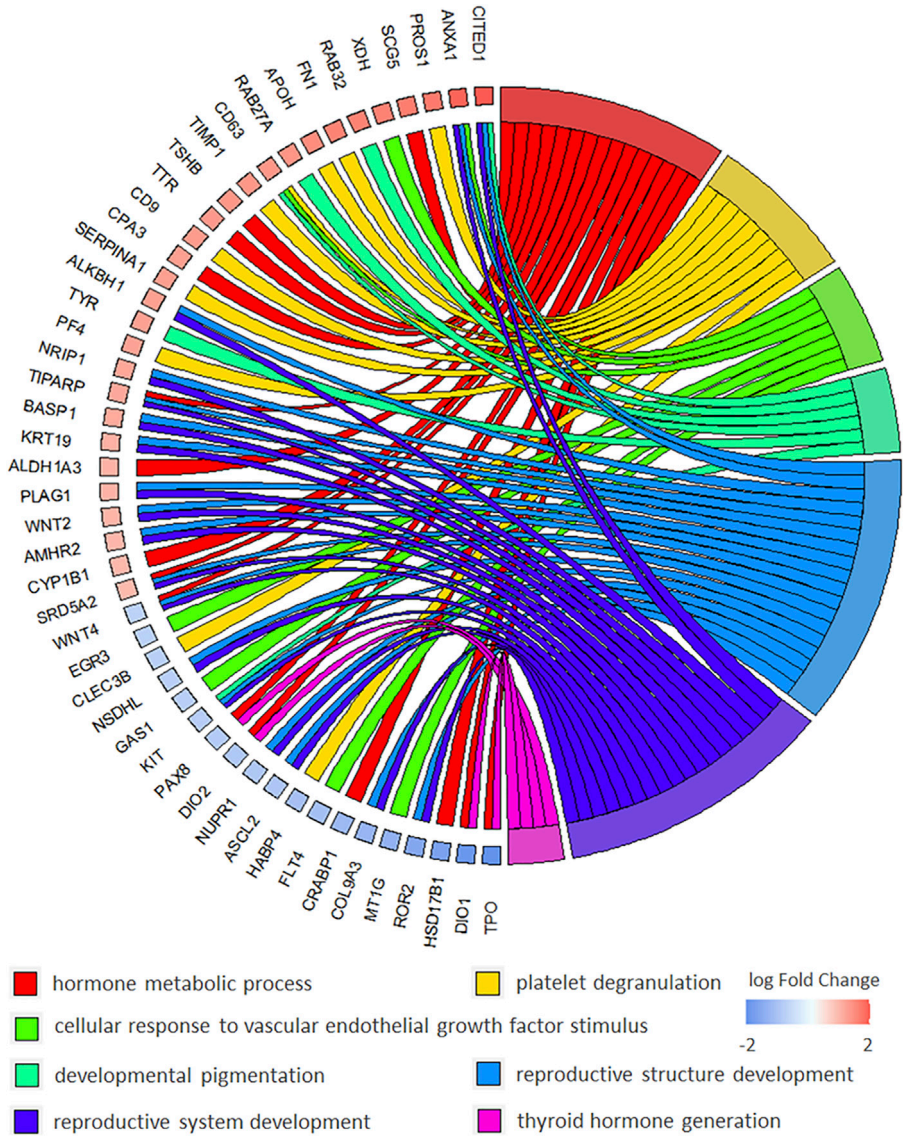


FIGURE 2 | Chord diagram of BP (biological process) of GO enrichment to explain the relationship between BP terms and DEGs in malignant *versus* benign thyroid nodules.

model. The accuracy of five-fold cross validation could reflect the quality of the model. In the second step, the comprehensive dataset was set as the training set, and the two independent datasets from GSE34289 were set as the test sets. The performance of the independent test set could accurately reflect the classification ability of the model. This high-performance classification model based on machine learning was constructed for discriminating benign and malignant thyroid nodules.

RESULTS AND DISCUSSION

Collection of Multiple Transcriptomic Data for Thyroid Nodules

A variety of microarray studies based on thyroid tissue were collected by searching the key word "thyroid nodules" in the GEO database. As a result, five independent transcriptomic studies were obtained, and each comprised a cohort of malignant samples and another cohort of benign samples. The detailed information of these independent datasets is provided in **Table 1**. Among these studies, the five datasets including 150 malignant and 93 benign thyroid nodules were combined as a comprehensive dataset. The boxplots of five datasets before and after batch effect removal are shown in **Supplementary Figure S1**. The intensity of all samples before batch effect removal was distributed in the range of 4–15 and fluctuated greatly. After batch effect removal, the intensity of all samples was roughly distributed in the range of -1–1. The stable distribution indicated that the batch effects were well removed in the combined dataset by *Z*-score transformation. After data preprocessing and batch effect removal, the comprehensive dataset with 7,265 genes from five independent studies was applied to discover the key genes of thyroid nodules.

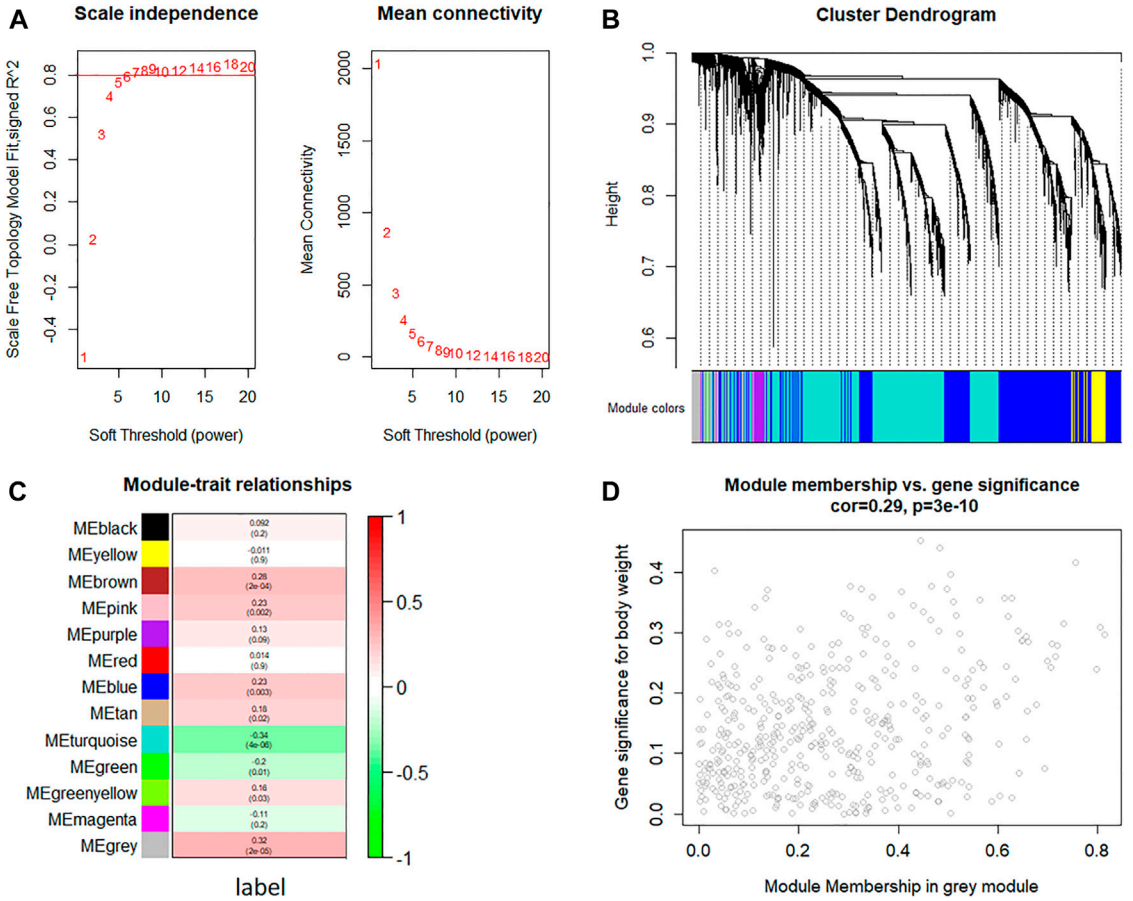


FIGURE 3 | Weighted gene co-expression network analysis of gene expression between malignant and benign thyroid nodules. **(A)** Analysis of the scale-free topology fit index and the mean connectivity for various soft threshold powers (β) for the genes, **(B)** dendrogram of all expressed genes clustered based on a dissimilarity measure, **(C)** heatmap of module–trait relationships depicting correlations between module eigengenes and phenotypic traits (the label of malignant and benign thyroid nodules). Numbers correspond to the correlation and the *p*-value in parentheses. The degree of correlation is illustrated with the color legend, and **(D)** identification of hub genes using the scatterplot of module eigengenes in the gray co-expression module.

DEGs of Thyroid Nodules Identified Using the Combined Dataset

Based on this comprehensive dataset, the DEGs were discovered using feature selection methods (both Student's *t* test and fold change). The volcano plot (as shown in **Figure 1**) illuminated the variation of DEGs in malignant *versus* benign thyroid nodules. The horizon line was the cutoff (adjusted *p*-value < 0.05) of Student's *t* test. The cutoff levels for the vertical line were set to logFC >0.58 (FC > 1.5) or logFC < -0.58 (FC < 0.67) of fold change. The blue and red dots were used to indicate the upregulated (logFC >0.58) and downregulated (logFC < -0.58) genes, respectively. In this study, 279 DEGs were finally identified by both Student's *t* test and fold change. The total number of upregulated genes (172 genes) was larger than that of the downregulated ones (107 genes). The top 25 upregulated and downregulated DEGs are shown in **Table 2**, including the information of entrez ID, gene symbol, adjusted *p*-value, and fold change for each gene. The information of all DEGs is shown in **Supplementary Table S1**.

GO Enrichment Analysis Using DEGs of Thyroid Nodules

GO enrichment analysis is ubiquitously used for interpreting high throughput molecular data and underlying biological phenomena of experiments (Tomczak et al., 2018). For a set of genes, an enrichment analysis will find which GO terms are overrepresented using annotations for the gene set. GO enrichment analysis for the DEGs was performed in this study. Using the DEGs between malignant and benign thyroid nodules, the enrichment analysis included the BP (biological process), MF (molecular function), and CC (cell component) terms. The detailed information of GO ID, description, *p*-value, name, and the number of genes is shown in **Supplementary Table S2**.

Particularly, multiple biological processes were enriched to interpret the biological mechanism of malignant thyroid nodules. The chord diagram of BP enrichment (as interpreted in **Figure 2**) was applied to explain the relationship between DEGs and BP terms. It was reported that these BP terms were associated with the biological mechanism of thyroid nodules. For example,

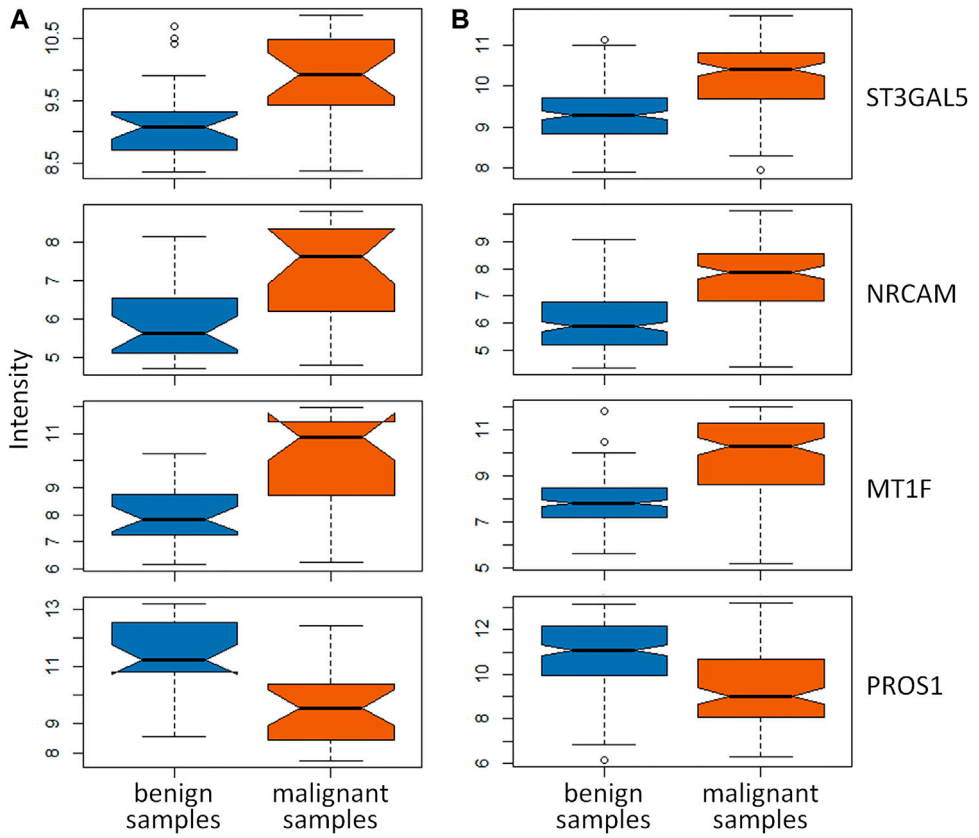


FIGURE 4 | Validation of key genes identified by both DEGs identified by the feature selection method and hub genes in the gray module using WGCNA. The boxplots of these key genes between malignant and benign thyroid nodules were validated in the independent dataset detected by **(A)** GPL5175 and **(B)** GPL14961 platforms.

there were 15 DEGs enriched in the hormone metabolic process, and the association with thyroid cancer has been reported (Han et al., 2018). The platelet degranulation enriched by 10 DEGs was discovered in papillary thyroid carcinoma using the biomarkers (Wu et al., 2018). The concentration of the vascular endothelial growth factor was increased and stimulated endothelial cell proliferation in the cyst fluid of enlarging and recurrent thyroid nodules (Sato et al., 1997). It was reported that patients with spotty skin pigmentation had a predisposition toward the development of thyroid abnormalities (Courcoutsakis et al., 2009). It was found that low thyroid hormones might have implications for reproductive health, so the reproductive structure development and reproductive system development might be affected in thyroid nodules (Medda et al., 2017). The thyroid hormone generation reported that the significant biologic process was involved in thyroid cancers (Durante et al., 2018).

Construction of the WGCNA Network and Identification of the Gene Co-Expression Module

The WGCNA network was constructed to identify the gene co-expression module (as shown in **Figure 3**). The value of power (10) was selected as the soft-threshold power to ensure scale-free (R^2 = 0.8) networks using the *WGCNA* package (**Figure 3A**) because it reached the plateau at power 10 from the scale-free topology plot and mean connectivity plot. Genes with similar expression patterns were clustered into co-expression modules. Different modules were shown in different colors, and 13 modules were identified totally (**Figure 3B**). The heatmap of module–trait relationships was applied for depicting correlations between module eigengenes and phenotypic traits (the label of malignant and benign thyroid nodules). As shown in **Figure 3C**, the numbers correspond to the correlation, and the *p*-values were set in parentheses. Moreover, the degree of correlation was illustrated with the color legend. Here, the gray module was the most correlated one with malignant thyroid nodules (R = 0.32, p-value = 2×10^{-5}). Hence, the gray module was used for the identification of the hub genes. Hub genes in the co-expression network were characterized by high intra-modular connectivity measured by the value of gene significance and module membership. The scatterplot of module eigengenes related to malignant thyroid nodules in the gray co-expression module ($R = 0.29$, p-value = 3×10^{-10}) is shown in **Figure 3D**. As a result, 454 genes in the gray module

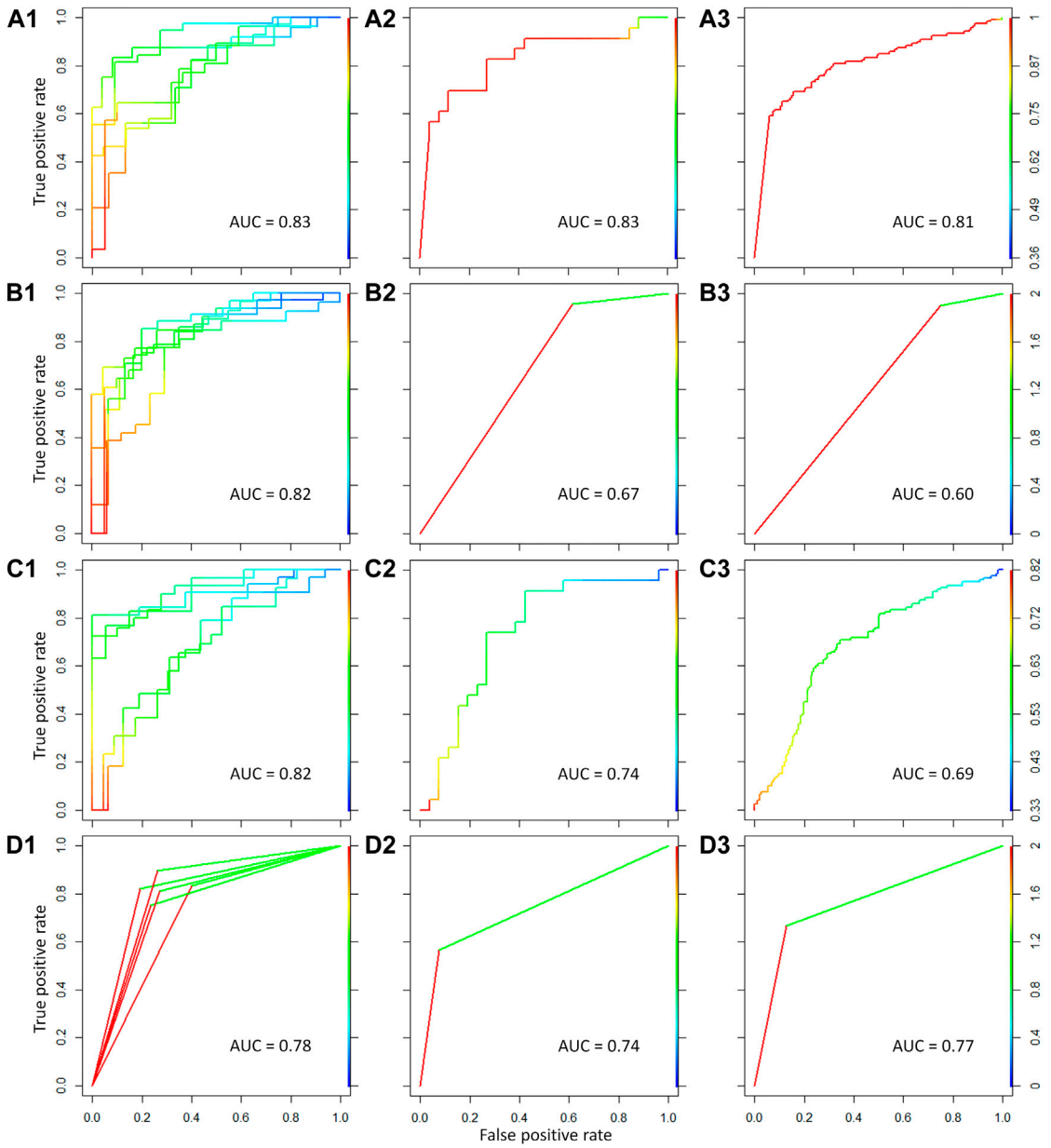


FIGURE 5 | Classification model constructed for discriminating malignant from benign thyroid nodules using four different machine learning methods. The four methods referred to support vector machine, linear discriminate analysis, partial least squares, and random forest algorithm from top to bottom. The ROC curves and AUC values for the five-fold cross validation were shown in **(A1–D1)** for the comprehensive dataset using the four methods. The ROC curves and AUC values for the first independent test set were shown in **(A2–D2)**. The ROC curves and AUC values for the second independent test set were shown in **(A3–D3)**.

highly correlated with gene significance were identified as hub genes using WGCNA.

Validation of the Key Genes Using the Independent Datasets

In this study, there were 19 overlapping genes in the intersection between 279 DEGs identified by the feature selection method and 454 hub genes in the gray module totally. To validate these overlapping genes, two independent datasets from GSE34289 were applied to perform the systematic validation (Alexander et al., 2012). In this validation dataset, there were 23 malignant with 26 benign samples and 120 malignant with 198 benign samples form GPL5175 and GPL14961 platforms, respectively. The boxplots (as shown in **Figure 4**) were used to demonstrate the key genes between malignant and benign thyroid nodules. Among the 19 overlapping genes, there were four key genes expressed in the

independent dataset, and the dysregulation of these key genes was validated. As shown in **Figure 4**, the significant differences of three upregulated genes (ST3GAL5, NRCAM, and MT1F) and one downregulated gene (PROS1) were indicated in these boxplots obviously for the independent data detected from GPL5175 (**Figure 4A**) and GPL14961 platforms (**Figure 4B**), respectively.

As a result, these four key genes were effectively validated as the important ones participated in the pathogenesis of thyroid nodules. It was reported that the specific genetic variants of ST3GAL5 in patients with thyroid-associated ophthalmopathy were discovered (Park et al., 2017). Górka et al. provided the first evidence that NRCAM is overexpressed in papillary thyroid carcinomas, and the upregulation of NRCAM was implicated in the pathogenesis and behavior of papillary thyroid cancers (Gorka et al., 2007). It was reported that MT1F might contribute to thyroid carcinogenesis and potentially serve as a diagnostic marker in distinguishing benign from malignant lesions (Kim et al., 2010; Wojtczak et al., 2017). In the previous studies, PROS1 was reported as the biomarker significantly related to thyroid nodules' malignancy (Griffith et al., 2006; Wu et al., 2020). In this study, these four key genes (ST3GAL5, NRCAM, MT1F, and PROS1) were discovered for distinguishing malignant from benign thyroid nodules.

Construction of the High-Performance Classification Model Using the Key Genes

To distinguish malignant from benign thyroid nodules, four popular machine learning methods were applied to construct the classification model in this study. These methods included support vector machine, linear discriminate analysis, partial least squares, and random forest algorithm. The key genes between benign and malignant thyroid nodules were used to discriminate different samples. For the comprehensive dataset in **Table 1**, the five-fold cross validation was first performed to validate the performance of this classification model. As shown in **Figure 5A1**, **5B1**, **5C1**, and **5D1**, the values of area under the ROC curve (AUC) were 0.83, 0.82, 0.82, and 0.78 for the five-fold cross validation using four different machine learning methods, respectively. Moreover, the high performance of the independent test sets could accurately reflect the ability of the classification model. The comprehensive dataset was set as the training set, and the test sets consisted of two parts detected by GPL5175 and GPL14961 platforms from the independent dataset (GSE34289). As displayed in **Figure 5A2**, **5B2**, **5C2**, and **5D2**, the AUC values of the ROC curve for the first independent test set were 0.83, 0.67, 0.74, and 0.74 by four machine learning methods, respectively. As shown in **Figure 5A3**, **5B3**, **5C3**, and **5D3**, the AUC values for the second independent test set were 0.81, 0.60, 0.69, and 0.77 by four machine learning methods, respectively.

As shown in **Figure 5**, for the five-fold cross validation, the performances (AUC >0.8) of the classification model were outstanding using support vector machine, linear discriminate analysis, and partial least squares. However, the classification models of support vector machine and random forest (AUC >0.7) have shown more excellent performances than the other methods for the two independent test sets. Therefore, the high-performance classification model using support vector machine was recommended for discriminating malignant from benign thyroid nodules based on both five-fold cross validation and independent test.

Until now, it fails to discriminate as benign or malignant in one-third of thyroid nodules using FNA with cytologic evaluation. To save medical costs and improve the diagnostic accuracy, the high-performance classification model constructed in this study could be applied before FNA. For the thyroid nodule patients, the expression of four key genes could be detected. Then, this sample could be classified as benign or malignant thyroid nodules based on the classification model. If the patient was classified as a malignant thyroid sample, it was highly necessary to make a definite diagnosis using FNA with cytologic evaluation. If the patient was classified as a benign sample based on the classification model, the necessity of the FNA could be determined depending on the specific conditions. In the future, selection method, the high-performance classification model is expected to be applied for clinical diagnosis and management for malignant and benign thyroid nodules.

CONCLUSION

In this study, a comprehensive dataset including 150 malignant and 93 benign samples was collected to discover the gene signature of thyroid nodules. Then, 279 DEGs were identified by the feature selection method (Student's *t* test and fold change). Then, the WGCNA network was performed to identify modules of highly co-expressed genes, and 454 genes were discovered as the hub genes. As a result, the intersection between the DEGs and the hub genes was identified as the key genes. Using the independent dataset, three upregulated genes (ST3GAL5, NRCAM, and MT1F) and one downregulated gene (PROS1) were effectively validated. Moreover, the high-performance classification model was constructed for discriminating malignant from benign thyroid nodules. However, certain limitations still exist in this study. The number of samples for identifying and validating key genes was still needed to be increased. In the future, the key genes and classification model could be further verified based on the experimental data.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

YG designed research. QY performed research and wrote the scripts. QY and YG wrote the manuscript.

FUNDING

This work was funded by the National Natural Science Foundation of Jiangsu (BK20210597) and the NUPTSF (Grant No. NY220169).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.791349/full#supplementary-material

REFERENCES

Alexander, E. K., Kennedy, G. C., Baloch, Z. W., Cibas, E. S., Chudova, D., Diggans, J., et al. (2012). Preoperative Diagnosis of Benign Thyroid Nodules with Indeterminate Cytology. *N. Engl. J. Med.* 367, 705–715. doi:10.1056/NEJMoa1203208

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Burman, K. D., and Wartofsky, L. (2015). Thyroid Nodules. *N. Engl. J. Med.* 373, 2347–2356. doi:10.1056/NEJMcp1415786

Chen, M., Yan, J., Han, Q., Luo, J., and Zhang, Q. (2020). Identification of Hub-methylated Differentially Expressed Genes in Patients with Gestational Diabetes Mellitus by Multi-omic WGCNA Basing Epigenome-wide and Transcriptome-wide Profiling. *J. Cel Biochem.* 121, 3173–3184. doi:10.1002/jcb.29584

Cho, Y. Y., Park, S. Y., Shin, J. H., Oh, Y. L., Choe, J.-H., Kim, J.-H., et al. (2020). Highly Sensitive and Specific Molecular Test for Mutations in the Diagnosis of Thyroid Nodules: a Prospective Study of BRAF-Prevalent Population. *Ijms* 21, 5629. doi:10.3390/ijms21165629

Cibas, E. S., and Ali, S. Z. (2009). The bethesda System for Reporting Thyroid Cytopathology. *Thyroid* 19, 1159–1165. doi:10.1089/thy.2009.0274

Courcoutsakis, N., Patronas, N., Filie, A. C., Carney, J. A., Moraitis, A., and Stratakis, C. A. (2009). Ectopic Thymus Presenting as a Thyroid Nodule in a Patient with the Carney Complex. *Thyroid* 19, 293–294. doi:10.1089/thy.2008.0404

Durante, C., Costante, G., Lucisano, G., Bruno, R., Meringolo, D., Paciaroni, A., et al. (2015). The Natural History of Benign Thyroid Nodules. *JAMA* 313, 926–935. doi:10.1001/jama.2015.0956

Durante, C., Grani, G., Lamartina, L., Filetti, S., Mandel, S. J., and Cooper, D. S. (2018). The Diagnosis and Management of Thyroid Nodules. *JAMA* 319, 914–924. doi:10.1001/jama.2018.0898

Giordano, T. J., Au, A. Y. M., Kuick, R., Thomas, D. G., Rhodes, D. R., Wilhelm, K. G., Jr., et al. (2006). Delineation, Functional Validation, and Bioinformatic Evaluation of Gene Expression in Thyroid Follicular Carcinomas with the PAX8-PPARG Translocation. *Clin. Cancer Res.* 12, 1983–1993. doi:10.1158/1078-0432.CCR-05-2039

Górka, B., Skubis-Zegadło, J., Mikula, M., Bardadin, K., Paliczka, E., and Czarnocka, B. (2007). NrCAM, a Neuronal System Cell-Adhesion Molecule, Is Induced in Papillary Thyroid Carcinomas. *Br. J. Cancer* 97, 531–538. doi:10.1038/sj.bjc.6603915

Griffith, O. L., Melck, A., Jones, S. J. M., and Wiseman, S. M. (2006). Meta-analysis and Meta-Review of Thyroid Cancer Gene Expression Profiling Studies Identifies Important Diagnostic Biomarkers. *Jco* 24, 5043–5051. doi:10.1200/JCO.2006.06.7330

Han, L.-o., Li, X.-y., Cao, M.-m., Cao, Y., and Zhou, L.-h. (2018). Development and Validation of an Individualized Diagnostic Signature in Thyroid Cancer. *Cancer Med.* 7, 1135–1140. doi:10.1002/cam4.1397

He, P., Mo, X.-B., Lei, S.-F., and Deng, F.-Y. (2019). Epigenetically Regulated Co-expression Network of Genes Significant for Rheumatoid Arthritis. *Epigenomics* 11, 1601–1612. doi:10.2217/epi-2019-0028

Heider, A., Arnold, S., and Jing, X. (2020). Bethesda System for Reporting Thyroid Cytopathology in Pediatric Thyroid Nodules: Experience of a Tertiary Care Referral center. *Arch. Pathol. Lab. Med.* 144, 473–477. doi:10.5858/arpa.2018-0596-OA

Hinsch, N., Frank, M., Döring, C., Vorländer, C., and Hansmann, M.-L. (2009). QPRT: a Potential Marker for Follicular Thyroid Carcinoma Including Minimal Invasive Variant; a Gene Expression, RNA and Immunohistochemical Study. *BMC Cancer* 9, 93. doi:10.1186/1471-2407-9-93

Ito, Y., Miyauchi, A., Kihara, M., Higashiyama, T., Kobayashi, K., and Miya, A. (2014). Patient Age Is Significantly Related to the Progression of Papillary Microcarcinoma of the Thyroid under Observation. *Thyroid* 24, 27–34. doi:10.1089/thy.2013.0367

Jasim, S., Baranski, T. J., Teefey, S. A., and Middleton, W. D. (2020). Investigating the Effect of Thyroid Nodule Location on the Risk of Thyroid Cancer. *Thyroid* 30, 401–407. doi:10.1089/thy.2019.0478

Kim, H. S., Kim, D. H., Kim, J. Y., Jeoung, N. H., Lee, I. K., Bong, J. G., et al. (2010). Microarray Analysis of Papillary Thyroid Cancers in Korean. *Korean J. Intern. Med.* 25, 399–407. doi:10.3904/kjim.2010.25.4.399

Knyazeva, M., Korobkina, E., Karizky, A., Sorokin, M., Buzdin, A., Vorobyev, S., et al. (2020). Reciprocal Dysregulation of MiR-146b and MiR-451 Contributes in Malignant Phenotype of Follicular Thyroid Tumor. *Ijms* 21, 5950. doi:10.3390/ijms21175950

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

Medda, E., Santini, F., De Angelis, S., Franzellin, F., Fiumalbi, C., Perico, A., et al. (2017). Iodine Nutritional Status and Thyroid Effects of Exposure to Ethylenebisdithiocarbamates. *Environ. Res.* 154, 152–159. doi:10.1016/j.envres.2016.12.019

Mistry, M., Gillis, J., and Pavlidis, P. (2013). Genome-wide Expression Profiling of Schizophrenia Using a Large Combined Cohort. *Mol. Psychiatry* 18, 215–225. doi:10.1038/mp.2011.172

Moncada, R., Barkley, D., Wagner, F., Chiodin, M., Devlin, J. C., Baron, M., et al. (2020). Integrating Microarray-Based Spatial Transcriptomics and Single-Cell RNA-Seq Reveals Tissue Architecture in Pancreatic Ductal Adenocarcinomas. *Nat. Biotechnol.* 38, 333–342. doi:10.1038/s41587-019-0392-8

Niemira, M., Collin, F., Szalkowska, A., Bielska, A., Chwialkowska, K., Reszec, J., et al. (2019). Molecular Signature of Subtypes of Non-small-cell Lung Cancer by Large-Scale Transcriptional Profiling: Identification of Key Modules and Genes by Weighted Gene Co-expression Network Analysis (WGCNA). *Cancers* 12, 37. doi:10.3390/cancers12010037

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using Support Vector Machine to Identify Imaging Biomarkers of Neurological and Psychiatric Disease: a Critical Review. *Neurosci. Biobehavioral Rev.* 36, 1140–1152. doi:10.1016/j.neubiorev.2012.01.004

Osborn, D., Burton, A., Hunter, R., Marston, L., Atkins, L., Barnes, T., et al. (2018). Clinical and Cost-Effectiveness of an Intervention for Reducing Cholesterol and Cardiovascular Risk for People with Severe Mental Illness in English Primary Care: a Cluster Randomised Controlled Trial. *The Lancet Psychiatry* 5, 145–154. doi:10.1016/S2215-0366(18)30007-5

Park, H. J., Kim, J. H., Yoon, J. S., Choi, Y. J., Choi, Y.-H., Kook, K. H., et al. (2017). Identification and Functional Characterization of ST3GAL5 and ST8SIA1 Variants in Patients with Thyroid-Associated Ophthalmopathy. *Yonsei. Med. J.* 58, 1160–1169. doi:10.3349/ymj.2017.58.6.1160

Roth, M. Y., Witt, R. L., and Steward, D. L. (2018). Molecular Testing for Thyroid Nodules: Review and Current State. *Cancer* 124, 888–898. doi:10.1002/cncr.30708

Sato, K., Miyakawa, M., Onoda, N., Demura, H., Yamashita, T., Miura, M., et al. (1997). Increased Concentration of Vascular Endothelial Growth Factor/Vascular Permeability Factor in Cyst Fluid of Enlarging and Recurrent Thyroid Nodules1. *J. Clin. Endocrinol. Metab.* 82, 1968–1973. doi:10.1210/jcem.82.6.3989

Schulten, H.-J., Al-Mansouri, Z., Baghallab, I., Bagatian, N., Subhi, O., Karim, S., et al. (2015). Comparison of Microarray Expression Profiles between Follicular Variant of Papillary Thyroid Carcinomas and Follicular Adenomas of the Thyroid. *BMC Genomics* 16, S7. doi:10.1186/1471-2164-16-S1-S7

Schwalbe, E. C., Lindsey, J. C., Nakjang, S., Crosier, S., Smith, A. J., Hicks, D., et al. (2017). Novel Molecular Subgroups for Clinical Classification and Outcome

Prediction in Childhood Medulloblastoma: a Cohort Study. *Lancet Oncol.* 18, 958–971. doi:10.1016/S1470-2045(17)30243-7

Sepulveda, J. L. (2020). Using R and Bioconductor in Clinical Genomics and Transcriptomics. *J. Mol. Diagn.* 22, 3–20. doi:10.1016/j.jmoldx.2019.08.006

Singh Ospina, N., Iñiguez-Ariza, N. M., and Castro, M. R. (2020). Thyroid Nodules: Diagnostic Evaluation Based on Thyroid Cancer Risk Assessment. *BMJ* 368, l6670. doi:10.1136/bmj.l6670

Tippmann, S. (2015). Programming Tools: Adventures with R. *Nature* 517, 109–110. doi:10.1038/517109a

Tomczak, A., Mortensen, J. M., Winnenburg, R., Liu, C., Alessi, D. T., Swamy, V., et al. (2018). Interpretation of Biological Experiments Changes with Evolution of the Gene Ontology and its Annotations. *Sci. Rep.* 8, 5115. doi:10.1038/s41598-018-23395-2

Wojtas, B., Pfeifer, A., Oczko-Wojciechowska, M., Krajewska, J., Czarniecka, A., Kukulska, A., et al. (2017). Gene Expression (mRNA) Markers for Differentiating between Malignant and Benign Follicular Thyroid Tumours. *Ijms* 18, 1184. doi:10.3390/ijms18061184

Wojtczak, B., Pula, B., Gomulkiewicz, A., Olbromski, M., Podhorska-Okolow, M., Domoslawski, P., et al. (2017). Metallothionein Isoform Expression in Benign and Malignant Thyroid Lesions. *Ar* 37, 5179–5185. doi:10.21873/anticanres.11940

Wong, R., Farrell, S. G., and Grossmann, M. (2018). Thyroid Nodules: Diagnosis and Management. *Med. J. Aust.* 209, 92–98. doi:10.5694/mja17.01204

Wu, C.-C., Lin, J.-D., Chen, J.-T., Chang, C.-M., Weng, H.-F., Hsueh, C., et al. (2018). Integrated Analysis of fine-needle-aspiration Cystic Fluid Proteome, Cancer Cell Secretome, and Public Transcriptome Datasets for Papillary Thyroid Cancer Biomarker Discovery. *Oncotarget* 9, 12079–12100. doi:10.18632/oncotarget.23951

Wu, D., Hu, S., Hou, Y., He, Y., and Liu, S. (2020). Identification of Potential Novel Biomarkers to Differentiate Malignant Thyroid Nodules with Cytological Indeterminate. *BMC Cancer* 20, 199. doi:10.1186/s12885-020-6676-z

Yan, H., Zheng, G., Qu, J., Liu, Y., Huang, X., Zhang, E., et al. (2019). Identification of Key Candidate Genes and Pathways in Multiple Myeloma by Integrated Bioinformatics Analysis. *J. Cel Physiol.* 234, 23785–23797. doi:10.1002/jcp.28947

Yang, Q., Hong, J., Li, Y., Xue, W., Li, S., Yang, H., et al. (2020a). A Novel Bioinformatics Approach to Identify the Consistently Well-Performing Normalization Strategy for Current Metabolomic Studies. *Brief. Bioinform.* 21, 2142–2152. doi:10.1093/bib/bbz137

Yang, Q., Li, B., Chen, S., Tang, J., Li, Y., Li, Y., et al. (2021). MMEASE: Online Meta-Analysis of Metabolomic Data by Enhanced Metabolite Annotation, Marker Selection and Enrichment Analysis. *J. Proteomics* 232, 104023. doi:10.1016/j.jprot.2020.104023

Yang, Q., Li, B., Tang, J., Cui, X., Wang, Y., Li, X., et al. (2020b). Consistent Gene Signature of Schizophrenia Identified by a Novel Feature Selection Strategy from Comprehensive Sets of Transcriptomic Data. *Brief. Bioinform.* 21, 1058–1068. doi:10.1093/bib/bbz049

Yang, Q., Wang, Y., Zhang, S., Tang, J., Li, F., Yin, J., et al. (2019a). Biomarker Discovery for Immunotherapy of Pituitary Adenomas: Enhanced Robustness and Prediction Ability by Modern Computational Tools. *Ijms* 20, 151. doi:10.3390/ijms20010151

Yang, Q., Wang, Y., Zhang, Y., Li, F., Xia, W., Zhou, Y., et al. (2020c). NOREVA: Enhanced Normalization and Evaluation of Time-Course and Multi-Class Metabolomic Data. *Nucleic Acids Res.* 48, W436–W448. doi:10.1093/nar/gkaa258

Yang, Q. X., Wang, Y. X., Li, F. C., Zhang, S., Luo, Y. C., Li, Y., et al. (2019b). Identification of the Gene Signature Reflecting Schizophrenia's Etiology by Constructing Artificial Intelligence-based Method of Enhanced Reproducibility. *CNS Neurosci. Ther.* 25, 1054–1063. doi:10.1111/cns.13196

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

Yu, H., Xing, S., Nierves, L., Lange, P. F., and Huan, T. (2020). Fold-change Compression: an Unexplored but Correctable Quantitative Bias Caused by Nonlinear Electrospray Ionization Responses in Untargeted Metabolomics. *Anal. Chem.* 92, 7011–7019. doi:10.1021/acs.analchem.0c00246

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

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

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 14 February 2022
doi: 10.3389/fmolb.2021.758742

Differential Expression of Serum TUG1, LINC00657, miR-9, and miR-106a in Diabetic Patients With and Without Ischemic Stroke

Omayma O Abdelaleem[1], Olfat G. Shaker[2], Mohamed M. Mohamed[3], Tarek I. Ahmed[4], Ahmed F. Elkhateeb[5], Noha K. Abdelghaffar[6], Naglaa A. Ahmed[7], Abeer A. Khalefa[7], Nada F. Hemeda[8] and Rania H. Mahmoud[1]*

[1]Department of Medical Biochemistry and Molecular Biology, Faculty of Medicine, Fayoum University, Fayoum, Egypt, [2]Department of Medical Biochemistry and Molecular Biology, Faculty of Medicine, Cairo University, Giza, Egypt, [3]Department of Internal Medicine, Faculty of Medicine, Cairo University, Giza, Egypt, [4]Department of Internal Medicine, Faculty of Medicine, Fayoum University, Fayoum, Egypt, [5]Department of Critical Care, Faculty of Medicine, Fayoum University, Fayoum, Egypt, [6]Department of Clinical Pathology, Faculty of Medicine, Fayoum University, Fayoum, Egypt, [7]Department of Physiology, Faculty of Medicine, Zagazig University, Zagazig, Egypt, [8]Department of Genetics, Faculty of Agriculture, Fayoum University, Fayoum, Egypt

OPEN ACCESS

Edited by:
Jie Li,
Harbin Institute of Technology, China

Reviewed by:
Yinan Jiang,
University of Pittsburgh, United States
Vikram Dalal,
Washington University in St. Louis, United States

***Correspondence:**
Omayma O Abdelaleem
dr.omayma@yahoo.com

Specialty section:
This article was submitted to Molecular Diagnostics and Therapeutics, a section of the journal Frontiers in Molecular Biosciences

Received: *14 August 2021*
Accepted: *24 December 2021*
Published: *14 February 2022*

Citation:
Abdelaleem OO, Shaker OG, Mohamed MM, Ahmed TI, Elkhateeb AF, Abdelghaffar NK, Ahmed NA, Khalefa AA, Hemeda NF and Mahmoud RH (2022) Differential Expression of Serum TUG1, LINC00657, miR-9, and miR-106a in Diabetic Patients With and Without Ischemic Stroke. Front. Mol. Biosci. 8:758742. doi: 10.3389/fmolb.2021.758742

Background: Ischemic stroke is one of the serious complications of diabetes. Non-coding RNAs are established as promising biomarkers for diabetes and its complications. The present research investigated the expression profiles of serum TUG1, LINC00657, miR-9, and miR-106a in diabetic patients with and without stroke.

Methods: A total of 75 diabetic patients without stroke, 77 patients with stroke, and 71 healthy controls were recruited in the current study. The serum expression levels of TUG1, LINC00657, miR-9, and miR-106a were assessed using quantitative real-time polymerase chain reaction assays.

Results: We observed significant high expression levels of LINC00657 and miR-9 in the serum of diabetic patients without stroke compared to control participants. At the same time, we found marked increases of serum TUG1, LINC00657, and miR-9 and a marked decrease of serum miR-106a in diabetic patients who had stroke relative to those without stroke. Also, we revealed positive correlations between each of TUG1, LINC00657, and miR-9 and the National Institutes of Health Stroke Scale (NIHSS). However, there was a negative correlation between miR-106a and NIHSS. Finally, we demonstrated a negative correlation between LINC00657 and miR-106a in diabetic patients with stroke.

Conclusion: Serum non-coding RNAs, TUG1, LINC00657, miR-9, and miR-106a displayed potential as novel molecular biomarkers for diabetes complicated with stroke, suggesting that they might be new therapeutic targets for the treatment of diabetic patients with stroke.

Keywords: TUG1, LINC00657, miR-9, miR-106a, stroke

INTRODUCTION

Diabetes mellitus (DM) is a complex, multisystem disease and is one of the risk factors of stroke. Oxidative stress occurs due to an elevated blood glucose level, which is associated with increased glycated end products, resulting in endothelial dysfunction, cerebrovascular atherosclerosis, and thrombosis, which are the main causes of ischemic stroke in diabetic patients that is associated with high mortality and poor prognosis (Nakagami et al., 2005; Ferreiro et al., 2010).

It is important to understand the molecular mechanisms of cerebral stoke associated with DM to facilitate the development of new effective potential biomarkers and therapeutic targets for diabetic patients with stroke.

Non-coding RNAs, including microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), have been proven to have necessary roles in regulating gene expression (Kitagawa et al., 2013; Iyengar et al., 2014; Khoshnam et al., 2017). LncRNAs are an important group of non-coding RNAs that are of long transcripts (>200 bp). Numerous studies have elucidated the significant roles of lncRNAs in various diseases, including ischemic stroke (Mercer and Mattick, 2013; Bao et al., 2018a).

Taurine-upregulated gene 1 (TUG1), a lncRNA, has been shown to be related to the pathogenesis of many diseases. Recently, TUG1 has gained significant attention in ischemic injuries (Long et al., 2016; Chen et al., 2017; Wang et al., 2017), although little has been identified regarding its role in DM complicated with stroke.

LINC00657 is a lncRNA that is highly conserved and profusely expressed in endothelial cells (Michalik et al., 2014). Accumulating evidence has demonstrated that LINC00657 might play an oncogenic role, and it is upregulated in many cancers (Liu H. et al., 2016; Liu S. et al., 2016). However, its role in DM or diabetes-related complications has not been investigated yet.

MicroRNAs (miRNAs) are tiny non-coding RNAs (20–25 nucleotides long). Recently, promising research studies have explained the importance of miRNAs in the pathogenesis of diabetes and its cardiovascular complications (Meng et al., 2012; Koutsis et al., 2013). However, the expressions of miR-9 and miR-106a in diabetic patients with ischemic stroke have not been examined.

Bioinformatics has reported that TUG1 has complementary sequences of miR-9. Additionally, LINC00657 contains binding sites for miR-106a (Li et al., 2014). However, a study of their relationship in DM complicated with cerebral stroke remains to be conducted.

In this study, we aimed to assess the serum expression levels of TUG1, LINC00657, miR-9, and miR-106a in diabetic patients who had stroke and those without cerebral stroke and to explore any association between these non-coding RNAs and clinico-laboratory data.

Subjects and Methods

Study Population

A total of 152 diabetic patients (with type 2 diabetes) were recruited among those admitted to the outpatient and inpatient clinics of the Internal Medicine and Intensive Care Unit, Fayoum University Hospital, Fayoum, in the period from November 2019 to December 2020. Diabetic patients were selected based on the American Diabetes Association 2015 diagnostic criteria (Pinsker et al., 2015). Diabetic patients were divided into two groups: diabetic patients with stroke (30females and 47males, with a mean = 57.08 ± 16.31 years) and diabetic patients without stroke (26 females and 49 males, mean age = 53.19 ± 17.78 years) (**Figure 1**).

Ischemic stroke diagnosis was assessed according to clinical symptoms and physical examinations, and this diagnosis was confirmed by computed tomography (CT) or magnetic resonance imaging (MRI). The National Institutes of Health Stroke Scale (NIHSS) was used by experienced neurologists to evaluate the neurological deficits.

All patients with brain tumors, intracerebral hemorrhage, recurrent stroke, history of hypertension, recent head injuries, immune system disorders, liver or renal diseases, blood diseases, acute infectious diseases, or a family history of stroke were excluded from the study. Furthermore, 71 healthy individuals (age and sex matched to the patients) who did not have systemic or neurologic diseases were considered as the control group in this study.

Written informed consent was signed by all enrolled participants after a detailed explanation of the study. The study protocol was performed in agreement with the Declaration of Helsinki. The Ethics Committee of the Faculty of Medicine, Fayoum University, approved this research protocol.

Serum Collection

From each participant, 5 ml of venous blood was collected into plain tubes using a Vacutainer system following a 12-h fast. Serum separator tubes were used to collect the samples that were left for 15 min to clot. Centrifugation at 4,000 × *g* for 10 min was performed, separating the serum that was stored at −80°C until the time of use. An extra blood sample was taken 2 h after a meal (2 h post-prandial, 2hPP) into tubes containing fluoride.

Fasting blood glucose (FBG), 2hPP blood glucose, cholesterol, triglycerides, HbA1C, creatinine, and low-density lipoprotein (LDL) were assessed using standard methods on cobas c311 (Roche, Mannheim, Germany) in accordance with the instructions in the kit. Serum samples were used for the quantification of TUG1, LINC00657, miR-9, and miR-106a using real-time PCR.

LncRNA and miRNA Extraction and Reverse Transcription

According to the manufacturer's protocol, total RNA (including lncRNAs and miRNAs) was extracted from serum samples using the miRNeasy extraction kit (Qiagen, Hilden, Germany) after adding the QIAzol lysis reagent. Quantitation and the purity of the RNA samples were assessed using the NanoDrop® (ND)-1000 spectrophotometer (NanoDrop Technologies, Inc., Wilmington, DE, USA).

Complementary DNA (cDNA) was generated from the extracted RNA in a total volume of 20 µl/reaction using the

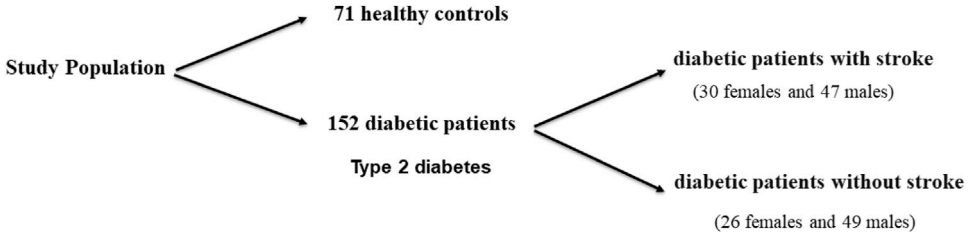


FIGURE 1 | Schematic diagram of the outline of the work performed in this study.

TABLE 1 | Baseline characteristics of the enrolled groups

	Control (*n* = 71)	Diabetes		*p*-value[a]	*p*-value[b]
		Without stroke (*n* = 75)	With stroke (*n* = 77)		
Sex, *n* (%)					
Female	25 (35.21%)	26 (34.67%)	30 (38.96%)	0.441	0.352
Male	56 (64.79%)	49 (65.33%)	47 (61.04%)	0.553	0.907
Age (years)	54.58 ± 18.75	53.19 ± 17.78	57.08 ± 16.31	0.894	0.559
BMI (kg/m^2)	29.37 ± 1.82	30.07 ± 2.46	31.89 ± 2.09	0.389	0.604
FBG (mg/dl)	83.25 ± 8.97	154.58 ± 28.11	185.41 ± 40.85	**<0.001***	**0.04***
2hPP (mg/dl)	111.85 ± 10.24	255.75 ± 47.08	309.15 ± 53.29	**<0.001***	**0.04***
HbA1c (%)	4.27 ± 1.65	7.87 ± 2.34	9.07 ± 3.12	**<0.001***	**0.02***
ALT (IU/L)	18.74 ± 3.89	37.25 ± 9.27	40.09 ± 8.97	**<0.001***	0.07
AST (IU/L)	17.92 ± 8.17	32.25 ± 6.87	35.51 ± 8.71	**0.002***	0.425
Urea (mg/dl)	24.71 ± 8.74	55.19 ± 11.31	58.18 ± 9.47	**<0.001***	0.108
Creatinine (mg/dl)	0.70 ± 0.19	2.72 ± 0.34	3.09 ± 0.17	**0.02***	0.094
Hb (gm/dl)	12.13 ± 3.24	11.89 ± 2.98	12.01 ± 3.07	0.498	0.571
MCV	33.12 ± 2.19	33.09 ± 2.01	32.97 ± 2.05	0.608	0.580
MCH	28.11 ± 2.31	27.98 ± 3.07	29.01 ± 2.13	0.333	0.231
Cholesterol (mg/dl)	138.15 ± 23.19	168.25 ± 18.52	198.16 ± 34.08	**0.008***	0.06
LDL (mg/dl)	49.57 ± 19.87	86.17 ± 15.79	101.71 ± 25.55	**<0.001***	**0.03***
HDL (mg/dl)	41.22 ± 8.99	35.12 ± 7.58	30.09 ± 8.88	**0.03***	0.064
Triglycerides (mg/dl)	65.13 ± 9.13	137.32 ± 35.62	149.85 ± 44.73	**<0.001***	0.091
NIHSS	–	–	11.38 ± 5.12	–	–
Disease duration (years)	–	13.35 ± 1.87	15.729 ± 1.97	–	0.09

Data are shown as the mean ± (SD, median (range), or n *(%).*

BMI, *body mass index;* FBG, *fasting blood glucose;* 2hPP, *2 h post-prandial;* HbA1c, *glycated hemoglobin A1c;* ALT, *alanine transaminase;* AST, *aspartate transaminase;* Hb, *hemoglobin;* MCV, *mean corpuscular volume;* MCH, *mean corpuscular hemoglobin;* LDL, *low-density lipoprotein;* HDL, *high-density lipoprotein;* NIHSS, *National Institutes of Health Stroke Scale*

**Significant at* p < *0.05*

[a]*Comparison of diabetic patients (with and without stroke)* versus *the healthy control group*

[b]*Comparison of diabetic patients with stroke* versus *diabetic patients without stroke*

RT^2 First Strand Kit (Qiagen, Germantown, MD, USA) according to the manufacturer's protocol for lncRNA expression analysis. Moreover, the miScript II RT Kit (Qiagen, Valencia, CA, US) was used for miRNA expression analysis in a 20-μl reverse transcription (RT) reaction according to the instructions in the pamphlet.

LncRNA and miRNA Expression by Real-Time Quantitative PCR

Real-time PCR amplification reactions were performed using the RT^2 SYBR Green PCR Kit (Qiagen, Germantown, MD, USA) for the detection of lncRNA. However, the miScript SYBR Green PCR Kit (Qiagen, Valencia, CA, USA) was used in the quantification of miRNAs with the aid of the Rotor-Gene Q System (Qiagen).

The RefSeq accession no. of TUG1 was NR_002323.2 and that of LINC00657was NR_027451.1. GAPDH was used as an endogenous control for the evaluation of TUG1 and LINC00657 according to the manufacturer's instructions. Numerous studies have used GAPDH as an internal reference for serum lncRNAs (Duan et al., 2016; Shaker et al., 2019). The primer sequences of GAPDH were as follows: forward: 5′-CCCTTCATTGACCTCAAC TA-3′; reverse: 5-′TGGAAGATGGTGATGGGATT-3′.

Moreover, the catalog number of miR-9 was MS00010752 and that of miR-106a was MS00008393. SNORD68 was used as the internal reference for the evaluation of the gene expression levels of miR-9 and miR-106a. The catalog number of *SNORD68* was MS00033712.

The PCR cycling program for the quantification of lncRNAs consists of an initial incubation at 95°C for 10 min, followed by 40

TABLE 2 | Expression levels of serum TUG1, LINC00657, miR-9, and miR-106a in all groups

Variables	Diabetes without stroke (n = 75)	Diabetes with stroke (n = 77)	p-value
	Median (intraquartile range)		
TUG1	0.71 (0.01–1.97)	2.90 (0.31–13.25)	0.125[a] < **0.001**[*b,c]
LINC00657	3.18 (0.87–20.98)	11.85 (0.50–53.85)	**0.001**[*a] < **0.001**[*b,c]
miR-9	1.45 (0.12–8.07)	4.40 (0.35–12.25)	**0.001**[*a] < **0.001**[*b]**0.003**[*c]
miR-106a	0.760 (0.13–1.62)	0.03 (0.01–0.41)	0.09[a] < **0.001**[*b]0.05[c]

Fold change levels represent non-coding RNA expression relative to controls that were calculated using $2^{-\Delta\Delta CT}$. Control fold change levels are equivalent to 1. Data are expressed as the median and intraquartile range. Adjusted p-values for multiple comparisons of the studied groups were estimated using the Bonferroni correction method.

*****Significant at* p < 0.017*

[a]*Comparison of diabetes without stroke* versus *healthy controls*

[b]*Comparison of diabetes with stroke* versus *healthy controls*

[c]*Comparison of diabetes with stroke* versus *diabetes without stroke*

cycles at 95°C for 15 s and 60°C for 60 s. That for the detection of miRNAs consists of 95°C for 30 min, followed by 40 cycles at 94°C for 15 s, 55°C for 30 s, and 70°C for 30 s.

The relative expression levels of TUG1, LINC00657, miR-9, and miR-106a were calculated using $2^{-\Delta\Delta Ct}$. Fold change (FC) values less than 1 indicated downregulation, while values more than 1 indicated upregulation of non-coding RNAs (Livak and Schmittgen, 2001). Control FC values were set as 1.

Statistical Analyses

Statistical analysis was performed using Statistical Package for Social Sciences (SPSS) version 24. The mean, standard deviation (SD), median, and interquartile range (IQR) were utilized to represent quantitative data. A chi-square test was performed for categorical data. However, the Mann–Whitney U test was used for continuous variables, which were presented as median (interquartile range). To determine the relation of the expressions of non-coding RNAs with the study parameters, Spearman's correlation was run. A multivariate stepwise logistic regression was constructed to identify the significant predictors of cerebral stroke among the four markers.

Analyses of the receiver operating characteristic (ROC) curves were conducted to determine the sensitivity and specificity of TUG1, LINC00657, miR-9, and miR-106a as predictors in differentiating between different groups. Statistical significance was considered at a p-value <0.05. Adjusted p-values for multiple comparisons of the studied groups were estimated using the Bonferroni correction method. The p-value (of 0.05) was divided by the number of comparisons, i.e., 3 (0.05/3). Therefore, the test results were considered to be statistically significant at p-values <0.017.

RESULTS

Clinical and Laboratory Features of the Enrolled Participants

A total of 75 diabetic patients without stroke, 77 diabetic patients with stroke, and 71 healthy individuals were included in the current study.

There were marked differences between the total diabetic patients (with and without stroke) and the healthy group regarding FBG, 2hPP, HbAIc, alanine transaminase (ALT), aspartate transaminase (AST), urea, creatinine, cholesterol, LDL, high-density lipoprotein (HDL), and triglycerides (all p < 0.05). However, no significant differences in age, sex, and other clinical and laboratory data were observed between all diabetic patients and control participants (p > 0.05) (**Table 1**). Moreover, there were significant differences concerning FBG, 2hPP, HbAIc, and LDL when comparing diabetic patients with stroke to those without stroke (all p < 0.05). On the other hand, there were no marked differences regarding age, sex, and all other data when comparing diabetic patients with stroke to those without stroke (p > 0.05) (**Table 1**).

Comparison of the Serum Expression Levels of TUG1, LINC00657, miR-9, and miR-106a in the Different Studied Groups

As clarified in **Table 2**, the serum expression levels of LINC00657 and miR-9 were increased significantly in diabetic patients without stroke when compared to healthy individuals (p = 0.001 for LINC00657 and miR-9).

We next compared the expressions of TUG1, LINC00657, miR-9, and miR-106a in the sera of diabetic patients with stroke relative to healthy controls. The results showed significant upregulation of TUG1, LINC00657, and miR-9 (p < 0.001 for TUG1, LINC00657, and miR-9). In contrast, the level of miR-106a in serum was markedly decreased in diabetic patients who had stroke relative to the control subjects (p < 0.001).

Furthermore, we revealed a marked elevation of the serum expressions of TUG1, LINC00657, and miR-9 in diabetic patients with stroke relative to those without stroke (p < 0.001 for TUG1 and LINC00657; p = 0.003 for miR-9). Meanwhile, a non-significant decrease of miR-106a was detected between diabetic patients with stroke and those without stroke (p = 0.05).

Correlation of TUG1, LINC00657, miR-9, and miR-106a With Stroke Severity and Clinical Characteristics

NIHSS scoring was performed to evaluate stroke severity. We used Spearman's analysis to assess the correlation between the

TABLE 3 | Correlation between the expression levels of serum non-coding RNAs and clinical parameters in diabetic patients with stroke

Variables	TUG1	LINC00657	miR-9	miR-106a
Disease duration	**0.754 (<0.001)***	**0.720 (<0.001)***	**0.675 (<0.001)***	**−0.600 (<0.001)***
NIHSS	**0.802 (<0.001)***	**0.709 (<0.001)***	**0.681 (<0.001)***	**−0.569 (0.001)***
Age	0.097 (0.821)	−0.074 (0.893)	0.107 (0.275)	−0.197 (0.104)
BMI	0.099 (0.752)	0.055 (0.708)	0.122 (0.564)	−0.189 (0.262)
FBG	0.017 (0.920)	0.213 (0.206)	0.109 (0.523)	0.269 (0.107)
2hPP	0.094 (0.578)	0.172 (0.310)	0.102 (0.548)	0.114 (0.501)
HbA1c	−0.014 (0.779)	0.098 (0.587)	0.073 (0.669)	−0.107 (0.527)
AST	0.034 (0.839)	0.098 (0.565)	0.013 (0.941)	0.277 (0.096)
ALT	−0.123 (0.467)	−0.187 (0.267)	−0.085 (0.618)	−0.264 (0.114)
Urea	−0.076 (0.653)	−0.146 (0.388)	−0.123 (0.467)	−0.138 (0.416)
Creatinine	−0.127 (0.454)	−0.115 (0.499)	−0.125 (0.460)	0.069 (0.686)
Cholesterol	0.101 (0.552)	0.136 (0.421)	0.187 (0.268)	−0.131 (0.441)
LDL	0.037 (0.830)	−0.053 (0.755)	−0.052 (0.684)	0.171 (0.244)
HDL	−0.012 (0.890)	0.113 (0.474)	0.124 (0.466)	−0.165 (0.328)
Triglycerides	0.051 (0.765)	0.011 (0.872)	0.029 (0.865)	−0.097 (0.524)

BMI, *body mass index;* FBG, *fasting blood glucose;* 2hPP, *2 h post-prandial;* HbA1c, *glycated hemoglobin A1c;* ALT, *alanine transaminase;* AST, *aspartate transaminase;* Hb, *hemoglobin;* LDL, *low-density lipoprotein;* HDL, *high-density lipoprotein;* NIHSS, *National Institutes of Health Stroke Scale*
**Significant at* $p < 0.05$

aforementioned ncRNAs and stroke severity, as well as clinical and laboratory data.

As demonstrated in **Table 3**, TUG1, LINC00657, and miR-9 were positively correlated with NIHSS ($r = 0.802$, $p < 0.001$; $r = 0.709$, $p < 0.001$; and $r = 0.681$, $p < 0.001$, respectively). At the same time, a negative correlation was shown between miR-106a and NIHSS ($r = -0.569$, $p = 0.001$). In addition, TUG1, LINC00657, and miR-9 were positively correlated with disease duration ($r = 0.754$, $p < 0.001$; $r = 0.720$, $p < 0.001$; and $r = 0.675$, $p < 0.001$, respectively). On the other hand, a negative correlation was observed between miR-106a and years since the occurrence of DM ($r = -0.600$, $p < 0.001$).

However, no significant correlation was detected between TUG1, LINC00657, miR-9, and miR-106a and the laboratory parameters in the present study (all $p > 0.05$).

Correlation of TUG1 With miR-9 and of LINC00657 With miR-106a

Interestingly, the current results reported a negative correlation between LINC00657 and miR-106a ($r = -0.507$, $p = 0.002$) in diabetic patients with stroke. However, no significant correlation was shown between the serum levels of TUG1 and miR-9 ($r = 0.251$, $p = 0.10$).

ROC Analysis to Determine the Diagnostic Performance of Serum TUG1, LINC00657, miR-9, and miR-106a in Distinguishing Diabetic Patients With Stroke From Control Subjects

An ROC curve was assembled to estimate the diagnostic value of TUG1, LINC00657, miR-9, and miR-106a as novel biomarkers for DM with stroke relative to healthy subjects. For TUG1, the AUC was 0.758 (95% CI = 0.669–0.846, $p < 0.001$), with a sensitivity of 48.50% and a specificity of 100%. Moreover, the AUC of LINC00657 was 0.892 (95% CI = 0.834–0.950, $p < 0.001$), with a sensitivity of 73.50% and a specificity of 100%. Also, the AUC of miR-9 was 0.755 (95% CI = 0.677–0.834, $p < 0.001$), with a sensitivity of 39.5% and a specificity of 100%. Regarding miR-106a, its AUC was 0.674 (95% CI = 0.583–0.765, $p < 0.001$), and the sensitivity and specificity were 38.4% and 100%, respectively. On the other hand, the AUC of LDL was 0.979 (95% CI = 0.962–0.996, $p < 0.001$), with sensitivity of 87.5% and specificity of 45.8% (**Table 4** and **Figure 2**).

ROC curve analysis revealed that serum TUG1, LINC00657, miR-9, and miR-106a have good value as prognostic markers in discriminating diabetic patients with stroke from those without stroke.

The current results demonstrated that using TUG1 to diagnose diabetes with stroke yielded an AUC of 0.954 (95% CI = 0.915–0.994, $p < 0.001$), with a sensitivity of 87.9% and a specificity of 98.5%. In addition, the AUC value for LINC00657 was 0.902 (95% CI = 0.847–0.957, $p < 0.001$), with a sensitivity of 35.1% and a specificity of 98.5%. Also, miR-9 had an AUC of 0.661 (95% CI = 0.571–0.752, $p < 0.001$) and sensitivity and specificity values of 39.0% and 93.5%, respectively. For miR-106a, the AUC was 0.747 (95% CI = 0.661–0.832, $p = 0.01$), with a sensitivity of 39.0% and a specificity of 100%, while the AUC of LDL was 0.736 (95% CI = 0.654–0.819, $p < 0.001$) and the sensitivity and specificity were 20% and 90.2%, respectively (**Table 5** and **Figure 3**).

Multiple Logistic Regression Analysis

Multivariate regression analysis (considering NIHSS as the dependent variable) confirmed that TUG1 and LINC00657 were independent predictors for diabetes with stroke ($p = 0.04$ and $p = 0.01$, respectively) (**Table 6**).

DISCUSSION

Type 2 diabetes mellitus (T2DM) has emerged as a cause of serious concern worldwide and has been established as a risk factor for ischemic stroke (Nakagami et al., 2005). It is important

TABLE 4 | Receiver operating characteristics (ROC) curve analysis using serum TUG1, LINC00657, miR-9, miR-106a, and LDL for discriminating diabetic patients with stroke from control subjects

Variable	AUC (95% CI)	*p*-value	Sensitivity (%)	Specificity (%)	Total accuracy
TUG1	0.758 (0.669–0.846)	<0.001*	48.50	100	74.25
LINC00657	0.892 (0.834–0.950)	<0.001*	73.50	100	86.75
miR-9	0.755 (0.677–0.834)	<0.001*	39.5	100	69.75
miR-106a	0.674 (0.583–0.765)	<0.001*	38.4	100	69.20
LDL	0.979 (0.962–0.996)	<0.001*	87.5	45.8	66.65

AUC, *area under the curve;* CI, *confidence interval;* LDL, *low-density lipoprotein*
**Significant at* $p < 0.05$.

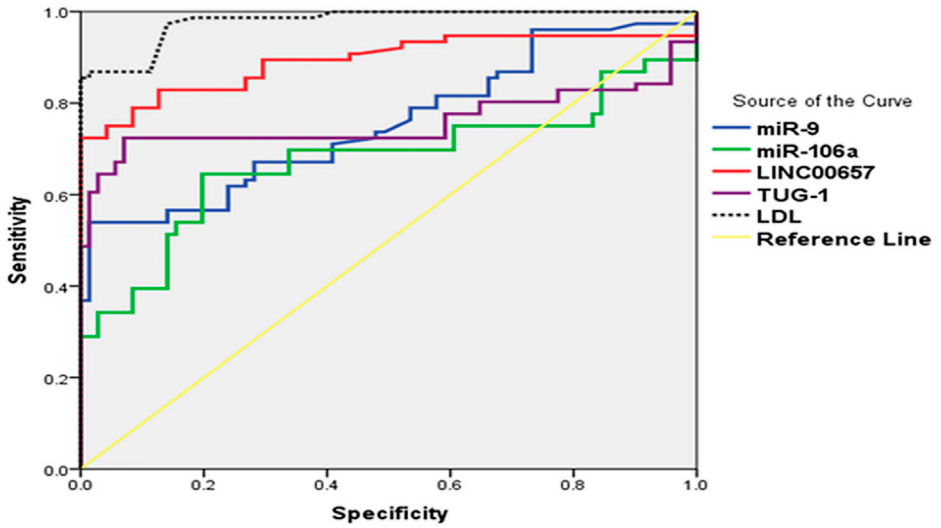


FIGURE 2 | (ROC) curve analysis of serum TUG1, LINC00657, miR-9, and miR-106a for distinguishing diabetic patients without stroke from control subjects.

to discover new sensitive and easily detected biomarkers for the diagnosis and prognosis of T2DM and its related complications.

Recently, it has been shown that non-coding RNAs (including lncRNAs and microRNAs) may be used as probable biomarkers for T2DM and its associated complications due to their stability and differential expression in a variety of body fluids, such as plasma and serum (Mastropasqua et al., 2014; Shaker et al., 2019). However, no previous reports have investigated the role of TUG1, LINC00657, miR-9, and miR-106a in stroke associated with DM. Thus, in this article, we assessed the serum expression levels of TUG1, LINC00657, miR-9, and miR-106a in diabetic patients with and without stroke.

We observed low TUG1 and miR-106a and significantly high LINC00657 and miR-9 expression levels in the serum of diabetic patients without stroke compared to control participants. At the same time, we verified marked increases of serum TUG1, LINC00657, and miR-9 and a marked decrease of serum miR-106a in diabetic patients who had stroke relative to those without stroke. Previous studies reported that the expression of TUG1 was decreased in rats with diabetes and in mesangial cells induced with high-level glucose through inhibition of the PI3K/AKT pathway (Zang et al., 2019). Furthermore, Wang et al. showed that TUG1 was downregulated in NRK-52E cells (high-glucose-stimulated) in mice *via* targeting miR-21 (Wang et al., 2019). Similarly, Li et al. documented a low expression level of TUG1 in high-glucose-stimulated podocytes by hindering the expression of miR-27a-3p (Li et al., 2019).

Our results regarding the upregulation of TUG1 in diabetic patients who had stroke are in line with recent studies showing that TUG1 was overexpressed in ischemic stroke by regulating miR-9 and decreasing Bcl-2-like 11 protein [25]. Also, in atherosclerosis, the elevated expression level of TUG1 increased endothelial cell apoptosis through miR-26a sponging (Chen et al., 2016). Similarly, many recent studies have discussed the role of TUG1 in atherosclerosis. For example, Li et al. found that TUG1, *via* regulating the miR-21/PTEN axis, increased the proliferation of vascular smooth muscle (Li et al., 2018). In addition, Yan et al. documented the role of TUG1 in the migration and proliferation of endothelial cells by the Wnt pathway (Yan et al., 2018). Moreover, Zhang et al. reported that TUG1 knockdown ameliorated atherosclerotic lesion and inhibited inflammation and hyperlipidemia *via* the upregulation of fibroblast growth factor 1 (Zhang et al., 2018). Besides, Yang et al. noted an increased expression level of TUG1 in ischemic heart exposed to oxidative stress *via* increasing cardiomyocyte apoptosis (Yang et al., 2019).

TABLE 5 | Receiver operating characteristics (ROC) curve analysis using serum TUG1, LINC00657, miR-9, and miR-106a for discriminating diabetic patients with stroke from diabetic patients without stroke

Variable	AUC (95% CI)	*p*-value	Sensitivity (%)	Specificity (%)	Total accuracy
TUG1	0.954 (0.915–0.994)	<0.001*	87.9	98.5	93.2
LINC00657	0.902 (0.847–0.957)	<0.001*	35.1	98.5	66.8
miR-9	0.661 (0.571–0.752)	0.001*	39.0	93.5	51.25
miR-106a	0.747 (0.661–0.832)	<0.001*	39.0	100	69.5
LDL	0.736 (0.654–0.819)	<0.001*	20	90.2	55.1

AUC, *area under the curve;* CI, *confidence interval;* LDL, *low-density lipoprotein*
**Significant at* $p < 0.05$.

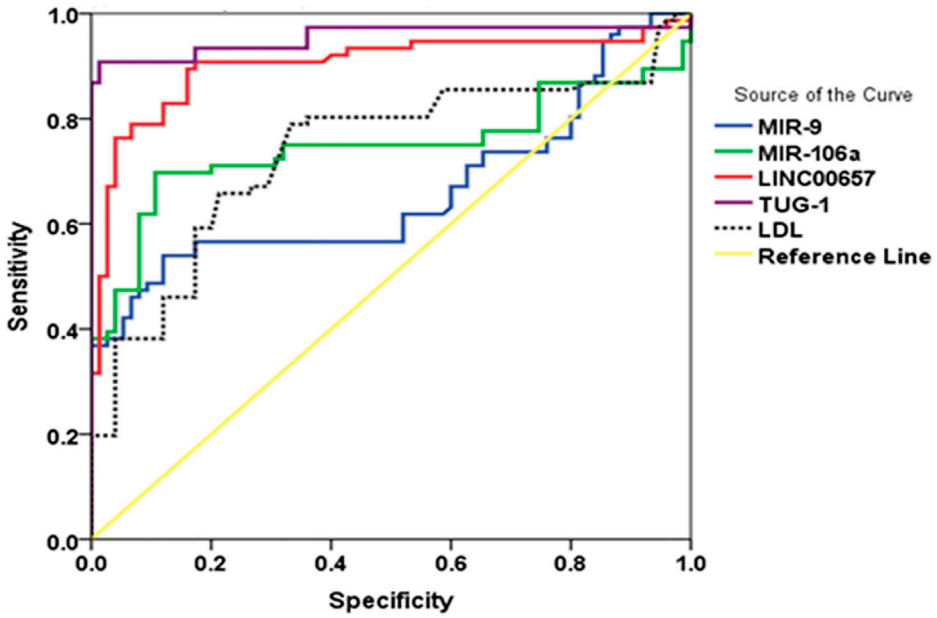


FIGURE 3 | (ROC) curve analysis of serum TUG1, LINC00657, miR-9, and miR-106a for discriminating diabetic patients with stroke from those without stroke.

However, there are no reports on the relation between TUG1 and stoke associated with DM.

Regarding LINC00657 (NORAD), our results are in line with a study which found that LINC00657, which is expressed in vascular endothelial cells, induced angiogenesis during atherosclerosis through the upregulation of VEGF, MMP-2, and MMP-9 (Wan et al., 2020). Also, Michalik et al. revealed that LINC00657 was markedly elevated during hypoxia (Michalik et al., 2014). Of note is that Bao et al. reported that oxidized LDL (oxLDL) treatment, which promotes oxidative stress and is implicated in atherosclerosis, resulted in the overexpression of LINC00657 (Bao et al., 2018b). Since hypoxia and atherosclerosis are predisposing factors of ischemic stroke, we therefore assumed that LINC00657 might contribute to the pathogenesis of stroke.

In the current research, we assessed the expression level of miR-9, which is a target gene of TUG1 and miR-106a, which are target genes of LINC00657. It was revealed in previous studies that the serum expression level of miR-9 increased significantly in T2DM (Kong et al., 2011), which is in accordance with our results. Furthermore, miR-9 was found to decrease insulin secretion *via* targeting syntaxin-binding protein 1, Onecut 2, and sirtuin 1 (Sirt1) (Plaisance et al., 2006; Ramachandran et al., 2011; Hu et al., 2018). However, Jiménez-Lucena et al. reported a low plasma level of miR-9 in patients at risk of T2DM (Jiménez-Lucena et al., 2018).

More importantly, previous studies also demonstrated the role of miR-9 in ischemic stroke, such as Ji et al. who found that miR-9 was upregulated in the serum exosomes of patients with acute ischemic stroke and was strongly associated with interleukin 6 (IL-6) production (Ji et al., 2016). In addition, the serum expression level of miR-9 was verified to be elevated significantly in acute ischemic stroke patients and was positively correlated with inflammatory markers, infarct volume, and the NIHSS score (Ji et al., 2016). Besides, another study has considered miR-9 to be a new biomarker of neurotoxicity and neural damage (Xue et al., 2018). At the same time, an increasing number of studies have explained the role of miR-9 in neuronal apoptosis after ischemic stroke (Wei et al., 2016).

TABLE 6 | Multiple logistic regression analysis

	B	SE	*p*-value	95% CI for *B*	
				Lower	Upper
TUG1	–5.980	1.071	**0.02***	2.82	11.2
LINC00657	–19.248	2.09	**0.03***	1.02	5.54
miR-9	–0.633	0.115	0.124	0.91	2.11
miR-106a	4.135	3.294	0.247	0.44	5.52
LDL	0.012	0.007	0.078	–0.001	0.025
Disease duration	0.243	0.082	**0.004***	0.080	0.406
FBG	–0.005	0.017	0.786	–0.038	0.029
HbA1c	–0.311	–0.196	0.117	–0.702	0.080
Constant	9.464	3.688	0.013	2.100	16.828

CI, *confidence interval;* LDL, *low-density lipoprotein;* FBG, *fasting blood glucose;* 2hPP, *2 h post-prandial;* HbA1c, *glycated hemoglobin A1c*
**Significant at* $p < 0.05$

Concerning miR-106a, our findings are in line with the study by Wu et al., which determined that miR-106a was decreased in diabetic peripheral neuropathy *via* the regulation of 12/15-lipoxygenase of oxidative/nitrative stress (Wu et al., 2017). Previous findings demonstrated the role of miR-106a in numerous risk factors of ischemic stroke. For example, under oxidative stress, miR-106-5p was documented to be decreased, causing premature senescence by suppressing the G1/S-phase transition of the cell cycle through modulating the expression of E2F1 (Tai et al., 2020). Similarly, increased levels of reactive oxygen species (ROS) resulted to the decreased expression of miR-106a (Wang et al., 2010). On the other hand, elevated levels of miR-106a prevented oxidative stress injury and inflammation in hepatic mouse with gestational hypertension (Wang Z. et al., 2019), resulting to repression of the expressions of HIF1-α and VEGF in diabetic retina (Ling et al., 2013). In addition, miR-106a has been associated with macrophage activation, suggesting its involvement in inflammation (Zhu et al., 2013).

In the present work, it was interesting to find a negative correlation between LINC00657 and miR-106a in diabetic patients who had stroke. A number of recent studies have hypothesized that lncRNAs could affect the progression of diseases through regulating miRNAs. It was reported that LINC00657 could influence tumorigenesis in hepatocellular carcinoma by regulating miR-106a (Hu et al., 2017).

Notably, an ROC curve was constructed in our study. The results implied that serum TUG1, LINC00657, miR-9, and miR-106a could discriminate diabetic patients without stroke from healthy subjects. More importantly, the aforementioned non-coding RNAs may be used to differentiate diabetic patients with stroke from those without stroke.

Some limitations of this work should be addressed. First is the relatively small sample size. Therefore, further works with larger sample sizes in various populations are needed. Moreover, further experiments are necessary to explain the detailed mechanisms of the role of these non-coding RNAs in diabetic patients with and without stroke.

CONCLUSION

The current study, for the first time, revealed that serum TUG1, LINC00657, miR-9, and miR-106a may serve as novel potential indicators of stroke associated with diabetes and correlated significantly with NIHSS. Furthermore, they might be used as new targets of treatment for diabetic patients with stroke.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available due to patient confidentiality. Requests to access the datasets should be directed to dr.omayma@yahoo.com.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the Faculty of Medicine, Fayoum University. The patients/participants provided written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MM, AE, and TA performed the patient examination and treatments. OS, RM, OA, NKA, and NH performed the biochemical assays. AK and NAA interpreted the data. OA and RM were major contributors to the writing of the manuscript. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank all the medical and paramedical staff who helped in the achievement of this work.

REFERENCES

Bao, M.-H., Szeto, V., Yang, B. B., Zhu, S.-z., Sun, H.-S., and Feng, Z.-P. (2018a). Long Non-coding RNAs in Ischemic Stroke. *Cell Death Dis* 9, 281. doi:10.1038/s41419-018-0282-x

Bao, M. H., Li, G. Y., Huang, X. S., Tang, L., Dong, L. P., and Li, J. M. (2018b). Long Noncoding RNA LINC00657 Acting as a miR-590-3p Sponge to Facilitate Low Concentration Oxidized Low-Density Lipoprotein-Induced Angiogenesis. *Mol. Pharmacol.* 93, 368–375. doi:10.1124/mol.117.110650

Chen, C., Cheng, G., Yang, X., Li, C., Shi, R., and Zhao, N. (2016). Tanshinol Suppresses Endothelial Cells Apoptosis in Mice with Atherosclerosis via lncRNA TUG1 Up-Regulating the Expression of miR-26a. *Am. J. Transl. Res.* 8, 2981–2991.

Chen, S., Wang, M., Yang, H., Mao, L., He, Q., Jin, H., et al. (2017). LncRNA TUG1 Sponges microRNA-9 to Promote Neurons Apoptosis by Up-Regulated Bcl2l11

under Ischemia. *Biochem. Biophys. Res. Commun.* 485, 167–173. doi:10.1016/j.bbrc.2017.02.043

Duan, W., Du, L., Jiang, X., Wang, R., Yan, S., Xie, Y., et al. (2016). Identification of a Serum Circulating lncRNA Panel for the Diagnosis and Recurrence Prediction of Bladder Cancer. *Oncotarget* 7, 78850–78858. doi:10.18632/oncotarget.12880

Ferreiro, J. L., Gómez-Hospital, J. A., and Angiolillo, D. J. (2010). Review Article: Platelet Abnormalities in Diabetes Mellitus. *Diabetes Vasc. Dis. Res.* 7, 251–259. doi:10.1177/1479164110383994

Hu, B., Cai, H., Zheng, R., Yang, S., ZhouZand Tu, J. (2017). Long Non-coding RNA 657 Suppresses Hepatocellular Carcinoma Cell Growth by Acting as a Molecular Sponge of miR-106a-5p to Regulate PTEN Expression. *Int. J. Biochem. Cel Biol* 92, 34–42. doi:10.1016/j.biocel.2017.09.008

Hu, D., Wang, Y., Zhang, H., and Kong, D. (2018). Identification of miR-9 as a Negative Factor of Insulin Secretion from Beta Cells. *J. Physiol. Biochem.* 74, 291–299. doi:10.1007/s13105-018-0615-3

Iyengar, B. R., Choudhary, A., Sarangdhar, M. A., Venkatesh, K. V., Gadgil, C. J., and Pillai, B. (2014). Non-coding RNA Interact to Regulate Neuronal Development and Function. *Front. Cel. Neurosci.* 8, 47. doi:10.3389/fncel.2014.00047

Ji, Q., Ji, Y., Peng, J., Zhou, X., Chen, X., Zhao, H., et al. (2016). Increased Brain-specific MiR-9 and MiR-124 in the Serum Exosomes of Acute Ischemic Stroke Patients. *PLoS ONE* 11, e0163645. doi:10.1371/journal.pone.0163645

Jiménez-Lucena, R., Rangel-Zúñiga, O. A., Alcalá-Díaz, J. F., López-Moreno, J., Roncero-Ramos, I., Molina-Abril, H., et al. (2018). Circulating miRNAs as Predictive Biomarkers of Type 2 Diabetes Mellitus Development in Coronary Heart Disease Patients from the CORDIOPREV Study. *Mol. Ther. - Nucleic Acids* 12, 146–157. doi:10.1016/j.omtn.2018.05.002

Khoshnam, S. E., Winlow, W., Farbood, Y., Moghaddamand, H. F., and Farzaneh, M. (2017). Emerging Roles of microRNAs in Ischemic Stroke: As Possible Therapeutic Agents. *J. Stroke* 19, 166–187. doi:10.5853/jos.2016.01368

Kitagawa, M., Kitagawa, K., Kotake, Y., Niida, H., and Ohhata, T. (2013). Cell Cycle Regulation by Long Non-coding RNAs. *Cell. Mol. Life Sci.* 70, 4785–4794. doi:10.1007/s00018-013-1423-0

Kong, L., Zhu, J., Han, W., Jiang, X., Xu, M., Zhao, Y., et al. (2011). Significance of Serum microRNAs in Pre-diabetes and Newly Diagnosed Type 2 Diabetes: a Clinical Study. *Acta Diabetol.* 48, 61–69. doi:10.1007/s00592-010-0226-0

Koutsis, G., Siasos, G., and Spengos, K. (2013). The Emerging Role of microRNA in Stroke. *Curr. Top. Med. Chem.* 13, 1573–1588. doi:10.2174/15680266113139990106

Li, F. P., Lin, D. Q., and Gao, L. Y. (2018). LncRNA TUG1 Promotes Proliferation of Vascular Smooth Muscle Cell and Atherosclerosis through Regulating miRNA-21/PTEN axis. *Eur. Rev. Med. Pharmacol. Sci.* 22, 7439–7447. doi:10.26355/eurrev_201811_16284

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and Protein-RNA Interaction Networks from Large-Scale CLIP-Seq Data. *Nucleic Acids Res.* 42 (Database issue), D92–D97. doi:10.1093/nar/gkt1248

Li, Y., Huang, D., Zheng, L., Cao, H., Gao, Y., Yang, Y., et al. (2019). Retracted Article: Long Non-coding RNA TUG1 Alleviates High Glucose Induced Podocyte Inflammation, Fibrosis and Apoptosis in Diabetic Nephropathy via Targeting the miR-27a-3p/E2F3 axis. *RSC Adv.* 9, 37620–37629. doi:10.1039/c9ra06136c

Ling, S., Birnbaum, Y., Nanhwan, M. K., Thomas, B., Bajaj, M., and Ye, Y. (2013). MicroRNA-dependent Cross-Talk between VEGF and HIF1α in the Diabetic Retina. *Cell Signal.* 25, 2840–2847. doi:10.1016/j.cellsig.2013.08.039

Liu, H., Li, J., Koirala, P., Ding, X., Chen, B., Wang, Y., et al. (2016). Long Non-coding RNAs as Prognostic Markers in Human Breast Cancer. *Oncotarget* 7, 20584–20596. doi:10.18632/oncotarget.7828

Liu, S., Zou, L., Xie, J., Xie, W., Wen, S., Xie, Q., et al. (2016). LncRNA NONRATT021972 siRNA Regulates Neuropathic Pain Behaviors in Type 2 Diabetic Rats through the P2X7 Receptor in Dorsal Root Ganglia. *Mol. Brain* 9, 44. doi:10.1186/s13041-016-0226-2

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods* 25, 402–408. doi:10.1006/meth.2001.1262

Long, J., Badal, S. S., Ye, Z., Wang, Y., Ayanga, B. A., Galvan, D. L., et al. (2016). Long Noncoding RNA Tug1 Regulates Mitochondrial Bioenergetics in Diabetic Nephropathy. *J. Clin. Invest.* 126, 4205–4218. doi:10.1172/JCI87927

Mastropasqua, R., Toto, L., Cipollone, F., Santovito, D., Carpineto, P., and Mastropasqua, L. (2014). Role of microRNAs in the Modulation of Diabetic Retinopathy. *Prog. Retin. Eye Res.* 43, 92–107. doi:10.1016/j.preteyeres.2014.07.003

Meng, S., Cao, J. T., Zhang, B., Zhou, Q., Shen, C. X., and Wang, C. Q. (2012). Downregulation of microRNA-126 in Endothelial Progenitor Cells from Diabetes Patients, Impairs Their Functional Properties, via Target Gene Spred-1. *J. Mol. Cel Cardiol* 53, 64–72. doi:10.1016/j.yjmcc.2012.04.003

Mercer, T. R., and Mattick, J. S. (2013). Structure and Function of Long Noncoding RNAs in Epigenetic Regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi:10.1038/nsmb.2480

Michalik, K. M., You, X., Manavski, Y., Doddaballapur, A., Zörnig, M., Braun, T., et al. (2014). Long Noncoding RNA MALAT1 Regulates Endothelial Cell Function and Vessel Growth. *Circ. Res.* 114, 1389–1397. doi:10.1161/circresaha.114.303265

Nakagami, H., Kaneda, Y., Ogihara, T., and Morishita, R. (2005). Endothelial Dysfunction in Hyperglycemia as a Trigger of Atherosclerosis. *Cdr* 1, 59–63. doi:10.2174/1573399052952550

Pinsker, J. E., Shank, T., Dassau, E., and Kerr, D. (2015). Comment on American Diabetes Association. Approaches to glycemic treatment. Sec. 7. In Standards of Medical Care in Diabetes-2015. Diabetes Care 2015; 38(Suppl. 1):S41-S48. *Diabetes Care* 38 (10), e174. doi:10.2337/dc15-0839

Plaisance, V., Abderrahmani, A., Perret-Menoud, V., Jacquemin, P., LemaigreandRegazzi, F. R., and Regazzi, R. (2006). MicroRNA-9 Controls the Expression of Granuphilin/Slp4 and the Secretory Response of Insulin-Producing Cells. *J. Biol. Chem.* 281, 26932–26942. doi:10.1074/jbc.M601225200

Ramachandran, D., Roy, U., Garg, S., Ghosh, S., PathakandKolthur-Seetharam, S. U., and Kolthur-Seetharam, U. (2011). Sirt1 and Mir-9 Expression Is Regulated during Glucose-Stimulated Insulin Secretion in Pancreatic β-islets. *FEBSJ* 278, 1167–1174. doi:10.1111/j.1742-4658.2011.08042.x

Shaker, O. G., Abdelaleem, O. O., Mahmoud, R. H., Abdelghaffar, N. K., Ahmed, T. I., Said, O. M., et al. (2019). Diagnostic and Prognostic Role of Serum miR-20b, miR-17-3p, HOTAIR, and MALAT1 in Diabetic Retinopathy. *IUBMB Life* 71, 310–320. doi:10.1002/iub.1970

Tai, L., Huang, C. J., Choo, K. B., Cheong, S. K., and Kamarul, T. (2020). Oxidative Stress Down-Regulates MiR-20b-5p, MiR-106a-5p and E2F1 Expression to Suppress the G1/S Transition of the Cell Cycle in Multipotent Stromal Cells. *Int. J. Med. Sci.* 17, 457–470. doi:10.7150/ijms.38832

Wan, W., Wan, W., Long, Y., Li, Q., Jin, X., Wan, G., et al. (2020). RETRACTED ATICLE: Physcion 8-O-β-Glucopyranoside Exerts Protective Roles in High Glucose-Induced Diabetic Retinopathy via Regulating lncRNA NORAD/miR-125/STAT3 Signalling. *Artif. Cell Nanomedicine, Biotechnol.* 48, 463–472. doi:10.1080/21691401.2019.1709861

Wang, W.-Y., Wang, Y.-F., Ma, P., Xu, T.-P., and Shu, Y.-Q. (2017). Taurine-upregulated Gene 1: A Vital Long Non-coding RNA Associated with Cancer in Humans. *Mol. Med. Rep.* 16, 6467–6471. doi:10.3892/mmr.2017.7472

Wang, F., Gao, X., Zhang, R., Zhao, P., Sun, Y., and Li, C. (2019). LncRNA TUG1 Ameliorates Diabetic Nephropathy by Inhibiting miR-21 to Promote TIMP3-Expression. *Int. J. Clin. Exp. Pathol.* 12, 717–729.

Wang, Z., Bao, X., Song, L., Tian, Y., and Sun, P. (2019). Role of miR-106-mediated Mitogen-activated Protein Kinase Signaling Pathway in Oxidative Stress Injury and Inflammatory Infiltration in the Liver of the Mouse with Gestational Hypertension. *J. Cel Biochem* 122, 958–968. doi:10.1002/jcb.29552

Wang, Z., Liu, Y., Han, N., Chen, X., Yu, W., Zhang, W., et al. (2010). Profiles of Oxidative Stress-Related microRNA and mRNA Expression in Auditory Cells. *Brain Res.* 1346, 14–25. doi:10.1016/j.brainres.2010.05.059

Wei, N., Xiao, L., Xue, R., Zhang, D., Zhou, J., au, H., et al. (2016). MicroRNA-9 Mediates the Cell Apoptosis by Targeting Bcl2l11 in Ischemic Stroke. *Mol. Neurobiol.* 53, 6809–6817. doi:10.1007/s12035-015-9605-4

Wu, Y., Xu, D., Zhu, X., Yang, G., and Ren, M. (2017). MiR-106a Associated with Diabetic Peripheral Neuropathy through the Regulation of 12/15-LOX-Meidiated Oxidative/Nitrative Stress. *Curr. Neurovasc Res.* 14, 117–124. doi:10.2174/1567202614666170404115912

Xue, Y., Li, M., Liu, D., Zhu, Q., and Chen, H. (2018). Expression of miR-9 in the Serum of Patients with Acute Ischemic Stroke and its Effect on Neuronal Damage. *Int. J. Clin. Exp. Pathol.* 11, 5885–5892.

Yan, H. Y., Bu, S. Z., Zhou, W. B., and Mai, Y. F. (2018). TUG1 Promotes Diabetic Atherosclerosis by Regulating Proliferation of Endothelial Cells via Wnt

Pathway. *Eur. Rev. Med. Pharmacol. Sci.* 22, 6922–6929. doi:10.26355/eurrev_201810_16162

Yang, D., Yu, J., Liu, H.-B., Yan, X.-Q., Hu, J., Yu, Y., et al. (2019). The Long Non-coding RNA TUG1-miR-9a-5p axis Contributes to Ischemic Injuries by Promoting Cardiomyocyte Apoptosis via Targeting KLF5. *Cel Death Dis* 10, 908. doi:10.1038/s41419-019-2138-4

Zang, X. J., Li, L., Du, X., Yang, B., and Mei, C. L. (2019). LncRNA TUG1 Inhibits the Proliferation and Fibrosis of Mesangial Cells in Diabetic Nephropathy via Inhibiting the PI3K/AKT Pathway. *Eur. Rev. Med. Pharmacol. Sci.* 23, 7519–7525. doi:10.26355/eurrev_201909_18867

Zhang, L., Cheng, H., Yue, Y., Li, S., Zhang, D., and He, R. (2018). TUG1 Knockdown Ameliorates Atherosclerosis via Up-Regulating the Expression of miR-133a Target Gene FGF1. *Cardiovasc. Pathol.* 33, 6–15. doi:10.1016/j.carpath.2017.11.004

Zhu, D., Pan, C., Li, L., Bian, Z., Lv, Z., Shi, L., et al. (2013). MicroRNA-17/20a/106a Modulate Macrophage Inflammatory Responses through Targeting Signal-Regulatory Protein α. *J. Allergy Clin. Immunol.* 132, 426–436. doi:10.1016/j.jaci.2013.02.005

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2022
doi: 10.3389/fgene.2021.793629

Feature Selection of OMIC Data by Ensemble Swarm Intelligence Based Approaches

Zhaomin Yao[1,2], *Gancheng Zhu*[3], *Jingwei Too*[4], *Meiyu Duan*[3] *and Zhiguo Wang*[1,2]*

[1]Department of Nuclear Medicine, General Hospital of Northern Theater Command, Shenyang, China, [2]College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China, [3]Key Laboratory of Symbolic Computation, College of Computer Science and Technology, Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China, [4]Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, Melaka, Malaysia

OMIC datasets have high dimensions, and the connection among OMIC features is very complicated. It is difficult to establish linkages among these features and certain biological traits of significance. The proposed ensemble swarm intelligence-based approaches can identify key biomarkers and reduce feature dimension efficiently. It is an end-to-end method that only relies on the rules of the algorithm itself, without presets such as the number of filtering features. Additionally, this method achieves good classification accuracy without excessive consumption of computing resources.

Keywords: swarm intelligence (SI), feature selection (FS), transcriptome data, methylation data, intersection and union combination

OPEN ACCESS

Edited by:
Lin Hua,
Capital Medical University, China

Reviewed by:
Yushan Qiu,
Shenzhen University, China
Collins Leke,
University of Johannesburg, South Africa
Nebojsa Bacanin,
Singidunum University, Serbia

***Correspondence:**
Zhiguo Wang
wangzhiguo5778@163.com

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

Received: *12 October 2021*
Accepted: *22 December 2021*
Published: *08 March 2022*

Citation:
Yao Z, Zhu G, Too J, Duan M and Wang Z (2022) Feature Selection of OMIC Data by Ensemble Swarm Intelligence Based Approaches. Front. Genet. 12:793629. doi: 10.3389/fgene.2021.793629

1 INTRODUCTION

The OMIC data includes genomes, transcriptomes, metabolomes, and proteomes. (Karczewski and Snyder 2018). Its quantity and quality have been improved significantly during the rapid development and continuous innovation of high-throughput sequencing and mass spectrum technologies (Margolis et al., 2014). Generally, biomedical data has the characteristics of "large p and small n," that is, the species of features is far larger than the species of samples (Liao and Chin 2007). Thus, it is necessary for biomedical dataset dimension reduction to protect against potential dimension disaster.

Feature selection has been proven with excellent performance in data preprocessing, especially for high dimensional data (Dash and Liu 1997; Bolón-Canedo, Sánchez-Maroño, and Alonso-Betanzos 2015). Its goals consist of cleaning out understandable and analyzable data, constructing simple and efficient models, and improving the efficiency of data mining (Li et al., 2017). It has achieved prominent results in the bioinformation field (Fu et al., 2018; Qiu, Ching, and Zou 2021). Swarm intelligence (SI) is the decentralized self-organizing collective behavior at the collective level (Hu et al., 2021b). It usually consists of a group of simple agents that interact with each other locally and with their environment. The agents follow very simple rules, and there is no centralized control structure to specify the behavior of a single agent. However, the interaction among these agents will lead to the emergence of "intelligent" global behavior (Hu et al., 2021a). Therefore, the whole problem-solving process will not be affected by the failure of one or several agents, so this method has good robustness and potential global search ability. Additionally, SI can transmit and coordinate information through indirect communication. With the increase in the number of individuals, the increase in communication overhead is small. Thus, it also has good scalability. Because of these advantages, SI is widely used in feature selection; its combination with machine learning has especially proven to be able to obtain outstanding results. Through the research and development of

the genetic algorithm (Malakar et al., 2019) and the firefly algorithm (Bacanin et al., 2021), the features extracted from each handwritten word image have been significantly optimized so that the performance of the handwritten word recognition technique has been increased visibly.

Various computational feature selection models have been proposed to reduce the dimension of OMIC datasets (Ge et al., 2016; Liu et al., 2017; Yuanyuan, Lan, and Fengfeng 2021). However, these algorithms need to design the number of features in advance as an intervention. Meanwhile, the heuristic rules applied are almost mathematical principles. Thus, this study was intended to investigate the performance of the features screened based on biological or natural rules, instead of traditional mathematical principles, and manually specify the number.

This article is organized as follows: details of the datasets and overview of the methods are described in **Section 2**. Experimental results and a corresponding analysis of these results are presented in **Section 3**. Finally, a brief conclusion is drawn in **Section 4**.

2 MATERIALS AND METHODS

As shown in **Figure 1**, this study involved six major stages: Dataset curation, data preprocessing, feature selection, model training and validation, feature intersection and union combination, and prediction. First, a large number of OMIC datasets are collected, including transcriptome datasets (Dataset 1) and methylation datasets (Dataset 2). Then, all the features with missing values in the collected datasets will be deleted. Next, all the transcriptome datasets will have features extracted by twelve advanced swarm intelligent algorithms, and then these features will be input into five different representative classifiers and finally classification performance will be obtained. According to these results, the best classifier and the top three algorithms that use this classifier to get the best results will be selected to apply to methylation datasets. Later, these subsets will generate different combinations through union and intersection. Finally, the classification performance of these combinations will be evaluated by the best classifiers. The details of each process are described in the following sections.

2.1 Summary of Datasets

This study concentrated on binary classification and analyzed the relevant publicly available OMIC databases. As shown in **Supplementary Table S1**, these data include 17 transcriptome datasets and 10 methylation datasets. Methylation is an important modification of proteins and nucleic acids; it reveals the influence of genetic and environmental factors on the occurrence and development of complex diseases (Barros and Offenbacher 2009). Compared with transcriptome data, methylation data usually have more feature dimension and are more challenging in classification.

First, all transcriptome datasets (Dataset 1) were used to test the performance of the algorithm. As shown in **Supplementary Table S1**, they were DLBCL (Shipp et al., 2002), Pros (Aalinkeel et al., 2004), Colon (Alon et al., 1999), Leuk (Golub et al., 1999), Mye (Tian et al., 2003), All (All1/All2/All3/All4) (Chiaretti et al., 2004), CNS (Pomeroy et al., 2002), Lym (Alizadeh et al., 2000), Adeno (Notterman et al., 2001), Gas (Wu et al., 2013), Gas1/Gas2 (Wang et al., 2013) , T1D (Levy et al., 2012), and Stroke (Krug et al., 2012). These datasets were obtained and preprocessed as similar in Mctwo (Ge et al., 2016).

Additionally, ten methylation datasets (Dataset 2) were used to demonstrate the binary classification performances, as shown in **Supplementary Table S1**. The dataset GSE74845 profiled 110 Fimbria and 106 proximal tubal DNA samples of fallopian tube

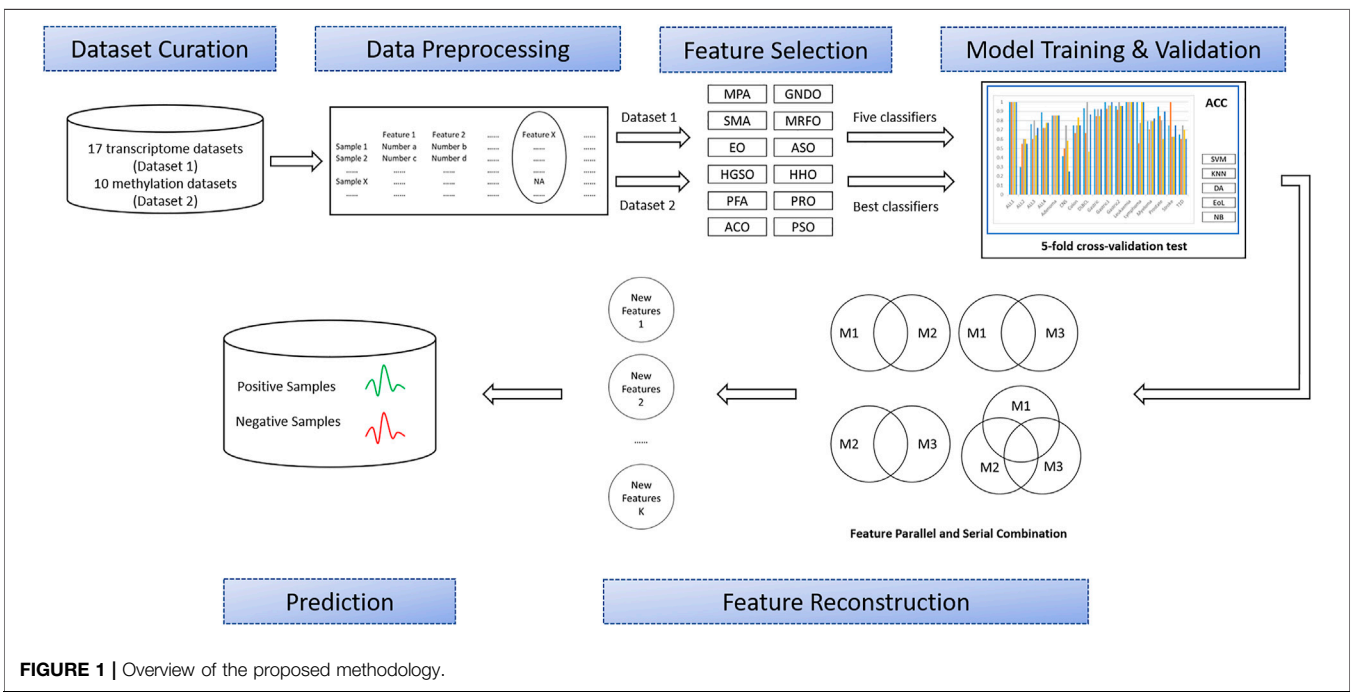


FIGURE 1 | Overview of the proposed methodology.

fimbriae in BRCA mutation carriers (Bartlett et al., 2016). The dataset GSE80970 provided the methylomes of 148 Alzheimer's disease samples and 138 controls (Smith et al., 2018). The dataset GSE103186 illustrated 130 gastric light or mild intestinal metaplasia and 61 gastric normal samples (Huang et al., 2017). The dataset GSE139032 investigated 77 lung adenocarcinomas and 77 matched non-malignant lung samples (Enfield et al., 2019). The dataset GSE139404 compared 40 low-grade adenoma and high-grade adenoma in colorectal and 20 normal tissues (Fan et al., 2020). The dataset GSE144910 collected a total of 88 genomic DNA samples taken from the postmortem superior temporal gyrus of the human brain with 44 schizophrenia and paired non-psychiatric controls (Mckinney et al., 2020). The dataset GSE164269 generated 33 discovery and 46 independent validation cohorts of malignant pleural mesothelioma samples (Bertero et al., 2021). The dataset GSE166787 contrasted DNA methylation data throughout human muscle cell differentiation in 28 individuals with type 2 diabetes and 28 controls (Davegårdh et al., 2021). The dataset GSE173330 supplied DNA methylation data from several tissues in toothed whales ($N = 254$) and dolphin ($N = 291$) (Robeck et al., 2021). The last dataset GSE174613 analyzed samples of non-malignancy obtained from prostatectomy specimens ($n = 12$) and of bone metastasis tissue samples obtained from separate prostate cancer patients ($n = 70$) (Ylitalo et al., 2021).

2.2 Data Preprocessing

Due to various experimental reasons, gene expression data universally suffer from the missing value problem. The features with missing values can adversely affect the classifiers (Varsha et al., 2016). Considering the number of features with missing values in the datasets accounts for less than 0.1% of the total number of features, direct removal also has little impact on the overall datasets. Thus, these features affected by missing values are removed directly. For example, for a feature X, the value of X is missing in only one sample, but there is a definite value in all other samples. The X must be removed from all samples.

2.3 Summary of Swarm Intelligence Methods in Feature Selection

Twelve swarm intelligence methods are used in the study, including ten state-of-the-art methods from the last 2 years and two classic methods. The methods are briefly described below.

2.3.1 Marine Predators Algorithm

Marine predator algorithm (MPA) is a natural heuristic optimization algorithm. It follows the rule of natural dominance in the optimal foraging strategy and encounters the rate strategy between predator and prey in the marine ecosystem. This algorithm is inspired by the predator–prey strategy in nature and considers that the top predator has the greatest search ability, that is, the decision of a top predator is a solution of the problem (Faramarzi et al., 2020a).

2.3.2 Generalized Normal Distribution Optimization

Generalized normal distribution optimization (GNDO) is a novel metaheuristic algorithm inspired by normal distribution theory. It can solve optimization problems by natural phenomenon distribution and fitting minimum standard variance of the positions of all individuals. Generally speaking, GNDO consists of two main strategies: local exploitation and global exploration. The former focuses on building the generalized distribution model while the latter explores the search region based on three randomly selected individuals (Zhang et al., 2020).

2.3.3 Slime Mould Algorithm

Slime mould algorithm (SMA) is based on the diffusion and foraging behavior of slime mould in nature. It calculates the optimal path by simulating the relationship between morphological changes and contraction patterns of slime mould during foraging. SMA performs the search relying on three stages: Find approach, wrap food, and oscillation (Li et al., 2020).

2.3.4 Manta Ray Foraging Optimization

Manta ray foraging optimization (MRFO) mathematically models and mimics three unique foraging strategies of manta rays, including chain foraging, cyclone foraging, and somersault foraging, for solving global optimization problems. In chain foraging, the manta rays update their solutions by following the best solution and the solution in front of it. For cyclone foraging, the manta rays move toward the global optima along a spiral path. Last, in somersault foraging, the manta rays tend to update their position around the best solution in the population (Zhao et al., 2020).

2.3.5 Equilibrium Optimizer

Equilibrium optimizer (EO) is inspired by a physical phenomenon of controlling volume mass balance. It simulates the physical process of mass entering, leaving, and generating in the control volume to finally reach the equilibrium state as optimal results. In EO, there is an equilibrium pool that used to store the current four best-so-far solutions. Iteratively, these stored solutions will be applied to enhance the quality of solutions in the population. Additionally, EO integrates the particle memory saving to benefit the exploitation capability (Faramarzi et al., 2020b).

2.3.6 Atom Search Optimization

Atom search optimization (ASO) is a novel algorithm based on a basic molecular dynamics model. In a molecular system, there are interaction forces between neighboring atoms, and the globally optimal atoms constrain other atoms. Gravitation makes atoms explore the whole search space extensively, and repulsion makes them develop the potential region effectively. It simulates this phenomenon to find the global optimal solution (Zhao et al., 2019).

2.3.7 Henry Gas Solubility Optimization

Henry gas solubility optimization (HGSO) is a novel metaheuristic algorithm; it imitates the huddling behavior of gas described in Henry's law to balance the exploitation ability and the exploration ability of the algorithm for searching the global optimum and avoid trapping into local optima (Hashim et al., 2019).

2.3.8 Harris Hawks Optimization

Harris hawks optimization (HHO) is a novel population-based, natural heuristic optimization. Its main inspiration comes from Harris's eagle's cooperative behavior and pursuit in nature. It is unique because it has a unique cooperative foraging activity with other family members in the group. Because of this, it is very suitable to simulate the unique predatory behavior of Harris's hawk as a swarm intelligence optimization process (Heidari et al., 2019).

2.3.9 Path Finder Algorithm

Path finder algorithm (PFA) is inspired by the hunting behavior of group animals. The algorithm realizes the optimization process through the communication between pathfinder and follower from the population in the process of the population searching for food. Naturally, PFA stores the best-so-far solution (pathfinder), in which the pathfinder is used to enhance the exploitation and exploration capability (Yapici and Cetinkaya 2019).

2.3.10 Poor and Rich Optimization

Poor and rich optimization (PRO) is developed based on the real social phenomenon, that is, the attempt of the rich and the poor to improve their economic conditions. This social behavior can be regarded as a solution for complex optimization problems. In PRO, a mutation operator is designed to improve the compound population. Even though PRO is a promising algorithm, it suffers from the high computational complexity (Moosavi and Bardsiri 2019).

2.3.11 Ant Colony Optimization

Ant colony algorithm is inspired by the foraging behavior of ants in nature. In the process of ant foraging, an ant colony can always find an optimal path between the ant nest and food source. This is because the ants in the ant colony can transmit information through some information mechanism. After further research, it is found that ants will release a substance called "pheromone" on their path. Ants in the ant colony have the ability to perceive the "pheromone." They will walk along the path with high concentration of "pheromone," and each passing ant will leave "pheromone" on the road, which forms a mechanism similar to positive feedback; in this way, after a period of time, the whole ant colony will reach the food source along the shortest path (Dorigo et al., 2006).

2.3.12 Particle Swarm Optimization

Particle swarm optimization is inspired by the study of bird predation behavior. Specifically, birds find the optimal destination through collective information sharing. In PSO, the potential solution of each optimization problem is a bird in the search space, which is called a particle. All particles have a fitness value determined by the optimized function, and each particle also has a speed to determine their flying direction and distance. Then the particles follow the current optimal particle to search in the solution space (Kennedy and Eberhart 1995).

2.4 Model Training and Validation

2.4.1 Random 5-Fold Cross-Validation Strategy

K-fold cross-validation is one of the most commonly used evaluation strategies. This experimental procedure is performed by the 5-fold cross-validation, that is, the baseline dataset is randomly divided into five equal parts (the number and distribution of samples are the same) and the test processes are repeated five times; for each cross-validation test, one subset is used for testing while the remains are used for training the model. The final performance is represented by the average of five experimental results.

2.4.2 Leave-One-Out Cross-Validation Strategy

Leave one method cross-validation is to treat each data sample as an independent dataset, use one sample each time as the test set, and use all the remaining samples as the training set. The result obtained using this method is closest to the expected value of the whole test set, but the computing cost is excessively expensive.

2.4.3 Performance Evaluation of Various Classifiers

Higher classification accuracy and fewer features are the objectives of generating models; however, it is difficult to achieve both at the same time. Here, the first consideration in this study is the classification accuracy. For achieving a more comprehensive and stable performance, five widely used classifiers are applied to the models, that is, support vector machine (SVM), K-Nearest Neighbor (KNN), discriminant analysis (DA), ensemble of learners (EoL), and naive Bayes (NB). This study evaluates a feature subset through the best classification performance of multiple classifiers. Generally, prediction accuracy is defined as follows:

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

where TP, FP, TN, and FN represent the value of true positives, false positives, true negatives, and false negatives, respectively.

2.5 Feature Intersection and Union Combination

Intersection and union combination approaches were employed to ensemble the selected features. As shown in **Figure 2**, two or three different feature selection results were combined into eight subsets for performance comparison.

3 RESULTS AND DISCUSSIONS

3.1 The Result on Transcriptome Datasets

This study used these transcriptome datasets for testing the performance of baseline swarm intelligence algorithms and classifiers. Enough iterations are used to satisfy the fitness value. Here, the random 5-fold cross-validation and leave-one-out cross-validation are used to evaluate the performance, respectively. The results are shown in **Supplementary Tables S2, S3**. Both of the tables show that KNN can make most datasets achieve the best classification effect in most algorithms. Additionally, in the other three algorithms, where KNN cannot achieve the best results, the gap between KNN and the best classifier in the number of datasets for best performance is small, only one to three datasets.

Through the information combination of two tables, when using KNN, the number of best results obtained by PFA and SMA

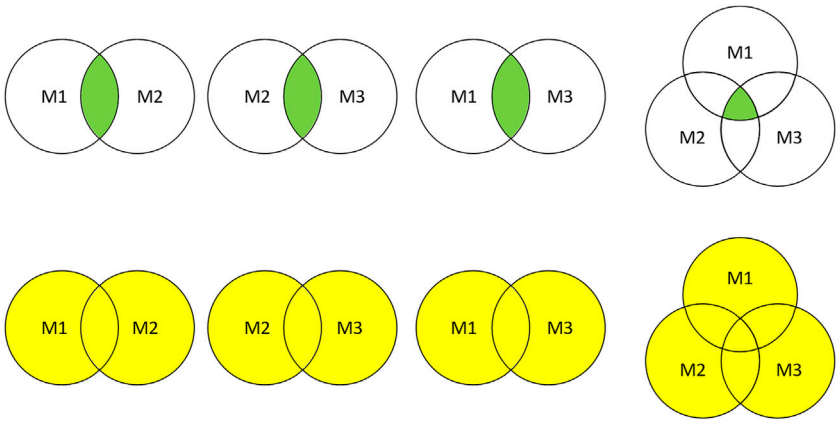


FIGURE 2 | Feature subsets combination. M1, M2, and M3 represent the feature subsets extracted by three different methods, respectively. The green part and yellow part represent the combination results obtained by intersection and union.

is 12 and 8, respectively, ranking first and second. ASO, GNDO, PSO, and HGSO all get 7 best results, and the number is equal. As shown in **Supplementary Table S4**, considering the average number of features used on each dataset, HGSO is chosen as the last algorithm to be applied to the next stage.

Because there is little performance difference between 5-fold cross-validation and leave-one-out cross-validation in these transcriptome datasets and the computing cost of leave-one-out cross-validation is relatively high, the subsequent evaluation is only based on the random 5-fold cross-validation.

3.2 Convergence of Top Three Swarm Intelligence Algorithms

In the FS phase, a fitness function is adopted to evaluate the quality of the initial and newly generated solutions. This study evaluates the solutions by considering the minimum classification error and minimum size of features (Emary et al., 2016a). Mathematically, the fitness function is defined as follows:

$$Fit = \beta ER + (1-\beta)\left(\frac{|SF|}{|AF|}\right)$$

where ER is the classification error rate computed by the k-nearest neighbor classifier (KNN, k-value = 5), $|SF|$ is the number of the selected features, $|AF|$ is the total number of features, and β is the weight factor between 0 and 1. This study adopts $\beta = 0.99$ since the classification performance is the most importance measurement (Emary, Zawbaa, and Hassanien 2016b; Mafarja et al., 2019). In the fitness evaluation stage, the dataset is partitioned into training and validation sets using the k-fold cross-validation method. Consequently, the dataset is divided into 5 folds, in which k-1 folds are used to build the training set while the rest is kept for accessing the selected features.

The T1D dataset is used as an example to show the convergence of the top three algorithms. As shown in **Figure 3**, PFA and HGSO converge in about 22 iterations, while SMA converges faster, and the convergence can be completed in about 10 iterations.

3.3 The Result of Top Three Swarm Intelligence Algorithms on Methylation Datasets

This section evaluated the performance of SMA, PFA, and HGSO on the methylation datasets, and the classifier is KNN.

Although methylome datasets may be a challenge for many feature selection algorithms, the swarm intelligence algorithm has achieved good results on many datasets. As shown in **Figure 4**, PFA achieves more than 90% accuracy on four datasets. Meanwhile, SMA obtains about 90% accuracy on the GSE139032 and GSE139404, where PFA does not get good results. In addition, the consumption of computing resources and time is also within an acceptable range; the average time consumption (CPU: i9-11900H) of SMA, PFA, and HGSO are 101.83, 415.21, and 312.31 s, respectively.

3.4 Other Evaluation Indexes of Top Three Swarm Intelligence Algorithms on Methylation Datasets

Besides accuracy, other evaluation indicators are also very important. They can reveal the characteristics of the algorithm in other aspects. Therefore, another four commonly used indicators for classification evaluation (precision, recall, F1-score, and AUC ROC) have also been tested, and the results are shown in the **Supplementary Table S5**. It can be seen from the results that there is little difference between precision and recall of most models. However, the precision of PFA reaches 100% but the corresponding recall just obtains about 12% on GSE164269. It may be caused by the insensitivity of the dataset to the algorithm, that is, the algorithm cannot filter the core features of the dataset. Thus, many positive samples are identified as negative samples.

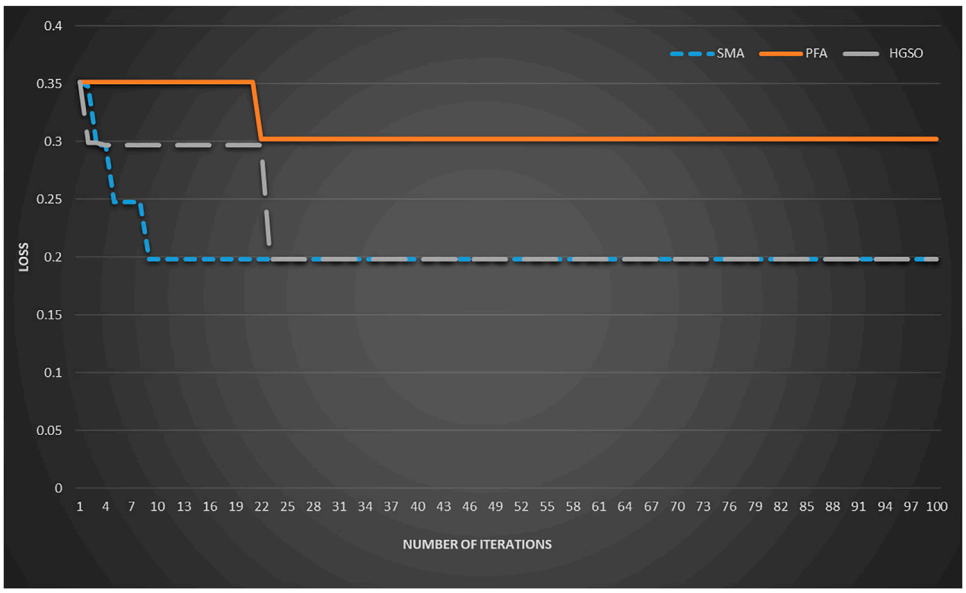


FIGURE 3 | The convergence speed of top three swarm intelligence algorithms on T1D.

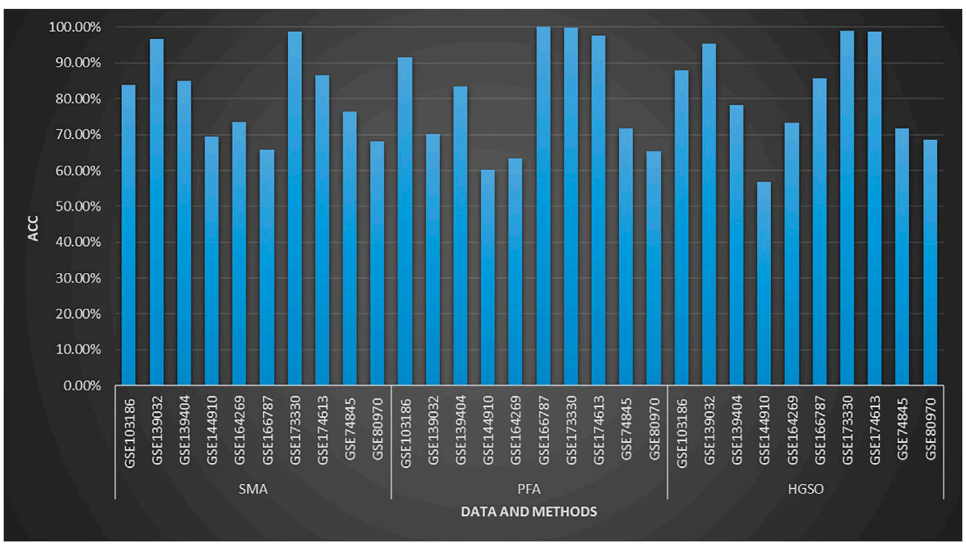


FIGURE 4 | Performance of three swarm intelligence algorithms on methylation datasets.

3.5 Statistical Tests of Obtained Results

Statistical tests on the results obtained using the three methods were performed. The statistics are described in **Table 1**. The result of Wilcoxon signed ranks test are shown in **Table 2**. Through the nonparametric test of paired samples, the *p*-values are greater than the significance level, indicating that there is no difference in the measurement accuracy of these 10 samples after three methods. Additionally, the Friedman test was also applied, and the chi-squared, df, and *p*-value are 0.2, 2, and 0.906, respectively. It also proved that there was no significant difference in accuracy.

3.6 The Result of Feature Intersection and Union Combination on Methylation Datasets

Generally, for a given dataset, the feature subsets for different feature selection are individually somewhat different due to the different theories. So, their different combinations will be more diverse. These subsets are evaluated in this section. What is more, there is no duplicate selection of the same features by different methods.

Figure 5 shows the classification performance obtained by intersection and union combination-based feature subset

TABLE 1 | Descriptive statistics of the results on methylation datasets.

Methods	Sample number	Average (%)	Standard deviation	Min (%)	Max (%)
SMA	10	80.44	11.62	65.91	98.72
PFA	10	80.30	15.98	60.13	100.00
HGSO	10	81.55	14.17	56.73	98.90

TABLE 2 | Wilcoxon signed ranks test.

Comparison	R^+	R^-	*p*-value
PFA versus SMA	4	6	0.721
HGSO versus SMA	5	5	0.959
HGSO versus PFA	5	5	0.878

ensemble methods. In some feature subset combinations, no classification accuracy is available because there is no repeat selection of the same features by the applied methods. As we can see, the performance of the union combination method with PFA is not obvious. The reason may be that PFA selects too many features, which is over 2000 times that of SMA and about 200 times that of HGSO. Additionally, the performance of union combination between SMA and HGSO is always better than just using HGSO but not always better than just using SMA. The reason may be that the number of features used by HGSO is ten times than that of SMA. Therefore, the characteristics of SMA can only be used for auxiliary adjustment. What is more, the performance of some intersection methods does not decrease so much. This may be because the features selected by all these algorithms are the core features of the datasets.

3.7 The Feature Selection Rates on Methylation Datasets

Table 3 shows the feature selection rates of single and different combination swarm intelligence methods on methylation datasets. Note that the feature selection rate is the percentage of the features that are extracted from the original features.

As we can see, SMA produces the lowest feature reduction rate in a single model, that is, the average is 0.0238%. This means that applying SMA as the embedded feature selection method may cause "over selection," with too many informative features filtered out. On the other hand, PFA not only allows selection of the most informative features but also avoids the risk of over selection. However, using the intersection combination with HGSO and PFA not only can reduce the number of features further but also not reduce the accuracy in many datasets. The results indicate that intersection combination method-based ensemble feature selection is likely to play a positive role in filtering out information redundancy among the feature selection methods that retain too much information after use.

In addition, using the combination among feature subsets with widely different feature numbers will not lead to excessive changes in classification performance, and most of the classification results will be the result of the feature subset with the highest number of features, because its feature distribution has not changed.

3.8 The Results of Multi-Classification on GSE103186

The internal metaplasia samples contained in GSE103186 can also be more finely divided into classic and mild. Therefore, GSE103186 is regarded as a three-category dataset for testing the multi-classification performance. The performance of SMA, PFA, and HGSO is 81.69, 80.63, and 83.78%, respectively. Although the proposed method mainly focuses on binary classification problems, the results show that it still has the potential to be used in multi-classification problems.

3.9 Biological Function Analysis of Selected Features on GSE144910

The dataset GSE144910 collected DNA samples from the superior temporal gyrus of the human brain for researching schizophrenia. The features detected by the union combination of SMA and HGSO as the classification biomarkers and these methylation features are related to 18 genes, which are C1orf168, CAMLG, SMOX, KCNIP4, MIR658, CENPA, ASRGL1, PISD, HNRNPL, EEF2K, GMDS, MPPED1, ANKRD54, PLEK2, ADA, RNF121, KRT6A, and EPHA2. In order to explore the biological functions of the selected genes, pathway analysis was conducted. **Figure 6** showed the mainly obtained four biological process pathways (GO: 0033627, 072657, 00488872, and 0044089). We found that schizophrenia may be related to the function of cell adhesion.

4 CONCLUSION

This study focuses on examining the binary classification performance of swarm intelligence algorithms on OMIC datasets. The experimental results suggest that swarm intelligence algorithms can achieve high accuracy on the collected OMIC datasets, significantly reduce feature dimensions, and identify key features. Meanwhile, this study finds some rules to improve ensemble feature subset performance through intersection and union combination methods. However, there are still some limitations in the proposed study. For example, the methodology framework has not been improved, and there is no methodological fusion of different swarm intelligence algorithms. Our future research will focus on combining machine learning and swarm intelligence approaches for reducing the feature dimension and improve the accuracy further in OMIC data and other biological data.

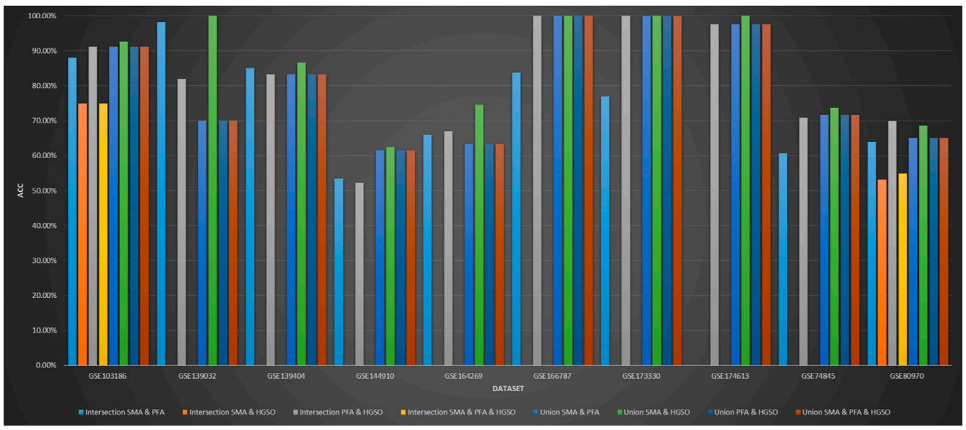


FIGURE 5 | Performance of feature intersection and union combination on methylation datasets.

TABLE 3 | Feature selection rates of all used feature subsets on methylation datasets.

Data	Solo			Intersection				Union			
	SMA (%)	PFA (%)	HGSO (%)	SMA and PFA	SMA and HGSO	PFA and HGSO (%)	SMA and PFA and HGSO	SMA and PFA (%)	SMA and HGSO (%)	PFA and HGSO (%)	SMA and PFA and HGSO (%)
GSE103186	0.0338	49.7048	0.6381	0.0154%	0.0002%	0.3184	0.0002%	49.7232	0.6716	50.0245	50.0428
GSE139032	0.0218	49.8948	0.0181	0.0145%	—	0.0109	—	49.9021	0.0399	49.9021	49.9093
GSE139404	0.0009	49.7509	0.0328	0.0004%	—	0.0149	—	49.7513	0.0336	49.7688	49.7692
GSE144910	0.0004	49.9814	0.0046	0.0001%	—	0.0018	—	49.9816	0.0049	49.9841	49.9844
GSE164269	0.0044	49.9655	0.7131	0.0022%	—	0.3630	—	49.9677	0.7175	50.3156	50.3178
GSE166787	0.0017	49.6841	0.0111	0.0009%	—	0.0059	—	49.6849	0.0129	49.6893	49.6902
GSE173330	0.0160	48.7964	0.3728	0.0107%	—	0.1651	—	48.8017	0.3888	49.0041	49.0094
GSE174613	0.0008	49.4005	0.0066	—	—	0.0049	—	49.4014	0.0074	49.4022	49.4030
GSE74845	0.0023	49.9412	0.1849	0.0011%	—	0.0933	—	49.9425	0.1873	50.0328	50.0341
GSE80970	0.1564	49.9624	0.6070	0.0871%	0.0007%	0.3080	0.0005%	50.0317	0.7626	50.2614	50.3304
Average	0.0238	49.7082	0.2589	0.0147%	0.0005%	0.1286	0.0004%	49.7188	0.2827	49.8385	49.8491

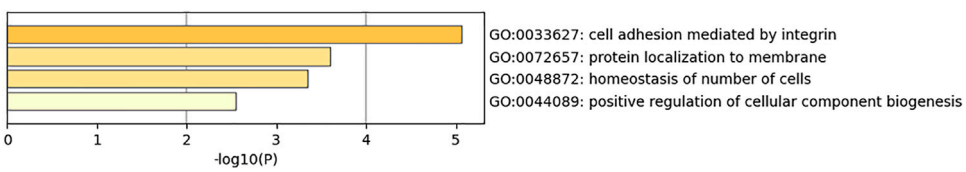


FIGURE 6 | Performance of feature intersection and union combination on methylation datasets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data can be found at: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi (Broad Institute Genome Data Analysis Center) and https://www.ncbi.nlm.nih.gov/geo/ [NCBI Gene Expression Omnibus (GEO) database].

AUTHOR CONTRIBUTIONS

ZW and ZY designed the project, GZ collected the datasets, ZY and JT carried out the coding of the computational analysis, ZY, JT, and MD drafted the manuscript, and ZW revised and polished the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Key Research and Development Program of Liaoning Province (2019JH2/10300010).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.793629/full#supplementary-material

REFERENCES

Aalinkeel, R., Nair, M. P. N., Sufrin, G., Mahajan, S. D., Chadha, K. C., Chawda, R. P., et al. (2004). Gene Expression of Angiogenic Factors Correlates with Metastatic Potential of Prostate Cancer Cells. *Cancer Res.* 64, 5311–5321. doi:10.1158/0008-5472.can-2506-2

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., et al. (2000). Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expressionprofiling. *Nature* 403, 503. doi:10.1038/35000501

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and normal colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci.* 96, 6745–6750. doi:10.1073/pnas.96.12.6745

Bacanin, N., Stoean, R., Zivkovic, M., Petrovic, A., Rashid, T. A., and Bezdan, T. (2021). Performance of a Novel Chaotic Firefly Algorithm with Enhanced Exploration for Tackling Global Optimization Problems: Application for Dropout Regularization. *Mathematics* 9, 2705. doi:10.3390/math9212705

Barros, S. P., and Offenbacher, S. (2009). Epigenetics: Connecting Environment and Genotype to Phenotype and Disease. *J. Dental Res.* 88, 400–408. doi:10.1177/0022034509335868

Bartlett, T. E., Chindera, K., Mcdermott, J., Breeze, C. E., Cooke, W. R., Jones, A., et al. (2016). Epigenetic Reprogramming of Fallopian Tube Fimbriae in BRCA Mutation Carriers Defines Early Ovarian Cancer Evolution. *Nat. Commun.* 7, 11620. doi:10.1038/ncomms11620

Bertero, L., Righi, L., Collemi, G., Koelsche, C., and Deimling, A. V. (2021). DNA Methylation Profiling Discriminates between Malignant Pleural Mesothelioma and Neoplastic or Reactive Histological Mimics. *J. Mol. Diagn.* 23, 834–846. doi:10.1016/j.jmoldx.2021.04.002

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2015). *Feature Selection for High-Dimensional Data*. Berlin/Heidelberg, Germany: Springer.

Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., et al. (2004). Gene Expression Profile of Adult T-Cell Acute Lymphocytic Leukemia Identifies Distinct Subsets of Patients with Different Response to Therapy and Survival. *Blood* 103, 2771–2778. doi:10.1182/blood-2003-09-3243

Dash, M., and Liu, H. (1997). Feature Selection for Classification. *Intell. Data Anal.* 1, 131–156. doi:10.1016/s1088-467x(97)00008-5

Davegårdh, C., Säll, J., Anna, B., Broholm, C., Volkov, P., Alexander, P., et al. (2021). VPS39-deficiency Observed in Type 2 Diabetes Impairs Muscle Stem Cell Differentiation via Altered Autophagy and Epigenetics. *Nat. Commun.* 12, 2431. doi:10.1038/s41467-021-22068-5

Dorigo, M., Birattari, M., and Stutzle, T. (2006). Ant colony Optimization. *IEEE Comput. intelligence Mag.* 1, 28–39. doi:10.1109/ci-m.2006.248054

Emary, E., Zawbaa, H. M., and Hassanien, A. E. (2016a). Binary Ant Lion Approaches for Feature Selection. *Neurocomputing* 213, 54–65. doi:10.1016/j.neucom.2016.03.101

Emary, E., Zawbaa, H. M., and Hassanien, A. E. (2016b). Binary Grey Wolf Optimization Approaches for Feature Selection. *Neurocomputing* 172, 371–381. doi:10.1016/j.neucom.2015.06.083

Enfield, K. S. S., Marshall, E. A., AndersonNg, C. K. W., and Wan, L. L. (2019). Epithelial Tumor Suppressor ELF3 Is a Lineage-specific Amplified Oncogene in Lung Adenocarcinoma. *Nat. Commun.* 10, 5438. doi:10.1038/s41467-019-13295-y

Fan, J., Li, J., Guo, S., Tao, C., and Zeng, C. (2020). Genome-wide DNA Methylation Profiles of Low- and High-Grade Adenoma Reveals Potential Biomarkers for Early Detection of Colorectal Carcinoma. *Clin. Epigenetics* 12, 56. doi:10.1186/s13148-020-00851-3

Faramarzi, A., Heidarinejad, M., Mirjalili, S., and Gandomi, A. H. (2020a). Marine Predators Algorithm: A Nature-Inspired Metaheuristic. *Expert Syst. Appl.* 152, 113377. doi:10.1016/j.eswa.2020.113377

Faramarzi, A., Heidarinejad, M., Stephens, B., and Mirjalili, S. (2020b). Equilibrium Optimizer: A Novel Optimization Algorithm. *Knowledge-Based Syst.* 191, 105190. doi:10.1016/j.knosys.2019.105190

Fu, G., Wang, J., Domeniconi, C., and Yu, G. (2018). Matrix Factorization-Based Data Fusion for the Prediction of lncRNA–Disease Associations. *Bioinformatics* 34, 1529–1537. doi:10.1093/bioinformatics/btx794

Ge, R., Zhou, M., Luo, Y., Meng, Q., and Zhou, F. (2016). McTwo: A Two-step Feature Selection Algorithm Based on Maximal Information Coefficient. *BMC Bioinformatics* 17, 142. doi:10.1186/s12859-016-0990-0

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., and Lander, E. S. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Monitoring. *Science* 286, 531–537. doi:10.1126/science.286.5439.531

Hashim, F. A., Houssein, E. H., Mai, S. M., Al-Atabany, W., and Mirjalili, S. (2019). Henry Gas Solubility Optimization: A Novel Physics-Based Algorithm. *Future Generation Comput. Syst.* 101, 646–667. doi:10.1016/j.future.2019.07.015

Heidari, A. A., Mirjalili, S., Faris, H., Aljarah, I., Mafarja, M., and Chen, H. (2019). Harris Hawks Optimization: Algorithm and Applications. *Future Generation Comput. Syst.* 97, 849–872. doi:10.1016/j.future.2019.02.028

Hu, J., Bhowmick, P., Jang, I., Arvin, F., and Lanzon, A. (2021a). A Decentralized Cluster Formation Containment Framework for Multirobot Systems. *IEEE Trans. Robotics* 37, 1. doi:10.1109/tro.2021.3071615

Hu, J., Turgut, A. E., Lennox, B., and Arvin, F. (2021b). Robust Formation Coordination of Robot Swarms with Nonlinear Dynamics and Unknown Disturbances: Design and Experiments. *IEEE Trans. Circuits Syst. Express Briefs* 69, 114–118. doi:10.1109/TCSII.2021.3074705

Huang, K. K., Ramnarayanan, K., Zhu, F., Srivastava, S., Xu, C., Tan, A. L. K., et al. (2017). Genomic and Epigenomic Profiling of High-Risk Intestinal Metaplasia Reveals Molecular Determinants of Progression to Gastric Cancer. *Cancer Cell* 33, 137–150. doi:10.1016/j.ccell.2017.11.018

Karczewski, K. J., and Snyder, M. P. (2018). Integrative Omics for Health and Disease. *Nat. Rev. Genet.* 19, 299–310. doi:10.1038/nrg.2018.4

Kennedy, J., and Eberhart, R. (1995). "Particle Swarm Optimization," in Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia, 27 Nov.-1 Dec. 1995.

Krug, T., Gabriel, J. P., Taipa, R., Fonseca, B. V., Domingues-Montanari, S., Fernandez-Cadenas, I., et al. (2012). TTC7B Emerges as a Novel Risk Factor for Ischemic Stroke through the Convergence of Several Genome-wide Approaches. *J. Cereb. Blood Flow Metab.* 32, 1061–1072. doi:10.1038/jcbfm.2012.24

Levy, H., Wang, X., Kaldunski, M., Shuang, J., and Hessner, M. J. (2012). Transcriptional Signatures as a Disease-specific and Predictive Inflammatory Biomarker for Type 1 Diabetes. *Genes Immun.* 13, 593–604. doi:10.1038/gene.2012.41

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., et al. (2017). Feature Selection: A Data Perspective. *ACM Comput. Surv. (Csur)* 50, 1–45. doi:10.1145/3136625

Li, S., Chen, H., Wang, M., Heidari, A. A., and Mirjalili, S. (2020). Slime Mould Algorithm: A New Method for Stochastic Optimization. *Future Generation Comput. Syst.* 111, 300–323. aliasgharheidari.com. doi:10.1016/j.future.2020.03.055

Liao, J. G., and Chin, K.-V. (2007). Logistic Regression for Disease Classification Using Microarray Data: Model Selection in a Large P and Small N Case. *Bioinformatics* 23, 1945–1951. doi:10.1093/bioinformatics/btm287

Liu, J., Cheng, X., Yang, W., Shu, Y., and Zhou., F. (2017). Multiple Similarly-Well Solutions Exist for Biomedical Feature Selection and Classification Problems. *Scientific Rep.* 7, 838. doi:10.1038/s41598-017-13184-8

Mafarja, M., Aljarah, I., Faris, H., Hammouri, A. I., Al-Zoubi, A.'M., and Mirjalili, S. (2019). Binary Grasshopper Optimisation Algorithm Approaches for Feature Selection Problems. *Expert Syst. Appl.* 117, 267–286. doi:10.1016/j.eswa.2018.09.015

Malakar, S., Ghosh, M., Bhowmik, S., Sarkar, R., and Nasipuri, M. (2019). A GA Based Hierarchical Feature Selection Approach for Handwritten Word Recognition. *Neural Comput. Appl.* 32 (7), 2533–2552. doi:10.1007/s00521-018-3937-8

Margolis, R., Derr, L., Dunn, M., Huerta, M., Larkin, J., Sheehan, J., et al. (2014). The National Institutes of Health's Big Data to Knowledge (BD2K) Initiative: Capitalizing on Biomedical Big Data. *J. Am. Med. Inform. Assoc.* 21, 957–958. doi:10.1136/amiajnl-2014-002974

Mckinney, B. C., Hensler, C. M., Wei, Y., Lewis, D. A., and Sweet, R. A. (2020). Schizophrenia-associated Differential DNA Methylation in the superior Temporal Gyrus Is Distributed to many Sites across the Genome and Annotated by the Risk Gene MAD1L1. *medRxiv*. doi:10.1101/2020.08.02.20166777

Moosavi, Shs., and Bardsiri, V. K. (2019). Poor and Rich Optimization Algorithm: A New Human-Based and Multi Populations Algorithm. *Eng. Appl. Artif. Intelligence* 86, 165–181. doi:10.1016/j.engappai.2019.08.025

Notterman, D. A., Alon, U. A., Sierk, A. J., and Levine, A. J. (2001). Transcriptional Gene Expression Profiles of Colorectal Adenoma, Adenocarcinoma, and Normal Tissue Examined by Oligonucleotide Arrays. *Cancer Res.* 61, 3124–3130. https://cancerres.aacrjournals.org/content/61/7/3124.long

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., and Sturia, L. M. (2002). Prediction of central Nervous System Embryonal Tumour Outcome Based on Gene Expression. *Nature* 415, 436–36. doi:10.1038/415436a

Qiu, Y., Ching, W.-K., and Zou, Q. (2021). Prediction of RNA-Binding Protein and Alternative Splicing Event Associations during Epithelial–Mesenchymal Transition Based on Inductive Matrix Completion. *Brief. Bioinform.* 22, bbaa440. doi:10.1093/bib/bbaa440

Robeck, T. R., Fei, Z., Lu, A. T., Haghani, A., Jourdain, E., Zoller, J. A., et al. (2021). Multi-species and Multi-Tissue Methylation Clocks for Age Estimation in Toothed Whales and Dolphins. *Commun. Biol.* 4, 1–11. doi:10.1038/s42003-021-02179-x

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., et al. (2002). Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning. *Nat. Med.* 8, 69–74. doi:10.1038/nm0102-68

Smith, R. G., Hannon, E., De Jager, P. L., Chibnik, L., Lott, S. J., Condliffe, D., et al. (2018). Elevated DNA Methylation across a 48-kb Region Spanning the HOXA Gene Cluster Is Associated with Alzheimer's Disease Neuropathology. *Alzheimer Demen.* 14, 1580–1588. doi:10.1016/j.jalz.2018.01.017

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B., et al. (2003). The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma. *N. Engl. J. Med.* 349, 2483–2494. doi:10.1056/nejmoa030847

Varsha, D., Gibbs, D. L., Theo, K., Roger, K., Joseph, V., John, N., et al. (2016). Using Incomplete Trios to Boost Confidence in Family Based Association Studies. *Front. Genet.* 7, 34. doi:10.3389/fgene.2016.00034

Wang, G., Hu, N., Yang, H. H., Wang, L., Hua, S., Wang, C., et al. (2013). Comparison of Global Gene Expression of Gastric Cardia and Noncardia Cancers from a High-Risk Population in China. *Plos One* 8, e63826. doi:10.1371/journal.pone.0063826

Wu, Y., Grabsch, H., Ivanova, T., Tan, I. B., Murray, J., Ooi, C. H., et al. (2013). Comprehensive Genomic Meta-Analysis Identifies Intra-tumoural Stroma as a Predictor of Survival in Patients with Gastric Cancer. *Gut* 62, 1100–1111. doi:10.1136/gutjnl-2011-301373

Yapici, H., and Cetinkaya, N. (2019). A New Meta-Heuristic Optimizer: Pathfinder Algorithm. *Appl. Soft Comput.* 74, 545–568. doi:10.1016/j.asoc.2019.03.012

Ylitalo, E. B., Thysell, E., Landfors, M., Brattsand, M., Jernberg, E., Crnalic, S., et al. (2021). A Novel DNA Methylation Signature Is Associated with Androgen Receptor Activity and Patient Prognosis in Bone Metastatic Prostate Cancer. *Clin. Epigenetics* 13, 1–15. doi:10.1186/s13148-021-01119-0

Yuanyuan, H., Huang, L., and Zhou, F. (2021). A Dynamic Recursive Feature Elimination Framework (dRFE) to Further Refine a Set of OMIC Biomarkers. *Bioinformatics*, 37 (15), 2183–2189. doi:10.1093/bioinformatics/btab055

Zhang, Y., Jin, Z., and Mirjalili, S. (2020). Generalized normal Distribution Optimization and its Applications in Parameter Extraction of Photovoltaic Models. *Energ. Convers. Manage.* 224, 113301. doi:10.1016/j.enconman.2020.113301

Zhao, W., Wang, L., and Zhang, Z. (2019). Atom Search Optimization and its Application to Solve a Hydrogeologic Parameter Estimation Problem. *Knowledge-Based Syst.* 163, 283–304. doi:10.1016/j.knosys.2018.08.030

Zhao, W., Zhang, Z., and Wang, L. (2020). Manta ray Foraging Optimization: An Effective Bio-Inspired Optimizer for Engineering Applications. *Eng. Appl. Artif. Intelligence* 87, 103300. doi:10.1016/j.engappai.2019.103300

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