COMPUTATIONAL METHODS IN INFERRING CANCER TISSUE-OF-ORIGIN AND CANCER MOLECULAR CLASSIFICATION, VOLUME I

EDITED BY: Min Tang, Cheng Guo, Ling Kui, Shuai Cheng Li and Jialiang Yang
PUBLISHED IN: Frontiers in Genetics, Frontiers in Bioengineering and Biotechnology and Frontiers in Oncology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-654-6
DOI 10.3389/978-2-88966-654-6

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

COMPUTATIONAL METHODS IN INFERRING CANCER TISSUE-OF-ORIGIN AND CANCER MOLECULAR CLASSIFICATION, VOLUME I

Topic Editors:
Min Tang, Jiangsu University, China
Cheng Guo, Columbia University, United States
Ling Kui, Harvard Medical School, United States
Shuai Cheng Li, City University of Hong Kong, Hong Kong
Jialiang Yang, Geneis (Beijing) Co. Ltd, China

Topic Editor Dr. Jialiang Yang is employed by Geneis Co. Ltd. The rest of the Topic Editors declare no conflicts of interest with regards to this Research Topic.

Citation: Tang, M., Guo, C., Kui, L., Li, S. C., Yang, J., eds. (2021). Computational Methods in Inferring Cancer Tissue-of-Origin and Cancer Molecular Classification, Volume I. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-654-6

Table of Contents

frontiers in Genetics

EDITORIAL
published: 23 February 2021
doi: 10.3389/fgene.2021.644542

Editorial: Computational Methods in Inferring Cancer Tissue-of-Origin and Cancer Molecular Classification

Ling Kui[1,2], *Cheng Guo*[3], *Shuai Cheng Li*[4], *Jialiang Yang*[5] *and Min Tang*[6*]

[1] School of Pharmacy, Jiangsu University, Zhenjiang, China, [2] Harvard Medical School, Dana-Farber Cancer Institute, Boston, MA, United States, [3] Center for Infection and Immunity, School of Public Health, Columbia University, New York, NY, United States, [4] Department of Computer Science, City University of Hong Kong, Kowloon, China, [5] Geneis (Beijing) Co. Ltd., Beijing, China, [6] School of Life Sciences, Jiangsu University, Zhenjiang, China

Keywords: cancer tissue-of-origin, cancer molecular classification, liquid biopsy, machine learning, single-cell

Editorial on the Research Topic

Computational Methods in Inferring Cancer Tissue-of-Origin and Cancer Molecular Classification

The development of cancer therapeutics increasingly relies on the results of tissue-of-origin and molecular classification. In the clinic, up to 5% of the cancer primary site is unclassified (CUP). For clinicians, it is important to identify the sensitive patients and determine treatment. The main option is empirical chemotherapy, which leads to a lower survival rate. Therefore, inferring cancer tissue-of-origin is an urgent need to be solved. The key point is to detect the exact genetic events associated with cancer formation, which usually contribute to cell proliferation and uncontrolled metabolic changes. However, using only experimental approaches cannot provide a full view of the genetic features in the era of big biomedical data. Although a series of computational methods have been developed in this area, the accuracy is often insufficient for clinical use.

The molecular classification in cancer is useful in optimizing treatment policies. With data accumulation, especially more and more single-cell sequencing data, the molecular classification will be improved for various cancer types. As better biomarkers evolve, more efficient treatments and new drugs will be developed.

This Research Topic gathered research articles and reviews representing not only the computational methods for inferring the origins and molecular classification but also translational studies for cancer treatment in hospitals. This collection of papers sheds light on the development of cancer therapeutics, with a focus on the most cutting-edge computational applications in cancer diagnosis.

The 19 published articles consist of 18 research papers and a regular review, which comprehensively illustrates the use of computational methods in inferring cancer Tissue-of-Origin and molecular classification in various cancer types, including but not limited to hepatocellular carcinoma (HCC), Pancreatic cancer (PC), ovarian cancer (OC), glioma, gastric cancer (GC), circulating tumor cells (CTCs), cervical cancer (CC), and endometrial cancer (EC).

Seven research articles introduce several different methods to capture gene signature (models) for similar purposes. Li et al. first employed the limma R package to the got the top 5,000 significant differentially expressed genes (DGEs) in HC. These DEGs were gathered into nine modules after they underwent a weighted correlation network analysis (WGCNA). Then, six genes were screened by univariate, LASSO, and multivariate Cox regression analysis, and they were validated as an independent prognostic factor in survival analysis (Li et al.). Most of the bioinformatic approaches in this study were implemented in the article of Zhang et al., whose aim was to develop a stemness

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

***Correspondence:**
Min Tang
mt3138@ujs.edu.cn

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

Received: *21 December 2020*
Accepted: *27 January 2021*
Published: *23 February 2021*

Citation:
Kui L, Guo C, Li SC, Yang J and Tang M (2021) Editorial: Computational Methods in Inferring Cancer Tissue-of-Origin and Cancer Molecular Classification. Front. Genet. 12:644542. doi: 10.3389/fgene.2021.644542

index-based gene signature for lower-grade glioma (LGG). Interestingly, the same research group developed an immune-related signature for prognosis prediction and risk stratification in LGG with data from The Cancer Genome Atlas (TCGA), Genome Tissue Expression (GTEx), and Chinese Glioma Genome Atlas (CGGA) (Zhang et al.). A similar study in CC and EC was completed by Ding et al. Importantly, they validated the gene signature with many methods, such as enrichment analyses through GO, KEGG, and GSEA pathways, Kaplan-Meier survival curve, ROC curves, and immune cell infiltration (Ding et al.). Moreover, Pan et al. also demonstrated that gene methylation can be utilized to classify gliomas as signatures. They used advanced computational methods of Monte Carlo feature selection (MCFS), incremental feature selection (IFS), and support machine vector (SVM) to detect methylation features related to glioma subclasses (Pan et al.). A back-to-back study performed by Hou et al. illustrated the functions and mechanisms of N6-methyladenosine (m6A) modification in the development of PC. A six-m6A-regulator-signature related to overall survival (OS) was identified by LASSO regression (Hou et al.). Furthermore, Kieffer et al. established gene signatures by combining transcriptomic and genomic data for high-grade OC.

Notably, three research articles elucidate the application of machine learning in gene feature captivation. Using DNA somatic mutation data, Liu et al. extracted genetic features using the random forest algorithm and established a logistic regression-based classifier. With the extracted matrix of features from the functional 300 genes, the prediction accuracy can reach up to 81% in 10-fold cross-validation. To reduce the workload of CTCs counting and improve the automation level, He et al. established a cell recognition program based on deep learning to identify the CTCs. In their project, the CTCs images of 600 in-house patients were analyzed with python's OpenCV scheme for segmentation. Then, convolutional neural network deep learning networks in machine learning algorithms were implemented on 1,300 cells for training, and the others were used for testing. The final specificity and sensitivity of recognition reached 91.3 and 90.3%, respectively (He et al.). Qian et al. provide a feature extraction algorithm based on Support Vector Machines (SVM) for cancer lectins prediction with a fusion of G-Gap dipeptide.

Three research articles focus on the development of computational approaches. Zhu et al. exploited a prediction model called MiRNA-Disease Association prediction (BHCMDA) based on the Biased Heat Conduction (BHC) algorithm to discover potentially associated miRNAs of diseases by integrating known miRNA-disease associations, the disease semantic similarity, the miRNA functional similarity, and the Gaussian interaction profile kernel similarity. Zhao et al. created a novel computational approach named multiplex biological network (MON) by integrating protein interaction networks (PINs), protein domains, and gene expression files. The new approach was able to detect the essential proteins by extending the random walk with a restart algorithm to the tensor (Zhao et al.). To predict lung cancer recurrence after surgical resection, Wu et al. established a convolutional neural network (CNN) framework called DeepLRHE by analyzing histopathological images of patients from the TCGA database, and the receiver operating characteristic (ROC) curve (AUC) was 0.79.

Finally, the systematic review demonstrates in detail that the CLDN18-ARHGAP fusion is a significant molecular characteristic of diffuse GC, which is also an independent prognostic risk factor (Zhang et al.).

All of the research articles and reviews in this Research Topic use state-of-the-art sources about the origin and gene signatures of different cancers, examining the available computational methods and providing a guide for physicians.

AUTHOR CONTRIBUTIONS

This editorial was designed by MT and written by LK and CG. SL and JY revised it. All authors made a direct and intellectual contribution to this topic and approved the article for publication.

FUNDING

This work was supported by grants from Jiangsu University (19JDG039 and 20JDG47) and an ARG project from CityU (9667204).

Conflict of Interest: JY was employed by the company Geneis Co. Ltd. (Beijing).

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

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 14 November 2019
doi: 10.3389/fbioe.2019.00339

Screening of Methylation Signature and Gene Functions Associated With the Subtypes of Isocitrate Dehydrogenase-Mutation Gliomas

XiaoYong Pan[1,2,3†], Tao Zeng[4†], Fei Yuan[5], Yu-Hang Zhang[6], Lei Chen[7,8], LiuCun Zhu[1], SiBao Wan[1], Tao Huang[6] and Yu-Dong Cai[1*]*

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Key Laboratory of System Control and Information Processing, Ministry of Education of China, Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, [3] IDLab, Department for Electronics and Information Systems, Ghent University, Ghent, Belgium, [4] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai, China, [5] Department of Science and Technology, Binzhou Medical University Hospital, Binzhou, China, [6] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, [7] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [8] Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai, China

Isocitrate dehydrogenase (IDH) is an oncogene, and the expression of a mutated IDH promotes cell proliferation and inhibits cell differentiation. IDH exists in three different isoforms, whose mutation can cause many solid tumors, especially gliomas in adults. No effective method for classifying gliomas on genetic signatures is currently available. DNA methylation may be applied to distinguish cancer cells from normal tissues. In this study, we focused on three subtypes of IDH-mutation gliomas by examining methylation data. Several advanced computational methods were used, such as Monte Carlo feature selection (MCFS), incremental feature selection (IFS), support machine vector (SVM), etc. The MCFS method was adopted to analyze methylation features, resulting in a feature list. Then, the IFS method incorporating SVM was applied to the list to extract important methylation features and construct an optimal SVM classifier. As a result, several methylation features (sites) were found to relate to glioma subclasses, which are annotated onto multiple genes, such as *FLJ37543*, *LCE3D*, *FAM89A*, *ADCY5*, *ESR1*, *C2orf67*, *REST*, *EPHA7*, etc. These genes are enriched in biological functions, including cellular developmental process, neuron differentiation, cellular component morphogenesis, and G-protein-coupled receptor signaling pathway. Our results, which are supported by literature reports and independent dataset validation, showed that our identified genes and functions contributed to the detailed glioma subtypes. This study provided a basic research on IDH-mutation gliomas.

Keywords: isocitrate dehydrogenase, methylation, IDH-mutation, gliomas, multi-class classification

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Xiao Chang,
Children's Hospital of Philadelphia, United States
Guang Wu,
Guangxi Academy of Sciences, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cail_yud@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *09 September 2019*
Accepted: *30 October 2019*
Published: *14 November 2019*

Citation:
Pan X, Zeng T, Yuan F, Zhang Y-H, Chen L, Zhu L, Wan S, Huang T and Cai Y-D (2019) Screening of Methylation Signature and Gene Functions Associated With the Subtypes of Isocitrate Dehydrogenase-Mutation Gliomas. Front. Bioeng. Biotechnol. 7:339. doi: 10.3389/fbioe.2019.00339

INTRODUCTION

Isocitrate dehydrogenase (IDH) exists in three different isoforms. IDH1 and DH2 catalyze the same reaction and use NADP+ as a cofactor instead of NAD+. IDH3 converts NAD+ to NADH in the mitochondria. IDH is an oncogene, and the expression of mutated IDH promotes cell proliferation and inhibits cell differentiation. Mutant IDH-derived (R)-2HG is a potential malignant substance and unwanted byproduct of cellular metabolism. 2HG dehydrogenase (2HGDH) prevents 2HG from accumulating in cells, and its intracellular levels in normal cells are maintained at <0.1 mM. The transformation induced by (R)-2HG is effective and reversible, suggesting that inhibiting 2HG has efficacy in the treatment of IDH mutant cancers. Mutations at Arg132 of IDH1 are present in five of six secondary glioblastoma (GBM) subtypes, and IDH mutations have been found in many other solid tumors (Losman and Kaelin, 2013).

Glioma in adults includes three main categories, namely, glioblastoma (GBM), astrocytoma, and oligodendroglioma. They are determined by genetic and histologic features. IDH1 and IDH2 mutations are generally detected in astrocytoma and oligodendroglioma but not in the GBM subtype. Thus, IDH-mutation is an important marker for glioma classification. Different subtypes of glioma have different mutation patterns. Mutations in ATRX and TP53 are usually identified in astrocytomas with mutant IDH, but TRET promoter variations and chromosome abnormality are generally identified in oligodendrogliomas (O-IDH) (Cancer Genome Atlas Research Network et al., 2015). Thus, A-IDH and O-IDH are two major subtypes of IDH-mutant gliomas distinguished by co-occurring genetic signatures and histopathology (Venteicher et al., 2017).

No effective method for classifying gliomas on genetic signatures is currently available. By contrast, DNA methylation is used to distinguish cancer cells from normal tissues (Delpu et al., 2013). DNA methylation is a part of the normal epigenetic modification with potential regulatory significance, such as regulating gene expression patterns. In this study, we focused on three subtypes of IDH-mutation gliomas by methylation data, including astrocytomas with IDH mutations (A-IDH), astrocytoma with IDH mutation and enriched HG (A-IDH-HG), and oligodendrogliomas with IDH mutations (O-IDH). Our analyzing procedures used several advanced computational methods, like Monte Carlo feature selection (MCFS; Draminski et al., 2008), incremental feature selection (IFS; Liu and Setiono, 1998), and support machine vector (SVM; Cortes and Vapnik, 1995), etc. A feature list was produced by applying the MCFS method on the methylation data. Then, the IFS method followed to extract important methylation features by evaluating the performance of SVM on different feature subsets that consisted of top features in the list. As a result, we accessed some key methylation features (sites) related to the classification of gliomas annotated onto multiple genes, such as *FLJ37543*, *LCE3D*, *FAM89A*, *ADCY5*, *ESR1*, *C2orf67*, *REST*, *EPHA7*, etc. Furthermore, we obtained several biological functions related to the classification of glioma subtypes, which are also related to gene methylation and corresponding functions, such as cellular developmental process, neuron differentiation, cellular component morphogenesis, and G-protein-coupled receptor signaling pathway. We then validated these methylation signatures, genes, and functions on an independent dataset. We identified a group of methylation sites, genes, and functions by using our screening analysis method. This study provided a basic research on the detailed classification of A-IDH and O-IDH cases.

MATERIALS AND METHODS

Data Sources

We downloaded the methylation profiles of patients with IDH-mutation glioma from GEO (Gene Expression Omnibus) under accession numbers GSE90496 and GSE109379, which were originally generated by Capper et al. (2018). The GSE90496 dataset was used as a training dataset, and the GSE109379 dataset was used as an independent test dataset. The training dataset had samples of 78 A-IDH subclasses, 46 high-grade astrocytoma (A-IDH-HG) subclasses, and 80 1p/19q co-deleted O-IDH subclasses. The test dataset had 94 A-IDH, 41 A-IDH-HG, and 83 O-IDH samples. The overlapped 42,383 methylation probes between training and test datasets were used to encode IDH-mutation glioma in each patient to investigate the methylation difference among different IDH-mutation glioma subclasses.

Feature Selection

In this study, we first used MCFS (Chen et al., 2018a, 2019a,b; Pan et al., 2018, 2019a,b; Li et al., 2019) to rank the input features, and the ranked features were further selected through IFS (Zhang et al., 2015; Zhou et al., 2015; Chen et al., 2017b,c, 2018b; Wang et al., 2017; Li and Huang, 2018; Zhang T. M. et al., 2018) with a supervised classifier SVM (Cortes and Vapnik, 1995).

MCFS is a supervised feature selection method based on multiple decision trees (Draminski et al., 2008). We used it to generate m bootstrap sample sets and t feature subsets from original data. One decision tree was grown on the basis of each combination of bootstrap sets and feature subsets. A total of $m \times t$ decision trees was obtained. According to these trees, we calculated relative importance (RI) score for each feature. The main criterion is that the more frequent a feature is involved in splitting nodes of growing the $m \times t$ trees, the more important the feature will be; the accuracy of each decision tree is also considered for evaluating the importance of this feature. In detail, the RI score for one feature f is computed by

$$RI_f = \sum_{\tau=1}^{m\times t} (wAcc)^u IG(n_f(\tau))\left(\frac{no.in\ n_f(\tau)}{no.in\ \tau}\right)^v,$$

where $wAcc$ stands for the weighted accuracy, $n_f(\tau)$ represents a node of f in decision tree τ, the information gain of $n_f(\tau)$ is denoted as $IG(n_f(\tau))$, $no.in\ n_f(\tau)$ stands for the number of samples in $n_f(\tau)$, $no.in\ \tau$ indicates the number of samples in τ. u and v are weighting factors, which were set to one in this study. After accessing the RI scores of all features, we ranked them in a list in terms of the decreasing order of their RI scores.

MCFS only ranked the input features but could not remove redundant features. The feature selection by an arbitrary cutoff

of RI score was not the best method. Thus, IFS, which is a feature selection method with a supervised classifier, was further used to identify the optimum number of features for classification. IFS first generated a series of feature subsets with a step of 10 based on the ranked features from MCFS. The first feature subset consisted of the top 10 features, the second feature subset comprised the top 20 features, and so on. A supervised classifier was built and evaluated on the samples consisting of the features from each feature subset through 10-fold cross-validation. Lastly, we selected the optimum feature subset with the best performance.

Supervised Classifiers

We integrated IFS with SVM. To compare the performance baseline, we also evaluated the IFS with random forest (RF; Ho, 1995) and repeated incremental pruning to produce error reduction (RIPPER; Cohen, 1995).

SVM is a supervised classification algorithm based on statistical theory (Cortes and Vapnik, 1995). It finds a hyperplane with the maximum margin between two classes. SVM can handle linear and non-linear data. For non-linear data, SVM first maps the original data into a high-dimensional space by using kernels in which new data can be linearly separable. SVM is designed for binary classification, and one-vs.-the-rest strategy is used for multi-class classification. Multiple SVMs are trained, and each SVM is trained on positive samples from one class and negative samples from the remaining classes. A new sample is assigned a predicted class label corresponding to the highest probability score from one SVM.

RF is a supervised meta-classifier based on multiple decision trees (Ho, 1995). It grows multiple decision trees from bootstrap sets, and each decision tree is trained on a randomly selected feature subset. In contrast to SVM, RF can be directly applied to multiclass classification.

RIPPER is a rule-based classifier that greedily produces classification rules (Cohen, 1995). It first finds a good rule to cover training samples as much as possible and then removes the covered samples from the training set for mining the next rule. RIPPER repeats the above process until all the samples are covered by the produced classification rules.

To quickly implement above-mentioned three classification algorithms, three tools "SMO," "RandomForest," and "JRip" in Weka (Witten and Frank, 2005) were employed. Their default parameters were used.

GO- and KEGG-Based Enrichment Analysis

To investigate whether the selected methylation probes were significantly enriched onto certain biological functions, we did the GO and KEGG enrichment analysis. The identified methylation probes were mapped onto genes based on the probe annotations of Illumina HumanMethylation450 BeadChip at GEO under the accession number GPL13534. The genes were enriched onto GO and KEGG terms by using hypergeometric test. We used R function phyper to perform the hypergeometric test. The KEGG database Release 86.0 was retrieved using R/Bioconductor package KEGGREST (https://bioconductor.org/packages/KEGGREST/) and the GO database with date stamp of 2017-Nov01 was provided in R/Bioconductor package org.Hs.eg.db (https://bioconductor.org/packages/org.Hs.eg.db/). The hypergeometric test *P*-values were adjusted to obtain their false discovery rate (FDR). The GO terms and KEGG pathways with FDR smaller than 0.05 were considered as significant and analyzed.

Performance Evaluation

We used a multiclass classifier to classify samples from A-IDH, A-IDH-HG, and O-IDH and evaluated the trained classifiers by using 10-fold cross-validation (Kohavi, 1995; Chen et al., 2017c, 2018b; Li et al., 2019; Zhang et al., 2019; Zhou et al., 2019) on the training set. To further demonstrate the generalization ability of model learning, we examined the trained classifiers on an independent test set. We also considered Matthews correlation coefficient (MCC; Matthews, 1975; Gorodkin, 2004; Chen et al., 2017a; Zhao et al., 2018, 2019; Cui and Chen, 2019), accuracies of individual classes, and overall accuracy to measure model performance.

RESULTS

In this study, we adopted several advanced computational methods to investigate the methylation profiles of patients with three IDH-mutation glioma subclasses. The entire procedures are illustrated in **Figure 1**.

We first ranked 42,383 features (e.g., methylation sites) as the input by using MCFS. The RI scores of the input features are given in **Table S1**. A total of 19,692 features have RI scores >0, and the remaining 22,691 features have no any discriminative ability to classify samples from A-IDH, A-IDH-HG, and O-IDH. Thus, only 19,692 features were used for the tasks below.

Next, we evaluated the IFS with an SVM on the training set by using 10-fold cross-validation. **Table 1** shows that we yielded the best MCC value of 0.977 when the top 750 features were used, with an overall accuracy of 0.985. The accuracies on three subclasses were 0.987, 0.957, and 1.000, respectively, indicating the good performance of SVM based on top 750 features. **Figure 2B** illustrates that the MCCs of SVMs changed with the number of the involved features. To justify why we selected SVM as the final classifier of IFS, we also evaluated the performance of IFS with RF and RIPPER. In **Table 1**, **Figures 2A,C**, IFS with RF yielded the best MCC value of 0.962 and an overall accuracy of 0.975 when the top 1,330 features were used. The accuracies on three subclasses were 0.987, 0.913, and 1.000, respectively. RF used more features but yielded a lower performance than SVM did. By contrast, the rule-based method RIPPER yielded lower performance than SVM and RF did, thereby achieving the MCC of 0.895 when the top 19,270 features were utilized. The accuracies on three subclasses were also lower than those of SVM and RF (see the last row of **Table 1**). RIPPER was worse than SVM and RF because RIPPER is a rule-based method that considers the balance between detecting interpretable classification rules and obtaining the high classification performance of "black-box." The performance corresponding to the number of features of SVM, RF, and RIPPER is given in **Table S2**.

To further demonstrate the generalizability of our learned models, we further evaluated the IFS with SVM, RF, and RIPPER

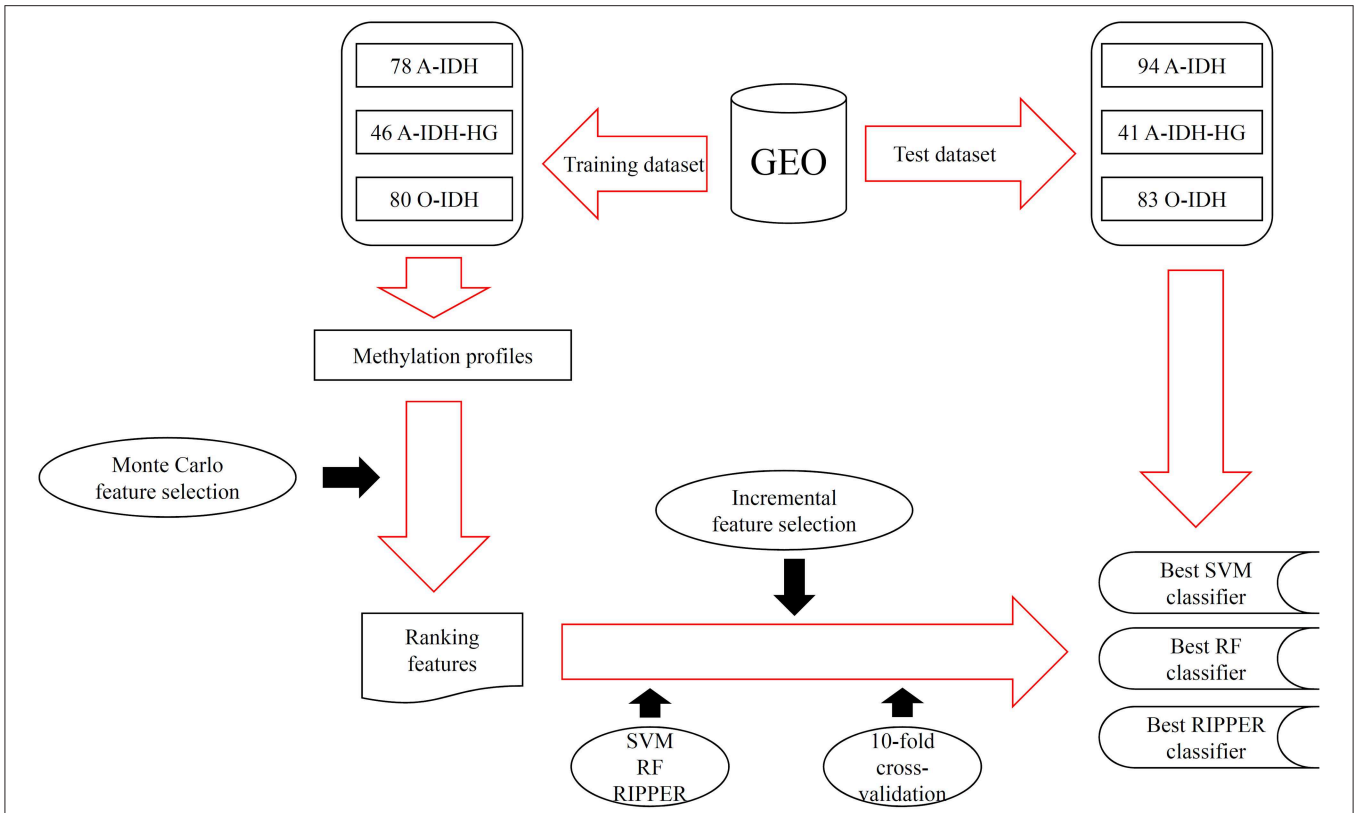


FIGURE 1 | The entire procedures for investigating the methylation profiles of patients with three IDH-mutation glioma subclasses.

TABLE 1 | The 10-fold cross-validation performance of IFS with different classifiers on the training set.

Classifier	Number of optimum features	Accuracy			Overall accuracy	MCC
		A-IDH	A-IDH-HG	O-IDH		
SVM	750	0.987	0.957	1.000	0.985	0.977
SVM	20	1.000	0.913	1.000	0.980	0.970
RF	1,330	0.987	0.913	1.000	0.975	0.962
RIPPER	19,270	0.962	0.848	0.950	0.931	0.895

on the independent test set. **Table 2** shows their performance on the independent test set, where the same number of optimum features identified on the training set was used for each classifier. The MCCs yielded by SVM, RF, and RIPPER were 0.899, 0.907, and 0.972, respectively. The three methods achieved a high performance, demonstrating the generalizability of the trained models. RIPPER yielded the lowest 10-fold cross-validation performance on the training set, but it yielded the highest performance on the independent test set. This result indicated that the simple rule-based method RIPPER might not easily suffer model overfitting compared with that of complicated classifiers SVM and RF, but too many features were used in this classifier.

As mentioned above, SVM with top 750 features yielded the best performance on the training set. However, when top 20 features were used, the SVM generated the MCC of 0.970, which was only 0.007 lower than that obtained by the SVM with top 750 features. Considering the efficiency of SVM, SVM with top 20 features was a more proper choice. Its performance on three classes is listed in **Table 1**, which was almost at the same level compared with that of the SVM with top 750 features. Furthermore, its performance on the test set is listed in **Table 2**, which was still acceptable.

DISCUSSION

We found 750 optimal features for distinguishing A-IDH, A-IDH-HG, and O-IDH with the help of SVM. However, considering the efficiency, SVM with top 20 features was a more suitable choice. Thus, it is believed that these 20 features were extremely important. Here, we gave an extensive discussion on these 20 features (**Table 3**), which were supported by previous studies. In addition, we further identified a group of detailed biological functions associated with different IDH-mutation glioma subclasses.

Genes Associated With Glioma Subclasses

The top probe was **cg04437966**, marking gene *FLJ37543*. Also known as *C5orf64*, such gene has been widely reported to participate in tumorigenesis (Aschebrook-Kilfoy et al., 2015). As for its potential contribution on distinguishing different IDH subtypes, it has been reported to participate in multiscale

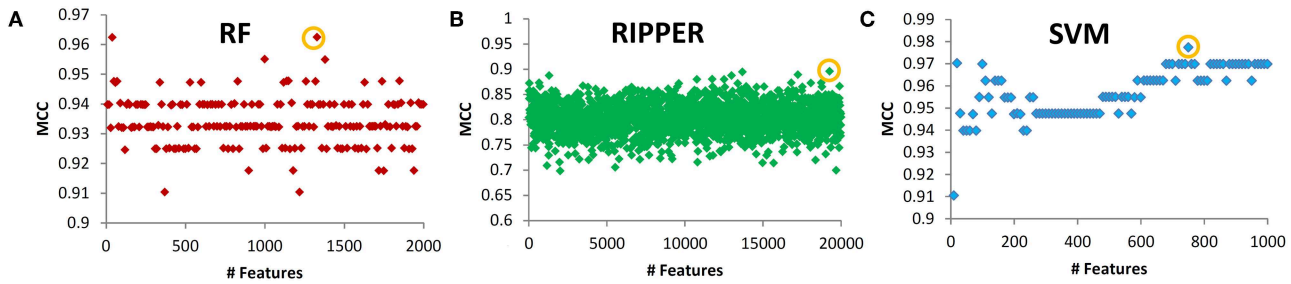


FIGURE 2 | Performance of SVM, RF, and RIPPER that changed with the corresponding number of features. **(A)** RF performance, **(B)** RIPPER performance, and **(C)** SVM performance.

TABLE 2 | The performance of IFS with different classifiers on the independent test set.

Classifier	Number of features	Accuracy			Overall accuracy	MCC
		A-IDH	A-IDH-HG	O-IDH		
SVM	750	0.947	0.780	1.000	0.936	0.899
SVM	20	0.926	0.756	0.964	0.908	0.855
RF	1,330	0.968	0.756	1.000	0.940	0.907
RIPPER	19,270	0.957	1.000	1.000	0.982	0.972

TABLE 3 | Top features (methylation probes) and their targeting genes.

Rank	Feature	Targeting gene	RI
1	cg04437966	*FLJ37543*	0.5637
2	cg14159026	*BVES*	0.4719
3	cg22519158	*LCE3D*	0.3781
4	cg12450347	*FAM89A*	0.3505
5	cg17482114	*ADCY5*	0.3397
6	cg08415493	*ESR1*	0.3244
7	cg12760041	*C2orf67*	0.3119
8	cg12930304	–	0.2875
9	cg26694713	*REST*	0.2846
10	cg04360458	*REST*	0.2591
11	cg17398252	*BVES*	0.2497
12	cg21552709	*EPHA7*	0.2374
13	cg20138711	*ARHGEF3*	0.2327
14	cg11902641	–	0.2271
15	cg03903398	*MIR1275*	0.2052
16	cg19681793	*THBS2*	0.1916
17	cg24215279	*TPO*	0.1889
18	cg05427966	*EPHA7*	0.1797
19	cg11235583	*CLCNKB*	0.1766
20	cg14158583	*PVRL4*	0.1739

modeling of oligodendrocytes in physical and pathological conditions, but not other neural cell subtypes (Mckenzie et al., 2017). Therefore, the expression level of such gene may actually contribution to the subtyping processes.

The next probe was **cg14159026**, identifying gene *BVES*. Encoding a specific member of the POP family of protein, such gene has been widely reported to participate in cell adhesion processes (Wada et al., 2001). As for its specific contribution on IDH-dependent glioma subtyping, it has been reported that such gene can participate in the development of different neural cells and functionally related to IDH (Lord et al., 1997; Ton et al., 2002). Therefore, although no direct reports confirmed its unique classification potentials for glioma subtyping, it is reasonable for us to regard such gene as a reference for IDH-dependent glioma subtyping. Apart from such probe, another effective probe named as **cg17398252** is also designed to detect the methylation status of such gene, further confirming above results.

The third probe was **cg22519158**, detecting the methylation status of gene *LCE3D*. *LCE3D* is also a specific development associated gene, participating in the formation of stratum corneum (Bergboer et al., 2011). As for its potential relationship with IDH and its contribution on such subtyping, it has been reported that such gene is related to the expression of IDH and different subtypes of glioma at methylation level, corresponding with our results (Zhang M. et al., 2018).

FAM89A, as the following identified target gene is marked by the fourth probe, named **cg12450347**. There are no detailed reports on the biological functions of *FAM89A*. However, the abnormal expression level of such gene has also been screened out on some glioma gene expression profiling studies (Mascelli et al., 2013; Xie et al., 2017). Therefore, our screened-out probe definitely contributes to the IDH-dependent subtyping of glioma.

The next gene *ADCY5*, detected by probe **cg17482114**, is an enzyme that interacts with *RGS2* in humans. *ADCY5* is associated with various neurological syndromes in non-cancer tissues and can cause chorea, a type of neurological syndrome (Walker, 2016). The SNPs of *ADCY5* are associated with elevated fasting glucose and increased type 2 diabetes risk. The DNA hypermethylation of *ADCY5* induces a low mRNA expression pattern in malignant tissue samples (Sato et al., 2013).

ESR1, detected by probe **cg08415493**, was also identified to participate in IDH-dependent glioma subtyping. Encoding an estrogen receptor, such gene has been widely reported to participate in hormone related cell proliferation and differentiation (Dalvai and Bystricky, 2010; Mascelli et al., 2013). In glioma, such gene has been reported to be a specific biomarker for glioma subtyping on expression and methylation

level (Uhlmann et al., 2003). Considering that such gene has also been identified to be functionally related to IDH, it is quite reasonable to regard such gene as a potential marker for such subtyping (Richardson et al., 2019).

C2orf67, as the target of probe **cg12760041,** was also identified in this study. According to recent publications, such gene has been reported to be effective as a serum metabolite measurement parameter (Ohyama et al., 2016; Aibara et al., 2018). As for the methylation status and expression pattern of such gene in different glioma subtypes, it has been identified as one of the potential markers reflecting the activation status of EGF signaling pathway (Trang et al., 2010). Considering that different IDH-dependent glioma subtypes have different EGF activation status (Roth and Weller, 2014; Thorne et al., 2016), it is reasonable to identify such gene and its targeted probe as one of the potential markers for such IDH-dependent subtyping.

REST, targeted by probes named as **cg26694713** and **cg04360458,** is also predicted to participate in IDH-dependent glioma subtyping. *REST* is actually a transcriptional regulatory factor for neuronal genes (Zuccato et al., 2003). Apart from that, *REST* has also been identified as a specific marker for glioma subtyping due to its epigenetic alteration pattern (Zuccato et al., 2003). In the same report, the mutation status of IDH has also been validated to be functionally related to such methylation alteration (Zuccato et al., 2003).

The next two probes, named as **cg21552709** and **cg05427966**, target Ephrin type-A receptor 7 (*EPHA7*). *EPHA7*, as a member of the ephrin receptor superfamily, mediates developmental events, particularly in the nervous system. During the embryonic development of the central nervous system, Ephs and ephrins have defined functions, such as axon mapping, neural crest cell migration, hindbrain segmentation, synapse formation, and physiological and abnormal angiogenesis. Eph and ephrins are frequently overexpressed in different tumor types, including GBM. An increased *EphA7* expression is correlated with adverse outcomes in patients with primary and recurrent glioblastoma multiforme (Wang et al., 2008).

The next probe **cg20138711** targeting *ARHGEF3* was screened out in our study, which were deemed to contribute to IDH-dependent glioma subtyping. *ARHGEF3* is a regulator for RhoA and RhoB GTPases (Hilgers and Webb, 2005). According to recent publications, mediating RhoA associated biological processes, *ARHGEF3* has been confirmed to interact with IDH (Okada et al., 2003; Kloth et al., 2005) and has unique methylation status in glioma (Northcott et al., 2009). Therefore, it is quite reasonable to summary that such probe actually targets an effective regulatory gene for IDH-dependent glioma subtyping.

Probe **cg03903398** is another informant feature targeting effective microRNA, coding gene named as *MIR1275*. *MIR1275* is a functional microRNA coding gene, which has been directly reported to participate in multiple sclerosis (MS; Angerstein et al., 2012). As for its specific role for glioma subtyping, similar with gene *ARHGEF3*, such microRNA participates in TGF-beta signaling pathway (Yan et al., 2013) and has been validated to have different methylation status together with expression pattern in different IDH expression glioma subtypes (Kondo et al., 2014).

The following four probes **cg19681793** (targeting *THBS2*), **cg24215279** (targeting *TPO*), **cg11235583** (targeting *CLCNKB*), and **cg14158583** (targeting *PVRL4*) have also been confirmed to target effective genes with different methylation status in different IDH-dependent glioma subtypes. Apart from above-discussed eighteen probes, **cg12930304** and **cg11902641** were also identified to be significant for subtyping. However, according to the annotation, no actual genes are presented in such region, which may be induced by incomplete annotation reference or prediction redundancy. All in all, most genes corresponding

TABLE 4 | The significantly enriched GO/KEGG functions with FDR < 0.05.

GO/KEGG function	FDR	*p*-value
GO:0048731 system development	5.02E-05	3.18E-09
GO:0030154 cell differentiation	9.78E-05	1.88E-08
GO:0032502 developmental process	9.78E-05	2.13E-08
GO:0048869 cellular developmental process	9.78E-05	2.48E-08
GO:0007275 multicellular organism development	0.0001	4.69E-08
GO:0048856 anatomical structure development	0.0001	4.33E-08
GO:0048513 animal organ development	0.0002	1.06E-07
GO:0009653 anatomical structure morphogenesis	0.0003	1.98E-07
GO:0032501 multicellular organismal process	0.0003	1.92E-07
GO:0007399 nervous system development	0.0004	2.52E-07
GO:0048518 positive regulation of biological process	0.0005	3.44E-07
GO:0030182 neuron differentiation	0.0009	7.14E-07
GO:0048699 generation of neurons	0.0010	7.99E-07
GO:0022008 neurogenesis	0.0011	9.80E-07
GO:0051239 regulation of multicellular organismal process	0.0028	2.61E-06
GO:0048468 cell development	0.0050	5.02E-06
GO:0009887 animal organ morphogenesis	0.0054	5.86E-06
GO:0048598 embryonic morphogenesis	0.0066	7.53E-06
GO:0000904 cell morphogenesis involved in differentiation	0.0084	1.01E-05
GO:0050793 regulation of developmental process	0.0088	1.11E-05
GO:0001501 skeletal system development	0.0094	1.25E-05
GO:0051240 positive regulation of multicellular organismal process	0.0108	1.51E-05
GO:0048534 hematopoietic or lymphoid organ development	0.0117	1.70E-05
GO:0002520 immune system development	0.0124	1.95E-05
GO:0035295 tube development	0.0124	1.96E-05
GO:0000902 cell morphogenesis	0.0129	2.13E-05
GO:0048522 positive regulation of cellular process	0.0160	2.73E-05
GO:0009790 embryo development	0.0224	3.97E-05
GO:0009888 tissue development	0.0253	4.64E-05
GO:0007187 G-protein coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	0.0352	6.91E-05
GO:0032989 cellular component morphogenesis	0.0352	6.92E-05
GO:0032736 positive regulation of interleukin-13 production	0.0356	7.21E-05
GO:0048871 multicellular organismal homeostasis	0.0418	8.73E-05
GO:0030097 hemopoiesis	0.0459	9.88E-05
GO:0046703 natural killer cell lectin-like receptor binding	0.0481	1.04E-05

to top ranked probes can be confirmed to have differential methylation patterns and corresponding contributions to A-IDH and O-IDH cases, validating the reliability of our findings.

GO and KEGG Enrichment Associated With Glioma Subclasses

The SVM with top 750 features yielded the best performance. These 750 features (methylation probes) were mapped onto genes, on which a GO and KEGG enrichment analysis was performed. **Table 4** lists the significantly enriched GO/KEGG functions with FDR < 0.05. This section analyzed some of them.

Cellular development with hypergeometric test *p*-value of 2.48E-8 and FDR of 9.78E-5, is an important biological function that can be a marker to classify different glioma subclasses. The tyrosine kinase Fyn is an Src kinase family member essential for normal myelination and implicated in oligodendrocyte development (Ma et al., 2005). Fyn regulates oligodendroglial cell development in oligodendroglioma, considering that the neurogenesis of an adult brain is generally regulated by glial cells.

Neuron differentiation with hypergeometric test *p*-value of 7.14E-8 and FDR of 0.0009, can be another marker for classifying different glioma subclasses. The suppression of NSC (neural stem cells) differentiation and the promotion of its self-renewal capacity are controlled by the upregulation of PLAGL2. The inhibition of Wnt signaling partially restores *the differentiation* capacity of PLAGL2-expressing *NSC* (Zheng et al., 2010). These functions are consistent with a well-known hallmark of glioblastoma, e.g., strong self-renewal potential and immature differentiation state.

Cellular component morphogenesis with hypergeometric test *p*-value of 6.92E-5 and FDR of 0.0352, varies in different types of gliomas. Tumor cell metastasis mediated by abnormal extracellular matrix (ECM) regulations contributes to the rapid progression of GBM. As such, ECM may play an irreplaceable role during the invasion of GBM (Ulrich et al., 2009). Thus, cellular component morphogenesis may be a functional signature for characterizing different subtypes of gliomas.

G-protein-coupled receptor signaling pathway with hypergeometric test *p*-value of 6.91E-5 and FDR of 0.0352, coupled to a cyclic nucleotide second messenger, is an important pathway related to GBM. This pathway regulates glioma cells by interfering with calcium signaling processes. Its components, namely, P2Y1 and P2Y2 receptors, coexist in glioma C6 cells as an effective molecular identity of P2Y receptors (Ulrich et al., 2009). In terms of the specific role of this pathway in malignant diseases, Rho GTPase activation and angiogenesis are two typical pathological processes of the identified pathway to trigger tumorigenesis. Therefore, our enriched pathway may be effective and significant for the identification of different glioma subtypes (O'hayre et al., 2014).

The qualitatively analyzed genes help distinguish different glioma subclasses, and all the identified genes are supported by recent literature and related independent expression profiles. The functional enrichment of these genes further validates the differential functional characteristics of gliomas. Therefore, our new analysis method can help determine (methylation) signatures for glioma subclasses and establish a basis for further studying the detailed pathological mechanisms of these glioma subtypes at multiple omics levels.

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=GSE90496, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109379.

AUTHOR CONTRIBUTIONS

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

FUNDING

This study was supported by Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), and Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2019.00339/full#supplementary-material

Table S1 | RI scores of all input features ranked by MCFS.

Table S2 | Ten-fold cross-validation performance of IFS with SVM, RF, and RIPPER that changed with the number of features.

REFERENCES

Aibara, N., Ohyama, K., Hidaka, M., Kishikawa, N., Miyata, Y., Takatsuki, M., et al. (2018). Immune complexome analysis of antigens in circulating immune complexes from patients with acute cellular rejection after living donor liver transplantation. *Transpl. Immunol.* 48, 60–64. doi: 10.1016/j.trim.2018.02.011

Angerstein, C., Hecker, M., Paap, B. K., Koczan, D., Thamilarasan, M., Thiesen, H. J., et al. (2012). Integration of MicroRNA databases to study MicroRNAs associated with multiple sclerosis. *Mol. Neurobiol.* 45, 520–535. doi: 10.1007/s12035-012-8270-0

Aschebrook-Kilfoy, B., Argos, M., Pierce, B. L., Tong, L., Jasmine, F., Roy, S., et al. (2015). Genome-wide association study of parity in Bangladeshi women. *PLoS ONE* 10:e0118488. doi: 10.1371/journal.pone.0118488

Bergboer, J. G., Tjabringa, G. S., Kamsteeg, M., Van Vlijmen-Willems, I. M., Rodijk-Olthuis, D., Jansen, P. A., et al. (2011). Psoriasis risk genes of the late cornified envelope-3 group are distinctly expressed compared with genes of other LCE groups. *Am. J. Pathol.* 178, 1470–1477. doi: 10.1016/j.ajpath.2010.12.017

Cancer Genome Atlas Research Network, Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., et al. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N. Engl. J. Med.* 372, 2481–2498. doi: 10.1056/NEJMoa1402121

Capper, D., Jones, D. T. W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* 555, 469–474. doi: 10.1038/nature26000

Chen, L., Chu, C., Zhang, Y.-H., Zheng, M.-Y., Zhu, L., Kong, X., et al. (2017a). Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* 12, 526–534. doi: 10.2174/1574893611666160618094219

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

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

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

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

Chen, L., Zhang, S., Pan, X., Hu, X., Zhang, Y. H., Yuan, F., et al. (2019b). HIV infection alters the human epigenetic landscape. *Gene Ther.* 26, 29–39. doi: 10.1038/s41434-018-0051-6

Chen, L., Zhang, Y.-H., Lu, G., Huang, T., and Cai, Y.-D. (2017c). Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. *Artif. Intell. Med.* 76, 27–36. doi: 10.1016/j.artmed.2017.02.001

Cohen, W. W. (1995). "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning* (Tahoe City, CA), 115–123. doi: 10.1016/B978-1-55860-377-6.50023-2

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cui, H., and Chen, L. (2019). A binary classifier for the prediction of EC numbers of enzymes. *Curr. Proteomics* 16, 381–389. doi: 10.2174/1570164616666190126103036

Dalvai, M., and Bystricky, K. (2010). Cell cycle and anti-estrogen effects synergize to regulate cell proliferation and ER target gene expression. *PLoS ONE* 5:e11011. doi: 10.1371/journal.pone.0011011

Delpu, Y., Cordelier, P., Cho, W. C., and Torrisani, J. (2013). DNA methylation and cancer diagnosis. *Int. J. Mol. Sci.* 14, 15029–15058. doi: 10.3390/ijms140715029

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

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

Hilgers, R. H., and Webb, R. C. (2005). Molecular aspects of arterial smooth muscle contraction: focus on Rho. *Exp. Biol. Med.* 230, 829–835. doi: 10.1177/153537020523001107

Ho, T. K. (1995). "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition* (Montreal, QC).

Kloth, J. N., Fleuren, G. J., Oosting, J., De Menezes, R. X., Eilers, P. H., Kenter, G. G., et al. (2005). Substantial changes in gene expression of Wnt, MAPK and TNFalpha pathways induced by TGF-beta1 in cervical cancer cell lines. *Carcinogenesis* 26, 1493–1502. doi: 10.1093/carcin/bgi110

Kohavi, R. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence* (Montreal, QC: Lawrence Erlbaum Associates Ltd.), 1137–1145.

Kondo, Y., Katsushima, K., Ohka, F., Natsume, A., and Shinjo, K. (2014). Epigenetic dysregulation in glioma. *Cancer Sci.* 105, 363–369. doi: 10.1111/cas.12379

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

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

Liu, H. A., and Setiono, R. (1998). Incremental feature selection. *Appl. Intell.* 9, 217–230. doi: 10.1023/A:1008363719778

Lord, K. A., Wang, X. M., Simmons, S. J., Bruckner, R. C., Loscig, J., O'connor, B., et al. (1997). Variant cDNA sequences of human ATP:citrate lyase: cloning, expression, and purification from baculovirus-infected insect cells. *Protein Expr. Purif.* 9, 133–141. doi: 10.1006/prep.1996.0668

Losman, J. A., and Kaelin, W. G. Jr. (2013). What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. *Genes Dev.* 27, 836–852. doi: 10.1101/gad.217406.113

Ma, D. K., Ming, G. L., and Song, H. (2005). Glial influences on neural stem cell development: cellular niches for adult neurogenesis. *Curr. Opin. Neurobiol.* 15, 514–520. doi: 10.1016/j.conb.2005.08.003

Mascelli, S., Barla, A., Raso, A., Mosci, S., Nozza, P., Biassoni, R., et al. (2013). Molecular fingerprinting reflects different histotypes and brain region in low grade gliomas. *BMC Cancer* 13:387. doi: 10.1186/1471-2407-13-387

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

Mckenzie, A. T., Moyon, S., Wang, M., Katsyv, I., Song, W. M., Zhou, X., et al. (2017). Multiscale network modeling of oligodendrocytes reveals molecular components of myelin dysregulation in Alzheimer's disease. *Mol. Neurodegener.* 12:82. doi: 10.1186/s13024-017-0219-3

Northcott, P. A., Nakahara, Y., Wu, X., Feuk, L., Ellison, D. W., Croul, S., et al. (2009). Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. *Nat. Genet.* 41, 465–472. doi: 10.1038/ng.336

O'hayre, M., Degese, M. S., and Gutkind, J. S. (2014). Novel insights into G protein and G protein-coupled receptor signaling in cancer. *Curr. Opin. Cell Biol.* 27, 126–135. doi: 10.1016/j.ceb.2014.01.005

Ohyama, K., Baba, M., Tamai, M., Yamamoto, M., Ichinose, K., Kishikawa, N., et al. (2016). Immune complexome analysis of antigens in circulating immune complexes isolated from patients with IgG4-related dacryoadenitis and/or sialadenitis. *Mod. Rheumatol.* 26, 248–250. doi: 10.3109/14397595.2015.1072296

Okada, K., Katagiri, T., Tsunoda, T., Mizutani, Y., Suzuki, Y., Kamada, M., et al. (2003). Analysis of gene-expression profiles in testicular seminomas using a genome-wide cDNA microarray. *Int. J. Oncol.* 23, 1615–1635. doi: 10.3892/ijo.23.6.1615

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

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

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

Richardson, T. E., Patel, S., Serrano, J., Sathe, A. A., Daoud, E. V., Oliver, D., et al. (2019). Genome-wide analysis of glioblastoma patients with unexpectedly long survival. *J. Neuropathol. Exp. Neurol.* 78, 501–507. doi: 10.1093/jnen/nlz025

Roth, P., and Weller, M. (2014). Challenges to targeting epidermal growth factor receptor in glioblastoma: escape mechanisms and combinatorial treatment strategies. *Neuro Oncol.* 16(Suppl. 8), viii14–viii19. doi: 10.1093/neuonc/nou222

Sato, T., Arai, E., Kohno, T., Tsuta, K., Watanabe, S., Soejima, K., et al. (2013). DNA methylation profiles at precancerous stages associated with recurrence of lung adenocarcinoma. *PLoS ONE* 8:e59444. doi: 10.1371/journal.pone.0059444

Thorne, A. H., Zanca, C., and Furnari, F. (2016). Epidermal growth factor receptor targeting and challenges in glioblastoma. *Neuro Oncol.* 18, 914–918. doi: 10.1093/neuonc/nov319

Ton, C., Stamatiou, D., Dzau, V. J., and Liew, C. C. (2002). Construction of a zebrafish cDNA microarray: gene expression profiling of the zebrafish during development. *Biochem. Biophys. Res. Commun.* 296, 1134–1142. doi: 10.1016/S0006-291X(02)02010-7

Trang, S. H., Joyner, D. E., Damron, T. A., Aboulafia, A. J., and Randall, R. L. (2010). Potential for functional redundancy in EGF and TGFalpha signaling in desmoid cells: a cDNA microarray analysis. *Growth Factors* 28, 10–23. doi: 10.3109/08977190903299387

Uhlmann, K., Rohde, K., Zeller, C., Szymas, J., Vogel, S., Marczinek, K., et al. (2003). Distinct methylation profiles of glioma subtypes. *Int. J. Cancer* 106, 52–59. doi: 10.1002/ijc.11175

Ulrich, T. A., De Juan Pardo, E. M., and Kumar, S. (2009). The mechanical rigidity of the extracellular matrix regulates the structure, motility, and proliferation of glioma cells. *Cancer Res.* 69, 4167–4174. doi: 10.1158/0008-5472.CAN-08-4859

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 355:eaai8478. doi: 10.1126/science.aai8478

Wada, A. M., Reese, D. E., and Bader, D. M. (2001). Bves: prototype of a new class of cell adhesion molecules expressed during coronary artery development. *Development* 128, 2085–2093.

Walker, R. H. (2016). The non-Huntington disease choreas: five new things. *Neurol. Clin. Pract.* 6, 150–156. doi: 10.1212/CPJ.0000000000000236

Wang, L. F., Fokas, E., Juricko, J., You, A., Rose, F., Pagenstecher, A., et al. (2008). Increased expression of EphA7 correlates with adverse outcome in primary and recurrent glioblastoma multiforme patients. *BMC Cancer* 8:79. doi: 10.1186/1471-2407-8-79

Wang, S., Zhang, Y. H., Zhang, N., Chen, L., Huang, T., and Cai, Y. D. (2017). Recognizing and predicting thioether bridges formed by lanthionine and beta-methyllanthionine in lantibiotics using a random forest approach with feature selection. *Comb. Chem. High Throughput Screen* 20, 582–593. doi: 10.2174/1386207320666170310115754

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

Xie, L., Liao, Y., Shen, L., Hu, F., Yu, S., Zhou, Y., et al. (2017). Identification of the miRNA-mRNA regulatory network of small cell osteosarcoma based on RNA-seq. *Oncotarget* 8, 42525–42536. doi: 10.18632/oncotarget.17208

Yan, K., Yang, K., and Rich, J. N. (2013). The evolving landscape of glioblastoma stem cells. *Curr. Opin. Neurol.* 26, 701–707. doi: 10.1097/WCO.0000000000000032

Zhang, M., Pan, Y., Qi, X., Liu, Y., Dong, R., Zheng, D., et al. (2018). Identification of new biomarkers associated with IDH mutation and prognosis in astrocytic tumors using nanostring ncounter analysis system. *Appl. Immunohistochem. Mol. Morphol.* 26, 101–107. doi: 10.1097/PAI.0000000000000396

Zhang, P. W., Chen, L., Huang, T., Zhang, N., Kong, X. Y., and Cai, Y. D. (2015). Classifying ten types of major cancers based on reverse phase protein array profiles. *PLoS ONE* 10:e0123147. doi: 10.1371/journal.pone.0123147

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

Zhang, X., Chen, L., Guo, Z.-H., and Liang, H. (2019). Identification of human membrane protein types by incorporating network embedding methods. *IEEE Access* 7, 140794–140805. doi: 10.1109/ACCESS.2019.2944177

Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* 14:1. doi: 10.2174/1574893614666190220114644

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

Zheng, H., Ying, H., Wiedemeyer, R., Yan, H., Quayle, S. N., Ivanova, E. V., et al. (2010). PLAGL2 regulates Wnt signaling to impede differentiation in neural stem cells and gliomas. *Cancer Cell* 17, 497–509. doi: 10.1016/j.ccr.2010.03.020

Zhou, J.-P., Chen, L., and Guo, Z.-H. (2019). iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical (ATC) classes of drugs. *Bioinformatics*. doi: 10.1093/bioinformatics/btz757. [Epub ahead of print].

Zhou, Y., Zhang, N., Li, B. Q., Huang, T., Cai, Y. D., and Kong, X. Y. (2015). A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. *J. Biomol. Struct. Dyn.* 33, 2479–2490. doi: 10.1080/07391102.2014.1001793

Zuccato, C., Tartari, M., Crotti, A., Goffredo, D., Valenza, M., Conti, L., et al. (2003). Huntingtin interacts with REST/NRSF to modulate the transcription of NRSE-controlled neuronal genes. *Nat. Genet.* 35, 76–83. doi: 10.1038/ng1219

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 January 2020
doi: 10.3389/fgene.2019.01323

An Integrated Model Based on a Six-Gene Signature Predicts Overall Survival in Patients With Hepatocellular Carcinoma

Wenli Li[1,2†], *Jianjun Lu*[3,4,5†], *Zhanzhong Ma*[1], *Jiafeng Zhao*[6] *and Jun Liu*[1,5*]

[1] Department of Clinical Laboratory, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China, [2] Department of Reproductive Medicine Center, The Affiliated Yue Bei People's Hospital of Shantou University Medical College, Shaoguan, China, [3] The Second School of Clinical Medicine, Southern Medical University, Guangzhou, China, [4] Department of Medical Services, First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [5] Morning Star Academic Cooperation, Shanghai, China, [6] Department of Hepatobiliary Surgery, Yue Bei People's Hospital, Shantou University Medical College, Shaoguan, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Juan Ye,
National Institutes of Health (NIH), United States
Jinyuan Ma,
Boston University, United States
Xinglei Liu,
Albert Einstein College of Medicine, United States

***Correspondence:**
Jun Liu
liuyu8566@126.com

[†]These authors share first authorship

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

Received: 30 September 2019
Accepted: 05 December 2019
Published: 14 January 2020

Citation:
Li W, Lu J, Ma Z, Zhao J and Liu J (2020) An Integrated Model Based on a Six-Gene Signature Predicts Overall Survival in Patients With Hepatocellular Carcinoma. Front. Genet. 10:1323. doi: 10.3389/fgene.2019.01323

Background: Nowadays, clinical treatment outcomes of patients with hepatocellular carcinoma (HCC) have been improved. However, due to the complexity of the molecular mechanisms, the recurrence rate and mortality in HCC inpatients are still at a high level. Therefore, there is an urgent need in screening biomarkers of HCC to show therapeutic effects and improve the prognosis.

Methods: In this study, we aim to establish a gene signature that can predict the prognosis of HCC patients by downloading and analyzing RNA sequencing data and clinical information from three independent public databases. Firstly, we applied the limma R package to analyze biomarkers by the genetic data and clinical information downloaded from the Gene Expression Omnibus database (GEO), and then used the least absolute shrinkage and selection operator (LASSO) Cox regression and survival analysis to establish a gene signature and a prediction model by data from the Cancer Genome Atlas (TCGA). Besides, messenger RNA (mRNA) and protein expressions of the six-gene signature were explored using Oncomine, Human Protein Atlas (HPA) and the International Cancer Genome Consortium (ICGC).

Results: A total of 8,306 differentially expressed genes (DEGs) were obtained between HCC (n = 115) and normal tissues (n = 52). Top 5,000 significant genes were selected and subjected to the weighted correlation network analysis (WGCNA), which constructed nine gene co-expression modules that assign these genes to different modules by cluster dendrogram trees. By analyzing the most significant module (red module), six genes (SQSTM1, AHSA1, VNN2, SMG5, SRXN1, and GLS) were screened by univariate, LASSO, and multivariate Cox regression analysis. By a survival analysis with the HCC data in TCGA, we established a nomogram based on the six-gene signature and multiple clinicopathological features. The six-gene signature was then validated as an independent prognostic factor in independent HCC cohort from ICGC. Receiver operating

characteristic (ROC) curve analysis confirmed the predictive capacity of the six-gene signature and nomogram. Besides, overexpression of the six genes at the mRNA and protein levels was validated using Oncomine and HPA, respectively.

Conclusion: The predictive six-gene signature and nomograms established in this study can assist clinicians in selecting personalized treatment for patients with HCC.

Keywords: hepatocellular carcinoma, overall survival, risk score, mRNA signature, weighted gene co-expression network analysis

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most common malignancies worldwide. The mortality rate of HCC ranks second among all cancers, and HCC has a higher rate in developing countries compared to developed countries (El-Serag and Rudolph, 2007). Approximately 70% of HCC relapse within 5 years after receiving resection or ablation (Cancer Genome Atlas Research Network Electronic Address and Cancer Genome Atlas Research, 2017). The main causes leading to the poor prognosis are tumor metastasis and postoperative recurrence (Budhu et al., 2006). Abnormal expression of messenger RNAs (mRNAs) plays critical roles in a variety of biological processes. Recent studies have documented that mRNAs can function as potential biomarkers in cancer prognosis (Wang et al., 2018). Therefore, there is an urgent need in screening biomarkers of HCC to show therapeutic effects, reduce mortality, and improve the prognosis. A routine prognostic assessment tool for HCC patients was clinical pathological staging. However, HCC is always with clinical heterogeneity. For example, the clinical heterogeneity caused by the simultaneous presence of two life-threatening diseases, cancer and cirrhosis, often affects the effect of routine prognosis assessment. In order to provide more clinically beneficial treatment strategies for high-risk populations, there is an urgent need to develop a new prognostic prediction model as a supplement to the prediction outcomes of clinical staging.

During the last decades, gene sequencing and bioinformatic analysis have been widely used to screen genetic alterations at the genome level, which have helped us identify the differentially expressed genes (DEGs) and functional pathways involved in the progression of HCC. It was reported that epithelial cell adhesion molecule (Yamashita et al., 2008), CD24 (Woo et al., 2008), and TGF-beta (Coulouarn et al., 2008) were associated with the overall survival (OS) of HCC inpatients. However, false-positive rates in a single cohort analysis make it difficult to obtain reliable results. Thus, in the present study, we identify biomarkers of HCC by extracting a dataset of HCC patients from the Gene Expression Omnibus database (GEO). Then, we established a gene signature for HCC in Cancer Genome Atlas (TCGA) and established an integrated nomogram by combining multiple clinicopathological factors including the gene signature. Subsequently, the six-gene signature was verified in an independent external HCC cohort in International Union of Cancer Genome (ICGC). Besides, expression status of the six-gene signature in human HCC tissues at the mRNA and protein levels was explored using the Oncomine and the Human Protein Atlas (HPA) databases, respectively. In summary, we aim to establish a genetic marker and prognostic model that can predict the OS of HCC patients by bioinformatics methods. And this model could assist physicians to develop more individualized treatment plans.

MATERIALS AND METHODS

Data Source

The mRNA expression profile of HCC patients used to identify differentially expressed genes was derived from GEO, which was calculated on the Illumina HiSeq RNA sequencing (RNA-seq) platform and contained 115 HCC tissues and 52 adjacent non-tumor tissues (ANTTs) as of August 13, 2018 (GSE76427). The training dataset with HCC mRNA expression profiles and clinical information used to construct multi-gene signature was obtained from TCGA. The validation dataset with mRNA expression profile and clinical information used to verify the multi-gene signature was downloaded from ICGC. The above three databases are publicly available and open-access, and the present study followed the data access policy and publishing guidelines of these databases. Therefore, no local ethics committee is required to approve this study.

Identification of DEGs Between HCC and Non-Cancerous Tissues

Firstly, we obtained raw sequencing data for HCC mRNA including 41,718 mRNA expression profiles from the GEO database. Then, the DEG was calculated using the limma R package (Ritchie et al., 2015). DEGs with absolute log2 fold change (FC) > 1 and adjusted P value <0.05 were considered to be included for subsequent analysis.

Co-Expression Gene Network Based On RNA-Seq Data

The weighted correlation network analysis (WGCNA) was used to construct the gene co-expression network (Langfelder and Horvath, 2008). Firstly, to construct a gene expression similarity matrix, we calculate the absolute value of the Pearson's correlation coefficient between gene i and gene j:

$$S_{ij} = \left|(1 + \mathrm{cor}(x_i + y_j))/2\right|,$$

where i and j represent the amount of expression of the i and j genes, respectively. Then, the gene expression similarity matrix

was converted into an adjacency matrix, and the network type is signed. β is a soft threshold, which is actually the Pearson's correlation coefficient β of each pair of genes (Horvath et al., 2006). This step can strengthen strong correlation and weaken weak correlation from the index level:

$$a_{ij} = \left|(1 + \mathrm{cor}(x_i + y_j))/2\right|^{\beta}.$$

The next step was to convert the adjacency matrix into a topological matrix. The topological overlap measure (TOM) was used to describe the degree of association between genes:

$$\mathrm{TOM} = \left(\sum_{\mu \neq ij} \alpha_{i\mu}\alpha_{\mu j} + \alpha_{ij}\right) / \left(\min\left(\sum_{\mu}\alpha_{i\mu} + \sum_{\mu}\alpha_{j\mu}\right) + 1 - \alpha_{ij}\right).$$

TOM indicates the degree of dissimilarity between gene *i* and gene *j*. We conducted hierarchical clustering of genes using 1-TOM as a distance, and then used the method of dynamic cut tree for module identification. The most representative gene in each module was called the eigenvector gene, referred to as ME, which represents the overall level of gene expression within the module:

$$\mathrm{ME} = \mathrm{princomp}\left(x_{ij}^{q}\right),$$

where *i* represents the gene in module *q* and *j* represents the chip sample in module *q*. We use Pearson's correlation between the expression profile of a gene in all samples and the ME expression profile of an eigenvector gene to measure the identity of the gene in the module. We called it module membership (MM):

$$\mathrm{MM}_i^q = \mathrm{cor}(x_i, \mathrm{ME}^q)$$

where ME represents the expression profile of the *i* gene.

Functional Enrichment Analysis

Enrichment analysis of Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway for genes in the most significant modules of the WGCNA analysis was performed using the clusterProfiler R package (Yu et al., 2012).

Definition of the Gene-Related Prognostic Model

Univariate, the least absolute shrinkage and selection operator (LASSO), and multivariate Cox regression analyses were used to study the correlation between patient OS and gene expression levels (Tibshirani, 1997). Firstly, we used univariate Cox regression analysis to identify genes associated with OS, and then applied LASSO Cox regression to further narrow the range of HCC marker genes. After that, multiple Cox regression analysis was applied to assess whether marker genes could be an independent prognostic factor for patient survival. A multi-gene marker-based prognostic risk score was established based on a combination of regression coefficients from the multivariate Cox regression model (β) multiplied by their expression levels. Prognostic index (Pi) = (β * expression level of SQSTM1) + (β * expression level of AHSA1) + (β * expression level of VNN2) + (β * expression level of SMG5) + (β * expression level of SRXN1) + (β * expression level of GLS). Taking the median risk score as a cutoff value, 365 HCC patients from TCGA were divided into high- and low-risk groups. Kaplan–Meier (KM) survival curves and time-dependent receiver operational feature (ROC) curve analyses were made to assess the predictive capacity of the model. Decision curve analysis (DCA) curves were used to visually assess the clinical benefit of the model. Besides, the prognostic model was validated in an independent cohort from ICGC.

Prognostic Model Based on Six-Gene Signature as an Independent Predictor for OS

We used univariate and multivariate Cox regression analysis to assess whether the prognostic model could be independent of other clinicopathological variables (including age, gender, tissue registration, pathological stage, T staging, and risk score) for HCC patients. Clinical features were selected as an independent variable, and OS was selected as the dependent variable to calculate the hazard ratio (HR) and the 95% confidence interval, two-sided *P* value.

Validation of the Six-Gene Signature Using Multiple Databases

We used an online microarray database called Oncomine (http://www.oncomine.org) to analyze the mRNA expression of the gene signature between HCC tissues and normal liver tissues (Rhodes et al., 2004). The threshold settings were as follows: *P* value: 0.01; fold change: 2; gene rank: 10%. The datasets, sample size, fold change, *t* test, and *P* value were all derived from studies with statistical differences. In addition, immunohistochemical images were downloaded from publicly available human protein maps (http://www.proteinatlas.org) for comparison of protein expression levels related to the gene signature (Uhlen et al., 2010). We obtained an independent HCC cohort from ICGC, extracted the expression levels of six-gene signature, and compared the expression levels of six-gene signature between HCC and non-tumor tissues using Wilcoxon signed-rank test (two-sided *P* values, and $P < 0.05$ indicates significant statistical differences).

Establishment and Evaluation of the Nomograms for HCC Survival Prediction

Nomogram is an effective method for predicting the prognosis of cancer patients by simplifying the complex statistical prediction model into a profile chart for assessing the probability of OS in individual patients (Park, 2018). In this study, we included all independent clinical pathological prognostic factors selected from Cox regression analysis to construct a nomogram which can assess the OS probability of 1, 3, and 5 years in HCC patients. The prediction probability of the nomogram was compared with the observed actual probability by the calibration curve to verify the accuracy of the nomogram. Overlapping the reference line indicates that the model is accurate. ROC analysis was used to compare the prediction accuracy between the nomogram of combined model and the nomogram for each single clinical pathological prognostic factor.

RESULTS

Study Process and Summary of Patients' Information

Figure 1 is a flowchart for the entire work of this study. The detailed construction process of the OS prediction model for patients with HCC was shown in this chart. Patients' information in the GEO, TCGA, and ICGC cohorts was shown in **Table 1**.

Identification of DEGS with Prognosis Value in HCC

As shown in the volcano map (**Figure 2A**), a comparative analysis of mRNA expression profiles between HCC tissues (n = 115) and ANTTs (n = 52) identified 8,306 significantly differentially expressed mRNAs (logFC > 1 or logFC < −1, adjusted P < 0.05). Then, all DEGs were sorted in ascending order according to the adjusted P value, and the top 5,000 genes were selected and subjected to WGCNA, which constructed gene co-expression modules that assign these genes to different modules by cluster dendrogram trees (**Figure 2B**). Gene numbers of each module in WGCNA are shown in **Table 2**. The correlation coefficients between each co-expressed gene module and the clinical features of HCC are shown in **Figure 2C**, and the module membership vs. gene significance analysis of the nine HCC-related modules is shown in **Figure 2D**. **Figures 2C, D** show that the red module was not only with the largest correlation coefficient regarding to OS time (0.25) but also with the most significant module membership relevance to gene significance (module membership vs. gene significance: cor = 0.59, P = 1.2e−27). Thus, the red module was considered as the most important module related to the prognosis of HCC. And genes of the red module were extracted for GO and KEGG analysis. GO analysis that showed the most significant biological process (BP), molecular function (MF), and cellular component (CC) were I-kappa B kinase/NF-kappa B signaling, mitochondrial matrix, and cofactor binding, respectively (**Figure 2E**). And KEGG analysis showed the key pathways correlated with the HCC samples: carbon metabolism, fluid shear stress and atherosclerosis, biosynthesis of amino acids, arginine biosynthesis, and alanine–aspartate–glutamate metabolism (P_{adjust} < 0.05) (**Figure 2F**).

Constructing the Six-Gene Signature for Risk Scoring and Survival Prediction

A differential gene expression analysis was conducted (**Figure 2G**), and 61 key genes were selected for further analysis in TCGA. The entire process of extracting stable genes from the 61 prognostic-related genes in the HCC dataset from the TCGA to build a survival prediction model is presented in **Figure 3A**. To build a clinical survival prognostic model for HCC, we used TCGA as a training dataset and applied the LASSO Cox regression analysis to identify stable markers from 61 survival-related candidates. By forcing the sum of the absolute values of regression coefficients to be less than a fixed value, some coefficients were reduced to zero, and then we used relative regression coefficients to identify the most stable prognostic

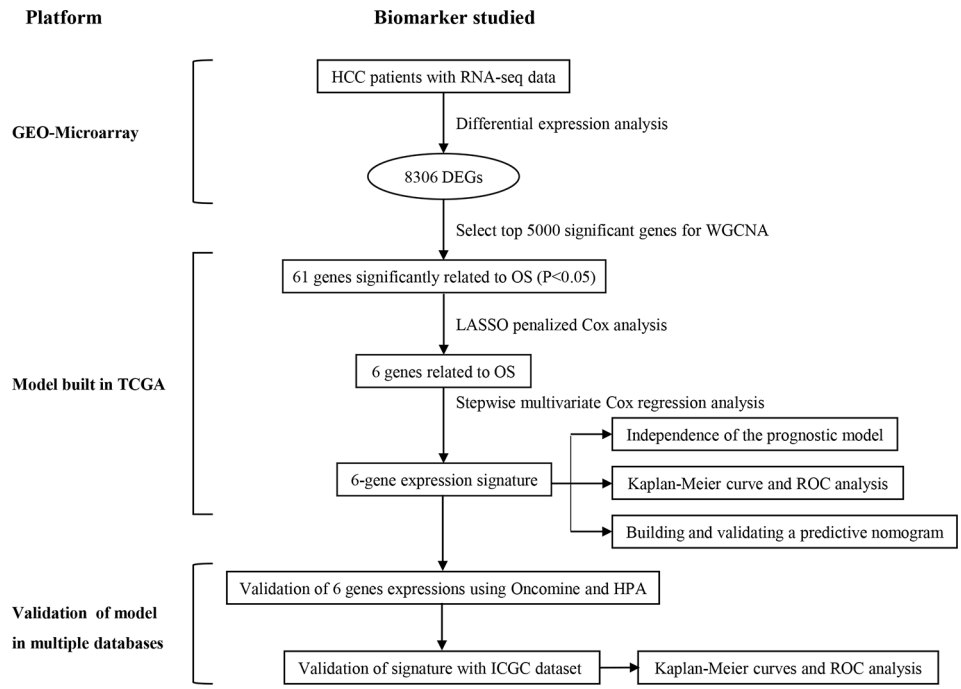


FIGURE 1 | Overall flowchart of this study.

TABLE 1 | Patients' information in the GEO, TCGA, and ICGC cohorts.

Clinical characteristics		Total	%
GSE 76427 in GEO		115	100
Survival status	Survival	92	80
	Death	23	20
Age	≤65 years	65	56.5
	>65 years	50	43.5
Gender	Female	22	19.1
	Male	93	80.9
Stage	I	55	47.8
	II	35	30.4
	III	21	18.3
	IV	3	2.5
TCGA		365	100
Survival status	Survival	239	65.48
	Death	126	34.52
Age	≤65 years	227	62.19
	>65 years	138	37.81
Gender	Male	246	67.40
	Female	119	32.60
Histological grade	G1	55	15.07
	G2	175	47.95
	G3	118	32.33
	G4	12	3.29
Stage	I	170	46.56
	II	84	23.01
	III	83	22.74
	IV	4	1.10
T classification	T1	180	49.32
	T2	91	24.93
	T3	78	21.37
	T4	13	3.56
ICGC		232	100
Survival status	Survival	189	81.47
	Death	43	18.53
Age	≤65 years	90	38.79
	>65 years	142	61.21
Gender	Male	171	73.71
	Female	61	26.29
Stage	I	36	15.52
	II	106	45.69
	III	71	30.60
	IV	19	8.19
Prior malignancy	No	202	87.07
	yes	30	12.93

markers and apply cross-validation to avoid overfitting of the LASSO Cox model. Parameters for building multivariate COX model are shown in **Table 3**, and six filter markers—SQSTM1, AHSA1, VNN2, SMG5, SRXN1, and GLS—are associated with high risk (HR > 1).

Then six genes were then applied to build a polygenic signature for prognostic prediction based on the minimum criteria. Subsequently, the risk score of each HCC patient from the training set was calculated using the coefficients obtained from the LASSO algorithm. To test the relationship between six identified genes and the prognosis of HCC patients, we constructed a prognostic model based on six-gene signature. Then, 365 HCC patients with follow-up information were divided into low-risk group and high-risk group according to the median value of risk scores among all HCC patients in the training set. Comparing the survival status and the six-gene expressions of the two groups, we found that the high-risk group was with poor prognosis and with higher expression of the six identified genes (**Figure 3B**).

Next, we proved our findings in the training set by validating the prognostic prediction function of the six-gene signature in an independent dataset from ICGC. We extracted microarray data from 243 HCC patients with follow-up information from the validation set and then calculated the risk score for each patient by using the same formula in the training set. Taking the median risk score as a cutoff value, the HCC patients in the validation set were divided into high- (n = 122) and low-risk (n = 121) groups, and the survival status and six-gene expressions were compared between the two groups. A similar result to the training set was obtained: the high-risk group was with poor prognosis and with higher six-gene expression level than the low-risk group (**Figure 3C**).

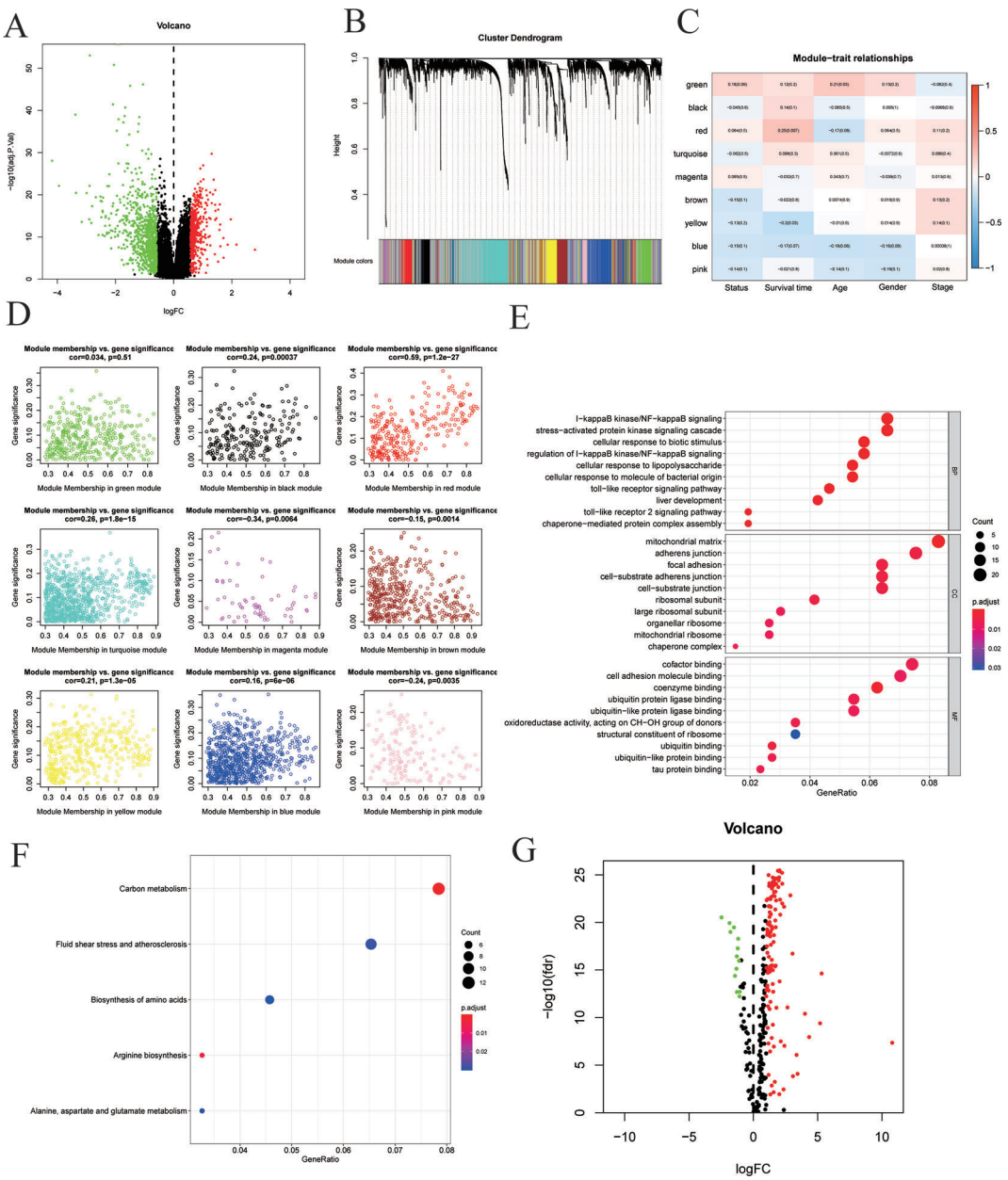


FIGURE 2 | Identification of prognostic genes in hepatocellular carcinoma patients. **(A)** Volcano plot showing differentially expressed genes (DEGs) in hepatocellular carcinoma samples. **(B)** Clustering dendrogram of genome-wide genes in hepatocellular carcinoma samples. **(C)** Correlation between modules and traits. Absolute values of correlation coefficients between hepatocellular carcinoma status and modules greater than 0.15 were considered as hepatocellular carcinoma-related modules. **(D)** Module membership in nine hepatocellular carcinoma-related modules. The *red* module was the most significant module. **(E, F)** GO and KEGG analysis revealed the most significant biological process (BP), molecular function (MF), cellular component (CC), and pathways correlated with the high-risk group genes in the red module. **(G)** Volcano plot revealed DEGs in the red module.

Kaplan–Meier and Time-Dependent ROC Curves of Six-Gene Signature

The Kaplan–Meier survival curve was applied to present a comparison of the OS of the two groups divided by the median risk score. Besides, the area under the ROC curve (AUC) of the time-dependent ROC curve was used to assess the prognostic ability of the six-gene signature, and a higher AUC means the better the model performance. We found that there was a significant difference on OS between the high- and low- risk groups in the TCGA dataset ($P < 0.0001$) (**Figure 4A**). The AUCs of the six-gene signature corresponding to 0.5, 1, 2, 3, and 5 years of survival were 0.759, 0.761, 0.708, 0.681, and 0.692,

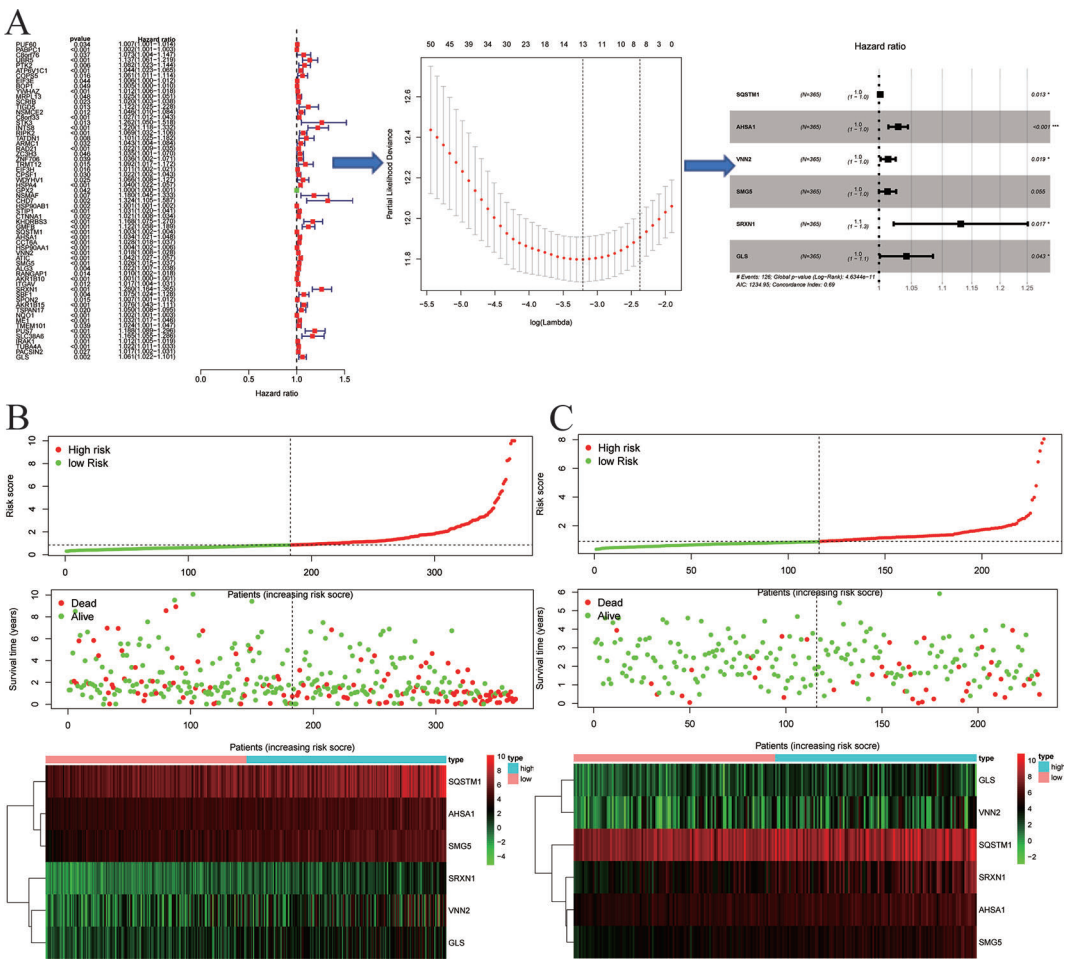


FIGURE 3 | Signature-based risk score is a promising marker in the training and validation cohorts. **(A)** The process of building the signature containing six genes most correlated with overall survival (OS) in the training set. The hazard ratios (HRs), 95% confidence intervals (CIs) calculated by univariate Cox regression, and the coefficients calculated by multivariate Cox regression using LASSO are shown. **(B, C)** Risk score distribution, survival overview, and heatmap for patients in the TCGA **(B)** and ICGC **(C)** datasets assigned to high- and low-risk groups based on the risk score.

TABLE 2 | Gene numbers of each module in WGCNA.

Module	Number
Black	216
Blue	792
Brown	449
Green	379
Gray	1,345
Magenta	63
Pink	146
Red	280
Turquoise	907
Yellow	423

respectively, suggesting that the prediction model had high sensitivity and specificity (**Figure 4C**). As shown in the other Kaplan–Meier curve (**Figure 4B**), the OS was significantly increased in the low-risk group compared to the high-risk group in the independent validation dataset from the ICGC dataset ($P < 0.001$). This result was consistent with our previous findings in the training cohort in TCGA dataset. As shown in **Figure 4D**, the AUCs of the six-gene signature model corresponding to 0.5, 1, 2, 3, and 5 years of survival were 0.637, 0.681, 0.690, 0.700, and 0.684, respectively, further confirming that the six-gene signature had high sensitivity and specificity and can be used as a reliable predictor of OS in HCC patients.

Prognostic Risk Scores were an Independent Prognostic Factor from the Other Clinicopathological Features

As shown in **Figure 5**, the risk score can be used as an independent factor in predicting OS. Univariate and multivariate Cox regression analyses were applied to assess independent predictive values for the six-gene signature in HCC patients. In the TCGA dataset, univariate Cox regression suggested that risk scores, pathological staging, and T staging

TABLE 3 | Parameters for building multivariate COX model.

Gene	Co-ef	HR	HR.95%L	HR.95%H	*P* value
SQSTM1	0.001803	1.001805	1.000376	1.003236	0.013287
AHSA1	0.028803	1.029222	1.014274	1.044391	0.000114
VNN2	0.013683	1.013777	1.002275	1.025411	0.018759
SMG5	0.012785	1.012867	0.999704	1.026203	0.055423
SRXN1	0.122996	1.13088	1.022251	1.251051	0.016983
GLS	0.041246	1.042108	1.001332	1.084545	0.042835

Co-ef : co-efficient, HR: hazard ratio.

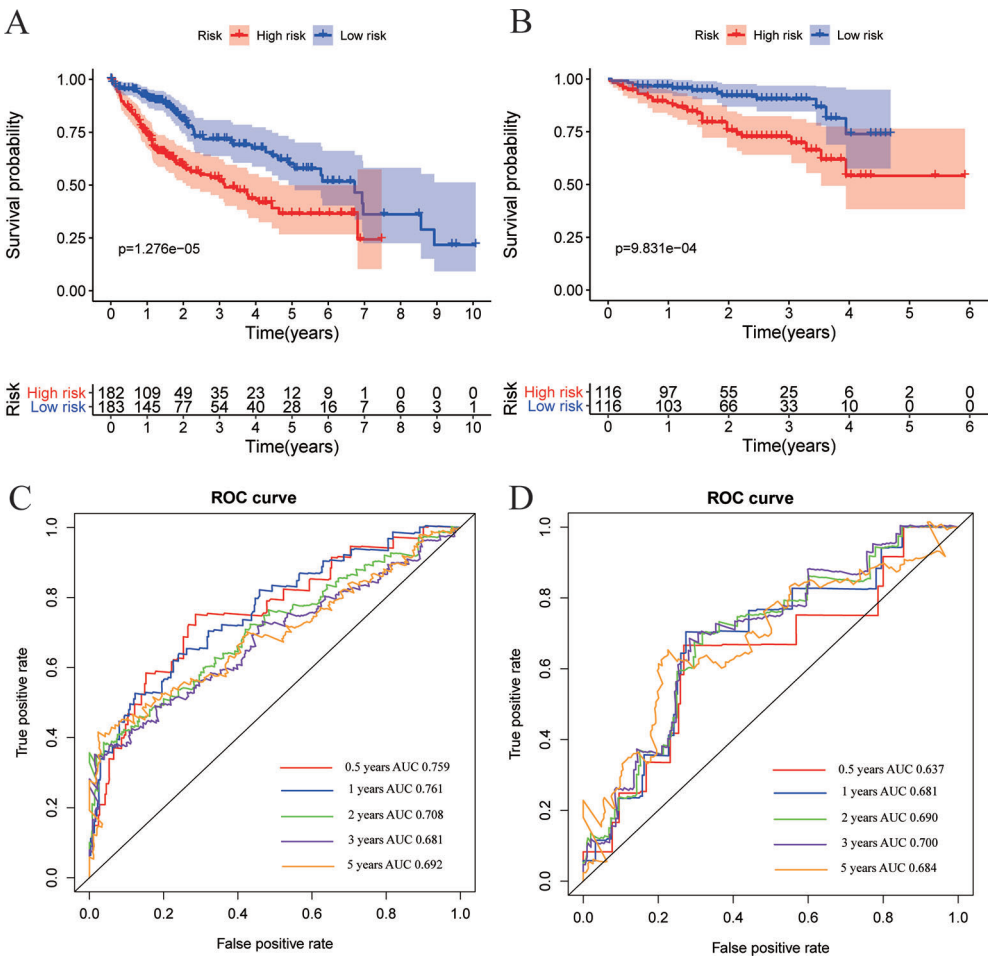


FIGURE 4 | Expression and survival analysis in training and validation datasets. **(A, B)** Kaplan–Meier overall survival (OS) curves for patients in the TCGA **(A)** and ICGC **(B)** datasets assigned to high- and low-risk groups based on the risk score. Patients with a high risk score exhibited poorer OS in the training and validation cohorts. **(C, D)** ROC curves showed the predictive efficiency of the risk signature for patients in the TCGA **(C)** and ICGC **(D)** datasets on the survival rate.

had a prognostic value, while age, gender, and histological grades were not associated with survival (**Figure 5A**). Then, multivariate Cox regression analysis suggested that only risk score was an independent prognostic factor associated with OS (**Figure 5C**). Next, we again used univariate and multivariate Cox regression analysis to validate whether the risk score can be used as an independent prognostic indicator in an independent HCC cohort from ICGC. Univariate Cox analysis suggested that risk score and pathological stage were associated with OS ($P < 0.05$; **Figure 5B**). Multivariate Cox regression analysis showed that risk scores, prior malignancy, and pathological stage were associated with OS ($P < 0.05$; **Figure 5D**). These results confirmed that risk scores based on six-gene signature can be used as an independent predictor of prognosis in HCC patients. Shown in the heat map are the expression levels of the six-gene signature in low-risk

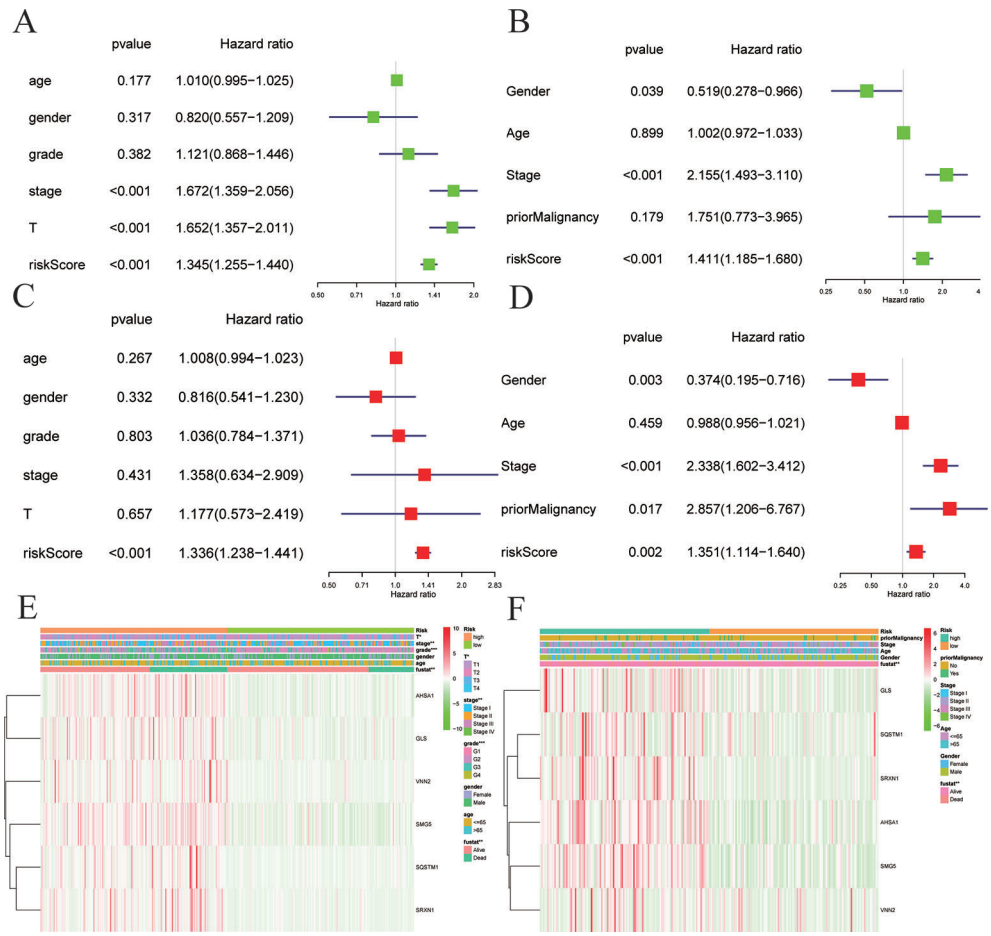


FIGURE 5 | Cox regression analyses of the association between clinicopathological factors and OS. **(A–F)** Univariate/multivariate Cox regression analyses and heatmaps of the association between clinicopathological factors (including the risk score) and overall survival (OS) of patients in the TCGA **(A**, **C**, **E)** and ICGC **(B**, **D**, **F)** datasets.

and high-risk HCC patients and the distribution of clinicopathological features between the low-risk and high-risk groups. It is suggested that there were significant differences in six-gene signature expression and OS between the high-risk group and the low-risk group both in the TCGA (**Figure 5E**) and ICGC (**Figure 5F**) datasets.

Subgroup Analysis of OS Based on Multiple Classification Methods

As shown in **Figure 6**, the survival analysis was conducted after risk score grouping based on six-gene signature expressions. We explored the expression profiles of six genes in different TNM stages, histological grades, viral hepatitis infection, BMI, and age in TCGA. Risk score based on six-gene signature was proven to be a potential marker for predicting OS in different subgroups, including stages I–II of TNM (P = 0.012), stages III–IV (P < 0.0001), G1 and G2 (P = 0.009), G3 and G4 (P < 0.0001), viral infection (P < 0.0001), BMI < 14 (P = 0.005), BMI(14–25) (P = 0.003), BMI > 25 (P = 0.047), age < 65 (P = 0.003), and age > 65 (P < 0.001).

Validation of the Six mRNA Expressions

In the TCGA HCC cohort, all six genes were highly expressed in HCC compared to that in adjacent non-tumor liver tissues. Next, we aimed to further confirm the expression patterns of these six genes in HCC tissues in the Oncomine database. Consistent with our results in TCGA, the average expression levels of SQSTM1, AHSA1, VNN2, SMG5, SRXN1, and GLS in HCC tissues were significantly higher than those in normal liver tissues (**Figures 7A–F**). To determine the clinical relevance of the six genes' expression, we analyzed the expression of the proteins encoded by these six genes using clinical specimens from the HPA. Relative to its expression level in normal liver tissue, SQSTM1 was strongly positive, while AHSA1 and GLS were moderately positive in HCC tissues (**Figures 7G–L**). However, VNN2, SMG5, and SRXN1 were not found on the website.

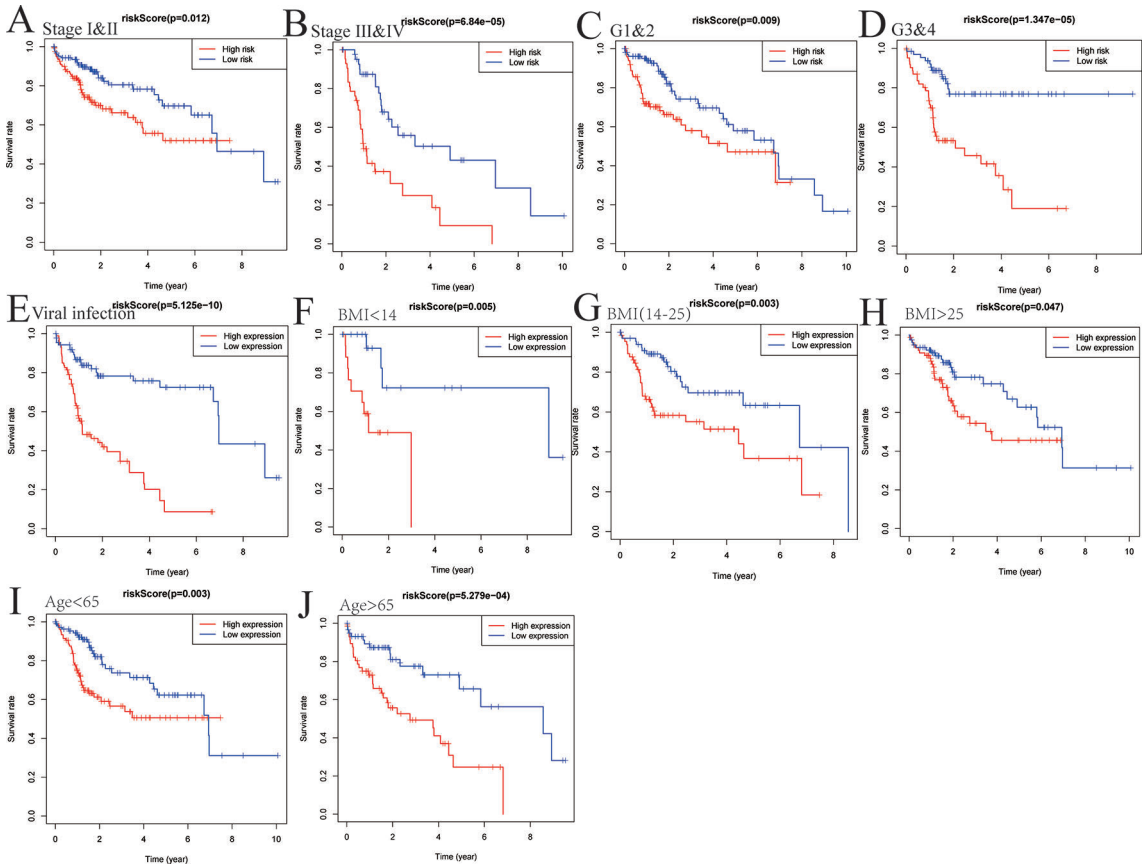


FIGURE 6 | The six-gene-based risk score is a promising marker for overall survival (OS) in subgroups. Subgroup analysis of OS based on pathological staging **(A**, **B)**, grading **(C**, **D)**, viral hepatitis **(E)**, BMI **(F**–**H)**, and age **(I**, **J)** of hepatocellular carcinoma (HCC) patients.

Building a Nomogram to Predict OS in HCC Patients

To establish a clinically applicable method for predicting the survival probability of patients with HCC, we developed a nomogram to predict the probability of the 1-, 3-, and 5-year OS in the TCGA cohort. The predictors of the nomogram included four independent prognostic factors (age, gender, pathologic stage, and six-gene signature). Subsequently, we constructed a nomogram that integrates clinical pathology features with six-gene signature to predict survival probabilities in HCC patients (**Figure 8A**). By calibration curve analysis, we found that the 1-, 3-, and 5-year survival probabilities predicted by the nomogram were closely related to the observed survival probability, which confirmed the reliability of the nomogram (**Figure 8B**).

Assessing the Accuracy of the Nomograms by ROC Curves

Time-dependent ROC curve analysis was used to evaluate the prediction accuracy of the integrated nomogram. The solid red line represents the integrated nomogram. In **Figures 9A–C**, the AUC of the integrated nomogram is the largest. Besides, all of AUCs of the integrated nomogram in **Figure 9** were above 0.77, suggesting that nomograms constructed by integrated factors are the best way to predict survival in HCC patients both for short-term and long-term survival compared to models constructed by a single prognostic factor. However, we also found that integrated predictions of the 3- and 5-year AUC of the integrated model are lower than that of 1 year, suggesting that the short-term prediction ability of the nomogram may be stronger than the long-term prediction ability. Besides, as shown in **Figures 9D–F**, the net benefits as calculated are plotted against the threshold probabilities of patients having 1-, 3-, and 5-year survival, and the results suggest that the net benefits of the integrated model were better than other models.

DISCUSSION

Due to the complex molecular mechanisms, HCC remains one of the most life-threatening malignancies in the world. Therefore, prognostic biomarkers are urgently needed to predict the

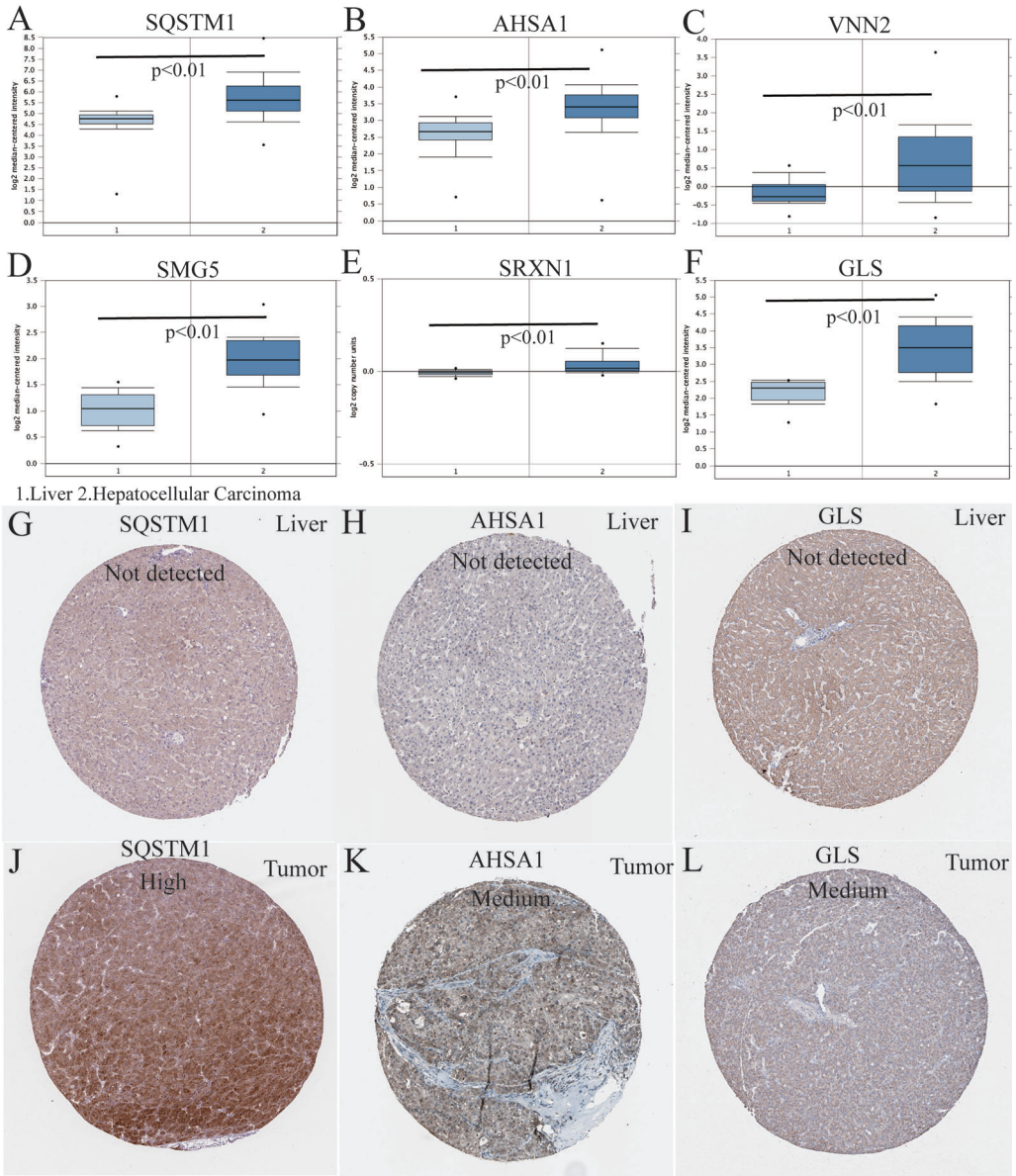


FIGURE 7 | Differences in protein expression induced by six genes were verified in human tissue samples. **(A–F)** The mRNA expression levels of the six-gene signature in human cancers (conducted in Oncomine database). **(G–L)** Human Protein Atlas immunohistochemistry using anti-SQSTM1, anti-AHSA1, and anti-GLS antibodies. Normal liver **(G–I)** vs. tumor tissues **(J–L)**.

outcome and to outline an individualized treatment plan for HCC patients. With the development of gene sequencing technology, some potential gene markers with predictive value for HCC patients have been identified. However, the number of such markers is still limited. In order to improve the prognosis of HCC, it is urgent to screen out more biomarkers with higher prediction accuracy in predicting prognosis.

In the present study, we identified potential gene biomarkers by analyzing the gene expression profiles of a HCC cohort in GEO. The DEGs between HCC samples and ANTTs were identified. Then, univariate, LASSO, and multivariate Cox analysis were used to further narrow the marker range and establish a risk model for predicting HCC prognosis. Our study found that high expression levels of six genes, including SQSTM1, AHSA1, VNN2, SMG5, SRXN1, and GLS, were associated with poor prognosis in HCC patients. We evaluated the model performance using the ROC curve of the six-gene signature. The results showed that the AUCs of the ROC curves for 0.5-, 1-, 2-, 3-, and 5-year survival prediction models were 0.637, 0.681, 0.690, 0.700, and 0.684, respectively, suggesting the

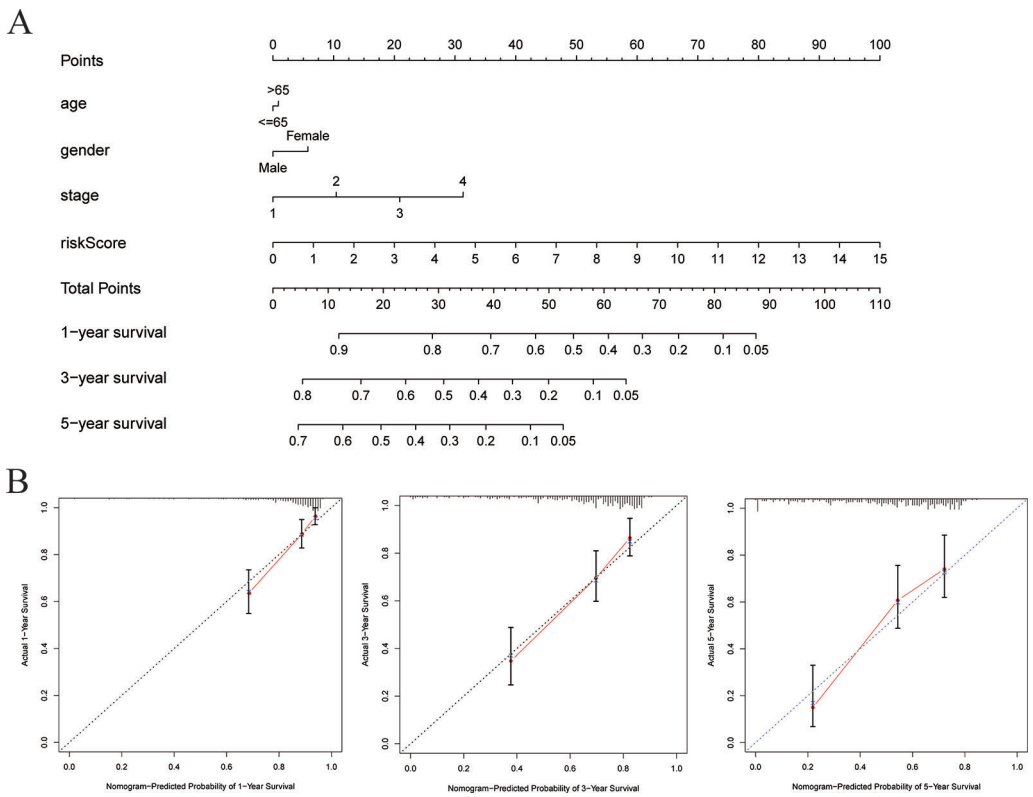


FIGURE 8 | Construction of a nomogram for survival prediction. **(A)** Nomogram combining signature with clinicopathological features. **(B)** Calibration plot showing that nomogram-predicted survival probabilities corresponded closely to the actual observed proportions.

six-gene signature was with good survival prediction performance. Then, we not only demonstrated that the six-gene signature was an independent prognostic factor for HCC patients superior to traditional clinicopathological factors but also verified their survival prediction ability in an external HCC cohort in ICGC. Thus, we believe that dividing HCC patients into high-risk group and low-risk group by the six-gene-based risk scoring model can be used for early prevention or detection of HCC recurrence in high-risk population.

Nomograms are a tool commonly used for tumor disease assessment to provide probabilistic predictions for individual patients. In our study, we constructed a nomogram that can predict the OS in HCC patients. The calibration curve indicates that the survival rate predicted by the nomogram is basically consistent with the actual observed survival rate in the dataset, indicating that the nomogram had good predictive performance. At the same time, we also proved that the use of the nomogram constructed by the combined model has better predictive performance than the nomogram constructed by a single HCC risk factor.

There were six genes identified for constructing the predictive model in this study. SQSTM1 is primarily involved in TNF signaling and the innate immune system. AHSA1 is primarily involved in ATPase activator activity. VNN2 is primarily involved in hydrolase activity. SMG5 is primarily involved in protein phosphatase 2A binding. And SRXN1 and GLS are involved in oxidoreductase activity and glutaminase activity, respectively. Combined with the results of GO and KEGG analysis, these perceptions suggested that abnormalities in energy metabolism and amino acid metabolism may play an important role in HCC.

HCC is a heterogeneous tumor that occurs through multiple pathway activations and molecular changes. Therefore, molecular heterogeneity affects the efficacy of prognostic evaluation by a single molecular marker. At the same time, some studies found that low survival rates of HCC were associated with strong cell proliferation and anti-apoptotic gene expression. These processes often involved multiple genes. And compared with single gene markers, multi-gene markers were always with more accurate prediction capacity for HCC (Lee et al., 2004). Bioinformatics methods were usually used to establish multi-gene signature for predicting the prognosis of HCC (Ye et al., 2003), and multi-gene signature is usually established by strategies including training, testing, and independent cross-validation (Roessler et al., 2010). Prediction capacity of a gene signature was significantly improved by the above strategies. It was reported that multi-gene signatures had a good predictive effect on venous metastasis (Budhu et al., 2006),

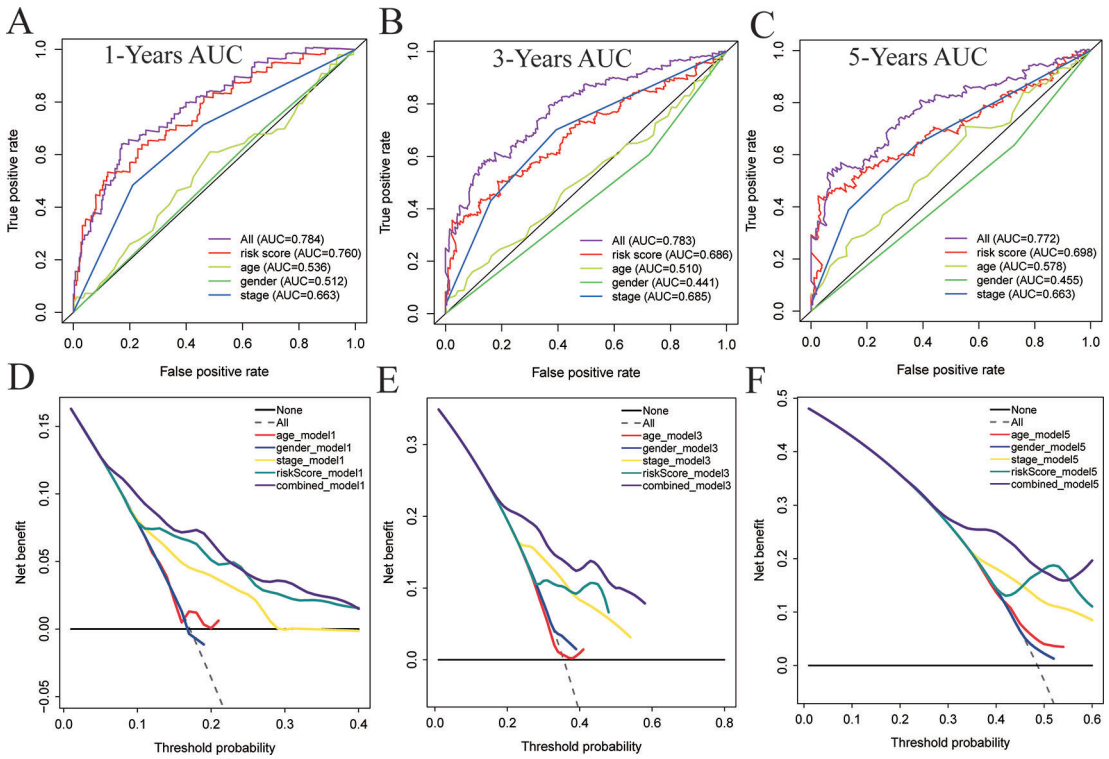


FIGURE 9 | The time-dependent receiver operating characteristic (ROC) and decision curve analysis (DCA) curves of the nomograms. Time-dependent ROC curve analysis evaluates the accuracy of the nomograms **(A–C)**. The *purple*, *red*, *yellow*, *green*, or *blue solid line* represents the nomogram. The DCA curves can intuitively evaluate the clinical benefit of the nomograms and the scope of application of the nomograms to obtain clinical benefits **(D–F)**. The net benefits (*Y*-axis) as calculated are plotted against the threshold probabilities of patients having 1, 3-, and 5-year survival on the *X*-axis. The *gray dotted line* represents the assumption that all patients have 1-, 3-, and 5-year survival. The *black solid line* represents the assumption that no patients have 1-, 3-, or 5-year survival. The *red*, *blue*, *yellow*, *green*, or *purple solid line* represents the nomograms.

progression (Roessler et al., 2012), recurrence (Kurokawa et al., 2004; Ho et al., 2006), and survival (Hoshida et al., 2013; Lim et al., 2013; Villa et al., 2016) for HCC. Initial multi-gene signatures often involved a large number of genes, and it affected the clinical application of the signature. It is believed that more user-friendly risk score models with a limited number of genes should be established to predict the prognosis of HCC patients (Kim et al., 2012). Recent study found a five-gene-based signature for HCC including HN1, RAN, AMP3, KRT19, and TAF9 (Nault et al., 2013). Now, it is believed that a combination model based on clinical, pathological, and gene signature will be more practical (Villanueva et al., 2011). At the same time, microRNAs (miRNAs) such as miR-517a (Toffanin et al., 2011), miR-125b (Li et al., 2008), and miR-26 (Ji et al., 2009) have been found to be associated with prognosis of HCC. There are also multiple gene markers based on multiple miRNAs and lncRNA (Budhu et al., 2008; Jiang et al., 2008).

In recent years, the identification of prognostic gene signature for HCC has been noted in many studies. For example, an eight-gene signature with a 5-year survival prediction AUC of 0.770 containing eight protein-coding genes (DCAF13, FAM163A, GPR18, LRP10, PVRIG, S100A9, SGCB, and TNNI3K) was established (Qiao et al., 2019). Subsequently, a six-gene signature (CSE1L, CSTB, MTHFR, DAGLA, MMP10, and GYS2) with a 5-year survival prediction AUC of 0.718 was established (Liu et al., 2019). In this study, we established a prognostic model with higher 5-year survival prediction AUC (0.772) based on a novel six-gene signature and further improved the predictive power of the HCC survival prediction model. To our knowledge, survival prediction models based on this six-gene signature have not been reported yet. Compared with traditional pathological staging and tissue grading, multi-gene signature of HCC has the advantages of higher prediction accuracy, more individualized test results, and reasonable sequencing costs. Therefore, six-gene signature has good prospects in clinical practice. In our study, we constructed and verified this six-gene signature by three independent datasets. More reasonable use of the biometric methods and mutual verification of multiple independent datasets make our study have more reliable results.

However, there were some limitations in this study. For example, the racial factors associated with sequencing samples and some potential prognostic factors may be not included in the model limited the predictive power of this model. In the future, we plan to use more rational bioinformatics strategies to improve

the model. In summary, our results suggest that the six-gene-based prognosis model is a reliable tool for predicting OS in patients with HCC, and the nomogram containing six-gene signature can help to develop personalized HCC treatments in clinical practice. The challenge in the future is how to apply various genes signature reasonably in a particular stage of HCC.

DATA AVAILABILITY STATEMENT

The datasets analyzed in the current study are available in the TCGA repository (http://cancergenome.nih.gov/), the ICGC (https://icgc.org/), and GEO (https://www. ncbi.nlm.nih.gov/geo/).

ETHICS STATEMENT

The usage of NIH controlled-access datasets was approved by the NCBI dbGaP.

AUTHOR CONTRIBUTIONS

JLiu designed and supervised the study and was a major contributor in editing the manuscript. WL and JLu analyzed and interpreted the data and were major contributors in writing the manuscript. ZM and JZ performed analysis and contributed to writing the manuscript. All authors read and approved the final manuscript.

REFERENCES

Budhu, A., Forgues, M., Ye, Q. H., Jia, H. L., He, P., Zanetti, K. A., et al. (2006). Prediction of venous metastases, recurrence, and prognosis in hepatocellular carcinoma based on a unique immune response signature of the liver microenvironment. *Cancer Cell* 10 (2), 99–111. doi: 10.1016/j.ccr.2006.06.016

Budhu, A., Jia, H. L., Forgues, M., Liu, C. G., Goldstein, D., Lam, A., et al. (2008). Identification of metastasis-related microRNAs in hepatocellular carcinoma. *Hepatology* 47 (3), 897–907. doi: 10.1002/hep.22160

Cancer Genome Atlas Research Network Electronic Address and Cancer Genome Atlas Research. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* 169 (7), 1327–1341, e1323. doi: 10.1016/j.cell.2017.05.046

Coulouarn, C., Factor, V. M., and Thorgeirsson, S. S. (2008). Transforming growth factor-beta gene expression signature in mouse hepatocytes predicts clinical outcome in human cancer. *Hepatology* 47 (6), 2059–2067. doi: 10.1002/hep.22283

El-Serag, H. B., and Rudolph, K. L. (2007). Hepatocellular carcinoma: epidemiology and molecular carcinogenesis. *Gastroenterology* 132 (7), 2557–2576. doi: 10.1053/j.gastro.2007.04.061

Ho, M. C., Lin, J. J., Chen, C. N., Chen, C. C., Lee, H., Yang, C. Y., et al. (2006). A gene expression profile for vascular invasion can predict the recurrence after resection of hepatocellular carcinoma: a microarray approach. *Ann. Surg. Oncol.* 13 (11), 1474–1484. doi: 10.1245/s10434-006-9057-1

Horvath, S., Zhang, B., Carlson, M., Lu, K. V., Zhu, S., Felciano, R. M., et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl. Acad. Sci. U. S. A.* 103 (46), 17402–17407. doi: 10.1073/pnas.0608396103

Hoshida, Y., Villanueva, A., Sangiovanni, A., Sole, M., Hur, C., Andersson, K. L., et al. (2013). Prognostic gene expression signature for patients with hepatitis C-related early-stage cirrhosis. *Gastroenterology* 144 (5), 1024–1030. doi: 10.1053/j.gastro.2013.01.021

Ji, J., Shi, J., Budhu, A., Yu, Z., Forgues, M., Roessler, S., et al. (2009). MicroRNA expression, survival, and response to interferon in liver cancer. *N Engl. J. Med.* 361 (15), 1437–1447. doi: 10.1056/NEJMoa0901282

Jiang, J., Gusev, Y., Aderca, I., Mettler, T. A., Nagorney, D. M., Brackett, D. J., et al. (2008). Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. *Clin. Cancer Res.* 14 (2), 419–427. doi: 10.1158/1078-0432.CCR-07-0523

Kim, S. M., Leem, S. H., Chu, I. S., Park, Y. Y., Kim, S. C., Kim, S. B., et al. (2012). Sixty-five gene-based risk score classifier predicts overall survival in hepatocellular carcinoma. *Hepatology* 55 (5), 1443–1452. doi: 10.1002/hep.24813

Kurokawa, Y., Matoba, R., Takemasa, I., Nagano, H., Dono, K., Nakamori, S., et al. (2004). Molecular-based prediction of early recurrence in hepatocellular carcinoma. *J. Hepatol.* 41 (2), 284–291. doi: 10.1016/j.jhep.2004.04.031

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559

Lee, J. S., Chu, I. S., Heo, J., Calvisi, D. F., Sun, Z., Roskams, T., et al. (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* 40 (3), 667–676. doi: 10.1002/hep.20375

Li, W., Xie, L., He, X., Li, J., Tu, K., Wei, L., et al. (2008). Diagnostic and prognostic implications of microRNAs in human hepatocellular carcinoma. *Int. J. Cancer* 123 (7), 1616–1622. doi: 10.1002/ijc.23693

Lim, H. Y., Sohn, I., Deng, S., Lee, J., Jung, S. H., Mao, M., et al. (2013). Prediction of disease-free survival in hepatocellular carcinoma by gene expression profiling. *Ann. Surg. Oncol.* 20 (12), 3747–3753. doi: 10.1245/s10434-013-3070-y

Liu, G. M., Zeng, H. D., Zhang, C. Y., and Xu, J. W. (2019). Identification of a six-gene signature predicting overall survival for hepatocellular carcinoma. *Cancer Cell Int.* 19, 138. doi: 10.1186/s12935-019-0858-2

Nault, J. C., De Reynies, A., Villanueva, A., Calderaro, J., Rebouissou, S., Couchy, G., et al. (2013). A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. *Gastroenterology* 145 (1), 176–187. doi: 10.1053/j.gastro.2013.03.051

Park, S. Y. (2018). Nomogram: An analogue tool to deliver digital knowledge. *J. Thorac. Cardiovasc. Surg.* 155 (4), 1793. doi: 10.1016/j.jtcvs.2017.12.107

Qiao, G. J., Chen, L., Wu, J. C., and Li, Z. R. (2019). Identification of an eight-gene signature for survival prediction for patients with hepatocellular carcinoma based on integrated bioinformatics analysis. *PeerJ* 7, e6548. doi: 10.7717/peerj.6548

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6 (1), 1–6. doi: 10.1016/s1476-5586(04)80047-2

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi: 10.1093/nar/gkv007

Roessler, S., Jia, H. L., Budhu, A., Forgues, M., Ye, Q. H., Lee, J. S., et al. (2010). A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res.* 70 (24), 10202–10212. doi: 10.1158/0008-5472.CAN-10-2607

Roessler, S., Long, E. L., Budhu, A., Chen, Y., Zhao, X., Ji, J., et al. (2012). Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. *Gastroenterology* 142957-966 (4), e912. doi: 10.1053/j.gastro.2011.12.039

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

Toffanin, S., Hoshida, Y., Lachenmayer, A., Villanueva, A., Cabellos, L., Minguez, B., et al. (2011). MicroRNA-based classification of hepatocellular carcinoma and oncogenic role of miR-517a. *Gastroenterology* 1401618-1628 (5), e1616. doi: 10.1053/j.gastro.2011.02.009

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., et al. (2010). Towards a knowledge-based human protein atlas. *Nat. Biotechnol.* 28 (12), 1248–1250. doi: 10.1038/nbt1210-1248

Villa, E., Critelli, R., Lei, B., Marzocchi, G., Camma, C., Giannelli, G., et al. (2016). Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. *Gut* 65 (5), 861–869. doi: 10.1136/gutjnl-2014-308483

Villanueva, A., Hoshida, Y., Battiston, C., Tovar, V., Sia, D., Alsinet, C., et al. (2011). Combining clinical, pathology, and gene expression data to predict recurrence of hepatocellular carcinoma. *Gastroenterology* 140 (5), 1501–1512 e1502. doi: 10.1053/j.gastro.2011.02.006

Wang, Z., Teng, D., Li, Y., Hu, Z., Liu, L., and Zheng, H. (2018). A six-gene-based prognostic signature for hepatocellular carcinoma overall survival prediction. *Life Sci.* 203, 83–91. doi: 10.1016/j.lfs.2018.04.025

Woo, H. G., Park, E. S., Cheon, J. H., Kim, J. H., Lee, J. S., Park, B. J., et al. (2008). Gene expression-based recurrence prediction of hepatitis B virus-related human hepatocellular carcinoma. *Clin. Cancer Res.* 14 (7), 2056–2064. doi: 10.1158/1078-0432.CCR-07-1473

Yamashita, T., Forgues, M., Wang, W., Kim, J. W., Ye, Q., Jia, H., et al. (2008). EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. *Cancer Res.* 68 (5), 1451–1461. doi: 10.1158/0008-5472.CAN-07-6013

Ye, Q. H., Qin, L. X., Forgues, M., He, P., Kim, J. W., Peng, A. C., et al. (2003). Predicting hepatitis B virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat. Med.* 9 (4), 416–423. doi: 10.1038/nm843

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16 (5), 284–287. doi: 10.1089/omi.2011.0118

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 March 2020
doi: 10.3389/fgene.2020.00219

Clinical Interest of Combining Transcriptomic and Genomic Signatures in High-Grade Serous Ovarian Cancer

Yann Kieffer[1,2†], Claire Bonneau[1,2†], Tatiana Popova[2,3], Roman Rouzier[4], Marc-Henri Stern[2,3] and Fatima Mechta-Grigoriou[1,2]*

[1] Institut Curie, Stress and Cancer Laboratory, Equipe labelisée Ligue Nationale Contre le Cancer, PSL University, Paris, France, [2] Inserm, U830, Paris, France, [3] Genomics and Biology of Hereditary Cancers, Institut Curie, Paris, France, [4] Department of Surgery, Institut Curie Hospital Group, René Huguenin Hospital, Saint-Cloud, France

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong,
Hong Kong

Reviewed by:
Yan Zhang,
The Ohio State University,
United States
Wenji Ma,
Columbia University, United States

***Correspondence:**
Fatima Mechta-Grigoriou
fatima.mechta-grigoriou@curie.fr

[†] These authors have contributed equally to this work

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

Received: 03 December 2019
Accepted: 24 February 2020
Published: 17 March 2020

Citation:
Kieffer Y, Bonneau C, Popova T, Rouzier R, Stern M-H and Mechta-Grigoriou F (2020) Clinical Interest of Combining Transcriptomic and Genomic Signatures in High-Grade Serous Ovarian Cancer. Front. Genet. 11:219. doi: 10.3389/fgene.2020.00219

High-grade serous ovarian cancer is one of the deadliest gynecological malignancies and remains a clinical challenge. There is a critical need to effectively define patient stratification in a clinical setting. In this study, we address this question and determine the optimal number of molecular subgroups for ovarian cancer patients. By studying several independent patient cohorts, we observed that classifying high-grade serous ovarian tumors into four molecular subgroups using a transcriptomic-based approach did not reproducibly predict patient survival. In contrast, classifying these tumors into only two molecular subgroups, fibrosis and non-fibrosis, could reliably inform on patient survival. In addition, we found complementarity between transcriptomic data and the genomic signature for homologous recombination deficiency (HRD) that helped in defining prognosis of ovarian cancer patients. We also established that the transcriptomic and genomic signatures underlined independent biological processes and defined four different risk populations. Thus, combining genomic and transcriptomic information appears as the most appropriate stratification method to reliably subgroup high-grade serous ovarian cancer patients. This method can easily be transferred into the clinical setting.

Keywords: HGSOC, fibrosis, mesenchymal, BRCA1/2, homologous recombination deficiency, prognosis

INTRODUCTION

Epithelial ovarian cancer is the fifth leading cause of cancer-related death among women, with only 40% of patients achieving an average 5-year survival (Berns and Bowtell, 2012). Ovarian cancers are predominantly classified by histological subtype (serous, endometrioid, mucinous, clear cell or squamous), grade (low or high) and stage (early or advanced). Approximately 75% of ovarian cancers are high-grade serous ovarian cancers. Standard treatment consists of surgical cytoreduction combined with Taxanes- and platinum salts-based chemotherapy. Recently, targeted therapies have also been included in treatment plans, such as anti-angiogenic drugs or poly-ADP-ribose polymerase (PARP) inhibitors indicated for certain patients with BRCA1/2 mutations

(Fong et al., 2009, 2010; Tutt et al., 2010; Gelmon et al., 2011; Kaye et al., 2012; Ledermann et al., 2012; Liu et al., 2014; Oza et al., 2015; Konecny and Kristeleit, 2016; McLachlan et al., 2016; Miller and Ledermann, 2016). Novel targeted therapies are being developed but their use remains limited, in part due to their cost (Raja et al., 2012; Kmietowicz, 2015; Monk et al., 2016; The Lancet, 2017). To increase the effectiveness of targeted therapies, there is a need to develop accurate methods to define novel patient stratifications that can be easily translated to the clinical environment.

Ovarian cancers have a high frequency of homologous recombination deficiency (HRD) due to germline or somatic mutations in the BRCA1 or BRCA2 genes, methylation of the BRCA1 or RAD51C promoter regions or other genetic alterations (Rigakos and Razis, 2012; Muggia and Safra, 2014). Patients carrying BRCA1/2 mutations have increased sensitivity to platinum salts and longer survival than patients with no BRCA1/2 mutations (Fong et al., 2009, 2010; Audeh et al., 2010; Goundiam et al., 2015) and HRD sensitizes cells to PARP inhibitors (Pujade-Lauraine et al., 2017). To assess HRD in breast and ovarian cancer, the large-scale state transition (LST) genomic signature can be used (Popova et al., 2012; Goundiam et al., 2015). In addition to genomic characterization, previous studies have identified distinct molecular subgroups of high-grade serous ovarian cancers based on transcriptomic profiling (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Sabatier et al., 2011; Bentink et al., 2012; Verhaak et al., 2013; Konecny et al., 2014). Importantly, all currently published studies observed one molecular subgroup, referred to as Stromal, Fibrotic, Mesenchymal or Angiogenic, that is invariably associated with poor patient survival (Tothill et al., 2008; Mateescu et al., 2011; Bentink et al., 2012; Verhaak et al., 2013; Konecny et al., 2014). The first mechanism that explains the Fibrotic/Mesenchymal subgroup, at least in part, is regulation by the miR-200 family of microRNAs (Mateescu et al., 2011; Batista et al., 2013, 2016). Genes inversely correlated with expression of the miR-200 family constitute the fibrosis signature that classifies ovarian cancers with mesenchymal features (Mateescu et al., 2011; Batista et al., 2013, 2016). Conversely, genes positively-correlated with miR-200 expression constitute an oxidative stress signature that classifies the oxidative stress ovarian cancer subgroup. This stress subgroup is associated with a better prognosis and increased cancer cell chemosensitivity (Leskela et al., 2011; Mateescu et al., 2011; Batista et al., 2013, 2016; Brozovic et al., 2015). Notably, the accumulation of miR-200 family members in ovarian tumors could be used for early detection of the pathology, but determining patient outcome through miR-200 expression remains highly controversial, and a consensus is far from being achieved (Batista et al., 2013; Muralidhar and Barbolina, 2015; Shi and Zhang, 2016). The ability to provide information on patient survival remains a priority in the field but the number of molecular subgroups required to define patient survival effectively is unknown, impeding their use in clinical practice. In this study, we address this question and define the optimal number of ovarian cancer molecular subgroups for prognostic stratification of patients.

MATERIALS AND METHODS

Clinical and Transcriptomic Data of Ovarian Cancer Patients

Three cohorts of patients with high-grade serous ovarian cancer were included in this study: Curie, AOCS and TCGA. Curie cohort: Ovarian tumors were obtained from a cohort of 107 patients treated at the Institut Curie between 1989 and 2005. Clinical characteristics of the cohort have already been described in Mateescu et al. (2011). For each patient, a surgical specimen was taken, prior to any chemotherapeutic treatment, for pathological analysis and tumor tissue cryopreservation. The median patient age was 58 years old (with a range of 31–87 years). Ovarian carcinomas were classified according to the World Health Organization histological classification of gynecological tumors. The Curie transcriptomic dataset is from Affymetrix Human Genome U133 Plus 2.0 arrays and is freely available in the Gene Expression Omnibus[1] under the accession number, GSE26193. AOCS cohort: Clinical characteristics of the 285 patients included in the AOCS cohort have been previously described in Tothill et al. (2008), and transcriptomic data, generated using Affymetrix Human Genome U133 Plus2.0 arrays, are freely available under the accession number, GSE9899. TCGA cohort: Clinical characteristics of the 557 patients included in the TCGA cohort, as well as transcriptomic data generated using Affymetrix Human Genome U133A arrays, have been previously described in Cancer Genome and Atlas Research (2011) and can be downloaded from the NIH Genomic Data Commons (GDC) data portal[2]. Most patients treated at Institut Curie are from Caucasian origin and 91% of the patients, for which the ethnicity variable is known in the TCGA cohort, are also from Caucasian origin.

Description of Transcriptomic Signatures

Transcriptomic signatures defining the molecular classification of ovarian cancers were retrieved from four original publications. First, Tothill et al. (2008) identified 478 Affymetrix HG U133 Plus 2.0 probe sets up-regulated in the C1 signature and 2,230 probe sets up-regulated in the C2–C6 signature. Second, Mateescu et al. (2011) identified 22 genes up-regulated in the Stress/non-Fibrosis signature and 16 genes up-regulated in the Fibrosis signature. Third, Bentink et al. (2012) identified 100 Illumina probes up-regulated in the M1 signature and 300 Illumina probes up-regulated in the M2–M4 signature. Lastly, Verhaak et al. (2013) identified 37 genes up-regulated in the Mesenchymal signature and 63 genes up-regulated in the Differentiated/Immunoreactive/Proliferative signature. The different transcriptomic signatures coming from these distinct studies are not overlapping in terms of genes (as shown **Supplementary Figures S1B,D**), enabling us to compare these different signatures as distinct entities.

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26193

[2] https://portal.gdc.cancer.gov

TABLE 1 | Comparative description of clinical parameters in the AOCS, Curie, and TCGA cohorts.

		AOCS	Curie	TCGA
Total number of patients		285	107	557
Age (year)	Median	59	58	59
	Range	22–80	31–87	26–89
Histotype				
	Serous	264 (92.6%)	82 (76.6%)	557 (100%)
	Endometrioid	20 (7.02%)	8 (7.5%)	
	Adenocarcinoma	1 (0.4%)		
	Mucinous		8 (7.5%)	
	Other		9 (8.4%)	
Figo Substage				
	I	24 (8.4%)	21 (19.6%)	
	II	18 (6.3%)	10 (9.35%)	24 (4.3%)
	III	217 (76.1%)	59 (55.14%)	381 (68.4%)
	IV	22 (7.7%)	17 (15.9%)	79 (14.2%)
	Not applicable	4 (1.4%)		73 (13.1%)
Grade				
	1	19 (6.7%)	7 (6.5%)	
	2	97 (34%)	34 (31.5%)	57 (10.2%)
	3	164 (57.5%)	66 (62%)	420 (75.4%)
	Not applicable	5 (1.8%)		80 (14.4%)
Surgery				
	Full	84 (29.5%)	38 (35.5%)	90 (16.2%)
	Partial	164 (57.5%)	69 (64.5%)	342 (61.4%)
	Not applicable	37 (13%)		125 (22.4%)
Clinical response	RC – complete response		51 (47.7%)	276 (49.6%)
	RP – Partial response		22 (20.6%)	57 (10.2%)
	S – Stability		7 (6.5%)	25 (4.5%)
	P – Progression		11 (10.3%)	37 (6.6%)
	Not applicable	285 (100%)	16 (15%)	162 (29.1%)
Signature D-I-M-P				
	Differentiated		30 (28%)	148 (26.6%)
	Immunoreactive		26 (24.3%)	129 (23.2%)
	Mesenchymal		31 (29%)	118 (21.2%)
	Proliferative		20 (18.7%)	138 (24.8%)
	Not applicable	285 (100%)		24 (4.3%)
Mateescu's Signature				
	Stress	150 (52.6%)	51 (47.7%)	326 (58.5%)
	Fibrosis	135 (47.4%)	56 (52.3%)	220 (39.5%)
	Not applicable			11 (2%)
Tothill's Signature				
	C1	83 (29.1%)		107 (19.2%)
	C2–C6	168 (58.9%)		443 (79.5%)
	Not applicable	34 (11.9%)		7 (1.3%)
Bentink's Signature				
	M1			128 (23%)
	M2–M4			422 (75.8%)
	Not applicable	285 (100%)		7 (1.3%)
Lst Signature				
	Low LST			238 (42.7%)
	High LST			303 (54.4%)
	Not applicable	285 (100%)		16 (2.9%)

AOCS, Curie, and TCGA cohorts have previously been described in Tothill et al. (2008), Cancer Genome and Atlas Research (2011), and Mateescu et al. (2011), respectively. For the Curie cohort, tumor samples were obtained from a cohort of ovarian carcinoma patients treated at the Institut Curie from 1989 to 2012. For each patient, a surgical specimen was taken, prior to any chemotherapeutic treatment, for pathological analysis and tumor tissue cryopreservation. The median patient age was 58 years old (with a range of 31–87 years). Ovarian carcinomas were classified according to the World Health Organization histological classification of gynecological tumors.

Enrichment of Biological Processes in Transcriptomic Signatures

Gene ontology (GO) enrichment analysis was performed using the DAVID bioinformatics resources (Version 6.7)[3]. For each signature tested, the 10 most significant biological processes (based on *p*-value) were selected. Reduce and Visualize Gene Ontology (REViGO) software (Supek et al., 2011; accessed January 2017)[4], with a parameter similarity of 0.5, was used to summarize information by removing redundant GO terms.

Classification of High-Grade Serous Ovarian Cancer From the TCGA Cohort According to Different Transcriptomic Signatures

High-grade serous ovarian cancers from the TCGA (Cancer Genome and Atlas Research, 2011) were studied (see **Table 1** for cohort description). Genes that comprise the C1–C6 (Tothill et al., 2008), Stress/Fibrosis (Mateescu et al., 2011), and M1–M4 (Bentink et al., 2012) signatures were applied to the TCGA transcriptomic data. This allowed us to classify high-grade serous ovarian cancers from the TCGA cohort according to Tothill's, Mateescu's, and Bentink's signatures, and compare them to the Differentiated/Immunoreactive/Mesenchymal/Proliferative (D-I-M-P) classification, initially generated from the TCGA dataset (Verhaak et al., 2013). Briefly, we first performed the hierarchical clustering shown in **Figure 1A** based on DIMP signature using Euclidean distance and Ward's agglomeration method. To compare this DIMP classification with the others, we next performed similar hierarchical clustering by applying each of the other signatures (Tothill et al., 2008; Mateescu et al., 2011; Bentink et al., 2012) on the TCGA transcriptomic dataset by using same parameters (Euclidean distance and Ward's agglomeration method). Only genes specific of each signatures were kept for the clustering. For the four signatures, each resulting dendrogram tree was next cut into two subgroups for classifying patients into two subgroups according to each signature (Stress/Fibrosis for Mateescu classification, C1/C2–C6 for Tothill classification and M1/M2–M4 for Bentink classification). By this way, for each of the four classifications studied, we have been able to determine to which subgroup each patient belongs, as show **Figures 1A,B**. The distribution of ovarian cancers from TCGA across the four signatures can be found in **Table 1**. Patient classification was thus independent of patient survival and strictly based on tumor molecular signature. We also aimed at comparing the association of patient clinical features with two distinct classifications, i.e., classification in two subgroups based on Mateescu's signature and classification in four subgroups based on Verhaac's signature using Fisher's exact test (as shown in **Table 2**). No correction was applied to *p*-values.

Expression of miR-200 Family Members

The predictive value of the miR-200 family was evaluated because this miRNA family was shown to be associated with the stress (non-Fibrosis)/Fibrosis classification (Mateescu et al., 2011; Batista et al., 2013, 2016). Indeed, genes that are inversely correlated with the miR-200 expression compose the "Fibrosis" signature and classify ovarian cancers with mesenchymal features. Conversely, genes positively-correlated with miR-200 expression constitute the non-Fibrosis (oxidative stress) signature and classify the "non-Fibrosis" ovarian cancer subgroup. Expression of the miR-200 family members (miR-141, miR-200a, miR-200b, miR-200c, and miR-429) was determined using the level 3 expression data from the TCGA data portal. Groups of low or high microRNA expression were defined using their median as a threshold to perform survival analysis.

Large-Scale State Transition (LST) Genomic Signature of HRD

Cytoscan HD SNP-array (Affymetrix) data were processed using the Genome Alteration Print (GAP) methodology to obtain absolute copy number profiles (Popova et al., 2009). DNA index was calculated as the averaged copy number. Based on the DNA index, tumor ploidy was set as near-diploid (DNA index < 1.3) or near-tetraploid (DNA index ≥ 1.3). Detection of HRD was determined by the number of LST, as previously described (Popova et al., 2012). Briefly, LST was defined as a chromosomal breakpoint (change in copy number or major allele counts) between adjacent regions of at least 10 Mb. The number of LST were then calculated after smoothing and filtering out copy number variant regions < 3 Mb. Tumors were segregated into near-diploid or near-tetraploid subgroups. Based on two ploidy-specific cut-offs (15 and 20 LST per genome in near-diploid and near-tetraploid tumors, respectively) tumors were classified as LST high (LST^{Hi}, equal or above the cut-off) or LST low (LST^{Lo}, below the cut-off). LST^{Hi} represents the HRD genomic pattern and LST^{Lo} corresponds to the non-HRD profile.

Statistical Analysis

All statistical analyses were performed in the R environment (Versions 3.3.2, 3.4.0, and 3.6.1)[5]. Fisher's exact test was used to determine any association between classes of ovarian cancers and clinical parameters. Overall survival (OS) and disease-free survival (DFS) were investigated using the Cox proportional hazards model and Kaplan-Meier curves through the R packages, *survival* and *survminer*. To identify differences between survival curves, *p*-values were assessed by the log-rank test. *P*-values ≤ 0.05 were considered to be statistically significant. To take into account multiple testing, *p*-values were adjusted using the Benjamini-Hochberg procedure using *pairwise*_survdiff function from R package *Survminer*.

Code Availability

R scripts used to generate panels of the Figures, Supplementary Figures and Tables are provided within the data source file of the paper, available with the doi: 10.6084/m9.figshare.11663232.

[3]https://david.ncifcrf.gov

[4]http://revigo.irb.hr

[5]https://cran.r-project.org

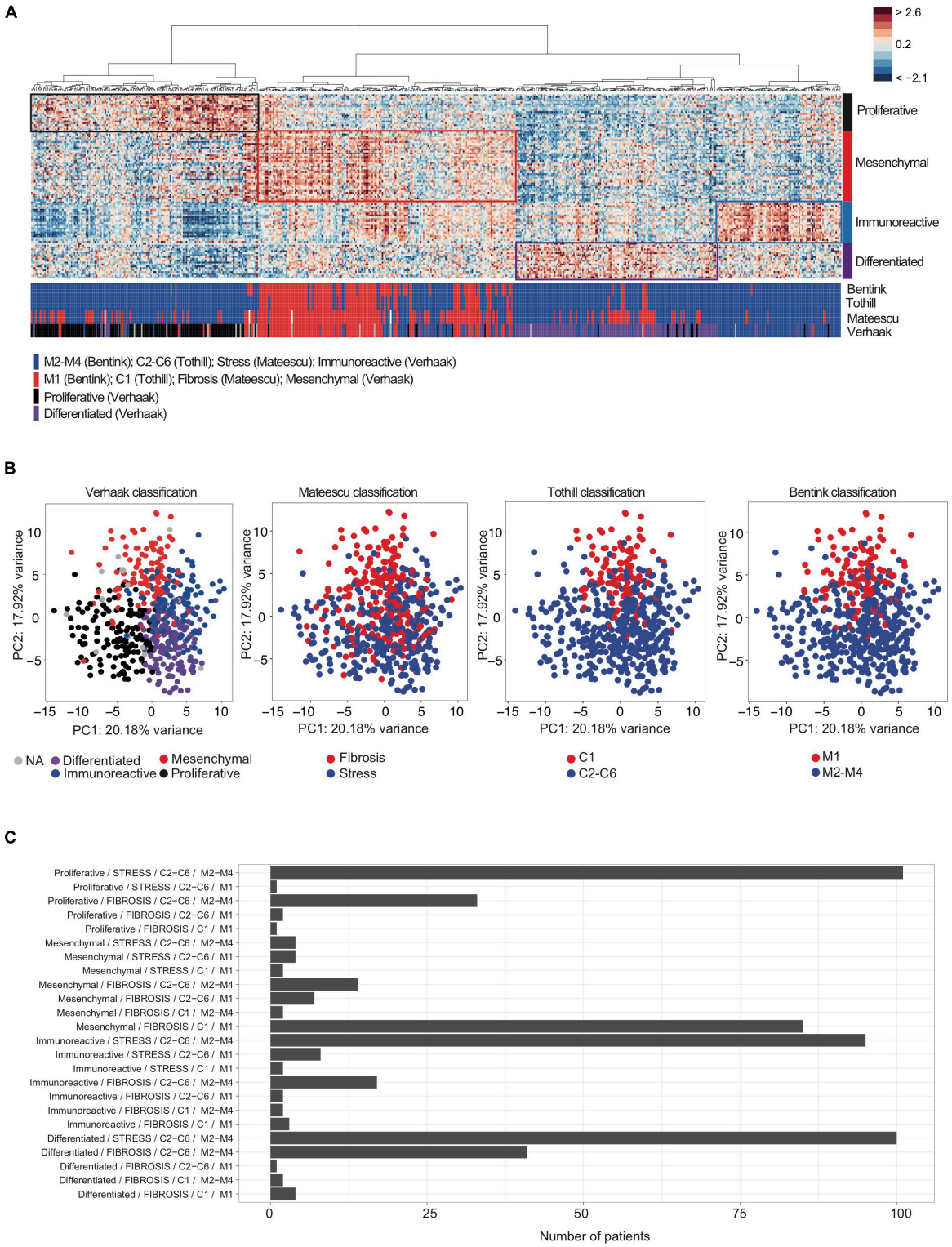


FIGURE 1 | Overlap between transcriptomic signatures used for classification of high-grade serous ovarian cancers. **(A)** A heatmap from hierarchical clustering applied on the TCGA cohort. Rows represent genes and columns represent patients. Clustering is based on the 100 genes of the D-I-M-P signature (Verhaak et al., 2013) using Pearson distance and Ward's agglomeration method. The color saturation shows the magnitude of the deviation from the mean for each gene, with red and blue indicating expression values above or below the mean, respectively. Colored bars below the heatmap represent tumor classifications obtained from the four

(Continued)

FIGURE 1 | Continued
transcriptomic signatures (Tothill et al., 2008; Mateescu et al., 2011; Bentink et al., 2012; Verhaak et al., 2013), as indicated. The red bars correspond to the Mesenchymal, C1, Angiogenic or Fibrosis subgroup, according to the classification considered. The blue bars correspond to C2–C6, non-Angiogenic and Stress subgroups. For the D-I-M-P signature, blue bars correspond to Immunoreactive, black correspond to Proliferative and purple bars correspond to Differentiated subgroups. **(B**, Left) Principal Component Analysis (PCA) applied on transcriptomic data from the TCGA cohort, using the 100 genes composing the D-I-M-P signature (Verhaak et al., 2013). The color code represents the four D-I-M-P molecular subgroups: Mesenchymal (red, *N* = 118), Differentiated (purple, *N* = 148), Proliferative (black, *N* = 138) and Immunoreactive (blue, *N* = 129). (Middle and Right) Further PCA with subgroups highlighted using Fibrosis (red, *N* = 220) or Stress (blue, *N* = 326) (Mateescu et al., 2011); C1 (red, *N* = 107) or C2–C6 (blue, *N* = 443) (Tothill et al., 2008); Angiogenic (M1, red, *N* = 128) or non-Angiogenic (M2–M4, blue, *N* = 422) (Bentink et al., 2012) signatures, as indicated. **(C)** Barplots showing the number of patients according to each combination of classes among the four classifications (Verhaak/Mateescu/Tothill/Bentink).

RESULTS

The Fibrosis Subgroup of High-Grade Serous Ovarian Cancers Exhibits Conserved Functional Pathways Across Studies

Although the genes defining ovarian cancer molecular subgroups were different across studies (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Sabatier et al., 2011; Bentink et al., 2012; Verhaak et al., 2013; Konecny et al., 2014), we observed that some of the identified functions were consistent across the fibrosis subgroups (**Supplementary Figures S1A,B**). Following a GO enrichment analysis on previously published ovarian cancer transcriptomic signatures (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Bentink et al., 2012; Verhaak et al., 2013), we found consistent

TABLE 2 | The association between transcriptomic signatures and clinical parameters.

		Fibrosis/non-Fibrosis classification		
		Non-fibrosis	**Fibrosis**	***p*-value**
Grade				$p = 0.67$
	G2	32 (11.6%)	25 (12.8%)	
	G3	245 (88.5%)	170 (87.2%)	
Stage				$\mathbf{p = 0.01}$
	II	20 (7.1%)	4 (2%)	
	III–IV	260 (92.9%)	195 (98%)	
Debulking				$\mathbf{p = 0.05}$
	Full	60 (24%)	28 (15.8%)	
	Partial	190 (76%)	149 (84.2%)	
Platinum resistance				$p = 0.38$
	Sensitive	153 (75.7%)	99 (71.3%)	
	Resistant	49 (24.3%)	40 (28.8%)	
Primary therapy outcome				$\mathbf{p = 0.02}$
	Complete response	172 (74.5%)	101 (63.1%)	
	Partial response	59 (25.6%)	59 (37%)	
BRCA1/2 mutation				$p = 0.11$
	No	269 (84.6%)	171 (78.8%)	
	Yes	49 (15.4%)	46 (21.2%)	
BRCA1 methylation				$p = 1$
	No	278 (87.4%)	190 (87.6%)	
	Yes	40 (12.6%)	27 (12.4%)	
RAD51C methylation				$p = 0.39$
	No	312 (98.1%)	210 (96.8%)	
	Yes	6 (1.9%)	7 (3.2%)	
LST signature (HRD)				$p = 0.29$
	Low	147 (46.2%)	89 (41.2%)	
	High	171 (53.8%)	127 (58.8%)	
Ploidy				$p = 0.20$
	2	104 (32.7%)	83 (38.4%)	
	$\geq$ 4	214(67.3%)	133 (61.6%)	

(Continued)

TABLE 2 | Continued

		D-I-M-P classification				
		D	**I**	**M**	**P**	***p*-value**
Grade						$p = 0.28$
	G2	13 (10.1%)	9 (8.8%)	17 (17%)	18 (13.4%)	
	G3	116 (89.9%)	93 (91.2%)	83 (83%)	116 (86.6%)	
Stage						***p* = 0.007**
	II	5 (3.7%)	12 (11.5%)	1 (1%)	6 (4.5%)	
	III–IV	129 (96.3%)	92 (88.5%)	101 (99%)	126 (95.5%)	
Debulking						***p* = 0.03**
	Full	33 (26.8%)	17 (19.5%)	10 (11%)	28 (23.5%)	
	Partial	90 (73.2%)	70 (80.5%)	81 (89%)	91 (76.5%)	
Platinum resistance						$p = 0.70$
	Sensitive	70 (70%)	54 (78.3%)	53 (74.7%)	73 (74.5%)	
	Resistant	30 (30%)	15 (21.7%)	18 (25.4%)	25 (25.5%)	
Primary therapy outcome						$p = 0.15$
	Complete response	78 (69.6%)	60 (69%)	49 (62%)	85 (77.3%)	
	Partial response	34 (30.3%)	27 (31%)	30 (38%)	25 (22.7%)	
BRCA1/2 mutation						***p* = 0.05**
	No	118 (79.7%)	103 (79.8%)	94 (79.7%)	124 (89.9%)	
	Yes	30 (20.3%)	26 (20.2%)	24 (20.3%)	14 (10.1%)	
BRCA1 methylation						$p = 0.15$
	No	127 (85.8%)	110 (85.3%)	101 (85.6%)	128 (92.8%)	
	Yes	21 (14.2%)	19 (14.7%)	17 (14.4%)	10 (7.2%)	
RAD51C methylation						$p = 0.38$
	No	144 (97.3%)	124 (96.1%)	115 (97.5%)	137 (99.3%)	
	Yes	4 (2.7%)	5 (3.9%)	3 (2.5%)	1 (0.7%)	
LST signature (HRD)						***p* = 0.0002**
	Low	62 (41.9%)	43 (33.3%)	48 (40.7%)	82 (59.4%)	
	High	86 (58.1%)	86 (66.7%)	70 (59.3%)	56 (40.6%)	
Ploidy						***p* = 4.6e-5**
	2	71 (48.0%)	39 (30.2%)	46 (39.0%)	31 (22.5%)	
	≥ 4	77 (52.0%)	90 (69.8%)	72 (61.0%)	107 (77.5%)	

Contingency table showing the association between Fibrosis/non-Fibrosis (two subgroups, Mateescu's classification), or D-I-M-P subgroups (four subgroups, Verhaac's classification), and clinical parameters. Data are from the TCGA cohort and the number of patients and frequencies in the population are indicated. Debulking status was defined as full when no macroscopic residue was detected after surgery or as partial otherwise. Response to primary therapy was considered as partial if the patient indicated with partial response, stable disease or progressive disease, considering both surgery efficiency and sensitivity to chemotherapies. P-values are calculated using Fisher's exact test without correction and significant p-values are indicated in bold.

enrichment in particular pathways, including cell adhesion, extracellular matrix organization, and response to wounding (**Supplementary Figure S1A**). It is important to note that this molecular ovarian cancer subgroup was named differently across studies, and referred to as C1 (Tothill et al., 2008), Fibrosis (Mateescu et al., 2011), Angiogenic (Bentink et al., 2012), or Mesenchymal (Verhaak et al., 2013; Konecny et al., 2014) subgroups, but they all possess similar biological features (mainly fibrosis and mesenchymal properties) (**Supplementary Figure S1A**). However, apart from Fibronectin 1 (FN1), the transcriptomic signatures did not show any overlap in gene expression (**Supplementary Figure S1B**). In contrast to the C1/Fibrosis/Angiogenic/Mesenchymal signature, none of the others signatures, defining C2–C6 (Tothill et al., 2008), Oxidative stress (Mateescu et al., 2011), Anti-angiogenic (M2–M4) (Bentink et al., 2012), or Differentiated-Immunoreactive-Proliferative (Verhaak et al., 2013; Konecny et al., 2014) high-grade serous ovarian cancer subgroups, showed overlap in either gene expression or pathways (**Supplementary Figures S1C,D**).

We next sought to test if the ovarian cancer patients identified by these different transcriptomic signatures were the same (**Figure 1**). To do so, we studied the TCGA cohort (Cancer Genome and Atlas Research, 2011; **Table 1** for cohort description) and classified each patient using the four transcriptomic signatures (Tothill et al., 2008; Mateescu et al., 2011; Bentink et al., 2012; Verhaak et al., 2013). Unsupervised analyses, including hierarchical clustering (**Figure 1A**) and Principal Component Analyses (**Figure 1B**), confirmed that there was a significant overlap between tumor classification in C1, Fibrosis, Angiogenic, and Mesenchymal subtypes. Indeed, patients classified

as Mesenchymal were also mainly classified as Fibrosis, C1 and M1, while patients classified as Stress, C2–C6 and M1–M4 can be equally classified as Proliferative, Immunoreactive or Differentiated (**Figure 1C**). Therefore, these different gene signatures not only identified the same biological characteristics (mesenchymal properties, accumulation of extra-cellular matrix components, and pro-angiogenic features) but also identified the same patients (**Figures 1A–C**). Almost all patients (91.5%) defined as Mesenchymal (using Verhaak's signature) were also classified as Fibrosis (using Mateescu's signature). However, 26% of non-Mesenchymal patients (using Verhaak's signature) were identified as Fibrosis (using Mateescu's signature), suggesting possible misclassifications. Interestingly, the overall survival and disease-free survival of those discordant patients (non-Mesenchymal/Fibrosis) were similar to the Mesenchymal/Fibrosis defined patients, but significantly different from non-Mesenchymal/non-Fibrosis patients (**Supplementary Figure S2**). This suggests that these patients could be classified as Fibrosis, as defined by Mateescu's signature. These observations show that high-grade serous ovarian cancers can be divided into two major molecular subtypes according to transcriptomic profiles: Fibrosis and non-Fibrosis.

High-Grade Serous Ovarian Cancers Stratified Into Two Subgroups Are Associated With Stage, Debulking, and Clinical Response to Treatment

We next questioned if stratification of ovarian cancers into four molecular subgroups (such as D-I-M-P, based on Verhaak's classification) could be more informative regarding clinical features than classification into two subgroups (Fibrosis and non-Fibrosis) (based on Mateescu's classification). The non-Fibrosis and Fibrosis subgroups were significantly associated with stages (p = 0.01), debulking (p = 0.05) and primary therapy outcome (p = 0.02) (**Table 2**). However, they were not associated with grade, ploidy, sensitivity to platinum, BRCA1/2 mutations or BRCA1 or RAD51C promoter methylation (**Table 2**). LST signature, which is linked to HRD status (Fong et al., 2009, 2010; Audeh et al., 2010; Popova et al., 2012; Goundiam et al., 2015), was also not significantly associated with the Fibrosis and non-Fibrosis subgroups (**Table 2**). The four D-I-M-P subgroups showed a significant association with stage (p = 0.007) and debulking (p = 0.03), but not with response to treatment. In addition, D-I-M-P was associated with ploidy (p = 4.6e-5), BRCA1/2 mutations (p = 0.05) and LST signature (p = 0.0002) but not with grade, platinum resistance and primary therapy outcome

TABLE 3 | Stratification of high-grade serous ovarian cancers into two subgroups provides a prognostic value, independent of stage and debulking.

	OS univariate analysis					OS multivariate analysis				
	HR	**CI 95% inf**	**CI 95% sup**	***p*-value**		**HR**	**CI 95% inf**	**CI 95% sup**	***p*-value**	
Signature										
Non-Fibrosis	Ref					Ref				
Fibrosis	1.43	1.13	1.82	**0.003**	**	1.22	0.95	1.57	0.12	
Stage										
II	Ref					Ref				
III	2.49	1.17	5.29	**0.02**	*	2.39	0.97	5.87	0.06	
IV	3.28	1.49	7.25	**0.003**	**	2.82	1.11	7.18	**0.03**	*
Debulking										
Full	Ref					Ref				
Partial	2.01	1.37	2.94	**0.0004**	***	1.90	1.27	2.82	**0.002**	**
Age										
<59 years	Ref									
>59 years	1.2	0.96	1.55	0.11						
Signature										
D	1.48	1.03	2.14	**0.04**	*	1.23	0.83	1.80	0.30	
I	Ref					Ref				
M	1.67	1.13	2.47	**0.01**	**	1.12	0.74	1.69	0.59	
P	1.40	0.96	2.03	0.08		1.17	0.79	1.72	0.43	
Stage										
II	Ref					Ref				
III	2.49	1.17	5.29	**0.02**	*	2.44	0.99	6.00	**0.05**	
IV	3.28	1.49	7.25	**0.003**	**	2.72	1.06	6.96	**0.04**	*
Debulking										
Full	Ref					Res				
Partial	2.01	1.37	2.94	**0.0004**	***	1.92	1.29	2.86	**0.001**	**

(Continued)

TABLE 3 | Continued

	DFS univariate analysis					DFS multivariate analysis				
	HR	**CI 95% inf**	**CI 95% sup**	***p*-value**		**HR**	**CI 95% inf**	**CI 95% sup**	***p*-value**	
Signature										
Non-Fibrosis	Ref					Ref				
Fibrosis	1.37	1.08	1.73	**0.01**	*	1.28	0.99	1.65	**0.05**	
Stage										
II	Ref					Ref				
III	1.93	1.12	3.31	**0.02**	*	1.67	0.89	3.12	0.11	
IV	2.48	1.35	4.54	**0.003**	**	2.03	1.02	4.04	**0.05**	*
Debulking										
Full	Ref					Ref				
Partial	1.69	1.23	2.32	**0.001**	**	1.54	1.11	2.13	**0.01**	*
Age										
<59 years	Ref									
>59 years	0.95	0.75	1.2	0.7						
Signature										
D	1.32	0.94	1.86	0.11		1.19	0.84	1.72	0.35	
I	Ref					Ref				
M	1.41	0.97	2.04	0.07		1.14	0.76	1.69	0.53	
P	1.23	0.87	1.74	0.24		1.12	0.78	1.63	0.54	
Stage										
II	Ref					Ref				
III	1.93	1.12	3.31	**0.02**	*	1.73	0.93	3.24	0.09	
IV	2.48	1.35	4.54	**0.003**	**	2.11	1.06	4.21	**0.03**	*
Debulking										
Full	Ref					Ref				
Partial	1.69	1.23	2.32	**0.001**	**	1.54	1.11	2.14	**0.01**	**

*Cox proportional hazards regression was performed on Fibrosis/non-Fibrosis or D-I-M-P subgroups and evaluated for overall survival (OS, Top) and disease-free survival (DFS, Bottom). These analyses were either: adjusted for stage and debulking status (multivariate, Right) or unadjusted (univariate, Left). Age at diagnosis was not taken into account for multivariate analysis as it was not significant at univariate level. HR, hazard ratio; CI 95% inf, lower limit of the 95% confidence interval; CI 95% sup, upper limit of the 95% confidence interval. Significant p-values are indicated in bold. *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.*

(**Table 2**). These results suggest that defining Mesenchymal ovarian cancers by applying the D-I-M-P signature is less informative than the Fibrosis classification.

Stratification Into Two Ovarian Cancer Subgroups Provides a Reliable Prognostic Value for Patient Survival

Taking into account the association between transcriptomic signatures and clinical parameters, we next investigated if these different transcriptomic signatures could be utilized as independent prognostic factors, compared to stage and debulking status, the two major variables of patient outcome used in clinics. Based on univariate analyses using the Cox regression model, we observed that the Fibrosis/non-Fibrosis signature was indicative of both overall survival (OS) and disease-free survival (DFS), with a shorter survival for the Fibrosis patients (**Table 3**). In contrast, while the D-I-M-P signature was indicative of overall survival, it had no prognostic value for disease-free survival in the univariate analysis (**Table 3**). In the multivariate analysis, none of the transcriptomic stratifications (either into two or four subgroups) were associated with overall survival, independent of stage and debulking (**Table 3**). Still, the Fibrosis – non-Fibrosis signature was the only one to be independent of stage and debulking and to provide additive prognostic value for disease-free survival (**Table 3**).

In the Kaplan-Meier survival analyses, Fibrosis patients exhibited significantly shorter overall survival (**Figure 2A**, Top) and disease-free survival (**Figure 2A**, Bottom) than non-Fibrosis patients in the three independent cohorts analyzed (Curie, AOCS, and TCGA). Classification using the D-I-M-P signature was initially only performed in the TCGA cohort (Verhaak et al., 2013). Therefore, we used unsupervised clustering to identify the four D-I-M-P subgroups in the Curie and AOCS cohorts (**Figure 2B**). The classification into those four subgroups was prognostic factor for overall survival and disease-free survival in the AOCS cohort, but not in the Curie and TCGA cohorts (**Figure 2C**). This shows that the D-I-M-P signature does not provide a systematic prognostic value for survival of ovarian cancer patients, but the division into two molecular subgroups, Fibrosis and non-Fibrosis, is discriminant and reliable. Because the Fibrosis/non-Fibrosis signature was defined by genes correlated- or anti-correlated with miR-200 expression (Mateescu et al., 2011; Batista et al., 2013, 2016), we also evaluated their prognostic value. No microRNA, separately or in

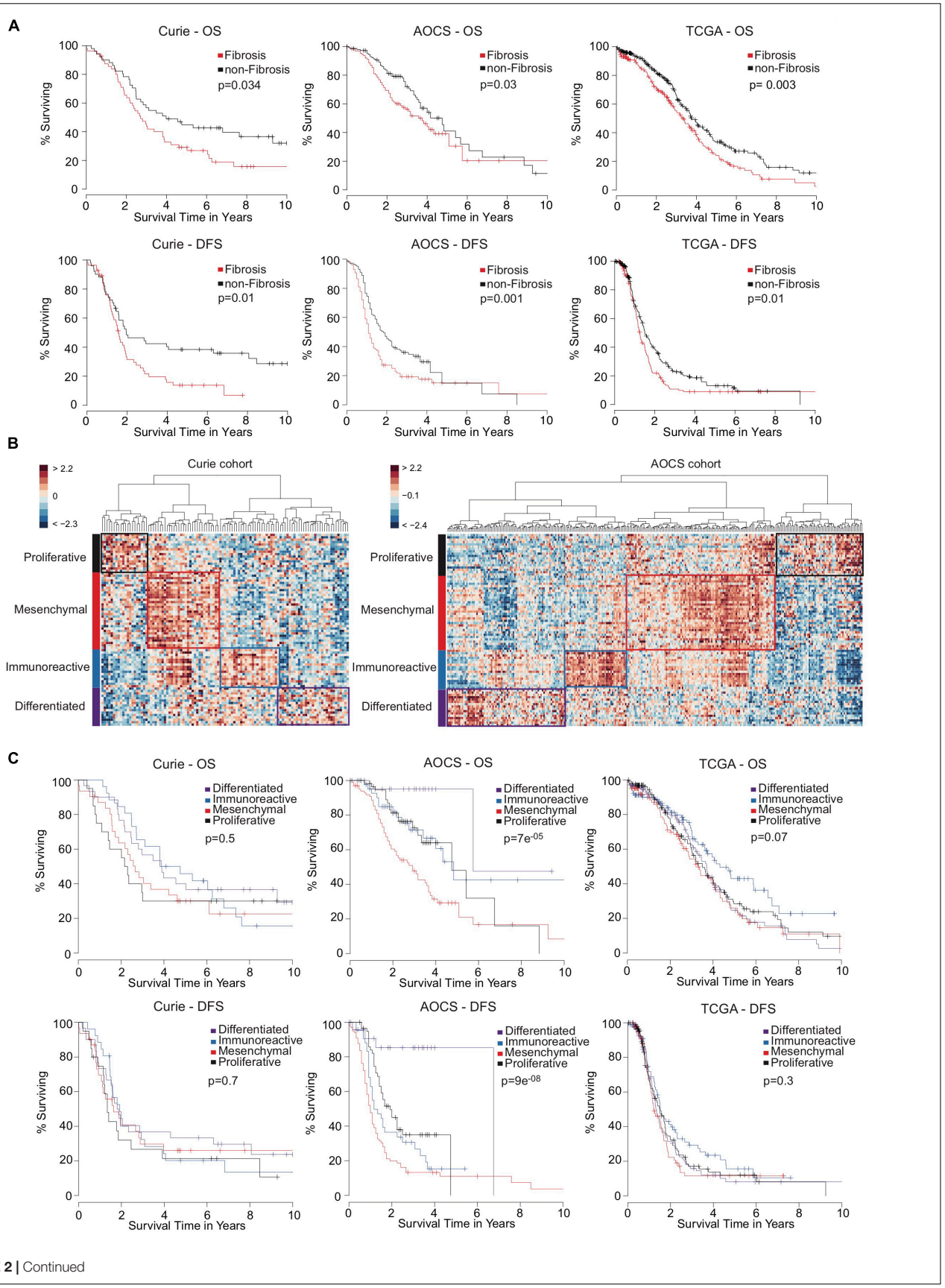


FIGURE 2 | Continued

FIGURE 2 | High-grade serous ovarian cancers stratified into two transcriptomic subgroups exhibit a reliable prognostic value of patient survival. **(A)** Kaplan-Meier curves showing 10-year overall survival (OS, Top) and disease-free survival (DFS, Bottom) of patients with Fibrosis (red) or non-Fibrosis (black) ovarian cancers. Patients from the Curie cohort (56 Fibrosis and 51 non-Fibrosis), the AOCS cohort (135 Fibrosis and 150 non-Fibrosis) and the TCGA cohort (220 Fibrosis and 326 non-Fibrosis) were analyzed, as indicated. *P*-values were calculated using the Log-rank test. **(B)** Heatmap and hierarchical clustering applied on the Curie **(**Left) and AOCS (Right**)** cohorts. Rows represent genes and columns represent patients. Clustering is based on the 100 genes of the D-I-M-P signature (Verhaak et al., 2013) using Euclidean distance and Ward's agglomeration method. The color saturation shows the magnitude of the deviation from the mean for each gene, with red and blue indicating expression values above or below the mean, respectively. **(C)** Kaplan-Meier curves showing 10-year overall survival (OS, Top) and disease-free survival (DFS, Bottom) of ovarian cancer patients according to D-I-M-P classification. Patients from the Curie (*N* = 30 Differentiated, *N* = 26 Immunoreactive, *N* = 31 Mesenchymal, and *N* = 20 Proliferative), the AOCS (*N* = 25 Differentiated, *N* = 42 Immunoreactive, *N* = 102 Mesenchymal, and *N* = 60 Proliferative) and the TCGA (*N* = 148 Differentiated, *N* = 129 Immunoreactive, *N* = 118 Mesenchymal, and *N* = 138 Proliferative) cohorts were analyzed. *P*-values were calculated using the Log-rank test.

combination, was sufficient as a prognostic marker for patient survival (**Supplementary Figure S3**), indicating that expression of the miR-200 family is not an applicable surrogate marker of patient outcome. In conclusion, stratification of ovarian cancer patients into two subgroups using the Fibrosis/non-Fibrosis signature provides a reliable prognostic value for patient survival, but using the D-I-M-P signature or miR-200 family member expression levels do not.

LST Genomic Signature Identifies Ovarian Cancer With HRD

High-grade serous ovarian cancers were analyzed according to the LST genomic signature allowing us to stratify patients into two subgroups: high LST (LST^{Hi}) for HRD tumors (303 tumors, 56%) or low LST (LST^{Lo}) for non-HRD tumors (238 tumors, 44%). As expected, LST^{Hi} ovarian cancers were associated with BRCA1/2 mutations, BRCA1 or RAD51C promotor methylation and showed increased sensitivity to platinum-based chemotherapy (**Table 4**). In contrast, the LST signature was not significantly associated with grade, debulking status or primary therapy outcome (**Table 4**). Univariate analyses, using the Cox regression model, showed that LST signature was indicative of better survival for LST^{Hi} patients ($p = 5.4 \times 10^{-10}$ for overall survival and $p = 1.7 \times 10^{-5}$ for disease-free survival). BRCA1/2 mutations were also associated with better patient outcome ($p = 1.8 \times 10^{-4}$ for overall survival and $p = 0.01$ for disease-free survival), but methylation of BRCA1 and RAD51C promoter regions were not. In multivariate Cox analyses adjusted for BRCA1/2 mutations, LST^{Lo} patients remained significantly associated with shorter disease-free survival (HR = 1.6, CI95% [1.2–2.1], $p = 3.9 \times 10^{-4}$, with HR, Hazard Ratio and CI, Confidence Interval) and overall survival (HR = 1.95, CI 95% [1.5–2.5], $p = 6.7 \times 10^{-7}$), whereas the presence of a BRCA mutation was not associated with disease-free survival ($p = 0.34$) and much less associated with overall survival ($p = 0.04$). This shows that using the LST signature is more efficient for predicting survival of ovarian cancer patients than testing the presence of BRCA1/2 mutations.

Genomic and Transcriptomic Signatures Provide Additive Prognostic Values for Ovarian Cancer Patient Survival

As shown above, the LST signature was significantly associated with HRD and platinum-sensitivity. In contrast, the Fibrosis/non-Fibrosis signature was linked to stage and clinical response to treatment, suggesting these signatures could be complementary. Performing Principal Component Analyses (PCA) on the TCGA transcriptomic data (Verhaak et al., 2013), we observed that the Fibrosis/non-Fibrosis signature did not overlap with the LST signature (**Figure 3A**, Top) and this lack of association was also statistically confirmed ($p = 0.29$, **Table 2**). Interestingly, the two signatures were not associated with the same principal components (PC): Fibrosis/non-Fibrosis signature was found associated with PC2 ($p < 2.2 \times 10^{-16}$) while the LST signature was found associated with PC1 ($p = 2.1 \times 10^{-6}$) (**Figure 3A**, Bottom). We then investigated if, together, they could provide additive value regarding prognosis. Interestingly, the genomic (LST) and transcriptomic (Fibrosis-/non-Fibrosis) signatures were complementary and defined four distinct patient subgroups with significantly different survival (**Figure 3B**). In other words, Fibrosis and non-Fibrosis patients could be subdivided into LST^{Hi} and LST^{Lo} subgroups. As expected, the Fibrosis subtype was associated with poor prognosis, in particular when combined to LST^{Lo}, the non-HRD status (**Figure 3B**). Reciprocally, the non-Fibrosis patients were characterized by a better outcome, especially when associated with the LST^{Hi} subgroup (**Figure 3B**). Pairwise comparison showed that each subgroup was significantly different from each other, in term of overall survival and disease-free survival (apart from the LST^{Lo}/non-Fibrosis subgroup in the disease-free survival analyses) (**Table 5**). These data show that combining genomic and transcriptomic signatures improved stratification of high-grade serous ovarian cancers and provided a significant additive prognostic value. These two signatures (genomic and transcriptomic) were independent for predicting disease-free survival (LST: HR = 1.7, CI 95% [1.3–2.2], $p = 10^{-5}$; Fibrosis/non-Fibrosis: HR = 1.4, CI 95% [1.1–1.8], $p = 6 \times 10^{-3}$ by multivariate Cox regression analysis) and overall survival (LST: HR = 2.2, CI 95% [1.7–2.8]; $p = 4 \times 10^{-10}$; Fibrosis/non-Fibrosis: HR = 1.5, CI 95% [1.2–2], $p = 1 \times 10^{-3}$). In contrast to the Fibrosis/non-Fibrosis signature, the D-I-M-P and LST signatures were significantly associated ($p = 0.02$, **Table 2**). The multivariate Cox analysis adjusted for the D-I-M-P and LST signatures showed that the two signatures were independent for predicting overall survival (LST: HR = 2.16, CI 95% [1.7–2.8], $p = 2.3 \times 10^{-9}$; DIMP: HR = 1.57, CI 95% [1.1–2.3], $p = 0.02$). In contrast, only the LST signature was significantly associated with the disease-free survival (LST: HR = 1.71, CI 95% [1.3–2.2], $p = 2.3 \times 10^{-5}$), while D-I-M-P was not. In conclusion, these

TABLE 4 | Association between genomic signature and clinical parameters.

		LST^{Lo}	LST^{Hi}	*p*-value
Grade				$p = 0.89$
	G2	27 (12.5%)	30 (11.8%)	
	G3	189 (87.5%)	224 (88.2%)	
Stage				$p = 0.29$
	II	8 (3.7%)	16 (6.2%)	
	III–IV	210 (96.3%)	243 (93.8%)	
Debulking				$p = 0.55$
	Full	39 (19.8%)	51 (22.4%)	
	Partial	158 (80.2%)	177 (77.6%)	
Platinum resistance				$\mathbf{p = 0.0001}$
	Sensitive	71 (56.3%)	124 (78.5%)	
	Resistant	55 (43.7%)	34 (21.5%)	
Primary therapy outcome				$p = 0.07$
	Complete response	111 (65.3%)	164 (73.9%)	
	Partial response	59 (34.7%)	58 (26.1%)	
BRCA1/2 mutation				$\mathbf{p = 2.0 \times 10^{-15}}$
	No	229 (96.2%)	217 (71.6%)	
	Yes	9 (3.8%)	86 (28.4%)	
BRCA1 methylation				$\mathbf{p = 1.9 \times 10^{-19}}$
	No	238 (100%)	235 (77.6%)	
	Yes	0 (0%)	68 (22.4%)	
RAD51C methylation				$\mathbf{p = 0.0008}$
	No	238 (100%)	290 (95.7%)	
	Yes	0 (0%)	13 (4.3%)	
Transcriptomic signature				$p = 0.29$
	Non-Fibrosis	147 (62.3%)	171 (57.4%)	
	Fibrosis	89 (37.7%)	127 (42.6%)	
Ploidy				$\mathbf{p = 2.7 \times 10^{-15}}$
	2	41 (17.2%)	150 (49.5%)	
	≥ 4	197(82.8%)	153 (50.5%)	

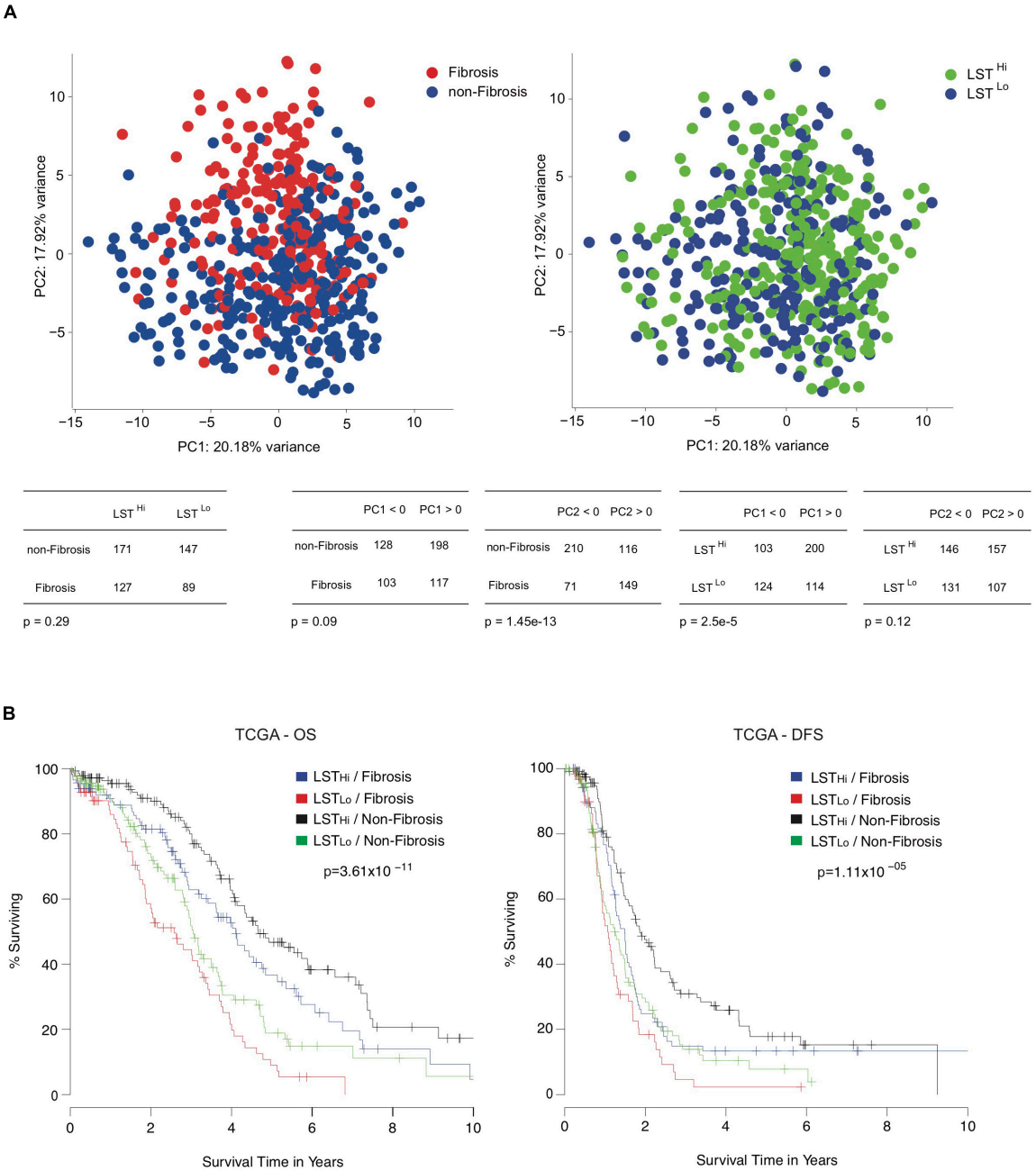
Contingency table showing associations between the LST^{Lo}/LST^{Hi} subgroups and clinical parameters. Debulking status and response to primary therapy were defined as in ***Table 2****. These analyses were performed on the TCGA cohort. P-values were calculated using Fisher's exact test and significant p-values are indicated in bold.*

results demonstrate that combining genomic and transcriptomic information is the most reliable method for stratifying high-grade serous ovarian cancer patients.

DISCUSSION

Stratification of high-grade serous ovarian cancer patients remains unclear. Previously, several ovarian cancer molecular subgroups were identified according to transcriptomic signatures (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Sabatier et al., 2011; Bentink et al., 2012; Verhaak et al., 2013; Konecny et al., 2014) or genomic (Fong et al., 2009, 2010; Audeh et al., 2010; Goundiam et al., 2015). In this study, we define the optimal number of ovarian cancer molecular subgroups with reproducible prognostic value. The study of several independent cohorts showed that classifying ovarian tumors into four molecular subgroups, based on D-I-M-P signatures, does not reproducibly inform on patient survival. In contrast, the subdivision of patients into two molecular subgroups (Fibrosis/non-Fibrosis) provided reliable prediction of patient survival. We also identified a novel complementarity between transcriptomic and genomic data. Indeed, transcriptomic profiling and HRD status characterize specific biological processes and could accurately reflect the different key components in ovarian tumors. Furthermore, combining both genomic and transcriptomic data identified four ovarian cancer patient subgroups with distinct prognostic values and is, therefore, currently the most appropriate method for stratifying high-grade serous ovarian cancer patients.

Although several transcriptomic signatures in ovarian cancers have been proposed (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Sabatier et al., 2011; Bentink et al., 2012; Verhaak et al., 2013; Konecny et al., 2014), there is no clear consensus for choosing a specific one. This is mainly due to the lack of overlap in the gene sets of these transcriptomic signatures (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Sabatier et al., 2011; Bentink et al., 2012; Verhaak et al., 2013;



	LST Hi	LST Lo
non-Fibrosis	171	147
Fibrosis	127	89

p = 0.29

	PC1 < 0	PC1 > 0
non-Fibrosis	128	198
Fibrosis	103	117

p = 0.09

	PC2 < 0	PC2 > 0
non-Fibrosis	210	116
Fibrosis	71	149

p = 1.45e-13

	PC1 < 0	PC1 > 0
LST Hi	103	200
LST Lo	124	114

p = 2.5e-5

	PC2 < 0	PC2 > 0
LST Hi	146	157
LST Lo	131	107

p = 0.12

FIGURE 3 | Combining genomic and transcriptomic signatures provide additive prognostic values for high-grade serous ovarian cancer patients. (**A**, Up) Principal component analyses (PCA) applied on transcriptomic data from the TCGA cohort using Verhaak's signature. On the left panel, the color code shows the non-Fibrosis (blue, *N* = 326) and Fibrosis (red, *N* = 220) subgroups, using Mateescu's signature (Mateescu et al., 2011). The right panel shows the same PCA representation but subgroups are highlighted using the LST genomic signature (Popova et al., 2012). The color code represents LST^{Lo} (blue, *N* = 238) and LST^{Hi} (green, *N* = 303) ovarian cancers. (Down) Contingency tables showing the repartition of patients regarding Mateescu or LST classification and the repartition against the two first principal components. **(B)** Kaplan-Meier curves showing 10-year overall survival (OS, Left) and disease-free survival (DFS, Right), after stratification into four groups: LST^{Lo}/Fibrosis (red, *N* = 89), LST^{Lo}/non-Fibrosis (green, *N* = 147), LST^{Hi}/Fibrosis (blue, *N* = 127), and LST^{Hi}/non-Fibrosis (black, *N* = 171). *P*-values are calculated using the Log-rank test.

Konecny et al., 2014). The lack of overlap could be explained, at least in part, by the heterogeneity in the techniques and platforms used for detecting gene expression, and by the diversity of unsupervised algorithms applied to molecular classifications. There is a clearer consensus of molecular classifications in breast cancer (Perou et al., 2000; Sorlie et al., 2001; Coates et al., 2015) that could be due to the presence of confirmed biomarkers (for example, hormonal receptors, and HER2 expression).

TABLE 5 | Pairwise comparison of transcriptomic and genomic signatures for overall and disease-free survival.

	LST^{Hi}/Fibrosis	LST^{Lo}/Fibrosis	LST^{Hi}/Non-Fibrosis	LST^{Lo}/Non-Fibrosis
Overall Survival				
LST^{Hi}/Fibrosis	–			
LST^{Lo}/Fibrosis	$1.60 \times 10^{-0.5}$	–		
LST^{Hi}/Non-Fibrosis	0.03	1.00×10^{-11}	–	
LST^{Lo}/Non-Fibrosis	0.04	0.03	1.10×10^{-05}	–
Disease Free Survival				
LST^{Hi}/Fibrosis	–			
LST^{Lo}/Fibrosis	0.03	–		
LST^{Hi}/Non-Fibrosis	0.03	$2.60 \times 10^{-0.6}$	–	
LST^{Lo}/Non-Fibrosis	0.31	0.14	$7.1 \times 10^{-0.4}$	–

Differences in overall survival (Top) and disease-free survival (Bottom) between the four groups: LST^{Hi}/Fibrosis, LST^{Lo}/Fibrosis, LST^{Hi}/Non-Fibrosis and LST^{Lo}/Non-Fibrosis. P-values are calculated using the Log-rank test and corrected for multiple testing using the Benjamini-Hochberg procedure.

The lack of consistency found in ovarian cancer classifications highlights the importance of using appropriate methods for stratifying high-grade serous ovarian cancer patients. Here, we demonstrate that among the four transcriptomic signatures analyzed (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Mateescu et al., 2011; Bentink et al., 2012; Verhaak et al., 2013), patient stratification into two subgroups, defined as the Fibrosis/non-Fibrosis signature (Mateescu et al., 2011), exhibits the most reliable prognostic value for patient survival compared to the others. We did observe a significant overlap in patient classification by applying the different transcriptomic signatures analyzed, but we also detected some differences between classifications. Indeed, Mesenchymal patients defined by the D-I-M-P signature (Verhaak et al., 2013) were all identified as Fibrosis using Mateescu's signature (Mateescu et al., 2011). In contrast, some patients defined as non-Mesenchymal by the D-I-M-P signature were defined as Fibrosis using Mateescu's signature, and they also exhibited poor survival. Based on the survival-data analyses, these observations suggest that some non-Mesenchymal patients should be considered Mesenchymal, as determined by the Fibrosis signature. This could also be explained by the non-exclusive attribution to a subtype using the D-I-M-P signature (40% of the tumor samples could be assigned to two distinct subtypes in Konecny's study) (Konecny et al., 2014) and/or by the spatial heterogeneity of signatures caused by the different geographic areas of sampling. Importantly, classifications tested in these studies (Tothill et al., 2008; Cancer Genome and Atlas Research, 2011; Bentink et al., 2012; Verhaak et al., 2013) were defined using a similar methodology (non-supervised analysis), but the Fibrosis/non-Fibrosis signature was identified through mechanistic studies based on miR-200-dependent profiling (Mateescu et al., 2011; Batista et al., 2016). This may explain the heterogeneity seen between our signature and others. In addition, our observations indicated that expression of miR-200 family members, either separately or combined, was not sufficient to predict patient survival. There has been a long-lasting controversy about the prognostic value of miR-200 with a number of studies displaying divergent results (Batista et al., 2013; Muralidhar and Barbolina, 2015). Recently, a meta-analysis including 7 articles with available data (553 patients) was conducted (Shi and Zhang, 2016). It is important to note that the populations included in those studies were quite small (from 55 to 100 patients) compared to the TCGA cohort studied here (557 patients). In that meta-analysis, higher expression of the miR-200 family was significantly associated with improved survival, predominantly due to the impact of miR-200c. This association was stronger in the Asian population. The discrepancies between this meta-analysis and our findings may be due to several reasons: inclusion of less Asian patients in the TCGA cohort, multiple small studies using different microarray protocols and significant heterogeneity across studies in the meta-analysis. This indicates that the prognostic value of using expression of the miR-200 family lacks reliability. Nonetheless, circulating miR-200s could still be good indicators for early detection of ovarian cancers or dynamic markers to follow-up during chemotherapy, as suggested in previous studies (Taylor and Gercel-Taylor, 2008; Kan et al., 2012; Sarojini et al., 2012; Kapetanakis et al., 2015; Pendlebury et al., 2017).

In addition to transcriptomic data, we have here provided new insight into genomic signatures of ovarian cancers. LST, defined as chromosomal breaks between adjacent regions of at least 10 Mb, constitute a robust indicator of HRD status (Popova et al., 2012; Goundiam et al., 2015). This classification was initially defined in breast cancers (Popova et al., 2012). Triple-negative breast carcinomas and high-grade serous ovarian cancers have some genomic instability patterns in common, providing a strong rationale for applying this LST signature on ovarian cancers. We and others have shown the impact of HRD on favorable response to platinum salts and overall survival (Fong et al., 2009, 2010; Audeh et al., 2010; Popova et al., 2012; Goundiam et al., 2015; Manie et al., 2016). Here, we confirm the clear prognostic value of the LST signature in high-grade serous ovarian cancers with better survival demonstrated for LST^{Hi} patients. Moreover, the interest for this classification will probably increase with the inclusion of PARP-inhibitors in routine clinical practice. Currently, the same therapeutic strategy, a combination of platinum and taxane-based chemotherapy, is used for all patients suffering from high-grade ovarian cancers. In the last decade, anti-angiogenic therapies and PARP-inhibitors were approved for

treatment of high-grade ovarian cancers, with a significant but limited impact on survival. This benefit on survival may be hidden by the molecular heterogeneity in tumors that drives either beneficial or deleterious response to treatments. Recent findings suggest that transcriptomic signatures could help in the identification of patients who will benefit from anti-angiogenic therapies (Gourley et al., 2014; Kommoss et al., 2017). In that context, we propose stratification of ovarian cancer patients that could help identify different sensitivity to treatment. The duality of our signature considering both the genomic HRD profile (LST signature) and the transcriptomic microenvironment features (Fibrosis/non-Fibrosis signature) provides compelling data for new therapies targeting the microenvironment (Thibault et al., 2014). There is a tendency to limit reimbursement of expansive new therapies if there is no biomarker predicting treatment response. We provide a reliable method to identify and subgroup high-grade serous ovarian cancer patients by combining genomic and transcriptomic information. Thus, our proposition of stratification could be used as a biomarker for some therapies that may help clinicians define the most appropriate therapeutic strategy.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE26193 and GSE9899, Clinical characteristics of the 557 patients included in the TCGA cohort, as well as transcriptomic data generated using Affymetrix Human Genome U133A arrays, have been previously described in Cancer Genome and Atlas Research (2011) and can be downloaded from the NIH Genomic Data Commons (GDC) data portal (https://gdc-portal.nci.nih.gov).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Institutional Review Board and Ethics committee of the Institut Curie Hospital Group approved all analyses realized in this study. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

FM-G and YK participated in the conception and design of the study. YK performed bioinformatic and statistical analyses of the data, with participation from CB. TP and M-HS defined the LST status of high-grade serous ovarian cancers. RR provided human samples from the Curie cohort. FM-G supervised the entire project and wrote the manuscript with YK and CB.

FUNDING

The results presented here are, in part, based upon data generated by the TCGA Research Network. YK was supported by funding from the Foundation pour la Recherche Medicale (FRM, ING20130526797), the SiRIC-Curie program (INCa-DGOS-4654), and CB by the Institut National de la Santé et de la Recherche Médicale (Inserm, Poste d'aide à la Recherche Translationnelle en Cancérologie). The laboratory has also received grants from Inserm, Institut Curie, and the Ligue Nationale Contre le Cancer (Labelisation), the Institut National du Cancer (INCa-DGOS-9963), the Foundation ARC (PJA 20151203364), and the ICGex (ANR-10-EQPX-03). We are very grateful to our funders for providing support over the years.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00219/full#supplementary-material

REFERENCES

Audeh, M. W., Carmichael, J., Penson, R. T., Friedlander, M., Powell, B., Bell-McGuinn, K. M., et al. (2010). Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet* 376, 245–251. doi: 10.1016/S0140-6736(10)60893-8

Batista, L., Bourachot, B., Mateescu, B., Reyal, F., and Mechta-Grigoriou, F. (2016). Regulation of miR-200c/141 expression by intergenic DNA-looping and transcriptional read-through. *Nat. Commun.* 7:8959. doi: 10.1038/ncomms9959

Batista, L., Gruosso, T., and Mechta-Grigoriou, F. (2013). Ovarian cancer emerging subtypes: role of oxidative stress and fibrosis in tumour development and response to treatment. *Int. J. Biochem. Cell Biol.* 45, 1092–1098. doi: 10.1016/j.biocel.2013.03.001

Bentink, S., Haibe-Kains, B., Risch, T., Fan, J. B., Hirsch, M. S., Holton, K., et al. (2012). Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PLoS ONE* 7:e30269. doi: 10.1371/journal.pone.0030269

Berns, E. M., and Bowtell, D. D. (2012). The changing view of high-grade serous ovarian cancer. *Cancer Res.* 72, 2701–2704. doi: 10.1158/0008-5472.CAN-11-3911

Brozovic, A., Duran, G. E., Wang, Y. C., Francisco, E. B., and Sikic, B. I. (2015). The miR-200 family differentially regulates sensitivity to paclitaxel and carboplatin in human ovarian carcinoma OVCAR-3 and MES-OV cells. *Mol. Oncol.* 9, 1678–1693. doi: 10.1016/j.molonc.2015.04.015

Cancer Genome and Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Coates, A. S., Winer, E. P., Goldhirsch, A., Gelber, R. D., Gnant, M., Piccart-Gebhart, M., et al. (2015). Tailoring therapies–improving the management of early breast cancer: St Gallen international expert consensus on the primary therapy of early breast cancer 2015. *Ann. Oncol.* 26, 1533–1546. doi: 10.1093/annonc/mdv221

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., et al. (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.* 361, 123–134. doi: 10.1056/NEJMoa0900212

Fong, P. C., Yap, T. A., Boss, D. S., Carden, C. P., Mergui-Roelvink, M., Gourley, C., et al. (2010). Poly(ADP)-ribose polymerase inhibition: frequent durable

responses in BRCA carrier ovarian cancer correlating with platinum-free interval. *J. Clin. Oncol.* 28, 2512–2519. doi: 10.1200/JCO.2009.26.9589

Gelmon, K. A., Tischkowitz, M., Mackay, H., Swenerton, K., Robidoux, A., Tonkin, K., et al. (2011). Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. *Lancet Oncol.* 12, 852–861. doi: 10.1016/S1470-2045(11)70214-5

Goundiam, O., Gestraud, P., Popova, T., De la Motte Rouge, T., Fourchotte, V., Gentien, D., et al. (2015). Histo-genomic stratification reveals the frequent amplification/overexpression of CCNE1 and BRD4 genes in non-BRCAness high grade ovarian carcinoma. *Int. J. Cancer* 137, 1890–1900. doi: 10.1002/ijc.29568

Gourley, C., McCavigan, A., Perren, T., Paul, J., Michie, C. O., Churcman, M., et al. (2014). Molecular subgroup of high-grade serous ovarian cancer (HGSOC) as a predictor of outcome following bevacizumab. *J. Clin. Oncol.* 32(15 Suppl. 5502), 5502–5502. doi: 10.1200/jco.2014.32.15_suppl.5502

Kan, C. W., Hahn, M. A., Gard, G. B., Maidens, J., Huh, J. Y., Marsh, D. J., et al. (2012). Elevated levels of circulating microRNA-200 family members correlate with serous epithelial ovarian cancer. *BMC Cancer* 12:627. doi: 10.1186/1471-2407-12-627

Kapetanakis, N. I., Uzan, C., Jimenez-Pailhes, A. S., Gouy, S., Bentivegna, E., Morice, P., et al. (2015). Plasma miR-200b in ovarian carcinoma patients: distinct pattern of pre/post-treatment variation compared to CA-125 and potential for prediction of progression-free survival. *Oncotarget* 6, 36815–36824. doi: 10.18632/oncotarget.5766

Kaye, S. B., Lubinski, J., Matulonis, U., Ang, J. E., Gourley, C., Karlan, B. Y., et al. (2012). Phase II, open-label, randomized, multicenter study comparing the efficacy and safety of olaparib, a poly (ADP-ribose) polymerase inhibitor, and pegylated liposomal doxorubicin in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer. *J. Clin. Oncol.* 30, 372–379. doi: 10.1200/JCO.2011.36.9215

Kmietowicz, Z. (2015). NICE rejects trastuzumab emtansine for use on NHS. *BMJ* 351:h6837. doi: 10.1136/bmj.h6837

Kommoss, S., Winterhoff, B., Oberg, A. L., Konecny, G. E., Wang, C., Riska, S. M., et al. (2017). Bevacizumab may differentially improve ovarian cancer outcome in patients with proliferative and mesenchymal molecular subtypes. *Clin. Cancer Res.* 23, 3794–3801. doi: 10.1158/1078-0432.CCR-16-2196

Konecny, G. E., and Kristeleit, R. S. (2016). PARP inhibitors for BRCA1/2-mutated and sporadic ovarian cancer: current practice and future directions. *Br. J. Cancer* 115, 1157–1173. doi: 10.1038/bjc.2016.311

Konecny, G. E., Wang, C., Hamidi, H., Winterhoff, B., Kalli, K. R., Dering, J., et al. (2014). Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *J. Natl. Cancer Inst.* 106:dju249. doi: 10.1093/jnci/dju249

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., et al. (2012). Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *N. Engl. J. Med.* 366, 1382–1392. doi: 10.1056/NEJMoa1105535

Leskela, S., Leandro-Garcia, L. J., Mendiola, M., Barriuso, J., Inglada-Perez, L., Munoz, I., et al. (2011). The miR-200 family controls beta-tubulin III expression and is associated with paclitaxel-based treatment response and progression-free survival in ovarian cancer patients. *Endocr. Relat. Cancer* 18, 85–95. doi: 10.1677/ERC-10-0148

Liu, J. F., Barry, W. T., Birrer, M., Lee, J. M., Buckanovich, R. J., Fleming, G. F., et al. (2014). Combination cediranib and olaparib versus olaparib alone for women with recurrent platinum-sensitive ovarian cancer: a randomised phase 2 study. *Lancet Oncol.* 15, 1207–1214. doi: 10.1016/S1470-2045(14)70391-2

Manie, E., Popova, T., Battistella, A., Tarabeux, J., Caux-Moncoutier, V., Golmard, L., et al. (2016). Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas. *Int. J. Cancer* 138, 891–900. doi: 10.1002/ijc.29829

Mateescu, B., Batista, L., Cardon, M., Gruosso, T., de Feraudy, Y., Mariani, O., et al. (2011). miR-141 and miR-200a act on ovarian tumorigenesis by controlling oxidative stress response. *Nat. Med.* 17, 1627–1635. doi: 10.1038/nm.2512

McLachlan, J., George, A., and Banerjee, S. (2016). The current status of PARP inhibitors in ovarian cancer. *Tumori* 102, 433–440. doi: 10.5301/tj.5000558

Miller, R. E., and Ledermann, J. A. (2016). The status of poly(adenosine diphosphate-ribose) polymerase (PARP) inhibitors in ovarian cancer, part 2: extending the scope beyond olaparib and BRCA1/2 mutations. *Clin. Adv. Hematol. Oncol.* 14, 704–711.

Monk, B. J., Minion, L. E., and Coleman, R. L. (2016). Anti-angiogenic agents in ovarian cancer: past, present, and future. *Ann. Oncol.* 27(Suppl. 1), i33–i39. doi: 10.1093/annonc/mdw093

Muggia, F., and Safra, T. (2014). 'BRCAness' and its implications for platinum action in gynecologic cancer. *Anticancer. Res.* 34, 551–556.

Muralidhar, G. G., and Barbolina, M. V. (2015). The miR-200 family: versatile players in epithelial ovarian cancer. *Int. J. Mol. Sci.* 16, 16833–16847. doi: 10.3390/ijms160816833

Oza, A. M., Cibula, D., Benzaquen, A. O., Poole, C., Mathijssen, R. H., Sonke, G. S., et al. (2015). Olaparib combined with chemotherapy for recurrent platinum-sensitive ovarian cancer: a randomised phase 2 trial. *Lancet Oncol.* 16, 87–97. doi: 10.1016/S1470-2045(14)71135-0

Pendlebury, A., Hannan, N. J., Binder, N., Beard, S., McGauran, M., Grant, P., et al. (2017). The circulating microRNA-200 family in whole blood are potential biomarkers for high-grade serous epithelial ovarian cancer. *Biomed. Rep.* 6, 319–322. doi: 10.3892/br.2017.847

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Popova, T., Manie, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., et al. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.* 72, 5454–5462. doi: 10.1158/0008-5472.CAN-12-1470

Popova, T., Manie, E., Stoppa-Lyonnet, D., Rigaill, G., Barillot, E., and Stern, M. H. (2009). Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol.* 10:R128. doi: 10.1186/gb-2009-10-11-r128

Pujade-Lauraine, E., Ledermann, J. A., Selle, F., Gebski, V., Penson, R. T., Oza, A. M., et al. (2017). Olaparib tablets as maintenance therapy in patients with platinum-sensitive, relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* 18, 1274–1284. doi: 10.1016/S1470-2045(17)30469-2

Raja, F. A., Hook, J. M., and Ledermann, J. A. (2012). Biomarkers in the development of anti-angiogenic therapies for ovarian cancer. *Cancer Treat. Rev.* 38, 662–672. doi: 10.1016/j.ctrv.2011.11.009

Rigakos, G., and Razis, E. (2012). BRCAness: finding the Achilles heel in ovarian cancer. *Oncologist* 17, 956–962. doi: 10.1634/theoncologist.2012-0028

Sabatier, R., Finetti, P., Bonensea, J., Jacquemier, J., Adelaide, J., Lambaudie, E., et al. (2011). A seven-gene prognostic model for platinum-treated ovarian carcinomas. *Br. J. Cancer* 105, 304–311. doi: 10.1038/bjc.2011.219

Sarojini, S., Tamir, A., Lim, H., Li, S., Zhang, S., Goy, A., et al. (2012). Early detection biomarkers for ovarian cancer. *J. Oncol.* 2012:709049. doi: 10.1155/2012/709049

Shi, C., and Zhang, Z. (2016). The prognostic value of the miR-200 family in ovarian cancer: a meta-analysis. *Acta Obstet. Gynecol. Scand.* 95, 505–512. doi: 10.1111/aogs.12883

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.191367098

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* 6:e21800. doi: 10.1371/journal.pone.0021800

Taylor, D. D., and Gercel-Taylor, C. (2008). MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. *Gynecol. Oncol.* 110, 13–21. doi: 10.1016/j.ygyno.2008.04.033

The Lancet (2017). Trastuzumab emtansine and cost-based decision making. *Lancet* 389:2. doi: 10.1016/S0140-6736(17)30006-5

Thibault, B., Castells, M., Delord, J. P., and Couderc, B. (2014). Ovarian cancer microenvironment: implications for cancer dissemination and chemoresistance

acquisition. *Cancer Metastasis Rev.* 33, 17–39. doi: 10.1007/s10555-013-9456-2

Tothill, R. W., Tinker, A. V., George, J., Brown, R., Fox, S. B., Lade, S., et al. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* 14, 5198–5208. doi: 10.1158/1078-0432.CCR-08-0196

Tutt, A., Robson, M., Garber, J. E., Domchek, S. M., Audeh, M. W., Weitzel, J. N., et al. (2010). Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and advanced breast cancer: a proof-of-concept trial. *Lancet* 376, 235–244. doi: 10.1016/S0140-6736(10)60892-6

Verhaak, R. G., Tamayo, P., Yang, J. Y., Hubbard, D., Zhang, H., Creighton, C. J., et al. (2013). Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J. Clin. Invest.* 123, 517–525. doi: 10.1172/JCI65833

Conflict of Interest: TP and M-HS are named inventors of a patent for the genomic signature of BRCAness. Current exploitation of the patent is on-going by Myriad Genetics.

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

frontiers in Genetics

METHODS
published: 03 April 2020
doi: 10.3389/fgene.2020.00275

Identification of Cancerlectins Using Support Vector Machines With Fusion of G-Gap Dipeptide

*Lili Qian†, Yaping Wen† and Guosheng Han**

Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan, China

The cancerlectin plays an important role in the initiation, survival, growth, metastasis, and spread of cancer. Therefore, to study the function of cancerlectin is greatly significant because it can help to identify tumor markers and tumor prevention, treatment, and prognosis. However, plenty of studies have generated a large amount of protein data. Traditional prediction methods have been unable to meet the needs of analysis. Developing powerful computational models based on these data to discriminate cancerlectins and non-cancerlectins on a large scale has been treated as one of the most important topics. In this study, we developed a feature extraction method to identify cancerlectins based on fusion of g-gap dipeptides. The analysis of variance was used to select the optimal feature set and a support vector machine was used to classify the data. The rigorous nested 10-fold cross-validation results, demonstrated that our method obtained the prediction accuracy of 83.91% and sensitivity of 83.15%. At the same time, in order to evaluate the performance of the classification model constructed in this work, we constructed a new data set. The prediction accuracy of the new data set reaches 83.3%. Experimental results show that the performance of our method is better than the state-of-the-art methods.

Keywords: cancerlectins, g-gap dipeptide, feature selection, analysis of variance, support vector machine

OPEN ACCESS

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

Reviewed by:
Yuansheng Liu,
University of Technology Sydney, Australia

***Correspondence:**
Guosheng Han
hangs@xtu.edu.cn

†These authors have contributed equally to this work

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

Received: *25 November 2019*
Accepted: *06 March 2020*
Published: *03 April 2020*

Citation:
Qian L, Wen Y and Han G (2020) Identification of Cancerlectins Using Support Vector Machines With Fusion of G-Gap Dipeptide. Front. Genet. 11:275. doi: 10.3389/fgene.2020.00275

INTRODUCTION

Cell recognition is the central event of various biological phenomena. The combination of cell surface molecular selectivity with other molecules is an important link in cell development and differentiation, such as fertilization, embryogenesis, immune defense, pathogen infection, and pathogenicity. Abnormal cell recognition may lead to diseases, such as defects in leukocyte and platelet adhesion, which can lead to the recurrence of bacterial infections and mucosal bleeding, respectively. In addition, abnormal cell recognition is considered to be the basis of uncontrolled cell growth and movement, which is the characteristic of tumor transformation and metastasis (Sharon and Lis, 1989).

Lectin is one of the cell recognition molecules. It is a biological molecule that specifically recognizes and binds the carbohydrate components existing in other proteins (Kumar and Panwar, 2011). Most lectins have high specificity and selectivity in identifying sugar molecules present in other proteins (Lis and Sharon, 1998). According to their affinity with monosaccharides, these glycoproteins can be divided into five categories: mannose, N-acetylglucosamine, galactose/N-acetylgalactosamine, fucose, and sialic acid, which represent a group of heterogeneous oligomeric

proteins (Kumar and Panwar, 2011). It has been found that lectins are involved to a variety of biological processes, such as maintaining the dynamic balance of cell proliferation and apoptosis, cell differentiation, cell adhesion and migration, cell-extracellular matrix interaction, host-pathogen interaction, cell-cell recognition, complement activation pathway, immune defense, and regulation of inflammatory response (Lin et al., 2015). Lectin molecules provide biological scripts to decipher complex codes in sugar groups (Damodaran et al., 2008). Therefore, lectins are often used as diagnostic and therapeutic tools in many fields such as cell biology, biochemistry, and immunology.

Cancerlectins are a group of lectins which are closely related to cancer (Kumar and Panwar, 2011). Lectin participates in serum-glycoprotein transformation and innate immune response, and has a special correlation with the growth and metastasis of tumors (Damodaran et al., 2008). Some evidences suggest that tumor cell agglutinin is involved in cell interactions, such as adhesion, cell growth, differentiation, metastasis and infection of cancer cells (Lis and Sharon, 1998). Whether basic research or clinical application, cancerlectins has been widely used in cancer research (Lai et al., 2017). For example, sialic acid-bound immunoglobulin lectin-9 is a neutrophil-specific expression that binds to sugar molecules on the surface of cancer cells, regulates immune response and promotes or inhibits tumor progression; spiral hemagglutinin is an effective prognostic indicator of colorectal cancer, etc. (Kumar and Panwar, 2011). The effect of lectins on the immune system by altering the production of various interleukins has been well-documented. There is also data showing that some lectins down-regulate the activity of telomere, thereby inhibiting angiogenesis (Choi et al., 2004; De Mejía and Prisecaru, 2005). Cancerlectins can induce cytotoxicity, apoptosis, and inhibit tumor growth by binding to receptors on the surface of cancer cells. It can be used as a therapeutic method for cancer treatment. Cancer is the second leading cause of death in the world. Therefore, the screening of specific lectins from a large number of lectins is of great significance not only for the discovery of tumor markers and cancer treatment, but also for better understanding and conquering cancer (Balachandran et al., 2017).

A plenty of studies have generated a large amount of protein data, using traditional biological experiments to predict and analyze the function of proteins is not only time-consuming but also laborious. Based on these data, it is one of the most important topics to predict a cancerous substance by establishing a powerful computational model to identify cancerous and non-cancerous substances on a large scale. The description of the characteristics of the protein sequence method contains a lot of information, such as the chemical and physical properties of amino acids, sequence characteristics, feature extraction algorithm for classification algorithm which has great impact on the design and the classification of results. Too few protein sequence characteristics will result in the loss of important information of protein sequence and affect the classification results, and therefore dimension disaster, conversely, there is no guarantee of the classification efficiency of the model. Therefore, how to conduct efficient feature fusion and establish appropriate mathematical expression methods and similarity measurement standards is an important problem.

Feature Extraction Based on Sequence Information

Nakashima et al. (1986) proposed amino acid composition to study protein folding. One of the most basic algorithms for extracting features of protein sequence is amino acid composition, which represents the occurrence frequency of each of the 20 common amino acids in the protein sequence and converts the protein sequence into a 20-dimensional feature vector. Yu et al. (2004) proposed using k peptide component information to represent protein sequences. Feng et al. (2013) proposed a Naïve Bayes-based method to predict antioxidant proteins using amino acid compositions and dipeptide compositions.

Feature Extraction Based on Physical and Chemical Properties of Amino Acids

Bu et al. (1999) proposed an autocorrelation function algorithm, which is a description method based on Amino Acid Residue Index (Kawashima et al., 1999), for the study of protein structure predetermination. Chou (2001) proposed the pseudo-amino acid composition method, including sequence order information other than amino acid composition.

Feature Extraction Based on Protein Evolution Information

Evolutionary information is one of the most important information of protein functional annotation in biological analysis, reflecting the sequence conservation of amino acids at each site of protein sequence in the evolutionary process (Xu et al., 2015). Evolutionary information of proteins mainly relies on positional specificity score matrix (PSSM) (An et al., 2016).

In the published research work, Kumar and Panwar (Kumar and Panwar, 2011) integrated PROSITE domain information with PSSM, developed a support vector machine model, and obtained MCC value of 0.38 with an accuracy of 69.09%; Lin et al. (2015) developed a sequence-based method to distinguish cancerlectins from non-cancerlectins, and used ANOVA to select the optimal feature subset. The accuracy of the method is 75.19%; Zhang et al. (2016) proposed a classification model based on random forest, the accuracy of the method is 70%; Lai et al. (2017) proposed a new method of feature expression based on amino acid sequence, and binomized it. In the jackknife cross-validation, the accuracy is 77.48%. Han et al. (2014) proposed a two-stage multi-class support vector machine combined with a two-step optimal feature selection process for predicting membrane protein types. Anh et al. (2014) propose a kernel method, named as SSEAKSVM, predicting protein structural classes for low-homology data sets based on predicted secondary structures. Balachandran et al. (2018) proposed a support vector machine (SVM)-based PVP predictor, called PVP-SVM, which was trained with 136 optimal features. Runtao et al. (2018) proposed a computational method based on the RF (Random Forest) algorithm for identifying cancerlectins, and achieves a

sensitivity of 0.779, a specificity of 0.717, an accuracy of 0.748. These methods have obtained quite good results, but the accuracy still needs to be improved. In this work, we constructed a new classification system of protein sequences, and the relatively better result was obtained on the benchmark dataset and the independent test dataset.

METHODS

Dataset

Data acquisition is the first step of data analysis. The benchmark dataset is not only the database of algorithm learning, but also the cornerstone of classification model. Constructing a good benchmark data set also plays an important role in the performance of classification model (Lin and Chen, 2010). In order to compare objectively with the existing research results, the dataset used in this work was widely used which was constructed by Kumar and Panwar (Kumar and Panwar, 2011).

The benchmark dataset contains both positive and negative samples. The original data were downloaded from the CancerLectinDB database (Damodaran et al., 2008), removing duplicated sequences and sequences without experimental evidence, or containing non-standard amino acids, and 385 proteins were obtained to form a positive subset (Lin et al., 2015). Using the keyword "lectins" search in UniProt database, deleting the sequences labeled "similarity," "fragment," "hypothesis," and "possibility," a negative subset containing 820 proteins was constructed (Kumar and Panwar, 2011; Lin et al., 2015). If the designed data sets contain highly similar sequences, misleading results with high prediction accuracy will be obtained, thus reducing the generalization ability of the model. In order to remove homologous sequences from the benchmark dataset, the CD-HIT program was employed with 50% as the sequence identity cutoff to exclude any protein/peptide sequences with more than 50% paired sequence in the benchmark dataset (Lin et al., 2015). The benchmark datasetScan be formulated as follows:

$$S = S_+ \cup S_-$$

where the positive subset S_+ contains 178 cancerlectin samples, the negative subset S_contains 226 non-cancerlectin samples, thus, the benchmark dataset Scontains 404 samples. The benchmark dataset is available at https://github.com/hangslab/cancerlectins.

Feature Extraction Method

When using the machine learning method, protein sequences need to be transformed into numerical vectors representing the characteristics of protein sequence. The extracted features need not only to retain the sequence information of proteins to the greatest extent, but also to have a greater correlation with protein classification.

The sequence of amino acids in protein sequence is the basis of protein biological function. The dipeptide composition is the condition of $k = 2$ in the feature extraction method of k-peptide composition (Yu et al., 2004; Lin and Chen, 2010). The dipeptide composition can only reflect the correlation of adjacent amino acids in protein sequence. Generally speaking, the intrinsic properties of protein sequences may be precipitated in higher-level residue relationships. In the tertiary structure of proteins, the two amino acids separated from the original sequence may be very close in space, which means that the g-gap dipeptide composition (Sharma and Paliwal, 2008; Lin et al., 2015) contains more information about protein sequences than the dipeptide composition. In this paper, we developed a feature extraction method of fusion g-gap dipeptide component, **Figure 1** is the flow chart of the model construction.

The g-gap dipeptide composition transforms each protein sequence into a feature vector. For each g value, a 400-dimensional feature vector (20*20) will be generated. The range of g is [0,9]. $g = g_h, g_h = h, h \in [0, 9]$ is used to distinguish the frequency of g-gap dipeptides with different values of g. We transformed a cancerlectin or non-cancerlectin protein sample P with L amino acids into an input vector of 4,000 dimensions, defined as follows:

$$F_{4000} = \left[f_1^0, \cdots, f_{400}^0, f_1^1, \cdots f_{400}^1, \cdots, f_u^{g_h}, \cdots, f_1^9, \cdots, f_{400}^9 \right]^T$$

where the $f_u^{g_h}$ is the frequency of the u-th ($u = 1, 2, \cdots, 400$) g_h-gap dipeptide and calculated by

$$f_u^{g_h} = \frac{n_u^{g_h}}{\sum_{u=1}^{400} n_u^{g_h}}$$

where $n_u^{g_h}$ denote the number of the u-th g_h-gap dipeptide in a protein. Note that when $g = 0$, the g-gap dipeptide will degenerate to the adjoining dipeptide composition.

The class labels corresponding to each feature vector are represented by t, $t \in \{0, 1\}$,1 represents positive sample and 0 represents negative samples. Finally, a 404*4,000 feature matrix was obtained.

Feature Selection

When the number of features is large, there may be unrelated features, or interdependence between features, which easily leads to the time-consuming process of analyzing features and training models. The more the number of features, the more likely it is to cause "dimension disaster," the more complex the model will be, and its generalization ability will decline. Feature selection can eliminate irrelevant or redundant features, reduce the number of features, improve the accuracy of the model and reduce the running time. On the other hand, the model is simplified by selecting truly relevant features, which makes it easy for researchers to understand the process of data generation.

Influenced by the collinearity of sample features, the results of linear discriminant analysis are poor (Lin et al., 2013), and the use of binomial distribution will lead to a high-dimensional feature vector (Yanyuan et al., 2018), which consumes a lot of computing time and may lead to over-fitting. After comparison, the feature selection method used in this paper is variance analysis (Lin

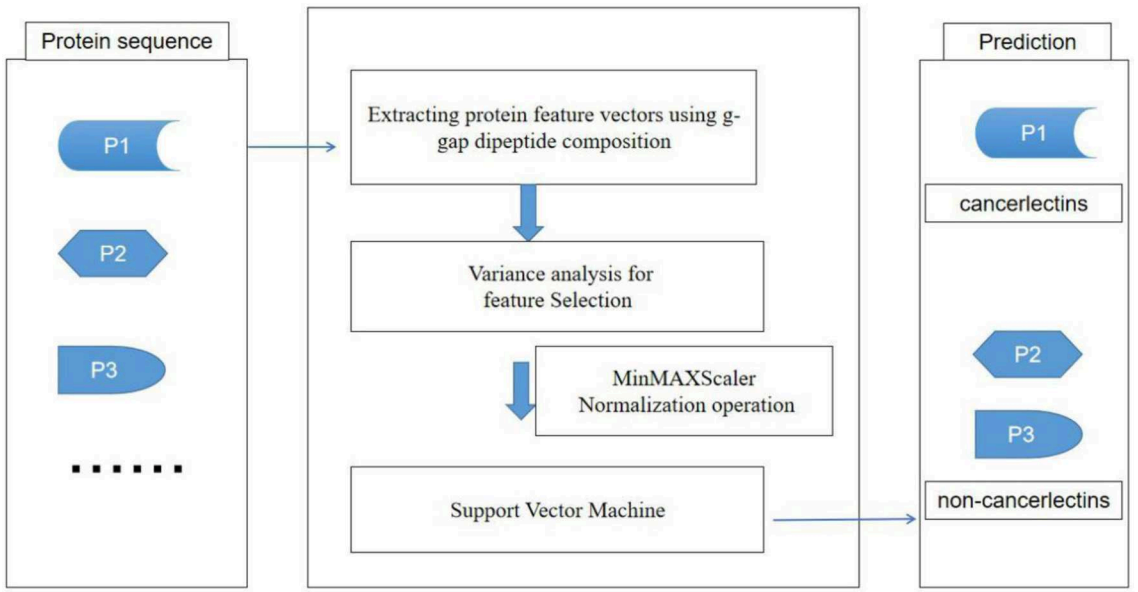


FIGURE 1 | The flowchart of our method.

et al., 2015). The variance analysis decomposes the difference of samples at the level of known influencing factors into intra-group variance and inter-group variance. The intra-group variance is not affected by the level of influencing factors, but mainly sampling error. The variance between groups is influenced by the level of factors, which is the essential difference between samples. The characteristic variance is measured by calculating the ratio F of variance between feature groups and variance within the group. The *F*-value of the *u*-th feature in the benchmark dataset is defined as follows:

$$F(u) = \frac{S_A^2(u)}{S_E^2(u)}$$

where $S_A^2(u)$ is the sample variance between groups, $S_E^2(u)$ is the sample variance within groups. They are given by:

$$\begin{cases} S_A^2(u) = \frac{SS_A(u)}{df_A} \\ S_E^2(u) = \frac{SS_E(u)}{df_E} \end{cases}$$

where $SS_A(u)$ is sum of squares between groups and $SS_E(u)$ is sum of squares within groups, which can be calculated by:

$$\begin{cases} SS_A(u) = \sum_{i=1}^{K} m_i \left(\frac{\sum_{j=1}^{m_i} f_u^{g_h}(i,j)}{m_i} - \frac{\sum_{i=1}^{K}\sum_{j=1}^{m_i} f_u^{g_h}(i,j)}{\sum_{i=1}^{K} m_i} \right)^2 \\ SS_E(u) = \sum_{i=1}^{K} \sum_{j=1}^{m_i} \left(f_u^{g_h}(i,j) - \frac{\sum_{j=1}^{m_i} f_u^{g_h}(i,j)}{m_i} \right)^2 \end{cases}$$

where $f_u^{g_h}(i,j)$ is the frequency of the *u*-th g_h-gap dipeptide of the *j*-th sample in the *i*-th group; $m_i$denotes the number of samples in the *i*-th group (here $m_1 = 178, m_2 = 226$).

df_A and df_E are degrees of freedom for the sample variance between groups and the sample variance within groups, respectively. They can be calculated by:

$$\begin{cases} df_A = K - 1 \\ df_E = N - K \end{cases}$$

where K and N are the number of groups ($K = 2$) and total number of samples ($N = 404$), respectively.

When $F < 1$, the smaller the F value is, the smaller the difference of the feature between the two groups is, the worse the ability of the feature to recognize two kinds of proteins is; when $F > 1$, the larger the F value is, the greater the difference of the feature between the two groups is, the better the ability of the feature to recognize proteins is. Each F value corresponds to a P-value. The larger the F-value is, the smaller the P-value, that is, the greater the difference of the feature between groups.

The larger the F value is, the better the discriminant ability of the feature is. Therefore, all features can be sorted according to their F values, and the number of optimal feature subsets can be determined by incremental feature selection. The first feature subset is the feature with the highest median value in ranking. When the second highest value is added, a new feature subset is generated. This process was repeated from the higher F to the lower F value until all candidate features were added, therefore, for each sample, 4,000 feature subsets will be generated. The ε-th feature subset is composed of ε ranked g_h-gap dipeptides and can be expressed as (Lin et al., 2015):

$$P_\varepsilon = \left[f_1^{g_h}, f_2^{g_h}, \cdots, f_\varepsilon^{g_h} \right]^T,\ 1 \le \varepsilon \le 4,000,\ 1 \le g_h \le 9$$

Normalization

In machine learning, normalization of feature data is an important step. Because the characteristic information of protein sequence transformation is dimensionless, data normalization is used to facilitate the comparison and weighting of indicators of different scales. The data normalization can improve the convergence speed and the prediction accuracy of the model. The data normalization method used in this paper is MinMAXScaler, which normalizes each feature into [0,1] interval. The normalization function as follows:

$$f_u^{gh^*} = \frac{f_u^{gh} - f_u^{gh}{}_{\min}}{f_u^{gh}{}_{\max} - f_u^{gh}{}_{\min}}$$

Support Vector Machine

In order to facilitate the comparison with the existing work, support vector machine (SVM) (Kumar and Panwar, 2011; Lin et al., 2015; Lai et al., 2017) is selected as the classifier in this work. The basic idea of SVM is to find an optimal classification hyperplane, which maximizes the interval between different types of samples. Kernel functions include linear and Gaussian kernels. In this paper, we use the radial basis function (RBF) (Cai et al., 2002; Yu et al., 2003; An et al., 2016). In this work, the parameters are tuned by the method of grid search-GridSearchCV (Liu et al., 2014). Grid search finds the optimal parameter combination by searching the specified parameter range exhaustively and gets the model performance results of each group of parameters combination. The search spaces for C is $[10^{-3}, 10^{4}]$. The search spaces for γ is $[10^{-4}, 10^{5}]$. Finally, the optimal combination of parameters $[C, \gamma]$ is [1,1].

Nested Cross-Validation Test

An important purpose of model validation is to select the most suitable model. A good model needs strong generalization ability to unknown data. This step of model validation can reflect the performance of different models for unknown data. In our method, we select the cross-validation model (Metfessel et al., 1993). Cross-checking divides the data set into two parts: training set and test set. Training set is used for model training, and test set is used to measure the prediction ability of the model. It can effectively prevent model over-fitting, and effectively evaluate the generalization ability of the model for data sets independent of training data.

Because the feature dimension in this paper is higher than 4,000, we chose nested cross-validation to prevent model overfitting. The samples are randomly divided into 10 equal and disjoint subsets in the external cycle of cross-validation. Nine of them are in turn selected as training sets, and one test subset is left, and then 10-fold cross-validation is carried out on the training set in the internal cycle. The internal loop performs feature selection and parameter optimization, and the external loop test set performs model performance evaluation. In nested cross-validation, the estimated true error is almost the same as the result obtained on the test set.

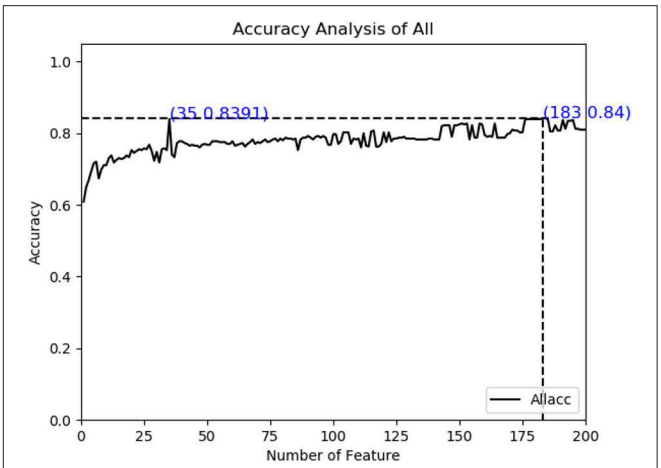


FIGURE 2 | Prediction accuracy curve of feature subset.

Performance Assessment

The following indicators are used to evaluate the classification performance of the model.

1. Accuracy: Correctly identify the proportion of samples in the total sample.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Sensitivity: The proportion of cancerlectins samples correctly identified as cancerlectins.

$$S_n = \frac{TP}{TP + FN}$$

3. Specificity: The proportion of non-cancerlectins samples correctly identified as non-cancerlectins.

$$S_p = \frac{TN}{TN + FP}$$

4. ROC curve

ROC curve is called "receiver operating characteristic curve". The ROC curve takes FPR as the horizontal axis and TPR as the vertical axis.

The area under the ROC curve is AUC. AUC value is between 0 and 1, and the closer the AUC value is to 1, the better the performance of the classifier is.

$$FPR = \frac{FP}{FP + TN}$$
$$TPR = \frac{TP}{TP + FN}$$

TABLE 1 | *F*-value and *P*-value of features in optimal feature subset.

	F-value	*P*-value
L_R1	26.76446	3.63E-07
R_L4	24.81686	9.38E-07
Q_E0	20.28248	8.77E-06
I_D0	16.70216	5.28E-05
N_K3	16.34925	6.32E-05
N_D6	15.78628	8.40E-05
Q_P9	15.52386	9.61E-05
I_D4	15.23462	0.000111
D_N0	14.73123	0.000144
P_A1	14.28921	0.000181
N_D1	14.13802	0.000195
P_L5	13.87658	0.000223
L_P7	13.82865	0.000229
S_N5	13.69697	0.000245
A_L2	13.26494	0.000306
A_R2	12.96445	0.000357
L_P5	12.90963	0.000367
R_Q3	12.90722	0.000368
L_R8	12.702	0.000409
N_D3	12.40946	0.000476
N_G8	12.37352	0.000485
D_N7	12.2193	0.000526
D_N8	12.09143	0.000562
L_C0	11.94945	0.000605
N_V1	11.87518	0.000629
E_L5	11.79776	0.000655
Q_P1	11.78632	0.000659
Q_A0	11.54244	0.000748
L_E6	11.50195	0.000764
R_P4	11.4276	0.000794
P_L6	11.23968	0.000877
Q_M7	11.22643	0.000883
D_G0	11.22351	0.000884
S_P2	11.17902	0.000905
Q_L1	11.06357	0.000961

where TP (True positive) and TN (True negative) denote the number of correctly predicted cancerlectins and the number of correctly predicted non-cancerlectins, respectively; FN is the number of the cancerlectins incorrectly predicted as the non-cancerlectins and FP is the number of the non-cancerlectins incorrectly predicted as the cancerlectins, respectively.

RESULTS

Prediction Performance

The protein sequence is represented by the fusion of g-gap dipeptide features. After feature transformation, all protein sequences are converted into a 404*4,000 feature matrix. After variance analysis, *F*-values of features are sorted in descending order, and then feature selection and parameter optimization are carried out in a nested cross validation.

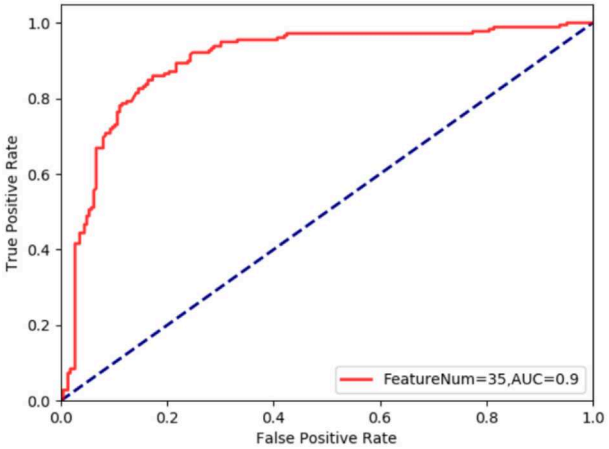


FIGURE 3 | The ROC curve for cancerlectin prediction using the optimal 35 g-gap dipeptide.

As described in the feature extraction section, each sample sequence is transformed into a 4,000-dimensional dipeptide vector. Using too many low variance features to train prediction models will be relatively time-consuming, and it is possible to build over-fitting models. On the contrary, if the number of characteristic peptides is too small, they can only describe some properties of cancerlectins, even though each property may have a high variance and contain extremely rich information. Both of these conditions will lead to poor prediction results. The total number of protein sequence samples in data sets is 404. In order to build a reliable robust model, the number and accuracy of features need to be considered simultaneously. From **Figure 2**, it can be seen that the accuracy of feature subset increases slowly after 35 dimensions, until the number of feature subsets increases to 183 dimensions, the accuracy of model has small change from the feature subset of 35 dimensions. The accuracy of the first 183- dimensional model is 84% and that of the first 35-dimensional model of feature subset is 83.91%. Finally, the top 35 g-gap dipeptides are selected. Therefore, 35 g-gap dipeptides are selected as the optimal feature subset of the final classifier.

Feature Description

As can be seen from **Table 1**, the variance of L_R1 is the largest, and the larger the variance, the smaller the *P*-value generally accompanied. The variance of L_R1 is 26.76446, *P*-value is 3.63E-07, Q_L1 variance is 11.06357, *P*-value is 0.000961. It can be seen that each feature in the optimal feature subset is significant and may play an important role in the classification and prediction of cancerlectins.

As can be seen from **Figure 3**, the AUC of cancerlectin prediction using the optimal 35 g-gap dipeptide is 0.9, it means the classification performance of this classification model is good.

Comparison With Existing Methods

In order to verify whether the classification model constructed in this work is over-fitting, 30 cancerlectins sequences were selected

TABLE 2 | Classification of new data.

ID	Prediction results
1016841179	1
1016841154	1
1016841024	1
1016841005	1
560189093	1
720063203	1
727346123	1
469469047	0
403420575	1
385719187	1
384367986	1
388890228	1
1508736536	1
873090602	1
1022943309	1
974005177	1
392996940	0
385719190	1
1391723745	1
400260732	1
1370479176	1
1370451719	1
1034557774	1
768011769	1
768007991	1
768006291	0
1258501064	0
1272616377	1
1272616369	1
859066280	0

TABLE 3 | Comparison of classification results of new data.

Methods	Acc (%)
CancerPred (Amino acid composition) (Kumar and Panwar, 2011)	70
CancerPred (Dipeptide composition) (Kumar and Panwar, 2011)	76.67
CancerPred [Split composition (2-part)] (Kumar and Panwar, 2011)	56.67
CancerPred [Split composition (4-part)] (Kumar and Panwar, 2011)	60
Our Method	83.3

from NCBI database which were newly stored after 2012. From **Table 2**, prediction result 1 means correct classification, 0 means wrong classification. We can see there are 25 cancerlectins in new data were correctly predicted, the prediction accuracy of the new data is 83.3%.

As can be seen from **Table 3**, the model in this work has better classification performance on new data, that is, the model generalization ability in this work is stronger.

Comparing our method with other published methods, as shown in **Table 4**, the accuracy of the model obtained by our method is higher than that of previous studies. Though the specificity of our method is not much improved compared with Lin et al. (2015) and Lai et al. (2017), the sensitivity is greatly improved compared with the other three methods. The classification model improves the ability of correct recognition of cancer agglutinin samples, which shows that the classification model in this paper is effective.

TABLE 4 | Comparison with the results of existing classification models.

Method	S_n (%)	S_p (%)	Acc (%)
Kumar and Panwar (2011)	68.00	69.90	69.09
Lin et al. (2015)	69.10	80.10	75.19
Damodaran et al. (2008)	75.28	80.53	77.48
Our method	83.15	80.87	83.91

DISCUSSION AND CONCLUSIONS

Accumulated experimental evidences have shown that the classification of cancerlectins has important theoretical and practical significance for understanding its structural and functional characteristics, identifying drug targets, discovering tumor markers, and cancer treatment. More and more evidences show that it is crucial to propose an effective computational model to identify cancerlectins. In this paper, we developed a method based on the feature extraction algorithm of fusing g-gap dipeptide components to extract protein sequence features. Our method improve the feature extraction algorithm of protein sequence in cancerlectins prediction. We use the feature extraction algorithm of fusing g-gap dipeptide components to extract protein sequence features, which obtain an optimal feature subset containing 35 features. The accuracy, sensitivity and specificity are 83.91, 83.15, and 80.87% respectively. The results are better than those of the published methods. We also collect 30 new data form NCBI for predicted the performance of our method, and the prediction accuracy is 83.3%. Experimental results demonstrate that the performance of our method is better than the state-of-the-art methods for predicting cancerlectins.

Although our method can improve the prediction accuracy, it still has some limitations. Firstly, the benchmark dataset we used is relatively small, so there are some gaps in the data, and some specific attributes may be missing. Secondly, the extraction of protein sequence feature information is a key step in protein prediction. How to construct a better feature extraction algorithm remains to be further studied. Third, we only focus on the prediction of cancerlectin classification, how to choose a better classifier is our future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://proline.physics.iisc.ernet.in/cgi-bin/cancerdb/input.cgi; http://www.uniprot.org/.

AUTHOR CONTRIBUTIONS

LQ and GH contributed to the conception and design of the study and developed the method.

LQ and YW implemented the algorithms and analyzed the data and results. GH gave the ideas and supervised the project. LQ wrote the manuscript. GH and YW reviewed the final manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by the Natural Science Foundation of China (Grant No. 11401503), Outstanding Youth Foundation of Hunan Educational Committee (Grant No. 16B256), and Key Project of Hunan Educational Committee (Grant No. 19A497).

REFERENCES

An, J., You, Z., and Chen, X. (2016). Identification of self-interacting proteins by exploring evolutionary information embedded in PSI-BLAST-constructed position specific scoring matrix. *Oncotarget* 7, 82440–82449. doi: 10.18632/oncotarget.12517

Anh, V., Yu, Z. G., and Han, G. S. (2014). Secondary structure element alignment kernel method for prediction of protein structural classes. *Curr. Bioinform*. 3:9. doi: 10.2174/1574893609999140523124847

Balachandran, M., Shaherin, B., and Hwan, S. T. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Balachandran, M., Shin, T. H., and Gwang, L. (2018). PVP-SVM: sequence-based prediction of phage virion proteins using a support vector machine. *Front. Microbiol.* 9:476. doi: 10.3389/fmicb.2018.00476

Bu, W. S., Feng, Z. P., Zhang, Z. D., and Zhang, C. T. (1999). Prediction of protein (domain) structural classes based on amino acid index. *Eur. J. Biochem.* 266, 1043–1046. doi: 10.1046/j.1432-1327.1999.00947.x

Cai, Y. D., Liu, X. J., Xu, X. B., and Chou, K. C. (2002). Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. *J. Cell. Biochem*. 84, 343–345. doi: 10.1002/jcb.10030

Choi, S. H., Lyu, S. Y., and Park, W. B. (2004). Mistletoe lectin induces apoptosis and telomerase inhibition in human A253 cancer cells through dephosphorylation of akt. *Arch. Pharm. Re.*. 27, 68–76. doi: 10.1007/BF02980049

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255. doi: 10.1002/prot.1035

Damodaran, D., Jeyakani, J., and Chauhan, A. (2008). CancerLectinDB: a database of lectins relevant to cancer. *Glycoconjugate J.* 5, 191–198. doi: 10.1007/s10719-007-9085-5

De Mejía, E. G., and Prisecaru, V. I. (2005). Lectins as bioactive plant proteins: a potential in cancer treatment. *Crit. Rev. Food Sci. Nutr.* 45, 425–445. doi: 10.1080/10408390591034445

Feng, P. M., Hao, L., and Wei, C. (2013). Identification of antioxidants from sequence information using naïve Bayes. *Comp. Math. Methods Med*. 2013:567529. doi: 10.1155/2013/567529

Han, G. S., Yu, Z. G., and Anh, V. (2014). A two-stage SVM method to predict membrane protein types by incorporating amino acid classifications and physicochemical properties into a general form of Chou's PseAAC. *J. Theor. Biol*. 344, 31–39. doi: 10.1016/j.jtbi.2013.11.017

Kawashima, S., Ogata, H., and Kanehisa, M. (1999). AAindex: amino acid index database. *Nucleic Acids Res*. 27, 368–370. doi: 10.1093/nar/27.1.368

Kumar, R., and Panwar, B. (2011). Analysis and prediction of cancerlectins using evolutionary and domain information. *BMC Res. Notes* 4:237. doi: 10.1186/1756-0500-4-237

Lai, H. Y., Chen, X. X., and Chen, W. (2017). Sequence-based predictive modeling to identify cancer-lectins. *Oncotarget* 8, 28169–28175. doi: 10.18632/oncotarget.15963

Lin, H., and Chen, W. (2010). Prediction of thermophilic proteins using feature selection technique. *J. Microbiol. Methods*. 84, 67–70. doi: 10.1016/j.mimet.2010.10.013

Lin, H., Chen, W., Yuan, L. F., Li, Z. Q., and Ding, H. (2013). Using over-represented tetrapeptides to predict protein submitochondria locations. *Acta Biotheor*. 61, 259–268. doi: 10.1007/s10441-013-9181-9

Lin, H., Liu, W. X., and He, J. (2015). Predicting cancerlectins by the optimal g-gap dipeptides. *Sci. Rep*. 5:16964. doi: 10.1038/srep16964

Lis, H., and Sharon, N. (1998). Lectins: carbohydrate-specific proteins that mediate cellular recognition. *Chem. Rev*. 98, 674–637. doi: 10.1021/cr940413g

Liu, C., Yin, S. Q., Zhang, M., Zeng, Y., and Liu, J. Y. (2014). An improved grid search algorithm for parameters optimization on SVM. *Appl. Mech. Mater*. 644–50, 2216–9. doi: 10.4028/www.scientific.net/AMM.644-650.2216

Metfessel, B. A., Saurugger, P. N., Connelly, D. P., and Rich, S. S. (1993). Cross validation of protein structural class prediction using statistical clustering and neural networks. *Protein Sci.* 1993, 1171–1183. doi: 10.1002/pro.5560020712

Nakashima, H., Nishikawa, K., and Ooi, T. (1986). The folding type of a protein is relevant to the amino acid composition. *J. Biochem.* 99, 153–62. doi: 10.1093/oxfordjournals.jbchem.a135454

Runtao, Y., Chengjin, Z., Lina, Z., and Gao, R. (2018). A two-step feature selection method to predict cancerlectins by multiview features and synthetic minority oversampling technique. *Bio Med Res Int*. 2018, 1–10. doi: 10.1155/2018/9364182

Sharma, A., and Paliwal, K. K. (2008). Rotational linear discriminant analysis technique for dimensionality reduction. *IEEE Trans. Knowl. Data Eng.*. 20, 1336–1347. doi: 10.1109/TKDE.2008.101

Sharon, N., and Lis, H. (1989). Lectins as cell recognition molecules. *Science* 246, 227–234. doi: 10.1126/science.2552581

Xu, R., Zhou, J., Wang, H., He, Y., Wang, X., and Liu, B. (2015). Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol*. 9(Suppl. 1):S10. doi: 10.1186/1752-0509-9-S1-S10

Yanyuan, P., Hui, G., and Hao, L. (2018). Identification of bacteriophage virion proteins using multinomial naive Bayes with g-gap feature tree. *Int. J. Mol. Sci.*. 19:1779. doi: 10.3390/ijms19061779

Yu, C. S., Lin, C. J., and Hwang, J. K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide composition. *Protein Sci*. 13, 1402–1406. doi: 10.1110/ps.03479604

Yu, C. S., Wang, J. Y., and Yang, J. M. (2003). Fine-grained protein fold assignment by support vector machines using generalized n-peptide coding schemes and jury voting from multiple parameter sets. *Proteins* 50, 531–536. doi: 10.1002/prot.10313

Zhang, J., Ju, Y., Lu, H., Xuan, P., and Zou, Q. (2016). Accurate identification of cancerlectins through hybrid machine learning technology. *Int. J. Genom*. 2016, 1–11. doi: 10.1155/2016/7604641

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fgene.2020.00363

Novel Immune-Related Gene Signature for Risk Stratification and Prognosis of Survival in Lower-Grade Glioma

Mingwei Zhang[1,2,3,4,5†], Xuezhen Wang[1†], Xiaoping Chen[6†], Qiuyu Zhang[2] and Jinsheng Hong[1,3,4*]*

[1] Department of Radiation Oncology, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China, [2] Institute of Immunotherapy, Fujian Medical University, Fuzhou, China, [3] Key Laboratory of Radiation Biology (Fujian Medical University), Fujian Province University, Fuzhou, China, [4] Fujian Key Laboratory of Individualized Active Immunotherapy, Fuzhou, China, [5] Fujian Medical University Union Hospital, Fuzhou, China, [6] Department of Statistics, College of Mathematics and Informatics & FJKLMAA, Fujian Normal University, Fuzhou, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Xin Wang,
Houston Methodist Hospital, United States
Hong Zheng,
Stanford University, United States
Juan Ye,
National Institutes of Health (NIH), United States
Min He,
Leiden University, Netherlands

***Correspondence:**
Qiuyu Zhang
qiuyu.zhang@fjmu.edu.cn
Jinsheng Hong
13799375732@163.com

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

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

Received: *29 December 2019*
Accepted: *25 March 2020*
Published: *15 April 2020*

Citation:
Zhang M, Wang X, Chen X, Zhang Q and Hong J (2020) Novel Immune-Related Gene Signature for Risk Stratification and Prognosis of Survival in Lower-Grade Glioma. Front. Genet. 11:363. doi: 10.3389/fgene.2020.00363

Objective: Despite several clinicopathological factors being integrated as prognostic biomarkers, the individual variants and risk stratification have not been fully elucidated in lower grade glioma (LGG). With the prevalence of gene expression profiling in LGG, and based on the critical role of the immune microenvironment, the aim of our study was to develop an immune-related signature for risk stratification and prognosis prediction in LGG.

Methods: RNA-sequencing data from The Cancer Genome Atlas (TCGA), Genome Tissue Expression (GTEx), and Chinese Glioma Genome Atlas (CGGA) were used. Immune-related genes were obtained from the Immunology Database and Analysis Portal (ImmPort). Univariate, multivariate cox regression, and Lasso regression were employed to identify differentially expressed immune-related genes (DEGs) and establish the signature. A nomogram was constructed, and its performance was evaluated by Harrell's concordance index (C-index), receiver operating characteristic (ROC), and calibration curves. Relationships between the risk score and tumor-infiltrating immune cell abundances were evaluated using CIBERSORTx and TIMER.

Results: Noted, 277 immune-related DEGs were identified. Consecutively, 6 immune genes (*CANX*, *HSPA1B*, *KLRC2*, *PSMC6*, *RFXAP*, and *TAP1*) were identified as risk signature and Kaplan–Meier curve, ROC curve, and risk plot verified its performance in TCGA and CGGA datasets. Univariate and multivariate Cox regression indicated that the risk group was an independent predictor in primary LGG. The prognostic signature showed fair accuracy for 3- and 5-year overall survival in both internal (TCGA) and external (CGGA) validation cohorts. However, predictive performance was poor in the recurrent LGG cohort. The CIBERSORTx algorithm revealed that naïve $CD4^+$ T cells were significant higher in low-risk group. Conversely, the infiltration levels of

M1-type macrophages, M2-type macrophages, and $CD8^+$T cells were significant higher in high-risk group in both TCGA and CGGA cohorts.

Conclusion: The present study constructed a robust six immune-related gene signature and established a prognostic nomogram effective in risk stratification and prediction of overall survival in primary LGG.

Keywords: lower grade glioma, The Cancer Genome Atlas, Chinese Glioma Genome Atlas, immune-related signature, prognosis

INTRODUCTION

Lower-grade gliomas (LGG) constitute the prevalent primary malignances of the central nervous system, demonstrating great intrinsic heterogeneity in terms of their biological behavior (Ostrom et al., 2013; Zeng et al., 2018). So far, maximum surgical resection combined with postoperative radiotherapy and chemotherapy is the standard treatment for LGG. Despite numerous efforts to improve the clinical outcome, more than half of the LGG cases evolve and progress to therapy-resistant high-grade aggressive glioma over time (Claus et al., 2015). Thus, it is imperative to identify novel prognostic factors for LGG. Several biomarkers, including the isocitrate dehydrogenase (IDH) mutation, co-deletion of chromosome arms 1p and 19q (1p/19q codeletion), and O-6-methylguanine-DNA methyltransferase (MGMT) methylation have been integrated to the 2016 WHO classification, to illustrate the histological features and guide the therapeutic strategy (Hartmann et al., 2010; Wick et al., 2013; Hainfellner et al., 2014; Louis et al., 2016). However, these widely utilized biomarkers do not fully elucidate the individual variants and properly address risk stratification in LGG. Thus, it would only be reasonable to attempt to integrate various methods, including gene expression profiles that have gathered enormous attention, to further improve stratification of LGG.

The immune microenvironment has been identified as playing a critical role in tumor biology (Hanahan and Weinberg, 2011), and recently, numerous promising preclinical and clinical immunotherapeutic treatments, including immune-checkpoint inhibitors, active or passive immunotherapy, and gene therapy, have been achieved in malignant gliomas (Mahmoodzadeh Hosseini et al., 2015; Xu et al., 2015; Reznik et al., 2018; Simonelli et al., 2018; Vismara et al., 2019), further establishing the vital role of immunotherapy in the management of gliomas. Hence, the molecular profiles of the immune components within the tumor microenvironments represent tremendous value in serving as prognostic biomarkers. Recently, several studies have proposed immune gene expression-based signatures for risk stratification and for predicting clinical outcomes in breast, gastric, thyroid, and ovarian cancers (Ascierto et al., 2012; Kim et al., 2018; Shen et al., 2019; Yang et al., 2019). In terms of the prognostic value of an immune-related risk signature in glioma, Cheng et al. (2016) revealed that not only did the immune-related risk signature had prognostic significance in the stratified patients for glioblastoma, but moreover the immune status and local immune response could be illustrated by the risk signature. However, implementation of an immune gene expression-based signature has not been fully elucidated in LGG.

In a previous study, Li and Meng (2019) identified an immune-related long non-coding RNA (lncRNA) signature based on 529 low-grade glioma cases. It was found that the 8-lncRNAs model could serve as an independent predictor in low-grade glioma, not enrolling cases of grade III glioma. However, the predictive accuracy of the lncRNA-based model needed to be enhanced and the external validation was warranted. Furthermore, the correlation between the immune-related model and immune cell phenotypes was not illustrated. To our knowledge, the latest version of Cell type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORTx) has been investigated as a highly sensitive and specific algorithm set to reveal the immune landscape of 22 human immune cell compositions in solid tumors (Newman et al., 2019) and thus might provide new insights into potential therapeutic candidates for the management of LGG.

In the present study, a large cohort of patients with primary LGG from The Cancer Genome Atlas (TCGA) database and normal control cases from the Genome Tissue Expression (GTEx) database were employed to screen differentially expressed immune-related genes (IRGs). After construction of the risk signature based on the immune related genes, patients with primary LGG with gene sequencing data from the Chinese Glioma Genome Atlas (CGGA) database were adopted as the external validation. In addition, the CIBERSORTx and Tumor Immune Estimation Resource (TIMER) algorithm were utilized to clarify the correlation between the risk signature and the abundances of the infiltrative immune cells in primary LGG samples.

MATERIALS AND METHODS

Acquisition of LGG Expression Profiles From TCGA Datasets

The RNA-seq data (level 3) and clinical information of LGG samples were collected from UCSC Xena[1]. Expression of genes analyzed in normal tissues was collected using the Genome Tissue Expression (GTEx) (Consortium, 2015; Gentles et al., 2015) tool. Normalized gene expression was measured as fragments per kilobase of transcript per million mapped reads (FPKM)

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

and log2-based transformation. Then, the "sva" package of R software was utilized for the normalization of RNA expression profiles and to remove the batch effects. Principal component analysis (PCA) was used for detecting batch effects from the GTEx and TCGA datasets.

Acquisition of Immune-Related Genes

A comprehensive list of IRGs was downloaded from the Immunology Database and Analysis Portal (ImmPort) database[2]. The list comprised a total of 2,498 IRGs, covering 17 immune categories (Bhattacharya et al., 2014).

Inclusive and Exclusive Criteria of Enrolled Patients for the Construction of Risk Signature

The inclusive criteria of patients with LGG for model construction were as follows: (1) only patients with primary glioma were enrolled, (2) pathologic types of WHO II or III grade, (3) complete clinicopathological parameters, (4) only samples with RNA-sequencing data, (5) overall survival (OS) as the primary endpoint, (6) minimum follow-up of 90 days. The exclusive criteria included (1) patients with recurrent LGG, (2) pathologic type was glioblastoma, (3) incomplete survival status and clinical information.

Establishment of the Immune-Related Risk Signature

Using the "survival" package in R, we employed univariate Cox regression on IRGs and OS of primary LGG in the TCGA database to identify survival-associated IRGs. Next, using the "glmnet" package in R, the least absolute shrinkage and selection operator (Lasso) regression model was selected to minimize the over-fitting and identify the most significant survival-associated IRGs in primary LGG. After testing for collinearity, stepwise multivariate Cox regression analysis was performed to establish the IRG-derived risk signature in primary LGG. The following formula based on a combination of Cox coefficient and gene expression was used to calculate the risk score (Lossos et al., 2004; Chen et al., 2007; Hu et al., 2019):

$$Model{:}Risk\ score = \sum_{i=1}^{k} \beta i Si$$

where k, β_i, S_i represent the number of signature genes, the coefficient index, and the gene expression level, respectively.

To stratify patients into low- and high-risk groups, the optimum cutoff value for the risk score was determined using the "survminer" package in R. In order to ensure the comparability of the sample size between two groups, we set the *min.prop* parameter = 0.3 in applying the "survminer" package. Next, the Kaplan Meier survival curve and log-rank test was performed to evaluate the survival rates between low- and high-risk groups. The area under the receiver operating characteristic (ROC) curve (AUC) was calculated using the "survival ROC" package in R. In addition, the risk plot was illustrated using the "pheatmap" package in R.

Identification of the Prognostic Factors for OS in Primary LGG

All patients with primary LGG in TCGA were randomly divided into the training and testing groups at a ratio of 7:3 using the "caret" package. Seven predominant clinical and prognostic factors, including age, gender, grade, radiotherapy, chemotherapy, IDH status, and the risk scores of the immune-related signature were evaluated using univariate and multivariate Cox regression analyses. Before that, we tested the proportional hazards assumption (Therneau, 1994) by Schoenfeld residuals analysis (Schoenfeld, 1982), using the statistical script language R (R Development Core Team, 2014). By employing "rms," "foreign," and "survival" R packages, we formulated a nomogram consisting of relevant clinical parameters and independent prognostic factors based on the multivariate Cox regression analysis. The performance of the prognostic nomogram was assessed by calculating Harrell's concordance index (C-index) (Harrell et al., 1996), the AUC of the time-dependent ROC curve, and calibration curves of the nomogram for 3-, and 5-year OS plotted to estimate the accuracy of actual observed rates with the predicted survival probability. Time-dependent ROC analyses were conducted by "timeROC" R package.

External Validation of the Signature in CGGA Datasets for Primary LGG

The prognostic capability of the immune-related risk signature was externally validated using CGGA database. The RNA-seq data and corresponding clinicopathological information were obtained from the CGGA database[3]. The specific risk score for each patient was calculated with the use of the prognostic gene signature. Similarly, patients were divided into low- and high-risk groups based on the constructed formula in TCGA database. The optimal cutoff of risk scores for CGGA dataset kept the same as that in primary TCGA cohorts. Survival curves for the low- and high-risk groups were plotted using Kaplan-Meier analysis. Next, the predictive accuracy of the signature was investigated using ROC curves, and the performance of the nomogram was also assessed by the time-dependent ROC curve and calibration.

Investigation of the Signature in Patients With Recurrent LGG

For testing the prediction model in patients with recurrent LGG, the main inclusion criteria were: (1) patients suffering from recurrent glioma with histologically confirmed WHO II or III grade, (2) evidence of tumor recurrence and complete clinicopathological factors, (3) available recurrent glioma RNA-sequencing profiling, (4) minimum follow-up of 90 days. The exclusive criteria were as follows: (1) incomplete survival status and clinical information, (2) primary LGG samples. Time-dependent ROC curve and calibration plots were created to

[2]https://immport.niaid.nih.gov

[3]http://www.cgga.org.cn

investigate whether the built model could effectively predict survival in recurrent LGG.

Tumor-Infiltrating Immune Cell Analysis

To characterize the abundance of 22 immune cell types based on the RNA-seq data in lower grade glioma tissues, the CIBERSORTx web tool was applied[4]. Using a deconvolution algorithm (Newman et al., 2019), CIBERSORTx computed that the 22 cell types encompassed among others B cells, T cells, natural killer (NK) cells, macrophages, and dendritic cells (DCs). CIBERSORTx derived an empirical *P*-value for the deconvolution of each case using Monte Carlo sampling, and samples with $P < 0.05$ were adopted for analysis because of high reliability of the inferred cell composition (Ali et al., 2016). Therefore, cases with a *P* value of $\geq$0.05 were not retained for subsequent analysis. For validating the accuracy of the CIBERSORTx, TIMER (Tumor Immune Estimation Resource) database was also employed to illustrate the abundance of six immune cells containing B cells, $CD4^+$ T cells, $CD8^+$ T cells, macrophages, neutrophils, and dendritic cells[5]. Subsequently, the box plots were utilized to present the difference of infiltrative immune cells, T cell activated and inhibitory receptors, and macrophage associated molecules between high and low risk groups using the "ggplot2" package. In addition, the Cox regression model was also applied to calculate the hazard ratios (HRs) of the abundance of immune cells between high-and low-risk groups and illustrated by the forest plot.

Validation of Gene Expression in Cell Lines and Glioma Tissues

The Cancer Cell Line Encyclopedia (CCLE) was generated to provide a compilation of mRNA expression, copy number variation, and preclinical datasets for mutations in various cancer types. Details regarding the acquisition of mRNA expression of six genes profiled by RNA-Seq were downloaded from the data portal[6] (Barretina et al., 2012). The genomic data were utilized to analyze the mRNA expression status of the six immune genes in LGG cell lines. Cell lines of LGG were identified through six dedicated websites[7,8,9,10,11,12]. We only retained the consistent LGG cell lines across six websites. Furthermore, the level of protein expression for these six IRGs were confirmed using immunohistochemistry data publicly available at http://www.proteinatlas.org/. This database was explored to verify the gene-specific expression information across normal human tissues, as well as LGG.

[4]https://cibersortx.stanford.edu/

[5]https://cistrome.shinyapps.io/timer/

[6]https://portals.broadinstitute.org/ccle

[7]https://web.expasy.org/cellosaurus/

[8]https://www.atcc.org/

[9]https://www.phe-culturecollections.org.uk/products/celllines/generalcell/search.jsp

[10]http://igrcid.ibms.sinica.edu.tw

[11]https://cansarblack.icr.ac.uk/

[12]https://www.dsmz.de/

Statistical Analysis

All statistical analyses were conducted using R (version 3.6.0). The Wilcox test was used to screen statistically differentially expressed genes and infiltrative immune cells. Pearson's chi-square tests were executed for the comparison of categorical variables. Kaplan–Meier curve using the log-rank test was used to evaluate the statistical significance of the survival rates between different risk groups. The predictive accuracy of the risk signatures were determined by ROC curves. The proportional-hazards assumption was tested with Schoenfeld residuals. Then, univariate and multivariate Cox regression analysis were performed to evaluate significantly prognostic factors. Finally, results of multivariate Cox regression analyses were visualized with nomogram. Concordance index, time-dependent ROC, and calibration were also important indicators used to assess the nomogram. *P* value < 0.05 was considered statistically significant.

RESULTS

Preparation of Glioma Datasets

The workflow of our study is delineated in **Supplementary Figure S1**. A total of 916 patients who met the inclusion criteria, including 432 patients with primary LGG from the TCGA database, 353 patients with primary LGG from the CGGA database, and 131 patients with recurrent LGG from the CGGA database were obtained for further analysis. The clinicopathological characteristics of patients from the two databases are listed in **Table 1**.

Identification of DEGs

Before the identifying of DEGs, the normalization and batch effects removal from GTEx and TCGA datasets was conducted by "sva" package. As shown in **Supplementary Figures S2A,C**, the normalization of the data was performed well by the "sva" package. Additionally, the PCA plot found that TCGA and GTEx datasets separated obviously (**Supplementary Figures S2B,D**). To identify DEGs between the TCGA and GTEx databases, we considered the absolute value of the log2-transformed fold change (FC) > 1 and the adjusted *P*-value (adj.P) < 0.05 as the threshold levels of significance. Compared to non-tumor tissues, a total of 5,490 DEGs consisting of 2,718 upregulated and 2,772 downregulated genes were identified. The heatmap and volcano plot of the DEGs are shown in **Supplementary Figures S3A,B**. IMMPORT[13] is a web server for acquiring immune gene lists. From this set of DEGs, a total of 277 differentially expressed IRGs were extracted. The heatmap of 277 differentially expressed IRGs was shown in **Figure 1A**.

Identification of Prognostic IRGs

Based on the univariate Cox regression model ($P < 0.05$), a total of 36 IRGs were discovered to be significantly associated with OS. A forest plot of HR showed that 29 IRGs were risk factors, whereas 7 IRGs were protective factors (**Figure 2**).

[13]http://immport.org

TABLE 1 | Summary of risk scores and clinical pathological characteristics for different cohorts.

	Primary LGG			Recurrent LGG
	Training Cohort	Internal Validation Cohorts	External Validation Cohorts	Investigation
Characteristic	TCGA (*n* = 304)	TCGA (*n* = 128)	CGGA (*n* = 353)	CGGA (*n* = 131)
Age (y)[1]				
≤40	152 (50%)	53 (41%)	189 (54%)	69 (53%)
>40	152 (50%)	75 (59%)	164 (46%)	62 (47%)
Gender				
Male	175 (58%)	62 (48%)	205 (58%)	76 (58%)
Female	129 (42%)	66 (52%)	148 (42%)	55 (42%)
Grade				
II	139 (46%)	66 (52%)	196 (56%)	32 (24%)
III	165 (54%)	62 (48%)	157 (44%)	99 (76%)
Radiation				
No	109 (36%)	47 (37%)	59 (17%)	26 (20%)
Yes	195 (64%)	81 (63%)	294 (83%)	105 (80%)
Chemotherapy				
No	134 (44%)	61 (48%)	147 (42%)	34 (26%)
Yes	170 (56%)	67 (52%)	206 (58%)	97 (74%)
IDH[2] status				
Wild-type	53 (17%)	27 (21%)	94 (27%)	31 (24%)
Mutation	251 (83%)	101 (79%)	259 (73%)	100 (76%)
Risk score				
Low risk	209 (69%)	88 (69%)	248 (70%)	94 (72%)
High risk	95 (31%)	40 (31%)	105 (30%)	37 (28%)

[1]Age, Age at pathological diagnosis of glioma; [2]IDH, Isocitrate dehydrogenase.

Evaluation of IRGs With Prognostic Value

Considering collinearity and following refinement by the Lasso, only 11 genes were remained in Lasso regression from 36 significant prognosis associated IRGs in univariate Cox regression model. Ultimately, a prognostic signature comprisingsix IRGs, including calnexin (*CANX*), heat shock protein family A (HSP70) member 1B (*HSPA1B*), killer cell lectin like receptor C2 (*KLRC2*), proteasome 26S subunit, ATPase 6 (*PSMC6*), regulatory factor X associated protein (*RFXAP*), and transporter 1, ATP-binding cassette subfamily B member (*TAP1*) was selected to construct a prediction model by stepwise multivariate Cox regression analysis. Correspondingly, the coefficients of the six genes were 0.38625, 0.18073, −0.27702, −0.71285, −0.68077, and 0.34100. Ultimately, the hazard ratios of the six genes were 1.4714, 1.1981, 0.7580, 0.4902, 0.5062, and 1.4064, respectively. The comprehensive risk score was imputed as follows: (0.38625 × expression level of *CANX*) + (0.18073 × expression level of *HSPA1B*) + (−0.27702 × expression level of *KLRC2*) + (−0.71285 × expression level of *PSMC6*) + (−0.68077 × expression level of *RFXAP*) + (0.34100 × expression level of *TAP1*). Optimal cutoff values for the risk scores were calculated using the "survminer" package. Thus, patients were stratified into low- (risk score < 1.28) and high-risk (risk score ≥ 1.28) groups. In addition, the differential expression of six risk genes between normal brain and LGG tissues were shown in **Figure 1B**.

Performance of Risk Signature in Primary LGG From TCGA

Four hundred and thirty-two patients with primary LGG from the TCGA database were included in subsequent survival analyses and divided into low- and high-risk groups. Kaplan–Meier plots indicated that patients with high-risk scores presented a worse OS probability (**Figure 3A**). To verify the diagnostic competence of the immune-related risk signature, theAUC was calculated. The AUC of the ROC was 0.914, indicating that the risk score literally played a significant performance in the efficacy of this diagnosis (**Figure 3B**). The heatmap demonstrated that *KLRC2* exhibited the lowest expression in the high-risk group, whereas *CANX*, *HSPA1B*, *PSMC6*, *RFXAP*, and *TAP1* had medium and high expression levels (**Figure 3C**). Consecutively, patients appeared to have an increased mortality rate with an increase in risk scores according to the risk plot (**Figure 3D**).

Construction of Prognostic Signature in Primary LGG From TCGA

Using the "caret" package, the 432 patients with primary LGG in the TCGA dataset were randomly separated into training and testing cohorts at a ratio of 7:3. Seven clinicopathological parameters recorded as binary variables: age (≤40 vs. >40), gender (male vs. female), grade (grade II vs. grade III), radiotherapy (yes vs. no), chemotherapy (yes vs. no), risk (low vs. high), and IDH status (wild-type vs. mutation) were employed into further analyses, following testing of the proportional hazards assumption with Schoenfeld residual plots (**Supplementary Figure S4**). To evaluate the independent prognostic force of the signature, both the univariable and multivariable Cox proportion hazard regression models were applied (**Figures 4A,B**). Results from univariable analysis showed that risk (HR = 5.807, $P < 0.001$), age (HR = 3.029, $P < 0.001$), grade (HR = 3.455, $P < 0.001$), radiation therapy (HR = 2.841, $P < 0.001$), and IDH status (HR = 0.084, $P < 0.001$) had prognostic value for OS in primary LGG. Likewise, the risk group (HR = 2.383, $P = 0.008$), age (HR = 2.356, $P = 0.005$), grade (HR = 2.233, $P = 0.007$) and IDH status (HR = 0.189, $P < 0.001$) maintained their prognostic values in multivariable stepwise cox regression analysis. Next, risk, age, gender, grade, radiotherapy, chemotherapy, and IDH status were visualized in the nomogram. Nomograms of 3- or 5-year OS in the cohort are presented in **Figure 4C**. Then, the C-index for the training group was 0.8642. The AUC of the nomogram was up to 0.88, indicating the excellent ability to discriminate patients of poor from patients of favored prognosis (**Figure 4D**). Meanwhile, the calibration curve also manifested a satisfactory agreement between predictive and observational values at the probabilities of 3- and 5-year survival (**Figures 4E,F**). These results revealed that the nomogram signified good accuracy in predicting the 3- or 5-year survival of patient with LGG.

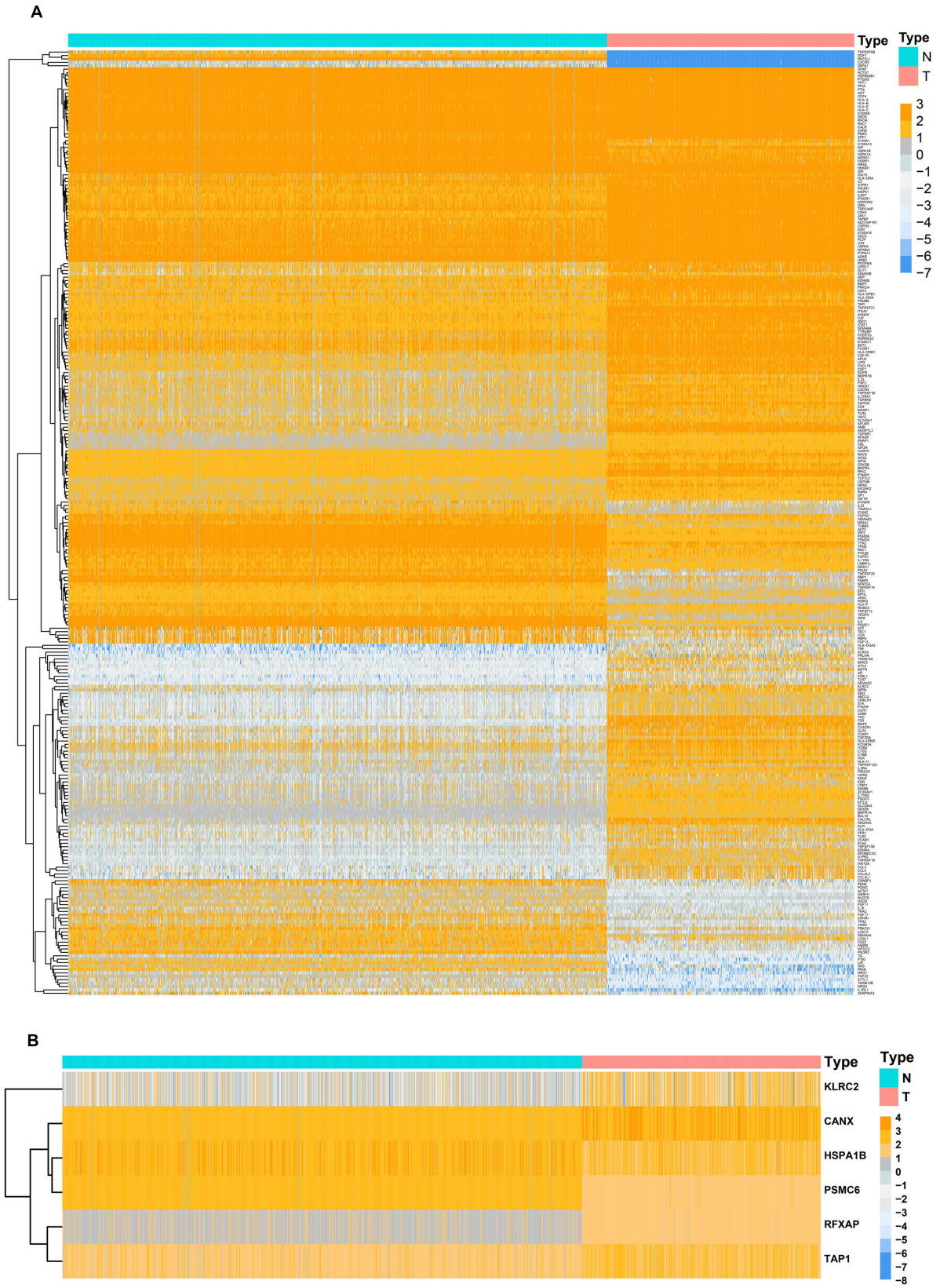


FIGURE 1 | Heatmaps of differentially expressed genes between normal tissue and lower-grade glioma. **(A)** Heatmap demonstrating the differential expressed 277 immune-related genes. **(B)** Heatmap demonstrating the differential expressed six immune-related risk genes.

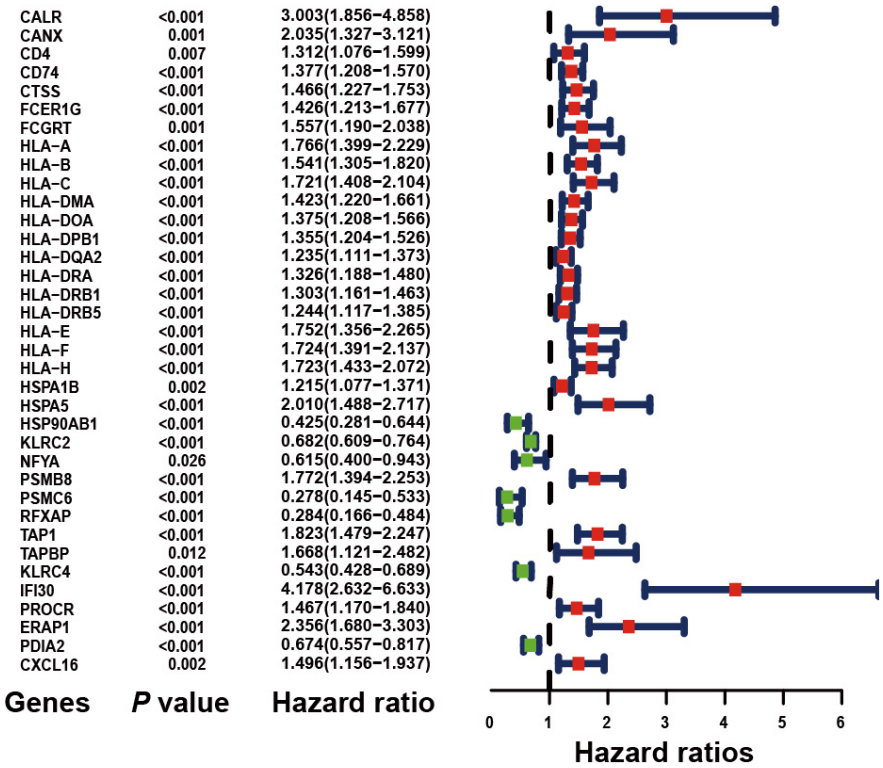


FIGURE 2 | Forest plot of hazard ratios demonstrating the prognostic values of immune-related genes (IRGs). The dash line was used to mark the location of HR = 1. The red box represents the adverse prognostic factor; Blue box represents the favorable prognostic factor.

Internal Validation of Prognostic Signature in Primary LGG From TCGA

A total of 128 patients with primary LGG in the TCGA dataset were randomly assigned in the internal cohort and the predictive power of the signature was accordingly confirmed. Each of the cases was divided into low- and high-risk groups. The C-index for the internal validation group was 0.8309. Time-dependent ROC analyses at 3- and 5-year were conducted to assess the prognostic accuracy of the six-gene-based classifier. The 3- and 5-year AUC were 0.836 and 0.761, respectively (**Figure 5A**). The calibration curve also manifested a satisfactory agreement between predictive values and observational values at the probabilities of 3- and 5-year survival (**Figures 5B,C**).

External Validation of Prognostic Signature in Primary LGG From CGGA

To determine whether the six-gene prognostic signature had similar prognostic value in different populations, its prediction performance was validated in another 353 primary LGG samples with RNA-seq transcriptome data and corresponding clinicopathological information from the CGGA database. The primary LGG samples were divided into two groups according to the cutoff value (<1.28 vs. ≥ 1.28). Consistent with the above findings, the Kaplan-Meier survival curves revealed a significant difference in OS between the low- and high-risk groups (**Figure 6A**). The AUC was 0.727, showing a fair prognostic power of the model (**Figure 6B**). To evaluate the prognostic accuracy of the model, time-dependent ROC analysis was conducted, with the AUC for 3, and 5-year survival being 0.836 and 0.798, respectively (**Figure 6C**). The C-index for the CGGA group was 0.7555. The calibrations plot for survival probability at 3- or 5-year showed an optimal consensus between the prediction and observation in both the external validation and training cohorts (**Figures 6D,E**).

Investigating the Application of Six Genes Based Signature in Recurrent LGG

Next, we investigated the feasibility of the six -immune-gene related risk signature in recurrent LGG. According to inclusive and exclusive criteria, 131 patients with recurrent LGG were enrolled for further analysis. Risk scores were calculated using the same formula and yielded similar results on Kaplan-Meier survival curves as those observed for primary LGG ($P < 0.05$; **Supplementary Figure S5A**). However, the AUC value was only 0.550, indicating a poor prognostic power in recurrent LGG (**Supplementary Figure S5B**). The C-index for the recurrent LGG group was 0.6135. Then, the AUC for 3-, and 5-y OS predictions for the recurrent cohort was 0.631, and 0.638, respectively (**Supplementary Figure S5C**). Meanwhile, the verification of the recurrent LGG cohort using the calibration plot was not satisfactory (**Supplementary Figures S5D,E**).

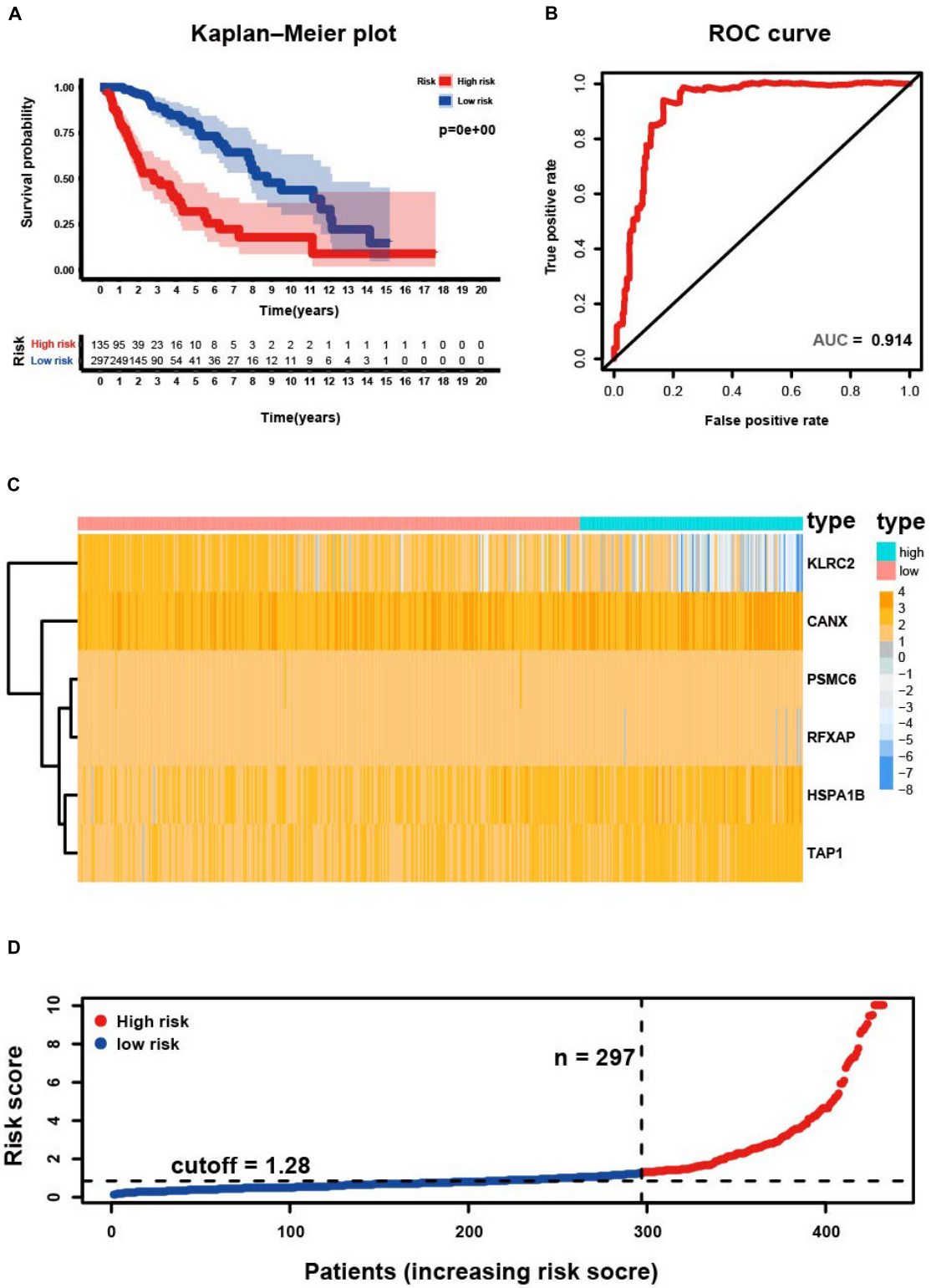


FIGURE 3 | Development of risk score based on the six immune-related gene signature of patients with primary LGG in TCGA. **(A)** Kaplan-Meier plot for overall survival (OS) based on risk score of the six gene based signature of patients with primary LGG in the TCGA cohort. **(B)** ROC curve with an AUC of 0.914, indicating that risk score plays a significant performance in the efficacy of this diagnosis. **(C)** Heatmap demonstrating the distribution of the six immune-related gene expression in the TCGA cohort. **(D)** Risk plot presenting each point sorted based on risk score, representing one patient. Blue, and red represent patients with low- and high-risk, respectively.

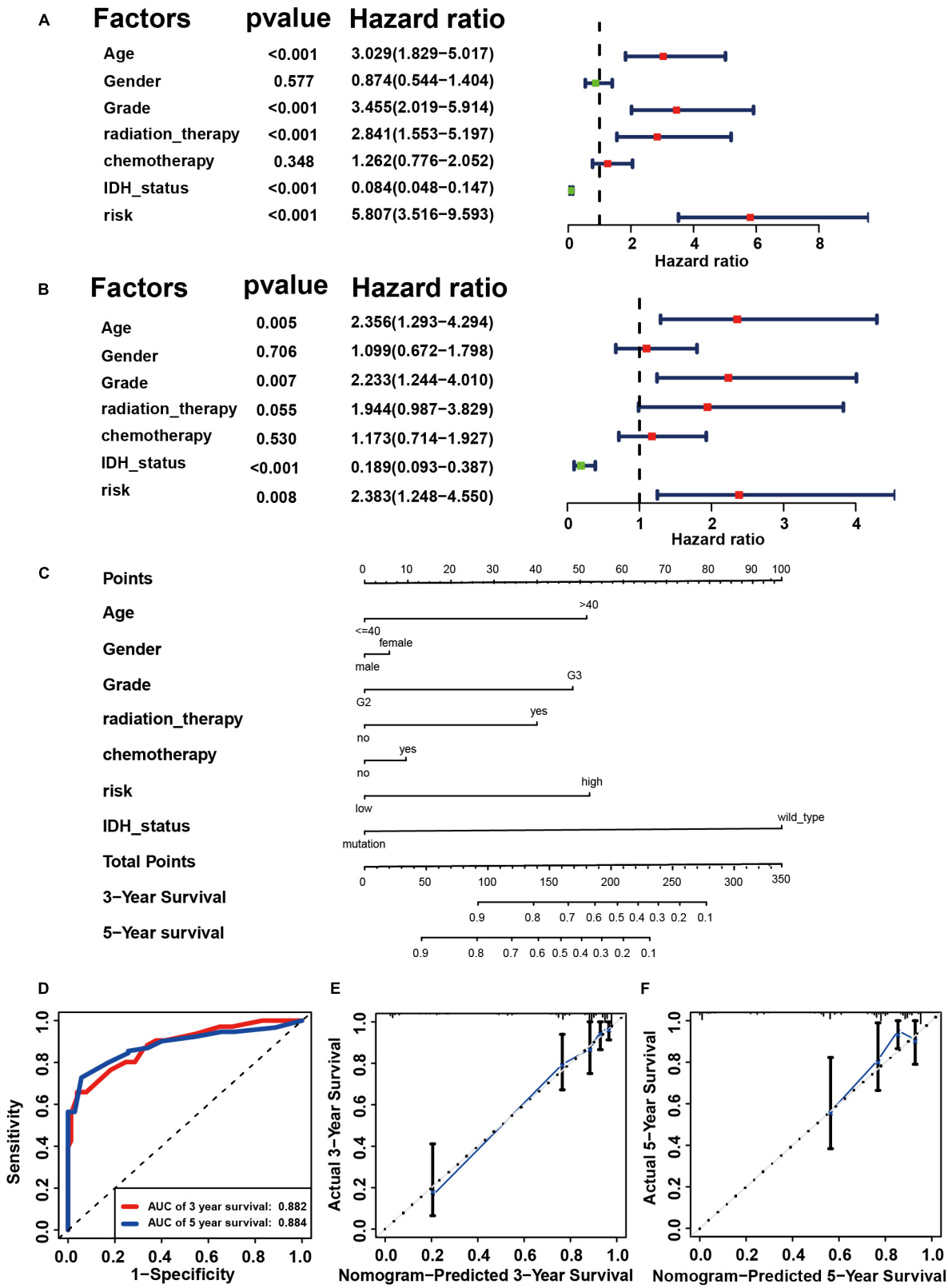


FIGURE 4 | Construction of prognostic signature in primary LGG from TCGA. Univariate **(A)** and multivariable Cox proportion hazard regression for OS **(B)** of primary LGG in training group. **(C)** A nomogram consisting of risk score and other clinical indicators for predicting 3-, and 5-year OS of primary LGG. **(D)** Time-dependent ROC for 3-, and 5-year OS predictions for the nomogram compared with actual observations. Calibration plot of nomogram for predicting probabilities of 3-year **(E)**, and 5-year **(F)** overall survival of patients. Blue line indicates actual survival.

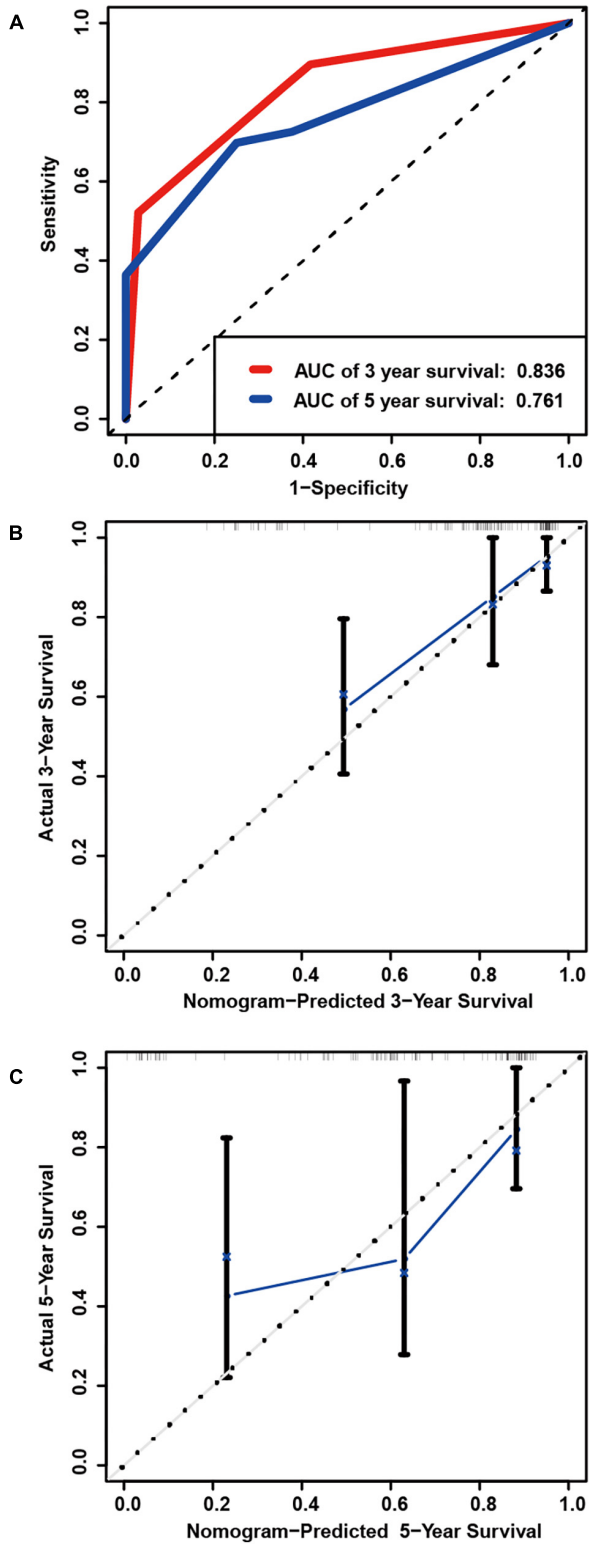


FIGURE 5 | Internal validation of prognostic signature in primary LGG from TCGA. **(A)** Time-dependent ROC curve based on the six genes based risk score for 3-, and 5-year OS probability in the internal validation cohort. Calibration plot for internal validation of 3-year **(B)**, and 5-year **(C)** OS of patients.

The Association Between Risk Score and Clinicopathological Parameters

Subsequently, we analyzed the relationship between the six-gene signature and clinicopathological parameters (age, gender, grade, radiotherapy, chemotherapy, and IDH mutation status) in LGG. In terms of grade and IDH status, patients of grade III or of the IDH wild type had higher risk scores than those with grade II or of the IDH mutant type, consistent with the findings in patients with primary LGG from CGGA. Moreover, data of patients with primary glioma from TCGA revealed that older patients had significantly higher risk scores than those of younger. Risk scores were also comparable across recurrent LGG in CGGA, with results revealing a preference for higher levels of risk scores in males. However, no significant difference was observed between the IDH wild and mutant groups in recurrent LGG (**Supplementary Figures S6A,C**).

Correlation of the Risk Score With Tumor-Infiltrating Immune Cells

By applying the CIBERSORTx algorithm to RNA-seq data, the relative proportions of 22 immune cell subsets of LGG were acquired. Consecutively, 432 cases of primary LGG in the TCGA dataset, 351 cases of primary LGG in the CGGA dataset were enrolled for further analysis after the filter criteria with P value < 0.05 via CIBERSORTx algorithms. As shown by bar plot in **Figure 7A**, the abundance of the 22 infiltrative immune cells by using CIBERSORTx were significantly different between high-risk and low-risk groups in primary LGG cohorts. Among them, the macrophage M2 was the most significant enrichment of immune cells. Subsequently, as shown in the box plots (**Figure 7B**), the infiltration levels of $CD8^+$T cells, resting memory $CD4^+$T cells, follicular helper T cells, regulatory T cells, activated NK cells, monocytes, macrophages (M0, M1, M2), activated DCs, resting mast cells, and neutrophils were significantly higher in high-risk group than that in low-risk group. On the contrary, the infiltration levels of naïve $CD4^+$T cells, and resting DCs were significantly higher in low-risk group. The differential abundance of the 22 infiltrative immune cells were summarized in **Table 2**. Furthermore, to validate the infiltrative abundance of immune cells in CIBERSORTx, the TIMER database was enrolled. As shown in **Figure 7C**, the B cells, $CD4^+$T cells, $CD8^+$T cells, DCs, macrophages, and neutrophils were all significantly higher in the high-risk group. To further investigate the prognostic values of the infiltrative immune cells, the univariate Cox proportion hazard regression models were applied. Results from Cox regression analysis showed that high abundance of Tregs, neutrophils, M2-type macrophages were significantly associated with unfavorable survival outcome ($P < 0.001$, $P < 0.001$, $P = 0.012$, respectively). Conversely, high abundance of macrophage M1 (HR = 0.203, $P < 0.001$), and activated DCs (HR = 0.416, $P < 0.001$) were identified as the protective factors in primary LGG (**Figure 7D**).

In addition, we also investigated the differential expressions of the T-cells activated and inhibitory receptors, and macrophage associated molecules between the high and low risk groups. As shown in **Supplementary Figure S7**, the T cells activation

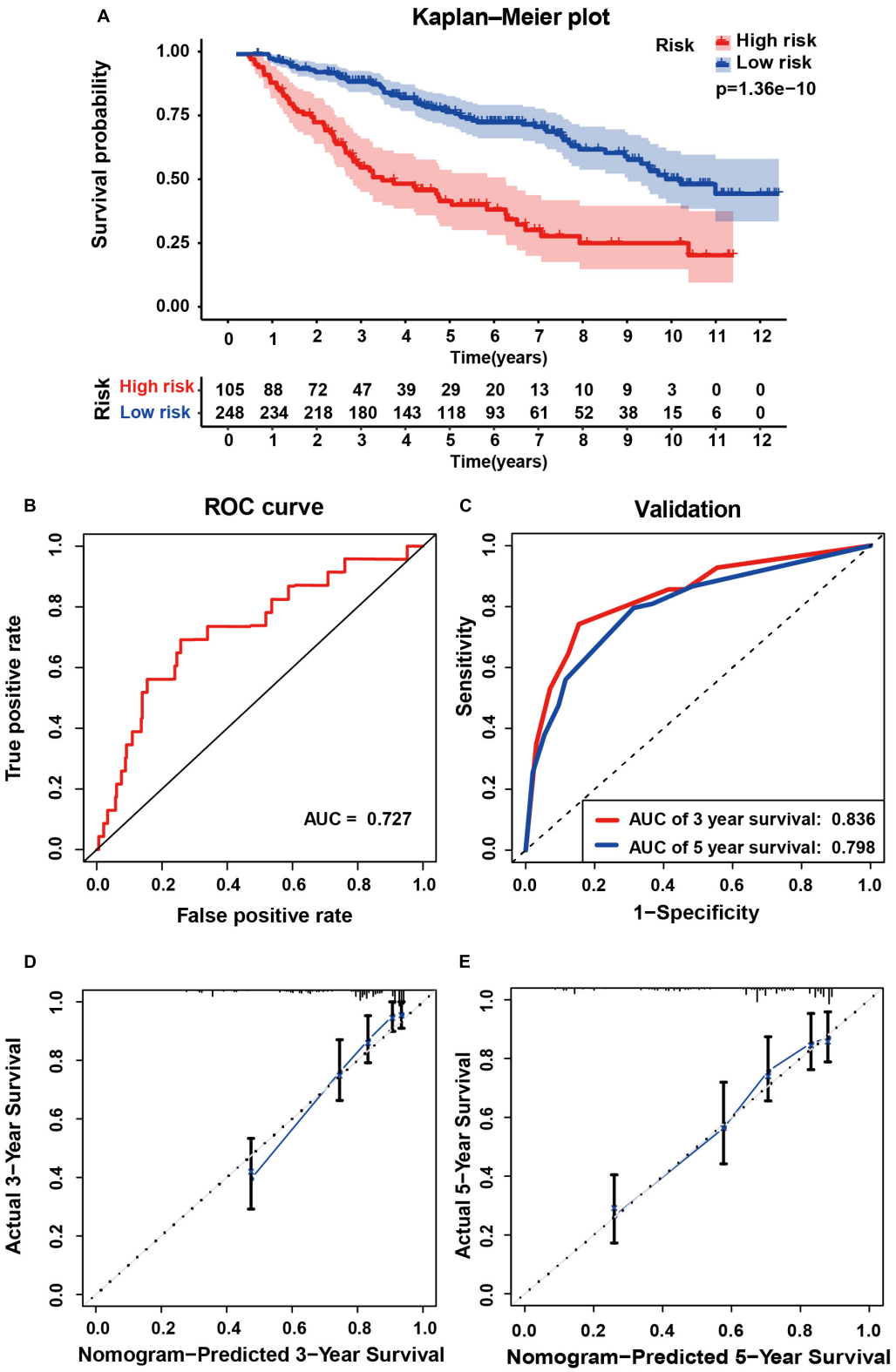


FIGURE 6 | External validation of the six gene signature in primary LGG inform the CGGA dataset. **(A)** Kaplan-Meier survival curves of the six gene signature of patients with primary LGG in the CGGA cohort. **(B)** ROC curve for assessing diagnostic competence of the risk score in the CGGA cohort. **(C)** ROC curves for 3-, and 5-year OS predictions for the six gene signature in the external validation cohort. Calibration curves for predicting probabilities of 3-year **(D)**, and 5-year **(E)** OS of patients in external validation.

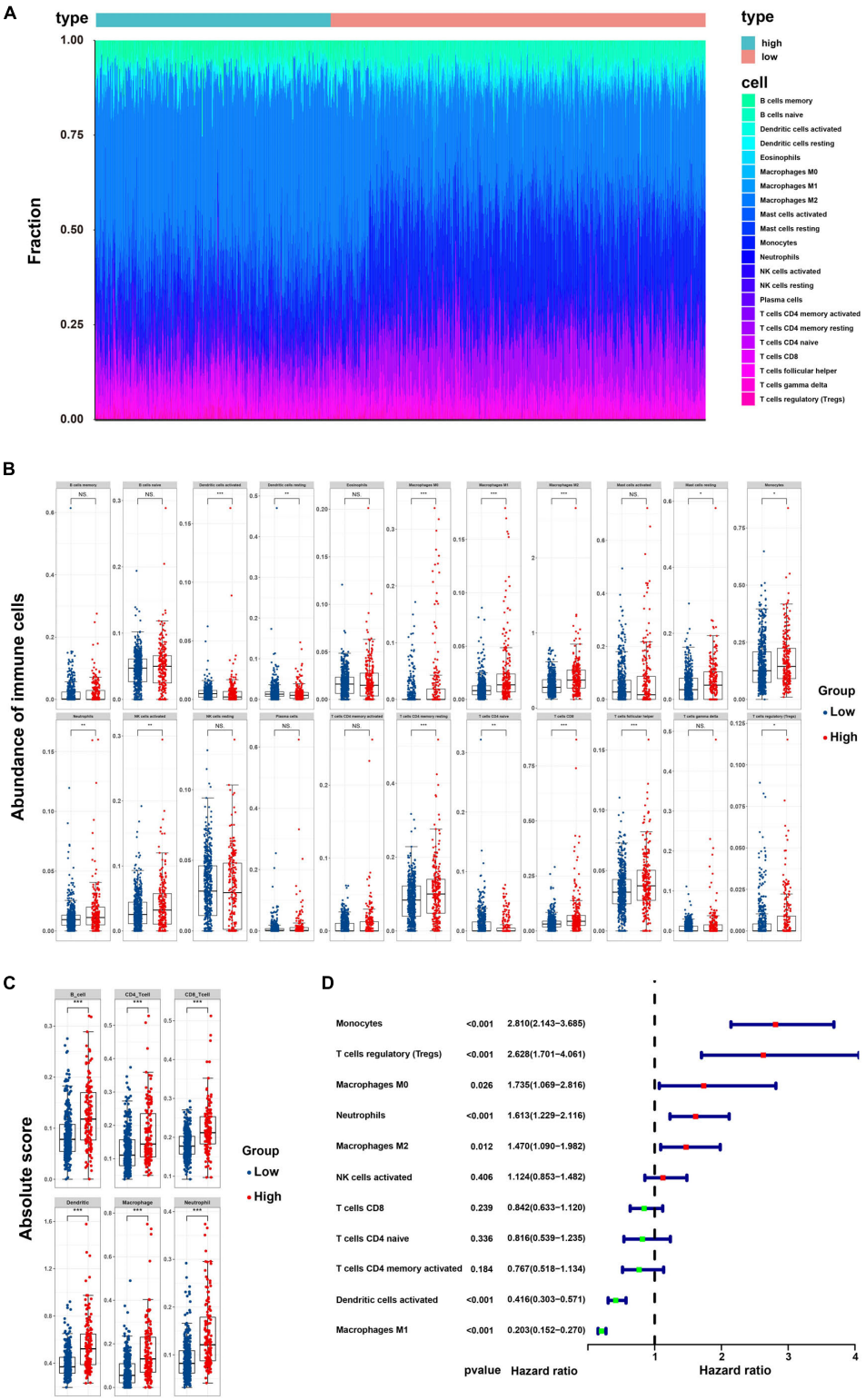


FIGURE 7 | Correlation of the risk score with infiltrative immune cells. **(A)** Barplot showing the percentages of 22 infiltrative immune cells calculated by CIBERSORTx between high-and low-risk groups in primary LGG from TCGA and CGGA cohorts (high-risk, 232 samples; low-risk, 551 samples); **(B)** Boxplot showing the differential abundance of 22 infiltrative immune cells calculated by CIBERSORTx between high-and low-risk group in primary LGG; **(C)** Boxplot showing the differential abundance of six infiltrative immune cells by TIMER database between high-and low-risk group in primary LGG; **(D)** Forest plot of hazard ratios demonstrating the prognostic values of 22 immune cells calculated by CIBERSORTx in primary LGG. The dash line was used to mark the location of HR = 1. The red box represents the adverse prognostic factor,and the blue box represents the favorable prognostic factor.

TABLE 2 | The differential abundances of 22 infiltrative immune cell types between high-and low-risk groups of with primary LGG as calculated by CIBERSORTx.

Immune cell type	Mean (high risk)	Mean (low risk)	Difference	*P* value
B cells naive	0.050	0.045	0.005	0.126
B cells memory	0.021	0.018	0.003	0.819
Plasma cells	0.015	0.007	0.008	0.179
T cells CD8	0.073	0.035	0.037	0.000
T cells CD4 naive	0.007	0.010	−0.003	0.001
T cells CD4 memory resting	0.107	0.083	0.024	0.001
T cells CD4 memory activated	0.011	0.006	0.004	0.539
T cells follicular helper	0.040	0.034	0.007	0.001
T cells regulatory (Tregs)	0.008	0.005	0.003	0.016
T cells gamma delta	0.016	0.007	0.009	0.167
NK cells resting	0.029	0.030	−0.001	0.298
NK cells activated	0.040	0.030	0.009	0.009
Monocytes	0.170	0.153	0.017	0.033
Macrophages M0	0.025	0.007	0.018	0.000
Macrophages M1	0.024	0.010	0.013	0.000
Macrophages M2	0.415	0.306	0.109	0.000
Dendritic cells resting	0.015	0.017	−0.003	0.001
Dendritic cells activated	0.006	0.005	0.000	0.000
Mast cells resting	0.055	0.039	0.015	0.021
Mast cells activated	0.069	0.054	0.015	0.170
Eosinophils	0.020	0.017	0.003	0.673
Neutrophils	0.016	0.011	0.005	0.001

associated genes containing *CD40L*, *GITR*, *4-1BB*, *OX40*, *CD27*, *ICOS*, and *CD28* were significant higher in high-risk group. T cells inhibition associated genes containing *CTLA4*, *PD-L1*, *PD-1*, *CD80*, *CD244*, *TIM3*, *BTLA*, *CD160* were also significant higher in high-risk group. Moreover, macrophage chemo-attractant and phagocytosis related genes containing *CSF1*, *CSF1R*, *CCL2*, *CCR2*, and *CXCR4* were also significant higher in high-risk group.

Six Genes Based Signature Expression Analysis in Databases

The expression of the six genes were queried from CCLE[14]. Results were sorted according to tumor type. The mRNA expression of *CANX*, *HSPA1B*, *PSMC6*, and *TAP1* was high in gliomas, whereas that of *KLRC2* was low (**Supplementary Figures S8A–F**). The expression of the six genes in 14 LGG cell lines is illustrated in **Table 3**. The Human Protein Atlas database was used to explore the protein expression levels of these six genes and results are shown in **Supplementary Figure S9**.

DISCUSSION

Emerging evidence has demonstrated that the immune microenvironment plays an essential role in tumor biology, and recently, numerous inspiring clinical trials have established the role of immunotherapy in gliomas. Thus, immune related biomarkers show great potential in risk stratification and in exerting prognostic value. In previous studies, immune-gene related signatures have been identified as independent prognostic factors in several solid tumors (Ascierto et al., 2012; Kim et al., 2018; Shen et al., 2019; Yang et al., 2019), revealing that the immune status and local immune response could be illustrated by the risk signatures employed. However, the prognostic value and the association between immune status and risk signatures have not been fully elucidated in LGG. In the current study, 277 immune-related DEGs were identified. After Lasso regression and multicox analysis, six immune genes (*CANX*, *HSPA1B*, *KLRC2*, *PSMC6*, *RFXAP*, and *TAP1*) were identified as components of the risk signature to divide LGGs into low- and high-risk groups. Subsequently, KM curve, ROC curve and risk plot analyses verified that the six-based risk signature performs well in stratifying the risk groups of primary LGG in TCGA and CGGA datasets. Furthermore, in univariable analysis, the risk group, age, grade, radiation therapy and IDH status exhibited their predictive value regarding OS in primary LGG. Correspondingly, in multivariable stepwise cox regression analysis, with the exception of radiation therapy showing borderline significance, all other factors retained their prognostic values. Consecutively, it was found that the prognostic signature showed fair accuracy regarding the 3- and 5-year OS in the internal (TCGA) and external (CGGA) validation cohorts. However, predictive performance was poor in the recurrent LGG cohort.

At first, it was shown that the IRG-based risk signature could function as a proper index in stratifying risk groups in LGG. Similar to our study, Shen et al. (2019) also found that an immune gene based signature could significantly stratify patients into different risk groups in ovarian cancer. Correspondingly, another study also revealed that the immune-related gene signature was capable of stratifying patients into responder and non-responder groups in human breast cancer, with the odds ratios of the immune-related risk signature making it the most significant predictor of pathological complete remission (odd ratio: 4.6, 95% confidence interval: 2.7 to 7.7, $P < 0.001$) (Sota et al., 2014). Second, we found that the risk group, age, grade, radiation therapy and IDH status had predictive values for OS in primary LGG. According to National Comprehensive Cancer Network guidelines, the prognostic values of age ($\leq$40 years vs. $>$40 years), tumor grade (II vs. III), and IDH status (wild-type vs. mutation) have been well-established in clinical practice (National Comprehensive Cancer Network, 2019). Compared with the above mentioned well-established clinicopathological prognostic factors, the risk group remained an independent prognostic value in univariate and multivariate cox regression analysis. In accordance with the present findings, Qian et al. (2018) also found that patients identified as high-risk by the IDH associated immune signature exhibited unfavorable prognosis in LGGs. The prognostic value of the local immune signature was also verified in glioblastomas. Risk scores were significantly associated with poor OS and progression-free survival (Cheng et al., 2016). Surprisingly, receiving or not radiation therapy was associated with OS in univariate analysis, but the relationship was borderline significant in multivariate analysis. In addition, the prognostic value of chemotherapy was also insignificant in our

[14]https://portals.broadinstitute.org/ccle

TABLE 3 | List the expression of the six genes in 14 LGG cell lines.

	Gene expression (TPM)						
Cell lines	***CANX***	***HSPA1B***	***KLRC2***	***PSMC6***	***RFXAP***	***TAP1***	***RRID***
H4	448.522	22.6035	0.0220068	15.7311	2.2642	34.2997	CVCL_1239
HS683	355.23	15.2679	0.278926	14.4031	3.26084	51.7626	CVCL_0844
KG1C	363.096	13.9961	0.113268	18.5451	3.64382	26.4942	CVCL_2971
LN215	288.476	26.108	2.48677	14.8853	4.68699	87.4974	CVCL_3954
LN235	235.629	27.8249	0.0447866	17.1381	4.57199	19.3359	CVCL_3957
LN319	271.245	21.0797	0	20.656	3.41301	17.9863	CVCL_3958
LNZ308	169.63	17.2783	0	22.7476	3.24888	17.519	CVCL_0394
NMCG1	264.062	16.9103	0.0049317	14.6614	2.81961	37.3184	CVCL_1608
SF268	272.357	27.68	0.0245476	9.43665	1.16183	31.9634	CVCL_1689
SNU738	144.254	17.2946	0.0770556	12.636	1.80947	31.244	CVCL_5087
SW1088	290.749	23.2654	0.0473503	17.416	4.36785	29.0014	CVCL_1715
SW1783	351.565	27.6688	0.0224892	18.6383	2.55953	33.8329	CVCL_1722
TM31	134.804	7.78124	0.0414412	21.8484	5.08505	60.5928	CVCL_6735
U178	220.836	8.10508	0.552537	17.0014	1.73231	37.2859	CVCL_A758

analysis. Our result is likely to be related to the undefined timing of radiation therapy (postoperative or palliative treatment), and differences in radiation dose or frequency. To our knowledge, numerous trials have investigated the prognostic values of chemotherapy and radiotherapy in gliomas, as well as their significant contribution in improving survival. The RTOG 9802 trial evaluated radiotherapy followed by adjuvant procarbazine, CCNU, and vincristine (PCV) chemotherapy in 251 patients with low-grade glioma and showed an improvement in median OS with the addition of PCV from 7.8 to 13.3 years (HR = 0.59; P = 0.002) (van den Bent, 2014). In the CATNON trial, the 5-year survival in patients with anaplastic glioma receiving combined chemo-radiotherapy was significant higher than that in patients receiving radiotherapy alone (55.9 vs. 44.1%, HR = 0.65; P = 0.0014) (van den Bent et al., 2017). The lack of prognostic values of chemotherapy and radiotherapy in our study, might be owing to several reasons: (1) undefined chemotherapy strategy (pre-radiotherapy or concurrent or adjuvant chemotherapy); (2) undefined chemotherapy regimens in the TCGA datasets; (3) undefined radiation regimens (postoperative or palliative treatment strategy, differences in radiation dose or frequency). Therefore, new trials are encouraged to further develop and verify our risk signature in standard treatment cohorts.

Emerging evidence have confirmed the prognostic values of immune genes in various cancers (Patel et al., 2013; Surmann et al., 2015; Yang et al., 2015; Ling et al., 2017; Ding et al., 2018). In current study, six IRGs were identified as the risk signature. Among them, *CANX*, *HSPA1B*, and *TAP1* were shown to be risk-associated genes, whereas *KLRC2*, *PSMC6*, and *RFXAP* were identified as protective genes. They have been reported to be involved in the regulation of immune response. Calnexin, an essential endoplasmic reticulum (ER) chaperone protein, plays a vital role in the synthesis of HLA class I surface antigen complex. Calnexin was revealed to inhibit the proliferation and activation of $CD4^+$T and $CD8^+$T cells, and it may impair the function of T cells by upregulating the expression of PD-1 in oral squamous cancer (Chen et al., 2019). Consistent with our results, it was found that decreased expression of *CANX* was associated with favorable survival outcome (Patel et al., 2013) and served as a biomarkers for tumor response in glioblastoma (Demeure et al., 2016). TAP1, an essential component of the major histocompatability complex (MHC) class I antigen-presenting pathway. It was found to be associated with tumor immune escape and prognosis (Leone et al., 2013). Ling et al. (2017) found that the expression of *TAP1* was significantly associated with infiltrative general T cells ($CD3^+$), $CD8^+$ cytotoxic T cells, M1-type macrophages, and M2-type macrophages, and the expression of *TAP1* could serve as an independent prognostic factor in colorectal cancer. In term of HSP70, encoded by *HSPA1B*, has emerged as a promising antitumor target in various cancer. Recently, it is also revealed that HSP70 may serve as a diverse immunoregulatory factors by acting as a cytokine in antigen presentation, DC maturation, the activities of NK cells, and myeloid-derived suppressor cells (Jego et al., 2019). Correspondingly, it was illustrated that up-regulation of *HSPA1B* was associated with poor outcomes in hepatocellular carcinoma (Yang et al., 2015). Comparatively, the investigations of *KLRC2* in cancer research is rare. To our knowledge, as a transmembrane activating receptor in NK cells, *KLRC2* is expressed in most NK cells and subsets of $CD8^+$T cells (Wischhusen et al., 2005; Borrego et al., 2006). PSMC6, as a critical component of 26S-proteasome complex, involving in numerous pathways: antigen presentation (Livneh et al., 2016), cell proliferation and migration (Guo and Dixon, 2016). Zhu et al. (2018) demonstrated that PSMC6 may involve in the downstream of silencing cat eye syndrome critical region protein-1 in targeting the proliferation of TAM in glioma. *RFXAP*, as a vital transcription factor for major histocompatibility complex (MHC) class II. It was revealed to downregulate the expression of MHC class II in DCs (Ding et al., 2015) and macrophages (Wu et al., 2019), resulting inhibition of $CD4^+$T cells infiltration (Surmann et al., 2015). It was associated with survival outcomes in solid tumors (Surmann et al., 2015; Ding et al., 2018). Overall, the prognostic values of the six risk genes have been exploited

in various cancers, and their contribution to immune regulations were mainly concentrated on antigen presenting cells and effector T lymphocytes. Hence, further investigation is warranted to illustrate the correlations between risk groups and infiltrative immune cells in primary LGG.

The immune microenvironment has been identified as playing a critical role in tumor biology (Hanahan and Weinberg, 2011). Numerous studies have exploited the critical roles of infiltrative immune cells in glioma (Perus and Walsh, 2019; Wang et al., 2020). In current study, it was found that the M2-type macrophage was significantly enriched in primary LGG. Despite the glioma was defined as "cold tumor" with very little infiltrative immune cells, the proportions of macrophage can still constitute up to 30–50% in the TME of glioma (Guadagno et al., 2018). Additionally, the predictive values of immune cells have been extensively investigated. It was demonstrated that high levels of M2-type macrophages (marked as CD204 or CD206) (Ding et al., 2014), neutrophils (Liang et al., 2014), Tregs (Iwata et al., 2019) were defined as the adverse prognostic factors in glioma. Conversely, high levels of M1-type macrophages (Ding et al., 2014), $CD8^+$T cells (Kmiecik et al., 2013) were identified as protective factors in glioma. Likewise, our results also revealed that elevated abundance of M2-type macrophages, neutrophils, and Tregs were associated with adverse survival outcomes. On the contrary, increased abundance of M1-type macrophages, and $CD8^+$T cells were associated with favorable survival outcomes. As mentioned above, the six risk genes can not only have intrinsic roles in tumor growth and apoptosis (i.e., Guo and Dixon, 2016; Chen et al., 2019; Jego et al., 2019), but also serve as the immune-regulatory factors via antigen-presenting cells (APCs) and effector T lymphocytes (Borrego et al., 2006; Surmann et al., 2015; Ling et al., 2017; Zhu et al., 2018; Chen et al., 2019; Wu et al., 2019). Hence, it is worthwhile to explore the relationship between the risk groups and infiltrative immune cells in primary LGG. Interestingly, it was found that the abundance of macrophages, activated DCs, NK cells, $CD8^+$T cells were significantly higher, while that of naïve $CD4^+$T cells were significantly lower in high-risk group. Moreover, our results also demonstrated that high riskscores were associated with aggressive tumor subtypes, rapid proliferation and shorter survival time. Therefore, we hypothesized that malignant proliferation in high-risk patients may be accompanied with elevated tumor mutation burden and increased necrosis and apoptosis, which lead to continuous exposure of neoantigens and subsequent activation of the immune response. Consequently, high levels of infiltrative APCs and effector cells (including NK, $CD4^+$T, and $CD8^+$T) were observed in TME of primary LGG. Correspondingly, our results in **Supplementary Figure S6** also illustrated that macrophage associated chemo-attractant molecules and T cell activating receptors were significant higher in high-risk group. Meanwhile, as a compensation response to increased immune activation (Perus and Walsh, 2019), the expressions of inhibitory molecules containing CTLA-4, PD-1, PD-L1, TIM-3, etc. (Wherry and Kurachi, 2015) were relatively higher in high-risk group. Noteworthy, it is necessary to clarify the positive relationship between riskscores and increased infiltrative immune cells. The aggressive phenotypes determined by the dysregulation of the six risk genes was fluctuated with the proportions of immune cells in TME, indicating that these genes may involve in the process of neoantigen presence and trigger the immune response. Considering that tumor cell is the large group of the antigen-presenting cells, 14 LGG cells lines were employed to validate the expression of six risk genes. It is obvious that all the six risk genes were commonly expressed, even some were high expressed in LGG cell lines. Further *in vivo* and *in vitro* experiments are warranted to investigate the mechanisms of six genes in LGG and the communications with immune cells in TME.

Our study, however had several limitations that should be addressed. First, because of the retrospective design and despite strict inclusive and exclusive criteria, selection and recall bias are unavoidable; Second, due to lack of complete chemotherapy and radiotherapy regimens in the current study, their prognostic values could not be fully elucidated. Third, although the 1p19q codeletion status constitutes a vital prognostic factor in clinical practice, such information was unavailable in the TCGA datasets and hence, was not employed in our prognostic signature. Fourth, although the six-based genes risk signature indicated a fair predictive ability for 5-year survival, more key factors are still needed to be brought into analysis. This is owing to the poor performance in predicting the survival outcome in recurrent LGG. Thus, it is reasonable to aim to utilize more factors into building a prognostic model that could enable risk stratification of recurrent LGG. Fifth, as molecular mechanism have not been investigated in the current study, it is necessary to explore the underlying mechanisms behind the risk scores and poor survival outcomes of LGG in further *in vitro* or *in vivo* experiments. Sixth, the "sva" package was applied in current study to remove the batch effects of Level 3 data from TCGA and GTEx. Despite the two groups separated obviously, however, several outliers can be found in the PCA plots. It should be noted that the reasons of several outliers may be caused by the insufficient batch effect removal of Level 3 data by "sva" (Wang et al., 2018) or others such as different parts of brain tissues or lacking reference of normal controls in TCGA, all of them warranting further investigations.

CONCLUSION

In this study, we demonstrated that a six immune-related genes based risk signature might be effective in risk stratification and in serving as an independent prognostic factor of the overall survival in patients with primary LGG. Further *in vitro* and *in vivo* experiments are warranted to explore the underlying mechanisms behind immune genes and survival outcome in primary LGG.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The RNA-seq data (level 3) and clinical information of LGG samples can be found in UCSC Xena (http://xena.ucsc.edu/), and the CGGA database (http://www.cgga.org.cn). The immune-related genes available at https://immport.niaid.nih.gov. The mRNA expression of genes profiled by RNA-Seq available at https://portals.broadinstitute.org/ccle.

ETHICS STATEMENT

All the information of patients was obtained from Chinese Glioma Genome Atlas (CGGA), and The Cancer Genome Atlas (TCGA). All the patients and treatments were complied with the principles laid down in the Declaration of Helsinki in 1964 and its later amendments or comparable ethical standards.

AUTHOR CONTRIBUTIONS

MZ analyzed the data. MZ, XW, and JH contributed materials or analysis tools. XW prepared the figures and tables. MZ, XC, and XW authored or reviewed drafts of the manuscript. QZ and JH conceived and designed the study. QZ revised the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (11601083 and U1805263), the program for Probability and Statistics: Theory and Application (IRTL1704), Innovative Research Team in Science and Technology in Fujian Province University (IRTSTFJ), the Science Foundation for Young Scientists of Fujian Health and Family Planning Commission (Grant No. 2018-2-17), the Key Research Program of Fujian Provincial Heath and Education Joint Committee (Grant No. WKJ2016-2-31), National Collaboration Center in Immuno-Oncology (Grant No. 2016sysbz02), and Qihang Foundation of Fujian Medical University (Grant No. 2018QH1088).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00363/full#supplementary-material

FIGURE S1 | The flowchart of the project.

FIGURE S2 | The normalization and batch effect removal from TCGA and GTEx datasets. **(A)** Box plots illustrated the data distributions from TCGA and GTEx datasets before normalization. **(B)** PCA plot illustrated the cluster of the samples from TCGA and GTEx datasets before batch effect removal. **(C)** Box plots illustrated the data distributions from TCGA and GTEx datasets after normalization. **(D)** PCA plot illustrated the cluster of the samples from TCGA and GTEx datasets after batch effect removal.

FIGURE S3 | (A) Heatmaps showing that the 5,490 differentially expressed genes (DEGs) can effectively distinguish tumors from non-tumor tissues after integrated analysis. **(B)** Volcano plot presenting DEGs between LGG and non-tumor tissues. Red dots, and green dots represent up-regulated genes, and down-regulated genes, respectively.

FIGURE S4 | Schoenfeld residual plots showing *P* value of all factors were greater to 0.05.

FIGURE S5 | Investigating the application of six genes based signature in recurrent LGG. **(A)** Kaplan-Meier plot for overall survival based on risk score of the six gene based signature of recurrent LGG patients in CGGA cohort. **(B)** ROC curve based on the risk score for diagnostic competence verification of recurrent LGG patients in CGGA cohort. **(C)** Time-dependent ROC curve based on the six genes based risk score for 3-, and 5-year overall survival probability of recurrent LGG patients in CGGA cohort. Calibration curve for predicting probabilities of patients' 3-year **(D)**, and 5-year **(E)** overall survival of recurrent LGG patients in CGGA cohort.

FIGURE S6 | Association between risk score and clinical-pathological parameters. Association between risk score and age, gender, grade, radiotherapy, chemotherapy, and IDH mutation status of primary LGG patients in TCGA cohort **(A)**, in CGGA cohort **(B)**, while patients of recurrent LGG patients in CGGA cohort are shown in **(C)**.

FIGURE S7 | The differential expressed T cell associated activated and inhibitory genes, macrophage chemo-attractant and phagocytosis related genes between high and low risk groups in primary LGG.

FIGURE S8 | Expression data were sorted by the tumor type. The expression of the CANX **(A)**, HSPA1B **(B)**, KLRC2 **(C)**, PSMC6 **(D)**, RFXAP **(E)**, and TAP1 **(F)** in Cancer Cell Line Encyclopedia.

FIGURE S9 | Number of patients with staining **(A)**. The typical protein expression of six genes of immunohistochemistry (IHC) images in LGG tissue and paired non-tumor samples **(B)**. Data was queried from the human protein atlas (https://www.proteinatlas.org/).

REFERENCES

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

Ascierto, M. L., Kmieciak, M., Idowu, M. O., Manjili, R., Zhao, Y., Grimes, M., et al. (2012). A signature of immune function genes associated with recurrence-free survival in breast cancer patients. *Breast Cancer Res. Treat.* 131, 871–880. doi: 10.1007/s10549-011-1470-x

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

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

Borrego, F., Masilamani, M., Marusina, A. I., Tang, X., and Coligan, J. E. (2006). The CD94/NKG2 family of receptors: from molecules and cells to clinical relevance. *Immunol. Res.* 35, 263–278. doi: 10.1385/ir:35:3:263

Chen, H. Y., Yu, S. L., Chen, C. H., Chang, G. C., Chen, C. Y., Yuan, A., et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.* 356, 11–20. doi: 10.1056/NEJMoa060096

Chen, Y., Ma, D., Wang, X., Fang, J., Liu, X., Song, J., et al. (2019). Calnexin impairs the antitumor immunity of CD4(+) and CD8(+) T cells. *Cancer Immunol. Res.* 7, 123–135. doi: 10.1158/2326-6066.cir-18-0124

Cheng, W., Ren, X., Zhang, C., Cai, J., Liu, Y., Han, S., et al. (2016). Bioinformatic profiling identifies an immune-related risk signature for glioblastoma. *Neurology* 86, 2226–2234. doi: 10.1212/WNL.0000000000002770

Claus, E. B., Walsh, K. M., Wiencke, J. K., Molinaro, A. M., Wiemels, J. L., Schildkraut, J. M., et al. (2015). Survival and low-grade glioma: the emergence of genetic information. *Neurosurg. Focus* 38:E6. doi: 10.3171/2014.10.FOCUS12367

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

Demeure, K., Fack, F., Duriez, E., Tiemann, K., Bernard, A., Golebiewska, A., et al. (2016). Targeted proteomics to assess the response to anti-angiogenic treatment in human glioblastoma (GBM). *Mol. Cell. Proteomics* 15, 481–492. doi: 10.1074/mcp.M115.052423

Ding, G., Zhou, L., Qian, Y., Fu, M., Chen, J., Chen, J., et al. (2015). Pancreatic cancer-derived exosomes transfer miRNAs to dendritic cells and inhibit RFXAP expression via miR-212-3p. *Oncotarget* 6, 29877–29888. doi: 10.18632/oncotarget.4924

Ding, G., Zhou, L., Shen, T., and Cao, L. (2018). IFN-gamma induces the upregulation of RFXAP via inhibition of miR-212-3p in pancreatic cancer cells: a novel mechanism for IFN-gamma response. *Oncol. Lett.* 15, 3760–3765. doi: 10.3892/ol.2018.7777

Ding, P., Wang, W., Wang, J., Yang, Z., and Xue, L. (2014). Expression of tumor-associated macrophage in progression of human glioma. *Cell Biochem. Biophys.* 70, 1625–1631. doi: 10.1007/s12013-014-0105-3

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21, 938–945. doi: 10.1038/nm.3909

Guadagno, E., Presta, I., Maisano, D., Donato, A., Pirrone, C. K., Cardillo, G., et al. (2018). Role of macrophages in brain tumor growth and progression. *Int. J. Mol. Sci.* 19:1005. doi: 10.3390/ijms19041005

Guo, X., and Dixon, J. E. (2016). The 26S proteasome: a cell cycle regulator regulated by cell cycle. *Cell Cycle* 15, 875–876. doi: 10.1080/15384101.2016.1151728

Hainfellner, J., Louis, D. N., Perry, A., and Wesseling, P. (2014). Letter in response to David N. Louis et al., international society of neuropathology-Haarlem consensus guidelines for nervous system tumor classification and grading. *Brain Pathol.* 24, 671–672. doi: 10.1111/bpa.12187

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

Harrell, F. E. Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models,evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15, 361–387. doi: 10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168<3.0.co;2-4

Hartmann, C., Hentschel, B., Wick, W., Capper, D., Felsberg, J., Simon, M., et al. (2010). Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than IDH1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas. *Acta Neuropathol.* 120, 707–718. doi: 10.1007/s00401-010-0781-z

Hu, F., Zeng, W., and Liu, X. (2019). A gene signature of survival prediction for kidney renal cell carcinoma by multi-omic data analysis. *Int. J. Mol. Sci.* 20:5720. doi: 10.3390/ijms20225720

Iwata, R., Lee, J. H., Hayashi, M., Dianzani, U., Ofune, K., Maruyama, M., et al. (2019). ICOSLG-mediated regulatory T cell expansion and IL-10 production promote progression of glioblastoma. *Neuro Oncol.* 22, 333–344. doi: 10.1093/neuonc/noz204

Jego, G., Hermetet, F., Girodon, F., and Garrido, C. (2019). Chaperoning STAT3/5 by heat shock proteins: interest of their targeting in cancer therapy. *Cancers (Basel)* 12:21. doi: 10.3390/cancers12010021

Kim, K., Jeon, S., Kim, T. M., and Jung, C. K. (2018). Immune gene signature delineates a subclass of papillary thyroid cancer with unfavorable clinical outcomes. *Cancers (Basel)* 10:494. doi: 10.3390/cancers10120494

Kmiecik, J., Poli, A., Brons, N. H., Waha, A., Eide, G. E., Enger, P. O., et al. (2013). Elevated CD3+ and CD8+ tumor-infiltrating immune cells correlate with prolonged survival in glioblastoma patients despite integrated immunosuppressive mechanisms in the tumor microenvironment and at the systemic level. *J. Neuroimmunol.* 264, 71–83. doi: 10.1016/j.jneuroim.2013.08.013

Leone, P., Shin, E. C., Perosa, F., Vacca, A., Dammacco, F., and Racanelli, V. (2013). MHC class I antigen processing and presenting machinery: organization, function, and defects in tumor cells. *J. Natl. Cancer Inst.* 105, 1172–1187. doi: 10.1093/jnci/djt184

Li, X., and Meng, Y. (2019). Survival analysis of immune-related lncRNA in low-grade glioma. *BMC Cancer* 19:813. doi: 10.1186/s12885-019-6032-3

Liang, J., Piao, Y., Holmes, L., Fuller, G. N., Henry, V., Tiao, N., et al. (2014). Neutrophils promote the malignant glioma phenotype through S100A4. *Clin. Cancer Res.* 20, 187–198. doi: 10.1158/1078-0432.CCR-13-1279

Ling, A., Lofgren-Burstrom, A., Larsson, P., Li, X., Wikberg, M. L., Oberg, A., et al. (2017). TAP1 down-regulation elicits immune escape and poor prognosis in colorectal cancer. *Oncoimmunology* 6:e1356143. doi: 10.1080/2162402x.2017.1356143

Livneh, I., Cohen-Kaplan, V., Cohen-Rosenzweig, C., Avni, N., and Ciechanover, A. (2016). The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. *Cell Res.* 26, 869–885. doi: 10.1038/cr.2016.86

Lossos, I. S., Czerwinski, D. K., Alizadeh, A. A., Wechser, M. A., Tibshirani, R., Botstein, D., et al. (2004). Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.* 350, 1828–1837. doi: 10.1056/NEJMoa032520

Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1

Mahmoodzadeh Hosseini, H., Halabian, R., Amin, M., and Imani Fooladi, A. A. (2015). Texosome-based drug delivery system for cancer therapy: from past to present. *Cancer Biol. Med.* 12, 150–162. doi: 10.7497/j.issn.2095-3941.2015.0045

National Comprehensive Cancer Network (2019). *NCCN Clinical Practice Guidelines in Oncology: Central Nervous System Cancers Version 1.2019*. Available online at: https://www.nccn.org/professionals/physician_gls/pdf/cns.pdf (accessed December 21, 2019).

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* 37, 773–782. doi: 10.1038/s41587-019-0114-2

Ostrom, Q. T., Gittleman, H., Farah, P., Ondracek, A., Chen, Y., Wolinsky, Y., et al. (2013). CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2006-2010. *Neuro Oncol.* 15(Suppl. 2), ii1–ii56. doi: 10.1093/neuonc/not151

Patel, V. N., Gokulrangan, G., Chowdhury, S. A., Chen, Y., Sloan, A. E., Koyuturk, M., et al. (2013). Network signatures of survival in glioblastoma multiforme. *PLoS Comput. Biol.* 9:e1003237. doi: 10.1371/journal.pcbi.1003237

Perus, L. J. M., and Walsh, L. A. (2019). Microenvironmental heterogeneity in brain malignancies. *Front. Immunol.* 10:2294. doi: 10.3389/fimmu.2019.02294

Qian, Z., Li, Y., Fan, X., Zhang, C., Wang, Y., Jiang, T., et al. (2018). Molecular and clinical characterization of IDH associated immune signature in lower-grade gliomas. *Oncoimmunology* 7:e1434466. doi: 10.1080/2162402X.2018.1434466

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing, 3.1.1*. Vienna: R Foundation for Statistical Computing.

Reznik, E., Smith, A. W., Taube, S., Mann, J., Yondorf, M. Z., Parashar, B., et al. (2018). Radiation and immunotherapy in high-grade gliomas: where do we stand? *Am. J. Clin. Oncol.* 41, 197–212. doi: 10.1097/COC.0000000000000406

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241. doi: 10.1093/biomet/69.1.239

Shen, S., Wang, G., Zhang, R., Zhao, Y., Yu, H., Wei, Y., et al. (2019). Development and validation of an immune gene-set based Prognostic signature in ovarian cancer. *EBioMedicine* 40, 318–326. doi: 10.1016/j.ebiom.2018.12.054

Simonelli, M., Persico, P., Perrino, M., Zucali, P. A., Navarria, P., Pessina, F., et al. (2018). Checkpoint inhibitors as treatment for malignant gliomas: "A long way to the top". *Cancer Treat. Rev.* 69, 121–131. doi: 10.1016/j.ctrv.2018.06.016

Sota, Y., Naoi, Y., Tsunashima, R., Kagara, N., Shimazu, K., Maruyama, N., et al. (2014). Construction of novel immune-related signature for prediction of pathological complete response to neoadjuvant chemotherapy in human breast cancer. *Ann. Oncol.* 25, 100–106. doi: 10.1093/annonc/mdt427

Surmann, E. M., Voigt, A. Y., Michel, S., Bauer, K., Reuschenbach, M., Ferrone, S., et al. (2015). Association of high CD4-positive T cell infiltration with mutations in HLA class II-regulatory genes in microsatellite-unstable colorectal cancer. *Cancer Immunol. Immunother.* 64, 357–366. doi: 10.1007/s00262-014-1638-4

Therneau, G. T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–526. doi: 10.2307/2337123

van den Bent, M. J. (2014). Practice changing mature results of RTOG study 9802: another positive PCV trial makes adjuvant chemotherapy part of standard of care in low-grade glioma. *Neuro Oncol.* 16, 1570–1574. doi: 10.1093/neuonc/nou297

van den Bent, M. J., Baumert, B., Erridge, S. C., Vogelbaum, M. A., Nowak, A. K., Sanson, M., et al. (2017). Interim results from the CATNON trial (EORTC study 26053-22054) of treatment with concurrent and adjuvant temozolomide for 1p/19q non-co-deleted anaplastic glioma: a phase 3, randomised, open-label

intergroup study. *Lancet* 390, 1645–1653. doi: 10.1016/S0140-6736(17)31442-3

Vismara, M. F. M., Donato, A., Malara, N., Presta, I., and Donato, G. (2019). Immunotherapy in gliomas: are we reckoning without the innate immunity? *Int. J. Immunopathol. Pharmacol.* 33:2058738419843378. doi: 10.1177/2058738419843378

Wang, H., Xu, T., Huang, Q., Jin, W., and Chen, J. (2020). Immunotherapy for malignant glioma: current status and future directions. *Trends Pharmacol. Sci.* 41, 123–138. doi: 10.1016/j.tips.2019.12.003

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., et al. (2018). Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* 5:180061. doi: 10.1038/sdata.2018.61

Wherry, E. J., and Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.* 15, 486–499. doi: 10.1038/nri3862

Wick, W., Meisner, C., Hentschel, B., Platten, M., Schilling, A., Wiestler, B., et al. (2013). Prognostic or predictive value of MGMT promoter methylation in gliomas depends on IDH1 mutation. *Neurology* 81, 1515–1522. doi: 10.1212/WNL.0b013e3182a95680

Wischhusen, J., Friese, M. A., Mittelbronn, M., Meyermann, R., and Weller, M. (2005). HLA-E protects glioma cells from NKG2D-mediated immune responses in vitro: implications for immune escape in vivo. *J. Neuropathol. Exp. Neurol.* 64, 523–528. doi: 10.1093/jnen/64.6.523

Wu, X., Fan, Z., Chen, M., Chen, Y., Rong, D., Cui, Z., et al. (2019). Forkhead transcription factor FOXO3a mediates interferon-gamma-induced MHC II transcription in macrophages. *Immunology* 158, 304–313. doi: 10.1111/imm.13116

Xu, Y. Y., Gao, P., Sun, Y., and Duan, Y. R. (2015). Development of targeted therapies in treatment of glioblastoma. *Cancer Biol. Med.* 12, 223–237. doi: 10.7497/j.issn.2095-3941.2015.0020

Yang, W., Lai, Z., Li, Y., Mu, J., Yang, M., Xie, J., et al. (2019). Immune signature profiling identified prognostic factors for gastric cancer. *Chin. J. Cancer Res.* 31, 463–470. doi: 10.21147/j.issn.1000-9604.2019.03.08

Yang, Z., Zhuang, L., Szatmary, P., Wen, L., Sun, H., Lu, Y., et al. (2015). Upregulation of heat shock proteins (HSPA12A, HSP90B1, HSPA4, HSPA5 and HSPA6) in tumour tissues is associated with poor outcomes from HBV-related early-stage hepatocellular carcinoma. *Int. J. Med. Sci.* 12, 256–263. doi: 10.7150/ijms.10735

Zeng, W. J., Yang, Y. L., Liu, Z. Z., Wen, Z. P., Chen, Y. H., Hu, X. L., et al. (2018). Integrative analysis of DNA methylation and gene expression identify a three-gene signature for predicting prognosis in lower-grade gliomas. *Cell. Physiol. Biochem.* 47, 428–439. doi: 10.1159/000489954

Zhu, C., Mustafa, D. A. M., Krebber, M. M., Chrifi, I., Leenen, P. J. M., Duncker, D. J., et al. (2018). Comparative proteomic analysis of cat eye syndrome critical region protein 1- function in tumor-associated macrophages and immune response regulation of glial tumors. *Oncotarget* 9, 33500–33514. doi: 10.18632/oncotarget.26063

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: 21 April 2020
doi: 10.3389/fgene.2020.00343

A Novel Computational Approach for Identifying Essential Proteins From Multiplex Biological Networks

Bihai Zhao[1,2,3†], *Sai Hu*[1†], *Xiner Liu*[1], *Huijun Xiong*[1], *Xiao Han*[1], *Zhihong Zhang*[1,2], *Xueyong Li*[1] *and Lei Wang*[1,2*]

[1] College of Computer Engineering and Applied Mathematics, Changsha University, Changsha, China, [2] Hunan Provincial Key Laboratory of Industrial Internet Technology and Security, Changsha University, Changsha, China, [3] Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals, Changsha University, Changsha, China

The identification of essential proteins can help in understanding the minimum requirements for cell survival and development. Ever-increasing amounts of high-throughput data provide us with opportunities to detect essential proteins from protein interaction networks (PINs). Existing network-based approaches are limited by the poor quality of the underlying PIN data, which exhibits high rates of false positive and false negative results. To overcome this problem, researchers have focused on the prediction of essential proteins by combining PINs with other biological data, which has led to the emergence of various interactions between proteins. It remains challenging, however, to use aggregated multiplex interactions within a single analysis framework to identify essential proteins. In this study, we created a multiplex biological network (MON) by initially integrating PINs, protein domains, and gene expression profiles. Next, we proposed a new approach to discover essential proteins by extending the random walk with restart algorithm to the tensor, which provides a data model representation of the MON. In contrast to existing approaches, the proposed MON approach considers for the importance of nodes and the different types of interactions between proteins during the iteration. MON was implemented to identify essential proteins within two yeast PINs. Our comprehensive experimental results demonstrated that MON outperformed 11 other state-of-the-art approaches in terms of precision-recall curve, jackknife curve, and other criteria.

Keywords: identification of essential proteins, protein interaction network, tensor, multiplex biological networks, random walk, Markov chain, gene expression, yeast

OPEN ACCESS

Edited by:
Ling Kui,
Harvard Medical School,
United States

Reviewed by:
Juan Ye,
National Institutes of Health (NIH),
United States
Weiyu Chen,
Stanford University, United States

****Correspondence:***
Lei Wang
wanglei@xtu.edu.cn

† These authors have contributed equally to this work

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

Received: *12 February 2020*
Accepted: *23 March 2020*
Published: *21 April 2020*

Citation:
Zhao B, Hu S, Liu X, Xiong H, Han X, Zhang Z, Li X and Wang L (2020) A Novel Computational Approach for Identifying Essential Proteins From Multiplex Biological Networks. Front. Genet. 11:343. doi: 10.3389/fgene.2020.00343

INTRODUCTION

Essential proteins are necessary for the survival of living organisms. The identification of essential proteins can help us to understand the basic requirements of living organisms, and it can also play an important role in drug design (Dubach et al., 2017), genetic disease diagnosis (Zeng et al., 2017), and drug synergy prediction in cancers (Li et al., 2018). Traditional experimental approaches, such as gene knockouts (Narasimhan et al., 2016), RNA interference (Inouye, 2016), and Knockout Sudoku (Baym et al., 2016), are time-consuming and costly. Over the last few decades, high-throughput technologies have produced a tremendous amount of protein interaction network (PIN) data that provide us with new opportunities to detect essential proteins through

the use of computational approaches. A number of network topology-based centrality approaches have been proposed to predict essential proteins, and these approaches include Degree Centrality (DC) (Hahn and Kern, 2004), Information Centrality (IC) (Stephenson and Zelen, 1989), Closeness Centrality (CC) (Wuchty and Stadler, 2003), Betweenness Centrality (BC) (Joy et al., 2005), Subgraph Centrality (SC) (Estrada and Rodriguez-Velazquez, 2005), and Neighbor Centrality (NC) (Wang et al., 2011).

Unfortunately, these approaches are often plagued by noise and errors**,** which can result in biases and low confidence in protein–protein interaction (PPI) networks. To provide accurate prediction results, the integration of different types of biological data has become an important and popular strategy. A number of approaches have been developed to facilitate the prediction of essential proteins by combining PINs with multisource biological data. For example, Gene Ontology (GO) annotations were used as a bioinformatics tool to predict essential proteins in several single-cell PINs, such as those from *Escherichia coli*, *Saccharomyces cerevisiae*, and *Drosophila melanogaster* (Hsing et al., 2008). A prediction model called integrating orthology with PPI network (ION) (Peng et al., 2012) was proposed to infer essential proteins by integrating orthologous information and the topological characteristics of PINs. In the United complex Centrality (UC) (Li et al., 2015) method, protein complexes were also combined with the topological features of PINs to detect essential genes. After analyzing the correlations between domain characteristics and essential proteins, Peng et al. (2015) designed an approach named unite domain and network centrality (UDoNC) for the prediction of essential proteins in yeast PINs. Li et al. (2012) and Zhang et al. (2013) developed two types of prediction models called prediction of essential proteins centrality (PeC) and co-expression weighted by clustering coefficient method (CoEWC) to infer essential proteins by fusing gene expressions and topological characteristics of PINs, respectively. In our previous studies, we proposed a prediction method called predictive model based on overlapping essential modules (POEM) (Zhao et al., 2014) to measure the essentiality of proteins by detecting overlapping essential modules based on the modularity of essential proteins. Lei et al. (2018) designed a method called AFSO_EP for the prediction of essential proteins based on the artificial fish-swarm algorithm. In this method, the network topology, gene expression, GO annotation, and subcellular localization information were utilized. Zhang W. et al. (2019) proposed a new method to discover essential proteins, named predicting essential proteins by integrating network topology, expression profile, GO annotation and subcellular localization (TEGS), based on integrating network topology, gene expression profiles, GO annotation information, and protein subcellular localization information. In the fusing the dynamic PPI networks (FDP) approach Zhang F. et al. (2019), active PINs were constructed first and then they were fused into a final network according to the networks' similarities. Finally, a new approach for identification of essential proteins was proposed by considering orthologous property and topological properties in the network.

A common characteristic and limitation of these approaches, however, is that they complete the prediction of essential proteins using only a single network of relationships between proteins. Currently, PINs are not the only large-scale network datasets, as protein–DNA interactions and signaling-regulatory pathway interaction data are also stored in dedicated databases (Valdeolivas et al., 2019). Additionally, other interactions such as the co-expression network established from gene expression profiles and the co-annotation network constructed from GO annotations can be derived. Each interaction data source has its own meaning or relevance and can play a different role in the prediction of essential proteins. These approaches mentioned above classically aggregated multiple interaction networks into a single and unique network, which tends to dismiss the topologies and features of the individual interaction networks. The convention of representing different types of interactions in a system with a single type of link is no longer a panacea for network science (De Domenico et al., 2015). The multiplex network offers us an alternative, in that it is a collection of networks sharing the same nodes; however, the edges belong to different categories or represent interactions of different natures (Didier et al., 2015). More recently, various applied studies have been adapted to multiplex networks. Valdeolivas et al. (2019) extended the Random walk algorithm to multiplex networks by building an $nL \times nL$ heterogeneous matrix in which n and L represent the number of nodes and layers of the multiplex network, respectively. Wang et al. (2018) compressed the multiple networks into two feature matrices and performed conserved functional modules detection by multi-view non-negative matrix factorization. In a newly proposed link prediction algorithm (Samei and Jalili, 2019) for multiplex networks, both intra-layer information and inter-layer information are combined based on layer relevance. In our previous work, we constructed a multilayer protein network and applied it for the detection of protein complexes (Li et al., 2016) and for the prediction of protein functions (Zhao et al., 2016a). In this study, we propose a tensorial framework to represent the newly constructed multiplex biological network, and we aim to apply it for the identification of essential proteins by extending the random walk with restart algorithm. Our experimental results demonstrated that our proposed MON approach outperformed six types of centrality approaches, including DC (Hahn and Kern, 2004), IC (Stephenson and Zelen, 1989), CC (Wuchty and Stadler, 2003), BC (Joy et al., 2005), SC (Estrada and Rodriguez-Velazquez, 2005), and NC (Wang et al., 2011) and five types of network topological features and biological data sources fusion-based approaches such as PeC (Li et al., 2012), CoEWC (Zhang et al., 2013), POEM (Zhao et al., 2014), ION (Peng et al., 2012), and FDP (Zhang F. et al., 2019).

MATERIALS

To estimate the performance of MON, we used it to identify essential proteins in the PIN of *Saccharomyces cerevisiae* that was derived from the database of interacting proteins (DIP) (Xenarios et al., 2002) and Gavin datasets (Gavin et al., 2006). The PINs from *Saccharomyces cerevisiae,* which have been well-characterized by a number of studies, are the most complete and comprehensive. After removing self-interactions and repeated interactions, the DIP dataset finally obtained 5,093 proteins and

TABLE 1 | Details of two yeast protein interaction networks.

Dataset	Proteins	Interactions	Essential proteins	Expressed proteins
DIP	5,093	24,753	1,167	4,985
Gavin	1,855	7,669	714	1,927

24,743 interactions, and the Gavin dataset consisted of 1,855 proteins and 7,669 interactions. The domain data for building the multiplex biological network was downloaded from the Pfam database (Punta et al., 2011). The gene expression profile (Tu et al., 2005) of the yeast was derived from GSE3431 in the GEO (Gene Expression Omnibus) that contained the expression values of 6,776 genes at 36 moments, where 4,985 and 1,827 of these genes were located in the DIP and Gavin PINs, respectively. The gene coverage rates of the two PINs in gene expression profile were all >95% (DIP: 4,985/5,093 = 97.88%, Gavin: 1,827/1,855 = 98.49%). Information on orthologous proteins was obtained from the InParanoid database (Östlund et al., 2009) (Version 7) that consisted of a collection of pairwise comparisons between 100 whole genomes. A benchmark set of essential proteins from *Saccharomyces cerevisiae* that consisted of 1,285 essential proteins was derived from the MIPS (MIPS: analysis and annotation of proteins from whole genomes in 2005) (Mewes et al., 2006), saccharomyces genome database (SGD) (Cherry et al., 2011), and database of essential genes (DEG) (Zhang and Lin, 2008) databases. Among the 5,093 proteins in the DIP network, 1,167 proteins were essential and 3,526 proteins were non-essential. In the Gavin dataset, the number of essential proteins and non-essential proteins was 714 and 1,141, respectively. **Table 1** lists the details of the two yeast PINs.

METHODS

The outline for the entire MON approach includes (1) establishing a multiplex biological network by integrating the topology of PINs, protein domains, and gene expression profile, (2) extending the random walk with restart algorithm to the tensor model corresponding to the multiplex biological network, and (3) sorting proteins in descending order, with the top K of these proteins being exported. The flowchart for the MON approach is provided in **Figure 1**.

Construction of Multiplex Biological Networks

For our purpose, we consider a multiplex biological network $G = (G^1, G^2, \ldots, G^L)$, where $G^i = (V, E^i)$ represents the network of the layer of i. $V = \{v_1, v_2, \ldots, v_n\}$ is a set of sharing proteins for all layers in G, and $E^i = \{e_{i1}, e_{i2}, \ldots, e_{im}\}$ is a set of interactions at i-th layer in the multiplex biological network G.

In this study, we constructed a multiplex biological network $G = (G^1, G^2, G^3)$ by integrating PINs, gene expression profiles, and protein domain information. In the first layer, a co-neighbor network (CN) was established through the analysis of the topology characteristics of PINs, while in the second layer, a co-structure network was constructed according to the correlation analysis based on the protein domain information. In the third layer, a co-expression network was related to the property of co-expression derived from time course gene expression profiles.

Co-neighbor Network G^1

The CN was established by exploring common neighbors between pairs of proteins. Intuitively, the greater number of common neighbors that the two proteins possess, the more credible the interactions between these two proteins will be. If two proteins p_i and $p_j$interact with each other in PINs and share at least one common neighbor, they will connect to each other within the CN. The weight of interaction between p_i and p_j can be calculated by the following formula:

$$e^1(i,j) = \begin{cases} \frac{|N_i \bigcap N_j|^2}{(|N_i|-1)\times(|N_j|-1)}, & if\,|N_i \bigcap N_j| > 0 \\ 0 \quad , & otherwise \end{cases} \tag{1}$$

where N_i and N_j represent the direct neighbors set of p_i and p_j, respectively, and $N_i \cap N_j$ denotes the common neighbors set for protein p_i and protein p_j.

Co-structure Network G^2

Domains are sequential and structural motifs that are found independently in different proteins and act as the stable functional blocks of proteins. Based on this, we created the co-structure network based on data from protein domains. First, we analyzed the importance of proteins relative to the domains based on the association between proteins and domains. Given a protein p_i, its domain score P_D can be calculated as follows:

$$P\_D(P_i) = \sum_{j=1}^{|D|} \frac{1}{NP_j} \times t_{ij} \tag{2}$$

In Equation (2), D is a list of distinct categories of domains related to all proteins. NP_j is the number of proteins that contain the domain d_j. If the protein p_i contains the domain d_j, t_{ij} is assigned the value of 1. Otherwise, t_{ij} is set to 0. Finally, the P_D score of p_i can be normalized and calculated as follows:

$$P\_D(p_i) = \frac{P\_D(p_i) - \min\limits_{1\le j\le |P|}(P\_D(p_j))}{\max\limits_{1\le j\le |P|}(P\_D(p_j)) - \min\limits_{1\le j\le |P|}(P\_D(p_j))} \tag{3}$$

From the above equation, we can easily determine that the value of P_D falls into the interval [0, 1]. From this perspective, the P_D score of a protein can be interpreted as its probability of becoming an essential protein. Moreover, previous studies (Stephenson and Zelen, 1989) have indicated that essential genes or proteins tend to form essential modules through their interactions. We assumed that the essential probabilities of proteins mentioned above were independent of each other. The probability (or weight) of interaction between two proteins p_i and p_j in the co-structure network can be calculated as follows.

$$e^2(i,j) = P\_D(p_i) \times P\_D(p_j) \tag{4}$$

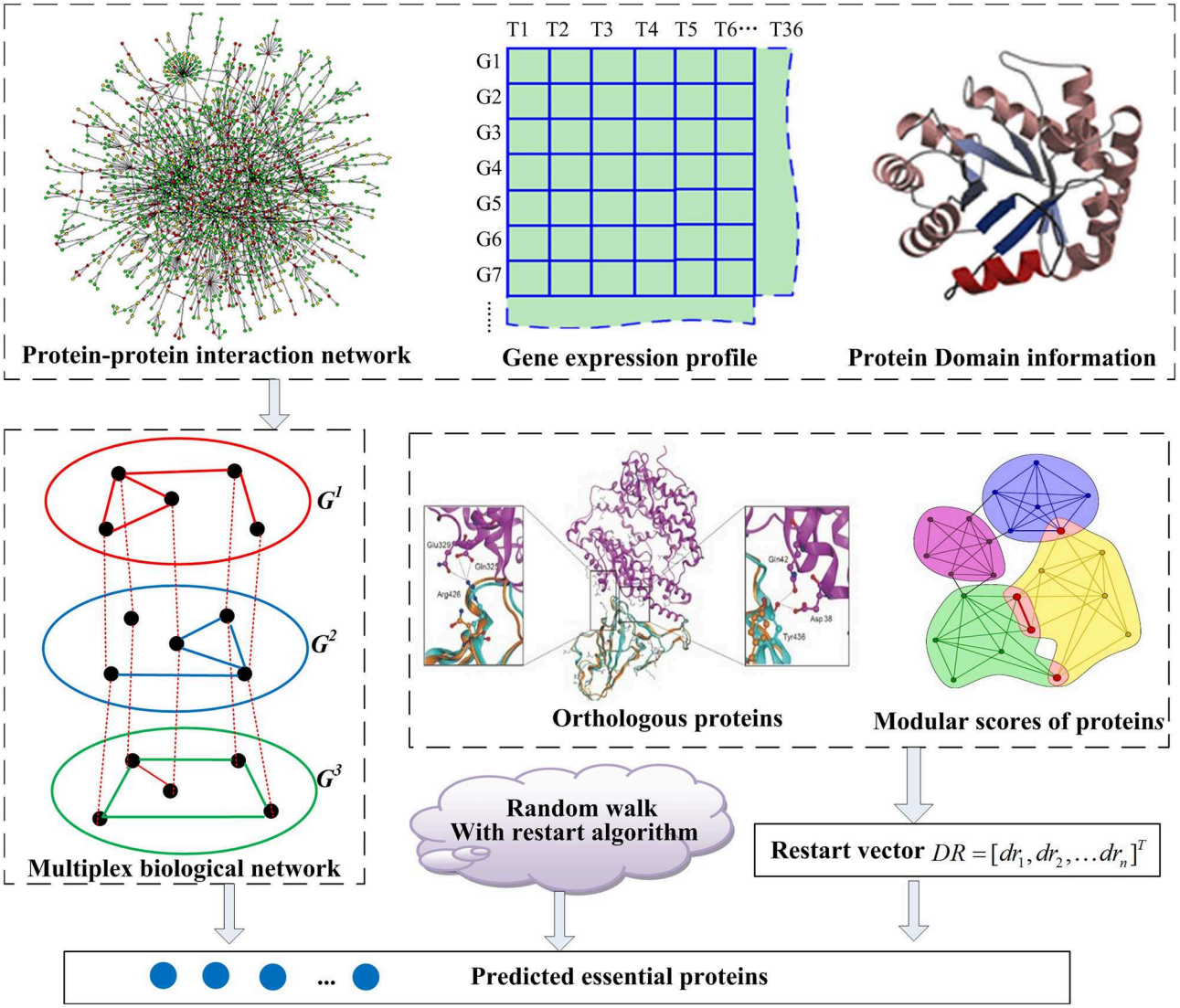


FIGURE 1 | The flowchart of MON method. A multiplex biological network $G = (G^1, G^2, G^3)$ is constructed with integration of protein interaction networks (PINs), gene expression profile, and protein domain information, firstly. And then, a restart vector is established according to orthologous proteins and module scores of proteins. Based on these, the random walk with restart algorithm is applied to score and rank essential proteins.

Co-expression Network G^3

The Pearson's correlation coefficient (PCC) was adopted to evaluate the co-expression probability of a pair of proteins based on gene expression profiles. Let $g(p_i, j)$ denote the expression value of the gene p_i at the j-th time point, and then for a pair of genes p_i and p_j, the correlation between them can be calculated as follows:

$$PCC(p_i, p_j) = \frac{n\sum g(p_i,k)g(p_j,k) - \sum g(p_i,k)\sum g(p_j,k)}{\sqrt{n\sum g(p_i,k)^2 - (\sum g(p_i,k))^2}\sqrt{n\sum g(p_j,k)^2 - (\sum g(p_j,k))^2}} \quad (5)$$

Two proteins were regarded as co-expressed if they interacted with each other in the original PINs and their correlation coefficient was not zero. The weight of interaction between p_i and p_j in the co-expression network was set to the absolute value of their correlation coefficient. Specifically, $e^3(i, j) = |PCC(p_i, p_j)|$.

Random Walk With Restart on Multiplex Biological Networks

To study the multiplex network systematically, it is necessary to develop a precise mathematical model and appropriate tools. In this paper, we represent the newly constructed multiplex biological network G using the tensor model and extend the random walk with restart algorithm.

Let $T = (t_{ijk}) \in \mathbb{R}^{n\times n\times m}$ denote the three-order adjacency tensor corresponding to the multiplex biological network $G = (G^1, G^2, G^3)$, where n and m are the number of proteins and categories of interactions between proteins, respectively. Each element of T is defined as follows:

$$t_{ijk} = \begin{cases} e^k(i,j) & , \ if(p_i, p_j) \in E^k \\ 0 & , \ otherwise \end{cases} \tag{6}$$

Here $1 \le i, j \le n$, $1 \le k \le m$ ($m = 3$) and e^k (i, j) represents the weight of interaction between p_i and p_j at the k-th layer. We can thus extend the random walk with restart algorithm from a two-dimensional matrix to the tensor for scoring proteins. Studies show that the structural characteristics of different layers in multiplex networks are indeed correlated to each other (Jalili et al., 2017). Based on this, we propose that considering the importance of different types of interactions can enhance the performance for the discovery of essential proteins. Our statistics revealed mutually reinforcing relationships between important or key nodes with different types of links pointed to them in multiplex biological networks. Let the vectors $x = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ and $y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$ denote important scores of proteins and different categories of interactions between proteins, respectively. We formally described the relationships between x and y based on the tensor T using the following equation:

$$x = f(T, x, y), y = g(T, x) \tag{7}$$

The most critical task for us was to design reasonable functions f and g and to calculate y and z, respectively. We now propose the idea to define a higher-order Markov chain by normalizing the tensor. This leads to two probability transition tensors $T^{(1)} = ({t^{(1)}}_{ijk}) \in \mathbb{R}^{n\times n\times l}$ and $T^{(2)} = ({t^{(2)}}_{ijk}) \in \mathbb{R}^{n\times n\times l}$ that are calculated as follows:

$$t^{(1)}_{i,j,k} = \begin{cases} \frac{t_{i,j,k}}{\sum_{i=1}^{n} t_{i,j,k}} \ if \sum_{i=1}^{n} t_{i,j,k} > 0 \\ 1/n \ otherwise \end{cases} \tag{8}$$

$$t^{(2)}_{i,j,k} = \begin{cases} \frac{t_{i,j,k}}{\sum_{k=1}^{m} t_{i,j,k}} \ if \sum_{k=1}^{m} t_{i,j,k} > 0 \\ 1/m \ otherwise \end{cases} \tag{9}$$

We can then easily obtain the following formulas:

$$0 \le t^{(1)}_{i,j,k} \le 1, \ \sum_{i=1}^{n} t^{(1)}_{i,j,k} = 1 \tag{10}$$

$$0 \le t^{(2)}_{i,j,k} \le 1, \ \sum_{k=1}^{m} t^{(2)}_{i,j,k} = 1 \tag{11}$$

Equations (8) and (9) can be interpreted as the transition probabilities of two third-order Markov chains $(X_t)_{t\in\mathbb{N}}$ and $(Y_t)_{t\in\mathbb{N}}$, respectively.

$$t^{(1)}_{i,j,k} = P[X_t = i | X_{t-1} = j, Y_t = k] \tag{12}$$

$$t^{(2)}_{i,j,k} = P[Y_t = k | X_t = i, X_{t-1} = j] \tag{13}$$

If the last state was the i-th node, then the next state is the j-th node through the k-th type of interaction with probability $t^{(1)}_{i,j,k}$. Similarly, $t^{(2)}_{i,j,k}$ can be considered as the probability of selecting the k-th type of interaction from the j-th node to the i-th node. For the calculation of the random variables X and Y, the above two equations are deduced according to the total probability formula as follows:

$$P[X_t = i] = \sum_{j=1}^{n}\sum_{k=1}^{m} t^{(1)}_{i,j,k} \times P[X_{t-1} = j, Y_t = k] \tag{14}$$

$$P[Y_t = k] = \sum_{i=1}^{n}\sum_{j=1}^{n} t^{(2)}_{i,j,k} \times P[X_t = i, X_{t-1} = j] \tag{15}$$

$P[X_{t-1} = j, Y_t = k]$ represents the joint probability distribution of X_{t-1} and Y_t, and $P[X_t = i, X_{t-1} = j]$ denotes the joint probability distribution of X_{t-1} and X_t. Considering the steady state of the Markov chain, we can obtain the following formulas:

$$x_i = \lim_{t\to\infty} P[X_t = i], \ (1 \le i \le n) \tag{16}$$

$$y_k = \lim_{t\to\infty} P[Y_t = k], \ (1 \le k \le m) \tag{17}$$

It is very difficult to calculate X and Y due to their coupling to each other and the observation that they contain two joint probability distributions in Equations (14) and (15). In this study, we assumed that the random variables X and Y were completely independent of each other. Thereafter, we could obtain these following formulas:

$$P[X_{t-1} = j, Y_t = k] = P[X_{t-1} = j]P[Y_t = k] \tag{18}$$

$$P[X_t = i, X_{t-1} = j] = P[X_t = i]P[X_{t-1} = j] \tag{19}$$

Based on the above assumption and the fact that t continues to infinity, Equations (16) and (17) could be deduced as:

$$x_i = \sum_{j=1}^{n}\sum_{k=1}^{m} t^{(1)}_{i,j,k} x_j y_k, \ i = 1, 2, \ldots n \tag{20}$$

$$y_k = \sum_{i=1}^{n}\sum_{j=1}^{n} t^{(2)}_{i,j,k} x_i x_j, \ k = 1, 2, \ldots m \tag{21}$$

Based on this, we designed the proper solutions for the functions f and g. Therefore, the random walk with restart algorithm in the

multiplex biological network case could be described as follows:

$$X_t = \alpha \times T^{(1)} \times X_{t-1} Y_{t-1} + (1-\alpha) \times RV \quad (22)$$

$$Y_t = T^{(2)} \times {X_t}^2 \quad (23)$$

The restart vector, RV, represents the initial probability distribution. α is the restart probability. The overall framework of random walk with restart on multiplex biological networks can be illustrated by Algorithm 1.

Algorithm 1 | Random walk with restart in multiplex biological networks

Input: A multiplex biological network G; Restart vector RV; Stopping threshold ∂

Output: A vector representing the score of nodes X

Step 1. Construct two transition probability tensors $T^{(1)}$ and $T^{(2)}$ using Equations (8) and (9)

Step 2. Initialize $X_0 = 1/n$, $Y_0 = 1/m$

Step 3. Let $t = 1$

Step 4. Calculate $X_t = \alpha \times T^{(1)} \times X_{t-1} \times Y_{t-1} + (1\text{-}\alpha) \times RV$

Step 5. Calculate $Y_t = T^{(2)} \times {X_t}^2$

Step 6. If $\|X_t - X_{t-1}\| + \|Y_t - Y_{t-1}\| < \partial$, then let $X = X_t$, $Y = Y_t$ and terminate the algorithm. Otherwise, let $t = t + 1$, and then go to Step 4.

Step 7. Output X

Identification of Essential Proteins

Thus far, the framework for assessing the importance of proteins in multiplex biological networks has been established. Now, we describe the MON approach that was designed for the identification of essential proteins from multiplex biological networks. Algorithm 2 details the MON approach.

Based on a user-specified output number of top-ranking proteins, K, our approach first constructed the multiplex biological network G by integrating PINs, gene expression, and protein domains. Then, considering the conservative and modular features of proteins, a vector $DR = [dr_1, dr_2, \ldots dr_n]^T$ was initialized using the follow equation:

Algorithm 2 | MON

Input: A PIN network, protein domain, gene expression, ortholog data sets, module scores of proteins, and parameter K

Output: Top K proteins sorted by pr in descending order

Step 1. Construct a multiplex biological network G according to Equations (1)–(5)

Step 2. Calculate initial vector DR

Step 3. pr = Algorithm1(G, dr, ϵ)

Step 4. Sort proteins by the value of pr in descending order

Step 5. Output top K of sorted proteins

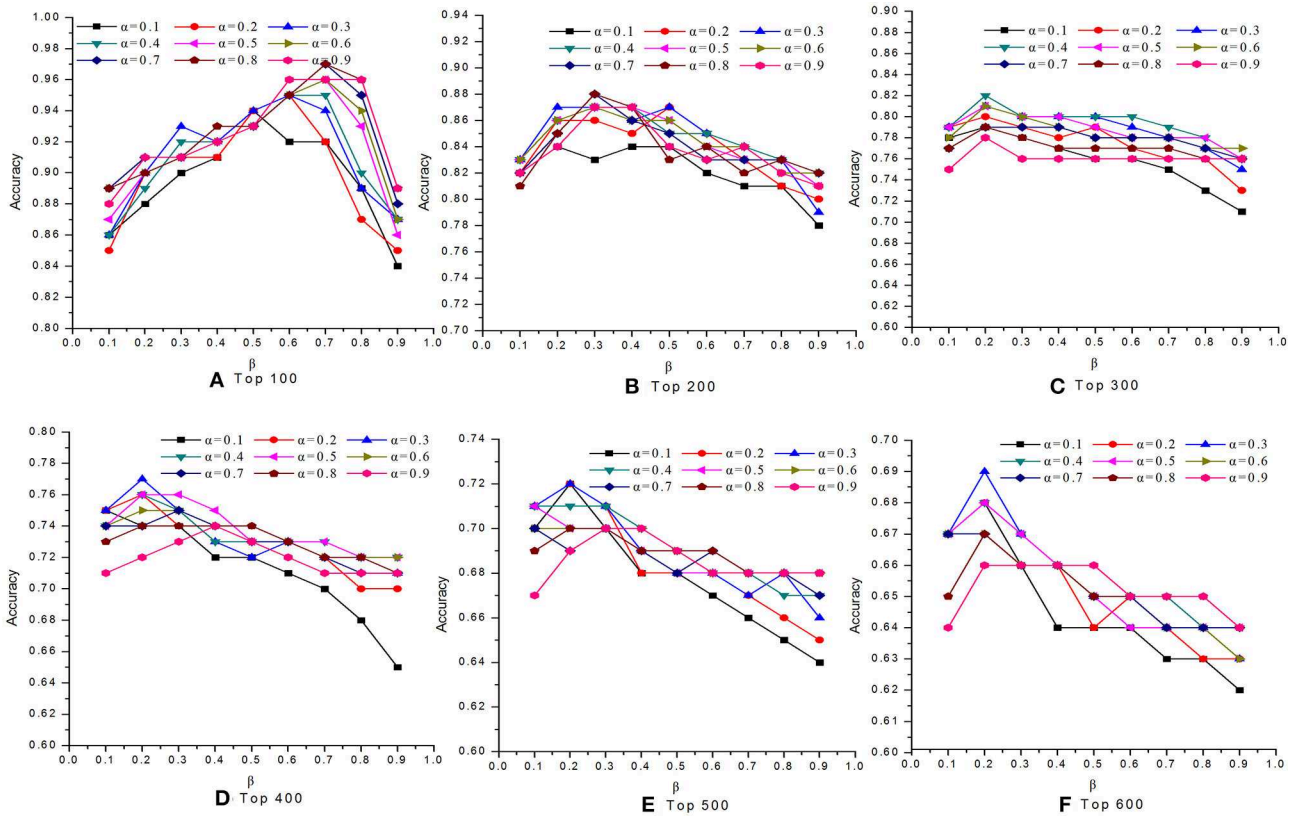


FIGURE 2 | The analysis of parameters α and β on DIP dataset. The figure shows the effect of parameter α and β on the performance of MON on DIP dataset. Six panels represent prediction accuracy of MON in each top percentage of ranked proteins by setting different values of α and β, ranging from 0 to 1. **(A)** Top 100, **(B)** Top 200, **(C)** Top 300, **(D)** Top 400, **(E)** Top 500, **(F)** Top 600.

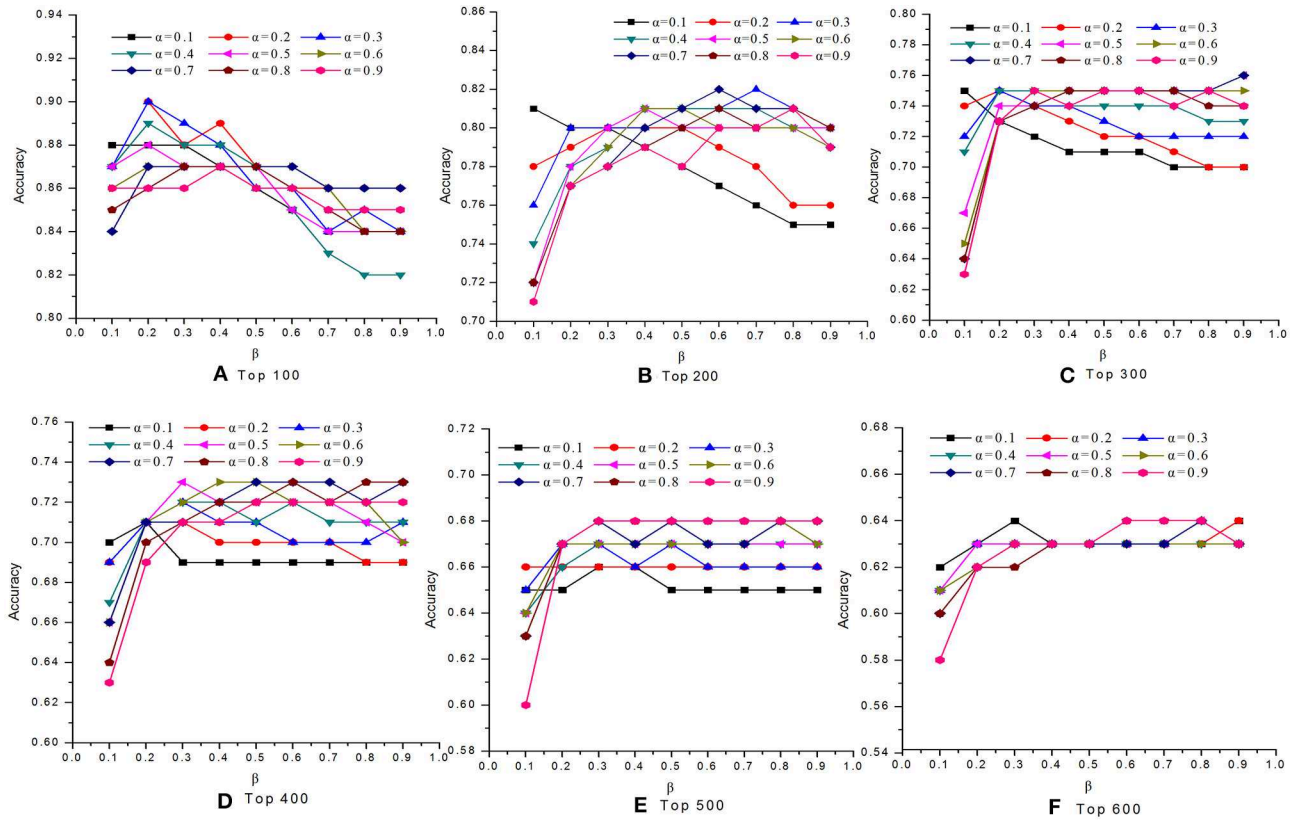


FIGURE 3 | The analysis of parameters α and β on Gavin dataset. The figure shows the effect of parameter α and β on the performance of MON on Gavin dataset. The optimum of α and β for Gavin dataset is 0.3 and 0.2, respectively. **(A)** Top 100, **(B)** Top 200, **(C)** Top 300, **(D)** Top 400, **(E)** Top 500, **(F)** Top 600.

$$dr(p_i) = \beta \times C\_S(p_i) + (1 - \beta) \times M\_S(p_i) \tag{24}$$

In the above equation, $C\_S(p_i)$ and $M\_S(p_i)$ represent conservative score and modular score of the protein p_i, respectively. Conservative score of the protein p_i is derived from information from orthologous proteins and is defined as follows (Zhao et al., 2016b):

$$C\_S(p_i) = \frac{N(p_i)}{\max\limits_{1 \le j \le |V|} (N(p_j))} \tag{25}$$

where $N(p_i)$ denotes the number of homologous proteins that p_i contains in reference organisms. The modular scores of proteins are output scores of the POEM approach with normalization processing (Zhao et al., 2014). Next, we applied the random walk with restart algorithm to the multiplex biological network G and generated a score vector pr. Finally, proteins were sorted in descending order according to pr, with the top K of them being exported.

RESULTS AND DISCUSSION

To evaluate the essential nature of proteins in PINs, they were ranked in descending order based on their ranking scores that were computed by our MON model and by the 11 other competing essential protein prediction approaches, which included DC (Hahn and Kern, 2004), IC (Stephenson and Zelen, 1989), CC (Wuchty and Stadler, 2003), BC (Joy et al., 2005), SC (Estrada and Rodriguez-Velazquez, 2005), NC (Wang et al., 2011), PeC (Li et al., 2012), CoEWC (Zhang et al., 2013), POEM (Zhao et al., 2014), ION (Peng et al., 2012), and FDP (Zhang F. et al., 2019). After this, the top 100, 200, 300, 400, 500, and 600 ranked proteins were selected as candidates for verification as essential proteins. According to the set of known essential proteins, the number of true essential proteins was determined to assess the performance of each approach. Here, we represent the results for the DIP dataset, in detail, and those for the Gavin dataset, in brief.

Effects of Parameters α and β

In this study, we introduced two self-defined parameters as α and β. The parameter α ($0 < \alpha < 1$) was used to control the weight of two scores at step 4 of Algorithm 1. The parameter β ($0 < \beta < 1$) was adopted to adjust the contribution of conservative scores and modular scores of proteins in Equation (24). To study the effects of parameters α and β on the performance of our MON approach, we evaluated the identification accuracy by setting different values for α and β. **Figures 2**, **3** reveal the comparative results in the DIP and Gavin datasets when the parameters α and β possessed different values between 0 and

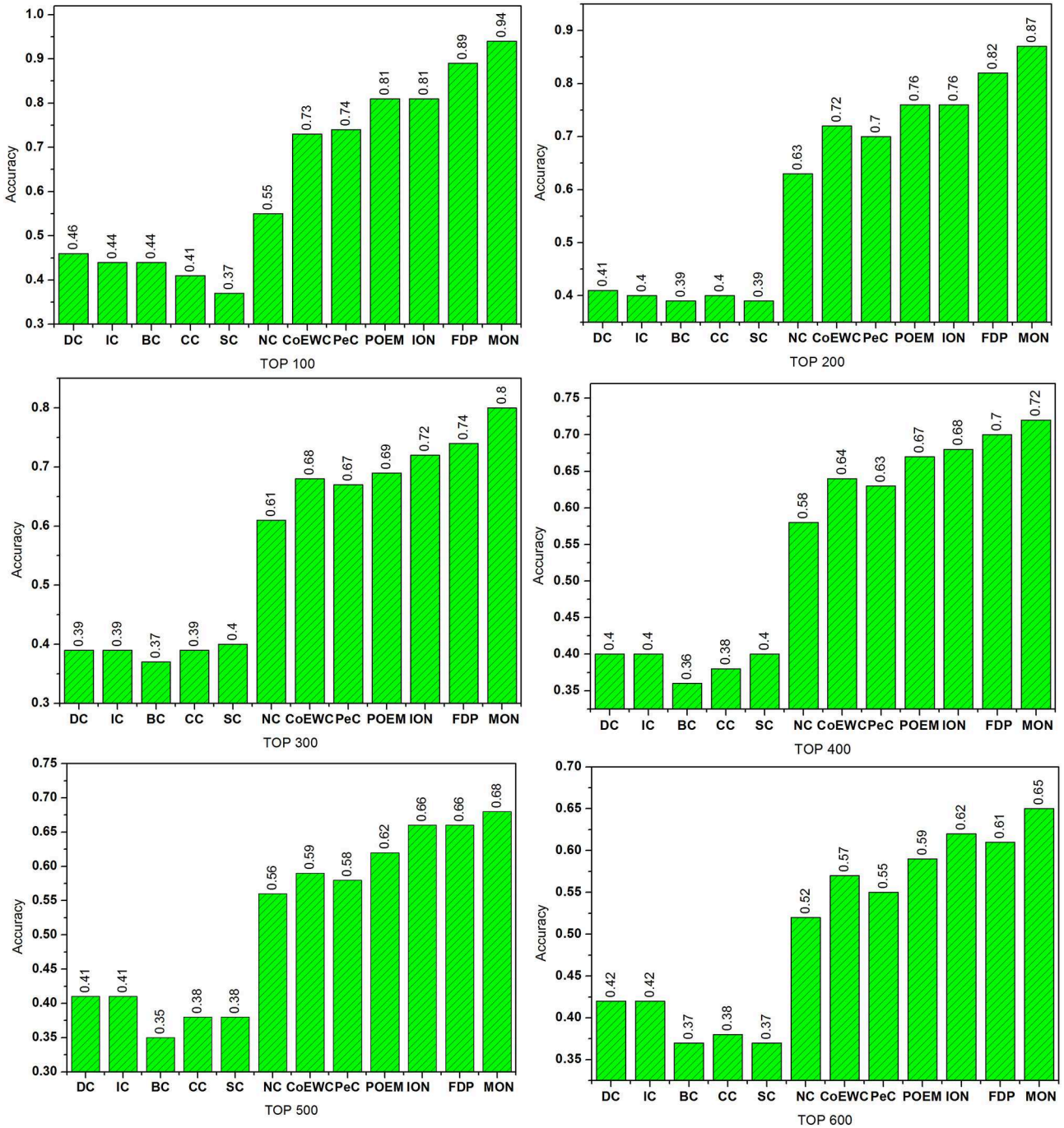


FIGURE 4 | Comparison of the percentage of essential proteins detected by MON and 11 other previously proposed methods. The proteins in protein–protein interaction (PPI) network are ranked in the descending order based on their ranking scores computed by MON, Degree Centrality (DC), Information Centrality (IC), Closeness Centrality (CC), Betweenness Centrality (BC), Subgraph Centrality (SC), Neighbor Centrality (NC), PeC, CoEWC, POEM, ION, and FDP. Then, top 100, 200, 300, 400, 500, and 600 of the ranked proteins are selected as candidates for essential proteins. According to the list of known essential proteins, the percentage of true essential proteins is used to judge the performance of each method. The figure shows the percentage of true essential proteins predicted by each method in each top percentage of ranked proteins. The digits on bars denote the percentage of proteins predicted by each method.

1, respectively. We selected top 100, top 200, top 300, top 400, top 500, and top 600 candidate proteins as detected by MON, respectively. The identification accuracy was evaluated by the percentage of true essential proteins in the top candidates. **Figure 2** indicates that MON achieves the highest prediction accuracy when α is 0.3 and β is 0.5. **Figure 3** shows that the

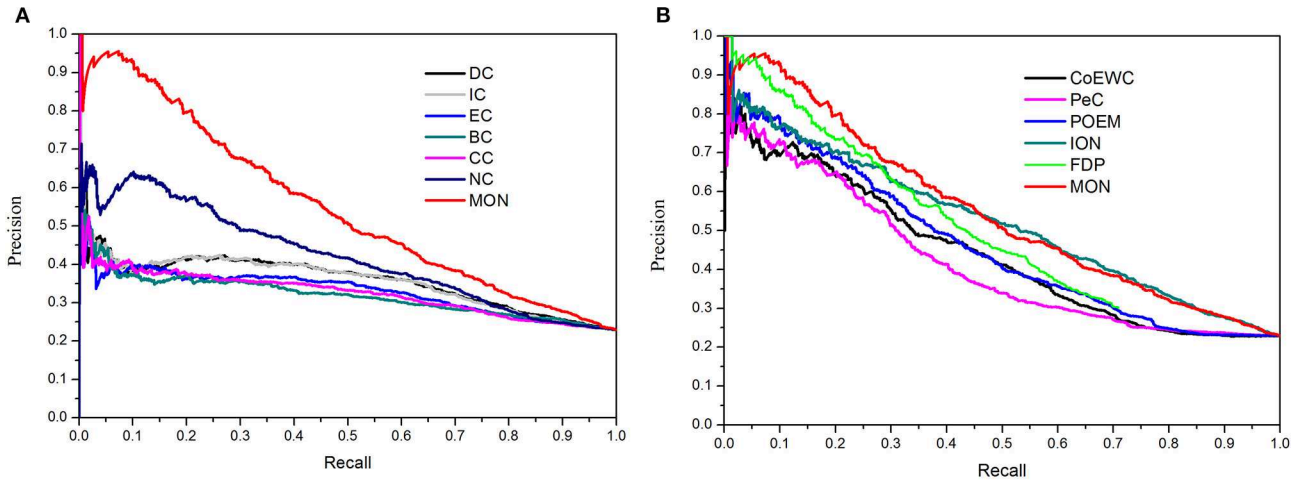


FIGURE 5 | Precision-recall (PR) curves of MON and 11 other existing centrality methods. The proteins ranked in top K (cutoff value) by each method (MON, DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION, and FDP) are selected as candidate essential proteins (positive data set), and the remaining proteins in PPI network are regarded as candidate non-essential proteins (negative data set). With different values of K selected, the values of precision and recall are computed for each method. The values of precision and recall are plotted in PR curves with different cutoff values. **(A)** Shows the PR curves of MON, DC, IC, SC, BC, CC, and NC. **(B)** Shows the PR curves of MON and other five methods: PeC, CoEWC, POEM, ION, and FDP.

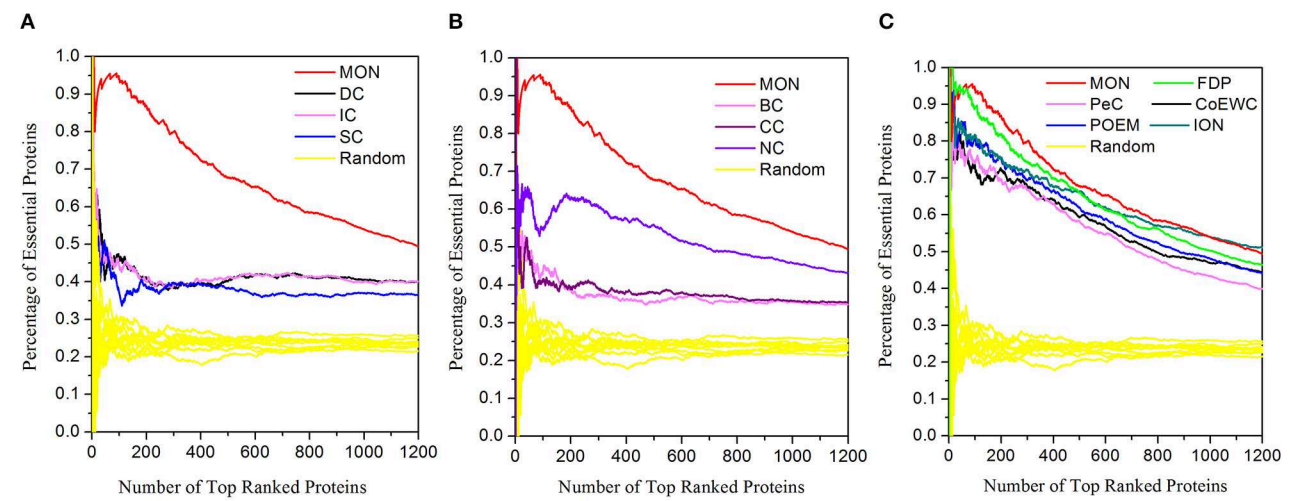


FIGURE 6 | Jackknife curves of the 12 methods. The x-axis represents the proteins in protein–protein interaction (PPI) network ranked by MON and 11 other methods, ranked from left to right as strongest to weakest identification of essentiality. The Y-axis is the percentage of essential proteins encountered moving left to right through the ranked. The areas under the curve for MON and the 11 other methods are used to compare their prediction performance. In addition, the 10 random assortments are also plotted for comparison. **(A)** Shows the comparison results of MON, DC, IC, SC, and DC. **(B)** Represents the comparison results of MON, BC, CC, and NC. **(C)** Illustrates the comparison results of MON and other five methods: PeC, CoEWC, POEM, ION, and FDP.

optimum values for α and β for the Gavin dataset are 0.3 and 0.2, respectively.

Comparison With 11 Other Approaches

To validate the performance of our MON approach, we made comprehensive comparisons of MON to the 11 other competing essential protein identification approaches. Proteins were ranked in descending order according to their scores obtained from each approach. Several of the top predicted proteins were viewed as essential proteins. Then, by comparing to the benchmark set, we determined how many of these candidate proteins were true essential proteins. **Figure 4** reveals the percentage of essential proteins detected by MON and the 11 other prediction approaches within the yeast PIN.

As shown in **Figure 4**, it is clear that MON allows for a higher predictive performance than that of the other competitive centrality methods. For the top 100 candidate proteins and the top 200 candidate proteins, the prediction accuracy of the MON approach was >86%. MON exhibited improvements of 70.91, 38.10, 31.87, 25.65, 21.51, and 26.45% compared to the values

achieved by NC, which possessed the highest prediction accuracy among the six network topology-based centrality methods (DC, IC, BC, CC, SC, and NC) when selecting from the top 100 to top 600 proteins. In particular, when selecting the top 200 proteins, the accuracy of MON in predicting essential proteins was still close to 90%, and this was higher than that of DC, IC, BC, CC, SC, NC, CoEWC, PeC, POEM, and ION for predicting the top 100 proteins. Compared to FDP, which obtained the best prediction accuracy of all 11 competitive approaches, the performance of MON was improved by 5.62, 6.10, 7.62, 3.21, 2.73, and 6.52% from the top 100 to top 600 proteins, respectively.

Table 2 | Common and different proteins predicted by MON and other competing methods ranked in top 100 proteins.

Methods (Mi)	|MON∩Mi|	Non-essential proteins in {Mi – MON}	Percentage of non-essential proteins in {Mi – MON}with low MON (%)
DC	8	54	88.89
IC	8	56	89.28
SC	8	63	92.06
BC	4	56	87.5
CC	7	59	89.83
NC	25	42	92.96
PeC	56	22	81.82
CoEWC	54	24	83.33
POEM	62	14	92.96
ION	54	19	52.63
FDP	48	8	75

The table shows the common and the difference between MON and 11 other competing methods (DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION, and FDP) when predicting top 100 proteins. |MON∩Mi | denotes the number of proteins predicted by both MON and one of the 11 other methods Mi. {Mi – MON} represents the set of proteins detected by Mi while ignored by MON. |Mi – MON| is the number of proteins in set {Mi – MON}. The last column describes the percentages of different non-essential proteins with low MON score (<0.45) in top 100 proteins.

Validated by Precision-Recall Curves

Additionally, the precision-recall (PR) curve was adopted to evaluate the overall performance of MON and the other 11 approaches. First, the proteins in PINs were ranked in a descending order based on the scores obtained from each approach. Next, the top K proteins were selected and placed into the positive set (candidate essential proteins), while the rest of the proteins were stored in the negative set (candidate non-essential proteins). The cutoff parameter of K ranged from 1 to 5,093. Based on different selected values of K, the values of precision and recall were calculated by each approach. Finally, the PR curves were plotted according to values of precision and recall when K changed from 1 to 5,093. **Figure 5A** shows the PR curves of MON and six topology-based centrality methods (DC, IC, BC, CC, SC, and NC). **Figure 5B** illustrates the PR curves for MON and the other five approaches (PeC, CoEWC, POEM, ION, and FDP). **Figure 5** indicates that the PR of MON is clearly higher than that of all competing approaches.

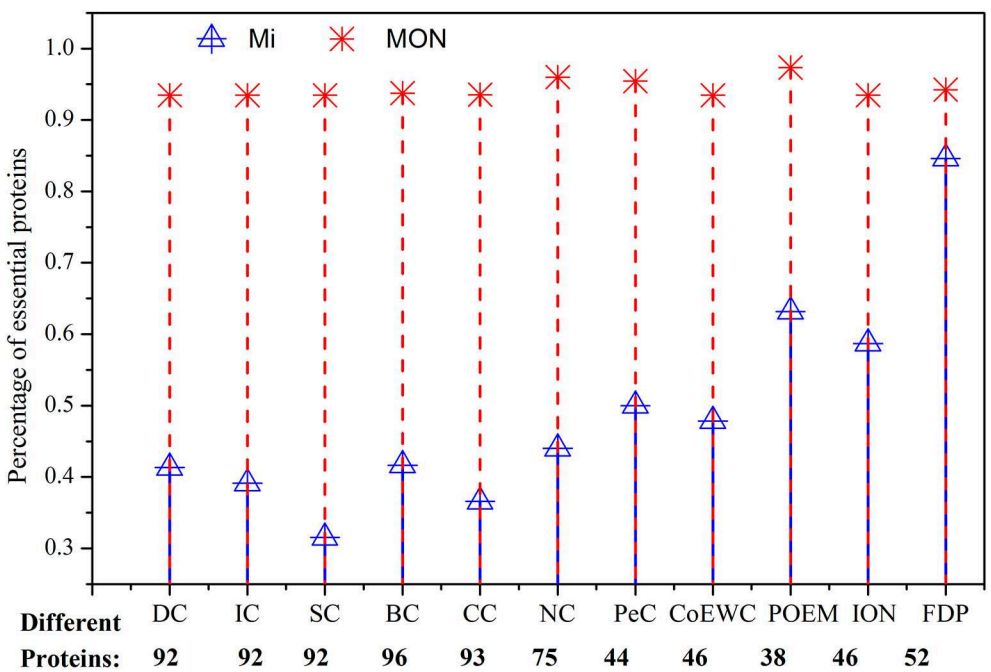


FIGURE 7 | Comparison of the percentage of essential proteins out of all the different proteins between MON and 11 other methods. Different proteins between two prediction methods are the proteins predicted by one method while neglected by the other method. The figure shows how many of the different proteins between MON and 11 other previously proposed methods: DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION, and FDP are essential. The red dash line represents the percentage of essential proteins detected by MON while ignored by Mi, and the blue solid line denotes the percentage of essential proteins predicted by Mi and not by MON.

Validated by Jackknife Methodology

A further comparison between the novel approach MON and the 11 other competing approaches (DC, BC, CC, SC, IC, NC, UDoNC, PeC, CoEWC, POEM, ION, and FDP) was performed by adopting the jackknife methodology (Holman et al., 2009). The areas under the jackknife curve for each approach were used to evaluate their accuracy in identifying essential proteins. Additionally, 10 random assortments were also depicted for this comparison. **Figure 6** illustrates the comparison results where the horizontal axis represents the proteins ranked in descending order according to their scores calculated by each approach and the vertical axis is the percentage of essential proteins related to ranked proteins. **Figure 6A** shows the comparison results between MON and three topology-based centrality methods (DC, IC, and SC). **Figure 6B** represents the comparison results between MON and three centrality methods (BC, CC, and NC). **Figure 6C** indicates the comparison results between MON and the remaining five approaches (PeC, CoEWC, POEM, ION, and FDP). As shown in **Figure 6**, it is clear that the jackknife curve for MON is evidently better than that of the 11 previously proposed approaches. Moreover, MON and the 11 other competing approaches had all achieved improved identification performance compared to that of randomized sorting.

Analysis of the Differences Between MON and Other Approaches

To analyze why and how MON obtains high performance for the identification of essential proteins, we investigated the relationship and differences between MON and the 11 other competitive approaches by detecting a small fraction of proteins. For each approach, the top 100 proteins were selected and compared. The number of top 100 identified proteins ranked by each approach is listed in **Table 2**.

First, we compared MON to DC, BC, CC, SC, IC, NC, PeC, CoEWC, POEM, ION, and FDP by statistically analyzing the number of proteins that were commonly detected by MON and any of the 11 other competitive approaches. The number of common and different proteins between MON and any of the other competing approaches is shown in **Table 2**. In **Table 2**, $|MON \cap Mi|$ represents the number of overlapping proteins identified by MON and by a centrality measure Mi. {Mi – MON} denotes the set of proteins predicted by Mi and not by MON, and |Mi–MON| is the number of proteins predicted by Mi and not by MON.

As illustrated in **Table 2**, among the top 100 proteins, the proportions of overlapping proteins identified by both MON and DC, BC, CC, SC, and IC are all <10%, while the proportions of overlapping proteins detected by both MON and NC and FDP are not more than 50%. The proportion of common proteins predicted by both MON and PeC, CoEWC, POEM, and ION is <65%. Such a small overlap between proteins identified by MON and the 11 other approaches indicates that MON provides a special approach that is different from that of the other

Table 3 | Functional annotations of top 10 predicted essential proteins by MON.

Proteins	Essentiality	Go Term	Categories
YDL147W	True	GO:0006511	BP
		GO:0008541, GO:0034515	CC
YFR004W	True	GO:0016579, GO:0043161	BP
		GO:0004843	MF
		GO:0005829, GO:0008541, GO:0034515	CC
YPR108W	True	GO:0006511	BP
		GO:0005198	MF
		GO:0008541	CC
YDL097C	True	GO:0043248, GO:0006511	BP
		GO:0005198	MF
		GO:0008541, GO:0034515	BP
YER012W	True	GO:0010499, GO:0043161	BP
		GO:0005789, GO:0034515	CC
YKL145W	True	GO:0006511, GO:0045899	BP
		GO:0016887	MF
		GO:0008540	CC
YFR052W	True	GO:0006511	BP
		GO:0008541, GO:0034515	CC
YHR200W	False	GO:0006511	BP
		GO:0005198	MF
		GO:0008540	CC
YOR261C	True	GO:0006511	BP
		GO:0008541, GO:0034515	CC
YGR232W	False	GO:0006508	BP
		GO:0005829	CC

The Table shows results of functional annotation for top 10 proteins predicted by the MON approach. BP, MF, and CC denote biological process, molecular function, and cellular component, respectively.

Table 4 | Percentage of essential proteins identified by MON and 11 other competitive methods based on Gavin dataset.

Methods	Top 100 (%)	Top 200 (%)	Top 300 (%)	Top 400 (%)	Top 500 (%)	Top 600 (%)
DC	46.00	41.00	38.33	39.50	40.20	41.83
IC	44.00	40.00	39.33	40.25	41.40	41.83
SC	37.00	38.50	39.67	39.50	38.40	36.83
BC	44.00	38.50	37.33	36.25	35.40	36.67
CC	41.00	39.50	39.00	38.25	37.80	38.00
NC	55.00	63.00	60.67	57.50	55.80	51.67
PeC	73.00	72.00	67.67	64.00	59.40	56.83
CoEWC	74.00	69.50	66.67	63.00	58.20	54.67
POEM	81.00	75.50	69.33	66.75	62.00	58.83
ION	77.00	77.00	73.67	70.50	65.80	62.83
FDP	89.00	81.50	75.67	70.25	67.00	63.17
MON	90.00	80.00	74.67	71.25	66.80	62.67

This table shows the comparison of the percentage of essential proteins predicted by MON and 11 other competitive methods (DC, IC, SC, BC, CC, NC, PeC, CoEWC, POEM, ION, and FDP) based on protein–protein interaction data from Gavin. Since the total number of ranked proteins in Gavin is 1,855.

approaches. The third column in **Table 2** denotes the number of non-essential proteins among different proteins predicted by Mi but not by MON. We further analyzed these non-essential proteins that were identified by the 11 other approaches, and we found that more than 87% of these non-essential genes that were predicted by six network topology-based centrality measures (DC, IC, BC, CC, SC, and NC) possessed very low MON ranking scores (<0.45). Similarly, more than 50% of the non-essential proteins predicted by PeC, CoEWC, POEM, and ION possessed very low MON ranking scores (<0.45).

Second, we analyzed the essentiality of different proteins detected by MON and by other competing approaches. **Figure 7** shows the percentage of essential proteins in all of the various predicted proteins that were detected by MON and the 11 other competitive approaches. In **Figure 7**, the red dash line represents the percentage of essential proteins detected by MON while ignored by Mi, and blue solid line denotes the percentage of essential proteins predicted by Mi and not by MON. The experimental results shown in **Figure 7** illustrate that among these different proteins, the proportion of essential proteins identified by the MON approach is significantly higher than that predicted by the other approaches. In this study, we chose two representative approaches (BC and POEM) as examples to analyze. The former exhibited the largest number of protein differences compared to our MON approach, and the POEM approach possessed the smallest difference compared to the MON approach. Compared to BC, for all of the top 100 predicted proteins, there were 96 different proteins identified by our MON approach. Among these 96 different proteins identified by MON, 93.75% were essential, while only 41.67% proteins predicted by BC were essential. As another example, there were 22 different proteins detected by either MON or by POEM. Among these different proteins, MON could predict more than 95% of the essential proteins, while POEM only discovered <64% of the essential proteins. The comparable results between MON and the other competitive approaches (DC, CC, SC, IC, NC, PeC, CoEWC, and ION) indicate that the proposed MON approach can identify more essential proteins than the other approaches.

Additionally, we selected top 10 identified candidate proteins by our approach as examples to analyze their functional annotations. To this purpose, GO Term (Ashburner et al., 2000) was adopted to characterize these candidate essential proteins, including molecular function (MF), biological process (BP), and cellular component (CC). **Table 3** shows the results of functional annotation for these 10 proteins. Out of all the 10 candidate proteins, eight proteins were true essential proteins. And all proteins were annotated in terms of BP, MC, and CC.

Prediction Performance of MON Based on the Gavin Dataset

To further test the performance of the proposed approach, we also performed discovery for essential proteins using the Gavin dataset. The ranking scores for proteins were computed using MON ($\alpha = 0.3$, $\beta = 0.2$) and 11 other existing competitive approaches (DC, BC, CC, SC, IC, NC, PeC, CoEWC, POEM, ION, and FDP). The percentage of essential proteins in the top 100, 200, 300, 400, 500, and 600 proteins ranked by these approaches are listed in **Table 4**. The jackknife curves of each approach are illustrated in **Figure 8**. All of these experimental results indicate that MON still outperforms the 11 other competitive approaches, using the Gavin dataset. Specifically, when selecting the top 100 ranked proteins, MON resulted in 95.65, 104.55, 143.24, 104.55, 119.51, 63.64, 23.29, 21.62, 11.11, 16.88, and 1.12% improvements compared to the results obtained from DC, IC, CC, BC, SC, NC, PeC, CoEWC, POEM, ION, and FDP, respectively.

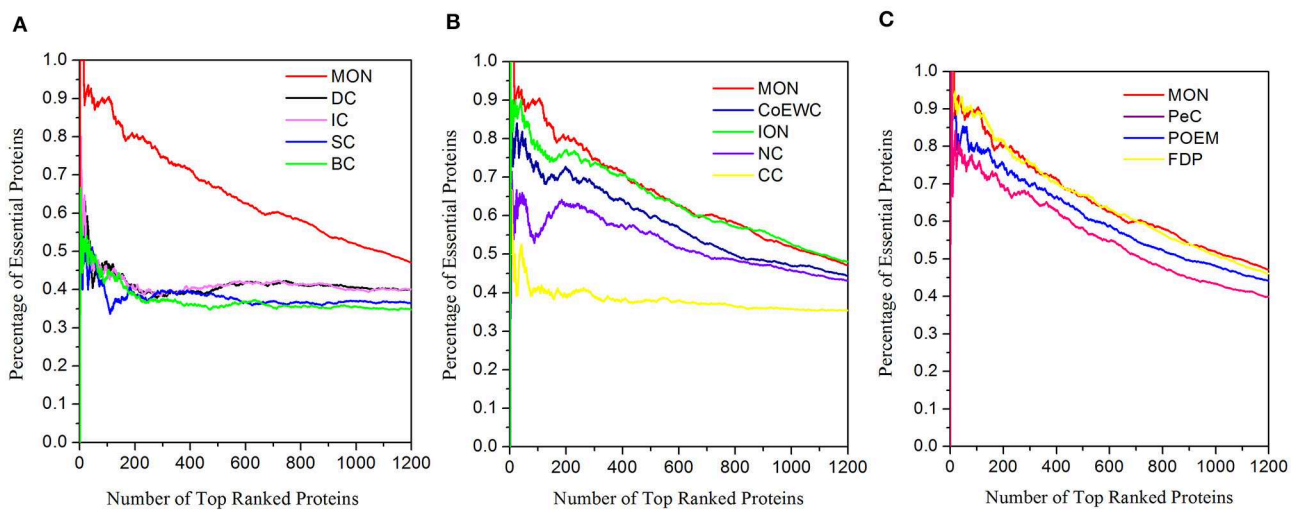


FIGURE 8 | Jackknife curves of MON and 11 other competitive methods based on Gavin dataset. The prediction performance of MON and 11 other existing competitive methods (DC, BC, CC, SC, IC, NC, PeC, CoEWC, POEM, ION, and FDP) based on protein–protein interaction data from Gavin are validated by the jackknife method. **(A)** Shows the Jackknife curves of MON, DC, IC, SC and BC. **(B)** Shows the Jackknife curves of MON, CoEWC, ION, NC and CC. **(C)** Shows the Jackknife curves of MON, PeC, POEM and FDP.

CONCLUSION

The detection of essential proteins is helpful for understanding the minimum requirements for cell survival and development. Many computational approaches have been proposed that integrate PINs and multi-omics data, and this has led to the identification of multiple interactions or links between proteins. Despite the advances in these approaches, designing efficient algorithms to fuse these multisource biological data remains challenging. A simple strategy is to aggregate a collection of heterogeneous data into a single network; however, this strategy can result in substantial information loss. Studies indicate that different types of biological data sources that possess inherent structural characteristics are correlated to each other. Moreover, high-throughput multi-omics biological data exhibit different degrees of quality and can play various roles in the prediction of essential proteins. The multiplex biological network provides an alternative means to address these problems. In this study, we constructed a multiplex biological network by combining PINs with multi-source biological information, and proposed a new essential proteins prediction approach named MON. In MON, we express the multiplex biological network in the tensor model and extend the random walk with restart algorithm by simulating a higher-order Markov chain. Additionally, the conservative and modular features of essential proteins are both taken into account to improve the performance of MON. The experimental results from two yeast PINs demonstrate that MON performs better than 11 other state-of-the-art approaches for predicting essential proteins.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: https://github.com/husaiccsu/MON.

AUTHOR CONTRIBUTIONS

BZ, SH, and LW obtained the protein interaction data, domain data, gene expression profile, and information on orthologous proteins and drafted the manuscript together. BZ and SH designed the new approach, MON, and analyzed the results. XLiu, HX, XH, XLi, and ZZ participated in revising the draft. All authors have read and approved the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (61772089, 61873221, 61672447), Natural Science Foundation of Hunan Province (Nos. 2019JJ40325, 2018JJ3566, 2018JJ3565, 2018JJ4058), National Scientific Research Foundation of Hunan Province (19A048), Major Scientific and Technological Projects for collaborative prevention and control of birth defects in Hunan Province (2019SK1010), Hunan Provincial Key Laboratory of Industrial Internet Technology and Security (2019TP1011), and Hunan Provincial Key Laboratory of Nutrition and Quality Control of Aquatic Animals (2018TP1027).

REFERENCES

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

Baym, M., Shaket, L., Anzai, I. A., Adesina, O., and Barstow, B. (2016). Rapid construction of a whole-genome transposon insertion collection for *Shewanella oneidensis* by Knockout Sudoku. *Nat. Commun.* 7:13270. doi: 10.1038/ncomms13270

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., et al. (2011). Saccharomyces genome database: the genomics resource of budding yeast. *Nucl. Acids Res.* 40, D700–D705. doi: 10.1093/nar/gkr1029

De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. (2015). Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* 5:011027. doi: 10.1103/PhysRevX.5.011027

Didier, G., Brun, C., and Baudot, A. (2015). Identifying communities from multiplex biological networks. *PeerJ* 3:e1525. doi: 10.7717/peerj.1525

Dubach, J. M., Kim, E., Yang, K., Cuccarese, M., Giedt, R. J., Meimetis, L. G., et al. (2017). Quantitating drug-target engagement in single cells *in vitro* and *in vivo*. *Nat. Chem. Biol.* 13:168. doi: 10.1038/nchembio.2248

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E* 71:056103. doi: 10.1103/PhysRevE.71.056103

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440:631. doi: 10.1038/nature04532

Hahn, M. W., and Kern, A. D. (2004). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* 22, 803–806. doi: 10.1093/molbev/msi072

Holman, A. G., Davis, P. J., Foster, J. M., Carlow, C. K., and Kumar, S. (2009). Computational prediction of essential genes in an unculturable endosymbiotic bacterium, Wolbachia of *Brugia malayi*. *BMC Microbiol.* 9:243. doi: 10.1186/1471-2180-9-243

Hsing, M., Byler, K. G., and Cherkasov, A. (2008). The use of gene ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks. *BMC Syst. Biol.* 2:80. doi: 10.1186/1752-0509-2-80

Inouye, M. (2016). The first demonstration of RNA interference to inhibit mRNA function. *Gene* 592, 332–333. doi: 10.1016/j.gene.2016.07.024

Jalili, M., Orouskhani, Y., Asgari, M., Alipourfard, N., and Perc, M. (2017). Link prediction in multiplex online social networks. *R. Soc. Open Sci.* 4:160863. doi: 10.1098/rsos.160863

Joy, M. P., Brock, A., Ingber, D. E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *BioMed Res. Int.* 2005, 96–103. doi: 10.1155/JBB.2005.96

Lei, X., Yang, X., and Wu, F. (2018). "Artificial fish swarm optimization based method to identify essential proteins," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Li, H., Li, T., Quang, D., and Guan, Y. (2018). Network propagation predicts drug synergy in cancers. *Cancer Res.* 78, 5446–5457. doi: 10.1158/0008-5472.CAN-18-0740

Li, M., Lu, Y., Niu, Z., and Wu, F. X. (2015). United complex centrality for identification of essential proteins from PPI networks. *IEEE ACM Trans. Comput. Biol. Bioinformatics* 14, 370–380. doi: 10.1109/TCBB.2015.2394487

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

Li, X., Wang, J., Zhao, B., Wu, F. X., and Pan, Y. (2016). Identification of protein complexes from multi-relationship protein interaction networks. *Hum. Genom.* 10:17. doi: 10.1186/s40246-016-0069-z

Mewes, H. W., Frishman, D., Mayer, K. F., Münsterkötter, M., Noubibou, O., Pagel, P., et al. (2006). MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucl. Acids Res.* 34, D169–D172. doi: 10.1093/nar/gkj148

Narasimhan, V. M., Hunt, K. A., Mason, D., Baker, C. L., Karczewski, K. J., Barnes, M. R., et al. (2016). Health and population effects of rare gene knockouts in adult humans with related parents. *Science* 352, 474–477. doi: 10.1126/science.aac8624

Östlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D. N., Roopra, S., et al. (2009). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucl. Acids Res.* 38, D196–D203. doi: 10.1093/nar/gkp931

Peng, W., Wang, J., Cheng, Y., Lu, Y., Wu, F., and Pan, Y. (2015). UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks. *IEEE ACM Trans. Comput. Biol. Bioinformatics* 12, 276–288. doi: 10.1109/TCBB.2014.2338317

Peng, W., Wang, J., 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

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2011). The Pfam protein families database. *Nucl. Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Samei, Z., and Jalili, M. (2019). Application of hyperbolic geometry in link prediction of multiplex networks. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-49001-7

Stephenson, K., and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Soc. Netw.* 11, 1–37. doi: 10.1016/0378-8733(89)90016-6

Tu, B. P., Kudlicki, A., Rowicka, M., and McKnight, S. L. (2005). Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes. *Science* 310, 1152–1158. doi: 10.1126/science.1120499

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

Wang, J., Li, M., Wang, H., and Pan, Y. (2011). Identification of essential proteins based on edge clustering coefficient. *IEEE ACM Trans. Comput. Biol. Bioinformatics* 9, 1070–1080. doi: 10.1109/TCBB.2011.147

Wang, P., Gao, L., Hu, Y., and Li, F. (2018). Feature related multi-view nonnegative matrix factorization for identifying conserved functional modules in multiple biological networks. *BMC Bioinformatics* 19:394. doi: 10.1186/s12859-018-2434-5

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., Salwinski, 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. *Nucl. Acids Res.* 30, 303–305. doi: 10.1093/nar/30.1.303

Zeng, X., Liao, Y., Liu, Y., and Zou, Q. (2017). Prediction and validation of disease genes using HeteSim Scores. *IEEE ACM Trans. Comput. Biol. Bioinformatics* 14, 687–695. doi: 10.1109/TCBB.2016.2520947

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* 10:31. doi: 10.3390/genes10010031

Zhang, R., and Lin, Y. (2008). DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucl. Acids Res.* 37, D455–D458. doi: 10.1093/nar/gkn858

Zhang, W., Xu, J., and Zou, X. (2019). Predicting essential proteins by integrating network topology, subcellular localization information, gene expression profile and GO annotation data. *IEEE ACM Trans. Comput. Biol. Bioinformatics*. doi: 10.1109/TCBB.2019.2916038

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., Hu, S., Li, X., Zhang, F., Tian, Q., and Ni, W. (2016a). An efficient method for protein function annotation based on multilayer protein networks. *Human Genom.* 10:33. doi: 10.1186/s40246-016-0087-x

Zhao, B., Wang, J., Li, M., Wu, F. X., and Pan, Y. (2014). Prediction of essential proteins based on overlapping essential modules. *IEEE Trans. nanobioscience*, 13, 415–424. doi: 10.1109/TNB.2014.2337912

Zhao, B., Wang, J., Li, X., and Wu, F. X. (2016b). Essential protein discovery based on a combination of modularity and conservatism. *Methods*, 110, 54–63. doi: 10.1016/j.ymeth.2016.07.005

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 April 2020
doi: 10.3389/fgene.2020.00384

BHCMDA: A New Biased Heat Conduction Based Method for Potential MiRNA-Disease Association Prediction

Xianyou Zhu[1†], Xuzai Wang[2†], Haochen Zhao[2], Tingrui Pei[2], Linai Kuang[1,2*] and Lei Wang[2,3*]*

[1] College of Computer Science and Technology, Hengyang Normal University, Hengyang, China, [2] Key Laboratory of Hunan Province for Internet of Things and Information Security, Xiangtan University, Xiangtan, China, [3] College of Computer Engineering & Applied Mathematics, Changsha University, Changsha, China

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Hui Peng,
National University of Singapore, Singapore
Khanh N. Q. Le,
Taipei Medical University, Taiwan

***Correspondence:**
Xianyou Zhu
zxy@hynu.edu.cn
Linai Kuang
kla@xtu.edu.cn
Lei Wang
wanglei@xtu.edu.cn

[†]These authors share first authorship

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

Received: 11 January 2020
Accepted: 27 March 2020
Published: 28 April 2020

Citation:
Zhu X, Wang X, Zhao H, Pei T, Kuang L and Wang L (2020) BHCMDA: A New Biased Heat Conduction Based Method for Potential MiRNA-Disease Association Prediction. Front. Genet. 11:384. doi: 10.3389/fgene.2020.00384

Recent studies have indicated that microRNAs (miRNAs) are closely related to sundry human sophisticated diseases. According to the surmise that functionally similar miRNAs are more likely associated with phenotypically similar diseases, researchers have proposed a variety of valid computational models through integrating known miRNA-disease associations, disease semantic similarity, miRNA functional similarity, and Gaussian interaction profile kernel similarity to discover the potential miRNA-disease relationships in biomedical researches. Taking account of the limitations of previous computational models, a new computational model based on biased heat conduction for MiRNA-Disease Association prediction (BHCMDA) was proposed in this paper, which can achieve the AUC of 0.8890 in LOOCV (Leave-One-Out Cross Validation) and the mean AUC of 0.9060, 0.8931 under the framework of twofold cross validation, fivefold cross validation, respectively. In addition, BHCMDA was further implemented to the case studies of three vital human cancers, and simulation results illustrated that there were 88% (Esophageal Neoplasms), 92% (Colonic Neoplasms) and 92% (Lymphoma) out of top 50 predicted miRNAs having been confirmed by experimental literatures, separately, which demonstrated the good performance of BHCMDA as well. Thence, BHCMDA would be a useful calculative resource for potential miRNA-disease association prediction.

Keywords: miRNA-disease association, bipartite graph network, biased heat conduction, clustering algorithm, integrated similarity

INTRODUCTION

MicroRNAs (miRNAs) are a class of endogenous regulatory non-coding RNAs found in eukaryotes which are about 20 to 25 nucleotides in length. They were normally considered to be negative gene regulators which suppressed the expression of messenger RNAs (mRNAs) and inhibited the protein translation of target genes (Meister and Tuschl, 2004). However, some studies had confirmed that miRNAs could also play a positive regulatory role (Jopling et al., 2005). In recent years, the studies about the miRNA-disease associations have attracted more and more attentions in consideration of

miRNAs having been identified to play a vital role in many important biological processes including cell proliferation, cell development, cell differentiation, cell apoptosis, cell metabolism, cell aging, cell signal transduction, cell viral infection and so on (Xu et al., 2004; Cheng et al., 2005; Miska, 2005; Cui et al., 2006; Bartel, 2009). For example, mir-31 and mir-335 were proved to be effective inhibitors of breast cancer (Tavazoie et al., 2008; Valastyan et al., 2009; Png et al., 2011). miR-122 inhibited cell proliferation and tumorigenesis in certain breast cancer patients by targeting IGF1R (Wang et al., 2012). In addition, researchers discovered that the expression of miR-126 in the blood of patients with Crohn's disease was significantly higher than normal people (Paraskevi et al., 2012). Moreover, the levels of miR-134 and mir-27b were found to be significantly lower in lung tumors than that in normal tissues, which demonstrated that they were associated with lung cancer (Hirota et al., 2012). Therefore, discovery of disease-related miRNAs is significant for the diagnosis, treatment and prevention of complex human diseases.

Up to now, based on the concept that functionally associated miRNAs are more likely related with phenotypically similar disease, a great number of computational models have been proposed to predict potential associations between diseases and miRNAs. For instance, Jiang et al. (2010) raised a hypergeometric distribution-based computational model through adopting miRNA-target interactions. Shi et al. (2013) developed a computational model by concentrating on the functional interlinkage between diseases and miRNAs and implementing random walk on the protein-protein interaction network. Mork et al. (2014) proposed a computational model called miRPD by integrating protein-disease associations and miRNA–protein associations for prediction of miRNA-Protein-Disease associations. Xuan et al. (2013) presented a computational method named HDMP to infer potential disease-related miRNAs based on weighted k most similar neighbors. Chen et al. (2012) developed the global network similarity-based prediction model called RWRMDA by applying random walk to the functional similarity network of miRNA-miRNA to search for potential associations between miRNAs and diseases. However, all these models mentioned above cannot be utilized to predict miRNAs associated new diseases while there are no known miRNA-target associations, since these models rely heavily on known miRNA-target interactions. In recent years, deep learning has been increasingly used to solve many problems, providing an important solution to improve related performance in the field of bioinformatics (Le et al., 2017, 2018). Therefore, in order to solve this problem, Chen and Yan (2014) developed a semi-supervised model called RLSMDA on the basis of regularized least squares, in which negative samples were not required. Zou et al. (2015) introduced two prediction models such as KATZ and CATAPULT to infer potential microRNA-disease associations based on machine learning method. Chen et al. (2016b) put forward a computational model called WBSMDA which was effective for both novel diseases without any known related miRNAs and novel miRNAs without any known associated diseases. Luo et al. (2017) proposed a prediction model named KRLSM to infer potential or missing miRNA-disease associations through integrating miRNA space and disease space into a total miRNA-disease space based on Kronecker product. Chen et al. (2018b) raised a decision tree learning-based model called EGBMMDA, which could serve as a valuable complement to the experimental approach for discovering potential miRNA-disease connections.

Different from above mentioned prediction models, in this paper, a new calculative model called BHCMDA based on Biased heat conduction (BHC) was developed for prediction of potential miRNA-disease association, in which, known miRNA-disease associations, disease semantic similarity, miRNA functional similarity and Gaussian interaction profile kernel similarity were integrated first, and then, the BHC algorithm was adopted to compute both the resources eventually received by miRNAs starting from the miRNA nodes and the resources eventually received by diseases starting from the disease nodes. BHC algorithm is a kind of personalized recommendation algorithm (Liu et al., 2011). Its process is like the transfer of heat in the binary network between the users and the objects. Because the influence of the user's degree and the object's degree are considered into the process of heat transfer, the accuracy of recommending the object that the user is interested in is improved. The transfer process is shown in **Figure 1**. **Figure 1A** shows a binary network of users and objects. **Figure 1B** shows the process of object O_1 and object O_2 receiving resources from users. **Figure 1C** shows the process of user U_1 receiving the resource from the objects. Finally, we averaged these two kinds of resources received by miRNAs and diseases to predict potential miRNA-disease associations. Moreover, in order to evaluate the performance of BHCMDA, twofold cross-validation (twofold CV), fivefold cross-validation (fivefold CV) and leave-one-out cross-validation (LOOCV) were implemented. As a result, BHCMDA could achieve reliable AUCs of 0.8890, 0.9060, and 0.8931 in LOOCV, twofold CV and fivefold CV separately. Furthermore, case studies of esophageal neoplasms, colonic neoplasms and lymphoma were taken to evaluate BHCMDA as well. The simulation results showed that there were 44, 46, and 46 out of top 50 predicted miRNA-disease associations for these three kinds of vital diseases, respectively. Hence, it is obvious that BHCMDA has good performance on prediction of potential miRNA-disease associations.

MATERIALS AND METHODS

MiRNA-Disease Associations

First, we downloaded the known miRNA-disease associations from the HMDD V2.0 database, which consisted of 5430 experimentally verified miRNA-disease associations including 383 diseases and 495 miRNAs (Li et al., 2013). Based on these known miRNAs-disease associations, an adjacency matrix A can be obtained according to the following formula:

$$a_{ij} = \begin{cases} 1: \text{if there is known assocaitionbetween the miRNA } m_i \\ \qquad \text{and the disease } d_j \\ 0: \qquad \text{otherwise} \end{cases} \quad (1)$$

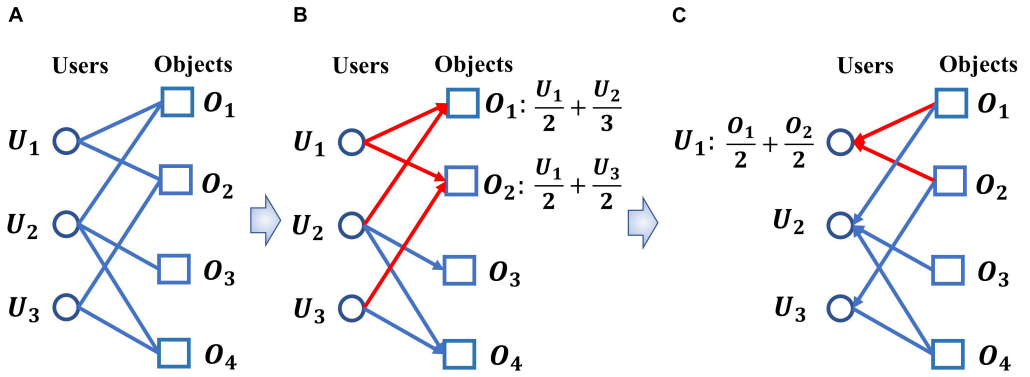


FIGURE 1 | The heat transfer process of biased heat conduction (BHC) algorithm. **(A)** A binary network of users and objects. **(B)** The process of objects receiving resources from users. **(C)** The process of users receiving resources from objects.

MiRNA Functional Similarity

Moreover, based on the assumption that functionally similar miRNAs are more likely associated with phenotypically similar diseases, the miRNA functional similarity scores can be obtained through adopting the modus put forward by Wang et al. (2010). For simplicity, we downloaded the miRNA functional similarity scores from http://www.cuilab.cn/files/images/cuilab/misim.zip directly and utilized these miRNA functional similarity scores to construct a miRNA functional similarity matrix FS, in which, the entity $FS(i, j)$ indicated the functional similarity between the miRNAs m_i and m_j.

Disease Semantic Similarity Model I

Furthermore, for all these 383 diseases obtained previously, we downloaded their *MeSH* descriptors from the *MeSH* database[1], and based on these *MeSH* descriptors, each disease D could be described by a Directed Acyclic Graph (DAG) such as $DAG(D) = (D,\ T(D),\ E(D))$ (Chen, 2015; Chen et al., 2016a; Huang et al., 2016), in which, $T(D)$ indicated the node set containing node D and its ancestor nodes, and $E(D)$ denoted the edge set involving the direct edges which linked the parent nodes to the child nodes. Hence, based on the concept of *DAG*, the semantic value of the disease D could be obtained according to the following formula:

$$DV1\left(D\right) = \sum_{d \in T(D)} D1_D\left(d\right) \tag{2}$$

Here, $D1_D(d)$ represented the contribution of the node d in $T(D)$ to the semantic value of the disease D, which could be obtained according to the following formula:

$$\begin{cases} D1_D\left(d\right) = 1 \quad \text{if} d = D \\ D1_D\left(d\right) = \max\left\{\Delta \times D1_D\left(d'\right) | d' \in \text{chiledren of } d\right\} \\ \qquad\qquad \text{if } d \neq D \end{cases} \tag{3}$$

Here, Δ denoted the semantic contribution factor. From formula (3), it is easy to see that for the disease D, its contribution to the semantic value of itself is equal to 1, while for any other disease d in $T(D)$, as the distance from d to D increases, the contribution of d to D will decrease. Hence, based on the assumption that similar diseases are inclined to share larger parts of their *DAGs*, the semantic similarity between two disease d_i and d_j could be obtained according to the following formula:

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

Disease Semantic Similarity Model II

From above formula (3), it is easy to see that the diseases in the same layer of $DAG(D)$ will make the same contribution to the semantic value of D. Moreover, for diseases in the same layer of $DAG(D)$, it is reasonable to assume that the diseases appeared in less *DAGs* will be more specific than those diseases appeared in more *DAGs* (Chen et al., 2018a). Hence, in order to protrude the contribution of these more specific diseases, the contribution of the node d in $T(D)$ to the semantic value of the disease D could be obtained according to the following formula as well (Chen et al., 2015):

$$D2_D(d) = -\log\left[\frac{\text{the number of } DAGs \text{ containing } d}{\text{the number of diseases}}\right] \tag{5}$$

Based on above formula, the semantic value of the disease D could be obtained according to the following formula as well:

$$DV2(D) = \sum_{d \in T(D)} D2_D\left(d\right) \tag{6}$$

Hence, the semantic similarity between two diseases d_i and d_j could be obtained according to the following formula as well:

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

Gaussian Interaction Profile Kernel Similarity for Diseases

According to the assumption that functionally similar miRNAs tend to be more associated with similar diseases, we can further

[1]http://www.ncbi.nlm.nih.gov/

construct the Gaussian interaction profile kernel similarity for diseases by using known miRNA-disease associations. For convenience, let $IP(d_i)$ denote the ith row of the matrix A, then the Gaussian interaction profile kernel similarity between two diseases d_i and d_j could be obtained according to the following formula:

$$KD\left(i,j\right)=\exp\left(-\gamma_d IP\left(d_i\right)-IP\left(d_j\right)^2\right) \tag{8}$$

Here, the parameter γ_d is utilized to control the kernel bandwidth and can be obtained through the normalization of the original bandwidth $\gamma_d^{'}$ as follows:

$$\gamma_d=\frac{\gamma_d^{'}}{\left(\frac{1}{n}\sum_{i=1}^{n}(IP(d_j)^2\right)} \tag{9}$$

Gaussian Interaction Profile Kernel Similarity for miRNAs

In a way similar to that of the Gaussian interaction profile kernel similarity for diseases, the Gaussian interaction profile kernel similarity between two miRNAs m_i and m_j could be obtained according to the following formula:

$$KM\left(i,j\right)=\exp\left(-\gamma_m IP\left(m_i\right)-IP\left(m_j\right)^2\right) \tag{10}$$

Here, $IP(m_i)$ denotes the ith column of the matrix A, and the parameter γ_m is utilized to control the kernel bandwidth and can be obtained through the normalization of the original bandwidth $\gamma_m^{'}$ as follows:

$$\gamma_m=\frac{\gamma_m^{'}}{\left(\frac{1}{m}\sum_{i=1}^{m}(IP(m_i)^2\right)} \tag{11}$$

Integrated Similarity for miRNAs and Diseases

Based on above formulas, for any two diseases d_i and d_j, we can obtain an integrated similarity between them according to the following formula:

$$SD(i,j)=\begin{cases}\frac{SS1(i,j)+S\,S2(i,j)}{2}\ d_i \text{ and } d_j \text{has semantic similarity}\\ KD\left(i,j\right) \qquad\qquad \text{otherwise}\end{cases} \tag{12}$$

Moreover, in a similar way, for any two miRNAs m_i and m_j, we can obtain an integrated similarity between them according to the following formula:

$$SM(i,j)=\begin{cases}FS\left(i,j\right)\ m_i \text{and} m_j \text{has functional similarity}\\ KM\left(i,j\right) \qquad\qquad \text{otherwise}\end{cases} \tag{13}$$

BHCMDA

According to the assumption that functionally similar miRNAs are more likely associated with phenotypically similar diseases (Liu et al., 2011), as illustrated in the following **Figure 2**, we developed a novel computational model called BHCMDA based on the BHC algorithm to predict potential miRNA-disease associations through combining the previously constructed adjacency matrix A, the integrated miRNA similarity matrix SM and the integrated disease similarity matrix SD according to the following steps:

Step 1: For convenience, let the $M=\{m_1, m_2, \ldots\ldots m_n\}$ and $D=\{d_1, d_2, \ldots\ldots d_q\}$ represent all the miRNAs and diseases collected previously, then we can obtain an $n\times q$ dimensional adjacency matrix A, an $q\times q$ dimensional integrated diseases similarity matrix SD, and an $n\times n$ dimensional integrated miRNAs similarity matrix SM according to the above formulas, respectively. Moreover, based on these newly obtained two kinds of matrices such as A and SM, we can further construct a new $n\times q$ dimensional miRNA-disease association adjacency matrix A' as follows:

$$a_{ij}^{'}=\begin{cases}1: & \text{If } a_{ij}=1\\ \max\limits_{m_t\in M_{ij}} SM(i,t): & \text{If } \max\limits_{m_t\in M_{ij}} SM(i,t)>0\\ 0: & \text{otherwise}\end{cases} \tag{14}$$

Here, M_{ij} is the set of miRNA nodes that satisfy: " $m_t\in M_{ij}$, there are $a_{tj}=1$ and $SM\left(i,t\right)>\delta$, where δ is a threshold parameter with value between 0 and 1. In this paper, we will set $\delta=0.29$ according to our simulation results. Thereafter, as illustrated in the following **Figure 3A**, based on the new adjacency matrix A', we can construct a bipartite miRNAs-diseases network.

Step 2: As illustrated in **Figure 3B**, let miRNAs and diseases represent the Object nodes and the User nodes respectively, then after implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network, for any given disease d_j in D, the final resources $f(d_j)$ received by d_j can be obtained according to the following formula while we started from the miRNA nodes:

$$f(d_j)=\sum_{i=1}^{n}\frac{a_{ij}^{'}\times f(m_i)}{d(m_i)} \tag{15}$$

Here, $f(m_i)$ is the initial resource of the miRNA m_i in M, which is set to 1, and $d(m_i)$ represents the degree of the miRNA node m_i in the newly constructed bipartite miRNAs-diseases network.

Step 3: As illustrated in **Figure 3C**, let diseases and miRNAs represent the Object nodes and the User nodes, respectively, then after implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network, for any given miRNA m_i in M the final resources $f(m_i)'$ received by m_i can be obtained according to the following formula while we started from the disease nodes:

$$f(m_i)'=\frac{1}{d(m_i)^{\gamma}}\times\sum_{j=1}^{q}\frac{a_{ij}^{'}\times f(d_j)}{d(d_j)} \tag{16}$$

Here, $d(d_j)$ represents the degree of the disease node d_j in the newly constructed bipartite miRNAs-diseases network, and γ is a parameter to adjust the impact of $d(d_j)$. In this paper, we set $\gamma=0.001$ according to our simulation results.

Step 4: Similar to above step 1, based on these newly constructed two kinds of matrices such as A and SD, we can

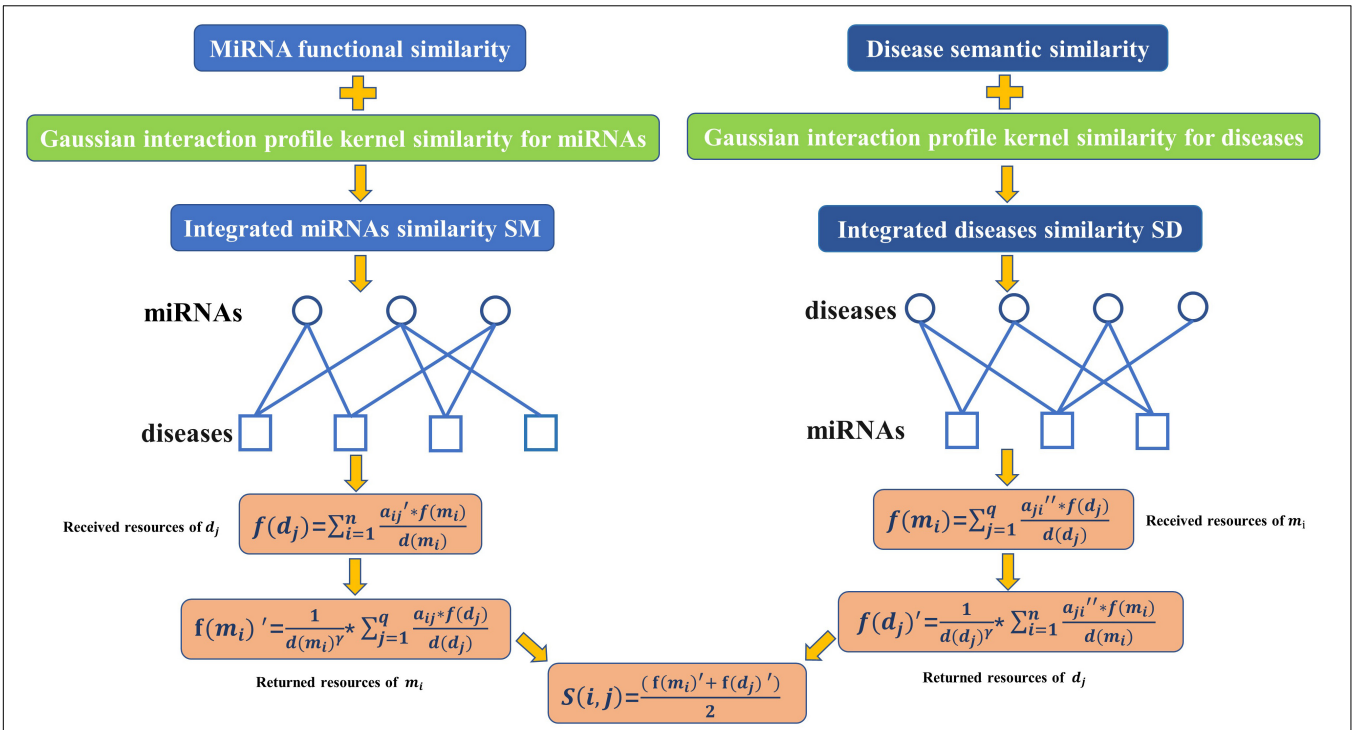


FIGURE 2 | Flow chart of BHCMDA model to predict the potential miRNA-disease associations.

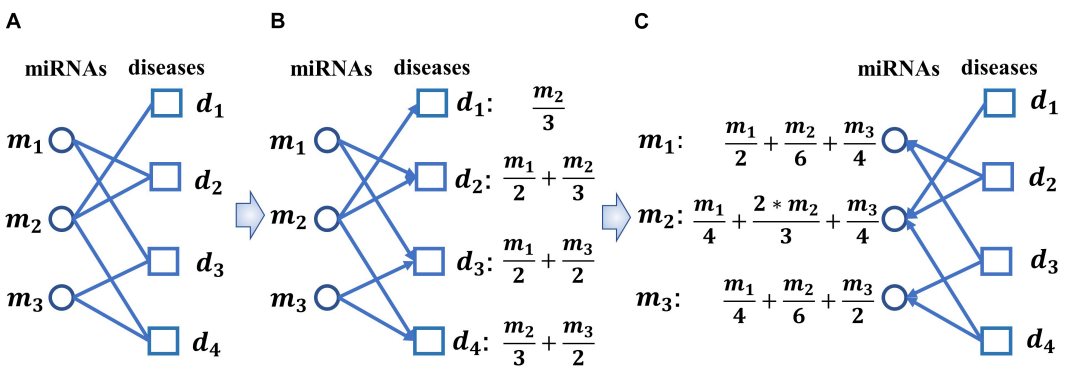


FIGURE 3 | Diagram of implementing the biased heat conduction (BHC) algorithm on the newly constructed bipartite miRNAs-diseases network. **(A)** The newly constructed bipartite miRNAs-diseases network, **(B)** let miRNAs and diseases represent the Object nodes and the User nodes respectively while implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network, **(C)** let diseases and miRNAs represent the Object nodes and the User nodes respectively while implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network.

also construct another new $n \times q$ dimensional miRNA-disease association adjacency matrix A'' as follows:

$$a_{ij}'' = \begin{cases} 1: & \text{If } a_{ij} = 1 \\ \max\limits_{d_t \in D_{ij}} SD(i,t): & \text{If } \max\limits_{d_t \in D_{ij}} SD(i,t) > 0 \\ 0: & \text{otherwise} \end{cases} \quad (17)$$

Here, D_{ij} is the set of disease nodes that satisfy: " $d_t \in D_{ij}$,there are $a_{jt} = 1$ and $SD(i,t) > \eta$, where η is a threshold parameter with value between 0 and 1. In this paper, we set $\eta = 0.13$ according to our simulation results. Thereafter, as illustrated in the following **Figure 3A**, based on the new adjacency matrix A'', we can construct another new bipartite miRNAs-diseases network.

Step 5: Similar to above step 2, after implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network, for any given miRNA m_i in M, the final resources $f(m_i)''$ received by m_i can be obtained according to the following formula while we started from the disease nodes:

$$f(m_i)'' = \sum_{j=1}^{q} \frac{a_{ji}'' \times f(d_j)'}{d(d_j)} \quad (18)$$

Here, $f(d_j)'$ is the initial resource of the disease d_j in D, which is set to 1, and $d(d_j)$ represents the degree of the disease node d_j in the newly constructed bipartite miRNAs-diseases network.

Step 6: Similar to above step 3, after implementing the BHC algorithm on the newly constructed bipartite miRNAs-diseases network, for any given

disease d_j in D, the final resources $f(d_j)''$ received by d_j can be obtained according to the following formula while we started from the miRNA nodes:

$$f(d_j)'' = \frac{1}{d(d_j)^{\gamma}} \times \sum_{i=1}^{n} \frac{a_{ji}'' \times f(m_i)''}{d(m_i)} \tag{19}$$

Here, $d(m_i)$ represents the degree of the miRNA node m_i in the newly constructed bipartite miRNAs-diseases network, and γ is a parameter to adjust the impact of $d(d_j)$. In this paper, we set $\gamma = 0.001$ according to our simulation results.

Step 7: Finally, based on above formulas, the association score between miRNA m_i and disease d_j can be calculated as follows:

$$S(i,j) = \frac{(f(m_i)' + f(d_j)'')}{2} \tag{20}$$

RESULTS

Performance Evaluation

In order to evaluate the predictive performance of BHCMDA, twofold cross-validation, fivefold cross-validation and LOOCV were implemented separately based on the known miRNA-disease associations downloaded from the HMDD V2.0 database. In LOOCV, every known miRNA-disease association takes turns to act as the test sample and the rest of known miRNA-disease associations serve as training samples. Moreover, all these miRNA-disease pairs having no known associations play the role of candidate samples, then we can obtain the ranking of each test sample with all candidate samples according to their predicted scores after implementing BHCMDA. If the rank of the test sample is higher than the given threshold, it will be considered as a correct prediction. In the framework of fivefold cross-validation, all known miRNA-disease associations are randomly divided into five equal groups without overlap first, then each group acts as test samples in turn and the other four groups serve as training samples. Besides, all these miRNA-disease pairs having no known associations play the role of candidate samples. After the scores of candidate samples and the test samples have been calculated, we take turns to compare the score of each test sample with the scores of candidate samples. If the rank of the test sample exceeds the given threshold, it will be thought as a successful prediction. Furthermore, the receiver-operating characteristics (ROC) curve can be painted to assess the performance of BHCMDA by computing false positive rate (FPR, 1-specificity) and true positive rate (TPR, sensitivity) on the basis of varying thresholds (Le et al., 2019). Here, sensitivity means the percentage of positive test samples whose rankings exceed the given threshold, while 1-specificity denotes the percentage of candidate samples with rankings under the given threshold. Then, area under the ROC curves (AUCs) can be calculated to evaluate the predictive performance of BHCMDA, the larger the value, the better the prediction performance of BHCMDA.

As a result, BHCMDA can achieve reliable AUCs of 0.8890, 0.9060, and 0.8931 under the frameworks of global LOOCV, twofold cross-validation and fivefold cross-validation respectively. Moreover, we compared BHCMDA with two kinds of state-of-the-art models such as RLSMDA (Chen and Yan, 2014) and WBSMDA (Chen et al., 2016b). As illustrated in the **Figure 4**, RLSMDA and WBSMDA can achieve AUCs of 0.8507 and 0.7802 under the frameworks of global LOOCV respectively, which are inferior to the BHCMDA's AUCs. Besides, as shown in the **Figure 5**, under the twofold cross-validation framework, the AUCs of RLSMDA and WBSMDA are 0.8470 and 0.6658 respectively, indicating that the AUCs of BHCMDA is higher than RLSMDA and WBSMDA. What's more, as illustrated in the **Figure 6**, RLSMDA and WBSMDA can achieve AUCs of 0.8498 and 0.7337 under the frameworks of fivefold cross-validation respectively, which are also lower than the BHCMDAs' AUCs. In conclusion, it is obvious that BHCMDA has better performance than RLSMDA and WBSMDA in miRNA-disease association prediction.

Case Studies

In order to further assess the predictive performance of BHCMDA, we conducted case studies of three kinds of human diseases such as esophageal neoplasms, colonic neoplasms and lymphoma, and the predicted results were verified by evidences illustrated in HMDD v3.0[2], dbDEMC 2.0[3], dbDEMC (Yang et al., 2010) and miR2Disease (Jiang et al., 2008), respectively.

Esophageal neoplasms is the eighth common cancer in the world according to the pathological characteristics (He et al., 2012). As the tumor grows, the patient may suffer from difficult or painful swallowing, coughing up blood and weight loss. The number of men having esophageal cancer are three to four times than that of women, and the survival rates are low (Enzinger and Mayer, 2003). The main treatment for esophageal neoplasms is cisplatin-based chemotherapy, but the chemotherapy reaction is difficult to detect. Therefore, the earlier the esophageal tumor is found, the more helpful it will be in the cancer treatment (Xie et al., 2013; Wan et al., 2016). A large number of miRNAs have been confirmed to be associated with esophageal neoplasms. For instance, the overexpression of hsa-miR-17 cluster can promote the growth of esophageal tumor cell. In addition, hsa-let-7 can server as the prognostic biomarker for weighing the response to chemotherapy (Liao et al., 2014; Xu et al., 2014). While implementing BHCMDA to predict associated miRNAs of esophageal neoplasms, there are 9 out of the top-10 and 44 out of the top-50 predicted miRNAs having been verified to be related with esophageal neoplasms according to confirmations provided by dbDEMC and dbDEMC 2.0, respectively (see **Table 1**).

Colonic neoplasms is a common malignant tumor which poses a huge threat to human lives in the world (Jemal et al., 2011; Ogata-Kawata et al., 2014). It is reported that

[2]http://www.cuilab.cn/hmdd

[3]http://www.picb.ac.cn/dbDEMC

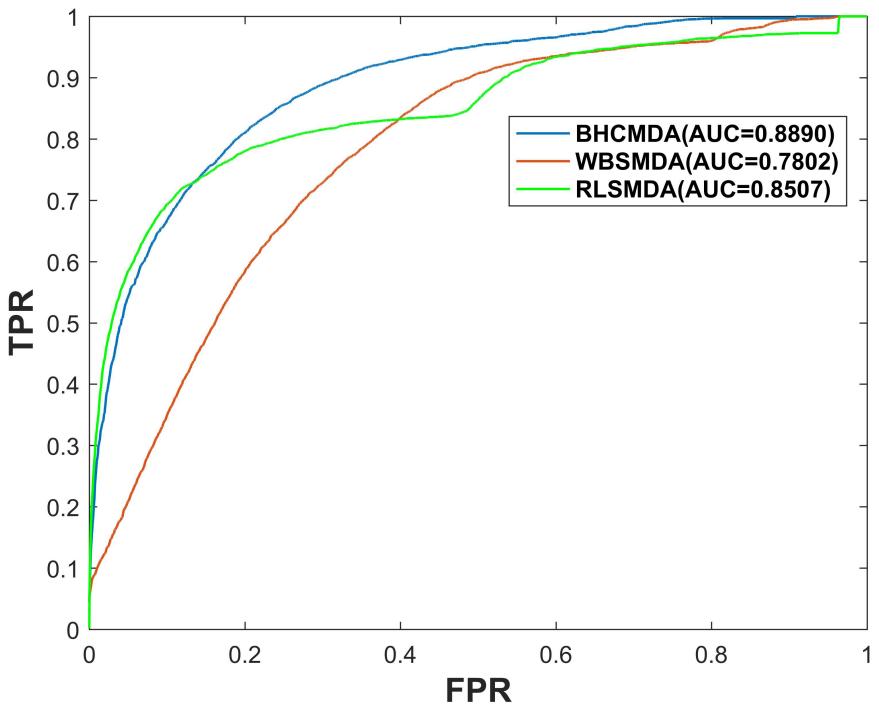


FIGURE 4 | Performance comparisons between BHCMDA, LRLSLDA, and WBSMDA in LOOCV.

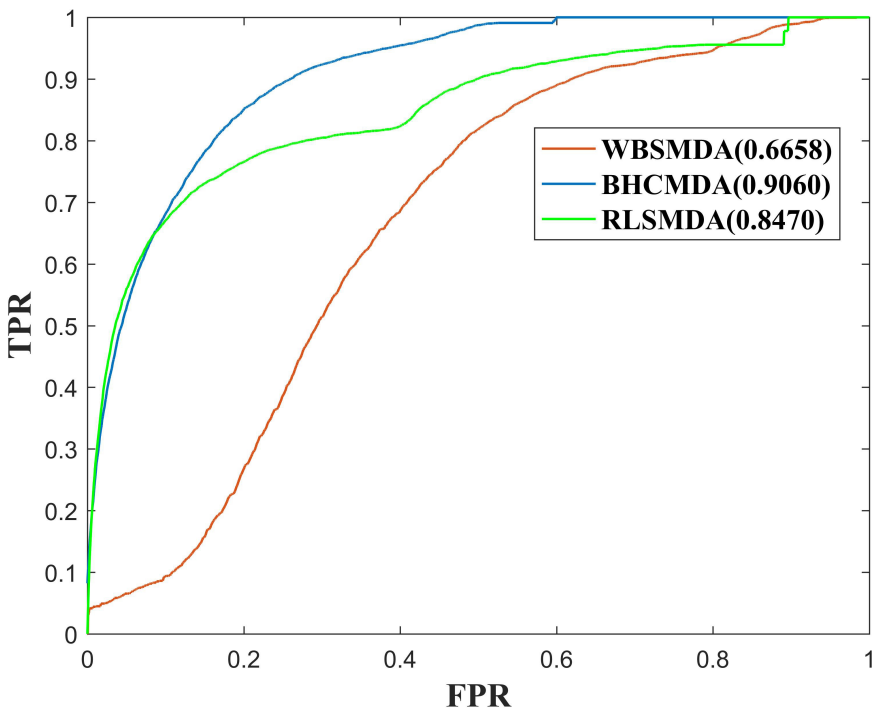


FIGURE 5 | Performance comparisons between BHCMDA, LRLSLDA, and WBSMDA in twofold cross-validation.

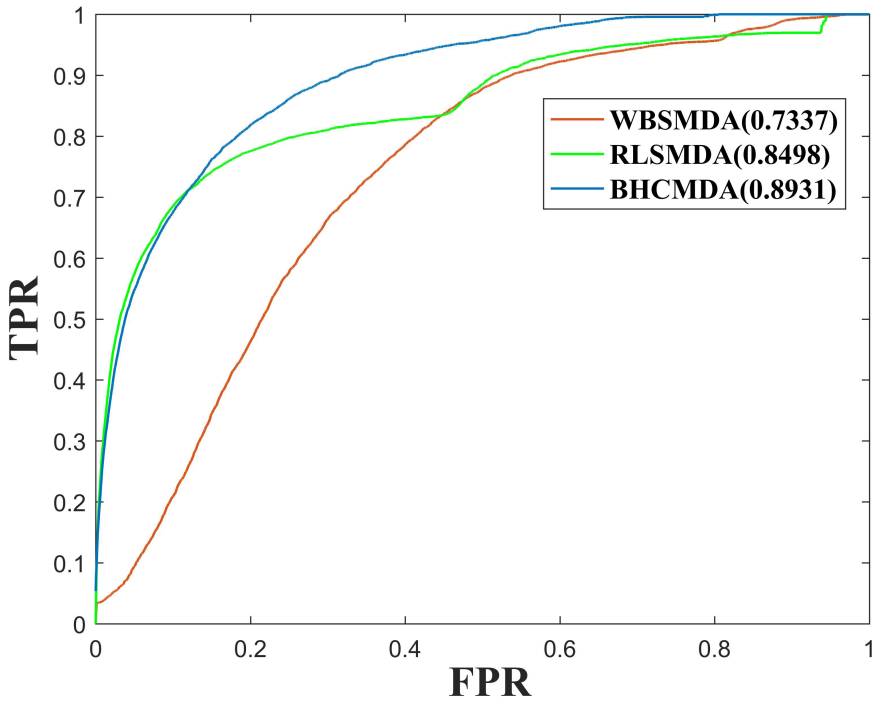


FIGURE 6 | Performance comparisons between BHCMDA, LRLSLDA, and WBSMDA in fivefold cross-validation.

TABLE 1 | Top 50 potential Esophageal Neoplasms-related miRNAs predicted by BHCMDA and confirmations for these predicted associations provided by the dbDEMC and dbDEMC 2.0.

miRNA	Evidence	miRNA	Evidence
hsa-mir-17	dbDEMC	hsa-mir-302c	dbDEMC
hsa-mir-18a	dbDEMC 2.0	hsa-mir-602	dbDEMC
hsa-mir-200b	dbDEMC	hsa-mir-612	dbDEMC 2.0
hsa-mir-629	dbDEMC 2.0	hsa-mir-657	unconfirmed
hsa-mir-93	dbDEMC 2.0	hsa-mir-376c	dbDEMC 2.0
hsa-mir-324	dbDEMC	hsa-mir-367	dbDEMC 2.0
hsa-mir-19b	dbDEMC 2.0	hsa-mir-153	dbDEMC
hsa-let-7d	dbDEMC	hsa-mir-302e	dbDEMC
hsa-mir-185	dbDEMC 2.0	hsa-mir-30c	dbDEMC 2.0
hsa-mir-638	unconfirmed	hsa-mir-302d	dbDEMC 2.0
hsa-let-7f	dbDEMC 2.0	hsa-mir-16	dbdemc 2.0
hsa-mir-601	unconfirmed	hsa-mir-429	dbDEMC 2.0
hsa-mir-1	dbDEMC 2.0	hsa-mir-106b	dbDEMC 2.0
hsa-let-7i	dbDEMC 2.0	hsa-mir-583	dbDEMC
hsa-let-7e	dbDEMC	hsa-mir-125b	dbDEMC 2.0
hsa-let-7g	dbDEMC	hsa-mir-660	dbDEMC
hsa-mir-637	dbDEMC 2.0	hsa-mir-557	dbDEMC 2.0
hsa-mir-218	dbDEMC 2.0	hsa-mir-600	unconfirmed
hsa-mir-608	unconfirmed	hsa-mir-611	unconfirmed
hsa-mir-596	dbDEMC 2.0	hsa-mir-654	dbDEMC 2.0
hsa-mir-615	dbDEMC	hsa-mir-662	dbDEMC 2.0
hsa-mir-622	dbDEMC	hsa-mir-769	dbDEMC
hsa-mir-518c	dbDEMC 2.0	hsa-mir-215	dbDEMC 2.0
hsa-mir-301a	HMDD3.0	hsa-mir-335	dbDEMC 2.0
hsa-mir-302b	dbDEMC	hsa-mir-221	dbDEMC 2.0

TABLE 2 | Top 50 potential Colonic Neoplasms-related miRNAs predicted by BHCMDA and confirmations for these predicted associations provided by the dbDEMC, dbDEMC 2.0, HMDD3.0 and miR2Disease.

miRNA	Evidence	miRNA	Evidence
hsa-mir-324	unconfirmed	hsa-mir-146b	dbDEMC 2.0
hsa-mir-222	dbDEMC 2.0	hsa-mir-601	dbDEMC 2.0
hsa-mir-301a	dbDEMC 2.0	hsa-mir-7	dbDEMC 2.0
hsa-mir-638	dbDEMC 2.0	hsa-mir-637	dbDEMC 2.0
hsa-mir-200a	unconfirmed	hsa-mir-526a	dbDEMC 2.0
hsa-mir-210	dbDEMC 2.0	hsa-mir-515	unconfirmed
hsa-mir-133a	dbDEMC 2.0	hsa-mir-27a	dbDEMC 2.0
hsa-mir-93	dbDEMC 2.0	hsa-mir-331	HMDD3.0
hsa-mir-185	dbDEMC 2.0	hsa-mir-148a	dbDEMC 2.0
hsa-mir-367	dbDEMC 2.0	hsa-mir-195	dbDEMC 2.0
hsa-mir-219	unconfirmed	hsa-mir-520h	dbDEMC 2.0
hsa-mir-520a	HMDD3.0	hsa-mir-153	dbDEMC 2.0
hsa-mir-196a	dbDEMC 2.0	hsa-mir-199b	dbDEMC 2.0
hsa-mir-199a	23292866	hsa-mir-30b	dbDEMC 2.0
hsa-mir-297	dbDEMC 2.0	hsa-mir-26a	dbDEMC
hsa-mir-608	dbDEMC 2.0	hsa-mir-181b	dbDEMC 2.0
hsa-mir-449b	dbDEMC 2.0	hsa-mir-520e	dbDEMC 2.0
hsa-mir-34c	miR2Disease	hsa-mir-602	dbDEMC 2.0
hsa-mir-215	dbDEMC 2.0	hsa-mir-512	HMDD3.0
hsa-mir-375	dbDEMC 2.0	hsa-mir-194	dbDEMC 2.0
hsa-mir-25	dbDEMC 2.0	hsa-mir-95	dbDEMC 2.0
hsa-mir-34b	dbDEMC	hsa-mir-612	dbDEMC 2.0
hsa-mir-429	dbDEMC 2.0	hsa-mir-526b	dbDEMC 2.0
hsa-mir-203	dbDEMC 2.0	hsa-mir-657	dbDEMC 2.0
hsa-mir-518b	dbDEMC 2.0	hsa-mir-135a	dbDEMC 2.0

about half of colonic neoplasms patients may die of metastatic disease in five years from diagnosis (Parkin et al., 2005; Drusco et al., 2014). Therefore, early diagnosis of colon cancer is of great significance in improving the patients' survival rate. In the recent years, investigators have verified a few miRNAs related with colonic neoplasms. Take Mir-199a-3p (the 3p arm of the pre-miRNA for miR-199a) as an example, it is highly expressed in colonic neoplasms tissues, resulting in significantly reduced survival rate of patients (Wan et al., 2013). In addition, tumor specimens illustrated highly significant and large multiple differential expressions of levels of some miRNAs, including mir-1, mir-31, mir-133a, mir-135b and others (Sarver et al., 2009). While implementing BHCMDA to discern the potentially relevant miRNAs of colonic neoplasms, there are 8 out of the top-10 and 46 out of the top-50 predicted miRNAs having been validated to be related with colonic neoplasms by confirmations provided by dbDEMC, dbDEMC 2.0, HMDD3.0 and miR2Disease, respectively (see **Table 2**).

There are two types of lymphoma, one is Hodgkin Lymphomas (HL) and the other is non-Hodgkin Lymphomas (NHL). HL is a more common form of lymphoma and it is difficult to be diagnosed at an early stage (Coiffier, 2006; Xie et al., 2012). NHL is a heterogeneous malignant tumor originating from lymphoid hematopoietic tissue and it is mainly treated by local radiotherapy and chemotherapy (Coiffier, 2006). An example of miRNAs related with lymphoma is miR-125b. By inhibiting miR-125b-5p (The 5p arm of the pre-miRNA for mir-125b), lymphoma cells will be sensitive to anticancer drugs such as bortezomib (Manfè et al., 2013). Besides, the overexpressed miR-142-5p (the 5p arm of the pre-miRNA for miR-142) which was found in gastric MALT lymphoma played a vital role in the pathogenesis of this cancer (Saito et al., 2012). Furthermore, the upregulation of miRNA hsa-mir-9, hsa-mir-34a, hsa-mir-183, hsa-mir-215 and down-regulation of hsa-mir-30b were all relevant to lymphoma's development based on experimental literatures. While implementing BHCMDA to infer the potentially relevant miRNAs of Lymphoma, there are 10 out of the top-10 and 46 out of the top-50 predicted miRNAs having been confirmed to be associated with Lymphomas by confirmations provided by dbDEMC 2.0 and the recent experimental literatures with relevant PMIDs, respectively (see **Table 3**).

DISCUSSION

In recent years, a growing number of computational models have been proposed to find underlying miRNA-disease associations. In this article, we put forward a prediction model called BHCMDA based on the BHC algorithm to discover potential associated miRNAs of the diseases by integrating known miRNA-disease associations, the disease semantic similarity, the miRNA functional similarity, and the Gaussian interaction profile kernel

TABLE 3 | Top 50 potential Lymphomas-related miRNAs predicted by BHCMDA and confirmations for these predicted associations provided by the dbDEMC 2.0 and the recent experimental literatures with relevant PMIDs.

miRNA	Evidence	disease	Evidence
hsa-mir-145	dbDEMC 2.0	hsa-mir-652	dbDEMC 2.0
hsa-mir-34a	dbDEMC 2.0	hsa-mir-221	dbDEMC 2.0
hsa-mir-29b	dbDEMC 2.0	hsa-mir-185	dbDEMC 2.0
hsa-mir-9	dbDEMC 2.0	hsa-mir-596	dbDEMC 2.0
hsa-mir-106b	dbDEMC 2.0	hsa-mir-608	dbDEMC 2.0
hsa-let-7a	dbDEMC 2.0	hsa-mir-223	dbDEMC 2.0
hsa-mir-125b	dbDEMC 2.0	hsa-mir-557	dbDEMC 2.0
hsa-mir-183	dbDEMC 2.0	hsa-mir-192	dbDEMC 2.0
hsa-mir-205	dbDEMC 2.0	hsa-mir-602	dbDEMC 2.0
hsa-mir-30b	dbDEMC 2.0	hsa-mir-181b	dbDEMC 2.0
hsa-mir-29a	dbDEMC 2.0	hsa-mir-214	dbDEMC 2.0
hsa-mir-93	dbDEMC 2.0	hsa-let-7c	dbDEMC 2.0
hsa-mir-199a	dbDEMC 2.0	hsa-let-7i	dbDEMC 2.0
hsa-mir-324	unconfirmed	hsa-mir-612	unconfirmed
hsa-mir-143	dbDEMC 2.0	hsa-mir-657	dbDEMC 2.0
hsa-mir-106a	dbDEMC 2.0	hsa-mir-142	23209550
hsa-let-7b	dbDEMC 2.0	hsa-mir-222	dbDEMC 2.0
hsa-mir-30e	dbDEMC 2.0	hsa-let-7d	dbDEMC 2.0
hsa-mir-638	dbDEMC 2.0	hsa-mir-153	dbDEMC 2.0
hsa-mir-215	dbDEMC 2.0	hsa-mir-367	dbDEMC 2.0
hsa-mir-637	dbDEMC 2.0	hsa-mir-518c	unconfirmed
hsa-mir-195	dbDEMC 2.0	hsa-mir-622	dbDEMC 2.0
hsa-mir-598	dbDEMC 2.0	hsa-mir-583	dbDEMC 2.0
hsa-let-7e	dbDEMC 2.0	hsa-mir-600	dbDEMC 2.0
hsa-mir-615	unconfirmed	hsa-mir-601	dbDEMC 2.0

similarity. In order to estimate the prediction performance of BHCMDA, LOOCV, twofold cross-validation and fivefold cross-validation were implemented, respectively. Moreover, three different kinds of case studies were conducted as well. Simulation results from both case studies and cross-validations demonstrated that BHCMDA had splendid performance in prediction of potential miRNA-disease associations.

There are a few reasons to explain the reliable performance of BHCMDA. In the first place, the data used to predict potential miRNA-disease associations obtained from HMDD V2.0 in this model is rich and reliable. In addition, BHCMDA not only integrates the disease semantic similarity and the miRNA functional similarity with the Gaussian interaction profile kernel similarity, but also applies a clustering algorithm based on the integrated data, which makes the basic data richer and more accurate. In the end, BHC algorithm has the ability to recommend unpopular products. We averaged the predicted data obtained by using BHC algorithm, which made the prediction more reliable.

Whereas there still exist some limitations in BHCMDA. For instance, the quantity of known miRNA-disease associations is still not adequate. In addition, we developed BHCMDA according to the assumption that functionally similar miRNAs are more likely associated with phenotypically similar diseases, which may bring about bias to miRNAs related with more known diseases. Obviously, all these limitations in BHCMDA deserve further study and need to be improved in the future.

DATA AVAILABILITY STATEMENT

Generated Statement: Publicly available datasets were analyzed in this study. These data can be found here: HMDD database (http://www.cuilab.cn/hmdd), miRNA functional similarity (http://www.cuilab.cn/files/images/cuilab/misim.zip), Mesh database (http://www.ncbi.nlm.nih.gov/),dbDEMC database (doi: 10.1186/1471-2164-11-s4-s5), dbDEMC 2.0 (http://www.picb.ac.cn/dbDEMC), HMDD 3.0 (http://www.cuilab.cn/hmdd), miR2Disease (doi: 10.1093/nar/gkn714).

AUTHOR CONTRIBUTIONS

XW and XZ conceived the study. XW, LK, and HZ improved the study based on the original model. XZ and TP implemented the algorithms corresponding to the study. LW, XZ, and LK supervised the study. XW and LW wrote the manuscript. All authors reviewed and improved the manuscript.

FUNDING

This research was partly supported by the National Natural Science Foundation of China (Nos. 61873221 and 61672447)

and the Natural Science Foundation of Hunan Province (Nos. 2018JJ4058, 2019JJ70010, and 2017JJ5036). Publication costs were funded by the National Natural Science Foundation of China (Nos. 61873221 and 61672447).

ACKNOWLEDGMENTS

The authors sincerely thank all the teachers and students who participated in this study for their guidance and help.

REFERENCES

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

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

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

Chen, X., Guan, N. N., Li, J. Q., and Yan, G. Y. (2018a). GIMDA: Graphlet interaction-based MiRNA-disease association prediction. *J. Cell Mol. Med.* 22, 1548–1561. doi: 10.1111/jcmm.13429

Chen, X., Huang, L., Xie, D., and Zhao, Q. (2018b). EGBMMDA: extreme gradient boosting machine for MiRNA-disease association prediction. *Cell Death Dis.* 9:3. doi: 10.1038/s41419-017-0003-x

Chen, X., Huang, Y.-A., Wang, X.-S., You, Z.-H., and Chan, K. C. C. (2016a). FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model. *Oncotarget* 7, 45948–45958. doi: 10.18632/oncotarget.10008

Chen, X., Yan, C. C., Zhang, X., You, Z.-H., Deng, L., Liu, Y., et al. (2016b). WBSMDA: within and between score for miRNA-disease association prediction. *Sci. Rep.* 6:21106. doi: 10.1038/srep21106

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

Chen, X., and Yan, G. Y. (2014). Semi-supervised learning for potential human microRNA-disease associations inference. *Sci. Rep.* 4:5501. doi: 10.1038/srep05501

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

Coiffier, B. (2006). Monoclonal antibody as therapy for malignant lymphomas. *C. R. Biol.* 329, 241–254. doi: 10.1016/j.crvi.2005.12.006

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

Drusco, A., Nuovo, G. J., Zanesi, N., Di Leva, G., Pichiorri, F., Volinia, S., et al. (2014). MicroRNA profiles discriminate among colon cancer metastasis. *PLoS One* 9:e96670. doi: 10.1371/journal.pone.0096670

Enzinger, P. C., and Mayer, R. J. (2003). Esophageal cancer. *New Engl. J. Med.* 349, 2241–2252. doi: 10.1056/NEJMra035010

He, B., Yin, B., Wang, B., Xia, Z., Chen, C., and Tang, J. (2012). MicroRNAs in esophageal cancer (review). *Mol. Med. Rep.* 6, 459–465. doi: 10.3892/mmr.2012.975

Hirota, T., Date, Y., Nishibatake, Y., Takane, H., Fukuoka, Y., Taniguchi, Y., et al. (2012). Dihydropyrimidine dehydrogenase (DPD) expression is negatively regulated by certain microRNAs in human lung tissues. *Lung Cancer* 77, 16–23. doi: 10.1016/j.lungcan.2011.12.018

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

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. doi: 10.3322/caac.20107

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4:S2. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2008). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* 37(Suppl._1), D98–D104. doi: 10.1093/nar/gkn714

Jopling, C. L., Yi, M., Lancaster, A. M., Lemon, S. M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 309, 1577–1581. doi: 10.1126/science.1113329

Le, N. Q., Ho, Q. T., and Ou, Y. Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N. Q., Ho, Q. T., and Ou, Y. Y. (2018). Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. *Anal. Biochem.* 555, 33–41. doi: 10.1016/j.ab.2018.06.011

Le, N. Q. K., Yapp, E. K. Y., Ho, Q. T., Nagasundaram, N., Ou, Y. Y., and Yeh, H. Y. (2019). iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* 571, 53–61. doi: 10.1016/j.ab.2019.02.017

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2013). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Liao, J., Liu, R., Yin, L., and Pu, Y. (2014). Expression profiling of exosomal miRNAs derived from human esophageal cancer cells by solexa high-throughput sequencing. *Intern. J. Mol. Sci.* 15:15530. doi: 10.3390/ijms150915530

Liu, J.-G., Zhou, T., and Guo, Q. (2011). Information filtering via biased heat conduction. *Phys. Rev. E* 84:037101. doi: 10.1103/PhysRevE.84.037101

Luo, J., Xiao, Q., Liang, C., and Ding, P. (2017). Predicting microRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data. *IEEE Access* 5, 2503–2513. doi: 10.1109/ACCESS.2017.2672600

Manfè, V., Biskup, E., Willumsgaard, A., Skov, A. G., Palmieri, D., Gasparini, P., et al. (2013). cMyc/miR-125b-5p signalling determines sensitivity to bortezomib in preclinical model of cutaneous T-cell lymphomas. *PLoS One* 8:e59390. doi: 10.1371/journal.pone.0059390

Meister, G., and Tuschl, T. (2004). Mechanisms of gene silencing by double-stranded RNA. *Nature* 431, 343–349. doi: 10.1038/nature02873

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

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

Ogata-Kawata, H., Izumiya, M., Kurioka, D., Honma, Y., Yamada, Y., Furuta, K., et al. (2014). Circulating exosomal microRNAs as biomarkers of colon cancer. *PLoS One* 9:e92921. doi: 10.1371/journal.pone.0092921

Paraskevi, A., Theodoropoulos, G., Papaconstantinou, I., Mantzaris, G., Nikiteas, N., and Gazouli, M. (2012). Circulating MicroRNA in inflammatory bowel disease. *J. Crohns. Colitis* 6, 900–904. doi: 10.1016/j.crohns.2012.02.006

Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P. (2005). Global cancer statistics, 2002. *CA Cancer J. Clin.* 55, 74–108. doi: 10.3322/canjclin.55.2.74

Png, K. J., Yoshida, M., Zhang, X. H., Shu, W., Lee, H., Rimner, A., et al. (2011). MicroRNA-335 inhibits tumor reinitiation and is silenced through genetic and epigenetic mechanisms in human breast cancer. *Genes Dev.* 25, 226–231. doi: 10.1101/gad.1974211

Saito, Y., Suzuki, H., Tsugawa, H., Imaeda, H., Matsuzaki, J., Hirata, K., et al. (2012). Overexpression of miR-142-5p and miR-155 in gastric mucosa-associated lymphoid tissue (MALT) lymphoma resistant to *Helicobacter* pylori eradication. *PLoS One* 7:e47396. doi: 10.1371/journal.pone.0047396

Sarver, A. L., French, A. J., Borralho, P. M., Thayanithy, V., Oberg, A. L., Silverstein, K. A. T., et al. (2009). Human colon cancer profiles show differential microRNA expression depending on mismatch repair status and are characteristic of undifferentiated proliferative states. *BMC Cancer* 9:401. doi: 10.1186/1471-2407-9-401

Shi, H., Xu, J., Zhang, G., Xu, L., Li, C., Wang, L., et al. (2013). Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes. *BMC Syst. Biol.* 7:101. doi: 10.1186/1752-0509-7-101

Tavazoie, S. F., Alarcon, C., Oskarsson, T., Padua, D., Wang, Q., Bos, P. D., et al. (2008). Endogenous human microRNAs that suppress breast cancer metastasis. *Nature* 451, 147–152. doi: 10.1038/nature06487

Valastyan, S., Reinhardt, F., Benaich, N., Calogrias, D., Szasz, A. M., Wang, Z. C., et al. (2009). A pleiotropically acting microRNA, miR-31, inhibits breast cancer metastasis. *Cell* 137, 1032–1046. doi: 10.1016/j.cell.2009.03.047

Wan, D., He, S., Xie, B., Xu, G., Gu, W., Shen, C., et al. (2013). Aberrant expression of miR-199a-3p and its clinical significance in colorectal cancers. *Med. Oncol.* 30:378.

Wan, J., Wu, W., Che, Y., Kang, N., and Zhang, R. (2016). Insights into the potential use of microRNAs as a novel class of biomarkers in esophageal cancer. *Dis. Esophagus.* 29, 412–420. doi: 10.1111/dote.12338

Wang, B., Wang, H., and Yang, Z. (2012). MiR-122 inhibits cell proliferation and tumorigenesis of breast cancer by targeting IGF1R. *PLoS One* 7:e47053. doi: 10.1371/journal.pone.0047053

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

Xie, L., Ushmorov, A., Leithäuser, F., Guan, H., Steidl, C., Färbinger, J., et al. (2012). FOXO1 is a tumor suppressor in classical Hodgkin lymphoma. *Blood* 119, 3503–3511. doi: 10.1182/blood-2011-09-381905

Xie, Z., Chen, G., Zhang, X., Li, D., Huang, J., Yang, C., et al. (2013). Salivary microRNAs as promising biomarkers for detection of esophageal cancer. *PLoS One* 8:e57502. doi: 10.1371/journal.pone.0057502

Xu, P., Guo, M., and Hay, B. A. (2004). MicroRNAs and the regulation of cell death. *Trends Genet.* 20, 617–624. doi: 10.1016/j.tig.2004.09.010

Xu, X.-L., Jiang, Y.-H., Feng, J.-G., Su, D., Chen, P.-C., and Mao, W.-M. (2014). MicroRNA-17, microRNA-18a, and microRNA-19a are prognostic indicators in esophageal squamous cell carcinoma. *Ann. Thorac. Surg.* 97, 1037–1045. doi: 10.1016/j.athoracsur.2013.10.042

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

Yang, Z., Ren, F., Liu, C., He, S., Sun, G., Gao, Q., et al. (2010). dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 11(Suppl. 4):S5. doi: 10.1186/1471-2164-11-S4-S5

Zou, Q., Li, J., Hong, Q., Lin, Z., Wu, Y., Shi, H., et al. (2015). Prediction of microRNA-disease associations based on social network analysis methods. *Biomed. Res. Int.* 2015:810514. doi: 10.1155/2015/810514

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: 05 May 2020
doi: 10.3389/fgene.2020.00441

Prognostic Value of a Stemness Index-Associated Signature in Primary Lower-Grade Glioma

Mingwei Zhang[1,2,3,4,5†], Xuezhen Wang[1†], Xiaoping Chen[6†], Feibao Guo[1] and Jinsheng Hong[1,3,4*]*

[1] Department of Radiation Oncology, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China, [2] Institute of Immunotherapy, Fujian Medical University, Fuzhou, China, [3] Key Laboratory of Radiation Biology (Fujian Medical University), Fujian Province University, Fuzhou, China, [4] Fujian Key Laboratory of Individualized Active Immunotherapy, Fuzhou, China, [5] Fujian Medical University Union Hospital, Fuzhou, China, [6] Department of Statistics, College of Mathematics and Informatics & FJKLMAA, Fujian Normal University, Fuzhou, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Zhipeng Tao,
Harvard Medical School, United States
Yajun Luo,
First Affiliated Hospital of Chongqing Medical University, China

***Correspondence:**
Feibao Guo
gfb803@fjmu.edu.cn
Jinsheng Hong
13799375732@163.com

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

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

Received: 09 January 2020
Accepted: 09 April 2020
Published: 05 May 2020

Citation:
Zhang M, Wang X, Chen X, Guo F and Hong J (2020) Prognostic Value of a Stemness Index-Associated Signature in Primary Lower-Grade Glioma. Front. Genet. 11:441. doi: 10.3389/fgene.2020.00441

Objective: As a prevalent and infiltrative cancer type of the central nervous system, the prognosis of lower-grade glioma (LGG) in adults is highly heterogeneous. Recent evidence has demonstrated the prognostic value of the mRNA expression-based stemness index (mRNAsi) in LGG. Our aim was to develop a stemness index-based signature (SI-signature) for risk stratification and survival prediction.

Methods: Differentially expressed genes (DEGs) between LGG in the Cancer Genome Atlas (TCGA) and normal brain tissue samples from the Genotype-Tissue Expression (GTEx) project were screened out, and the weighted gene correlation network analysis (WGCNA) was employed to identify the mRNAsi-related gene sets. Meanwhile, the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes enrichment analyses were performed for the functional annotation of the key genes. ESTIMATE was used to calculate tumor purity for acquiring the correct mRNAsi. Differences in overall survival (OS) between the high and low mRNAsi (corrected mRNAsi) groups were compared using the Kaplan Meier analysis. By combining the Lasso regression with univariate and multivariate Cox regression, the SI-signature was constructed and validated using the Chinese Glioma Genome Atlas (CGGA).

Results: There was a significant difference in OS between the high and low mRNAsi groups, which was also observed in the two corrected mRNAsi groups. Based on threshold limits, 86 DEGs were most significantly associated with mRNAsi via WGCNA. Seven genes (*ADAP2*, *ALOX5AP*, *APOBEC3C*, *FCGRT*, *GNG5*, *LRRC25*, and *SP100*) were selected to establish a risk signature for primary LGG. The ROC curves showed a fair performance in survival prediction in both the TCGA and the CGGA validation cohorts. Univariate and multivariate Cox regression revealed that the risk group was an independent prognostic factor in primary LGG. The nomogram was developed based on clinical parameters integrated with the risk signature, and its accuracy for

predicting 3- and 5-years survival was assessed by the concordance index, the area under the curve of the time-dependent receiver operating characteristics curve, and calibration curves.

Conclusion: The SI-signature with seven genes could serve as an independent predictor, and suggests the importance of stemness features in risk stratification and survival prediction in primary LGG.

Keywords: lower grade glioma, The Cancer Genome Atlas, Chinese Glioma Genome Atlas, stemness indices-related signature, prognosis

INTRODUCTION

Lower grade glioma is one of the prevalent and infiltrative types of primary malignant intracranial tumors in adults, the main components of which are diffuse low-grade and intermediate-grade gliomas (Ceccarelli et al., 2016; Ostrom et al., 2018). Despite comprehensive regimens that involve maximum surgical resection and subsequent radiotherapy and chemotherapy, the prognosis of LGG has not improved in the past four decades (Claus et al., 2015). Due to the great intrinsically biological and clinical heterogeneity, the overall survival (OS) of LGG estimates a range from 1 to 15 years, and the response to standard treatment varies from person to person (Cancer Genome Atlas Research et al., 2015). Although the histopathological classification of LGG has traditionally used to predict clinical outcomes, there remains a high intraobserver and interobserver variability, and is often hard to accurately predict outcomes even within the same grade (Coons et al., 1997; van den Bent, 2010). Therefore, it is imperative to search for novel molecular biomarkers for LGG genetic classification. Recently, the 2016 WHO brain tumor classification established the molecular markers for subclassification, including the chromosomal 1p and 19q (chr1p/19q) co-deletion, the isocitrate dehydrogenase (IDH) mutation, and the histone 3 mutational status. However, it seems that these widely utilized biomarkers have provided useful but insufficient prediction for risk stratification of patients with LGG, especially in genetically heterogeneous populations. Thus, novel prognostic parameters are urgently needed to develop and improve the stratification of LGG with the use of multiple advanced molecular platforms.

The complexity and heterogeneity of glioma cells is not only related to its genetic polymorphisms, but also to the characteristics of the microenvironment, such as stemness features and oncogenic and tumor suppressive pathways (Venteicher et al., 2017; Dirkse et al., 2019). Recent advancements have revealed that the populations of glioma stem-like cells are associated with the radio- and chemo-resistance, and with prognosis and tumor recurrence (Yi et al., 2016; Roos et al., 2017). To our knowledge, stemness features have been extracted by the novel stemness indices, including DNA methylation-based stemness index (mDNAsi), mRNA expression-based stemness index (mRNAsi) (Malta et al., 2018). Besides, Pan et al. (2019) developed a 13-gene prognostic signature based on mRNAsi, which suggested the stemness of cancer stem cells (CSCs) and the unfavorable prognosis. However, no study has previously attempted to identify the prognostic and predictive value of stem cell-related genes in LGG.

The scores of mRNAsi in LGG were computed using a one-class logistic regression machine learning algorithm (OCLR), and Tathiane et al. found a strong relationship between mRNAsi and prognosis of glioma, which provided new insights into stratification tumors with distinct clinical outcomes (Malta et al., 2018). However, that study mainly focused on comprehensive pan-cancer analysis. Despite the significant association observed between mRNAsi and OS, however, it was investigated based only on the level of bulky tumor. It is reasonable to take the tumor purity into account in order to further investigate the prognostic value of the stemness index in tumor parenchyma. In addition, a series of genes related to mRNAsi have not been analyzed in detail, and their biological function is also unknown. Meanwhile, the univariable and multivariable survival analyses of predominant clinicopathological factors (age, gender, IDH status, radiation, and chemotherapy status, etc.) and genes related to mRNAsi have not been explored in different cohorts. In order to identify the genes related to mRNAsi, the weighted gene correlation network analysis (WGCNA) was employed. This method takes the interrelation of genes into account for structure generation, instead of regarding genes as single entities. WGCNA has been applied to identify trait-related preserved modules for discovering the key genes (Zhang and Horvath, 2005; Langfelder and Horvath, 2008; Liang et al., 2019).

In addition, the ESTIMATE (Estimation of Stromal and Immune cells in MAlignant Tumor tissues using Expression data) algorithm is one of the most common methods to calculate the tumor purity, and is based on scores related to the level of immune cells infiltration and stromal cells in tumor tissues (Yoshihara et al., 2013). In the current study, the primary purpose was to identify the prognostic value of high- and low-score groups based on the mRNAsi or mRNAsi/purity in a Kaplan-Meier survival analysis. Next, differentially expressed genes (DEGs) were screened from The Cancer Genome Atlas (TCGA) database and the Genotype-Tissue Expression (GTEx) database. Subsequently, the WGCNA was applied for identifying the hub gene clusters and for selecting the stemness indices associated key genes in LGG. Meanwhile, the Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis was employed for function annotation. Finally, the stemness-index associated gene signature was established and validated in the TCGA database and the Chinese Glioma Genome Atlas (CGGA) database, which were used for internal and external validation, respectively.

MATERIALS AND METHODS

Data Source

The high-throughput RNA-seq data of 529 patients with LGG from the TCGA database and 1,152 normal brain tissue samples from the GTEx project were downloaded from the University of California Santa Cruz (UCSC) Xena website[1]. The gene expression profiles were quantified by fragments per kilobase of transcript per million mapped reads (FPKM) normalized estimation and log2-based transformation. Next, DEGs were selected by the "limma" package of R software under the threshold of absolute value of the log2-transformed fold change (FC) > 1 and the adjusted *P*-value (adj.*P*) < 0.05. Besides, the ComBat method was performed to remove the batch effects using the R package "sva."

Acquisition of Stemness Index Based on RNA-Seq

Malta et al. (2018) provided a novel analysis for an oncogenic dedifferentiation evaluation that considered the mRNAsi. The mRNAsi scores of the LGG samples were calculated when a one-class logistic regression machine learning algorithm (OCLR) was applied to LGG datasets from TCGA. The gene expression-based stemness index was represented using β values ranging from zero (no gene expression) to one (complete gene expression). The mRNAsi was obtained from the multiplatform analysis based on this previous research.

Weighted Gene Correlation Network Analysis for Building Stemness-Index Associated Preserved Modules

The WGCNA was developed to discover the correlations among genes by constructing significant modules. The WGCNA analysis was performed by the "WGCNA package" for R (version 1.61)[2] (Langfelder and Horvath, 2008).

Initially, the LGG transcriptome in the TCGA database was taken as a data source. The correlation of the expression levels of 5490 DEGs was analyzed with high precision and accuracy, which was a prerequisite for a WGCNA network development. Next, a parameter β was set based on the correlations of each DEG, which contributed to achieve a scale-free co-expression network. Next, the "blockwiseModules" function was carried out for constructing the network and detecting modules. Furthermore, the relationship between the modules and mRNAsi score was investigated, and the preserved module was determined by the top ranked modules with the strongest connections.

Finally, the key genes from the preserved module were explored. The Inclusive criterion for screening key genes was as follows: correlation (cor.) Gene GS > 0.5 and cor. Gene MM > 0.8 (Pan et al., 2019). Gene significance (GS) was calculated to measure the correlation between genes and sample traits (the values of mRNAsi), and Module Membership (MM) was used to assess the correlation between gene expression profiles and module eigengene. The associations among eigengenes, MM, and sample traits were assessed by Pearson's correlation.

Evaluation and Bioinformatics Analysis of Key Genes

The different expression levels of each key gene were visualized in a heatmap, which was retrieved from the normal tissue and tumor tissue. In addition, the interactions among key genes was visualized in a heatmap based on correlations. Moreover, the identification of the functional annotation was another vital step in the exploration of the potential mechanism of key genes. Thus, gene ontology (GO) enrichment analysis (Gene Ontology Constorium, 2015 and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Wanggou et al., 2016) signaling pathways were performed on a list of key genes. The visualization of results was implemented with the R "ggplot2" package. A *P* value < 0.05, and a false discovery rate (FDR) < 0.05 were considered to determine statistical significance.

Inclusive and Exclusive Criteria of Enrolled Patients for the Construction of the Risk Signature

Inclusion criteria included: (1) patients who suffered from primary LGG (except for recurrent LGG), (2) complete clinicopathological feature, (3) diagnosed with WHO grade II or III glioma, (4) the RNA-sequencing data of samples was available, (5) the OS was set as the primary endpoint, and (6) patients with a minimum follow-up of 90 days.

The exclusive criteria were as follows: (1) patients with a pathological diagnosis of recurrence LGG, (2) patients who suffered from brain tumors other than LGG, and (3) absent survival status and clinicopathological parameters.

Survival Analysis of mRNAsi

ESTIMATE, an algorithm based on a web tool[3] provided information for the purity of the tumor tissue calculation (Yoshihara et al., 2013). The data of mRNA expression-based stemness index was calculated for each sample, and the Kaplan Meier analysis for samples with the high and low mRNAsi set was carried out. In view of the effects of tumor purity on the corresponding mRNAsi, the corrected mRNAsi (mRNAsi/tumor purity) was included. From another perspective, the survival rate between the high and low mRNAsi groups was re-compared using a Kaplan Meier analysis based on the corrected mRNAsi scores.

Construction of a Prognostic Signature

A univariate Cox regression analysis was performed by the "survival" package in R to identify genes that are highly associated with and crucial for survival. The prognostic key genes were then further optimized by the least absolute shrinkage and selection operator (LASSO) regression model, using the R

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

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

[3]https://bioinformatics.mdanderson.org/estimate/

package "glmnet." After completing the variable selection and the shrinkage of prognostic key genes, a stepwise multivariate Cox regression analysis was performed to generate the risk score model. The following formula was built based on the coefficients and expression levels for each gene.

$$\text{Model}: Riskscore = \sum_{i=1}^{k} \beta i Si$$

Where k indicates the number of signature genes, β is equal to the coefficient index, and Si represents the expression level of key genes.

Afterward, using the "survminer" package in R (Li et al., 2019), the optimum cutoff value was obtained, and the primary LGG patients in the TCGA database were clustered into high-risk and low-risk groups. The gap of survival rates between the two groups was tested by the Kaplan–Meier analysis. The time dependent ROC was plotted in order to determine whether the risk score can accurately predict the survival status. Finally, the expression distributions of signature genes were shown in a heatmap using the "ComplexHeatmap" R package. The risk plot showed that the LGG patients in the TCGA database sorted by the rank of corresponding risk score.

Prognostic Value of the Seven-Gene-Based Signature

The patients suffering from primary LGG in the TCGA dataset were randomly categorized into the training group (accounting for 70%) and internal validation group (accounting for 30%) by using the "caret" package[4]. The risk scores and the corresponding clinical variants, including age, gender, grade, radiotherapy, chemotherapy, and IDH status were subjected to univariate and multivariate Cox model. Subsequently, proportional hazards assumption for different variables (Therneau, 1994) was examined by the scaled Schoenfeld residuals (Schoenfeld, 1982; R Development Core Team, 2014). In order to achieve the clinical application of survival prediction model, a prognostic nomogram was then constructed based on the outcomes of the multivariate Cox regression analysis (method = "enter"). Using the "rms," "foreign," and "survival" R packages, the nomogram was plotted based on the prognostic signature and six clinicopathology factors for the purpose of predicting 3-, and 5-OS of LGG. Furthermore, the concordance index (C-index) (Harrell et al., 1996) was employed to quantify predictive accuracies by using "survival" and "pec" package. Using the "timeROC" package of R, the time-dependent ROC curve was performed to estimate the prognostic power of the nomogram. To compare the accuracy and discrimination of different models (containing model 1: SI-risk signature; model 2: mRNAsi; model 3: corrected mRNAsi; model 4: six predominant clinic-pathological factors; model 5: model 4 + SI-risk signature), the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were applied by using "survIDINRI" package (Pencina et al., 2008). Calibration curves were employed to evaluate the agreement between the observed and the predicted probability (3- and 5-years OS) in the nomogram. The bootstrap method with 1,000 resamples were utilized to evaluate both discrimination and calibration.

[4]https://cran.r-project.org/web/packages/caret

External Validation of the Prognostic Signature

Another primary LGG of gene expression information and related predominant clinical and prognostic factors were downloaded from the CGGA platform[5]. A total of 353 samples were enrolled for external validation of the risk signature. The samples were uniformly divided into two distinct groups according to the same cutoff value (1.495), and the Kaplan–Meier analysis was employed to assess the high-risk and low-risk groups. Afterward, the ROC curve analysis was used to assess the discriminatory power of the risk score in the external validation set. Further, a heatmap was generated to show the gene expression distributions of signature genes in the CGGA database, and the risk plot showed the distribution of the LGG patients according to their individual risk score. Similarly, the C-index, the time-dependent ROC curves, and calibration curves (bootstrap method with 1,000 resamples) were compared to determine the performance of the risk signature.

Cancer Cell Line Encyclopedia (CCLE) and Protein Expression Verification

The mRNA expression of seven genes profiled by RNA-Seq extracted from database available at The Cancer Cell Line Encyclopedia (CCLE)[6] (Barretina et al., 2012). This portal covers genomic and expression data for more than 1000 cell lines from various tumors. The expression level of seven genes were analyzed in different types of cancer including LGG using CCLE. Cell lines of LGG were preliminary confirmed through six dedicated websites[7] and only the consistent LGG cell lines be retained. In addition, the protein expression levels of the seven genes between glioma tissue and normal control were analyzed using Human Protein Atlas database[8], and the data were visualized using immunohistochemistry staining.

Statistical Analysis

The statistical analysis in our exploratory study was carried out using the R software (version 3.6.0)[9]. For differentially expressed gene selection, the Wilcoxon test was performed. The OCLR method was implemented with the "gelnet" package1 with default parameters (Sokolov et al., 2016). Pearson's chi-square tests and Kruskal–Wallis tests were used to detect the variables difference. An analysis of the distinctness of survival between the two risk groups was illustrated by the Kaplan–Meier curve (Klein and Moeschberger, 1997) with the Wilcoxon logrank test using the

[5]http://cgga.org.cn/

[6]https://portals.broadinstitute.org/ccle

[7]https://web.expasy.org/cellosaurus/, https://www.atcc.org/, https://www.pheculturecollections.org.uk/products/celllines/generalcell/search.jsp, http://igrcid.ibms.sinica.edu.tw, https://cansarblack.icr.ac.uk/, https://www.dsmz.de/

[8]http://www.proteinatlas.org/

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

R package KMsurv. The univariate Cox regression analysis and multivariate Cox regression analysis were performed to assess the association between the factors and OS (Therneau, 2015). A $p < 0.05$ was deemed as statistically significant.

RESULTS

Data Processing

Identification of DEGs

The overview of the stemness index-related signature development and validation workflow is summarized in **Figure 1**. A total of 774 patients with primary LGG were enrolled in the generation of the stemness indices-associated risk signature, and the clinicopathological characteristics are listed in **Table 1**. The RNA-seq data (level 3) of 1,152 normal brain tissue samples and 529 LGG samples from GTEx projects and the TCGA were screened by the limma package. Before the identification of DEGs, the normalization and batch effect removal were tested. As illustrated in **Supplementary Figures S1A,C**, it performed well in normalization. Correspondingly, TCGA and GTEx samples separated obviously (**Supplementary Figures S1B,D**). Altogether, using the cutoff of significance of the absolute value of the log2-transformed fold change (FC) > 1 and the adjusted *P* value (adj.P) < 0.05, the differential expression analysis between 1,152 normal control samples and 529 LGG identified a cohort of 5,490 DEGs, of which 2,718 were upregulated and 2,772 were downregulated (**Supplementary Figures S2A,B**).

mRNAsi Mining

Gene expression-based stemness indices for LGG were extracted by the one-class logistic regression machine learning algorithm (OCLR) (Malta et al., 2018). A cohort of LGG samples stratified by the mRNAsi, which is based on the stemness index model, were utilized for the integrative analyses.

WGCNA: Construction the Correlation Matrix of mRNAsi and Module Eigengene Values

Data Acquisition

Using the TCGA database, a WGCNA network was constructed by the WGCNA package for the purpose of identifying stemness indices-related modules. The LGG transcriptome in the TCGA database was employed as the primary source for the analysis. Afterward, a global view of RNA-seq data analysis specific to LGG were provided by the WGCNA.

After data preprocessing, a correlation analysis of 5,490 DEGs was conducted, and the soft threshold power of β was 5 (scale-free $R^2 = 0.9$) to assure a scale-free topology model (**Supplementary Figure S3A**). A total of 5,490 DEGs were screened for further analysis according to the exclusion criteria.

Next, a clustering analysis on this basis for LGG identified a total of eleven diverse modules (module size ≥ 50 and cut height ≥ 0.25) in the network (purple, turquoise, black, brown, magenta, green, red, yellow, blue, pink, and gray). Genes in the same color module demonstrated common gene expression patterns (**Supplementary Figure S3B**).

Identification of Modules Associated With Stemness Indexes of LGG

Fold enrichment > 1 and $p < 0.05$ was regarded as the statistical threshold of significance for mRNAsi associated modules selection. There were ten sets of genes (modules) identified that were significantly associated with mRNAsi. The purple, brown, magenta, red, and gray modules were correlated negatively with mRNAsi (ME_{purple}:$r = -0.094$, $P = 0.04$, ME_{brown}:$r = -0.77$, $P = 3E^{-100}$, $ME_{magenta}$:$r = -0.27$, $P = 4E^{-10}$, ME_{red}:$r = -0.11$, $P = 0.01$, ME_{gray}:$r = -0.13$, $P = 0.005$). The turquoise, black, yellow, blue, and pink modules were correlated positively to mRNAsi ($ME_{turquoise}$:$r = 0.29$, $P = 3E^{-11}$, ME_{black}:$r = 0.15$, $P = 0.001$, ME_{yellow}:$r = 0.36$, $P = E^{-16}$, ME_{blue}:$r = 0.6$, $P = 4E^{-49}$, ME_{gray}:$r = 0.37$, $P = 2E^{-17}$) (**Figure 2** and **Supplementary Figure S3**).

The module-trait relationships showed that the brown module was most significantly related to mRNAsi, with the highest correlation value ($r = -0.77$, $P = 3E^{-100}$). Thus, the brown module was selected for subsequent analyses to explore key genes.

Based on the threshold limits (cor. gene GS > 0.5 and cor. gene MM > 0.8), 86 out of 748 hub genes were identified after selection in the brown module.

Analysis and Functional Annotation of Key Genes in the Brown Module

Analysis of Key Genes in the Brown Module

The expression values of each key gene were retrieved from the normal control tissue and tumor tissue, which were visualized as heatmap (**Supplementary Figure S4A**). The heatmap showed that most of the key genes had median expression levels in tumor tissue, whereas CD74 Molecule (*CD74),* major histocompatibility complex, class I, E (*HLA-E),* major histocompatibility complex, class II, DR Alpha *(HLA-DRA),* major histocompatibility complex, class II, DR Beta 1 *(HLA-DRB1),* complement C1q B chain (*C1QB*), complement C1q A chain (*C1QA*), and complement C1q C chain (*C1QC*) exhibited the higher expression in samples from cancer patients. The correlation analyses between key genes were also visualized as a heatmap (**Supplementary Figure S4B**).

Functional Annotation of Genes Related to mRNAsi

The Gene ontology enrichment analysis was executed for further describing the function of the key genes. In total 30 GO biological processes consisting of 10 biological processes (BP) terms (regulation of leukocyte activation, etc.), 10 cellular components (CC) terms (secretory granule membrane, etc.), and 10 molecular functions (MF) terms (peptide binding, etc.) were enriched (**Figure 3A**).

In addition, KEGG signaling pathway analysis indicated that the key genes were significantly enriched in 30 pathways, and several pathways were immune-related, such as antigen processing and presentation and cell adhesion molecules (CAMs) (**Figure 3B**). The above results suggest the potential regulatory mechanism of mRNAsi-associated genes in the development of LGG.

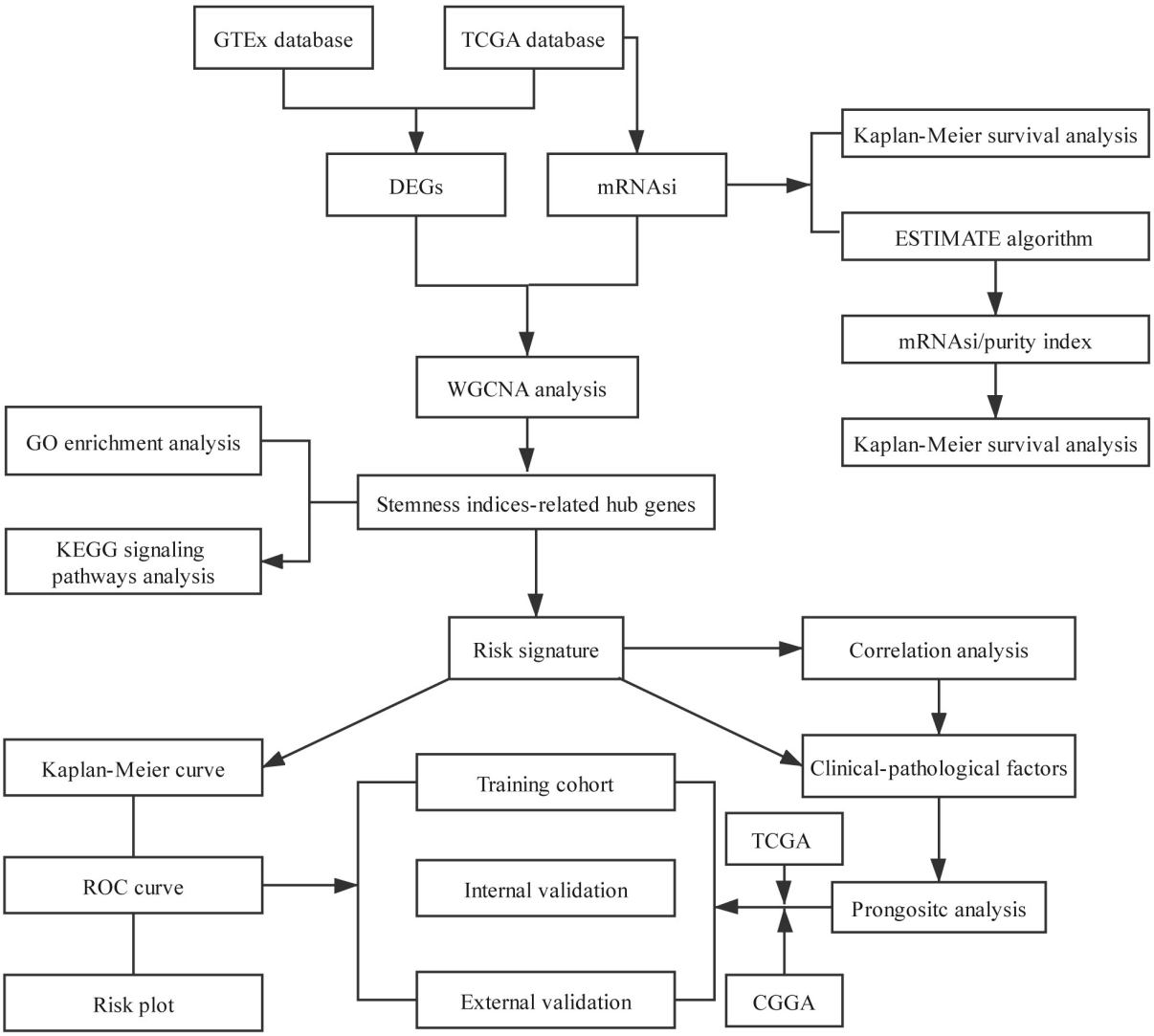


FIGURE 1 | Flowchart presenting the process of establishing the stemness index-related signature.

Survival Analysis of mRNAsi

After calculating the mRNAsi for all LGG samples, a cohort of 447 patients with LGG were classified into either a high mRNAsi score group or a low mRNAsi score group, using the optimum cutoff value of 0.354. The survival curves showed that the OS values were significantly different between the two groups ($P = 9.676E^{-4}$), based on Kaplan-Meier survival analysis (**Supplementary Figure S5A**).

Considering the interferences of tumor purity, the corrected mRNAsi (mRNAsi/tumor purity) was adopted. By applying ESTIMATE (Yoshihara et al., 2013), the tumor purity was calculated in any given LGG sample.

Similar results were also observed when the Kaplan-Meier survival analysis was applied to all the 463 samples based on corrected mRNAsi. There was a significant difference in OS between high mRNAsi score group and low mRNAsi score group ($P = 5.019E^{-4}$) (**Supplementary Figure S5B**).

Identification of Key Prognostic Genes in Primary LGG

To find out the prognostic value of stemness-index associated genes, 86 key genes were tested by univariate Cox regression analysis. It was found that 80 genes were significantly associated with OS in primary LGG. Surprisingly, all prognostic key genes were identified as risk factors (**Figure 4**).

Construction of Stemness-Index Associated Prognostic Signatures

Taking co-linearity into account, 80 key prognosis-related genes were subjected to LASSO Cox regression. A set of 11 key genes were then included in the subsequent analysis with non-zero regression coefficients. Next, 7 key genes were filtered and optimized for constructing a risk signature when implementing the stepwise multivariable Cox regression analysis (**Table 2**). The 7 key genes contained

TABLE 1 | Clinicopathological characteristics of primary LGG patients from the TCGA and CGGA databases.

Characteristic	Training cohort	Internal validation cohorts	External validation cohorts
	TCGA (*n* = 297)	TCGA (*n* = 124)	CGGA (*n* = 353)
Age (Y)[a]			
≤40	136 (46%)	63 (51%)	189 (54%)
>40	161 (54%)	61 (49%)	164 (46%)
Gender			
Male	168 (57%)	65 (52%)	205 (58%)
Female	129 (43%)	59 (48%)	148 (42%)
Grade			
I	144 (52%)	54 (44%)	196 (56%)
II	153 (48%)	70 (56%)	157 (44%)
Radiation			
No	99 (37%)	53 (43%)	59 (17%)
Yes	198 (63%)	71 (57%)	294 (83%)
Chemotherapy			
No	133 (45%)	58 (47%)	147 (42%)
Yes	164 (55%)	66 (53%)	206 (58%)
IDH[b] Status			
Wild-type	53 (18%)	26 (21%)	94 (27%)
Mutation	244 (82%)	98 (79%)	259 (73%)
Risk score			
Low risk	209 (70%)	80 (65%)	264 (75%)
High risk	88 (30%)	44 (35%)	89 (25%)

[a]*Age, Age at pathological diagnosis of glioma.* [b]*IDH, Isocitrate dehydrogenase.*

ArfGAP with dual PH domains 2 (*ADAP2*), arachidonate 5-lipoxygenase activating protein (*ALOX5AP*), apolipoprotein B mRNA editing enzyme catalytic subunit 3C (*APOBEC3C*), Fc fragment of IgG receptor and transporter (*FCGRT*), G protein subunit gamma 5 (*GNG5*), leucine rich repeat containing 25 (*LRRC25*), and SP100 nuclear antigen (*SP100*). Finally, a risk score formula was developed based on the seven key genes along with their individual coefficients and expression level, which was defined as follows: (−0.88603 × expression level of *ADAP2*) + (0.416964 × expression level of *ALOX-5AP*) + (0.914674 × expression level of *APOBE-C3C*) + (−0.73585 × expression level of *FCGRT*) + (0.631697 × expression level of *GNG5*) + (−0.64501 × expression level of *LRRC25*) + (0.745358 × expression level of *SP100*).

Evaluation of Survival Predicts the Accuracy of Seven-Gene-Based Signature

The robustness of the seven stemness-index associated genes was validated by evaluating the ability of stratifying the high- or low-risk group in TCGA datasets. Patients with primary LGG were dichotomized into high- (risk score ≥ 1.495) or low-risk group (risk score < 1.495) based on the optimal cutoff values. The Kaplan–Meier survival curve analysis showed that different risk groups by this risk scoring system were significantly linked with OS (**Figure 5A**). Next, the 1y-, 3y-, and 5y-AUC of the time-dependent ROC were 0.899, 0.875, and 0.778, respectively (**Figure 5B**), confirming the satisfactory prediction efficiency of the seven-gene stemness index-based signature in OS. Furthermore, as observed in the heatmap, *FCGRT* and *GNG5* had the highest expression levels, whereas *LRRC25*, *SP100*, *ALOX5AP*, *ADAP2*, and *APOBEC3C* exhibited low and medium expression levels (**Figure 5C**). Consecutively, the distribution of risk scores and survival status showed that patients with a risk score of 1.495 or higher generally had poorer survival when compared with another group (**Figure 5D**).

Prognostic Value of the Seven Gene-Based Signature

A cohort of 421 patients with primary LGG in the TCGA database were classified into training set (n = 297) and internal validation set (n = 124) randomly at a ratio of 7:3. In consideration of the prognostic value of the stemness-index associated signature, the risk score was set as a potential factor and explored by the univariable and multivariable Cox regression analysis. The forest plot of the univariable Cox regression analysis, based on 6 clinicopathologic features showed that risk group (HR = 6.648, $p < 0.001$), age (HR = 3.573, $p < 0.001$), grade (HR = 2.864, $p < 0.001$), radiation therapy (HR = 2.137, $p = 0.014$), and IDH status (HR = 0.143, $p < 0.001$) were prognostic elements associated with OS (**Figure 6A**). Next, the results revealed that risk (HR = 4.545, $p < 0.001$), age (HR = 3.399, $p < 0.001$), and IDH status (HR = 0.330, $p < 0.001$) were statistically significant in multivariable Cox regression analyses (**Figure 6B**).

Based on the above results, the nomogram was established for predicting primary LGG 3- and 5-years survival, which integrated both the unique risk score and clinicopathologic variables (**Figure 6C**). The C-index of the nomogram was 0.8701 (95% CI; 0.8358–0.9044). The area under the curves (AUCs) of the 3- and 5-years OS predictions for the constructed nomogram were 0.905, and 0.837 in the training set, respectively (**Figure 6D**). Meanwhile, the calibration curves for this nomogram were developed and plotted in **Figures 6E,F**.

In addition, the comparison of the accuracy and discrimination in five models were conducted. The c-indexes of five models were 0.775, 0.658, 0.615, 0.852, and 0.870, respectively (**Figure 7A**). Moreover, as shown in **Table 3**, when defined the model 1 as the reference, the continuous NRI for the 1y-, 3y-follow ups were significant lower in mRNAsi group (model 2) with NRIs were -0.598 (P = 0.01) and -0.548 (P = 0.022). Correspondingly, the continuous NRI for the 1y-, 3y-follow ups were also significant lower in corrected mRNAsi group (model 3), with NRIs were -0.663 ($P < 0.001$) and -0.508 ($p < 0.001$). Conversely, the 1y-, 3y-NRI were significantly improved in model 4 and model 5 with NRIs were 0.458 (P = 0.016), 0.317 (P = 0.028), 0.708 ($P < 0.001$), and 0.433 ($P < 0.001$). Furthermore, the comparison between the model 4 and model 5 was also conducted. The 1y-, 3y-NRIs were also significant higher in model 5 (comprising all the seven factors in nomogram).

Moreover, the 3y-, 5y IDI were significantly decreased in model 2 (IDI = -0.146 and -0.189). The 1y-, 3y-IDI were significantly decreased in model 3 (IDI = -0.063 and -0.178) with

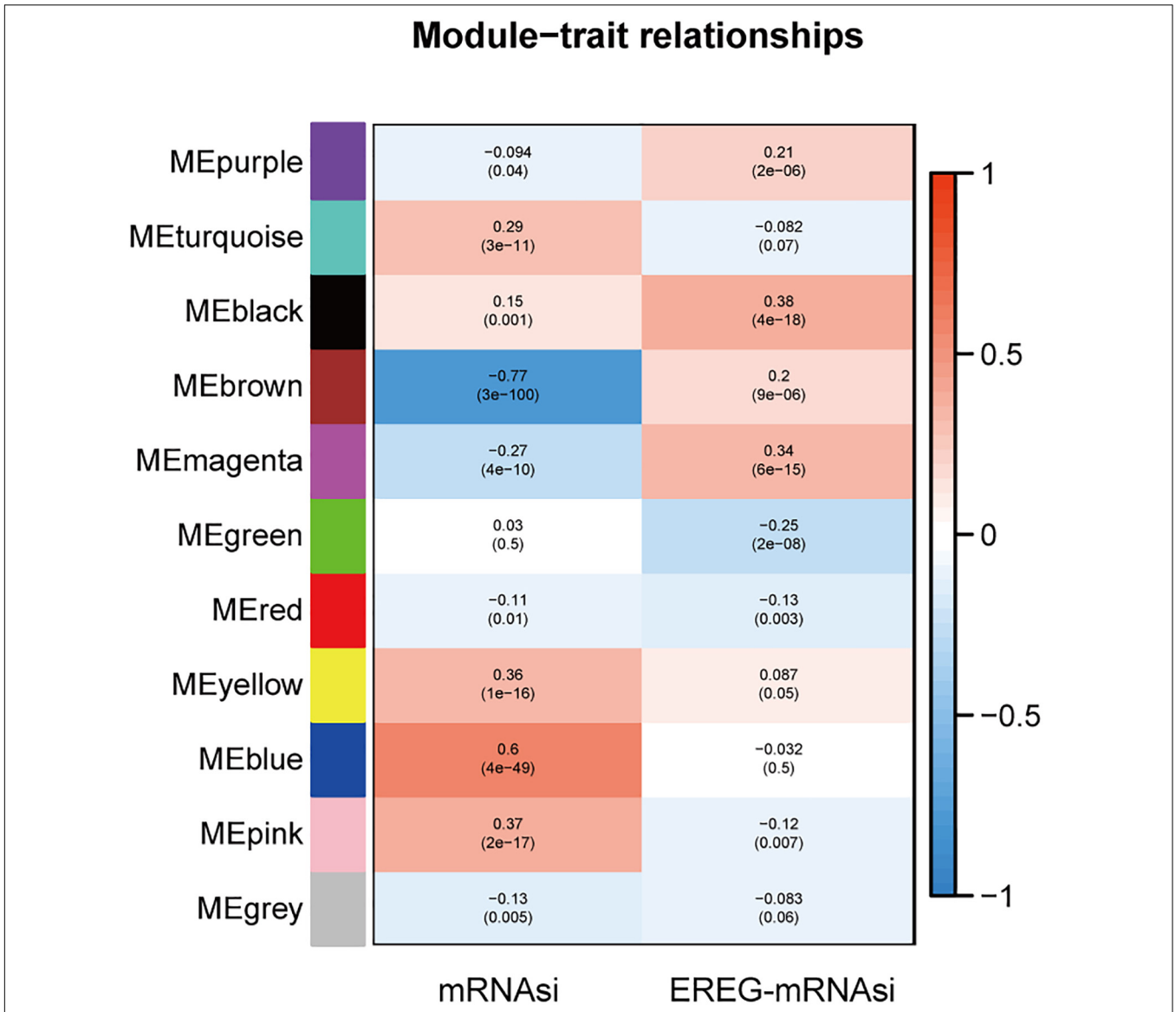


FIGURE 2 | The analysis of the weighted gene correlation network analysis (WGCNA) modules showed that there were ten sets of genes (modules) significantly associated with mRNAsi. The purple, brown, magenta, red, and gray modules were correlated negatively with mRNAsi.

borderline significance in 5y-IDI ($P = 0.056$). Conversely, the 1y-, 3y-IDI were significant higher in model 4 (IDI = 0.084 and 0.165). Interestingly, 1y-, 3y-, 5y-IDI were all significant improved in model 5. In terms of the comparison of IDI between model 4 and model 5, despite the IDI were all improved, however, the *P* values could not reach the levels of significance.

Internal Validation of Seven-Gene Stemness-Index Associated Prognostic Signature

Meanwhile, the clinical predictive model was evaluated in an internal validation set. The C-index was 0.8474 (95% CI; 0.7081–0.7971), the area under the curves (AUC) for 3 and 5-years-survival were 0.915 and 0.828, respectively (**Figure 7B**). Taking the calibration curves for the nomogram-probability of 3-years survival (**Figure 7C**) and 5-years survival (**Figure 7D**) together, the seven-gene signature was capable of predicting the OS of primary LGG patients with high efficiency.

Development and External Validation of the Prognostic Signature

According to the same cut-off value, the external validation set of 353 patients in the CGGA platform was employed and divided into high-risk cohort ($n = 89$) and low-risk cohort ($n = 264$). Similar procedures were conducted to assess the performance of the stemness-index associated signature. Using the Kaplan-Meier curve analysis, the high-risk cohort also showed a significantly poorer prognosis than the low-risk cohort ($P = 6.924E^{-13}$) (**Figure 8A**). The 1y-, 3y-,

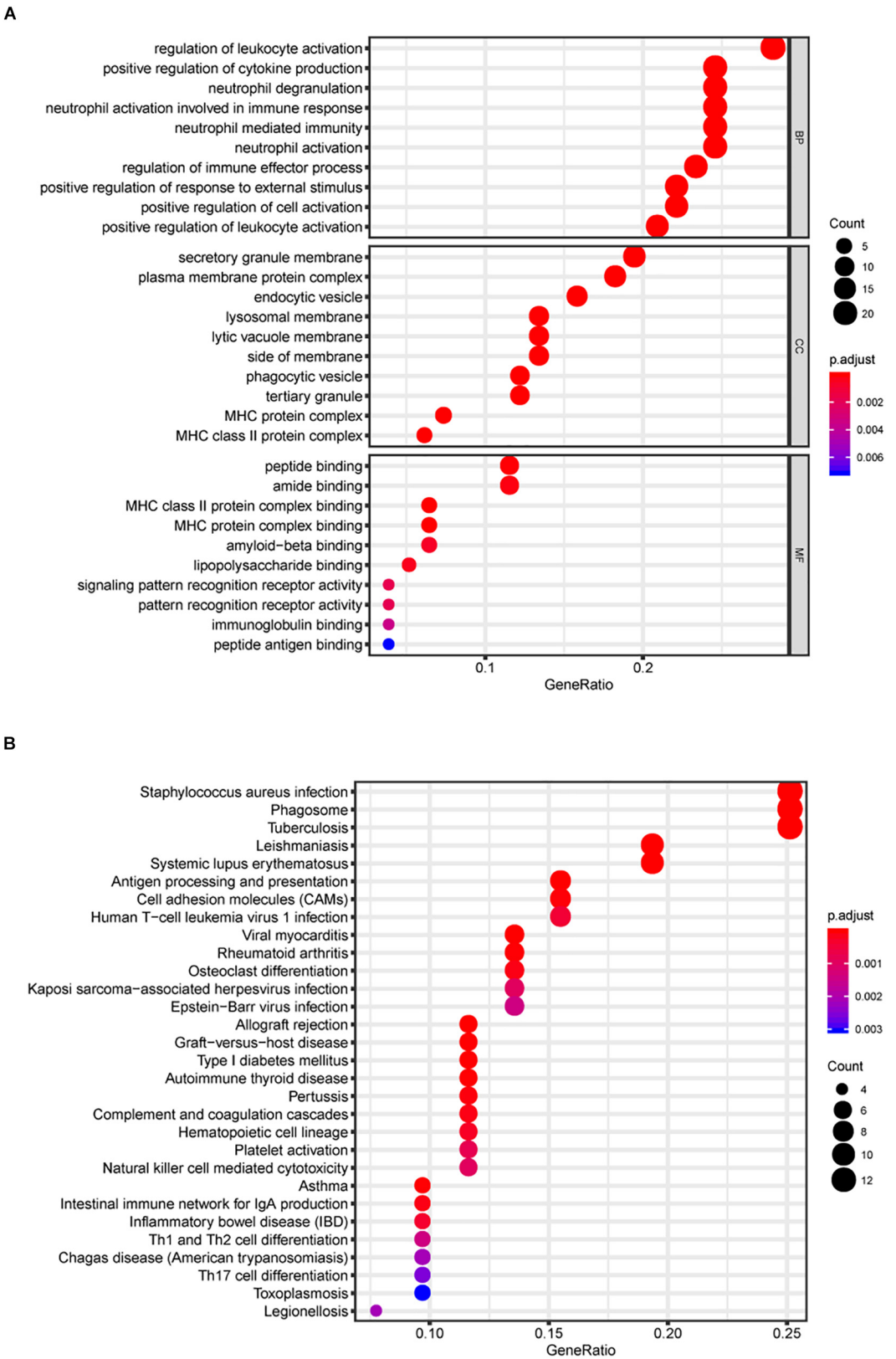


FIGURE 3 | Gene Ontology (GO) and KEGG pathway analyses. In total, 30 GO biological process consisting of 10 biological processes (BP) terms, 10 cellular components (CC) terms, and 10 molecular functions (MF) terms were enriched **(A)**. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis **(B)**.

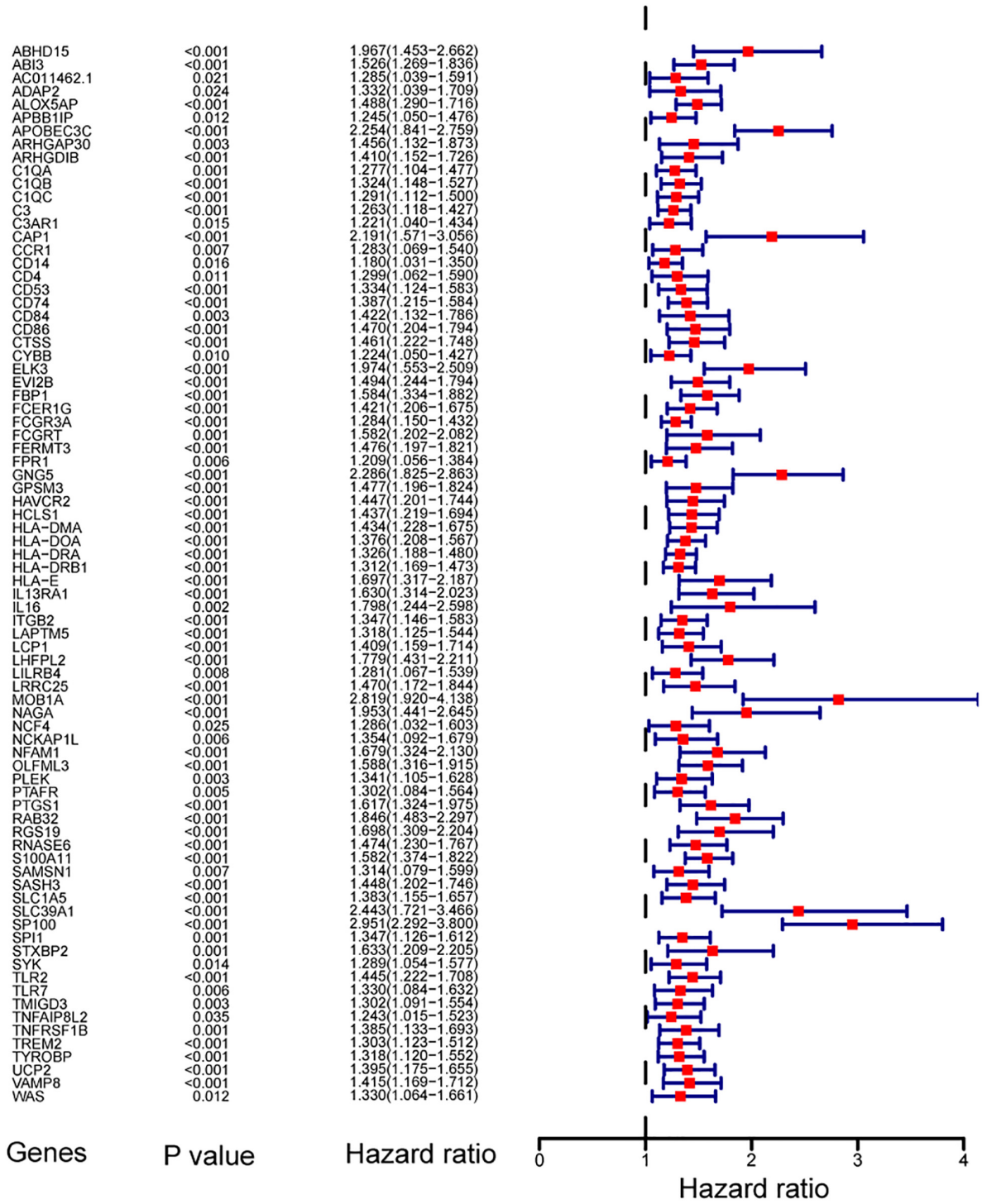


FIGURE 4 | Forest plot showing the hazard ratios from the univariate Cox regression analysis.

5y-AUC in the external validation set were 0.708, 0.727, and 0.725, respectively (**Figure 8B**). In accordance with the risk plot in the TCGA database, In accordance with the risk plot in the TCGA database, there was an inverse relationship between risk score and survival (**Figure 8C**). Subsequently, the AUCs for 3- and 5-years OS were 0.798, and 0.74, respectively (**Figure 8D**). The C-index in the external validation set was 0.7526. The calibration curves for the nomogram 3- and 5-year survival probabilities are shown in **Figures 8E,F**, respectively.

TABLE 2 | Results of the seven key genes in the multivariable Cox regression analysis.

Genes	Coef	HR	HR.95L	HR.95H	*P*-value
ADAP2	−0.886027724	0.412290235	0.22399138	0.758882942	0.004423098
ALOX5AP	0.416963664	1.517347377	1.10755058	2.07877013	0.009433367
APOBEC3C	0.914673555	2.495960324	1.783799483	3.492442955	9.47E-08
FCGRT	−0.735850888	0.479097627	0.294704586	0.778863129	0.002997479
GNG5	0.631697047	1.880799678	1.385572452	2.553029562	5.09E-05
LRRC25	−0.645008868	0.52465789	0.315056856	0.873702304	0.01318087
SP100	0.745358173	2.107196041	1.21584173	3.652017402	0.00789528

Evaluation of the Correlation Between Clinical Parameters and Signature

The relationship between the clinicopathological features (age, gender, grade, radiotherapy, chemotherapy, and IDH mutation status) and the seven-gene-based signature was explored. Older patients, patients of grade III, and IDH wild type tended to have higher risk scores than the younger, grade II, and the IDH mutant type patients, respectively in the TCGA database (**Supplementary Figure S6A**). As for the CGGA database, the risk scores of patients with IDH1 mutant type, and grade II were lower than IDH1 wild type, and grade III, respectively (**Supplementary Figure S6B**).

Expression Analysis of Seven Genes From Cancer Cell Line Encyclopedia (CCLE) and Human Protein Atlas Database

To validate the mRNA expression of seven genes, the expression levels of *ADAP2*, *ALOX5AP*, *APOBEC3C*, *FCGRT*, *GNG5*, *LRRC25*, and *SP100* in various human tumors and 14 LGG cell lines from the CCLE were determined (**Supplementary Figure S7** and **Table 4**). As shown in **Supplementary Figure S7**, the mRNA expression of *APOBEC3C*, *FCGRT*, *GNG5*, and *SP100* was elevated in glioma, whereas the expression of *ADAP2*, *ALOX5AP*, and *LRRC25* was low. To further explore the expression patterns of the seven genes in tissue level, the Human Protein Atlas database was employed to analyze the differential expression between glioma tissue and normal control, and the protein expression was evaluated using immunohistochemistry data as shown in **Supplementary Figure S8**. Consistent with the RNA-seq data, the protein expression levels of *FCGRT*, and *GNG5* were upregulated in tumor tissues when compared with the normal controls.

DISCUSSION

In previous studies, the risk stratification of the stemness index has been investigated in pan-cancer cohorts. However, the comprehensive prognostic value of the stemness index has not been exploited in LGG. In addition, the function annotation of the stemness index-associated genes and the prognostic value of the risk signature have not been investigated. In the current study, significant differences were found in survival between low- and high mRNAsi (mRNAsi/purity) score groups in the Kaplan Meier curve. Moreover, the detail of stemness indices-related modules and genes were identified after the application of WGCNA. A total of 86 key genes were screened according to the threshold limits, which were most significantly correlated with stemness-index. Next, for the enrichment analysis of the brown module, GO terms consisting of "regulation of leukocyte activation," "positive regulation of cytokine production," and "neutrophil degranulation" were ranked at the top of the list. In addition, KEGG pathway results such as CAMs, natural killer cell mediated cytotoxicity, and antigen processing and presentation were also obtained. Next, after the application of univariate Cox regression analysis, LASSO Cox regression model, and multiCox analysis, seven key genes (*ADAP2*, *ALOX5AP*, *APOBEC3C*, *FCGRT*, *GNG5*, *LRRC25*, and *SP100*) were enrolled as vital elements in stemness index-related signature. Furthermore, age, grade, radiotherapy, IDH status, and risk group were significantly associated with OS in the univariable Cox regression analysis; however, only age, IDH status, and risk group were significantly correlated with OS for primary LGG patients by applying the multivariate Cox regression analysis.

In the first part, it was found that the mRNAsi was significantly associated with OS in primary LGG, which was consistent with a previous study in pan-cancer cohorts (Malta et al., 2018). However, it should be noticed that the population of bulky tumor includes tumor cells, immune cells, and stromal cells. Taking the tumor purity into account may accurately reflect the actual stemness characteristic in tumor parenchyma. Moreover, ESTIMATE is one of the most common algorithms for quantifying tumor purity and composition of stromal and immune cells. Hence, the concept of the corrected mRNAsi (mRNAsi/tumor purity) was adopted to reduce the interference of non-tumor tissue (Malta et al., 2018; Lian et al., 2019; Pan et al., 2019). Of note, after employing the survival analysis, the significant survival difference in OS was still observed between high- and low- score groups based on the corrected mRNAsi (mRNAsi/tumor purity), which was consistent with the results from a previous study of bladder cancer (Pan et al., 2019). Additionally, the comparisons of the accuracy and discrimination among three models (model 1, model 2, and model 3) were conducted. Interestingly, the constructed risk signature in current study was superior to the mRNAsi and corrected mRNAsi in predicting the overall survival of LGG. To our knowledge, there is no previous study investigating the improvements of the accuracy and discrimination between mRNAsi and corrected mRNAsi. Further pan-cancer analyses are warranted.

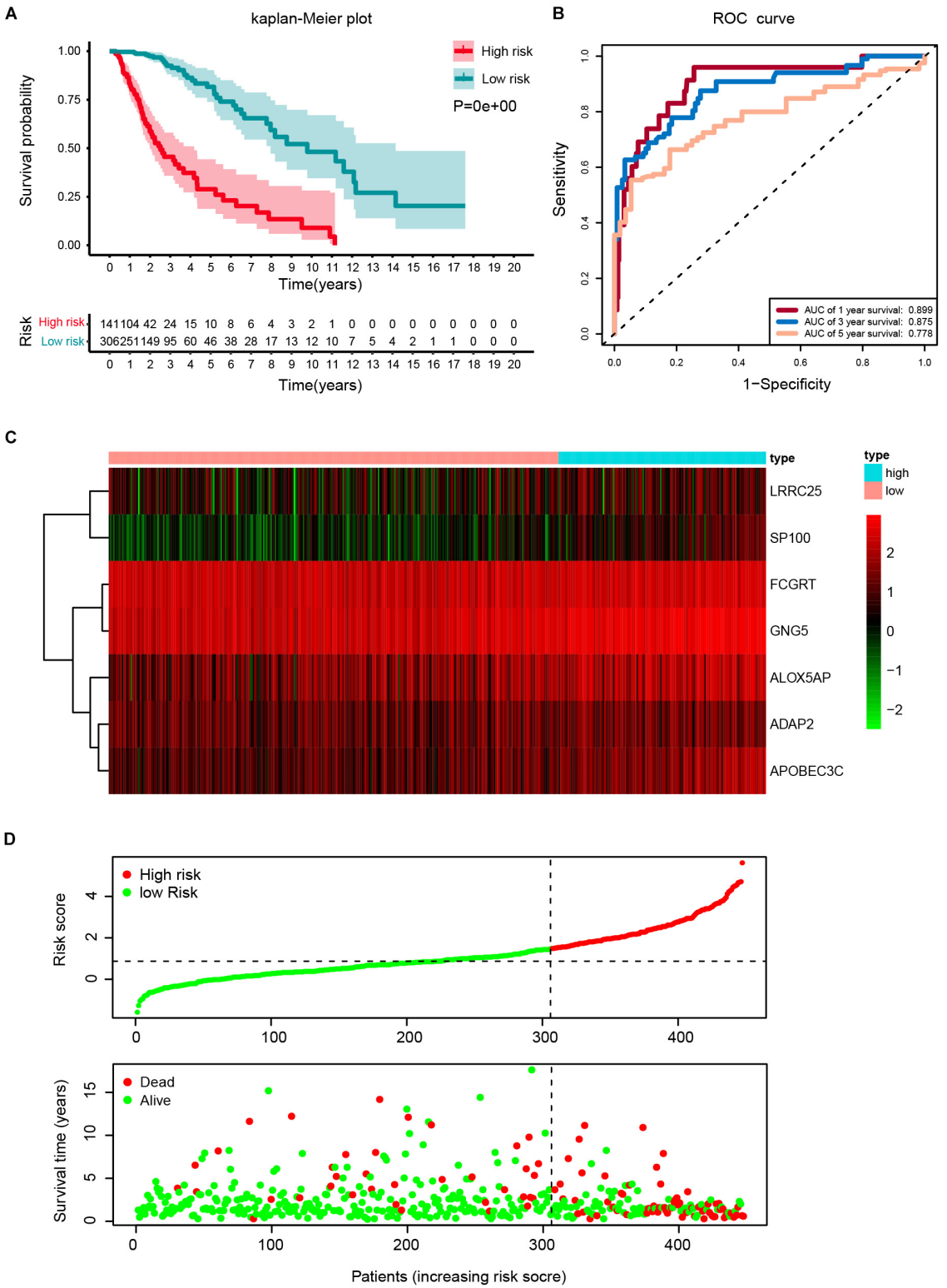


FIGURE 5 | Construction of a risk score based on the seven stemness indices-related gene signature in the TCGA cohort. **(A)** Kaplan-Meier analysis of OS for low-risk and high-risk patients in the training cohort. Additionally, the table indicating the number at risk for each group at corresponding time points. **(B)** The time-dependent receiver operating characteristics (ROC) curve for 1-, 3-, and 5-year OS predictions for stemness-index related risk signature. **(C)** Heatmap showing the distribution of the expression of the seven genes of the stemness index in the TCGA cohort. **(D)** Risk plot presenting each point sorted based on risk score, representing one patient. Green and red points represent patients with low- and high-risk, respectively.

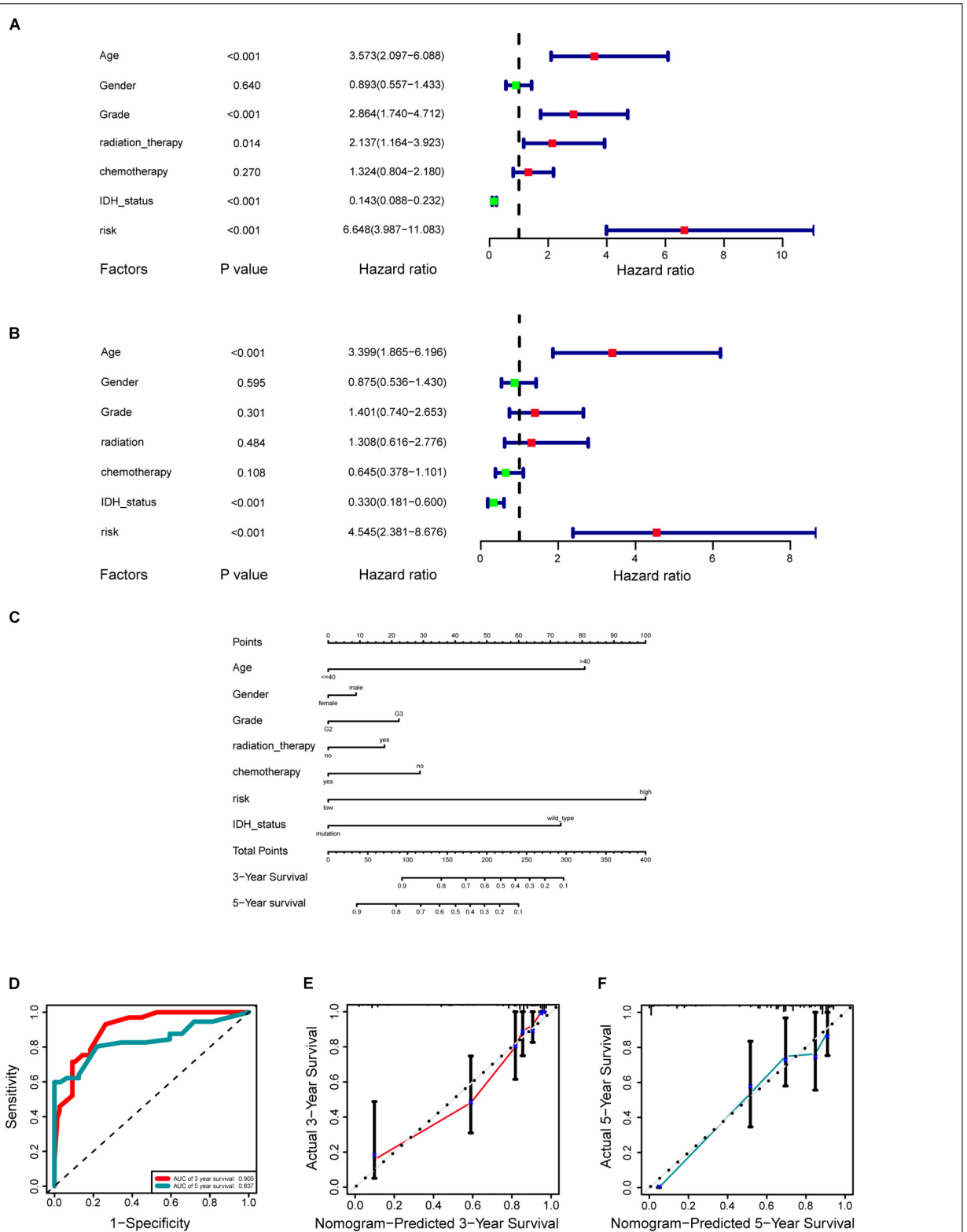


FIGURE 6 | Development of a prognostic stemness indices-related gene signature for primary low-grade glioma (LGG). **(A)** Univariable Cox regression analysis for the training cohort. **(B)** Multivariable Cox regression analysis for the training cohort. **(C)** A nomogram including risk score and other clinical features for predicting 3- and 5-years overall survival (OS) of primary LGG. **(D)** Time-dependent receiver operating characteristics (ROC) curve analysis for 3- and 5-years OS predictions for the nomogram compared with actual observations. Calibration plot of nomogram for predicting probabilities of 3- **(E)** and 5-year **(F)** OS of primary LGG patients in the TCGA database.

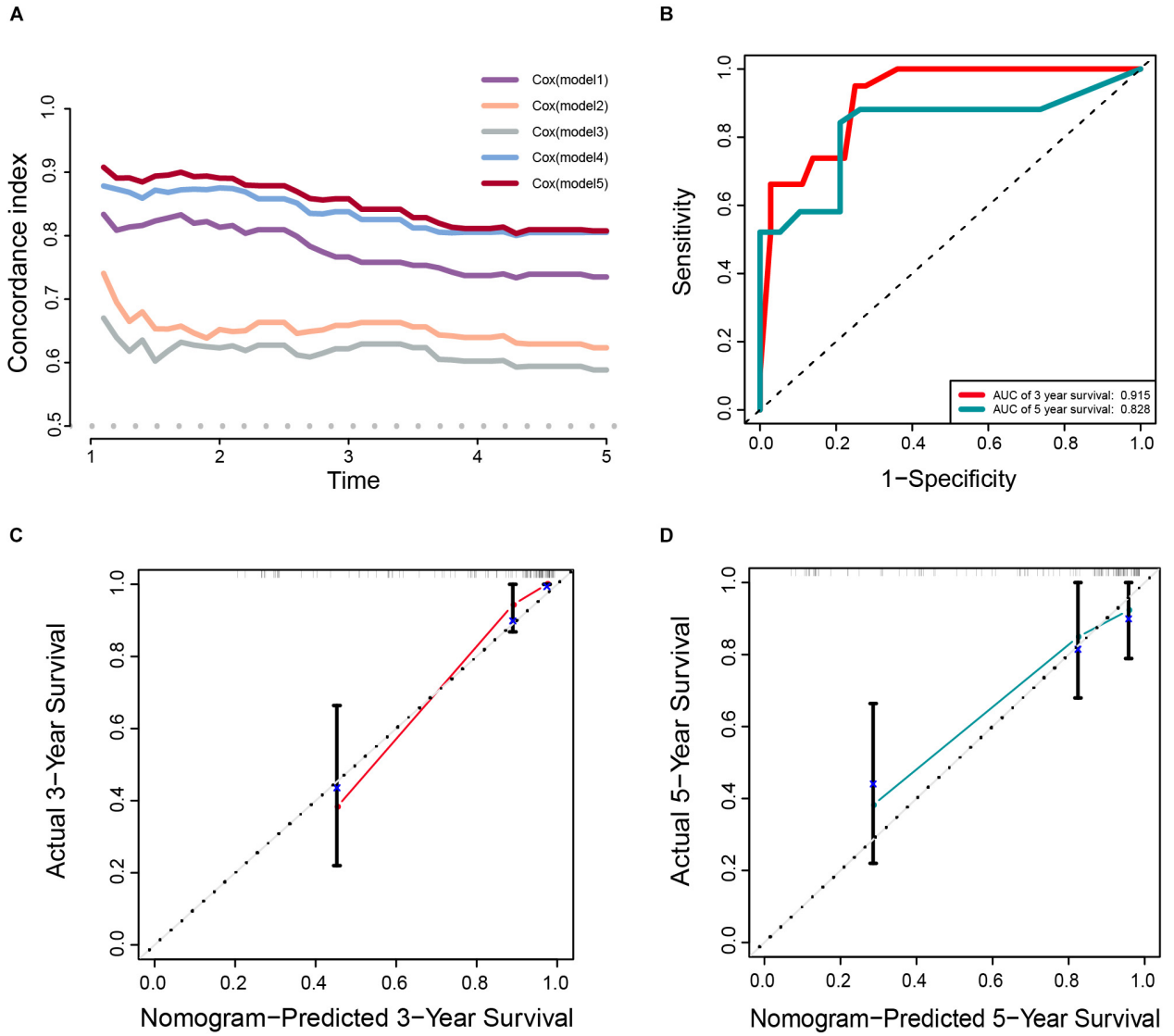


FIGURE 7 | (A) The calculation of the C-indexes in five models. **(B–D)** Internal validation of a prognostic stemness index-related gene signature for primary LGG. **(B)** Time-dependent ROC curve of the seven-gene-based risk score for 3- and 5-year OS probability in the internal validation cohort. Calibration plot for internal validation of 3- **(C)** and 5-year **(D)** OS of primary LGG patients.

TABLE 3 | Comprehensive comparison of the accuracy and discrimination in five models.

Index	Model 1 vs. Model 2	Model 1 vs. Model 3	Model 1 vs. Model 4	Model 1 vs. Model 5	Model4 vs. Model 5
IDI (1 year)	−0.037 (p = 0.274)	−0.063 (p = 0.02)	0.084 (p = 0.002)	0.108 (p < 0.001)	0.025 (p = 0.186)
Continuous NRI (1 year)	−0.598 (p = 0.010)	−0.663 (p < 0.001)	0.458 (p = 0.016)	0.708 (p < 0.001)	0.422 (p = 0.032)
IDI (3 year)	−0.146 (p = 0.040)	−0.178 (p = 0.006)	0.165 (p = 0.014)	0.214 (p < 0.001)	0.049 (p = 0.102)
Continuous NRI (3 year)	−0.548 (p = 0.022)	−0.508 (p < 0.001)	0.317 (p = 0.028)	0.433 (p < 0.001)	0.508 (p = 0.032)
IDI (5 year)	−0.189 (p = 0.044)	−0.211 (p = 0.056)	0.122 (p = 0.158)	0.177 (p = 0.018)	0.055 (p = 0.292)
Continuous NRI (5 year)	−0.530 (p = 0.078)	−0.366 (p = 0.058)	0.157 (p = 0.274)	0.410 (p = 0.036)	0.398 (p = 0.106)

Model 1: only the SI−risk signature was enrolled in the prognostic factor; Model 2: mRNAsi was enrolled in the prognostic factor; Model 3: corrected mRNAsi was enrolled in the prognostic factor; Model 4: age, gender, grade, radiation therapy, chemotherapy, and IDH status were enrolled in the prognostic factors; Model 5: age, gender, grade, radiation therapy, chemotherapy, IDH status, and risk group were enrolled in the prognostic factors. NRI, net reclassification improvement; IDI, integrated discrimination improvement.

To gain insights into the biological functions of key genes in WGCNA, it was found that the key genes were mainly enriched in infiltration, inflammation, and immune-related pathways, which were critically involved in the initiation and progression of glioma (Balkwill and Mantovani, 2001; Shacter and Weitzman, 2002; Mantovani et al., 2008; Michelson et al., 2016; Mostofa et al., 2017). Several studies have explored the prognostic value of host inflammatory cells, such as neutrophils in glioma. The role of neutrophils in glioma has two sides, mainly depending on the maturation and activation state. For example, the series of infiltrating neutrophils have the ability of contributing to glioma infiltration and pro-tumoral activity by secreting elastase (Iwatsuki et al., 2000). Circulating neutrophil-induced immunosuppression can promote tumor growth by secretion of arginase I (Sippel et al., 2011). On the other side, it has been found that the activation of neutrophils have an anti-tumor effect through antibody-dependent cellular cytotoxicity (Hafeman and Lucas, 1979; Fanger et al., 1989). Apart from making use of the migration of neutrophils, anti-cancer drugs can be delivered to the inflamed brain in glioma patients after surgery, which may reduce the recurrence of glioma (Xue et al., 2017). Moreover, recent evidence has revealed that the potential role of phagosomes in tumorigenesis via different mechanisms including its engagement in the autophagy pathway (Kim and Overholtzer, 2013).

Several studies have focused on the role that stemness features play in survival outcomes in human cancers. Similar to our study, using 763 primary medulloblastoma patients from the Gene Expression Omnibus (GEO) datasets, Lian and colleagues identified and validated a stemness-related gene expression signature to effectively stratify patients with Sonic hedgehog medulloblastoma into different OS groups (HR = 1.80, 95% confidence interval: 1.45–2.24, $P = 1.10E^{-07}$) (Lian et al., 2019). In terms of LGGs, age ($\leq$40 years vs. $>$ 40 years), tumor grade (II vs. III), and IDH status (wild-type vs. mutation) are well-established and widespread prognostic biomarkers in clinical practice (Ricard et al., 2012; Cancer Genome Atlas Research et al., 2015; Zeng et al., 2018; National Comprehensive Cancer Network, 2019).

In the present study, a seven-gene signature based on the mRNAsi was built to predict the prognosis of LGG. After the univariate and multivariate analysis, the stemness index-related gene signature, age, and IDH status were identified as independent prognostic markers for predicting OS in primary LGG patients. To our surprise, receiving radiation therapy and chemotherapy or not was not associated with OS in the multivariate analysis. The reason might be the undefined and inconsistency treatment protocols, including the duration of treatment, cycles of chemotherapy, total or fraction radiation dose, and combined treatment regimens. Moreover, numerous clinical trials have provided evidence for the adoption of chemotherapy and radiotherapy in gliomas and confirmed the OS benefit in adjuvant therapy. The Radiation therapy oncology group (RTOG) 9802 trial showed that radiotherapy combined with adjuvant procarbazine, 3 CCNU, and vincristine (PCV) chemotherapy substantially improves the median OS from 7.8 to 13.3 years (HR = 0.59; P = 0.002) in low-grade glioma patients older than 40 years or who did not undergo total tumor resection (van den Bent, 2014). Additionally, despite the six clinic-pathological factors comprised model performed fairly in predicting OS, however, integrating the risk signature further improve the c-index as well as the significant enhancements of 1y- and 3y-NRI. Thus, new prospective studies are necessary to further verify the prognostic value of the stemness index-associated risk signature in primary LGG patients who receive a combined standard approach of surgery, radiotherapy, and chemotherapy.

Among the seven genes, *ALOX5AP*, *APOBEC3C*, *GNG5*, and *SP100* were identified as risk-associated genes, whereas *ADAP2*, *FCGRT*, and *LRRC25* were confirmed as protective genes. Regarding the risk-associated genes, *APOBEC3C* was discovered as a vital member of the APOBEC family that encodes the APOBEC3C (apolipoprotein B mRNA editing enzyme catalytic subunit 3C, or A3C), clustered in the human chromosome 22 (Jarmuz et al., 2002). Some investigations have shown that the expression of *APOBEC3C* played a positive role in the invasiveness and prognosis of breast cancer (Zhang et al., 2015; Wang et al., 2019), hepatocellular carcinoma (Yang et al., 2015), and prostate cancer (Kawahara et al., 2019). Taking into account investigations on the role of GNG5 in carcinogenesis, Orchel et al. (2012) found that GNG5 may play a vital role in pathogenesis or progression of endometrial cancer. In addition, it has been revealed that GNG5 involved in PI3K-AKT and Wnt signaling pathway, and associated with reduced E-cadherin expression in invasive breast cancer (Alsaleem et al., 2019). *ALOX5AP* is one of the essential genes in the production of leukotrienes from arachidonic acid via encoding the ALOX5AP.Consistent with our results, Wu et al. (2018) found that high expression of *ALOX5AP* is associated with poor survival outcome in esophageal carcinoma. Additionally, ALOX5AP also involved in a risk model to serve as a prediction of osteosarcoma metastasis (Dong et al., 2019). It is known that the nuclear autoantigen SP100 participates in various biological processes, such as cellular gene expression, differentiation, and cell growth (Everett et al., 2006). It was found that high expression of *SP100* was associated with poor cell differentiation in laryngeal cancer (Li et al., 2010). Moreover, the expression of SP100 could regulate the transcriptional activity of ETS1 and further influence the cell invasion in breast cancer (Yordy et al., 2004). Regarding the role of *SP100* in glioma, previous study revealed that *SP100* was overexpressed in glioblastoma cells and involved in the regulation of glioblastoma cell proliferation and migration (Held-Feindt et al., 2011). With regards to the protective genes of *FCGRT*, it has been found to be responsible for encoding neonatal Fc receptor (FcRn), which participates in the transport and homeostasis of immunoglobulin as well as anti-tumor immunity (Roopenian and Akilesh, 2007; Ward and Ober, 2009). The expression of FcRn in immune cells, particularly in antigen presenting cells, is associated with its involvement in antigen presentation and cross-presentation that contributes to its shape anti-tumor properties. Studies showed that FcRn-expressed dendritic cells (DCs) are critical for the number and activation of CD8 + T-cells and are associated with prognosis in colorectal carcinoma (Baker et al.,

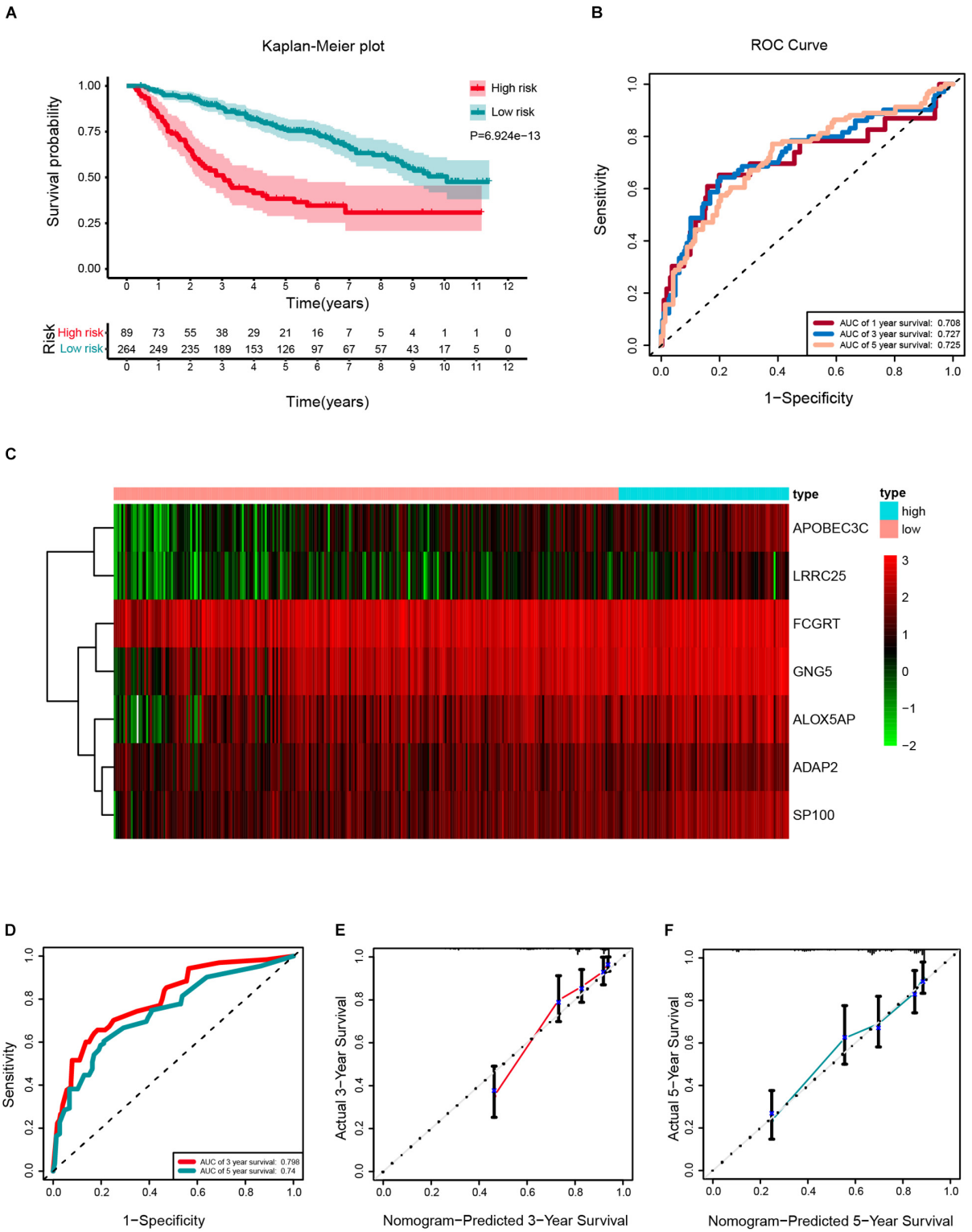


FIGURE 8 | External validation of a prognostic stemness index-related gene signature for primary low-grade glioma (LGG). **(A)** Kaplan-Meier analysis of OS for low-risk and high-risk patients in the external validation cohort. Additionally, the table indicating the number at risk for each group at corresponding time points. **(B)** The time-dependent receiver operating characteristics (ROC) curve for 1-, 3-, and 5-years OS predictions for the nomogram compared with actual observations. **(C)** The heatmap shows the expression of the seven genes between two risk groups in the CGGA cohort. **(D)** Time-dependent ROC for 3- and 5-years OS predictions for the nomogram compared with actual observations. Calibration plot of nomogram for predicting probabilities of 3- **(E)** and 5-year **(F)** OS of primary LGG patients in the CGGA database.

TABLE 4 | List the expression of the seven genes in 14 LGG cell lines.

Gene	*ADAP2*	*ALOX5AP*	*APOBEC3C*	*FCGRT*	*GNG5*	*LRRC25*	*SP100*	*ACTB*	RRID
Gene expression (TPM)									
H4	−3.234	−2.279	4.977	0.095	4.544	−4.032	3.775	10.664	CVCL_1239
HS683	−2.058	−0.673	4.300	2.101	6.059	−6.675	3.185	11.276	CVCL_0844
KG1C	−1.336	−1.609	6.160	0.334	5.218	−5.833	3.823	9.971	CVCL_2971
LN215	−3.879	−1.422	4.532	3.533	6.074	−7.049	4.243	10.537	CVCL_3954
LN235	−5.352	−1.489	4.216	3.967	6.028	−13.000	2.936	11.236	CVCL_3957
LN319	−4.063	−4.098	4.875	4.220	6.443	−13.000	1.443	9.752	CVCL_3958
LNZ308	−4.406	−3.171	2.872	0.317	7.113	−6.339	2.853	11.534	CVCL_0394
NMCG1	−3.648	−4.420	5.862	3.947	5.529	−8.266	2.392	11.488	CVCL_1608
SF268	−4.211	−1.512	0.554	3.739	6.236	−8.445	2.833	11.989	CVCL_1689
SNU738	−4.872	0.014	3.532	−3.016	6.417	−13.000	1.792	11.811	CVCL_5087
SW1088	−5.636	−2.916	5.441	1.140	6.274	−13.000	3.756	11.187	CVCL_1715
SW1783	−3.378	−1.839	4.975	2.773	5.975	−13.000	2.574	11.152	CVCL_1722
TM31	−2.523	−1.341	4.058	3.116	6.403	−8.164	1.793	10.397	CVCL_6735
U178	−3.724	−0.034	5.055	−3.648	5.367	−6.019	3.714	11.720	CVCL_A758

2013). The downregulation of FcRn is correlated with reducing maturation and activation of natural killer cells that in turn increase lung metastasis in an FcRn-depleted environment in mice (Castaneda et al., 2018). The protein encoded by ADAP2 is a GTPase-activating protein and increases the stability of microtubules. The investigation about the role of *ADAP2* in solid tumor is rare. Only one study found the expression of ADAP2 was markedly decreased in *in vivo* tumors without further validation about the function or mechanism (Laukkanen et al., 2015). Correspondingly, the prognostic value of *LRRC25* has not been investigated in solid tumors. Hoffman et al. found that the expression of *LRRC25* was significantly associated with the risk of breast cancer (Hoffman et al., 2017). Further investigations are warranted to explore the mechanisms of *LRRC25* in glioma.

Several limitations should be noticed in the current study. First, the stemness index-related signature and the nomogram developed were able to accurately predict survival outcome in primary LGG. Nonetheless, the validation in cellular experiments, and animal and tissue models warrants further investigation. Second, due to an absence of 1p19q characterization in the TCGA datasets, the status of 1p19q co-deletion was not investigated by the univariate and multivariate Cox regression analysis and was not employed for the establishment of prognostic nomogram. Third, considering a lack of standard treatment strategies in the TCGA and CGGA databases, the effectiveness of the seven-gene signature in primary LGG patients who received standard treatment needs to be further verified in well-designed prospective clinical investigations.

CONCLUSION

Our study identified a novel gene signature based on seven genes relevant to the stemness index and developed a prognostic nomogram composed of the gene signature and clinical prognostic factors that effectively predict overall survival in primary LGG patients. *ALOX5AP, APOBEC3C*, *GNG5*, *SP100, ADAP2*, *FCGRT*, and *LRRC25* might be candidate prognostic biomarkers in primary LGG.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The RNA-seq data (level 3) and clinical information of LGG samples can be found in the UCSC Xena browser (http://xena.ucsc.edu/), and the CGGA database (http://www.cgga.org.cn). The mRNA expression-based stemness index (mRNAsi) was provided by Tathiane M. Malta, et al.

ETHICS STATEMENT

All the information of patients was obtained from the Chinese Glioma Genome Atlas (CGGA), and The Cancer Genome Atlas (TCGA). All the patients and treatments complied with the principles laid down in the Declaration of Helsinki of 1964 and its later amendments or comparable ethical standards.

AUTHOR CONTRIBUTIONS

MZ analyzed the data. MZ, XC, and JH contributed materials and analysis tools. XW prepared figures and tables. MZ, XC, and XW authored or reviewed drafts of the manuscript. FG and JH conceived and designed the study. FG revised the manuscript.

FUNDING

This study was supported by the National Natural Science Foundation of China (11601083, U1805263), the program for Probability and Statistics: Theory and Application (IRTL1704), Innovative Research Team in Science and Technology in

Fujian Province University (IRTSTFJ), the Science Foundation for Young Scientists of Fujian Health and Family Planning Commission (Grant No. 2018-2-17), Qihang Foundation of Fujian Medical University (Grant No. 2018QH1088), and 2019 Fujian Provincial Science and Technology Department Gaoxiao Chanxueyuan Collaborative Project (Grant No. 2019Y4005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00441/full#supplementary-material

FIGURE S1 | The normalization and batch effect removal from TCGA and GTEx datasets. **(A)** Box plots illustrated the data distributions from TCGA and GTEx datasets before normalization. **(B)** PCA plot illustrated the cluster of the samples from TCGA and GTEx datasets before batch effect removal. **(C)** Box plots illustrated the data distributions from TCGA and GTEx datasets after normalization. **(D)** PCA plot illustrated the cluster of the samples from TCGA and GTEx datasets after batch effect removal.

FIGURE S2 | (A) Heatmaps showing that the 5,490 differentially expressed genes (DEGs) can effectively distinguish tumors from non-tumor tissues after integrated analysis. **(B)** Volcano plot presenting DEGs between LGG and non-tumor tissues. Red dots, and green dots represent up-regulated genes, and down-regulated genes, respectively.

FIGURE S3 | Weighted gene correlation network analysis for building stemness-index associated preserved Modules. **(A)** Determination of soft threshold for adjacency matrix, and plots of mean connectivity versus soft threshold. **(B)** Clustering results of WGCNA modules. The horizontal axis indicates modules with different colors.

FIGURE S4 | Analysis of key genes in the module brown. **(A)** The heatmap showing that the differentially expressed levels of the key genes between the normal control tissue and tumor tissue. **(B)** The heatmap of the correlation analysis among key genes.

FIGURE S5 | (A) Kaplan-Meier survival analysis of mRNAsi. **(B)** Kaplan-Meier survival analysis of corrected mRNAsi. Additionally, the table indicating the number at risk for each group at corresponding time points.

FIGURE S6 | Association between risk score and clinical-pathological parameters. Association between risk score and age, gender, grade, radiotherapy, chemotherapy, and IDH mutation status of primary LGG patients in TCGA cohort **(A)**, in CGGA cohort **(B)**.

FIGURE S7 | The mRNA expression level of *ADAP2* **(A)**, *ALOX5AP* **(B)**, *APOBEC3C* **(C)**, *FCGRT* **(D)**, *GNG5* **(E)**, *LRRC25* **(F)**, and *SP100* **(G)** in different types of human cancers.

FIGURE S8 | The protein expression level of immunohistochemistry (IHC) images collected from the Human Protein Atlas database of the risk genes between glioma tissue and normal control (ADAP2 was not available).

REFERENCES

Alsaleem, M., Toss, M. S., Joseph, C., Aleskandarany, M., Kurozumi, S., Alshankyty, I., et al. (2019). The molecular mechanisms underlying reduced E-cadherin expression in invasive ductal carcinoma of the breast: high throughput analysis of large cohorts. *Mod. Pathol.* 32, 967–976. doi: 10.1038/s41379-019-0209-9

Baker, K., Rath, T., Flak, M. B., Arthur, J. C., Chen, Z., Glickman, J. N., et al. (2013). Neonatal Fc receptor expression in dendritic cells mediates protective immunity against colorectal cancer. *Immunity* 39, 1095–1107. doi: 10.1016/j.immuni.2013.11.003

Balkwill, F., and Mantovani, A. (2001). Inflammation and cancer: back to Virchow? *Lancet* 357, 539–545. doi: 10.1016/S0140-6736(00)04046-0

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Cancer Genome Atlas Research, N., Brat, D. J., Verhaak, R. G., Aldape, K. D., Yung, W. K., Salama, S. R., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* 372, 2481–2498. doi: 10.1056/NEJMoa1402121

Castaneda, D. C., Dhommee, C., Baranek, T., Dalloneau, E., Lajoie, L., Valayer, A., et al. (2018). Lack of FcRn impairs natural killer cell development and functions in the tumor microenvironment. *Front. Immunol.* 9:2259. doi: 10.3389/fimmu.2018.02259

Ceccarelli, M., Barthel, F. P., Malta, T. M., Sabedot, T. S., Salama, S. R., Murray, B. A., et al. (2016). Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* 164, 550–563. doi: 10.1016/j.cell.2015.12.028

Claus, E. B., Walsh, K. M., Wiencke, J. K., Molinaro, A. M., Wiemels, J. L., Schildkraut, J. M., et al. (2015). Survival and low-grade glioma: the emergence of genetic information. *Neurosurg. Focus* 38:E6. doi: 10.3171/2014.10.FOCUS12367

Coons, S. W., Johnson, P. C., Scheithauer, B. W., Yates, A. J., and Pearl, D. K. (1997). Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas. *Cancer* 79, 1381–1393. doi: 10.1002/(sici)1097-0142(19970401)79:7(1381:aid-cncr16(3.0.co;2-w

Dirkse, A., Golebiewska, A., Buder, T., Nazarov, P. V., Muller, A., Poovathingal, S., et al. (2019). Stem cell-associated heterogeneity in Glioblastoma results from intrinsic tumor plasticity shaped by the microenvironment. *Nat. Commun.* 10:1787. doi: 10.1038/s41467-019-09853-z

Dong, S., Huo, H., Mao, Y., Li, X., and Dong, L. (2019). A risk score model for the prediction of osteosarcoma metastasis. *FEBS Open Biol.* 9, 519–526. doi: 10.1002/2211-5463.12592

Everett, R. D., Rechter, S., Papior, P., Tavalai, N., Stamminger, T., and Orr, A. (2006). PML contributes to a cellular mechanism of repression of herpes simplex virus type 1 infection that is inactivated by ICP0. *J. Virol.* 80, 7995–8005. doi: 10.1128/JVI.00734-06

Fanger, M. W., Shen, L., Graziano, R. F., and Guyre, P. M. (1989). Cytotoxicity mediated by human Fc receptors for IgG. *Immunol. Today* 10, 92–99. doi: 10.1016/0167-5699(89)90234-X

Gene Ontology Constorium (2015). Gene ontology consortium: going forward. *Nucleic Acids Res.* 43, D1049–D1056. doi: 10.1093/nar/gku1179

Hafeman, D. G., and Lucas, Z. J. (1979). Polymorphonuclear leukocyte-mediated, antibody-dependent, cellular cytotoxicity against tumor cells: dependence on oxygen and the respiratory burst. *J. Immunol.* 123, 55–62.

Harrell, F. E. Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models,evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* 15, 361–387. doi: 10.1002/(sici)1097-0258(19960229)15:4(361:aid-sim168(3.0.co

Held-Feindt, J., Hattermann, K., Knerlich-Lukoschus, F., Mehdorn, H. M., and Mentlein, R. (2011). SP100 reduces malignancy of human glioma cells. *Int. J. Oncol.* 38, 1023–1030. doi: 10.3892/ijo.2011.927

Hoffman, J. D., Graff, R. E., Emami, N. C., Tai, C. G., Passarelli, M. N., Hu, D., et al. (2017). Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. *PLoS Genet.* 13:e1006690. doi: 10.1371/journal.pgen.1006690

Iwatsuki, K., Kumara, E., Yoshimine, T., Nakagawa, H., Sato, M., and Hayakawa, T. (2000). Elastase expression by infiltrating neutrophils in gliomas. *Neurol. Res.* 22, 465–468. doi: 10.1080/01616412.2000.11740701

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., et al. (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* 79, 285–296. doi: 10.1006/geno.2002.6718

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kawahara, R., Recuero, S., Nogueira, F. C. S., Domont, G. B., Leite, K. R. M., Srougi, M., et al. (2019). Tissue proteome signatures associated with five grades of prostate cancer and benign prostatic hyperplasia. *Proteomics* 19:e1900174. doi: 10.1002/pmic.201900174

Kim, S. E., and Overholtzer, M. (2013). Autophagy proteins regulate cell engulfment mechanisms that participate in cancer. *Semin. Cancer Biol.* 23, 329–336. doi: 10.1016/j.semcancer.2013.05.004

Klein, J. P., and Moeschberger, M. L. (1997). *Survival Analysis: Techniques For Censored And Truncated Data. R package version 0.1-5*. Available online at: https://CRAN.R-project.org/package=KMsurv (accessed December 31, 2019).

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

Laukkanen, M. O., Cammarota, F., Esposito, T., Salvatore, M., and Castellone, M. D. (2015). extracellular superoxide dismutase regulates the expression of small GTPase regulatory proteins GEFs. GAPs, and GDI. *PLoS One* 10:e0121441. doi: 10.1371/journal.pone.0121441

Li, M., Spakowicz, D., Burkart, J., Patel, S., Husain, M., He, K., et al. (2019). Change in neutrophil to lymphocyte ratio during immunotherapy treatment is a non-linear predictor of patient outcomes in advanced cancers. *J. Cancer Res. Clin. Oncol.* 145, 2541–2546. doi: 10.1007/s00432-019-02982-4

Li, W., Shang, C., Guan, C., Zhang, Y., Sun, K., and Fu, W. (2010). Low expression of Sp100 in laryngeal cancer: correlation with cell differentiation. *Med. Sci. Monit.* 16, br174–br178.

Lian, H., Han, Y. P., Zhang, Y. C., Zhao, Y., Yan, S., Li, Q. F., et al. (2019). Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. *Mol. Oncol.* 13, 2227–2245. doi: 10.1002/1878-0261.12557

Liang, R., Zhi, Y., Zheng, G., Zhang, B., Zhu, H., and Wang, M. (2019). Analysis of long non-coding RNAs in glioblastoma for prognosis prediction using weighted gene co-expression network analysis. Cox regression, and L1-LASSO penalization. *Oncol. Targets Ther.* 12, 157–168. doi: 10.2147/OTT.S171957

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* 173:338-354.e15. doi: 10.1016/j.cell.2018.03.034

Mantovani, A., Allavena, P., Sica, A., and Balkwill, F. (2008). Cancer-related inflammation. *Nature* 454, 436–444. doi: 10.1038/nature07205

Michelson, N., Rincon-Torroella, J., Quinones-Hinojosa, A., and Greenfield, J. P. (2016). Exploring the role of inflammation in the malignant transformation of low-grade gliomas. *J. Neuroimmunol.* 297, 132–140. doi: 10.1016/j.jneuroim.2016.05.019

Mostofa, A. G., Punganuru, S. R., Madala, H. R., Al-Obaide, M., and Srivenugopal, K. S. (2017). The process and regulatory components of inflammation in brain oncogenesis. *Biomolecules* 7:34. doi: 10.3390/biom7020034

National Comprehensive Cancer Network (2019). *NCCN Clinical Practice Guidelines in Oncology: Central Nervous System Cancers*. Available online at: https://www.nccn.org/professionals/physician_gls/pdf/cns.pdf (accessed December 31, 2019)

Orchel, J., Witek, L., Kimsa, M., Strzalka-Mrozik, B., Kimsa, M., Olejek, A., et al. (2012). Expression patterns of kinin-dependent genes in endometrial cancer. *Int. J. Gynecol. Cancer* 22, 937–944. doi: 10.1097/igc.0b013e318259d8da

Ostrom, Q. T., Gittleman, H., Truitt, G., Boscia, A., Kruchko, C., and Barnholtz-Sloan, J. S. (2018). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2011-2015. *Neuro Oncol.* 20(Suppl. 4), 41–86. doi: 10.1093/neuonc/noy131

Pan, S., Zhan, Y., Chen, X., Wu, B., and Liu, B. (2019). Identification of biomarkers for controlling cancer stem cell characteristics in bladder cancer by network analysis of transcriptome data stemness indices. *Front. Oncol.* 9:613. doi: 10.3389/fonc.2019.00613

Pencina, M. J., D'Agostino, R. B., D'Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* 27, 157–172. doi: 10.1002/sim.2929

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing, 3.1.1*. Vienna: R Foundation for Statistical Computing.

Ricard, D., Idbaih, A., Ducray, F., Lahutte, M., Hoang-Xuan, K., and Delattre, J. Y. (2012). Primary brain tumours in adults. *Lancet* 379, 1984–1996. doi: 10.1016/S0140-6736(11)61346-9

Roopenian, D. C., and Akilesh, S. (2007). FcRn: the neonatal Fc receptor comes of age. *Nat. Rev. Immunol.* 7, 715–725. doi: 10.1038/nri2155

Roos, A., Ding, Z., Loftus, J. C., and Tran, N. L. (2017). Molecular and Microenvironmental Determinants of Glioma Stem-Like Cell Survival and Invasion. *Front. Oncol.* 7:120. doi: 10.3389/fonc.2017.00120

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241. doi: 10.1093/biomet/69.1.239

Shacter, E., and Weitzman, S. A. (2002). Chronic inflammation and cancer. *Oncology* 16, 217–226.

Sippel, T. R., White, J., Nag, K., Tsvankin, V., Klaassen, M., Kleinschmidt-DeMasters, B. K., et al. (2011). Neutrophil degranulation and immunosuppression in patients with GBM: restoration of cellular immune function by targeting arginase I. *Clin. Cancer Res.* 17, 6992–7002. doi: 10.1158/1078-0432.CCR-11-1107

Sokolov, A., Paull, E. O., and Stuart, J. M. (2016). One-class detection of cell states in tumor subtypes. *Pac. Symp. Biocomput.* 21, 405–416.

Therneau, G. T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–526. doi: 10.2307/2337123

Therneau, T. (2015). *A Package for Survival Analysis in S. version 2.38*. Available online at: https://cran.r-project.org/web/packages/survival/index.html (accessed December 31, 2019).

van den Bent, M. J. (2010). Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol.* 120, 297–304. doi: 10.1007/s00401-010-0725-7

van den Bent, M. J. (2014). Practice changing mature results of RTOG study 9802: another positive PCV trial makes adjuvant chemotherapy part of standard of care in low-grade glioma. *Neuro Oncol.* 16, 1570–1574. doi: 10.1093/neuonc/nou297

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., et al. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* 355:aai8478. doi: 10.1126/science.aai8478

Wang, K., Li, L., Fu, L., Yuan, Y., Dai, H., Zhu, T., et al. (2019). Integrated Bioinformatics Analysis the Function of RNA Binding Proteins (RBPs) and their prognostic value in breast cancer. *Front. Pharmacol.* 10:140. doi: 10.3389/fphar.2019.00140

Wanggou, S., Feng, C., Xie, Y., Ye, L., Wang, F., and Li, X. (2016). Sample level enrichment analysis of KEGG pathways identifies clinically relevant subtypes of glioblastoma. *J. Cancer* 7, 1701–1710. doi: 10.7150/jca.15486

Ward, E. S., and Ober, R. J. (2009). Chapter 4: Multitasking by exploitation of intracellular transport functions the many faces of FcRn. *Adv. Immunol.* 103, 77–115. doi: 10.1016/S0065-2776(09)03004-1

Wu, B., Bai, C., Du, Z., Zou, H., Wu, J., Xie, W., et al. (2018). The arachidonic acid metabolism protein-protein interaction network and its expression pattern in esophageal diseases. *Am. J. Transl. Res.* 10, 907–924.

Xue, J., Zhao, Z., Zhang, L., Xue, L., Shen, S., Wen, Y., et al. (2017). Neutrophil-mediated anticancer drug delivery for suppression of postoperative malignant glioma recurrence. *Nat. Nanotechnol.* 12, 692–700. doi: 10.1038/nnano.2017.54

Yang, Z., Lu, Y., Xu, Q., Zhuang, L., Tang, B., and Chen, X. (2015). Correlation of APOBEC3 in tumor tissues with clinico-pathological features and survival from hepatocellular carcinoma after curative hepatectomy. *Int. J. Clin. Exp. Med.* 8, 7762–7769.

Yi, Y., Hsieh, I. Y., Huang, X., Li, J., and Zhao, W. (2016). Glioblastoma stem-like cells: characteristics. microenvironment, and therapy. *Front. Pharmacol.* 7:477. doi: 10.3389/fphar.2016.00477

Yordy, J. S., Li, R., Sementchenko, V. I., Pei, H., Muise-Helmericks, R. C., and Watson, D. K. (2004). SP100 expression modulates ETS1 transcriptional activity and inhibits cell invasion. *Oncogene* 23, 6654–6665. doi: 10.1038/sj.onc.1207891

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

Zeng, W. J., Yang, Y. L., Liu, Z. Z., Wen, Z. P., Chen, Y. H., Hu, X. L., et al. (2018). Integrative analysis of DNA methylation and gene expression identify a three-gene signature for predicting prognosis in lower-grade gliomas. *Cell Physiol. Biochem.* 47, 428–439. doi: 10.1159/000489954

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

Zhang, Y., Delahanty, R., Guo, X., Zheng, W., and Long, J. (2015). Integrative genomic analysis reveals functional diversification of APOBEC gene family in breast cancer. *Hum. Genom.* 9:34. doi: 10.1186/s40246-015-0056-9

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

frontiers in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 08 May 2020
doi: 10.3389/fbioe.2020.00392

Exploring the Role of *SRC* in Extraocular Muscle Fibrosis of the Graves' Ophthalmopathy

*Mingyu Hao[1], Jingxue Sun[1], Yaguang Zhang[1], Dexin Zhang[1], Jun Han[2], Jirong Zhang[1] and Hong Qiao[1]**

[1] Endocrine and Metabolic Diseases, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, [2] Endocrine and Metabolic Diseases, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, China

The Graves' disease is an autoimmune disease highly associated with thyroid cancer. The Graves' ophthalmopathy (GO) is a special Graves' disease with inflammatory ophthalmopathy being a typical extrathymic complication. GO is caused by the formation of orbital fat and extraocular muscle fibrosis due to the inflammation of orbital connective tissues. Thus, controlling extraocular muscle fibrosis is critical for the prognosis of GO. The objective of this study is to identify and experimentally validate key genes associated with GO and explore their potential function mechanisms especially on extraocular muscle fibrosis. Specifically, we first created a GO mouse model, and performed RNA sequencing on the extraocular muscles of fibrotic GO mice and controls. *SRC* was identified as the most significant unstudied differentially expressed gene between GO mice and controls. Thus, we conducted a few *in vitro* analyses to explore the roles and functions of *SRC* in GO, for which we selected primary cultured orbital fibroblast (OF) as the *in vitro* cell line model. It is known that myofibroblast (MFB), which expresses α-SMA, is an important target cell in the process of fibrosis. Our experiment suggests that TGF-β can induce the transformation from OF to MFB, however, the transformation was inhibited by silencing the *SRC* gene in OF. In addition, we also inhibited TGF-β/Smad, NF-κB, and PI3K/Akt signaling pathways to analyze the interaction between these pathways and *SRC*. In conclusion, the silence of *SRC* in OF can inhibit the transformation from OF to MFB, which might be associated with the interaction between *SRC* and a few pathways such as TGF-β/Smad, NF-κB, and PI3K/Akt.

Keywords: *SRC* gene, extraocular muscle fibrosis, graves' ophthalmopathy (GO), orbital fibroblast (OF), myofibroblast (MFB), α-smooth muscle actin (α-SMA)

OPEN ACCESS

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

Reviewed by:
Guohua Huang,
Shaoyang University, China
Yongchao Dou,
Baylor College of Medicine, United States

****Correspondence:***
Hong Qiao
qiaoh0823@sina.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *10 December 2019*
Accepted: *07 April 2020*
Published: *08 May 2020*

Citation:
Hao M, Sun J, Zhang Y, Zhang D, Han J, Zhang J and Qiao H (2020) Exploring the Role of SRC in Extraocular Muscle Fibrosis of the Graves' Ophthalmopathy. Front. Bioeng. Biotechnol. 8:392. doi: 10.3389/fbioe.2020.00392

INTRODUCTION

The Graves' ophthalmopathy (GO), also called infiltrative exophthalmos, is one type of Graves' disease with great prevalence (Tsai et al., 2015). About 25–50% of the Graves' disease patients have varying degrees of GO (Jiskra, 2017). However, the pathogenesis of GO is still unclear. At present, many researches consider it as an autoimmune disease (Burch and Wartofsky, 1993). The symptoms in its early stage mainly include inflammation and edema, while that in the late stage is retrobulbar fibrosis (Heufelder, 1999). Fibrosis of extraocular muscles causes the loss of normal

contractile function of muscle tissue, which leads to the limitation of eyeball movement. Patients may suffer from diplopia, strabismus and even compression of optic nerve lead to blindness, which seriously affects their life quality. At present, there is no good treatment for GO and the medication usually cannot prevent the occurrence of advanced extraocular muscle fibrosis. Therefore, it is of great clinical importance to study the pathogenesis of extraocular fibrosis of GO and develop effective prevention and treatment strategies.

Previous studies have suggested that the thyrotropin receptor (TSHR) of orbital fibroblasts (OF), which can regulate thyroid function, plays a pivotal role in GO (Weetman, 2000). In addition to thyroid epithelial cells, TSHR can be detected in extraocular muscle tissue and fat tissue in orbit (Krieger et al., 2016), and the concentration of TSHR in extraocular muscle tissue of GO patients is significantly higher than that of healthy people (Gillespie et al., 2012). Thus, TSHR has been considered as important disease targets in GO (Iyer and Bahn, 2012). The acting mechanism of TSHR is related to various active factors in the process of orbital autoimmune response caused by thyroid orbital autoantigen, which may transform OF to myofibroblast (MFB), a type of cell expressing α-smooth muscle actin (α-SMA; Dik et al., 2016). A few previous studies suggest that the emergence of MFB is the key step in the process of fibrosis (Saika et al., 2016), and the continuous accumulation of MFB or the defect of apoptosis process will lead to the progressive development of fibrosis (Huang and Susztak, 2015).

As another important factor for transforming OF to MFB, transforming growth factor-β (TGF-β) also plays a critical role in the fibrotic diseases of various organs and tissues (Shen et al., 2015). In fact, TGF-β is recognized as the starting hub of the formation and development of fibrosis, which has been widely studied. For example, Steensel et al. (2009) found that the expression level of TGF-β1 mRNA in the orbital tissue of GO patients was twice that of normal people. In addition, TGF-β significantly promotes the proliferation and transformation OF into MFB (Heufelder and Bahn, 1994; Koumas et al., 2003), and regulates the expression of TSHR (Valyasevi, 2001).

At present, researches on GO mechanism are mainly focused on its immunological pathogenesis (Antonelli et al., 2014; Rapoport and McLachlan, 2014; Chen et al., 2015). Recent studies suggest that genes, oxidative stress and other factors may also affect the pathogenesis of GO (Chng et al., 2014; Wang et al., 2015). For example, many genes were abnormally expressed in GO (Chen et al., 2014; Pei et al., 2018), and research shows that gene polymorphism also affects the occurrence and development of GO (Hooshang et al., 2015; Yang et al., 2017). Studies on these aspects can provide a more comprehensive understanding of the pathogenesis of GO. However, a deep exploration on abnormally expressed genes, antioxidant stress, and their acting mechanisms on extraocular fibrosis is more or less ignored, especially at the late stage of the disease. In addition, though it is known that the genetic mechanism of translation and transcription of susceptibility genes in GO patients may cause the self-immune response to TSHR (Brand and Gough, 2010), the mechanism of extraocular fibrosis in the late stage of GO patients has not been clarified. Finally, although it has been found that some molecular mechanisms and signal pathways may be involved in the pathogenesis of GO (Wang, 2014; Tong et al., 2015; Xiao-Ling et al., 2017), the regulation and interaction of these molecular mechanisms need to be further studied, and the role of these molecular mechanisms and signal pathways in GO extraocular muscle fibrosis is unclear.

In this study, we established a GO mouse model by genetic immunization (Sajad et al., 2013), and selected OF as the cell model for *in vitro* study of GO extraocular muscle fibrosis (Lim et al., 2014). Specifically, we first used GO mouse extraocular muscle to screen out key genes for GO extraocular muscle fibrosis, among which *SRC* has not been studied according to literature mining. We then studied the role of *SRC* in GO extraocular muscle fibrosis using the *in vitro* cell model, and its interaction with a few signaling pathways including PI3K/Akt/NF-κB signaling pathway, TGF-β/Smad signaling pathway, and so on. This study provides a new direction for the mechanism of GO development. It also provides a new idea for finding effective intervention targets for the treatment of extraocular muscle fibrosis in GO, which might be of importance clinical significance.

MATERIALS AND METHODS

Mouse Model of Graves' Orbitopathy

BALB/c female mice (with age 8–10 weeks) were purchased from Animal Experimental Center of Harbin Medical University. Animals were housed under conventional conditions in cages with filter top lids and food made available *ad libitum*. For immunization, BALB/c mice were anesthetized for injection with 50 μL plasmid (1 mg/mL) into each biceps femoris (thigh) muscle. A single injection with the needle entered deep (3–4 mm) into the thigh muscle was performed, with slow release of plasmid into the muscle. Great care was taken to ensure reproducibility of the injection protocol in all immunizations. Injection and *in vivo* electroporation were performed four times at 3-week intervals using an ECM830 square wave electroporator with 7-mm caliper electrodes at 200 V/cm. Application of the current was in ten-20 ms square wave pulses at 1 Hz. All animals were maintained in Specific Pathogen Free and procedures were conducted under Harbin Medical University regulations of accepted standards of humane animal care.

Cloning and Preparation of Plasmid DNA

The multi-system expression plasmid pTriEx-1.1 Neo was used as a vector and purchased from the BioVector NTCC. Human TSHR A-subunit (amino acid residues 22–289) was amplified for cloning. TSHR A-subunit cDNA region was cloned into *Bam*HI and *Not*I restriction sites in pTriEx-1.1 Neo by amplification from pcDNA3.1-human TSHR plasmid using forward primer 5′-CGCGGATCCATGAGGCGATTTCGGAGG-3′. The cDNA was excised and subcloned into *Bam*HI and *Not*I-digested pTriEx-1.1 NeopTriEx-1.1 Neo vector. The plasmids, termed pTriEx-1.1 Neo-TSHR A-subunit was fully sequenced in strands. All plasmids were grown in *E. coli* XL-1 Blue cells in LB medium in cultures. Purified plasmid concentrations were measured using a

spectrophotometer, resuspended at 1 mg/mL in sterile water, and stored at −80°C. Single plasmid preparations were used for the entire set of injections for the group of animals.

Screening of Differentially Expressed Genes

Three groups of extraocular muscle tissues were entrusted to SeqHealth Tech Co., Ltd., Wuhan, China for RNA Sequencing. The cuff norm was used to quantify the expression levels for each gene normalized by reads per kb of RPKM reads (1). RPKM ≥ 0.5 was defined as a mapped gene. The mapped genes were then used to calculate the difference of RPKM values and the fold changes between GO mice samples and control group samples. A value of $p < 0.05$ and log FC > 1 were classified as a differentially expressed gene (DEG). Map all differentially expressed genes to the various terms of the Gene Ontology database[1], calculate the number of genes for each term, and then apply a hypergeometric test to find out the difference compared to the entire genome background. Significantly enriched GO entries in the expressed genes. Using Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway[2] saliency enrichment analysis, hypergeometric tests were used to find Pathway that was significantly enriched in differentially expressed genes compared to the entire genomic background. After making KO annotations for genes, statistics are based on the KEGG metabolic pathways they participate in. A value of $p < 0.01$ was defined as significant enrichment.

$$\mathrm{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads (millions)} \times \text{exon length (KB)}}$$

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

Cell Culture and Transfections

Primary culture was performed on OFs for BALB/c mice. The cells were cultured in Dulbecco's Modified Eagle's Medium (Corning, United States), supplemented with 10% fetal bovine serum (ExCell Bio, Uruguay), and maintained at 37°C in a humidified environment containing 5% CO_2. For transfection, the cells were plated into 6-well plates with 2.5×10^5cells/well. Once the cells were 70–90% confluent, SRC mimics or NC mimics were transfected into the OFs using Lipofectamine® 3000 reagent (Thermo Fisher Scientific, United States), according to the manufacturer's protocol. SRC mimics (5′-GCGGCUGCAGAUUGUCAAUTTAUUGACAAUCUGCAGCCGCTT-3′) and its scramble control (5′-ACGUGACACGUUCGGAGAATT-3′) were designed and chemically synthesized by Shanghai GenePharma Co. Ltd. For Inhibitor treatment, 24 h after seeding, the medium was removed and replaced with Oxymatrine (5 mg/mL, cat. no. HY-N0158), JSH-23 (50 μmol/l, cat. no. HY-13982) and PI3K-IN-1 (25 μmol/l, cat. no. HY-12068) or equal amounts of DMSO and incubated for 24h at 37°C. Inhibitors purchased from MedChemExpress, United States. Serum−starved for 24 h, and then treated with human recombinant TGF-β1 (PeproTech, United States). The concentration of TGF-β1 is 10 μg/L and treated for 24 h.

Immunofluorescent Staining

For Immunofluorescent (IF) staining, the cells grown on the slides were fixed with 4% paraformaldehyde for 30 min at 4°C, then blocked with 5% bovine serum album in for 1 h at room temperature and incubated with primary antibodies overnight at 4°C. The next day, the slides were washed with PBS, and incubated with goat anti-rabbit IgG H&L FICT secondary antibodies (1:1000; cat. no.ab6717; Abcam, United Kingdom) and DAPI (Beyotime Biotechnology, China) for 2 h at room temperature. Fluorescence microscopy images were obtained with a research fluorescence microscope equipped with a digital camera. The following primary antibodies were used: Anti-alpha smooth muscle antibody (1:200; cat. no. ab5694; Abcam, United Kingdom) and Anti-Collagen I antibody (1:500; cat. no. ab34710; Abcam, United Kingdom).

Real-Time Quantitative PCR

Total RNA was extracted from OFs with TRIzol® reagent (Thermo Fisher Scientific, United States). A total of 1 μg RNA was transcribed into cDNA using transcriptor first Strand cDNA Synthesis Kit (Roche Diagnostics GmbH, Germany) for mRNA according to the manufacturer's protocol. The expression levels of the genes were detected by qPCR. qPCR was performed using the SYBR Green (Roche Diagnostics GmbH, Germany) dye detection method. The thermocycling conditions were as follows: 95°C for 10 min; followed by 40 cycles of 95°C for 15 s; and 60°C for 60 s. The following primers were used: Acta2 forward, 5′-GACGCTGAAGTATCCGATAGAA-3′ and reverse, 5′-AATACCAGTTGTACGTCCAGAG-3′; Col1a1 forward, 5′-TGAACGTGGTGTACAAGGTC-3′ and reverse, 5′-CCATCTTTACCAGGAGAACCAT-3′; Src forward, 5′-CTATGTGGAGCGGATGAACTAT-3′ and reverse, 5′-ATTCGTTGTCTTCTATGAGCCG-3′; Pik3r1 forward, 5′-AAACAAAGCGGAGAACCTATTG-3′ and reverse, 5′-TAATGACGCAATGCTTGACTTC-3′; Nfkb1 forward, 5′-CAAAGACAAAGAGGAAGTGCAA-3′ and reverse, 5′-GATGGAATGTAATCCCACCGTA-3′; Smad3 forward,5′-ATTCCATTCCCGAGAACACTAA-3′ and reverse, 5′-TAGGTCCAAGTTATTGTGTGCT-3′. Primers were designed and chemically synthesized by Sangon Biotech Co. Ltd., Shanghai, China.

Western Blotting

Cells were lysed in RIPA buffer supplemented with complete Protease Inhibitor Cocktail tablets (Roche Diagnostics GmbH, Germany) and Phosphatase Inhibitor Cocktail tablets (Roche Diagnostics GmbH, Germany) for 30 min on ice. Protein lysates (30 μg) were subjected to 8% SDS-PAGE (Beyotime Biotechnology, China) and transferred to PVDF membrane. Subsequent to blocking with 5% non-fat milk in 0.05% TBS-Tween-20 (v/v) for 1 h at room temperature, the membranes were incubated with the appropriate primary antibodies overnight at 4°C. The secondary antibodies were horseradish peroxidase

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

[2]http://www.genome.jp/kegg/

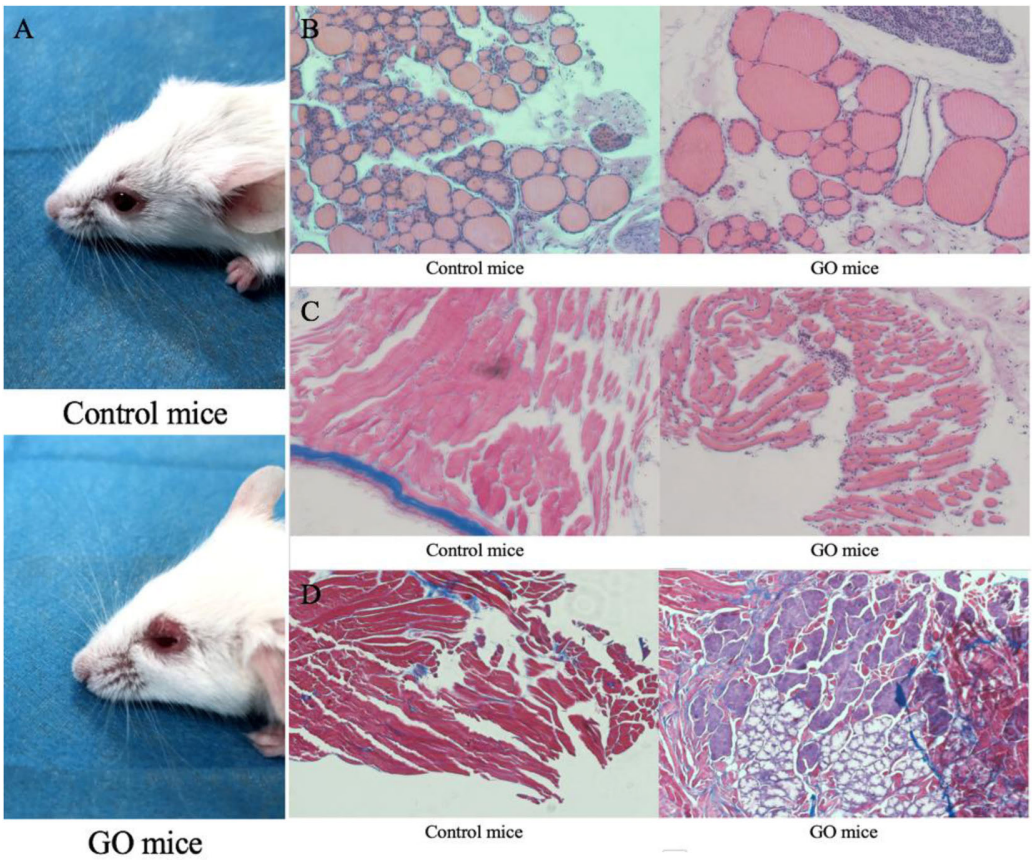


FIGURE 1 | GO mice model and Pathological images. **(A)** control mice and appearance of head region of hTSHR-A subunit plasmid-immunized mouse undergoing chemosis; **(B)** control mice normal thyroid and GO mice hypothyroid gland; **(C)** control mice H&E staining performed normal extraocular muscles and GO mice extraocular muscles showing interstitial inflammatory infiltrate; and **(D)** GO mice Masson's Trichrome-stained section of orbital muscle to show fibrosis in extraocular muscles and control mice did not show fibrosis.

(HRP)-conjugated goat anti-mouse IgG (cat. no. ZB-2305; 1:2500; ZSJQ-BIO, China) and HRP-conjugated goat anti-rabbit IgG (cat. no. ZB-2301; 1:2500; ZSJQ-BIO, China). The secondary antibodies were incubated for 1 h at room temperature. Protein detection was performed using an enhanced chemiluminescence substrate (Thermo Fisher Scientific, United States) prior to exposure to film. Primary antibodies used were as follows: Anti-alpha smooth muscle antibody (1:2500; cat. no. ab5694; Abcam, United Kingdom), Anti-Collagen I antibody (1:5000; cat. no. ab34710; Abcam, United Kingdom), TIMP-1 antibody (1:1000; cat. no. NB100-74551; Novus Biologicals, United States), Phospho-NF-κB p65 (ser536)(93H1) Rabbit mAb (1;1000; cat. no.#3033; Cell Signaling Technology, United States), NF-κB p65(D14E12)XP Rabbit mAb (1;1000; cat. no.#8242; Cell Signaling Technology, United States), Phospho-Smad2 (ser465/467)/Smad3(ser423/425)(D27F4) Rabbit mAb (1;1000; cat. no.#8828;Cell Signaling Technology, United States), Smad2/3(D7G7)XP Rabbit mAb (1;1000; cat. no.#8685;Cell Signaling Technology, United States), Anti-PI3 Kinase p85 alpha (phospho Y607) antibody (1:1000; cat. no. ab182651; Abcam, United Kingdom), PI3 Kinase p85 (19H8) Rabbit mAb (1;1000; cat. no.#8242; Cell Signaling Technology, United States).

Measurement of ROS Level in Cells

The cells (density: 2.5 × 10^5 cells/well.) were grown in a 6-well plate. After that, $DCFH_2$– (10 μM) and the culture were combined; they were subjected to incubation for 30 min at 37°C. Warm PBS was used to wash the cells, and generation of reactive oxygen species (ROS) was ascertained from intracellular 2′,7′-dichlorofluorescein (DCF) production that was the result of 2′,7′-dichlorodihydrofluorescein ($DCFH_2$) oxidation. A fluorescence enzyme-labeled instrument was employed to determine the level of DCF fluorescence in 488 nm excitation wavelength and 525 nm emission wavelength. ROS kit was purchased from Soleil Technology Co. Ltd.

RESULTS

HTSHR A-Subunit Plasmid-Immunized Mice

We initially challenged nine mice with hTSHR A-subunit plasmid by the Moshkelgosha protocol (Sajad et al., 2013). They were weighed weekly before the start of immunization. Animals

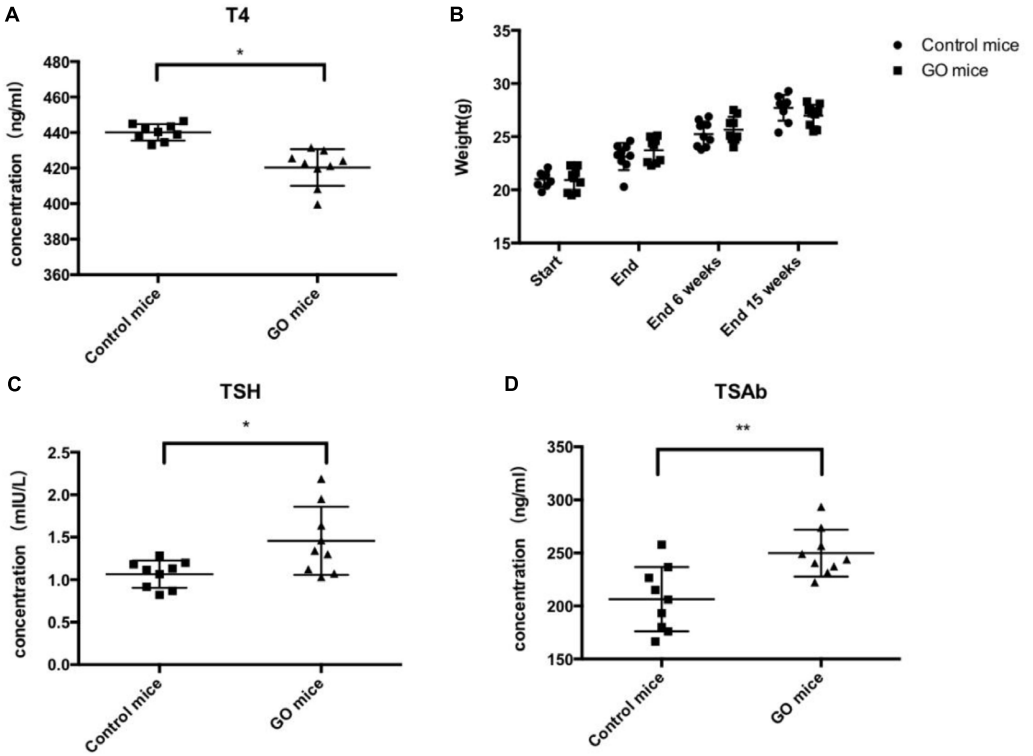


FIGURE 2 | Thyroid function and Antibody. **(A)** control mice and GO mice total serum T4 values; **(B)** control mice and GO mice killed at 6 and 15 weeks after end of immunization; **(C)** control mice and GO mice total serum TSH values; and **(D)** control mice and GO mice total serum TSAb values. *P value < 0.05 compared with control group; **P value < 0.01 compared with control group.

similarly immunized with pTRiEx1.1 neo ($n = 9$) did not lead to any visible changes in their health or to any histologic manifestations in thyroid and orbital tissue and were used as the control group. As a result, animals in the experiment group showed extraorbital changes with typical signs of acute orbital congestion (chemosis; **Figure 1A**). The finding of severe sickness prompted us to initially examine thyroid histology by hematoxylin–eosin (H&E) staining, which showed typical pattern of hypothyroidism, with most follicles characterized by thinning of epithelial cells (**Figure 1B**). In contrast, the H&E examination of thyroid glands of mice in the control group showed normal appearance. We next examined the H&E staining on orbital tissues. While the controlled mice showed normal appearance, the animals immunized with hTSHR A-subunit plasmid showed histologic signs of orbital pathology, and interstitial inflammatory infiltrate into extraocular muscle, which was extended into the muscle tissue and isolating individual fibers (**Figure 1C**). These symptoms are similar to those described in patients with active GO (Boschi et al., 2005). The hTSHR A-subunit plasmid-*in vivo* electroporation model in female BALB/c mice is recognized for robust antibody responses to TSHR, which persists for months after end of immunization (Zhao et al., 2011). The model therefore gave us the opportunity to evaluate the long-term effect of ongoing anti-TSHR immune response on orbital pathology. Finally, the H&E examination of orbital tissue was characterized predominantly by orbital muscle fibrosis, which by Masson's Trichrome staining exhibited extensive deposition of glycosaminoglycans with pericellular fibrosis in retrobulbar tissue (**Figure 1D**). Histologic analysis of the orbital tissue also showed disease heterogeneity in the experiment group with expansion of adipose tissue. None of the animals immunized with control plasmids showed any orbital pathology or disease.

We evaluated thyroid function in the above animals undergoing GO in serum obtained 15 weeks after the end of immunization. Total T4 measurements in mice undergoing experimental thyroid autoimmunity are commonly used for assessment of endocrine status during the course of disease (Gilbert et al., 2006). The animals showed a trend toward lower T4 values, correlating with the findings of hypothyroid glands by histology (**Figure 2A**). Importantly, the animals in the experiment group showed significant weight gain during the course of immunizations, conferring hypothyroid status (**Figure 2B**). In addition, the animals showed high levels of TSH (**Figure 2C**), and the determination of anti-TSHR antibody subtypes showed that the animals are highly positive for TSAbs (**Figure 2D**).

We performed immunohistochemical staining on the GO mice and control mice. Compared with controlled mice, we detected positive signals on TGF-β, α-SMA and Col-1, and immunohistochemical staining of extraocular muscles for GO mice (**Figure 3**).

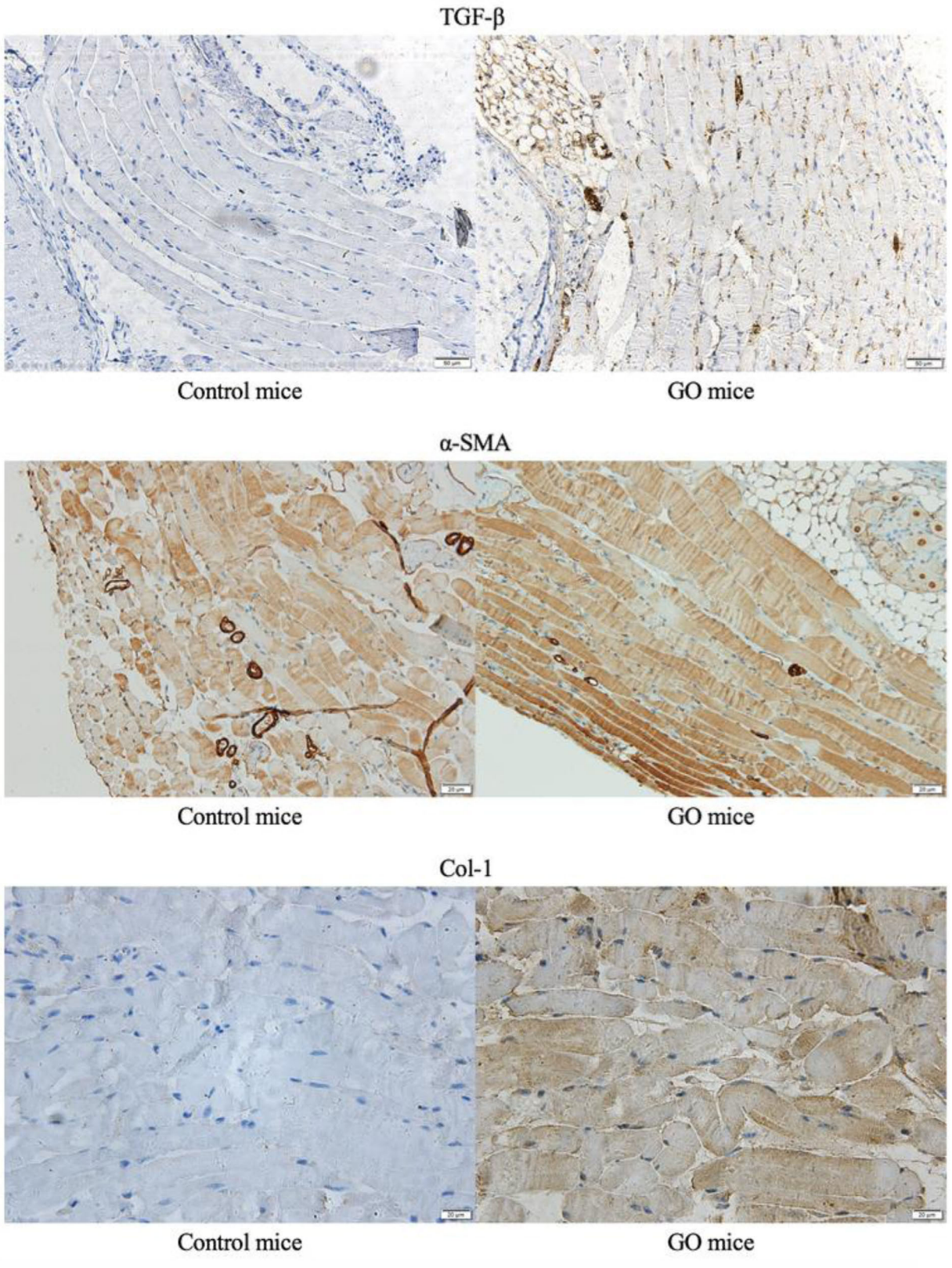


FIGURE 3 | Immunohistochemical staining. The immunohistochemical staining of extraocular muscles in GO mice and control mice.

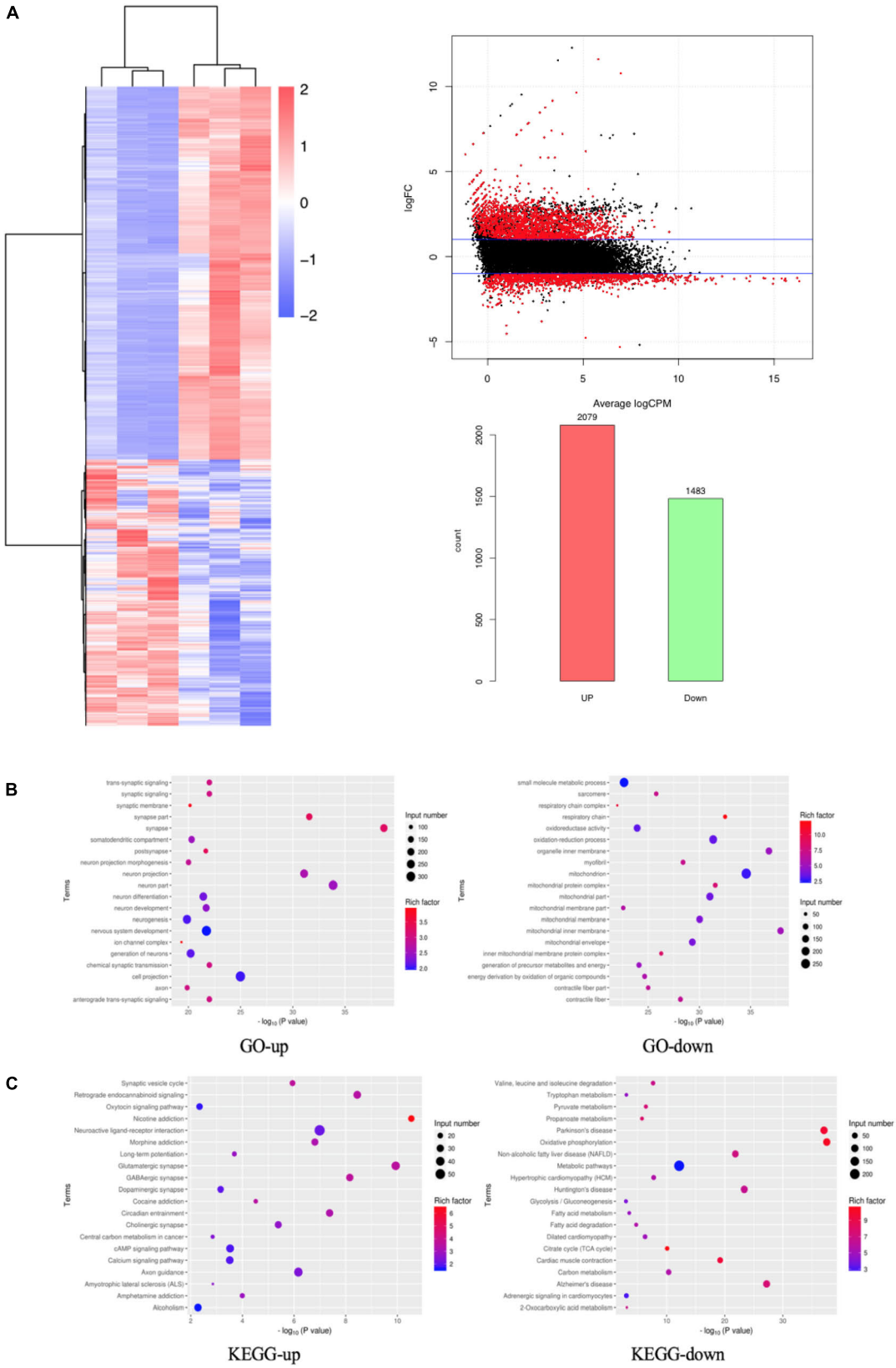


FIGURE 4 | DEGs based on RNA-Seq. **(A)** DEGs shown in a heat map, MA Plot of DEGs and Statistics of DEGs Up-Down; **(B)** GO enrichment map of up- and down-regulated genes; **(C)** KEGG enrichment map of up- and down-regulated genes.

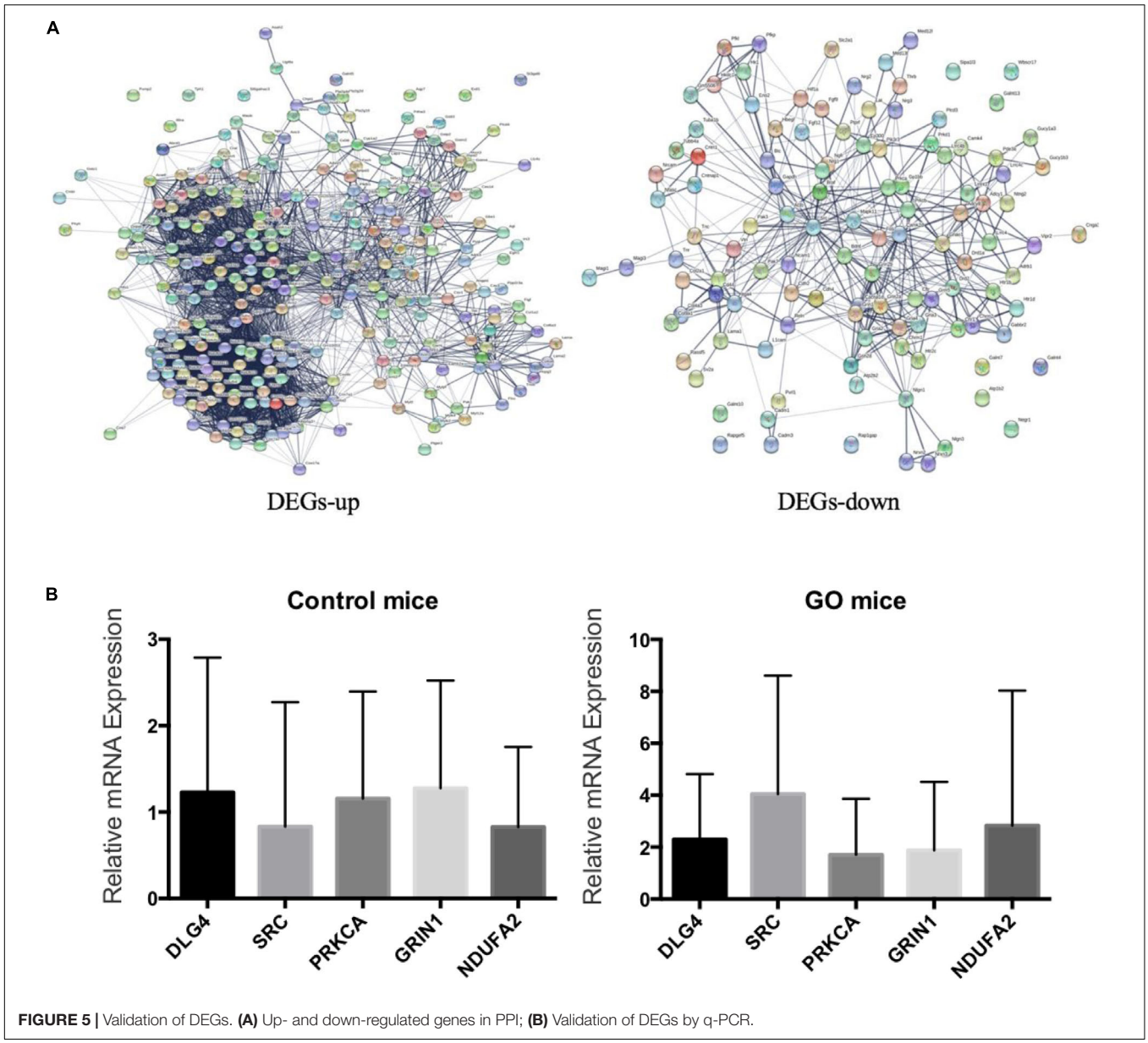


FIGURE 5 | Validation of DEGs. **(A)** Up- and down-regulated genes in PPI; **(B)** Validation of DEGs by q-PCR.

Screening of DEGs Based on RNA-Seq

We performed differentially expressed gene (DEG) analysis between GO mice and the controls. Specifically, the extraocular muscle tissues of three GO mice and three control mice were undergone RNA sequencing (RNA-Seq). The genes with transcript per million mapped (RPKM) values <0.5 were removed, and the remaining genes were kept for DEG analysis using DESeq2. A gene with adjusted $p < 0.05$ and $|\log FC| > 1$ was classified as a DEG. A DEG is up-regulated if its logFC > 1 and down-regulated if logFC < -1. Based on this definition, there were 2079 up-regulated and 1483 down-regulated genes. These 3562 DEGs were regarded as candidate genes for further study, and their expression levels were shown in a heat map in **Figure 4A**. We studied the functions of the DEGs by their enrichment on Gene Ontology terms and KEGG pathways using the hypergeometric test. A term/pathway of $p < 0.01$ was defined as significant enrichment. The significantly enriched Gene Ontology terms and KEGG pathways were shown in **Figures 4B,C**, respectively. As can be seen from **Figure 4B**, the top enriched Gene Ontology terms for up-regulated genes include synapse and neuron related function, while the top enriched Gene Ontology terms for down-regulated genes are mainly mitochondria-related functions. Similarly as shown in **Figure 4C**, the top enriched KEGG pathway for up-regulated genes is nicotine addition, while that for down-regulated genes is oxidative phosphorylation.

Validation of DEGs by q-PCR

We embedded the 3562 differentially expressed genes into the protein interaction network (STRING), and finally selected five

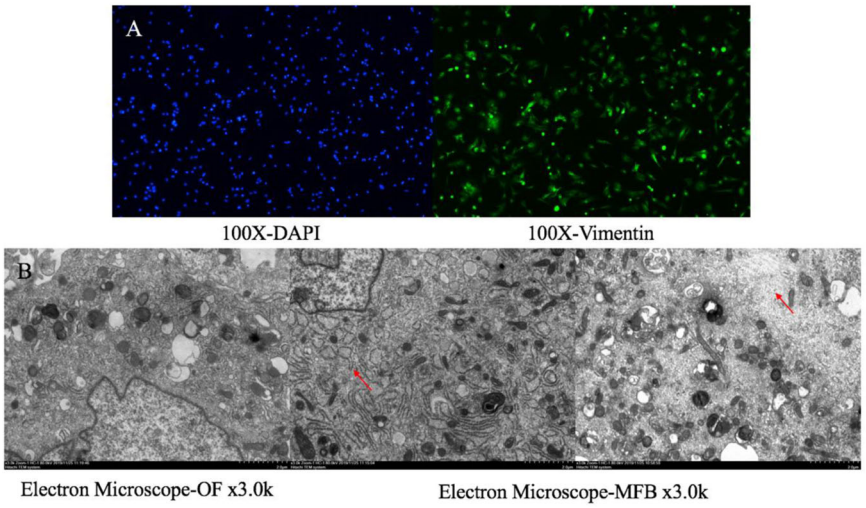


FIGURE 6 | Orbital fibroblasts and myofibroblast. **(A)** Primary culture orbital fibroblasts for BALB/c mice for IF staining; **(B)** OF and TGF-β to induce OF for Electron Microscope.

genes including *DLG4*, *SRC*, *PRKCA*, *GRIN1*, and *NDUFA2* with degree greater than 10 (**Figure 5A**). We then used q-PCR to validate the mRNA levels of five DEGs in extraocular muscle tissues of both GO mice and controls (**Figure 5B**). As can be seen, the expressions of four genes (*DLG4*, *SRC*, *PRKCA*, and *GRIN1*) were consistent between q-PCR and RNA-Seq, among which we could not find any GO-related research on *SRC* by PubMed search. Thus, we selected *SRC* as the candidate gene for further experiments.

TGF-β Induces OF and SRC Gene Silencing OF Transformation

The primary culture of OFs for BALB/c mice was shown in **Figure 6**. The serum was starved for 24 h, and then treated with human recombinant TGF-β1. We selected 5, 10, and 20 μg/L of TGF-β to induce OF cells for 6, 12, 24, and 48 h respectively, and tested the expression levels of α-SMA, Col-1and Timp-1. The expression was the highest when treated with 10 μg/L for 24 h (**Figure 7A**), so we chose 10 μg/L TGF-β1 treated for 24 h for further analysis. SRC mimics were transfected into the OFs using Lipofectamine® 3000 reagent, obtaining the OF with the *SRC* gene knockdown. The OF with *SRC* knockdown was further verified by Q-PCR. Furthermore, we can see that *Acta2* expression was also reduced in TGF-b-induced OF with *SRC* gene knockdown (**Figure 7B**). Western blot results also showed that the expression levels of α-SMA, Col-1, and Timp-1 were reduced in the *SRC* gene knockdown group compared with the control group, and TGF-β-induced OF transformation was inhibited (**Figure 7C**).

For inhibitor treatment, 24 h after seeding, the medium was removed and replaced with the TGF-β/Smad pathway inhibitor Oxymatrine (5 mg/mL), the NF-κB inhibitor JSH-23(50 μmol/L) and the PI3K inhibitor PI3K-IN-1(25μmol/L) to block the TGF-β/Smad, NF-κB, and PI3K/Akt signaling pathways, respectively. And an equal amount of DMSO was used as a control and incubated at 37°C for 24 h. Interestingly, we found that the *Acta2* expression was also reduced after TGF-β induction in OFs after inhibitor treatment (**Figure 8A**). The Western blot results also showed that compared with the control group, the expression of α-SMA, Col-1, and Timp-1 decreased, and TGF-β induced OF transformation was inhibited (**Figure 8C**). We further detected the expression of Smad3, Nfkb1, Pik3r1, and Akt1 in the *SRC* knockdown group by q-PCR (**Figure 8B**). The western blot results also suggested that *SRC* knockdown inhibited the phosphorylation of SMAD2/3, NF-κB p65, and PI3K p85 proteins (**Figure 8D**).

ROS Level in Cells

We tested the ROS generation in each group. $DCFH_2$–(10 μM) and the culture were combined, which were subjected to incubation for 30 min at 37°C. A fluorescence enzyme-labeled instrument was employed to determine the level of DCF fluorescence. As can be seen from **Figure 9**, the use of inhibitors and *SRC* gene knockdown can significantly inhibit ROS production during TGF-b-induced OF transformation.

DISCUSSION

We adopted the GO mouse model construction method proposed by Moshkelgosha, which is known to have a high GO mice formation rate of 75% and repeatable (Sajad et al., 2013). Fifteen weeks after the end of immunization, GO mice developed extraocular muscle fibrosis. We then killed the GO mice, obtained the extraocular muscle tissue and thyroid tissue of mice for pathological staining, and detected the serum T4, TSH, TSAb of the mice to evaluate the thyroid function. By comparing with the control group, we determined that the GO mice produced in this experiment had the changes of thyroid function and the

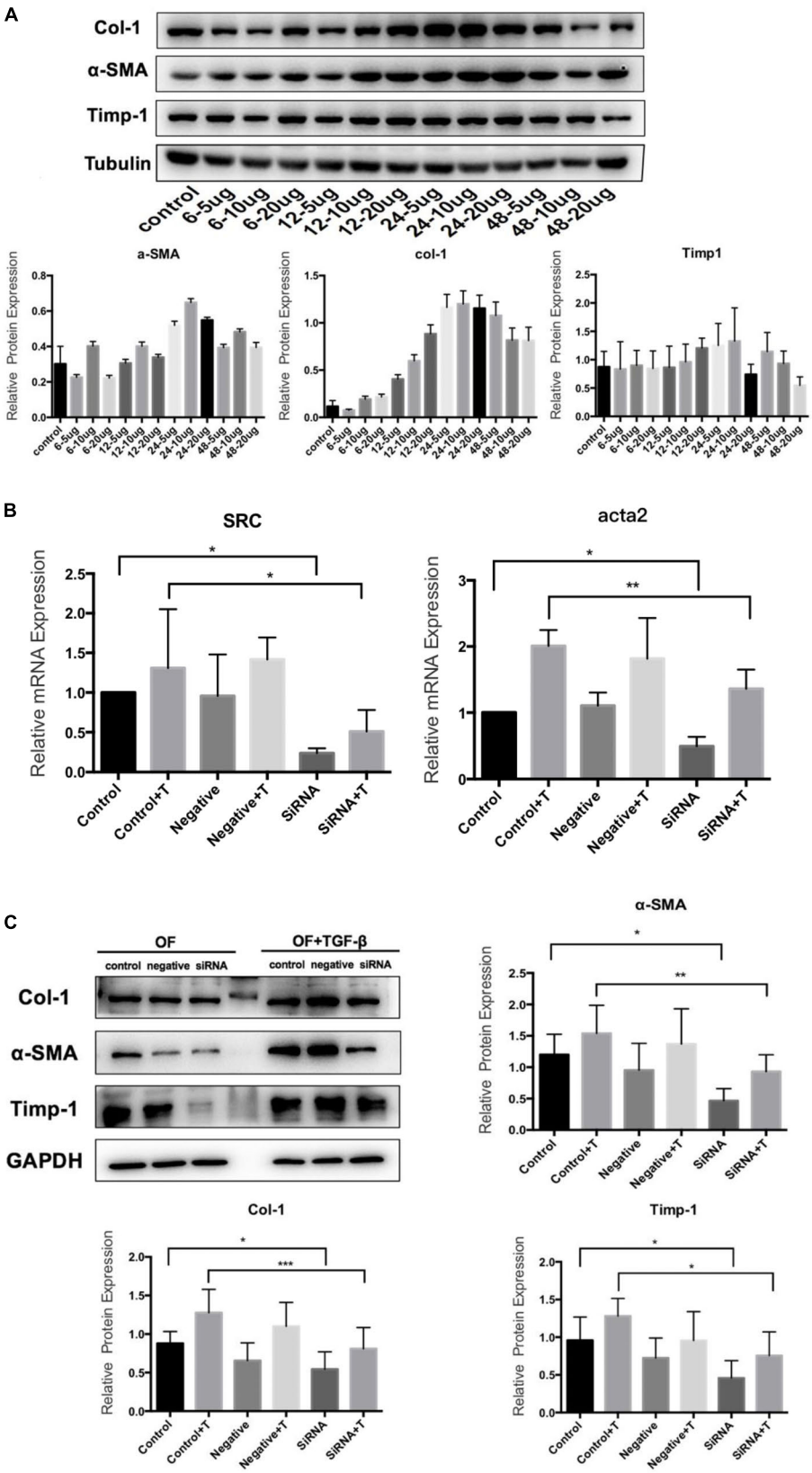


FIGURE 7 | TGF-β1 induces the transformation from OF to MFB. **(A)** Western blot for TGF-β1 concentration and time; **(B)** Src and Acta2 expression in SRC gene knockdown group by q-PCR; and **(C)** the expression levels of α-SMA, Col-1, and Timp-1 in the SRC gene knockdown group by Western blot.

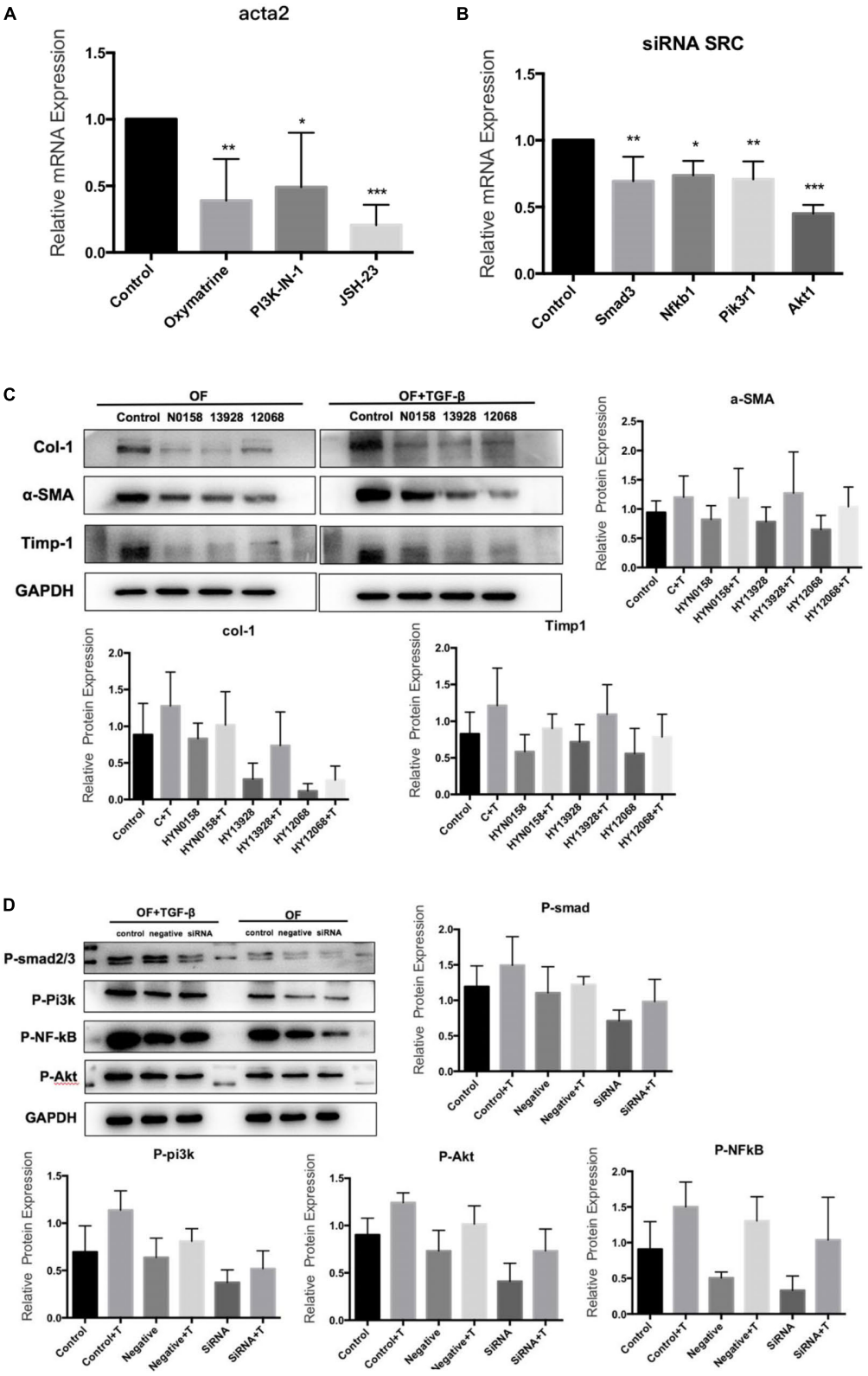


FIGURE 8 | Signaling inhibitors. **(A)** Acta2 expression after Signaling inhibitors by q-PCR; **(B)** the expression of Smad3, Nfkb1, Pik3r1, and Akt1 in the SRC gene knockdown group by q-PCR; **(C)** the expression of α-SMA, Col-1, and Timp-1 after Signaling inhibitors by western blot; **(D)** the phosphorylation of SMAD2/3, NF-κB p65 and PI3K p85 proteins in SRC gene knockdown group by western blot.

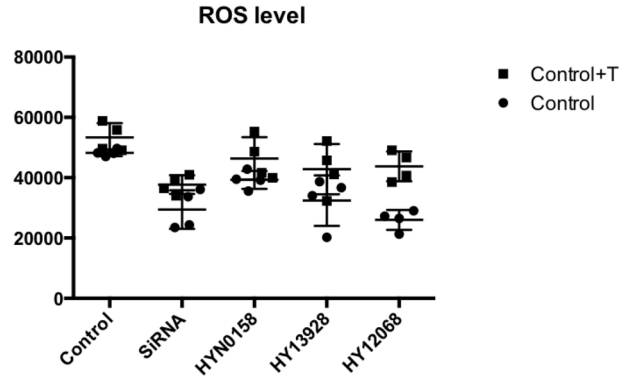


FIGURE 9 | Measurement of ROS level. The ROS generation of each group.

fibrosis of extraocular muscles. We identified the differentially expressed genes between the extraocular muscles of GO mice and those of the controls, among which *SRC* was selected as a candidate gene that may cause the extraocular muscle fibrosis. Steensel et al. (2009) showed that the expression of TGF-β1 mRNA in the orbital tissue of GO patients was twice that of normal people. In addition, by detecting the expression level of TGF-β1 in muscle, it was found that the expression level of TGF -β1 in mice with fibrosis was much higher than that in the control group. Thus, it was speculated that TGF-β1 could induce the differentiation of α-SMA protein into MFBs by inducing fibroblasts to express α-SMA, and then induces the occurrence of muscle fibrosis (Liu et al., 2016). Our results confirmed this speculation. Through immunohistochemical analysis, we can see that TGF-β1 expression was higher in extraocular muscle tissue of GO mice in controls. In addition, the expressions of α-SMA and Col-1 were also higher in GO mice than in controls, suggesting that the fibroblast in extraocular muscle of GO mice has begun to change in function and phenotype, and orbital tissue has a large number of extracellular matrix (ECM) accumulation. Moreover, the protrusion of the eyeball can be seen in all GO mice, and the Masson staining of the extraocular muscle can also confirm the fibrosis of the extraocular muscle tissue.

As the target and effector OF autoimmune response, OF can be cultured to serve as an *in vitro* GO model. The transformation from OF to MFB is a key step in the process of fibrosis (Saika et al., 2016), and the expression of α-SMA is a key marker for the transformation (Dik et al., 2016). In this study, we used TGF-β1 to induce the transformation from OF to MFB and to create a cell model of extraocular fibrosis for *in vitro* analyses. The expression of α-SMA was used to evaluate the fibrosis and the expression of COL-1 and TIMP-1 reflected the accumulation of ECM. Our results showed that the OF cell lines without *SRC* knockdown was transformed to MFB under the action of TGF-β1, which expresses a high level of α-SMA, and also produces a high level of ECM. In contrast, the transformation induced by TGF-β1was inhibited with *SRC* knockdown, as indicated by significant low expression of α-SMA, COL-1 and TIMP-1, and less accumulation of ECM. This indicates that SRC gene plays an important role in the fibrosis of go extraocular muscles.

It is worth noticing that we used an inhibitor Oxymatrine to inhibit TGF-β/Smad signaling pathways. In addition to restrain Smad2 and Smad3 phosphorylation, Oxymatrine also restrain the TGF-β1 induced transformation from OF to MFB, as indicated by the low express of α-SMA, Col-1, and TIMP-1. The results confirmed the role of TGF-β/Smad signaling pathway in the process of GO extraocular muscle fibrosis, consistent with the findings of Van Steensel L studies (Steensel et al., 2009). In addition, Smad2 and Smad3 phosphorylation were inhibited in OFs with *SRC* silencing, suggesting that *SRC* might be involved in the functioning of TGF-β/Smad signaling pathway in developing GO.

NF-κB is a kind of nuclear factor, which can promote direct or indirect activation of inflammatory factors, chemokines, inflammatory and TGF-β gene expression (Chen et al., 2012). Thus, it plays an important role in the development of extraocular muscle fibrosis. The results showed that JSH-23, an NF-κB inhibitor, could inhibit the TGF-β1 induced transformation from OF to MFB, which was indicated by relatively reduced expressions of α-SMA, COL-1 and TIMP-1. But the silencing of *SRC* can inhibit the phosphorylation of NF-κB, indicating that *SRC* could play a role in the process of GO extraocular muscle fibrosis by affecting the NF-κB signaling pathway. It is kown that PI3K/Akt signaling pathway plays a key role in TSH induced IL-1ra in GD (Li and Smith, 2014), and it is believed that PI3K/Akt pathway can increase the synthesis of HA by OFs (Xiao-Ling et al., 2017). Therefore, our study tried to determine the relationship between the *SRC* gene and the PI3K/Akt signaling pathway, which is a major upstream component of NF-κB. PI3K/Akt is thought to be involved in the pathogenesis of GO. By using the PI3K inhibitor PI3K-IN-1, we demonstrated that TGF-β1 induced transformation into MFB was also inhibited, with relatively reduced expressions of α-SMA, Col-1 and Timp-1. *SRC* silencing can also restrain the phosphorylation of PI3K and Akt, showing that *SRC* might affect the role of PI3K/Akt signal pathway in the process of external muscle fibrosis.

In addition, previous studies showed that the production of a large number of ROS can activate inflammatory signaling pathways like PI3K/Akt/NF-κB, and promote the expression of type I collagen fibers and TGF-β1, ultimately promoting the occurrence of fibrosis (Cohen-Naftaly and Friedman, 2011; Grochot-Przeczek et al., 2012). Interestingly, *SRC* can positively promote the production of ROS (Nakashima et al., 2017; Walter et al., 2017), and the rise of ROS induced by different stimuli can further promote the activity of *SRC* in cells (Kopetz et al., 2009). On the contrary, antioxidants can inhibit the activation of *SRC* activity by inhibiting ROS production (Fu et al., 2014). In our study, intracellular ROS production was detected, suggesting that the OF with *SRC* silencing produced fewer ROS during TGF-β1 induced transformation than the control group. The results further demonstrated the role of ROS in the process of GO extraocular muscle fibrosis.

Based on the studies, we believe that *SRC* is involved in the ROS mediated oxidative stress process, causing the activation of PI3K/Akt/NF-κB signaling pathway and leading to the

occurrence of GO extraocular muscle fibrosis. In addition, *SRC* also plays a role in the development of GO involved in TGF-β/Smad signaling pathway. The results provide a new direction for the study of mechanisms behind GO as well as a potential new intervention target for treating GO patients.

DATA AVAILABILITY STATEMENT

The data used in this study can be downloaded from https://submit.ncbi.nlm.nih.gov/subs/sra/SUB7185401.

ETHICS STATEMENT

The animal study was reviewed and approved by The Second Affiliated Hospital of Harbin Medical University.

AUTHOR CONTRIBUTIONS

HQ conceived the concept of the work. MH, JS, YZ, DZ, JH, and JZ performed the experiments. MH wrote the manuscript.

REFERENCES

Antonelli, A., Ferrari, S. M., Corrado, A., Franceschini, S. S., Gelmini, S., Ferrannini, E., et al. (2014). Extra-ocular muscle cells from patients with Graves' ophthalmopathy secrete α (CXCL10) and β (CCL2) chemokines under the influence of cytokines that are modulated by PPARγ. *Autoimmun. Rev.* 13, 1160–1166. doi: 10.1016/j.autrev.2014.08.025

Boschi, A., Daumerie, C., Spiritus, M., Beguin, C., and Many, M. C. (2005). Quantification of cells expressing the thyrotropin receptor in extraocular muscles in thyroid associated orbitopathy. *Br. J. Ophthalmol.* 89, 724–729. doi: 10.1136/bjo.2004.050807

Brand, O. J., and Gough, S. C. L. (2010). Genetics of thyroid autoimmunity and the role of the TSHR. *Mol. Cell. Endocrinol.* 322, 135–143. doi: 10.1016/j.mce.2010.01.013

Burch, H. B., and Wartofsky, L. (1993). Graves' ophthalmopathy: current concepts regarding pathogenesis and management. *Endocr. Rev.* 14, 747–793. doi: 10.1210/er.14.6.747

Chen, H. H., Yang, T., and Endocrinology, D. O. (2015). Research progresses in the pathogenesis of thyroid associated ophthalmopathy. *Chin. J. Pract. Inter. Med.* 7.

Chen, H. X., Shao, B. Z., Chen, X. C., Zhou, W. M., and Zhang, Y. (2014). Inhibition effect of B7-H1 gene-modified regulatory dendritic cells on thyroid-associated ophthalmopathy in mice. *Int. Eye Sci.* 14, 1765–1769.

Chen, X., Liu, C., Lu, Y., Yang, Z., and Lu, L. (2012). Paeoniflorin regulates macrophage activation in dimethylnitrosamine-induced liver fibrosis in rats. *BMC Complement. Alternat. Med.* 12:254. doi: 10.1186/1472-6882-12-254

Chng, C. L., Lai, O. F., Chew, S. M., Yu, P. P., Fook-Chong, M. C., Seah, L. L., et al. (2014). Hypoxia increases adipogenesis and affects adipocytokine production in orbital fibroblasts-a possible explanation of the link between smoking and Graves' ophthalmopathy. *Int. J. Ophthalmol.* 7:403. doi: 10.3980/j.issn.2222-3959.2014.03.03

Cohen-Naftaly, M., and Friedman, S. L. (2011). Current status of novel antifibrotic therapies in patients with chronic liver disease. *Ther. Adv. Gastroenterol.* 4, 391–417. doi: 10.1177/1756283X11413002

Dik, W. A., Virakul, S., and Van Steensel, L. (2016). Current perspectives on the role of orbital fibroblasts in the pathogenesis of Graves' ophthalmopathy. *Exp. Eye Res.* 142, 83–91. doi: 10.1016/j.exer.2015.02.007

Fu, Y., Yang, G., Zhu, F., Peng, C., Li, W., Li, H., et al. (2014). Antioxidants decrease the apoptotic effect of 5-Fu in colon cancer by regulating Src-dependent caspase-7 phosphorylation. *Cell Death Dis.* 5:e983. doi: 10.1038/cddis.2013.509

Gilbert, J. A., Gianoukakis, A. G., Salehi, S., Moorhead, J., Rao, P. V., Khan, M. Z., et al. (2006). Monoclonal pathogenic antibodies to the thyroid-stimulating hormone receptor in graves' disease with potent thyroid-stimulating activity but differential blocking activity activate multiple signaling pathways. *J. Immunol.* 176, 5084–5092. doi: 10.4049/jimmunol.176.8.5084

Gillespie, E. F., Papageorgiou, K. I., Fernando, R., Raychaudhuri, N., and Douglas, R. S. (2012). Increased expression of TSH receptor by fibrocytes in thyroid-associated Ophthalmopathy leads to Chemokine production. *J. Clin. Endocrinol. Metab.* 97, E740–E746. doi: 10.1210/jc.2011-2514

Grochot-Przeczek, A., Dulak, J., and Jozkowicz, A. (2012). Haem oxygenase-1: non-canonical roles in physiology and pathology. *Clin. Sci.* 122, 93–103. doi: 10.1042/CS20110147

Heufelder, A. (1999). Pathogenesis of Graves' ophthalmopathy. *Zeitschrift Arztliche Fortbildung Qualitatssicherung* 93(Suppl. 1), 35–39.

Heufelder, A. E., and Bahn, R. S. (1994). Modulation of Graves' orbital fibroblast proliferation by cytokines and glucocorticoid receptor agonists. *Invest. Ophthalmol. Vis. Sci.* 35, 120–127.

Hooshang, L., Daniele, C., Senarath, E., John, W., Leigh, D., Patrick, C., et al. (2015). Novel single-nucleotide polymorphisms in the calsequestrin-1 gene are associated with Graves' ophthalmopathy and Hashimoto’s thyroiditis. *Clin. Ophthalmol.* 9, 1731–1740. doi: 10.2147/OPTH.S87972

Huang, S., and Susztak, K. (2015). Epithelial Plasticity versus EMT in Kidney fibrosis. *Trends Mol. Med.* 22, 4–6. doi: 10.1016/j.molmed.2015.11.009

Iyer, S., and Bahn, R. (2012). Immunopathogenesis of Graves' ophthalmopathy: the role of the TSH receptor. *Best Pract. Res. Clin. Endocrinol. Metab.* 26, 281–289. doi: 10.1016/j.beem.2011.10.003

Jiskra, J. (2017). Endocrine orbitopathy: the present view of a clinical endocrinologist. *Vnitr Lek* 63, 690–696.

Kopetz, S., Lesslie, D. P., Dallas, N. A., Park, S. I., Johnson, M., Parikh, N. U., et al. (2009). Synergistic Activity of the Src Family Kinase Inhibitor Dasatinib and Oxaliplatin in Colon Carcinoma cells is mediated by oxidative stress. *Cancer Res.* 69, 3842–3849. doi: 10.1158/0008-5472.CAN-08-2246

Koumas, L., Smith, T. J., Feldon, S., Blumberg, N., and Phipps, R. P. (2003). Thy-1 expression in human fibroblast subsets defines Myofibroblastic or Lipofibroblastic Phenotypes. *Am. J. Pathol.* 163, 0–1300.

Krieger, C. C., Place, R. F., Bevilacqua, C., Marcussamuels, B., Abel, B. S., Skarulis, M. C., et al. (2016). TSH/IGF-1 receptor cross talk in Graves'. *Ophthal. Pathog.* 101:2340. doi: 10.1210/jc.2016-1315

Li, B., and Smith, T. J. (2014). PI3K/AKT pathway mediates induction of IL-1RA by TSH in Fibrocytes: modulation by PTEN. *J. Clin. Endocrinol. Metab.* 99, 3363–3372. doi: 10.1210/jc.2014-1257

Lim, H. S., Back, K. O., Kim, H. J., Choi, Y. H., Park, Y. M., and Kook, K. H. (2014). Hyaluronic acid induces COX-2 Expression via CD44 in orbital fibroblasts from patients with thyroid-associated ophthalmopathy. *Invest. Ophthalmol. Vis. Sci.* 55, 7441–7450. doi: 10.1167/iovs.14-14873

Liu, F., Tang, W., Chen, D., Li, M., Gao, Y., Zheng, H., et al. (2016). Expression of TGF-β1 and CTGF is associated with fibrosis of denervated Sternocleidomastoid muscles in mice. *Tohoku J. Exp. Med.* 238, 49–56. doi: 10.1620/tjem.238.49

Nakashima, K., Uekita, T., Yano, S., Kikuchi, J.-I., Nakanishi, R., Sakamoto, N., et al. (2017). Novel small molecule inhibiting CDCP1-PKCδ pathway reduces tumor metastasis and proliferation. *Cancer Sci.* 108, 1049–1057. doi: 10.1111/cas.13218

Pei, M., Ziyu, C., Lihong, J., Jinwei, C., and Ruili, W. (2018). PTX3: a potential biomarker in thyroid associated ophthalmopathy. *Biomed. Res. Int.* 2018:5961974. doi: 10.1155/2018/5961974

Rapoport, B., and McLachlan, S. M. (2014). Graves' hyperthyroidism is antibody-mediated but is predominantly a Th1-Type Cytokine disease.

J. Clin. Endocrinol. Metab. 99, 4060–4061. doi: 10.1210/jc.2014-3011

Saika, S., Yamanaka, O., Okada, Y., and Sumioka, T. (2016). Modulation of Smad signaling by non-TGFβ components in myofibroblast generation during wound healing in corneal stroma. *Exp. Eye Res.* 142, 40–48. doi: 10.1016/j.exer.2014.12.015

Sajad, M., Po-Wah, S., Neil, D., Salvador, D. C., and Paul, B. J. (2013). Cutting Edge: Retrobulbar Inflammation, Adipogenesis, and Acute Orbital Congestion in a Preclinical Female Mouse Model of Graves' Orbitopathy Induced by Thyrotropin Receptor Plasmid-in Vivo Electroporation. *Endocrinology* 154, 3008–3015. doi: 10.1210/en.2013-1576

Shen, M., Liu, X., Zhang, H., and Guo, S.-W. (2015). Transforming growth factor β1 signaling coincides with epithelial–mesenchymal transition and fibroblast-to-myofibroblast transdifferentiation in the development of adenomyosis in mice. *Hum. Reproduct.* 31, 355–369.

Steensel, L. V., Paridaens, D., Schrijver, B., Dingjan, G. M., and Dik, W. A. (2009). Imatinib Mesylate and AMN107 Inhibit PDGF-Signaling in Orbital Fibroblasts: a potential treatment for Graves'. Ophthalmopathy. *Invest. Ophthalmol. Vis. Sci.* 50, 3091–3098. doi: 10.1167/iovs.08-2443

Tong, B. D., Xiao, M. Y., Zeng, J. X., and Xiong, W. (2015). MiRNA-21 promotes fibrosis in orbital fibroblasts from thyroid-associated ophthalmopathy. *Mol. Vis.* 21, 324–334.

Tsai, C. C., Wu, S. B., Chang, P. C., and Wei, Y. H. (2015). Alteration of Connective Tissue Growth Factor (CTGF) Expression in Orbital Fibroblasts from Patients with Graves' Ophthalmopathy. *PLoS One* 10:e143514. doi: 10.1371/journal.pone.0143514

Valyasevi, R. W. (2001). Effect of Tumor Necrosis factor, interferon, and transforming growth factor on Adipogenesis and expression of Thyrotropin receptor in human orbital Preadipocyte Fibroblasts. *J. Clin. Endocrinol. Metab.* 86, 903–908. doi: 10.1210/jc.86.2.903

Walter, E., Vielmuth, F., Rotkopf, L., Sárdy, M., Horváth, O. N., Goebeler, M., et al. (2017). Different signaling patterns contribute to loss of keratinocyte cohesion dependent on autoantibody profile in pemphigus. *Sci. Rep.* 7:3579. doi: 10.1038/s41598-017-03697-7

Wang, R. (2014). Establishment and Regulation of the B lymphocytes and Orbital Fibroblasts Co-culture in thyroid associated ophthalmopathy. *Invest. Ophthalmol. Vis. Sci.* 55:1843.

Wang, W. Y., Zhou, W. M., Zhang, Y., and Chen, H. X. (2015). Inhibition effect of mouse orbital fibroblasts TLR4 gene silencing on the thyroid-associated ophthalmopathy. *Int. J. Ophthalmol.* 15, 1862–1866.

Weetman, P. A. (2000). Graves' Disease. *N. Engl. J. Med.* 343, 1236–1248.

Xiao-Ling, C., Wei-Min, H., Wei, L., Ophthalmology, D. O., Hospital, W. C., and University, S. (2017). Effects and its mechanism of IGF-1R on the Synthesis of Hyaluronic acid in Orbital Fibroblasts of thyroid associated ophthalmopathy. *J. Sichuan Univ.* 48, 727–731.

Yang, H. W., Wang, Y. X., Bao, J., Wang, S. H., and Sun, Z. L. (2017). Correlation of HLA-D Q and TNF-α gene polymorphisms with ocular myasthenia gravis combined with thyroid associated ophthalmopathy. *Biosci. Rep.* 37:BSR20160440.

Zhao, S.-X., Tsui, S., Cheung, A., Douglas, R. S., Smith, T. J., and Banga, J. P. (2011). Orbital fibrosis in a mouse model of Graves' disease induced by genetic immunization of thyrotropin receptor cDNA. *J. Endocrinol.* 210, 369–377. doi: 10.1530/JOE-11-0162

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 25 May 2020
doi: 10.3389/fbioe.2020.00348

Analysis of Gene Signatures of Tumor Microenvironment Yields Insight Into Mechanisms of Resistance to Immunotherapy

*Ben Wang[1], Mengmeng Liu[2], Zhujie Ran[3], Xin Li[4], Jie Li[5] and Yunsheng Ou[1]**

[1] Department of Orthopedics, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [2] Graduated School of Anhui University of Traditional Chinese Medicine, Hefei, China, [3] School of Public Health and Community Medicine, Chongqing Medical University, Chongqing, China, [4] Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [5] Department of Oncology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Hong Zheng,
Stanford University, United States
Lu Jiang,
Johns Hopkins University, United States
Weiyu Chen,
Stanford University, United States
Bogang Wu,
George Washington University, United States

***Correspondence:**
Yunsheng Ou
ouyunsheng2020@163.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology

Received: *06 March 2020*
Accepted: *30 March 2020*
Published: *25 May 2020*

Citation:
Wang B, Liu M, Ran Z, Li X, Li J and Ou Y (2020) Analysis of Gene Signatures of Tumor Microenvironment Yields Insight Into Mechanisms of Resistance to Immunotherapy. Front. Bioeng. Biotechnol. 8:348. doi: 10.3389/fbioe.2020.00348

Background: The recent clinical success of immunotherapy represents a turning point in cancer management. But the response rate of immunotherapy is still limited. The inflamed tumor microenvironment has been reported to correlate with response in tumor patients. However, due to the lack of appropriate experimental methods, the reason why the immunotherapeutic resistance still existed on the inflamed tumor microenvironment remains unclear.

Materials and Methods: Here, based on single-cell RNA sequencing, we classified the tumor microenvironment into inflamed immunotherapeutic responsive and inflamed non-responsive. Then, phenotype-specific genes were identified to show mechanistic differences between distant microenvironment phenotypes. Finally, we screened for some potential drugs that can convert an unfavorable microenvironment phenotype to a favorable one to aid current immunotherapy.

Results: Multiple signaling pathways were phenotypes-specific dysregulated. Compared to non-inflamed microenvironment, the expression of interleukin signaling pathways-associated genes was upregulated in inflamed microenvironment. Compared to inflamed responsive microenvironment, the PPAR signaling pathway-related genes and multiple epigenetic pathways-related genes were, respectively, suppressed and upregulated in the inflamed non-responsive microenvironment, suggesting a potential mechanism of immunotherapeutic resistance. Interestingly, some of the identified phenotype-specific gene signatures have shown their potential to enhance the efficacy of current immunotherapy.

Conclusion: These results may contribute to the mechanistic understanding of immunotherapeutic resistance and guide rational therapeutic combinations of distant targeted chemotherapy agents with immunotherapy.

Keywords: immunotherapy, tumor microenvironment, immunotherapeutic resistance, molecular targeted agents, personalized medicine

INTRODUCTION

Although immunotherapy has revolutionized tumor treatment, it still has some limitations (Larkin et al., 2015). For example, the success of adoptive cell therapy (ACT) on hematological malignancies cannot be reproduced on solid tumors (Newick et al., 2017). The responsive rate of immune checkpoint inhibitors (CPIs) varies by tumor type, from 45% for melanoma (Daud et al., 2016; Ribas et al., 2016) to only 12.2% for head neck squamous cancer (HNSC) (Abril-Rodriguez and Ribas, 2017; Darvin et al., 2018).

To better understand the reasons for these limitations, a number of studies tried to investigate the effect of tumor microenvironment (TME) phenotype on immunotherapy and suggested that TME phenotype (broadly categorized as being inflamed or non-inflamed) (Binnewies et al., 2018; Galon and Bruni, 2019) was a critical factor responsible for these limitations (Ji et al., 2012; Peng et al., 2015; Spranger et al., 2015; Chen et al., 2016; Kortlever et al., 2017). However, for the lack of appropriate experimental methods, a systematic understanding of how inflamed TME forms and why therapeutic resistance still exists on inflamed TME has been constrained. Here, to better understand the role of TME phenotypes to aid current immunotherapy, we systematically analyzed pan-cancer molecular characteristics of inflamed TME and further delved into the mechanistic differences between inflamed responsive TME and inflamed non-responsive TME. Importantly, part of our results has been supported in recent reports (Chowdhury et al., 2018; Wang J. et al., 2019).

Together, these results have profound prospects in clinical application, including identifying multiple potential immunotherapeutic targets, providing mechanistic insights into immunotherapeutic resistance in inflamed TME, and screening for some potential immunophenotypic regulation drugs to guide rational combination of chemotherapy agents with immunotherapy.

METHODS

Pan-Cancer Samples and Clinical Cohorts Treated by Immunotherapy

RNA sequencing data across 19 The Cancer Genome Atlas (TCGA) tumor types were downloaded from the Gene Expression Omnibus (GEO) database with accession number GSE62944 (Rahman et al., 2015). The updated clinical data were downloaded from *TCGAbiolinks* (Colaprico et al., 2016; Silva et al., 2016; Mounir et al., 2019). Published RNA sequencing data (Riaz et al., 2017) of 101 clinical tumor samples treated by anti-CTLA4 and anti-PD1 were downloaded from the GEO database with accession number GSE91061. The raw count data of RNA sequencing were normalized and quantitated by the edgeR package (Robinson et al., 2010).

Identifying Immune Cell Signature From Integrated Single-Cell RNA Sequencing Data

In order to analyze the TME of different tumor types and increase the diversity of non-immune cell to obtain robust immune cell markers, we applied the Seurat integration pipeline (Butler et al., 2018) to integrate two single-cell RNA sequencing data sets, respectively, from the Puram's HNSC cohort (GEO accession number: GSE103322) (Puram et al., 2017) and Tirosh's melanoma cohort (GSE72056) (Tirosh et al., 2016). A CCA algorithm (Butler et al., 2018) derived from machine learning was used to identify anchors of cells from different tumor types for the purpose of unbiased single-cell data integration (Stuart et al., 2019). Annotations of immune cells referred to the original literature and cell marker database (Tirosh et al., 2016; Puram et al., 2017; Zhang et al., 2019). Immune cell gene signatures (GSs) were defined based on the following criteria: (1) the proportion of signature expression in immune cells (CD8 T cell, CD4 T cell, B cells, macrophage, mast cell, dendritic cell, NK cell) should be >0.6; (2) the percent of GS expression in non-immune cells (myocytes, tumor cells, endothelial, fibroblast) should be <0.3; (3) adjusted $P < 0.001$; (4) log (fold change)>0 (compared to non-immune cells and other immune cell clusters).

Unsupervised Clustering Algorithm to Determine TME Subtypes of Tumor Samples

Immune cell markers identified in single-cell RNA sequencing analysis were used as an input for the gene set variation analysis (GSVA) algorithm (Hänzelmann et al., 2013) to calculate the immune score for each immune cell. Then, tumor samples were classified into high-immune score (inflamed), intermediate immune score, and low-immune score (non-inflamed) based on the unsupervised clustering pattern. This method has been proven as an efficient way to indirectly evaluate the phenotypes of TME (Wang et al., 2018). By using optCluster (Sekula et al., 2017) to evaluate the internal and stability indexes of the seven clustering algorithms (clara, diana, hierarchical, kmeans, model, pam, and sota), the optimal number and the algorithm of clustering were determined. Finally, the Clara algorithm and three groups were selected as the most robust clustering parameters. To avoid the unfavorable bias of confounding factors, we excluded intermediate immune score samples in further analysis.

Identification of Altered Signaling Pathways

Differentially expressed genes (DEGs) were identified by edgeR package (Robinson et al., 2010) with a negative binomial distribution algorithm; $P < 0.05$ and an absolute value of log2-fold change >1.5 were considered as statistically significant. Then, we annotated these DEGs with ClusterProfile (Yu et al., 2012) and RectomePA (Yu and He, 2016) package according to KEGG and Rectome pathway databases. Gene set enrichment analysis (GSEA) was used to provide a systematic view into

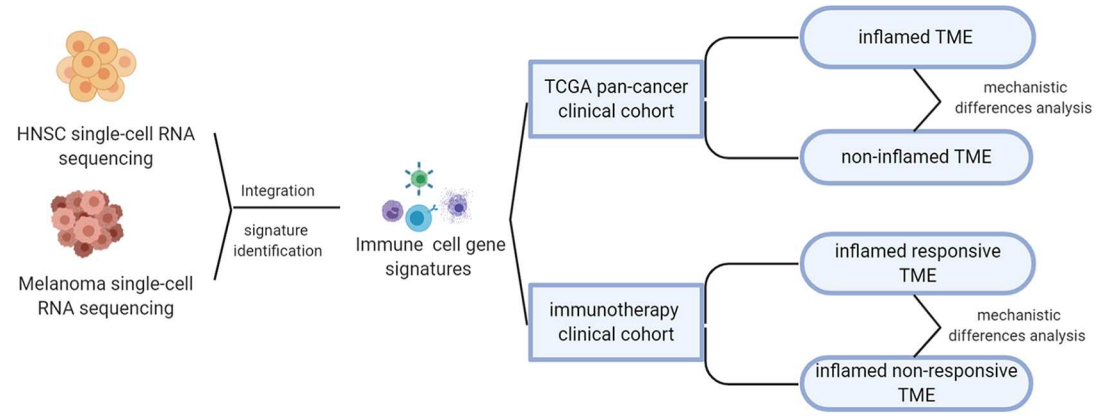


FIGURE 1 | Overall design of this study.

molecular pathway alternation (Subramanian et al., 2005). ToPASeq package was used to provide topology-based pathway analysis (Ihnatova and Budinska, 2015).

Screening for Potential Phenotype Transformation Drugs

To discover potential drugs aiding current immunotherapy, we calculated the connectivity score (Lamb et al., 2006) of multiple drugs to evaluate whether it is promising to promote the transformation of favorable TME phenotypes. This analysis was carried on *PharmacoGx* packages (Smirnov et al., 2015).

Statistical Analysis

To assess the prognostic significance of TME subtypes, we used a Cox test to calculate its hazard ratio. Then, Kaplan–Meier curves and log-rank test were used to assess the differences in the 5 years' and all years' overall survival times between inflamed and non-inflamed subtypes. Pearson's chi-square test and Fisher's exact test were used to calculate the *P*-value for the discrete variable. A $P < 0.05$ was regarded as statistically significant.

RESULTS

Integration of Single-Cell RNA Sequencing Data Sets

The overall design of this study was shown in **Figure 1**. As mentioned above, the responsive rate of immunotherapy varies by tumor type. To understand the factors that contribute to the differences in susceptibility to immunotherapy, we integrated two single-cell RNA sequencing datasets, respectively, from head and neck squamous carcinoma (HNSC) and melanoma, which were characterized by different immunotherapeutic sensitivity (∼45% response rate for melanoma Daud et al., 2016; Ribas et al., 2016, significantly higher than the 12.2% of HNSC Wang B. C. et al., 2019).

The integration result is shown in **Figures 2A–C**; tumor cells from HNSC and melanoma exhibited significant heterogeneity. Nevertheless, immune cells from different tumor types were integrated into corresponding immune cell clusters. These results suggested that immune cells from distant tumor types might have a relatively similar transcriptomic pattern, which may explain the reason why immunotherapy was always accompanied by a pan-cancer therapeutic effect. The heterogeneity of immunotherapeutic efficacy across distant tumor types may be mainly derived from different tumor cells and their tumor immune microenvironment characteristics, such as immune cell composition.

For instance, B cells are increasingly valued for their important role in immunotherapeutic resistance (Petitprez et al., 2020). As shown in **Figure 2D**, the proportion of B cells in melanoma was significantly higher than that of HNSC ($P < 0.001$, **Supplementary Table 1**).

Pan-Cancer Prognostic Significance of TME Subtypes

To classify TME phenotypes across distant tumor types, immune cell GSs were identified in the above single-cell data. Then, we classified TCGA pan-cancer samples into three TME subtypes based on the unsupervised clustering pattern of GS, each assigned as high-immune score (inflamed), intermediate immune score, or low-immune score (non-inflamed; **Figure 3A**). As shown in **Figure 3B**, the proportions of TME subtypes varied greatly among the different types of tumors. Next, we examined the association of this classification with the overall survival time of tumor patients. Consistent with previous reports from immunohistochemistry (Dubsky et al., 2019), favorable prognostic roles of inflamed TME were observed in most tumor types (such as SKCM, UCEC, etc.). Unexpectedly, as reported in a number of previous reports, an unfavorable prognostic role of inflamed TME was also observed in some tumor types, such as LGG (Zhang et al., 2017) (**Figures 3C–F**).

Molecular Characteristics of Inflamed or Non-inflamed TME Across Multiple Tumor Types

To further investigate mechanistic differences between inflamed and non-inflamed TME, we compared gene expression profiles between inflamed and non-inflamed TME. As shown in

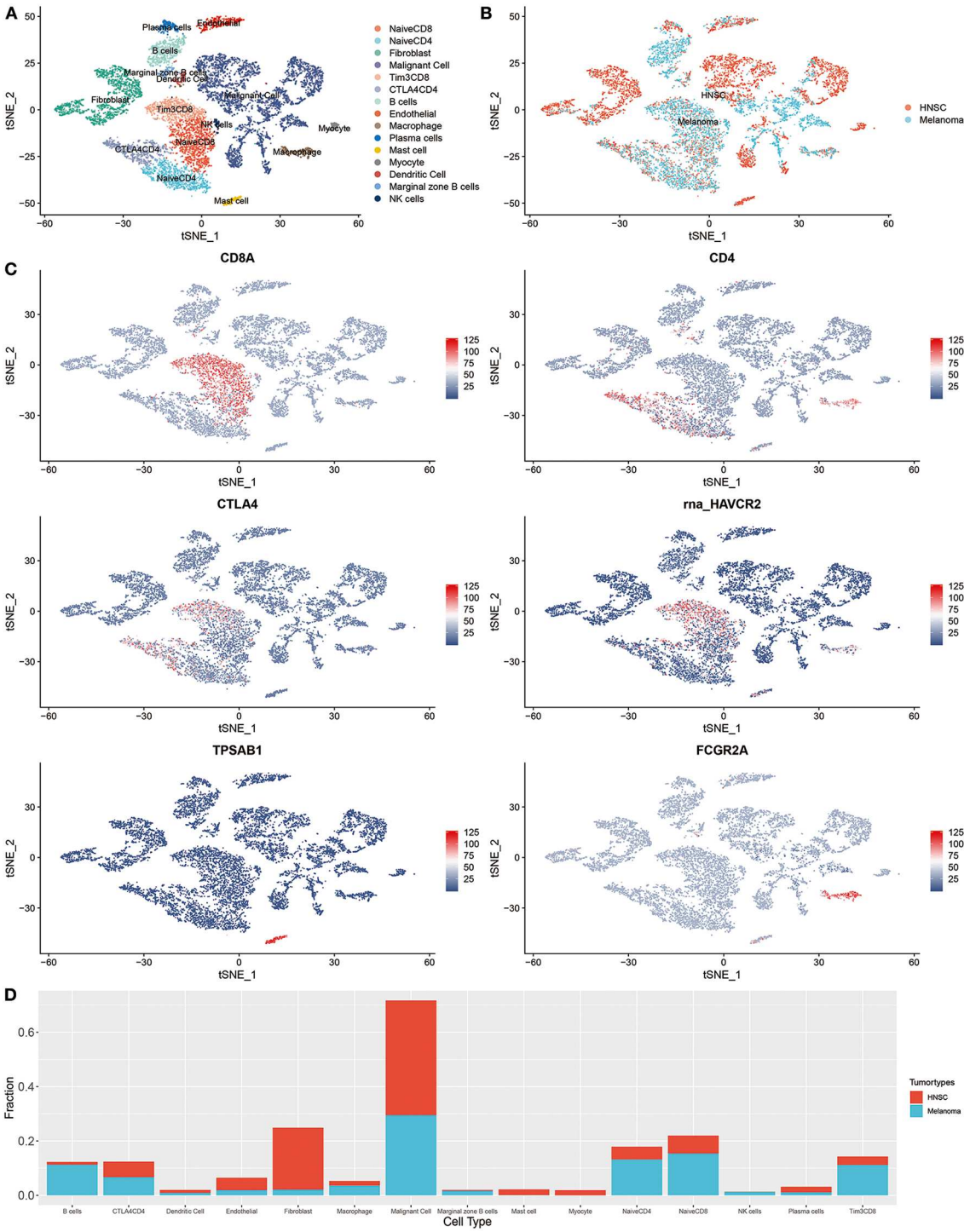
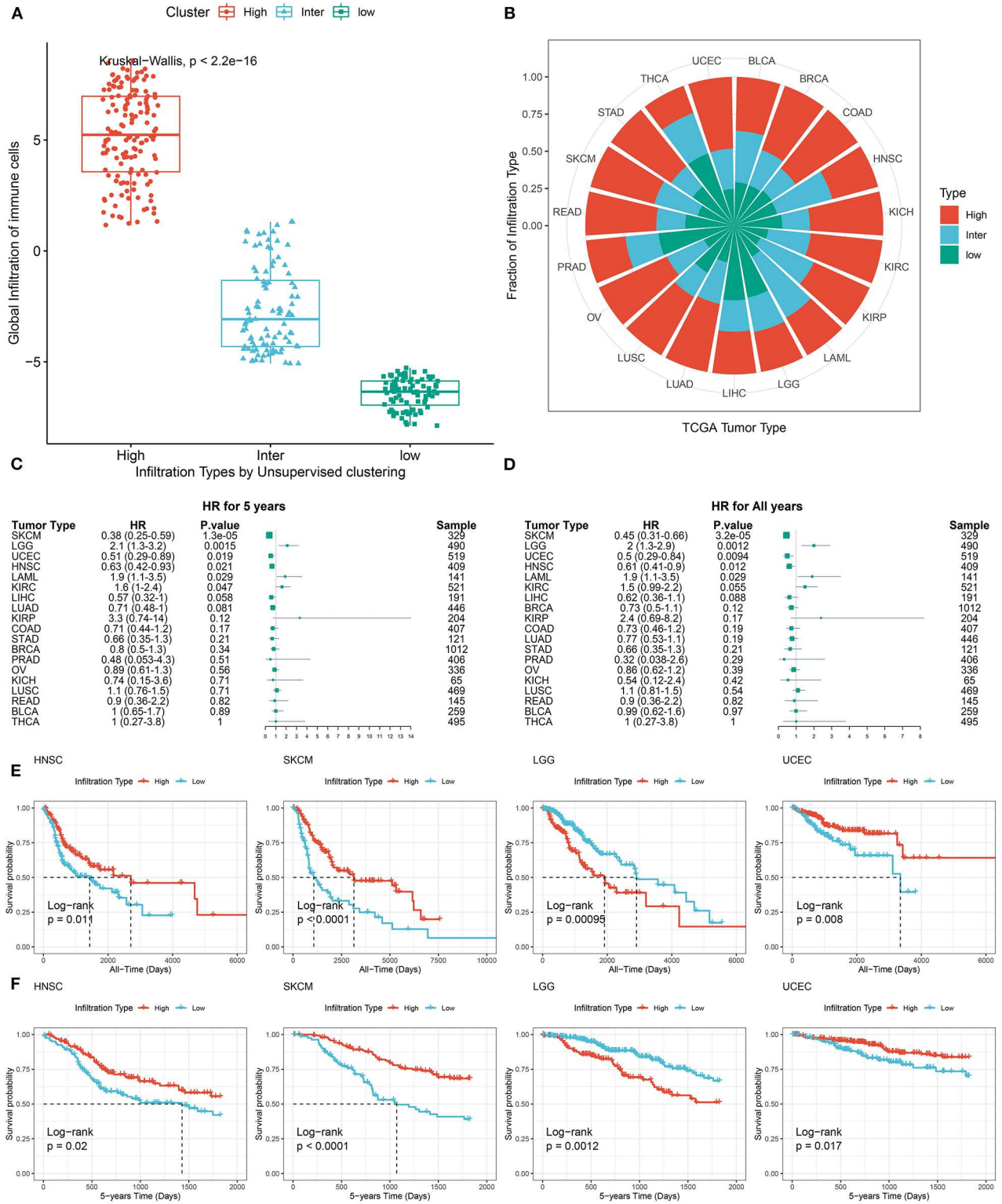


FIGURE 2 | Integrated single-cell RNA sequencing analysis revealed the microenvironment heterogeneity of distant tumor types. **(A)** The t-SNE plot displays immunological and non-immunological cells in the tumor microenvironment. Each dot represents a cell and color represents different types of cells. **(B)** The color was coded according to tumor types. **(C)** The expression of cell markers across different cell clusters. **(D)** The composition of cells in HNSC and melanoma.

Figure 4A, non-inflamed TME-specific genes (upregulated genes in non-inflamed TME) were related to the GPCR signaling pathway, neuronal system, and keratinization. Inflamed TME-specific genes (upregulated genes in inflamed TME) were related to interferon (IFN), multiple interleukin-related pathways including interleukin-4, interleukin-13,



Tumor Type	HR	P.value	Sample
SKCM	0.38 (0.25-0.59)	1.3e-05	329
LGG	2.1 (1.3-3.2)	0.0015	490
UCEC	0.51 (0.29-0.89)	0.019	519
HNSC	0.63 (0.42-0.93)	0.021	409
LAML	1.9 (1.1-3.5)	0.029	141
KIRC	1.6 (1-2.4)	0.047	521
LIHC	0.57 (0.32-1)	0.058	191
LUAD	0.71 (0.48-1)	0.081	446
KIRP	3.3 (0.74-14)	0.12	204
COAD	0.71 (0.44-1.2)	0.17	407
STAD	0.66 (0.35-1.3)	0.21	121
BRCA	0.8 (0.5-1.3)	0.34	1012
PRAD	0.48 (0.053-4.3)	0.51	406
OV	0.89 (0.61-1.3)	0.56	336
KICH	0.74 (0.15-3.6)	0.71	65
LUSC	1.1 (0.76-1.5)	0.71	469
READ	0.9 (0.36-2.2)	0.82	145
BLCA	1 (0.65-1.7)	0.89	259
THCA	1 (0.27-3.8)	1	495

Tumor Type	HR	P.value	Sample
SKCM	0.45 (0.31-0.66)	3.2e-05	329
LGG	2 (1.3-2.9)	0.0012	490
UCEC	0.5 (0.29-0.84)	0.0094	519
HNSC	0.61 (0.41-0.9)	0.012	409
LAML	1.9 (1.1-3.5)	0.029	141
KIRC	1.5 (0.99-2.2)	0.055	521
LIHC	0.62 (0.36-1.1)	0.088	191
BRCA	0.73 (0.5-1.1)	0.12	1012
KIRP	2.4 (0.69-8.2)	0.17	204
COAD	0.73 (0.46-1.2)	0.19	407
LUAD	0.77 (0.53-1.1)	0.19	446
STAD	0.66 (0.35-1.3)	0.21	121
PRAD	0.32 (0.038-2.6)	0.29	406
OV	0.86 (0.62-1.2)	0.39	336
KICH	0.54 (0.12-2.4)	0.42	65
LUSC	1.1 (0.81-1.5)	0.54	469
READ	0.9 (0.36-2.2)	0.82	145
BLCA	0.99 (0.62-1.6)	0.97	259
THCA	1 (0.27-3.8)	1	495

FIGURE 3 | Pan-cancer prognostic role of identified TME subtypes. **(A)** The global infiltration characteristics of distant TME subtypes. **(B)** The proportion of TME subtypes in different cancer types. **(C,D)** Forest plot for the association between identified TME subtypes and overall survival time of patients. **(E,F)** All years or 5 years overall survival time of inflamed (High) and non-inflamed (Low) TME. Tumor types were represented by TCGA. Standard abbreviations: LAML, acute myeloid leukemia; BLCA, bladder urothelial carcinoma; LGG, brain lower grade glioma; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; KICH, kidney chromophobe; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma; PRAD, prostate adenocarcinoma; READ, rectum adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma.

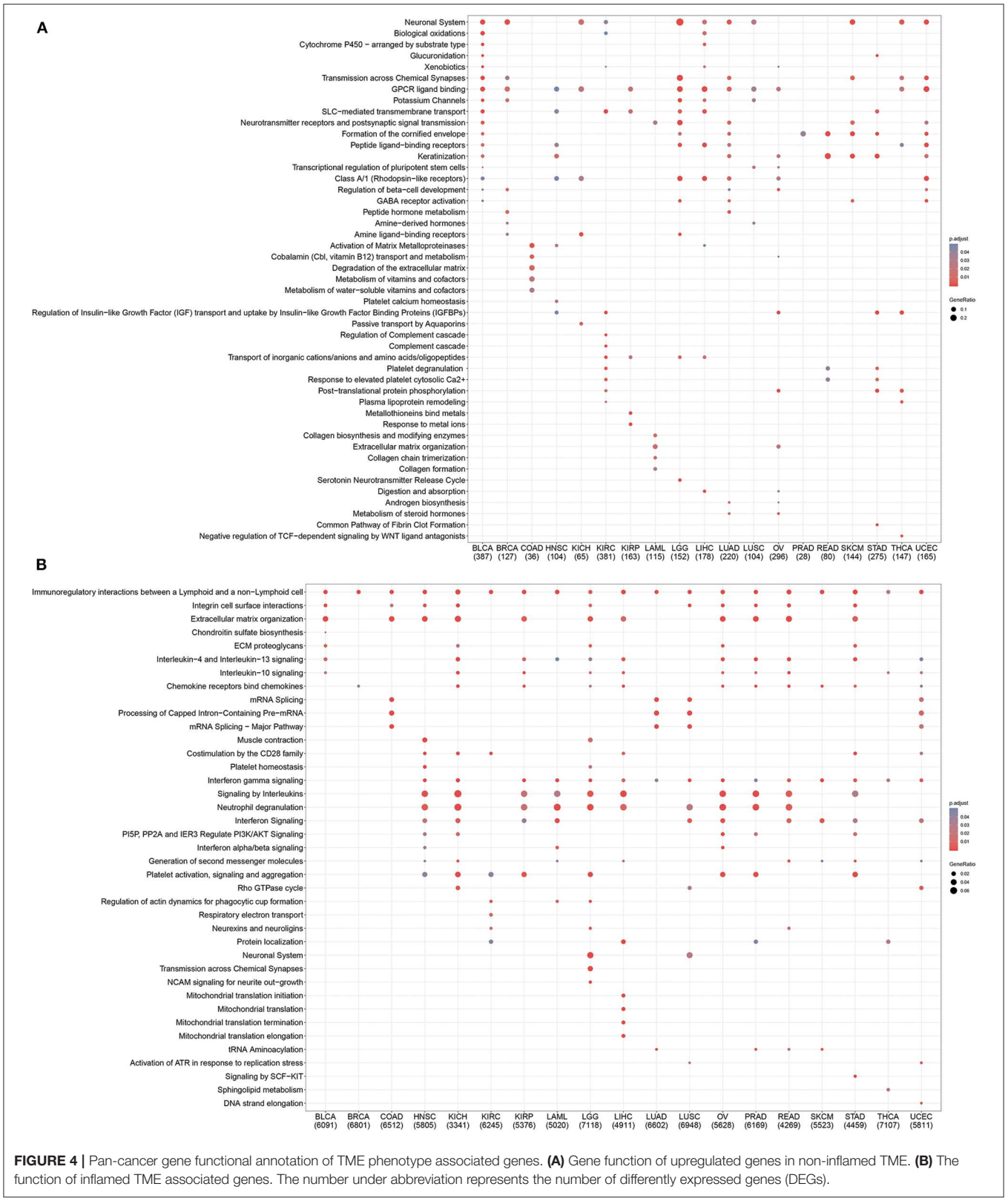


FIGURE 4 | Pan-cancer gene functional annotation of TME phenotype associated genes. **(A)** Gene function of upregulated genes in non-inflamed TME. **(B)** The function of inflamed TME associated genes. The number under abbreviation represents the number of differently expressed genes (DEGs).

and interleukin-10 signaling, CD28 costimulatory molecule family including PD-1, and CTLA-4-associated signaling pathways (**Figure 4B**). The topology-based pathway analysis demonstrated that interleukin-related pathways, interferon-related pathways, the NLRP3 inflammasome, Toll-like receptor, mitochondria, CD28 costimulation, and B cell

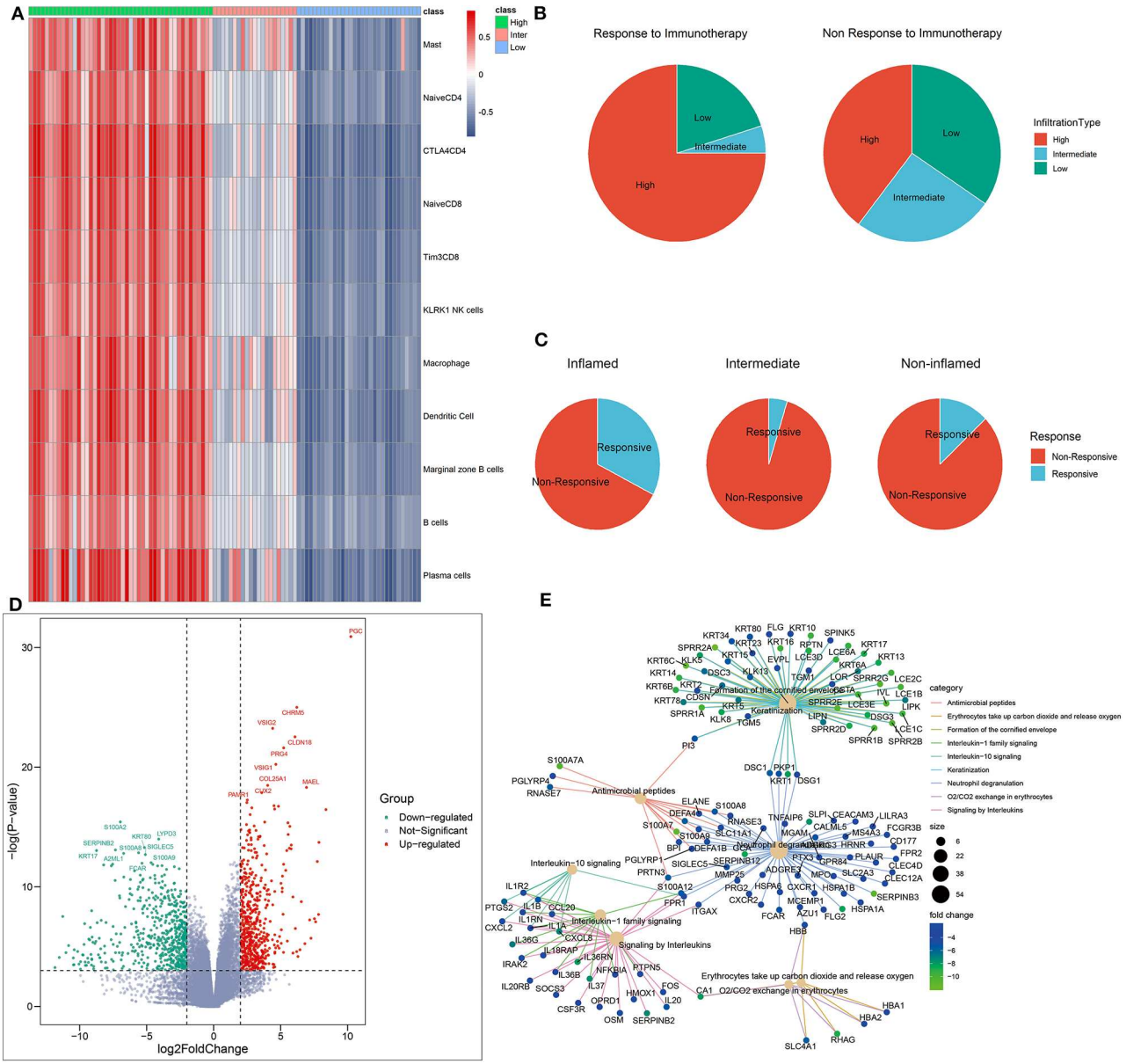


FIGURE 5 | Identification of phenotype-specific genes. **(A)** Heatmap of distant TME subtypes determined by unsupervised clustering algorithm. **(B,C)** TME subtypes correlated with the response to immunotherapy. **(D)** The volcano plot of differently expressed genes based on an RNA sequencing analysis of inflamed responders vs. inflamed non-responders. Red and green dots, respectively, represented upregulated genes and downregulated genes in inflamed responders. **(E)** The network plot showing common genes shared by top functional terms of upregulated genes in inflamed non-responders.

activation-related pathways were also activated in inflamed TME (**Supplementary Table 5**).

TME Phenotypes Correlated With the Immunotherapeutic Sensitivity

To better understand the association between TME phenotypes and the response to immunotherapy, we reproduced our TME classification in a published clinical melanoma cohort treated by immune CPIs (Riaz et al., 2017) (**Figure 5A**). This reproduction was performed based on immune GSs identified in the above single-cell RNA sequencing analysis with the same clustering parameter.

As expected, inflamed tumors were the most sensitive to CPI (CR+PR rate: 32.6% in inflamed vs. 3.2% in non-inflamed, $P = 0.015$, **Supplementary Table 2**) (**Figure 5B**), but only a percentage (CR rate: 8.7%, CR + PR rate: 32.6%) of these patients were responsive to CPI (**Figure 5C**).

To further offer mechanistic insights into CPI resistance in inflamed TME, we identified several DEGs in inflamed non-responders vs. inflamed responders (**Figure 5D**). These GSs of

TME phenotype may serve as potential targets for improving current immunotherapy.

For instance, CLDN18 was the signature of inflamed responsive TME. Therapy that directly targets on CLDN18 has shown its potential to improve the efficacy of ACT in treating solid tumors (Micke et al., 2014). On the other side, inhibiting the signature of inflamed non-responsive TME may be another promising way. Here, SIGLEC5 was significantly overexpressed in inflamed non-responders, and its family member SIGLEC15 has been proven as an efficient target to enhance antitumor immunity (Wang J. et al., 2019).

We also analyzed the correlation between TME status and known ICB response biomarkers. The inflamed TME was characterized by higher expression of PDCD1, CTLA4, CD28, (PD-L1) CD274, PD-L2, and lower tumor mutation burden than non-inflamed TME (**Supplementary Figures 3, 4**).

Mechanistic Differences Between Inflamed Responsive TME and Inflamed Non-responsive TME

Then, gene functional annotation analysis was used to understand the role of TME phenotype-specific genes. As shown in **Figure 6A**, genes upregulated in inflamed and responsive tumors enriched on complement cascade and bile metabolism. GSEA also confirmed that multiple metabolism associated pathways except for oxidative stress induced senescence were upregulated in this type of TME, including bile salt and bile acid metabolism, glucose metabolism, ethanol oxidation, glyoxylate metabolism, and glycine degradation (**Figure 6E**).

In terms of inflamed non-responsive tumors, signaling pathways, such as IL-13, IL-4, IL-10, and IL-1 cytokines-related signaling pathways and oxygen exchange pathway were upregulated, which are also downregulated in inflamed responders (**Figures 5E**, **6B**). Interestingly, the expression of CTLA-4 pathway-related genes did not differ between inflamed responders and inflamed non-responders (**Supplementary Table 4**).

The topology-based pathway analysis demonstrated that the B cell activation pathway, non-canonical NF-kB pathway, NOTCH signaling pathways, PD-1 signaling, bile acid, and bile salt metabolism-related pathways were inhibited in inflamed non-responders (**Supplementary Table 6**).

These results suggested that tumor hypermetabolism might confer resistance to immunotherapy.

Finally, for a more systematic understanding of the resistant mechanism, we applied GSEA to investigate the alternation of molecular pathways across four dimensions (epigenetic modification, immune or other associated signaling pathway, metabolism).

As shown in **Figure 6C**, multiple epigenetic signaling pathways were upregulated in inflamed non-responders, which suggested a mechanism of immunotherapeutic resistance as observed by others (Mondello et al., 2020; Olino et al., 2020).

In terms of inflamed responders, multiple carcinogenesis signaling pathways, except for the PPAR pathway, were downregulated (**Figures 6D,F**), which suggested a mechanism of therapeutic resistance and potential target for therapy. In line with this hypothesis, recent studies illustrated that PPAR agonists appeared to improve the therapeutic sensitivity of ACT and CPI therapy (Chowdhury et al., 2018; Saibil et al., 2019).

A deeper analysis of differentiating the patient population between different ICB treatments demonstrated that 135/980 (13.78%) pathways enriched on the CTLA-4 cohort were also enriched on the PDCD1 cohort (135/673, 20.06%) (**Supplementary Figure 1**). The shared pathways enriched on two cohorts were associated with glucuronidation, interleukin-10 signaling, O2/CO2 exchange in erythrocytes, post-translational phosphorylation, and metabolism of bile acids and bile salts (**Supplementary Figure 2A**). Genes dysregulated in the CTLA-4 cohort tended to be associated with epigenetic modification including epigenetic regulation of gene expression, HATs acetylate histones, HDAC deacetylates histones, transcriptional regulation by small RNAs, and gene silencing by RNA (**Supplementary Figure 2B**). Genes dysregulated in the PDCD1 cohort tended to be associated with PPAR active gene expression, glucose metabolism, extracellular matrix organization, GPCR ligand binding, and signaling by retinoic acid (**Supplementary Figure 2C**).

Screening for Potential Favorable TME Phenotype Transformation Drugs

Immunotherapy combined with chemotherapy is receiving increasing interest as a promising strategy to improve the deficiencies of current immunotherapy (Wargo et al., 2015). However, it is not completely clear how best to incorporate chemotherapy with immunotherapy. Here, we calculated the genomic connectivity score of 1,288 kinds of drugs to identify potential phenotype transformation drugs that could induce systemic favorable transcriptomic alternation, including from non-inflamed TME to inflamed TME, or from inflamed non-responsive TME to inflamed responsive TME. The drug-genomic perturbation database records genomic changes following multiple drug treatments. Analysis combining these drug-induced genomic changes with identified phenotypic genomic differences can help us find potential drugs that could convert unfavorable TME to favorable TME.

Mercaptopurine (6-MP) was identified as the most promising drug that might promote the transformation of inflamed responsive TME phenotype (**Table 1**). Interestingly, although some reports have shown that 6-MP can enhance the vaccine-dependent antitumor immunit y (Kataoka et al., 1984; Kataoka and Oh-hashi, 1985), it seems to be forgotten after that. But there are increasing interests trying to use 6-MP as a drug of immune disorders, such as autoimmune hepatitis (Hübener et al., 2016), inflammatory bowel disease (Present et al., 1989), etc. This may be because 6-mercaptopurine is widely recognized as an immunosuppressive agent, but our findings implicated that immunomodulatory may be a more accurate definition of such drugs. Our results indicated that further clinical studies are needed to assess the value of the combination of 6-MP with current immunotherapy.

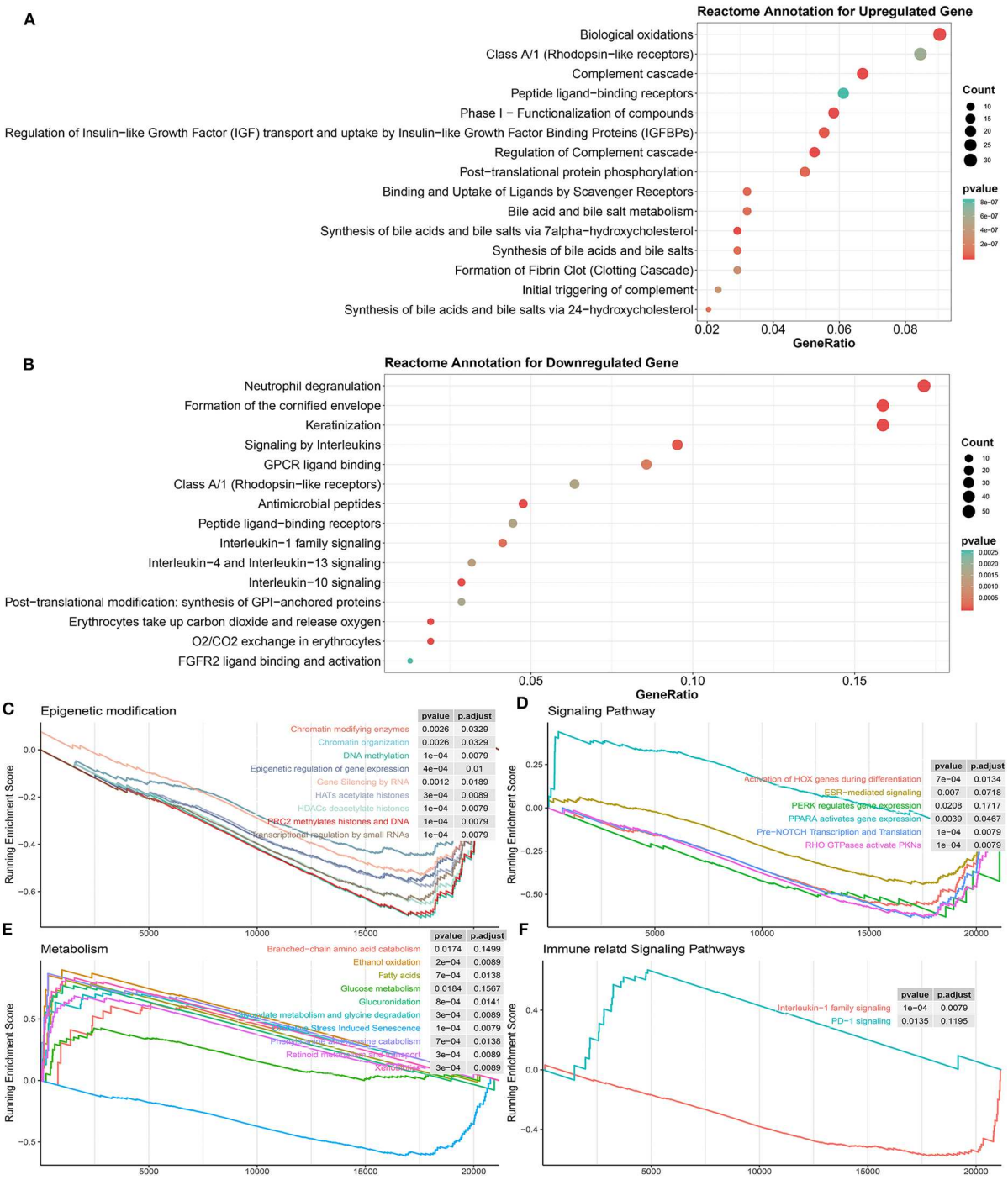


FIGURE 6 | Biological processes correlated with identified DEGs in inflamed responders. **(A)** Detailed function annotation of upregulated genes in inflamed responders (compared to inflamed non-responders). **(B)** Functional annotation of downregulated genes in inflamed responders (compared to inflamed non-responders). **(C–F)** GSEA shows four dimensions of the molecular function of DEGs across inflamed responders. Running enrichment score >0 means this pathway is upregulated in inflamed responders; running enrichment score <0 means this pathway is downregulated in inflamed responders.

DISCUSSION

Molecular stratification of TME phenotypes is paving the way for a better understanding of immunotherapeutic heterogeneity. Here, based on immune GSs developed from integrated single-cell RNA sequencing analysis, we systematically analyzed the molecular characteristics of inflamed TME across multiple cancer types and provided

TABLE 1 | Screening for potential favorable TME phenotype transformation drugs.

Potential drugs that convert "non-inflamed" TME to "inflamed" TME			Potential drugs that convert "inflamed nonresponsive" TME to "inflamed responsive" TME		
Drugs	**Connectivity**	***P*-Value**	**Drugs**	**Connectivity**	***P*-Value**
Clofibrate	0.803	0.021	Mercaptopurine	0.817	0.028
Metronidazole	0.637	0.002	Valdecoxib	0.309	0.040
Zalcitabine	0.605	0.028	5253409	0.292	0.006
Gabexate	0.585	0.012	Astemizole	0.290	0.039
S-propranolol	0.581	0.007	Isoconazole	0.283	0.003
Ifenprodil	0.580	0.044	Pizotifen	0.279	0.011
Sulfapyridine	0.578	0.027	Econazole	0.278	0.003
Succinylsulfathiazole	0.570	0.042	Orciprenaline	0.267	0.006

mechanistic insights into immunotherapeutic resistance in inflamed TME.

Some of the identified mechanistic differences have been supported by recent reports. Examples highlighted by these data include the upregulation of epigenetic signaling pathways and the downregulation of PPAR-signaling pathways in inflamed non-responsive tumors. These dysregulated pathways may be potential targets for improving the sensitivity to immunotherapy. Importantly, these results are in line with prior publications, which have provided some evidence that inhibition of epigenetic modification (Mondello et al., 2020) or activation of PPAR signaling pathways (Chowdhury et al., 2018; Saibil et al., 2019) might be a promising way to overcome therapeutic resistance to immune checkpoint blockade or ACT.

Our results also revealed the molecular characteristics of inflamed TME shared by different tumor types. These results demonstrated that inflamed TME was related to enhanced cytokine expression (interferon and IL-4,−13, and−10). Interestingly, these cytokines, except for interferon, were also upregulated in inflamed non-responders, which suggested a dual role of these interleukins. These results are in line with prior published reports (Mannino et al., 2015; Wang et al., 2016). For example, IL-10 is widely recognized as an immunosuppressive cytokine, but there is increasing evidence that it has a dual role in antitumor immunity. Blocking or activation of IL-10 has been proven as an efficient way to enhance antitumor immunity in different aspects (Ni et al., 2015; Naing et al., 2018). According to our results, we believe that TME phenotypes should be considered as a key factor in further study design to illuminate the remaining mysteries of IL-10.

In addition, our results have far-reaching clinical implications including the identification of multiple potential molecular targets for developing novel immunotherapy and combination therapeutic strategies. For instance, the success of ACT cannot be reproduced on solid tumors due to the obstacle of its microenvironment. Therefore, rather than directly targeting on whole solid tumors, selectively targeting the inflamed and responsive TME might be another easier therapeutic way. As expected, this hypothesis is supported by a recent report. CLDN18, a signature of inflamed and responsive TME, has been proven as an efficient target for improving the efficacy of current ACT on solid tumors (Micke et al., 2014).

Except for targeting on inflamed and responsive TME, examples highlighted by our data also included inhibiting the signature of inflamed non-responsive TME to reverse therapeutic resistance. For example, SIGLEC15, a signature of inflamed and non-responsive TME, has shown its power in blocking immune escape. Interestingly, its antitumor immunity enhancement effect is independent of the PD-1/PD-L1 axis, suggesting that it may be an ideal target to aid current anti-PD-1 therapy (Wang J. et al., 2019).

Finally, based on a drug-genomic perturbation database, we identified some drugs that were promising for promoting the transformation from an unfavorable TME phenotype to a favorable one.

In conclusion, our result provided an important view for understanding how inflamed TME and inflamed resistant TME form. This evidence has important clinical implications and may help guide rational combination immunotherapy.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GEO:https://www.ncbi.nlm.nih.gov/gds/.

AUTHOR CONTRIBUTIONS

BW and YO designed and supervised the study and was a major contributor in editing the manuscript. BW, ZR, and ML analyzed and interpreted the data and were major contributors in writing the manuscript. BW, XL, and JL performed analysis and contributed to writing the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Thanks to everyone who pushed the boundaries of human knowledge.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00348/full#supplementary-material

REFERENCES

Abril-Rodriguez, G., and Ribas, A. (2017). SnapShot: immune checkpoint inhibitors. *Cancer Cell* 31, 848–848.e1. doi: 10.1016/j.ccell.2017.05.010

Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* 24, 541–550. doi: 10.1038/s41591-018-0014-x

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

Chen, P. L., Roh, W., Reuben, A., Cooper, Z. A., Spencer, C. N., Prieto, P. A., et al. (2016). Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov.* 6, 827–837. doi: 10.1158/2159-8290.CD-15-1545

Chowdhury, P. S., Chamoto, K., Kumar, A., and Honjo, T. (2018). PPAR-induced fatty acid oxidation in T cells increases the number of tumor-reactive CD8(+) T cells and facilitates anti-PD-1 therapy. *Cancer Immunol Res.* 6, 1375–1387. doi: 10.1158/2326-6066.CIR-18-0095

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

Darvin, P., Toor, S. M., Sasidharan Nair, V., and Elkord, E. (2018). Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp. Mol. Med.* 50:165. doi: 10.1038/s12276-018-0191-1

Daud, A. I., Wolchok, J. D., Robert, C., Hwu, W. J., Weber, J. S., Ribas, A., et al. (2016). Programmed death-ligand 1 expression and response to the anti-programmed death 1 antibody pembrolizumab in melanoma. *J. Clin. Oncol.* 34, 4102–4109. doi: 10.1200/JCO.2016.67.2477

Dubsky, P., Montagna, G., and Ritter, M. (2019). Lymphocyte infiltration predicts survival in chemotherapy-naive, triple-negative breast cancer and identifies patients with intrinsically good prognosis: have we been bringing owls to Athens?. *Ann. Oncol.* 30, 1849–1850. doi: 10.1093/annonc/mdz444

Galon, J., and Bruni, D. (2019). Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* 18, 197–218. doi: 10.1038/s41573-018-0007-y

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

Hübener, S., Oo, Y. H., Than, N. N., Hübener, P., Weiler-Normann, C., Lohse, A. W., et al. (2016). Efficacy of 6-Mercaptopurine as Second-Line Treatment for Patients With Autoimmune Hepatitis and Azathioprine Intolerance. *Clin. Gastroenterol. Hepatol.* 14, 445–453. doi: 10.1016/j.cgh.2015.09.037

Ihnatova, I., and Budinska, E. (2015). ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. *BMC Bioinformatics* 16:350. doi: 10.1186/s12859-015-0763-1

Ji, R. R., Chasalow, S. D., Wang, L., Hamid, O., Schmidt, H., Cogswell, J., et al. (2012). An immune-active tumor microenvironment favors clinical response to ipilimumab. *Cancer Immunol. Immunother.* 61, 1019–1031. doi: 10.1007/s00262-011-1172-6

Kataoka, T., Akahori, Y., and Sakurai, Y. (1984). 6-Mercaptopurine-induced potentiation of active immunotherapy in L1210-bearing mice treated with concanavalin A-bound leukemia cell vaccine. *Cancer Res.* 44, 519–524.

Kataoka, T., and Oh-hashi, F. (1985). Suppressor macrophages in tumor-bearing mice and their selective inhibition by 6-mercaptopurine. *Cancer Res.* 45, 2139–2144.

Kortlever, R. M., Sodir, N. M., Wilson, C. H., Burkhart, D. L., Pellegrinet, L., Brown Swigart, L., et al. (2017). Myc cooperates with ras by programming inflammation and immune suppression. *Cell* 171, 1301–1315.e14. doi: 10.1016/j.cell.2017.11.013

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., et al. (2006). The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935. doi: 10.1126/science.1132939

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Cowey, C. L., Lao, C. D., et al. (2015). Combined nivolumab and ipilimumab or monotherapy in untreated melanoma. *N. Engl. J. Med.* 373, 23–34. doi: 10.1056/NEJMoa1504030

Mannino, M. H., Zhu, Z., Xiao, H., Bai, Q., Wakefield, M. R., and Fang, Y. (2015). The paradoxical role of IL-10 in immunity and cancer. *Cancer Lett.* 367, 103–107. doi: 10.1016/j.canlet.2015.07.009

Micke, P., Mattsson, J. S., Edlund, K., Lohr, M., Jirström, K., Berglund, A., et al. (2014). Aberrantly activated claudin 6 and 18.2 as potential therapy targets in non-small-cell lung cancer. *Int. J. Cancer* 135, 2206–2214. doi: 10.1002/ijc.28857

Mondello, P., Tadros, S., Teater, M., Fontan, L., Chang, A. Y., Jain, N., et al. (2020). Selective inhibition of HDAC3 targets synthetic vulnerabilities and activates immune surveillance in lymphoma. *Cancer Discov.* 10, 440–459. doi: 10.1158/2159-8290.CD-19-0116

Mounir, M., Lucchetta, M., Silva, T. C., Olsen, C., Bontempi, G., Chen, X., et al. (2019). New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.* 15:e1006701. doi: 10.1371/journal.pcbi.1006701

Naing, A., Infante, J. R., Papadopoulos, K. P., Chan, I. H., Shen, C., Ratti, N. P., et al. (2018). PEGylated IL-10 (pegilodecakin) induces systemic immune activation, CD8(+) T cell invigoration and polyclonal T cell expansion in cancer patients. *Cancer Cell* 34, 775–791.e3. doi: 10.1016/j.ccell.2018.10.007

Newick, K., O'Brien, S., Moon, E., and Albelda, S. M. (2017). CAR T cell therapy for solid tumors. *Annu. Rev. Med.* 68, 139–152. doi: 10.1146/annurev-med-062315-120245

Ni, G., Wang, T., Walton, S., Zhu, B., Chen, S., Wu, X., et al. (2015). Manipulating IL-10 signalling blockade for better immunotherapy. *Cell. Immunol.* 293, 126–129. doi: 10.1016/j.cellimm.2014.12.012

Olino, K., Park, T., and Ahuja, N. (2020). Exposing hidden targets: combining epigenetic and immunotherapy to overcome cancer resistance. *Semin. Cancer Biol.* 3:1044. doi: 10.1016/j.semcancer.2020.01.001

Peng, D., Kryczek, I., Nagarsheth, N., Zhao, L., Wei, S., Wang, W., et al. (2015). Epigenetic silencing of TH1-type chemokines shapes tumour immunity and immunotherapy. *Nature* 527, 249–253. doi: 10.1038/nature15520

Petitprez, F., de Reyniès, A., Keung, E. Z., Chen, T. W., Sun, C. M., Calderaro, J., et al. (2020). B cells are associated with survival and immunotherapy response in sarcoma. *Nature* 577, 556–560. doi: 10.1038/s41586-019-1906-8

Present, D. H., Meltzer, S. J., Krumholz, M. P., Wolke, A., and Korelitz, B. I. (1989). 6-Mercaptopurine in the management of inflammatory bowel disease: short- and long-term toxicity. *Ann. Intern. Med.* 111, 641–649. doi: 10.7326/0003-4819-111-8-641

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 171, 1611–1624.e24. doi: 10.1016/j.cell.2017.10.044

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., and Piccolo, S. R. (2015). Alternative preprocessing of RNA-sequencing data in the cancer genome atlas leads to improved analysis results. *Bioinformatics* 31, 3666–3672. doi: 10.1093/bioinformatics/btv377

Riaz, N., Havel, J. J., Makarov, V., Desrichard, A., Urba, W. J., Sims, J. S., et al. (2017). Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* 171, 934–949.e16. doi: 10.1016/j.cell.2017.09.028

Ribas, A., Hamid, O., Daud, A., Hodi, F. S., Wolchok, J. D., Kefford, R., et al. (2016). Association of pembrolizumab with tumor response and survival among patients with advanced melanoma. *JAMA* 315, 1600–1609. doi: 10.1001/jama.2016.4059

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–40. doi: 10.1093/bioinformatics/btp616

Saibil, S. D., St Paul, M., Laister, R. C., Garcia-Batres, C. R., Israni-Winger, K., Elford, A. R., et al. (2019). Activation of peroxisome proliferator-activated receptors alpha and delta synergizes with inflammatory signals to enhance adoptive cell therapy. *Cancer Res.* 79, 445–451. doi: 10.1158/0008-5472.CAN-17-3053

Sekula, M., Datta, S., and Datta, S. (2017). optCluster: an R package for determining the optimal clustering algorithm. *Bioinformation* 13, 101–103. doi: 10.6026/97320630013101

Silva, T. C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., et al. (2016). TCGA workflow: analyze cancer genomics and epigenomics data using bioconductor packages. *F1000Res.* 5:1542. doi: 10.12688/f1000research.8923.2

Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., et al. (2015). PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics* 32, 1244–1246. doi: 10.1093/bioinformatics/btv723

Spranger, S., Bao, R., and Gajewski, T. F. (2015). Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity. *Nature* 523, 231–235. doi: 10.1038/nature14404

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. doi: 10.1016/j.cell.2019.05.031

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

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H. II., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. doi: 10.1126/science.aad0501

Wang, B. C., Cao, R. B., Li, P. D., and Fu, C. (2019). The effects and safety of PD-1/PD-L1 inhibitors on head and neck cancer: a systematic review and meta-analysis. *Cancer Med*. 8, 5969–5978. doi: 10.1002/cam4.2510

Wang, J., Sun, J., Liu, L. N., Flies, D. B., Nie, X., Toki, M., et al. (2019). Siglec-15 as an immune suppressor and potential target for normalization cancer immunotherapy. *Nat. Med*. 25, 656–666. doi: 10.1038/s41591-019-0374-x

Wang, T., Lu, R., Kapur, P., Jaiswal, B. S., Hannan, R., Zhang, Z., et al. (2018). An empirical approach leveraging tumorgrafts to dissect the tumor microenvironment in renal cell carcinoma identifies missing link to prognostic inflammatory factors. *Cancer Discov*. 8, 1142–1155. doi: 10.1158/2159-8290.CD-17-1246

Wang, Y., Sun, S. N., Liu, Q., Yu, Y. Y., Guo, J., Wang, K., et al. (2016). Autocrine complement inhibits IL10-dependent T-cell-mediated antitumor immunity to promote tumor progression. *Cancer Discov*. 6, 1022–1035. doi: 10.1158/2159-8290.CD-15-1412

Wargo, J. A., Reuben, A., Cooper, Z. A., Oh, K. S., and Sullivan, R. J. (2015). Immune effects of chemotherapy, radiation, and targeted therapy and opportunities for combination with immunotherapy. *Semin. Oncol*. 42, 601–616. doi: 10.1053/j.seminoncol.2015.05.007

Yu, G., and He, Q. Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst*. 12, 477–479. doi: 10.1039/C5MB00663E

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

Zhang, C., Cheng, W., Ren, X., Wang, Z., Liu, X., Li, G., et al. (2017). Tumor purity as an underlying key factor in glioma. *Clin. Cancer Res*. 23, 6279–6291. doi: 10.1158/1078-0432.CCR-16-2598

Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res*. 47, D721–D728. doi: 10.1093/nar/gky900

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 10 July 2020
doi: 10.3389/fgene.2020.00522

Gene Signature and Identification of Clinical Trait-Related m^{6}A Regulators in Pancreatic Cancer

*Jie Hou, Zhan Wang, Hong Li, Hongzhi Zhang and Lan Luo**

The People's Hospital of Baoan Shenzhen, The 8th people's Hospital of Shenzhen, The Affiliated Baoan Hospital of Southern Medical University, Shenzhen, China

OPEN ACCESS

Edited by:
Ling Kui,
Harvard Medical School,
United States

Reviewed by:
Hong Zheng,
Stanford University, United States
Bogang Wu,
George Washington University,
United States
Lu Jiang,
Johns Hopkins University,
United States

***Correspondence:**
Lan Luo
luolan1066@163.com

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

Received: *23 February 2020*
Accepted: *29 April 2020*
Published: *10 July 2020*

Citation:
Hou J, Wang Z, Li H, Zhang H and Luo L (2020) Gene Signature and Identification of Clinical Trait-Related m^{6}A Regulators in Pancreatic Cancer. Front. Genet. 11:522. doi: 10.3389/fgene.2020.00522

Pancreatic cancer (PC) has a very poor prognosis and is usually diagnosed only at an advanced stage. The discovery of new biomarkers for PC will help in early diagnosis and a better prognosis for patients. Recently, N6-methyladenosine (m^{6}A) RNA modifications and their regulators have been implicated in the development of many cancers. To investigate the functions and mechanisms of m^{6}A modifications in the development of PC, 19 m^{6}A regulators, including m^{6}A-methyltransferases (ZC3H13, RBM15/15B, WTAP, KIAA1429, and METTL3/14), demethylases (FTO and ALKBH5), and binding proteins (YTHDF1/2/3, YTHDC1/2, IGF2BP1/2/3, HNRNPC, and HNRNPA2B1) were analyzed in 178 PC tissues from the cancer genome atlas (TCGA) database. The results were verified in PC cell lines Mia-PaCa-2, BXPC-3, and the control cell line HDE-CT. The m^{6}A regulators-based sample clusters were significantly related to overall survival (OS). Further, lasso regression identified a six-m^{6}A-regulator-signature prognostic model (KIAA1429, HNRNPC, METTL3, YTHDF1, IGF2BP2, and IGF2BP3). Model-based high-risk and low-risk groups were significantly correlated with OS and clinical traits (pathologic M, N, and clinical stages and vital status). The risk signature was verified as an independent prognostic marker for patients with PC. Finally, gene set enrichment analysis revealed m^{6}A regulators (KIAA1429, HNRNPC, and IGF2BP2) were related to multiple biological behaviors in PC, including adipocytokine signaling, the well vs. poorly differentiated tumor pathway, tumor metastasis pathway, epithelial mesenchymal transition pathway, gemcitabine resistance pathway, and stemness pathway. In summary, the m6A regulatory factors which related to clinical characteristics can be involved in the malignant progression of PC, and the constructed risk markers may be a promising prognostic biomarker that can guide the individualized treatment of PC patients.

Keywords: m^{6}A regulators, pancreatic cancer, prognostic model, biomarker, clinical traits

Abbreviations: FDR, false discovery rate; GSEA, gene set enrichment analysis; HPDE6-C7, human pancreatic duct epithelial cells; m6A, N6-methyladenosine; OS, overall survival; PC, pancreatic cancer; ROC, receiver operator characteristic curve; TCGA, the cancer genome atlas; YAP1, yes-associated protein 1.

INTRODUCTION

Pancreatic cancer (PC) is one of the most lethal malignant neoplasms and has become one of the leading causes of cancer-related deaths in developed countries (Ilic and Ilic, 2016). About 85% of PCs are adenocarcinoma, and less than 5% are pancreatic endocrine tumors (Wolfgang et al., 2013). There are usually no noticeable symptoms in the early stages of PC. When symptoms are specific enough to suggest PC, the disease might have reached an advanced stage. By the time of diagnosis, PC has often spread or metastasized to other parts of the body (Mohammed et al., 2014). With the development of medical techniques, PC can be diagnosed by ultrasound or computed tomography combined with blood tests and examination of tissue samples (biopsies). However, screening the general population for the early stage of the disease is not effective (Welinsky and Lucas, 2017). Pancreatic cancer can be treated with surgery, chemotherapy, targeted therapy, radiotherapy, palliative care, immunotherapy, or a combination of these based on the cancer stage (Mcguigan et al., 2018). With the current treatment methods pancreatic adenocarcinoma has a very poor prognosis: only 25% of patients with PC survive one year and 5 year-overall survival (OS) is lower than 5% (Ansari et al., 2015). Current treatment and diagnostic methods are not enough for the management of PC. Therefore, key goals for PC research are to develop novel prognostic markers, improve the early diagnostic rate, and find new targets for molecular targeted therapy. N6-methyladenosine (m^6A), a potential biomarker, is a chemical modification present in multiple RNA species, which take part in various biological processes in cancer (Liu et al., 2018).

N6-methyladenosine regulators are involved in more than 60% of all RNA [messenger (mRNA), transport RNA (tRNA), and ribosomal RNA (rRNA)] modifications, which is an intense area of research for post-transcriptional regulation including translation, mRNA splicing, and mRNA stability (Yu et al., 2018). The level of modification of transcripts with m^6A is regulated by methyltransferases, binding proteins and demethylases (Koh et al., 2019). The methyltransferases (including ZC3H13, RBM15, RBM15B, KIAA1429, METTL3/14, and WTAP), act as "writers," and add the methyl group to the nitrogen on the sixth carbon of the aromatic ring of an adenosine residue (Meyer and Jaffrey, 2017). The cellular m^6A status is reverted by demethylases (FTO, and ALKBH5; called "erasers"), and is recognized by m^6A-binding proteins (HNRNPC, YTHDF1/2/3, YTHDC1/2, IGF2BP1/2/3, and HNRNPA2B1; called "readers") (Zaccara et al., 2019). N6-methyladenosine, a potential biomarker, is a chemical modification present in multiple RNA species, which take part in various biological processes in cancer (Liu et al., 2018). The dysregulation of m^6A regulators is involved in the occurrence and development of multiple cancers, including bladder cancer, prostate cancer, head and neck squamous cell carcinoma, gastric cancer, breast cancer, hepatocellular carcinoma, and colorectal cancer (Hong, 2018). For example, METTL14 which suppresses colorectal cancer progression via regulating m^6A-dependent miR-375/yes-associated protein 1 (YAP1) pathway, is downregulated in colorectal cancer tissues and cell lines (Chen et al., 2020). FTO, a key m^6A demethylase, is up-regulated in human breast cancer and is significantly associated with poor survival rates (Niu et al., 2019). FTO mediates m^6A demethylation in the 3'UTR of BNIP3 mRNA and induces its degradation via an YTHDF2 independent mechanism, which indicates that FTO can serve as a novel potential therapeutic target for breast cancer (Niu et al., 2019). It has also been reported that IGF2BP2 regulates lncRNA DANCR through m6A modification, and IGF2BP2 and DANCR jointly promote the stemness-like characteristics of cancer and the pathogenesis of PC (Hu et al., 2019). Although more and more studies have shown that m6A regulatory factors play a crucial role in the pathogenesis and development of cancer, the fundamental relationship between m6A regulatory factors and PC remains unclear (Xia et al., 2019). The construction of prognostic signal based on m6A regulators that predicting the prognosis of PC will be helpful for prediction, prevention and personalized treatment.

This study used ConsensusClusterPlus to find that m^6A regulators were closely related to PC OS rates in different clusters. Furthermore, lasso regression was used to identify a six-gene signature model (KIAA1429, HNRNPC, METTL3, YTHDF1, IGF2BP2, and IGF2BP3). Most of the genes identified were consistent with previous data (Taketo et al., 2018). For example, the m^6A eraser ALKBH5, which was indicated as a potential therapeutic target for PC, was downregulated in PC cells and immortalized human pancreatic duct epithelial (HPDE6-C7) cells (He et al., 2018). Immunohistochemistry (IHC), western blots, and RT-qPCR were used to detect the expression of METTL3 in PC, and the results showed that METTL3 protein and mRNA levels were significantly higher in tumor samples than in paracancer samples. Down-regulation of METTL3 reduced the proliferation, invasion and migration of PC cell lines (Xia et al., 2019). While it is known that m^6A plays important roles in different types of cancers, the available clinical trait-related m^6A regulator studies in PC are insufficient. Single-gene analysis are used to predict prognosis and to guide therapy in cancer. However, RNA-Seq is helpful for the construction of a prediction model using multiple genes. Here, we analyzed the gene signatures in different PC cell lines and identified clinical trait-related m^6A regulators in PC. Additionally, potential related enrichment pathways of m^6A regulators might be useful to further study their mechanisms of action.

MATERIALS AND METHODS

Data Sources

RNA-seq transcriptome data, the corresponding clinical data, and large-scale cancer patient information for 178 patients with PC were obtained from the cancer genome atlas (TCGA) database[1]. The m^6A regulator genes include ZC3H13, RBM15/15B, KIAA1429, METTL14, YTHDC1/2, WTAP, METTL3, FTO, ALKBH5, YTHDF1/2/3, HNRNPA2B1, IGF2BP1/2/3, and HNRNPC. The corresponding clinical data include age at initial pathologic diagnosis (patients were aged 35–88), documented alcohol history (yes or no), alcoholic exposure category (daily

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

drinker, weekly drinker, occasional drinker, social drinker, and non-drinker), anatomic neoplasm subdivision (body of pancreas, head of pancreas, tail of pancreas, and other parts), family history of cancer (yes or no), gender (male and female), history of chronic pancreatitis (yes or no), history of diabetes (yes or no), count of lymph nodes examined (from 1 to 57), neoplasm histologic grade (G1, G2, and G3), pathologic M (M represents tumor metastasis, including M0, M1,and MX), pathologic N (N represents tumor lymph node metastasis, including N0, N1, N2, and NX), pathologic T (T represents tumor size, including T1, T2, T3,T4, and TX), pathologic stage (Stages I, II, III, and IV), vital status (alive or dead).

We systematically searched for PC gene expression datasets that were publicly available and reported full clinical annotations. we download data from GSE28735 "Microarray gene-expression profiles of 45 matching pairs of pancreatic tumor and adjacent non-tumor tissues from 45 patients with pancreatic ductal adenocarcinoma" to validate the reliability of the built model. The raw data from the microarray datasets generated by Affymetrix and Illumina were downloaded from the Gene Expression Omnibus[2].

Protein–Protein Interactions Network Construction and Correlation Analysis

The STRING database[3] was used for analyzing the protein–protein interactions (PPI) among m^6A regulators. The association among different m^6A regulators was revealed by Spearman correlation coefficient with R package.

Cell Lines and Cell Culture

Two PC cell lines (Mia-PaCa-2 and BXPC-3) and one control cell line (HDE-CT) were purchased from China Center for Type Culture Collection (CCTCC, Shanghai, China). HDE-CT is a normal human pancreatic cell line and is cultured in DMEM medium (Corning, NY, United States) supplemented with 10% fetal bovine serum (GIBCO, South America, NY, United States). Mia-PaCa-2 with a KRAS mutation and BXPC-3 with wild type KRAS are human PC cell lines. Mia-PaCa-2 was cultured in DMEM medium with 10% fetal bovine serum, and BXPC-3 was cultured in RPMI-1640 medium with 10% fetal bovine serum. All cell lines were maintained in 5% CO_2 atmosphere at 37°C.

RNA Extraction and qRT-PCR Verification

Total RNA of the four PC cell lines (Mia-PaCa-2 and BXPC-3) and HDE-CT were extracted with an RNA extraction kit (QIAGEN) according to the manufacturer's instructions. Briefly, 1×10^7 cells were collected and lysed for 10 min, genomic DNA was removed with an adsorption column, the samples were washed once with 75% ethyl alcohol and twice with wash buffer, and the samples were resuspended in RNA-grade enzyme-free water. Total RNA was reversely transcribed into cDNA and used to perform quantitative real-time PCR (qRT-PCR) with SYBR Premix ExTaq (TaKaRa). GAPDH was used as a reference gene. Primers (**Table 1**) were synthesized by Sangon Biotech (Shanghai, China).

Consensus Clustering for PC Tissues

Pancreatic cancer tissues with expression information for m^6A regulator genes (ZC3H13, RBM15, RBM15B, KIAA1429, YTHDC1, YTHDC2, METTL3, METTL14, WTAP, FTO, ALKBH5, YTHDF1, YTHDF2, YTHDF3, IGF2BP1, IGF2BP2, IGF2BP3, HNRNPA2B1, and HNRNPC) were clustered with a hierarchical agglomerative consensus. Clustering was based on Ward's linkage and Euclidean distance methods. Unsupervised

TABLE 1 | The list of RNA molecules that were assessed on the cell lines (note:F forward, R reverse).

Primer name	Primer sequence (from 5′ to 3′)
ZC3H13-F	GATCAGTTAAAGCGTGGAGAAC
ZC3H13-R	CTCTCTGTCGTGTTCATATCGA
FTO-F	GTTCACAACCTCGGTTTAGTTC
FTO-R	CATCATCATTGTCCACATCGTC
ALKBH5-F	GCAAGGTGAAGAGCGGCATCC
ALKBH5-R	GTCCACCGTGTGCTCGTTGTAC
KIAA1429-F	GCAACTTCAGGCATTAAGTTCA
KIAA1429-R	GTATTGCCTTGTCGAATCTGTC
METTL14-F	CAGGCTGGCTCACAGTTGGAC
METTL14-R	TTCCACCTCTTCCTCCACCTCTG
METTL3-F	CTTCAGCAGTTCCTGAATTAGC
METTL3-R	ATGTTAAGGCCAGATCAGAGAG
RBM15-F	GGCTGCCTGAGGAGAGTGGAG
RBM15-R	CGGCTACTGCTCAATTCTGGACTG
RBM15B-F	ATCTTTCAGAGTACGCTCAGAC
RBM15B-R	CTAGGATATGCATAGACGTGGG
WTAP-F	CTGACAAACGGACCAAGTAATG
WTAP-R	AAAGTCATCTTCGGTTGTGTTG
YTHDC1-F	AGTGACTCTGGTTCTGAATCTG
YTHDC1-R	CTGGTTTGATCTTTTCGGACAG
YTHDC2-F	GAGAATTGGGCTGTCGTTAAAG
YTHDC2-R	TGAAGCAGGATGAAATCGTACT
YTHDF2-F	ACTTCTCAGCATGGGGAAATAA
YTHDF2-R	TATTCATGCCAGGAGCCTTATT
YTHDF3-F	TCAACCACCACAACCACAGCAG
YTHDF3-R	TGAAGCACTGACAGGTACAACACC
IGF2BP1-R	GGGGTGGAATATTTCGGATTTG
IGF2BP1-F	GATGAAGGCCATCGAAACTTTC
IGF2BP2-F	GATGAACAAGCTTTACATCGGG
IGF2BP2-R	GATTTTCCCATGCAATTCCACT
IGF2BP3-F	GAGGCGCTTTCAGGTAAAATAG
IGF2BP3-R	AATGAGGCGGGATATTTCGTAT
YTHDF1-F	ATGACAATGACTTTGAGCCCTA
YTHDF1-R	AGGGAGTAAGGAAATCCAATGG
HNRNPA2B1-F	GCTTAAGCTTTGAAACCACAGA
HNRNPA2B1-F	GCTTAAGCTTTGAAACCACAGA
HNRNPC-F	ACAGATCCTCGCTCCATGAACTCC
HNRNPC-R	TTCTGCCATCCTCTCCTGCTACAG
GAPDH-F	CTGCACCACCAACTGCTT
GAPDH-R	TTCTGGGTGGCAGTGATG

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

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

clustering methods use the proportion of ambiguous clustering (PAC) to infer optimal K (K-means) in order to identify and classify patients for further analysis (Lock and Dunson, 2013). Cluster analysis was performed using the ConsensusClusterPlus R package with cycle computation for 1000 times to ensure the stability and reliability of the classification (Wilkerson and Hayes, 2010). The Kaplan–Meier method was used for the OS analysis in different clusters.

Lasso Regression for PC Tissues

Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. The best subset selection and the connections between lasso coefficient estimates can be identified to construct a prognostic model (Alhamzawi and Ali, 2018). Lasso regression was constructed to examine the relationship between gene signatures and PC risk. Further, clinical characteristics associated with OS were analyzed in patients with PC using Cox regression (including univariate and multivariate models) and the Kaplan–Meier method to evaluate the availability of the prognostic model. Pheatmap R package was used to correlate clinical data with the risk score (high or low).

Gene Set Enrichment Analysis for KIAA1429, HNRNPC, and IGF2BP2 in PC Tissues

Gene set enrichment analysis (GSEA) is widely used to analyze genome or proteome data, linking disease phenotypes with many different functional gene sets. The 178 patients with PC were divided into high expression groups and low expression groups according to the median expression values of KIAA1429, HNRNPC, and IGF2BP2. Two groups of TCGA data were analyzed by GSEA. Gene set enrichment analysis was also conducted in different sample risk groups based on the LASSO regression model. The 178 patients with PC were divided into high risk score group and low risk score group according to the median value of risk score.

Transient Transfection and Cell Proliferation Assay

The cells Mia-PaCa-2 and BXPC-3 were seeded in 6-well plates at 30–50% density. Transient transfection was performed with Lipo-fectamine 3000 reagents according to the manufacturer's instructions (Invitrogen, United States). For all the experiments, cells were collected at 24–48 h after transfection. After transfection, the cells were seeded in 96-well plates and cultured for 1–3 days according to 5000/well. On the indicated days, the CCK8 reagent (Sigma, St. Louis, MO, United States) was added, and the cells were incubated for 2 h at 37°C. The absorbance at 450 nm for each sample was measured using a microplate reader of Bio-Tek ELx800 (United States). For the colony formation assay, After transfection for 48 h, cells were used to measure DNA synthesis with a Cell-LightTM EdU imaging detecting kit (RiboBio, Guangzhou, China) according to the manufacturer's instructions.

Statistical Analysis

Gene expression data of FPKM form is used as input. WilcoxTest is used to get the p value for different expression between different clusters. The relationships between clusters or different risk score groups were analyzed using the Chi-square test. In all cases, $p < 0.05$ was considered statistically significant. Spearman correlation coefficient was calculated for the molecular pairing between m^{6}A regulator genes. The student's t-test in SPSS 13.0 (SPSS Inc., Chicago, United States) was used to assess the expression differences between HDE-CT and PC cancer cells. Each experiment was repeated at least three times. Benjamini-Hochberg for multiple testing, and false discovery rate (FDR) were calculated to correct the p-value in GSEA.

RESULTS

Consensus Clustering for PC Tissues Based on the Expression of m6A Regulators

To determine whether the expression levels of m^{6}A regulators were associated with PC prognosis, the TCGA PC cohort was clustered into different groups by consensus expression of m^{6}A regulators with the ConsensusClusterPlus R package. Gene signatures of m^{6}A regulators in PC are shown in **Supplementary Table S1**. When the consensus matrix k value was equal to 2, there was no crossover between PC samples (**Figure 1A**, **Supplementary Figure S1** and **Supplementary Table S2**). The OS difference between different clusters was calculated by the Kaplan–Meier method and log-rank test (**Figure 1B** and **Supplementary Table S2**). A heatmap was generated to visualize the expression pattern of m^{6}A regulators between different clusters (**Figure 1C**). The expression levels of RBM15B ($p = 0.037$), HNRNPC ($p = 0.001$), METTL14 ($p = 0.007$), METTL3 ($p = 0.005$), YTHDC1 ($p = 0.049$), KIAA1429 ($p = 0.010$), ALKBH5 ($p = 3.50E-06$), YTHF2 ($p = 0.038$), HNRN p A2B1 ($p = 0.003$), IGF2BP1 ($p = 1.22E-11$), IGF2BP2 ($p = 1.10E-05$), and IGF2BP3 ($p = 2.34E-27$) showed a significant dysregulation in tumor samples between different clusters.

The Interaction and Correlation Among the m^{6}A Regulators

The relationship between m^{6}A regulators were further supported by the correlation analysis. Some highly correlated (|correlation coefficient| ≥ 0.5, $p < 0.05$) m^{6}A regulator pairs were identified, including IGF2BP2 and IGF2BP3, IGF2BP2 and ALKBH5, YTHDC1 and YTHDC2, YTHDC1 and METTL14, YTHDC1 and ZC3H13, YTHDC2 and METTL14, YTHDC2 and ZC3H13, YTHDC2 and YTHDF3, METTL14 and FTO, METTL1 and ZC3H13, METTL14 and YTHDF3, FTO and ZC3H13 (**Figure 2** and **Supplementary Table S3**). The interactions among the 19 m^{6}A regulators are shown in **Figure 3A**. All m^{6}A regulators have interactions in the same network. The results of the interaction network showed that IGF2BP1 and IGF2BP3, WTAP and KIAA1429, HNRNPC and HNRNPA2B1, WTAP and ZC3H13, METTL14 and METTL3, KIAA1429 and ZC3H13, METTL14

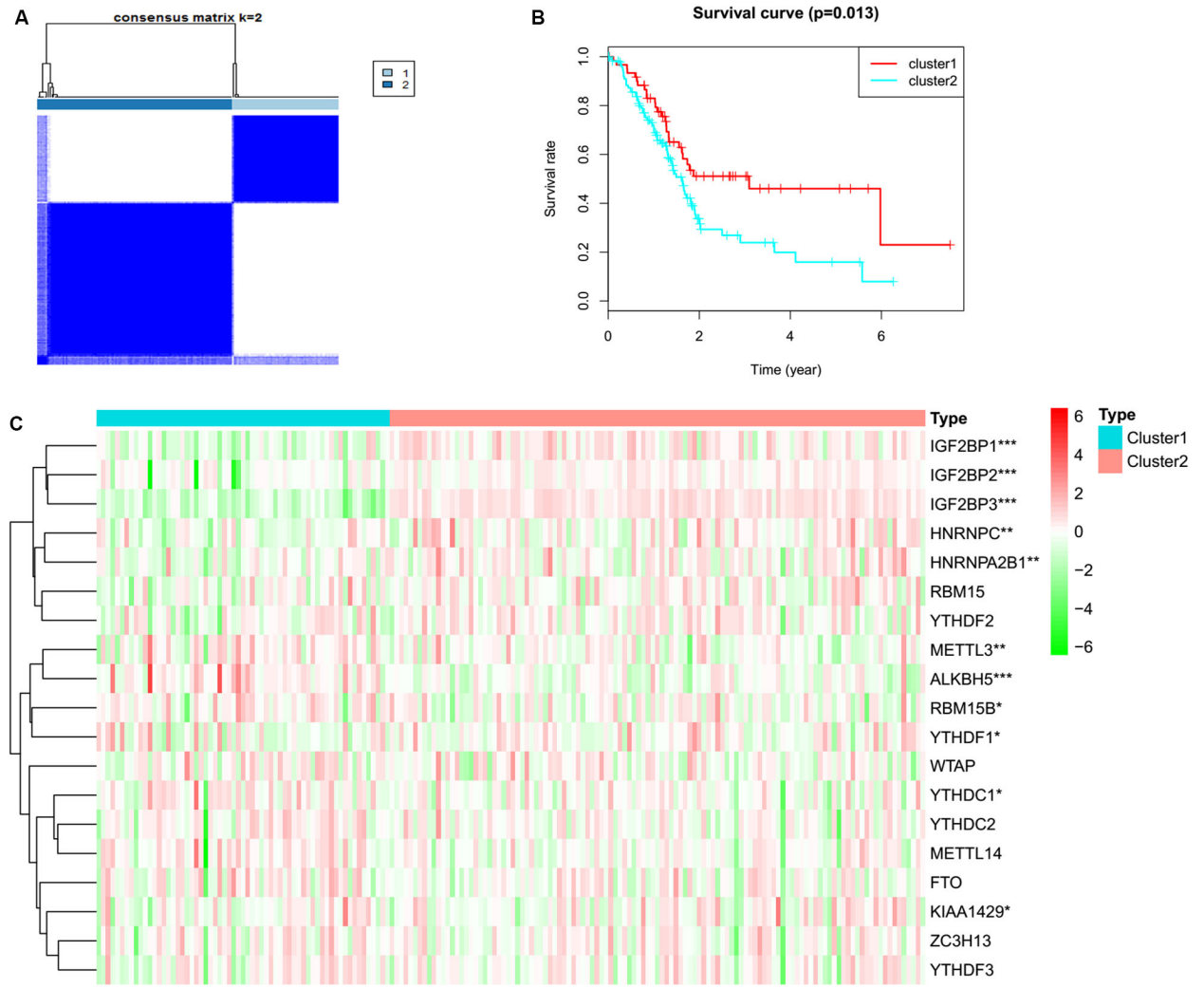


FIGURE 1 | Consensus clustering and heatmap. **(A)** Consensus clustering for PC tissues based on the expression of m^{6}A regulators (k = 2). **(B)** The overall survival in cluster 1 was significantly shorter than that in cluster 2. **(C)** Expression differences of m^{6}A RNA methylation regulators in pancreatic cancer based on TCGA data. Red and green represent relatively high or low expression, respectively. *p < 0.05, **p < 0.01, and ***p < 0.001.

and WTAP, WTAP and METTL3, METTL14 and KIAA1429, METTL3 and KIAA1429 have high combined score (>0.99).

Gene Signature of m^{6}A Regulators in PC Cell Lines

The expression of m^{6}A regulators, including the m^{6}A methyltransferases, the demethylases, and the m^{6}A-binding proteins were analyzed by qRT-PCR in the PC cell lines, Mia-PaCa-2 and BXPC-3, and the control cell line HDE-CT. The results showed that some m^{6}A regulators were differentially expressed in PC and control cell lines (**Figure 3B**).

Lasso Regression Identified the Six-Gene Signature Prognostic Model

In order to determine the optimal prognostic model, lasso regression was performed using the glmnet R package. Lasso regression is a generalized linear model, and the adjustment degree of lasso regression complexity is controlled by lambda. The optimal six-gene signature prognostic model was identified when log (lambda) was between −2 and −3 (**Supplementary Figures S2A,B**), where the coefficient of KIAA1429 was 0.28, the coefficient of HNRNPC was 0.34, the coefficient of METTL3 was −0.11, the coefficient of YTHDF1 was −0.37, the coefficient of IGF2BP2 was 0.28, and the coefficient of IGF2BP3 was 0.04. According to the median risk score, patients were divided into low- and high-risk groups (**Supplementary Table S4**). There was a significant difference in the OS rate between the two groups, and the OS rate of the high-risk group was significantly lower than that of the low-risk group (**Figure 4A**, p = 5.286e-04). A Receiver Operating Characteristiccurve (ROC) was used to evaluate the prediction efficiency of the prognostic signature. The prognostic signature model showed good prediction efficiency with the value of the area under the ROC curve (AUC) equal

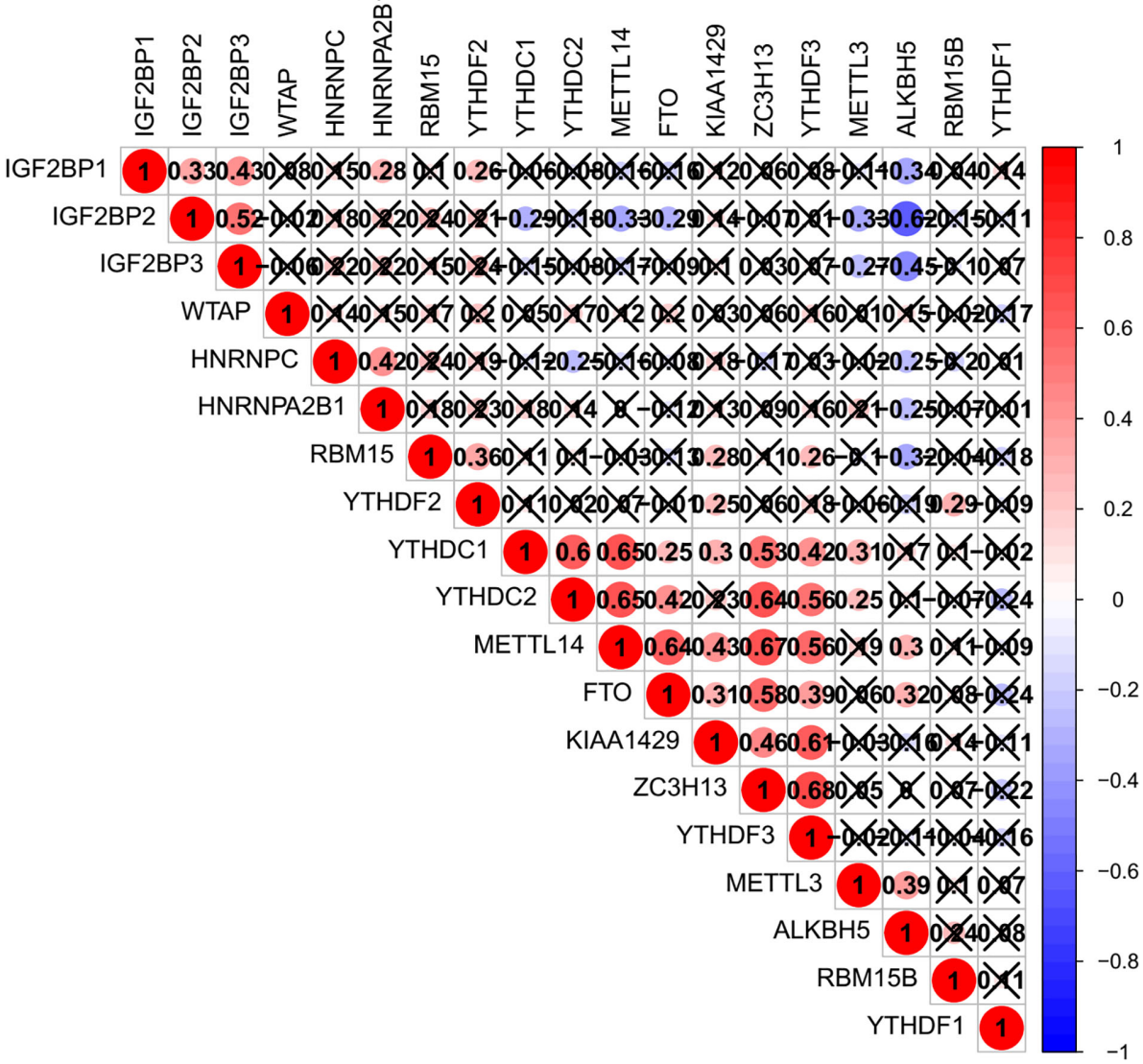


FIGURE 2 | Co-expression of m^{6}A regulator genes.

to 0.796 (**Figure 4B**). Additionally, the KM plotter showed that the six selected m^{6}A regulators were significantly related with OS according to OncoLnc[4] (**Figure 4C**). Importantly, the heat map shows the expression of the six selected m^{6}A regulators and clinicopathological variables in the high- and low-risk groups. Significant differences were found for the neoplasm histologic grade, pathologic M stage, pathologic N stage, pathologic stage, and vital status between high- and low-risk groups (**Figure 5** and **Supplementary Table S5**).

The Effect of m^{6}A Regulators on PC Prognosis

To investigate the effect of m^{6}A regulators on PC prognosis, we performed Cox univariate (**Figure 6A**) and multivariate analysis (**Figure 6B**). The six-gene signature was consistent with the single-factor analysis of genes using Cox regression. The univariate analysis revealed that age at initial pathologic diagnosis [hazard ratio (HR): 1.031; 95% confidence interval (CI): 1.009–1.053 p = 0.006], neoplasm histologic grade [hazard ratio (HR): 1.289; 95% confidence interval (CI): 1.000–1.662; p = 0.035], pathologic N stage [hazard ratio (HR): 631; 95% confidence interval (CI): 1.074–2.477; p = 0.022], pathologic T stage [hazard ratio (HR): 1.877; 95% confidence interval (CI): 1.174–3.002; p = 0.009] pathologic stage [hazard ratio (HR): 1.425; 95% confidence interval (CI): 0.983–2.064; p = 0.022], and risk score [hazard ratio (HR): 30.024; 95% confidence interval (CI): 8.884–171.416; p < 0.001] were correlated significantly with a poor OS (**Figure 6A**). The multivariate analysis revealed that age at initial pathologic diagnosis [hazard ratio (HR): 1.033; 95% confidence interval

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

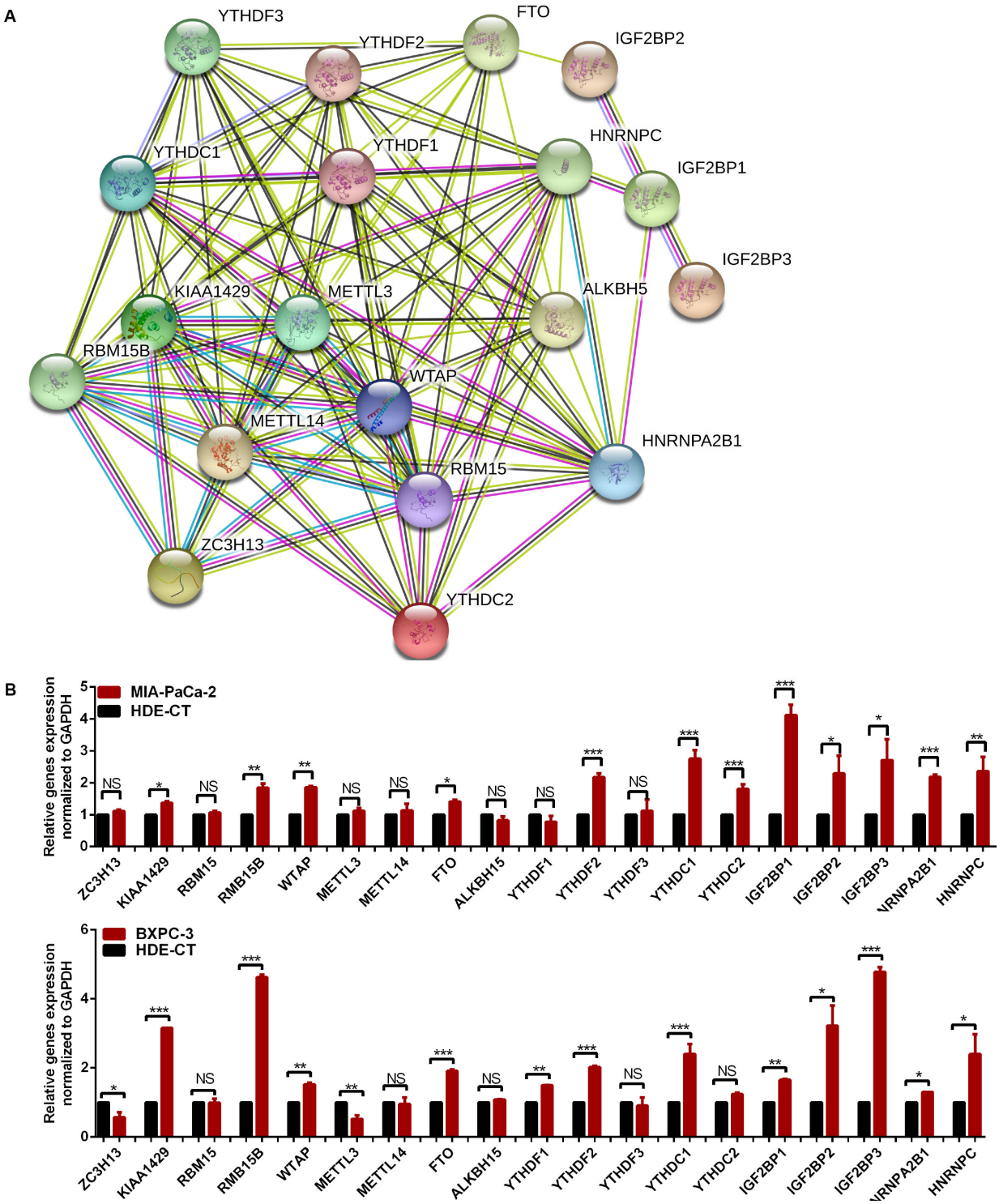


FIGURE 3 | The relationship of m^{6}A regulators in pancreatic cancer (PC) tissues. **(A)** Protein–protein interaction network of m^{6}A regulator proteins. **(B)** Gene signature of m^{6}A regulators in PC cell lines. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

(CI): 1.012–1.054; $p = 0.002$], pathologic N stage [hazard ratio (HR): 1.831; 95% confidence interval (CI): 1.045–3.210; $p = 0.035$], and risk score [hazard ratio (HR): 65.955; 95% confidence interval (CI): 13.308–326.879; $p < 0.001$] were correlated significantly with a poor OS (**Figure 6B**). The factor of risk score based on the optimal six-gene signature

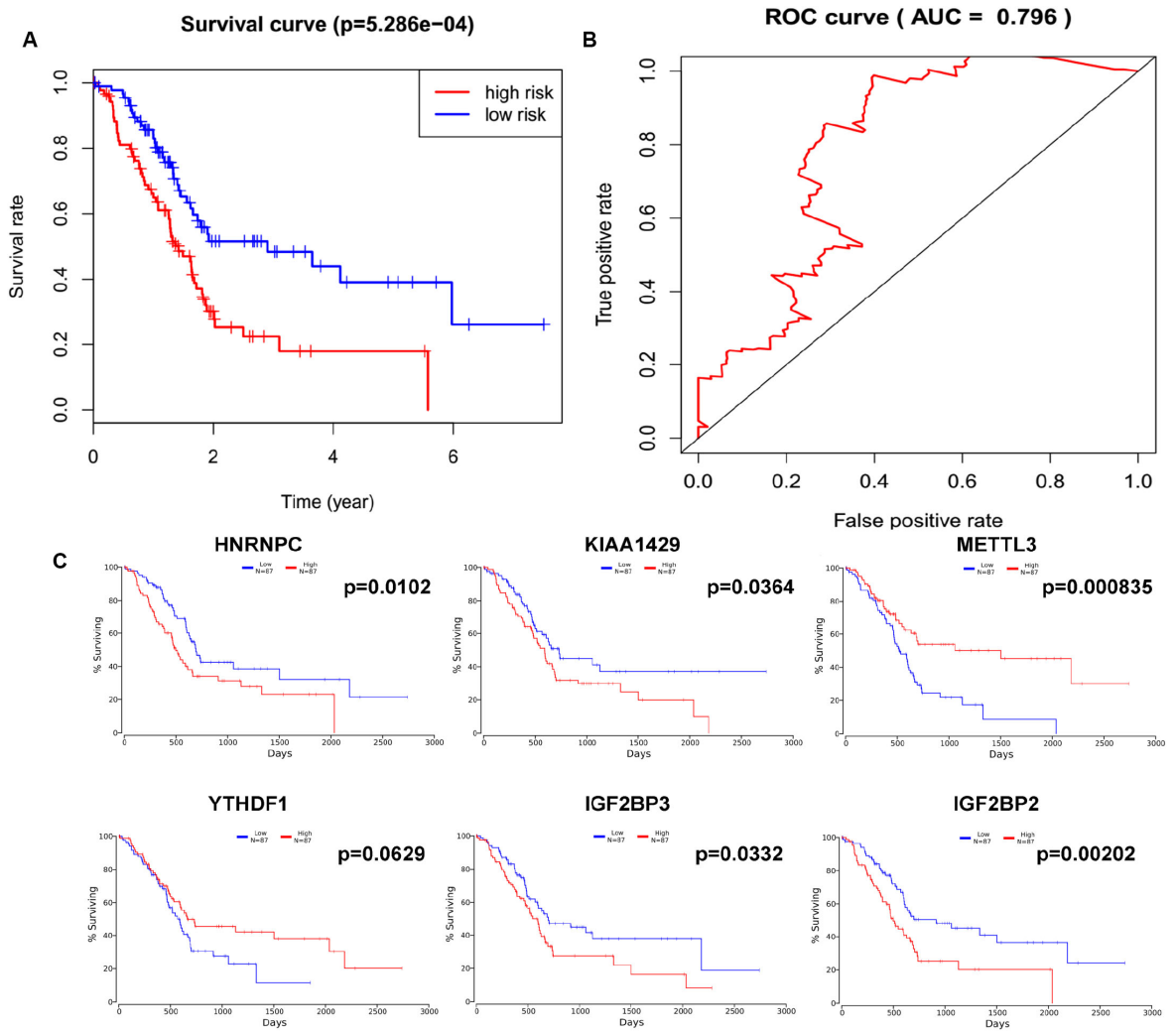


FIGURE 4 | Lasso regression identified a six-gene signature prognostic model. **(A)** Overall survival analysis of the high risk score and low risk score groups. **(B)** ROC curve was used to evaluate the prediction efficiency of the prognostic signature. **(C)** Kaplan–Meier (KM) survival curve of KIAA1429, HNRNPC, METTL3, YTHDF1, IGF2BP2, and IGF2BP3 in pancreatic cancer.

prognostic model was significant both at univariate and multivariate analyses.

GSEA Analysis Provided Insight Into Pathways of m^{6}A Regulators

According to the coefficient of m^{6}A regulators in the six-gene signature prognostic model and OS analysis, the GSEA result of HNRNPC showed that it is significantly related to phospholipase-c mediated cascade, type II diabetes mellitus, signaling by FGFR, downstream signaling of activated FGER, calcium signaling pathway, signaling by FGFR in disease, adipocytokine signaling pathway, vascular smooth muscle contraction, and metastasis. The GSEA result of IGF2BP2 showed that it is significantly related to metastasis, CREBBP targets, docetaxel resistance, hypoxia, BRCA1 targets, base excision repair, TAP63 pathway, etoposide sensitivity, epithelial mesenchymal transition, gemcitabine resistance, cisplatin resistance, gefitinib resistance, tumor differentiated well vs. poorly, and SFRP2 targets. The GSEA result of KIAA1429 showed that it is significantly related to CD5 targets, stemness, ubiquitin mediated proteolysis, YY1 targets, UV response via ERCC3, metastasis, EIF4 pathway, downregulation of SMAD2-SMAD3-SMAD4 transcriptional activity, EZH2 targets, ERBB1 receptor proximal pathway, BMI1 targets, signaling by hippo, and oncogenesis by Met (**Supplementary Table S6**). Some interesting pathways are shown in **Figure 7**. It's not containing GSEA analysis for METTL3, IGF2BP1, and IGF2BP3. We did it, but there were no significant results for METTL3, IGF2BP1, and IGF2BP3.

We conducted GSEA analysis in different sample risk score groups based on the LASSO regression model. The GSEA result showed that it is significantly related to cancer survival, oncogenesis by met, gemcitabine resistance, response to UV, HOXC6 targets cancer, recurrent liver cancer, WTAP targets,

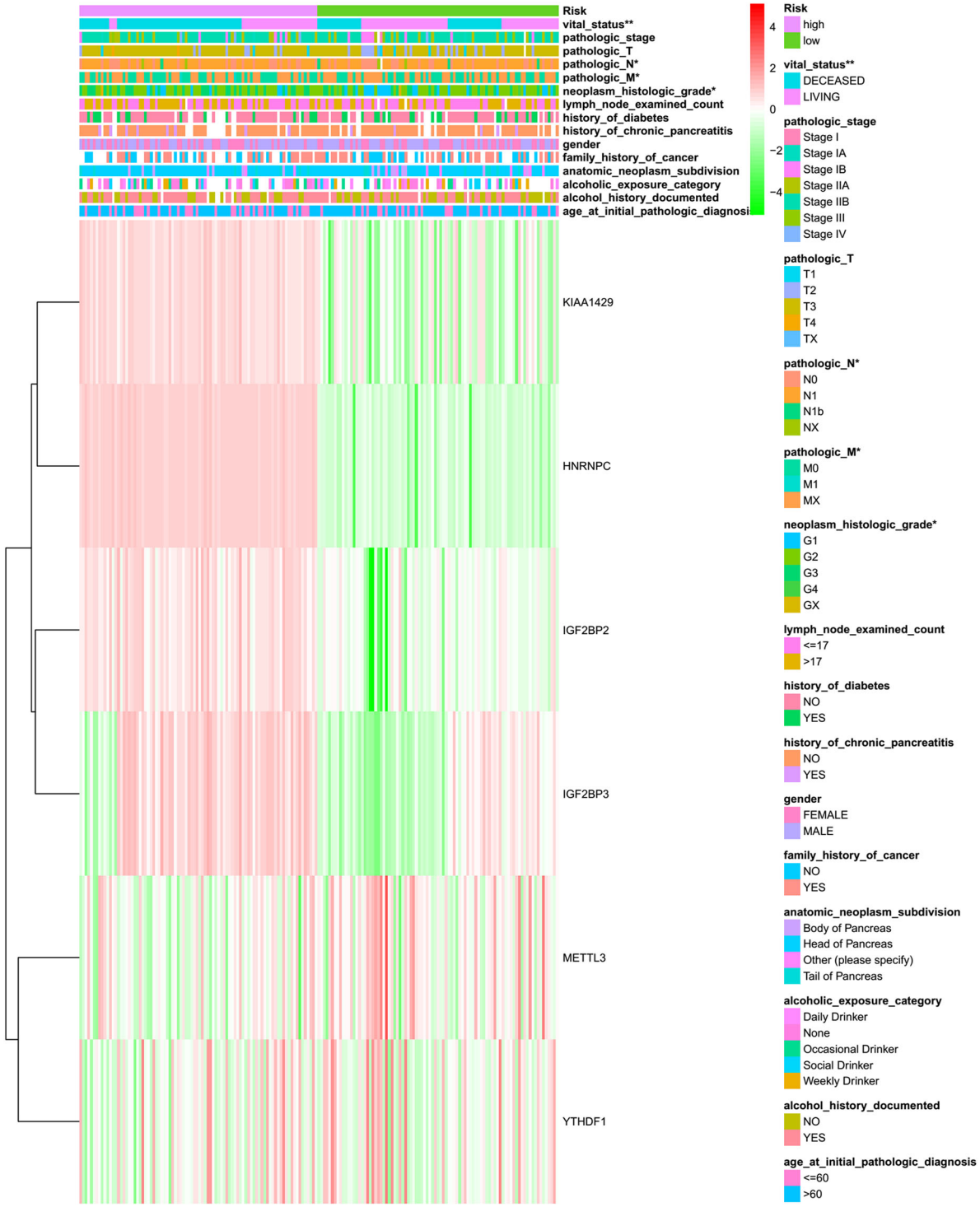


FIGURE 5 | The heatmap of sample risk groups and related pancreatic cancer clinical characteristics. Age at initial pathologic diagnosis (patients were aged 35–88), alcohol history documented (yes or no), alcoholic exposure category (daily drinker, weekly drinker, occasional drinker, social drinker, and non-drinker), anatomic neoplasm subdivision (body of pancreas, head of pancreas, tail of pancreas, and other parts), family history of cancer (yes or no), gender (male and female), history of chronic pancreatitis (yes or no), history of diabetes (yes or no), count of lymph nodes examined (from 1 to 57), neoplasm histologic grade (G1, G2, and G3), pathologic M (M represents tumor metastasis, including M0, M1,and MX), pathologic N (N represents tumor lymph node metastasis, including N0, N1, N2, and NX), pathologic T (T represents tumor size, including T1, T2, T3,T4, and TX), pathologic stage (Stages I, II, III, and IV), vital status (alive or dead). *$p < 0.05$ and **$p < 0.01$.

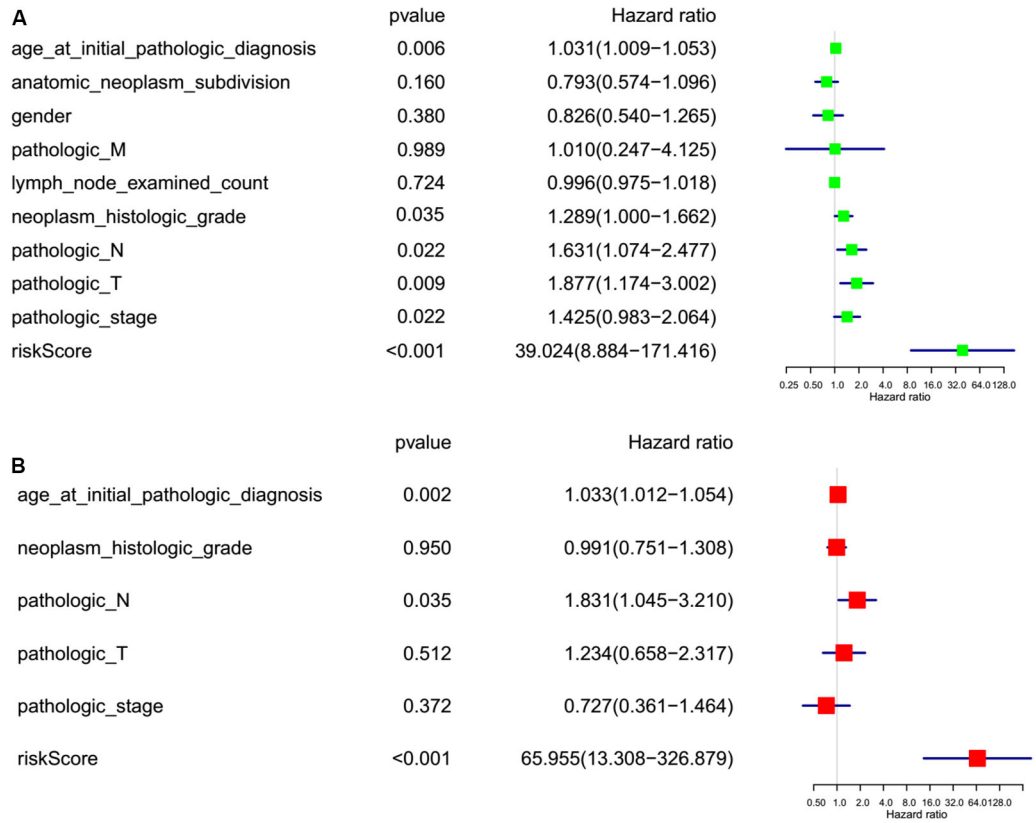


FIGURE 6 | Risk factor analyses for pancreatic cancer (PC). **(A)** Univariate analysis of risk factors for PC. **(B)** Multivariate analysis of risk factors for PC. Age at initial pathologic diagnosis (>60 vs. <60), anatomic neoplasm subdivision (body of pancreas, head of pancreas, tail of pancreas, and other parts), gender (male vs. female), count of lymph nodes examined (>17 vs. <17), neoplasm histologic grade (G1, G2, and G3), pathologic M (M represents tumor metastasis, including M0, M1,and MX), pathologic N (N represents tumor lymph node metastasis, including N0, N1, N2, and NX), pathologic T (T represents tumor size, including T1, T2, T3,T4, and TX), pathologic stage (Stages I, II, III, and IV), risk score (high risk score group vs. low risk score group).

tumor differentiated well vs. poorly, epithelial mesenchymal transition, hypoxia pathway, TGFB1 targets, cancer meta signature, and so on (**Supplementary Table S7**). Some interesting pathways are shown in **Supplementary Figure S3**, and those pathways closely related with tumorigenesis and development.

The Independent Verification by GEO

The different expression of m6A regulators between cancer tissue and normal tissue, including the m6A methyltransferases, the demethylases, and the m6A-binding proteins were analyzed based on the independent verification by GEO (**Supplementary Figure S4** and **Supplementary Table S8**). In view of some similarities of identified different genes in TCGA data and GEO data, it is believed that the prognostic m6A regulators might not just be due to chance. For example, the overlapping genes that are significant were including RBM15B, KIAA1429, ALKBH5, YTHDF1, IGF2BP 2/3, and HNRNPC. Furthermore, the testing dataset based on GEO showed the different expression of m6A regulators in PC and validate the reliability of the built model based on TCGA. The optimal six-gene signature prognostic model was validated. According to the median risk score, patients from GEO were divided into low- and high-risk score groups (**Supplementary Table S9**). There was a significant difference in the OS rate between the two groups, and the OS rate of the high-risk score group was significantly lower than that of the low-risk score group (**Supplementary Figure S5**, $p = 0.0012$).

Experimental Validation

The inhibition of KIAA1429, HNRNPC, and IGF2BP2, respectively, significantly suppressed the proliferation abilities of PC cells based on CCK8 (**Figure 8A**). The EdU assay further showed that KIAA1429, HNRNPC, and IGF2BP2 inhibitors reduced DNA replication in both Mia-PaCa-2 and BXPC-3 cells (**Figure 8B**).

DISCUSSION

Treatment for PC has improved considerably, for example surgery with high success and lower complication rate is better than ever before, novel drug combinations (chemotherapy, target therapy, and immunotherapy) have been shown to improve survival rate, and advances in radiation therapy have achieved less toxicity; however, many researchers are focused on early diagnosis and prompt treatment as PC is still one of the deadliest

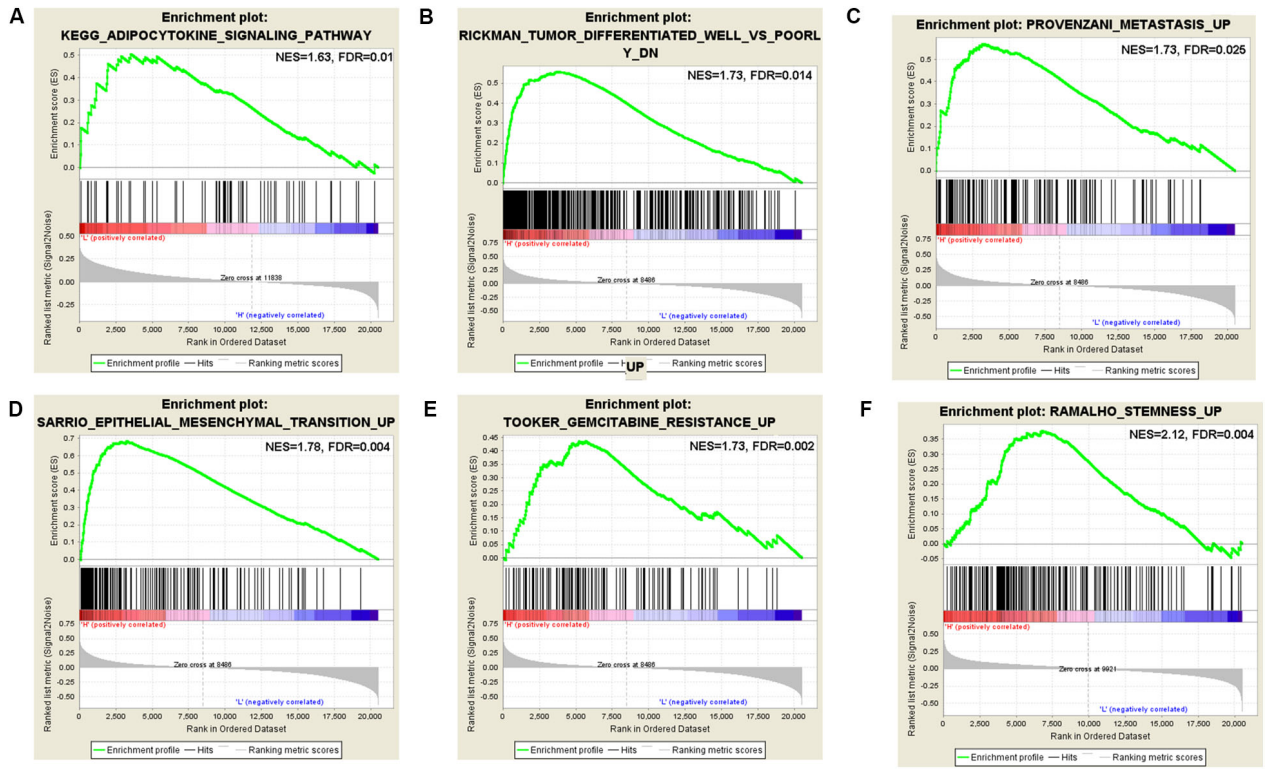


FIGURE 7 | Gene set enrichment analysis (GSEA) for KIAA1429, HNRNPC, and IGF2BP2. **(A)** GSEA enriched the adipocytokine signaling pathway of HNRNPC. **(B)** GSEA enriched the well vs. poorly differentiated tumor pathway of IGF2BP2. **(C)** GSEA enriched the tumor metastasis pathway of IGF2BP2. **(D)** GSEA enriched the epithelial mesenchymal transition pathway of IGF2BP2. **(E)** GSEA enriched the gemcitabine resistance pathway of IGF2BP2. **(F)** GSEA enriched the stemness pathway of KIAA1429.

solid malignancies (Chu et al., 2017). The development of multi-omics has given us a better understanding of the fundamental genetics of PC. These advancements provide hope, but the survival rate of patients with PC is still poor (Cid-Arregui and Juarez, 2015). Biological functions of m^6A were not studied extensively until around 2012, when major progress was made in the transcriptome profiling of m^6A through antibody-based immunoprecipitation and high-throughput sequencing (Gan et al., 2019). Moreover, m^6A regulators were shown to be related with the development of cancer (He et al., 2019). The process of m^6A modification is reversible through the regulation of m^6A methyltransferases, demethylases, and binding proteins. A series of m^6A regulators have been described (Dominissini et al., 2012), including ZC3H13, RBM15/15B, KIAA1429, METTL14, YTHDC1/2, WTAP, METTL3, FTO, ALKBH5, YTHDF1/2/3, HNRNPA2B1, IGF2BP1/2/3, and HNRNPC (Lee et al., 2014). Therefore, it is necessary to explore the influence of m^6A regulators on PC.

Recent studies have found that the m^6A modification, when the related enzyme is abnormal, plays various roles in a series of human diseases such as neurological disorders, cancer, and embryonic developmental retardation (Wu et al., 2019). Both coding RNAs and some non-coding RNAs, such as lncRNA, microRNA, tRNA, and rRNA and RNA splice body, were regulated by an m^6A modification before and after transcription (Yen et al., 2019). N6-methyladenosine modification is closely related to the metabolic processes of RNAs, for example, RNA processing, RNA transfer from the nucleus to the cytoplasm, RNA translation, RNA decay, and the biogenesis of RNA (Liang et al., 2020). The dynamic modification of RNA as a way of regulating genetic information is a new field of study, so there is still a lot of work to be done to understand the underlying mechanisms. Recently, a number of studies have found that m^6A modifications are associated with cancer, having functions such as helping tumor stem cells to self-renew, promoting the growth and proliferation of cancer cells, and resisting radiotherapy or chemotherapy (Mao et al., 2019). All this evidence indicates that m^6A regulators may be a target for cancer treatment (Bi et al., 2019; Ianniello et al., 2019). The regulation of m^6A modifications is a collaboration between methyltransferases, demethylases, and binding proteins. The functions of these proteins in stem cell differentiation, stomach cancer, lung cancer, osteosarcoma, liver cancer, colorectal cancer, leukemia, neuroblastoma, renal cell carcinoma, and breast cancer have been extensivelyreported (Feng et al., 2019; Jin et al., 2019). For example, YTHDF1-deficient mice show an elevated antigen-specific $CD8^+$ T cell antitumor response compared with wild-type mice, which indicated that durable neoantigen-specific immunity is regulated by mRNA m^6A methylation through the m^6A-binding protein YTHDF1 (Han et al., 2019). It was

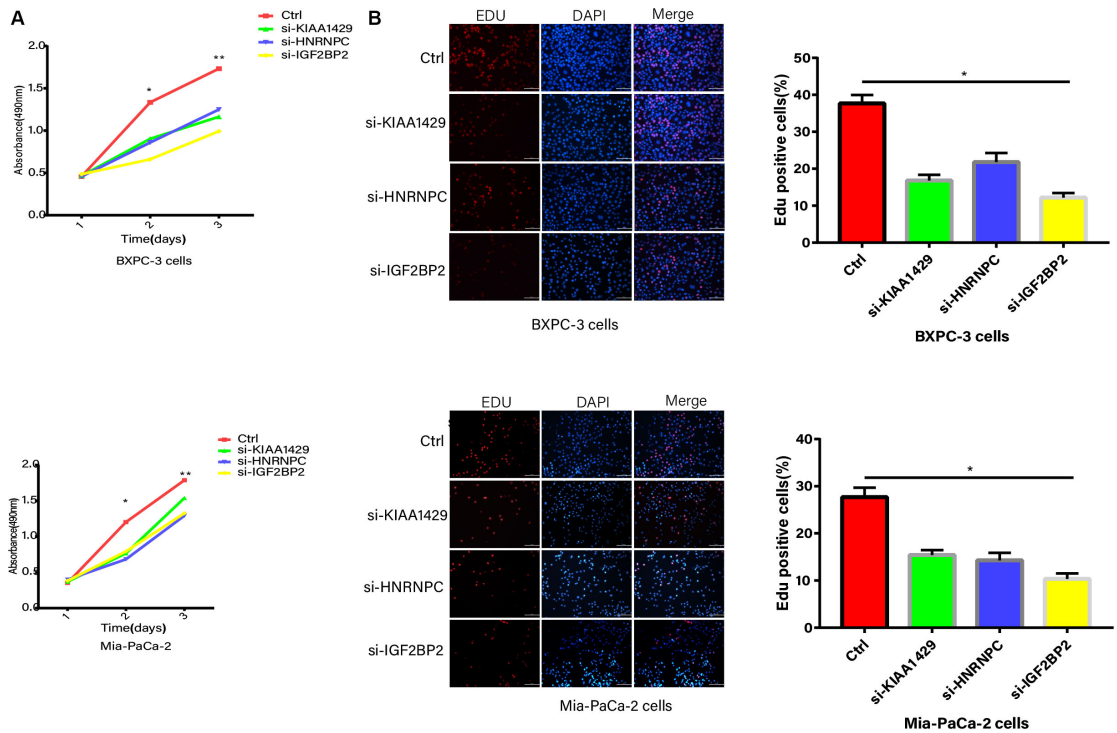


FIGURE 8 | Proliferation abilities validation. **(A)** CCK8 analysis of growth curve in Mia-PaCa-2 and BXPC-3 cells transfected with control and si-RNAs. **(B)** The EdU assay of DNA replication in both Mia-PaCa-2 and BXPC-3 cells transfected with control and si-RNAs. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

also reported that some drugs with antitumor activity, such as R-2-hydroxyglutarate (R-2HG), inhibited proliferation/survival of FTO-high cancer cells via targeting FTO/m^{6}A/MYC/CEBPA signaling (Su et al., 2018). METTL3 which is independently of METTL14, binds to chromatin, and locates the transcription initiation site of active genes. The promoter bounding METTL3 induces m6A modification in the coding region of the relevant mRNA transcription and enhances its translation by alleviating ribosomal stalling. The gene regulated by METTL3 in this way is necessary for acute myeloid leukemia, suggesting that METTL3 may be a therapeutic target for acute myeloid leukemia (Barbieri et al., 2017). The researchers also found that that m^{6}A mRNA demethylation by FTO increases melanoma growth and decreases response to anti-PD-1 blockade immunotherapy (Yang et al., 2019). Knockdown of FTO increased the methylation of m6A in the intrinsic genes of key primary melanoma cells such as PD-1 (PDCD1), CXCR4, SOX10, and so on, leading to increased attenuation of RNA in m6A reader YTHDF2, suggesting that FTO inhibition combined with anti-PD-1 blocking may abate the resistance of melanoma immunotherapy (Yang et al., 2019).

TCGA, a landmark cancer genomics project, described more than 20,000 primary cancers at the molecular level and matched normal samples of 33 cancer types. TCGA generated more than 2.5 petabytes of genome, epigenome, transcriptome and proteome data. The data has already lead to improvements in our ability to diagnose, treat, and prevent cancer (Blum et al., 2018). N6-methyladenosine RNA methylation regulators can lead to malignant progression and impact the prognosis of many kinds of cancer based on the TCGA database. For example, the lasso Cox regression model was applied to identify three m^{6}A regulators in bladder cancer. The risk signature was constructed as follows: 0.164FTO - (0.081YTHDC1 + 0.032WTAP), which indicated that the three m^{6}A regulators identified might be promising prognostic biomarkers to guide personalized treatment for patients with bladder cancer (Chen et al., 2019). Another study has built up a robust m^{6}A regulators-based molecular signature that predicts the prognosis of patients with head and neck squamous cell carcinoma with high accuracy, which might provide important guidance for therapeutic strategies. The results revealed that the expression levels of YTHDF1, METTL3, KIAA1429, YTHDF2, RBM15, METTL14, ALKBH5, FTO, WTAP, and HNRNPC were significantly upregulated in head and neck squamous cell carcinoma samples, while YTHDC2 was remarkably downregulated (Zhao and Cui, 2019). In addition, a study identified two subgroups of gastric cancer (cluster1 and 2) by applying consistency clustering to the m6A regulators. Compared with the cluster1 subgroup, the prognosis of the cluster2 subgroup was poorer, and most of the 13 major m6A regulators were highly expressed in cluster2. This finding provides clues to understand epigenetic modifications of RNA in gastric cancer (Su et al., 2019). However, the prognostic role of m^{6}A regulators in PC is poorly understood. In the present study, we are the first to show, by applying consensus clustering to m^{6}A regulators, that there are two subgroups of PC (cluster1 and 2). The cluster2 subgroup correlates with a poorer prognosis, which suggests that m^{6}A regulators may be promising

prognostic biomarkers for PC. Furthermore, the lasso regression analysis identified a six-gene signature prognostic model (KIAA1429, HNRNPC, METTL3, YTHDF1, IGF2BP2, and IGF2BP3). These results agree with the results of previous studies. The major function of IGF2BP2 is to regulate cell metabolism (Huang et al., 2018). However, our results suggest that lncRNA DANCR is a novel target for IGF2BP2 through m^{6}A modification in PC, and that it promotes cancer stemness-like properties and PC pathogenesis. Mechanistically, IGF2BP2 serves as a reader for the m^{6}A modified DANCR (at adenosine 664), and the definite interaction site provides a novel target for PC therapy (Hu et al., 2019).

We did GSEA for KIAA1429, HNRNPC, and IGF2BP2. Many enrichment pathways were significantly related to cancer pathogenesis. We focused on some important events, for example, pathways of oncogenesis by Met, EIF4 pathway, downregulation of SMAD2-SMAD3-SMAD4 transcriptional activity, EZH2 targets, stemness, well vs. poorly differentiated tumor, epithelial mesenchymal transition, UV response via ERCC3, and metastasis. The identified pathways were consistent with reported data. The importance of m^{6}A in the response to ultraviolet DNA damage was demonstrated, and the findings support that m^{6}A RNA serves as a beacon for the selective, rapid recruitment of DNA polymerase κ to damage sites to facilitate repair and cell survival (Xiang et al., 2017). Meanwhile, many studies show that m^{6}A-related genes work on stemness regulation in tumor relapse. For example, METTL3 was identified as a regulator for terminating murine naïve pluripotency. METTL3 knockout preimplantation epiblasts lead to early embryonic lethality, because it is associated with stability of key naïve pluripotency-promoting transcripts (Geula et al., 2015). Epithelial mesenchymal transition (EMT), as an important cellular program during tumor migration, invasion and metastasis, is also regulated by m^{6}A mRNA methylation. N6-methyladenosine-sequencing and functional studies confirm that YTHDF1 mediates m^{6}A-increased translation of Snail mRNA (a key transcription factor of EMT) (Lin et al., 2019). Interestingly, the process of m^{6}A mRNA methylation was also regulated by cytokines (Li et al., 2017). The TGFβ pathway plays roles in disease through the intracellular effectors SMAD2 and SMAD3. SMAD2/3 promotes binding of the m^{6}A methyltransferase complex to a subset of transcripts involved in early cell fate decisions. These aspects of m^{6}A methyltransferase signaling could have far-reaching implications in the treatment of many cancers (Bertero et al., 2018).

In conclusion, this study is the first to identify and profile the gene signatures of clinical trait-related m^{6}A regulatory genes in PC. We also developed a six-gene signature prognostic model, which might play a crucial role in determining the clinical progression of PC. With the development of m^{6}A-sequencing and methylated RNA immunoprecipitation, m^{6}A regulatory genes might serve as promising molecular biomarkers for monitoring many kinds of cancers and providing important guidance for selecting therapeutic strategies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in The Cancer Genome Atlas (TCGA), https://cancergenome.nih.gov/.

AUTHOR CONTRIBUTIONS

JH and LL designed the experiments. ZW and HL performed the experiments, data collection, and analysis. HZ wrote the manuscript. LL revised the manuscript and provided financial support. All authors approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00522/full#supplementary-material

FIGURE S1 | Consensus clustering for pancreatic cancer (PC) tissues. **(A)** Consensus clustering for PC tissues based on the expression of m^{6}A regulators (k = 3). **(B)** Consensus clustering for PC tissues based on the expression of m^{6}A regulators (k = 4). **(C)** Consensus clustering cumulative distribution function (CDF) for k = 2–4. **(D)** Relative change in area under CDF curve for k = 2–4.

FIGURE S2 | (A,B) Lasso regression complexity was controlled by lambda using the glmnet R package.

FIGURE S3 | Gene set enrichment analysis (GSEA) for high risk score vs. low risk score group. **(A)** GSEA enriched the liver cancer survival pathway. **(B)** GSEA enriched the cancer meta signature. **(C)** GSEA enriched the tumor differentiated well vs. poorly pathway. **(D)** GSEA enriched the epithelial mesenchymal transition pathway. **(E)** GSEA enriched the oncogenesis by MET. **(F)** GSEA enriched the hypoxia pathway.

FIGURE S4 | Expression differences of m^{6}A RNA methylation regulators in pancreatic cancer based on GEO data. Red and green represent relatively high or low expression, respectively. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

FIGURE S5 | Lasso regression validation. **(A)** Lasso regression complexity was controlled by lambda using the glmnet R package. **(B)** Overall survival analysis of the high risk score and low risk score group based on GEO data.

TABLE S1 | Gene signatures of m^{6}A regulators in pancreatic cancer.

TABLE S2 | Sample cluster based on m^{6}A regulators in pancreatic cancer.

TABLE S3 | PPI network of those m^{6}A regulators in pancreatic cancer.

TABLE S4 | Lasso regression was constructed examining the relationship between gene signature and pancreatic cancer risk.

TABLE S5 | The clinical features of pancreatic cancer and clusters based on consensus clustering method.

TABLE S6 | Gene sets enriched in pancreatic cancer by GSEA analysis based on expression of m^{6}A regulators (IGF2BP2,KIAA1429, and HNRNPC).

TABLE S7 | Gene sets enriched in pancreatic cancer by GSEA analysis in different sample risk groups based on the LASSO regression model.

TABLE S8 | Gene signatures of m6A regulators and different expression in pancreatic cancer using GEO.

TABLE S9 | Lasso regression was constructed examining the relationship between gene signature and pancreatic cancer risk verified by GEO data.

REFERENCES

Alhamzawi, R., and Ali, H. (2018). The Bayesian adaptive lasso regression. *Math. Biosci.* 303, 75–82. doi: 10.1016/j.mbs.2018.06.004

Ansari, D., Gustafsson, A., and Andersson, R. (2015). Update on the management of pancreatic cancer: surgery is not enough. *World J. Gastroenterol.* 21, 3157–3165. doi: 10.3748/wjg.v21.i11.3157

Barbieri, I., Tzelepis, K., Pandolfini, L., Shi, J., Millan-Zambrano, G., Robson, S. C., et al. (2017). Promoter-bound METTL3 maintains myeloid leukaemia by m(6)A-dependent translation control. *Nature* 552, 126–131. doi: 10.1038/nature24678

Bertero, A., Brown, S., Madrigal, P., Osnato, A., Ortmann, D., Yiangou, L., et al. (2018). The SMAD2/3 interactome reveals that TGFbeta controls m(6)A mRNA methylation in pluripotency. *Nature* 555, 256–259. doi: 10.1038/nature25784

Bi, Z., Liu, Y., Zhao, Y., Yao, Y., Wu, R., Liu, Q., et al. (2019). A dynamic reversible RNA N(6) -methyladenosine modification: current status and perspectives. *J. Cell. Physiol.* 234, 7948–7956. doi: 10.1002/jcp.28014

Blum, A., Wang, P., and Zenklusen, J. C. (2018). SnapShot: TCGA-analyzed tumors. *Cell* 173:530. doi: 10.1016/j.cell.2018.03.059

Chen, M., Nie, Z. Y., Wen, X. H., Gao, Y. H., Cao, H., and Zhang, S. F. (2019). m6A RNA methylation regulators can contribute to malignant progression and impact the prognosis of bladder cancer. *Biosci. Rep.* 39:BSR20192892.

Chen, X., Xu, M., Xu, X., Zeng, K., Liu, X., Sun, L., et al. (2020). METTL14 suppresses CRC progression via regulating N6-methyladenosine-dependent primary miR-375 processing. *Mol. Ther.* 28, 599–612. doi: 10.1016/j.ymthe.2019.11.016

Chu, L. C., Goggins, M. G., and Fishman, E. K. (2017). Diagnosis and detection of pancreatic cancer. *Cancer J.* 23, 333–342.

Cid-Arregui, A., and Juarez, V. (2015). Perspectives in the treatment of pancreatic adenocarcinoma. *World J. Gastroenterol.* 21, 9297–9316.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 485, 201–206. doi: 10.1038/nature11112

Feng, Y., Li, Y., Li, L., Wang, X., and Chen, Z. (2019). Identification of specific modules and significant genes associated with colon cancer by weighted gene coexpression network analysis. *Mol. Med. Rep.* 20, 693–700.

Gan, H., Hong, L., Yang, F., Liu, D., Jin, L., and Zheng, Q. (2019). [Progress in epigenetic modification of mRNA and the function of m6A modification]. *Sheng Wu Gong Cheng Xue Bao* 35, 775–783.

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., et al. (2015). Stem cells. *m*6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. *Science* 347, 1002–1006.

Han, D., Liu, J., Chen, C., Dong, L., Liu, Y., Chang, R., et al. (2019). Anti-tumour immunity controlled through mRNA m(6)A methylation and YTHDF1 in dendritic cells. *Nature* 566, 270–274. doi: 10.1038/s41586-019-0916-x

He, L., Li, H., Wu, A., Peng, Y., Shu, G., and Yin, G. (2019). Functions of N6-methyladenosine and its role in cancer. *Mol. Cancer* 18:176.

He, Y., Hu, H., Wang, Y., Yuan, H., Lu, Z., Wu, P., et al. (2018). ALKBH5 inhibits pancreatic cancer motility by decreasing long non-coding RNA KCNK15-AS1 methylation. *Cell. Physiol. Biochem.* 48, 838–846. doi: 10.1159/000491915

Hong, K. (2018). Emerging function of N6-methyladenosine in cancer. *Oncol. Lett.* 16, 5519–5524.

Hu, X., Peng, W. X., Zhou, H., Jiang, J., Zhou, X., Huang, D., et al. (2019). IGF2BP2 regulates DANCR by serving as an N6-methyladenosine reader. *Cell Death Differ.* [Epub ahead of print].

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., et al. (2018). Recognition of RNA N(6)-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nat. Cell Biol.* 20, 285–295. doi: 10.1038/s41556-018-0045-z

Ianniello, Z., Paiardini, A., and Fatica, A. (2019). N(6)-methyladenosine (m(6)A): a promising new molecular target in acute myeloid leukemia. *Front. Oncol.* 9:251. doi: 10.3389/fonc.2019.00251

Ilic, M., and Ilic, I. (2016). Epidemiology of pancreatic cancer. *World J. Gastroenterol.* 22, 9694–9705.

Jin, D., Guo, J., Wu, Y., Du, J., Yang, L., Wang, X., et al. (2019). m(6)A mRNA methylation initiated by METTL3 directly promotes YAP translation and increases YAP activity by regulating the MALAT1-miR-1914-3p-YAP axis to induce NSCLC drug resistance and metastasis. *J. Hematol. Oncol.* 12:135.

Koh, C., Goh, Y. T., and Goh, W. (2019). Atlas of quantitative single-base-resolution N(6)-methyl-adenine methylomes. *Nat. Commun.* 10:5636.

Lee, M., Kim, B., and Kim, V. N. (2014). Emerging roles of RNA modification: m(6)A and U-tail. *Cell* 158, 980–987. doi: 10.1016/j.cell.2014.08.005

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

Liang, Z., Riaz, A., Chachar, S., Ding, Y., Du, H., and Gu, X. (2020). Epigenetic Modifications of mRNA and DNA in plants. *Mol. Plant* 13, 14–30. doi: 10.1016/j.molp.2019.12.007

Lin, X., Chai, G., Wu, Y., Li, J., Chen, F., Liu, J., et al. (2019). RNA m(6)A methylation regulates the epithelial mesenchymal transition of cancer cells and translation of Snail. *Nat. Commun.* 10:2065.

Liu, Z. X., Li, L. M., Sun, H. L., and Liu, S. M. (2018). Link between m6A modification and cancers. *Front. Bioeng. Biotechnol.* 6:89. doi: 10.3389/fbioe.2018.00089

Lock, E. F., and Dunson, D. B. (2013). Bayesian consensus clustering. *Bioinformatics* 29, 2610–2616. doi: 10.1093/bioinformatics/btt425

Mao, Y., Dong, L., Liu, X. M., Guo, J., Ma, H., Shen, B., et al. (2019). m(6)A in mRNA coding regions promotes translation via the RNA helicase-containing YTHDC2. *Nat. Commun.* 10:5332.

Mcguigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., and Mccain, R. S. (2018). Pancreatic cancer: a review of clinical diagnosis, epidemiology, treatment and outcomes. *World J. Gastroenterol.* 24, 4846–4861. doi: 10.3748/wjg.v24.i43.4846

Meyer, K. D., and Jaffrey, S. R. (2017). Rethinking m(6)a readers, writers, and erasers. *Annu. Rev. Cell Dev. Biol.* 33, 319–342. doi: 10.1146/annurev-cellbio-100616-060758

Mohammed, S., Van Buren, G. N., and Fisher, W. E. (2014). Pancreatic cancer: advances in treatment. *World J. Gastroenterol.* 20, 9354–9360.

Niu, Y., Lin, Z., Wan, A., Chen, H., Liang, H., Sun, L., et al. (2019). RNA N6-methyladenosine demethylase FTO promotes breast tumor progression through inhibiting BNIP3. *Mol. Cancer* 18:46.

Su, R., Dong, L., Li, C., Nachtergaele, S., Wunderlich, M., Qing, Y., et al. (2018). R-2HG exhibits anti-tumor activity by targeting FTO/m(6)A/MYC/CEBPA signaling. *Cell* 172, 90–105.

Su, Y., Huang, J., and Hu, J. (2019). m(6)A RNA methylation regulators contribute to malignant progression and have clinical prognostic impact in gastric cancer. *Front. Oncol.* 9:1038. doi: 10.3389/fonc.2019.01038

Taketo, K., Konno, M., Asai, A., Koseki, J., Toratani, M., Satoh, T., et al. (2018). The epitranscriptome m6A writer METTL3 promotes chemo- and radioresistance in pancreatic cancer cells. *Int. J. Oncol.* 52, 621–629.

Welinsky, S., and Lucas, A. L. (2017). Familial pancreatic cancer and the future of directed screening. *Gut Liver* 11, 761–770. doi: 10.5009/gnl16414

Wilkerson, M. D., and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26, 1572–1573. doi: 10.1093/bioinformatics/btq170

Wolfgang, C. L., Herman, J. M., Laheru, D. A., Klein, A. P., Erdek, M. A., Fishman, E. K., et al. (2013). Recent progress in pancreatic cancer. *CA Cancer J. Clin.* 63, 318–348.

Wu, F., Cheng, W., Zhao, F., Tang, M., Diao, Y., and Xu, R. (2019). Association of N6-methyladenosine with viruses and related diseases. *Virol. J.* 16:133.

Xia, T., Wu, X., Cao, M., Zhang, P., Shi, G., Zhang, J., et al. (2019). The RNA m6A methyltransferase METTL3 promotes pancreatic cancer cell proliferation and invasion. *Pathol. Res. Pract.* 215:152666. doi: 10.1016/j.prp.2019.152666

Xiang, Y., Laurent, B., Hsu, C. H., Nachtergaele, S., Lu, Z., Sheng, W., et al. (2017). RNA m(6)A methylation regulates the ultraviolet-induced DNA damage response. *Nature* 543, 573–576. doi: 10.1038/nature21671

Yang, S., Wei, J., Cui, Y. H., Park, G., Shah, P., Deng, Y., et al. (2019). m(6)A mRNA demethylase FTO regulates melanoma tumorigenicity and response to anti-PD-1 blockade. *Nat. Commun.* 10:2782.

Yen, C. J., Yang, S. T., Chen, R. Y., Huang, W., Chayama, K., Lee, M. H., et al. (2019). Hepatitis B virus X protein (HBx) enhances centrosomal P4.1-associated protein (CPAP) expression to promote hepatocarcinogenesis. *J. Biomed. Sci.* 26:44.

Yu, J., Chen, M., Huang, H., Zhu, J., Song, H., Zhu, J., et al. (2018). Dynamic m6A modification regulates local translation of mRNA in axons. *Nucleic Acids Res.* 46, 1412–1423. doi: 10.1093/nar/gkx1182

Zaccara, S., Ries, R. J., and Jaffrey, S. R. (2019). Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* 20, 608–624. doi: 10.1038/s41580-019-0168-5

Zhao, X., and Cui, L. (2019). Development and validation of a m(6)A RNA methylation regulators-based signature for predicting the prognosis of head and neck squamous cell carcinoma. *Am. J. Cancer Res.* 9, 2156–2169.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 July 2020
doi: 10.3389/fgene.2020.00674

Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data

Xiaojun Liu[1†], Lianxing Li[2†], Lihong Peng[1], Bo Wang[3], Jidong Lang[3], Qingqing Lu[3], Xizhe Zhang[2], Yi Sun[2], Geng Tian[3], Huajun Zhang[4] and Liqian Zhou[1]**

[1] School of Computer Science, Hunan University of Technology, Zhuzhou, China, [2] Chifeng Municipal Hospital, Chifeng, China, [3] Genesis Beijing Co., Ltd., Beijing, China, [4] College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua, China

Patients with carcinoma of unknown primary (CUP) account for 3–5% of all cancer cases. A large number of metastatic cancers require further diagnosis to determine their tissue of origin. However, diagnosis of CUP and identification of its primary site are challenging. Previous studies have suggested that molecular profiling of tissue-specific genes could be useful in inferring the primary tissue of a tumor. The purpose of this study was to evaluate the performance somatic mutations detected in a tumor to identify the cancer tissue of origin. We downloaded the somatic mutation datasets from the International Cancer Genome Consortium project. The random forest algorithm was used to extract features, and a classifier was established based on the logistic regression. Specifically, the somatic mutations of 300 genes were extracted, which are significantly enriched in functions, such as cell-to-cell adhesion. In addition, the prediction accuracy on tissue-of-origin inference for 3,374 cancer samples across 13 cancer types reached 81% in a 10-fold cross-validation. Our method could be useful in the identification of cancer tissue of origin, as well as the diagnosis and treatment of cancers.

Keywords: somatic mutation, machine learning, random forest, patients with carcinoma of unknown primary, tissue of origin

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Fengbiao Mao,
University of Michigan, United States
Yongcui Wang,
Northwest Institute of Plateau Biology (CAS), China

****Correspondence:***
Huajun Zhang
huajunzhang@zjnu.cn
Liqian Zhou
zhoulq11@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: *24 February 2020*
Accepted: *02 June 2020*
Published: *14 July 2020*

Citation:
Liu X, Li L, Peng L, Wang B, Lang J, Lu Q, Zhang X, Sun Y, Tian G, Zhang H and Zhou L (2020) Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data. Front. Genet. 11:674. doi: 10.3389/fgene.2020.00674

INTRODUCTION

Researches have proved that hepatitis C virus (HCV) and hepatitis B virus (HBV) are the main causes of liver cancer, and liver cancer can be primary or metastatic, where metastatic liver cancer accounts for 5% (Hu and Ludgate, 2007; Lin et al., 2013). Studies have shown that Epstein–Barr virus (EBV) infection is one of the important causes of nasopharyngeal carcinoma (Hui et al., 1998; Krishna et al., 2006). Tsai et al. (1996) carried out numerous experiments and found that EBER1 expression is abundant in primary nasopharyngeal carcinoma, which may metastasize to lymph nodes. Numerous studies have shown that *Helicobacter pylori* (HP) is associated with gastric cancer (Farinati et al., 1993; Gonzaga et al., 2002; Geng and Zhang, 2017). Gastric cancer is one of the most common malignant diseases in the world, where metastasis often occurs, and there are histological differences between primary and metastatic gastric cancer (Wang et al., 2008). In most cases, viruses are a major cause of cancer. Metastatic cancer brings great adversity to the follow-up

diagnosis and treatment. Some biomarkers are related with metastasis of cancer. Chen et al. (2016) carried out researches on the differential expressed proteins and found two biomarkers related with lung adenocarcinoma. Xiuping et al. (2016) found that NTN4 is associated with breast cancer cell migration and invasion via regulation of epithelial–mesenchymal transition–related biomarkers. Differentially expressed genes between metastatic tissue samples and nonmetastatic tissue samples can be molecular biomarkers for gastric cancer metastasis (Li et al., 2016).

In clinical diagnosis, metastatic cancer is a common phenomenon and a great challenge for determination of the primary site of a tumor. In all cases of cancer diagnoses, 3–5% of patients are confirmed as carcinoma of unknown primary (CUP) (Shaw et al., 2007). Cases of CUP are usually heterogeneous and can make diagnosis and treatment of pathological and clinical cases difficult (Rizwan and Zulfiqar, 2010). In the recent years, immunohistochemistry was a crucial method for classification of cancer and identify the primary site of a tumor and made great contributions to CUP identification (Huebner et al., 2007; Voigt, 2008; Centeno et al., 2010; Kandalaft and Gown, 2015; Janick et al., 2018). However, immunohistochemistry is labor-intensive and applicable to small-scale sample data, and it is difficult to overcome the bottleneck in classification accuracy.

Computed tomography (CT) and positron emission tomography are good medical imaging tools for identifying cancer tissue and predicting the primary site of a tumor (Fencl et al., 2007; Kwee et al., 2010; Fu et al., 2019). CT and PET identify tumors with an accuracy of 20–27% and 24–40%, respectively (Ambrosini et al., 2006). Obviously, the prediction performance is too poor to reach a satisfying degree. Moreover, medical images usually generate large-scale data, and limitations of image processing technology also bring about great difficulty in application. Identification of tissue origin utilizing medical imaging still remains conservative.

Recently, the use of molecular profiling has become a popular method to infer the primary site of a tumor. In addition, the combination of machine learning method and molecular profiling has been proven to be better than the utilization of immunohistochemistry for undifferentiated or poorly differentiated tumors (Oien and Dennis, 2012). Combination of methylation and copy number variation can contribute to cancer classification and tissue origin identification (Hoadley et al., 2014). Küsters-Vandevelde et al. (2017) suggested that metastatic behavior of a tumor is closely associated with specific copy number variations, as the methylation profile of meningeal melanocytic metastatic tumor was found to be similar as to that of the primary site. Although metastasis of cancer occurs, methylation and copy number variation are still in accordance with those of the primary origin. Particularly, gene expression data were frequently used in identification of the primary site of a tumor (Erlander et al., 2004; Qu et al., 2007; Gross-Goupil et al., 2012; Greco, 2013; Hainsworth et al., 2013). Erlander et al. (2011) proved that the value of gene expression detected in metastasis is the same as that detected in the primary origin when metastatic cancer occurs. Centeno et al. (2010) carried out numerous experiments with the proposed hybrid model, which utilized immunohistochemistry and gene expression profiling, and obtained classifier accuracies of 89, 88, and 75% for cross-validation datasets, independent test sets, and institutional independent test sets, respectively. Rosenwald et al. (2010) gained an accuracy of 85% on prediction of the primary site of cancer with the use of the KNN algorithm and micro-RNA quantitative reverse transcription–polymerase chain reaction test. Bloom et al. (2004) explored a method based on the artificial neural network with gene expression profiling to infer the tumor origin and thus aid in making a correct pathological diagnosis.

Somatic mutation data can also be utilized to identify tissue origin. Sheffield et al. (2016) revealed that mutation of the *IDH1* gene in patients with cholangiocarcinoma can be used to infer the primary site of the malignant tumor. Dietlein and Eschner (2014) and Lawrence et al. (2014) explored a method using mutation spectra to predict the primary site of cancer and obtained a specificity of 79%, showing that the enrichment of mutation in tumor-specific genes can be effective for primary tissue tracing. Relatively comprehensive research was conducted by Marquard et al. (2016), using somatic mutation data, base substitution frequency, trinucleotide base substitution frequency, and copy number aberrations. The best results with accuracy of 87.6% were obtained using a combination of copy number status, trinucleotide context base substitution frequencies, and somatic point mutations. However, it is complicated that each cancer was trained with a classifier. Moreover, the best performance was achieved using three molecular profiling, in which data collection is challenging.

Use of copy number variation, methylation, and gene expression to predict the primary site of a tumor has been a hot spot. However, research of predicting tissue origin using mutation data has made little progress. This current study proposed a new method using somatic mutation data to

TABLE 1 | Distribution of samples with 13 cancers.

Cancer Types		Samples	
Type	**Abbreviation**	**Primary**	**Metastasis**
Biliary tract cancer	BTCA	310	0
Chronic myeloid disorders	CMDI	136	0
Colorectal cancer	COCA	317	4
Gastric cancer	GACA	708	0
Brain lower-grade glioma	LGG	508	0
Liver cancer	LIRI	258	0
Soft tissue cancer	LMS	67	0
Malignant lymphoma	MALY	152	89
Skin cancer	MELA	183	0
Nasopharyngeal cancer	NACA	21	0
Pancreatic endocrine neoplasms	PAEN	87	2
Renal cancer	RECA	432	0
Skin adenocarcinoma	SKCA	52	48
Total		3,219	155

predict the primary site of cancer. The International Cancer Genome Consortium (ICGC), together with machine learning methods could improve the predictive performance. Here, the random forest algorithm (Sandri and Zuccolotto, 2006) was selected as a gene selection algorithm, and the logistic regression algorithm (Zhang et al., 2014; Pranoto et al., 2015) was utilized to establish a classifier. Performance evaluation was judged by metrics, such as accuracy and specificity. Functional annotation and enrichment of specific gene set were settled by R packages.

MATERIALS AND METHODS

Data Preparation

We downloaded the somatic mutation data from ICGC database version 28[1]. The format of the gene name was Ensembl Gene ID. A total of 19,730 samples were obtained. We duplicated the samples according to chromosomal features, locus in chromosome, donor-id, and gene-affected. Sample data of 57 types of cancer were preliminarily extracted. Somatic mutation data cannot identify the primary site of some cancers. Samples with primary and metastasis of 13 types of common cancers were used to predict tissue origin (**Table 1**). Data were further filtered, and we generated an $S \times G$ matrix, where S represents the number of samples and G represents the number of genes included.

[1]https://dcc.icgc.org/releases/release_28/

Feature Selection

As mutation detection of tissue-specific gene is time consuming and costly, a balance between performance and number of genes used is necessary. Existing feature selection algorithms such as Lasso and Principal Component Analysis (PCA) (Malhi and Gao, 2005; Muthukrishnan and Rohini, 2016) have been largely used as a tool for feature processing. Here, we used the random

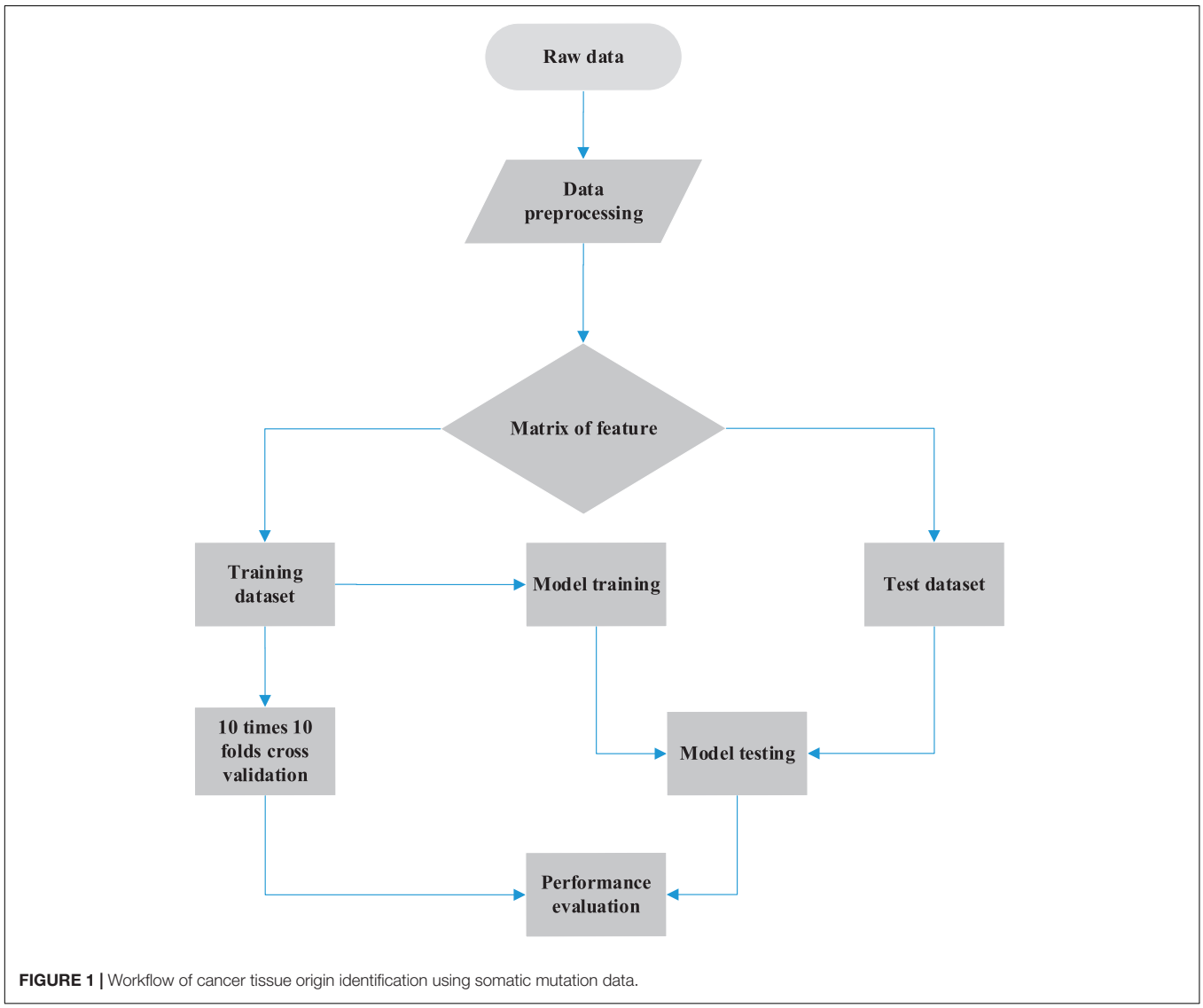


FIGURE 1 | Workflow of cancer tissue origin identification using somatic mutation data.

forest algorithm (Breiman, 2001; Sandri and Zuccolotto, 2006) for feature selection. It can handle a large number of input features and assess their importance, and its learning process is fast. Tt is a type of ensemble learning algorithm and is composed of a CART (classification and regression tree). In each tree, $\sqrt{g}$ was used, where g denotes the gene number. The process of feature selection was explained by the splitting of nodes. The Gini index was used to determine which feature should be selected as most important and was calculated by the following Eq. 1:

$$\text{Gini}\,(p) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2 \tag{1}$$

In a node, p denotes the weight represented as frequencies of cancers, k denotes the total cancer number, and the weight of k-th cancer is denoted by p_k. We calculated feature importance scores of the i-th gene in a node, which was represented by a decrease in the Gini index value. This was calculated by Eq. 2:

$$VIM_{im}^{(Gini)} = GI_m - GI_l - GI_r \tag{2}$$

M was used as the set of nodes. m denotes a node in M. Thereafter, we selected the i-th gene for splitting. Split subnodes have their own Gini index. We calculated the Gini index before node m splitting, denoted as $VIM_{im}^{(Gini)}$, and Gini index of two subnodes

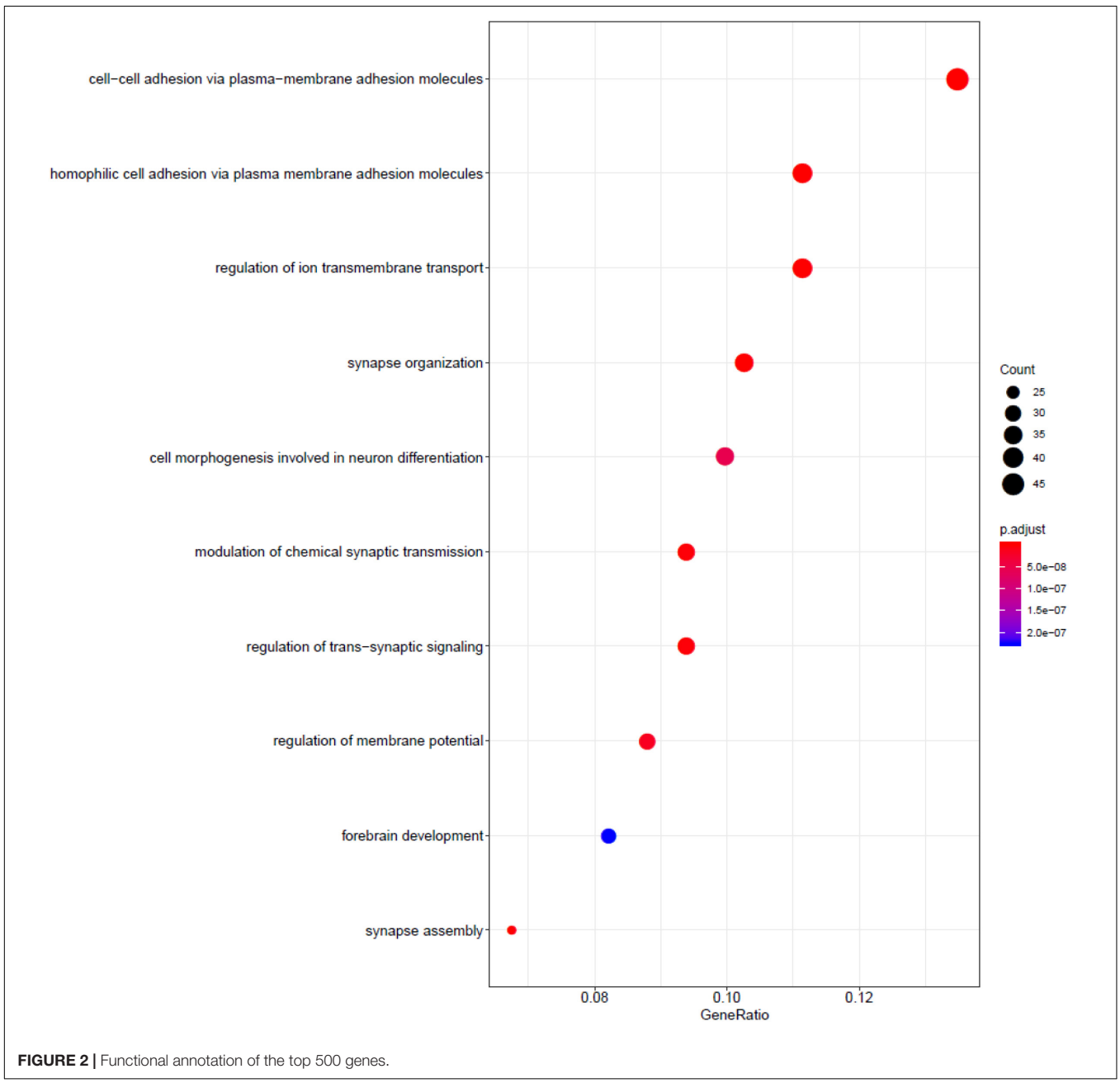


FIGURE 2 | Functional annotation of the top 500 genes.

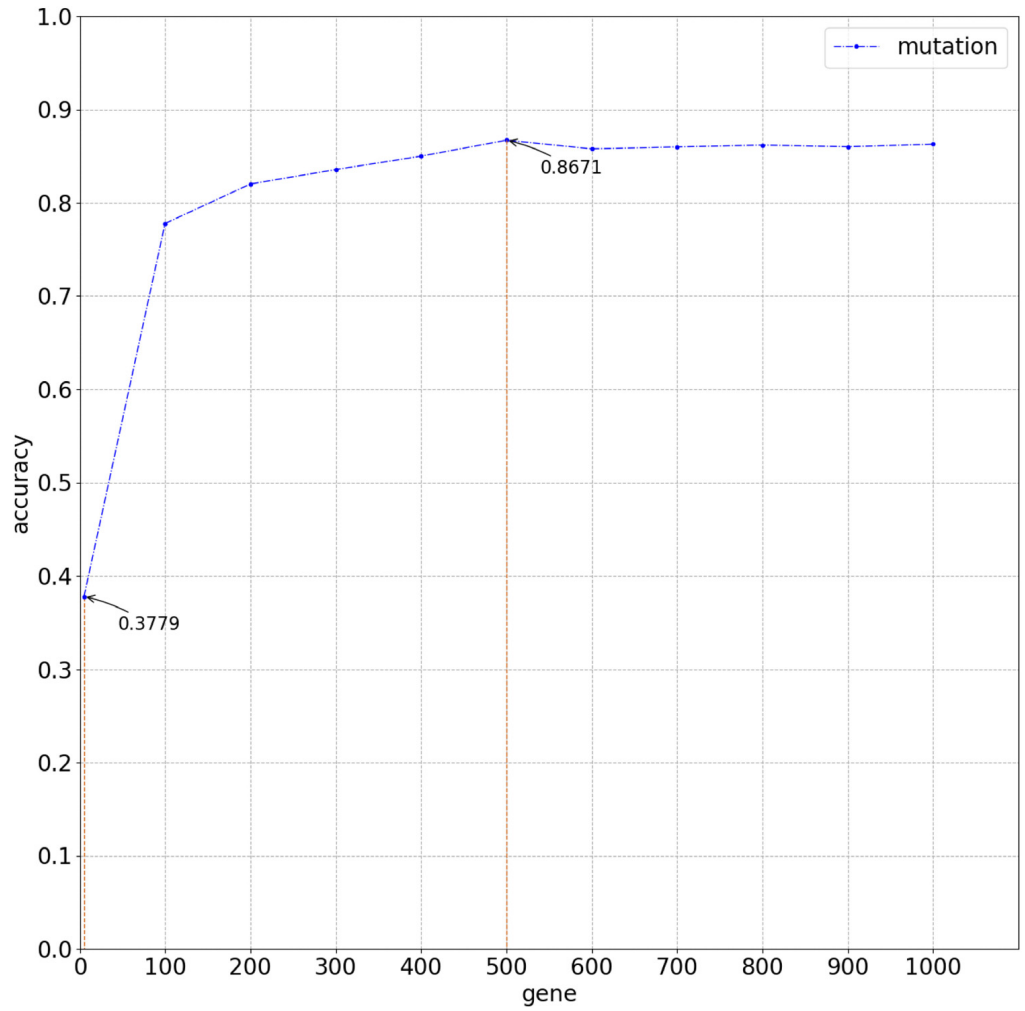


FIGURE 3 | Overall average accuracy using logistic regression classifier with 10-time 10-fold cross-validation.

TABLE 2 | Performance metric of training dataset using top 500 genes.

Cancer	Precision	Recall	F1 score	Support	Specificity
BTCA	0.6288	0.6331	0.6308	245.0000	0.9626
CMDI	0.9789	0.8921	0.9335	114.0000	0.9991
COCA	0.6479	0.7700	0.7036	250.0000	0.9573
GACA	0.8556	0.8265	0.8408	570.0000	0.9627
LGG	0.9315	0.9178	0.9246	400.0000	0.9883
LIRI	0.9390	0.9362	0.9376	207.0000	0.9949
LMS	0.9981	0.9796	0.9888	54.0000	1.0000
MALY	0.9944	0.9893	0.9918	196.0000	0.9996
MELA	0.8851	0.9147	0.8996	143.0000	0.9934
NACA	0.9018	0.6118	0.7275	17.0000	0.9996
PAEN	0.7150	0.7738	0.7431	80.0000	0.9906
RECA	0.9294	0.9077	0.9184	339.0000	0.9901
SKCA	0.9251	0.8259	0.8726	85.0000	0.9978
Average	0.8552	0.8445	0.8548	2, 700.0000	0.9883
Accuracy	0.8671	NA	NA	NA	NA

after splitting denoted as GI_l and GI_r, respectively. The bigger the $VIM_{im}^{(Gini)}$, the more important the i-th gene.

$$VIM_{ti}^{(Gini)} = \sum_{m \in M} VIM_{im}^{(Gini)} \tag{3}$$

T was used as a set of trees, and t denotes the t-th tree. Equation 3 shows the importance of the i-th gene in the t-th tree. Thereafter, we calculated the importance of the ith gene in all trees, and the sum was represented as Eq. 4 depicts:

$$VIM_{i}^{(Gini)} = \sum_{t=1}^{T} VIM_{ti}^{(Gini)} \tag{4}$$

Finally, importance scores of each feature in all trees were averaged by weight. The importance of each gene sorted according to their averaged importance score. We selected the top n genes by importance score, where n was a flexible value set to obtain the best classification performance.

Logistic Regression Classifier

We used the logistic regression algorithm to construct a classifier (Zhang et al., 2014; Pranoto et al., 2015). Logistic regression uses the sigmoid function to represent the probability of a sample being labeled as a certain category, and prediction of tissue origin can be explained as a one-to-many classification problem. In this process, one type of cancer was considered positive, and other types were considered negative. Thereafter, the probability of the sample was predicted as one cancer type and other cancer types, respectively. After a series of similar procedures, we obtained the probability of a sample being predicted as each cancer. The prediction function was calculated by Eq. 5:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (5)$$

where $h_\theta(x)$ denotes the probability of a sample being predicted as one cancer type (positive), or other cancer types, (negative). θ^T is a matrix of parameters used to determine the best model. θ is computed by the negative log-likelihood loss function. The loss function was calculated by Eq. 6:

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} log h_\theta\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad (6)$$

where $log h_\theta\left(x^{(i)}\right)$ and $\log\left(1 - h_\theta\left(x^{(i)}\right)\right)$ represent the log loss when a sample is labeled positive and negative, respectively. m represents the number of samples, and n denotes the number

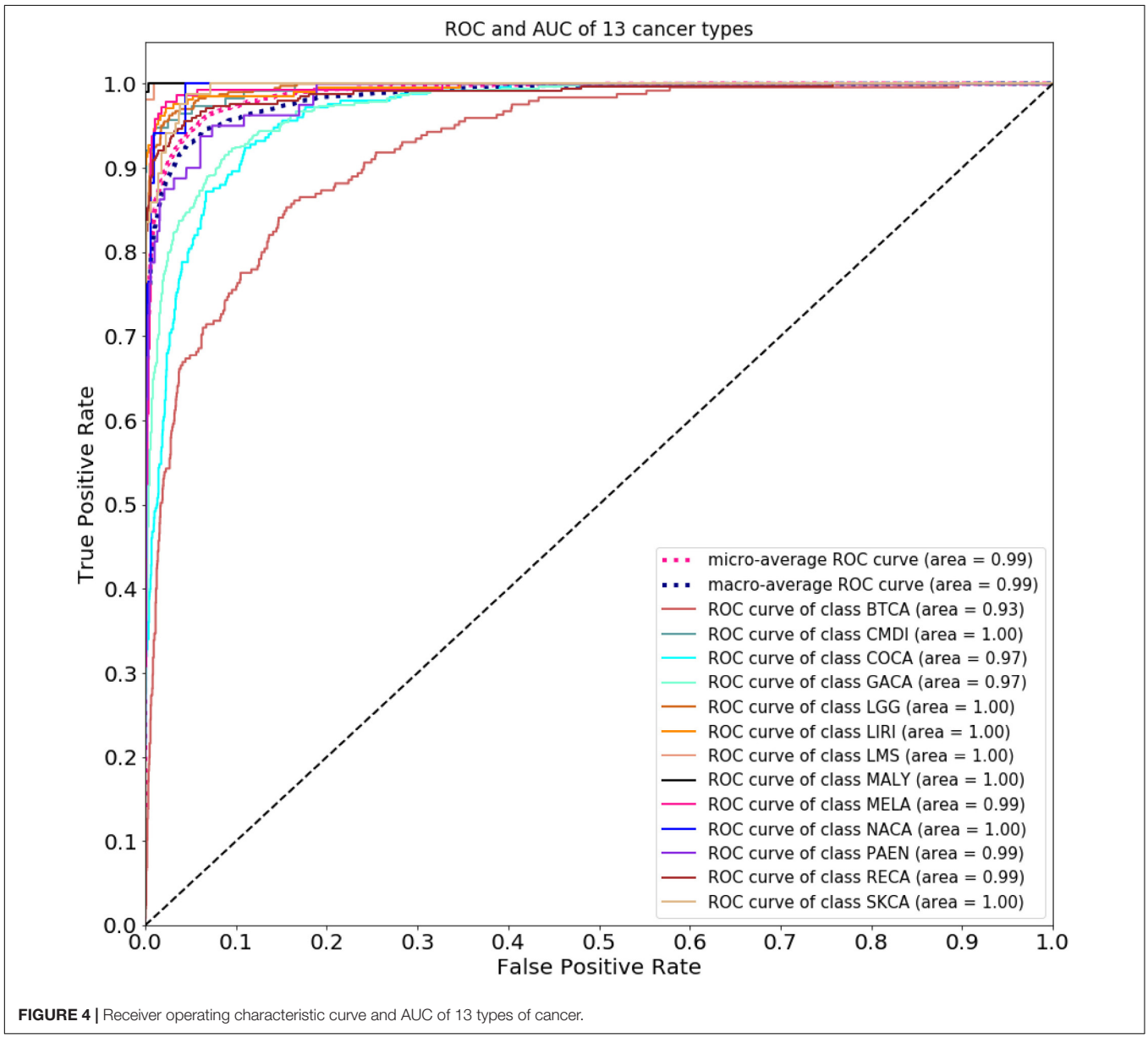


FIGURE 4 | Receiver operating characteristic curve and AUC of 13 types of cancer.

of features. And L1 regularization term was also used. The best θ was determined by minimizing the loss function based on gradient descent.

Evaluation Metric

We used accuracy, precision, recall, and F1 score as the metric for performance evaluation. True positive (TP) and false positive (FP) represent samples whose true label are positive and negative, respectively, were predicted as positive, whereas true negative (TN) and false negative (FN) represent samples, whose true label was negative and positive, respectively. These were predicted as negative. Accuracy was used to measure the overall performance and was calculated by Eq. 7. Precision demonstrates the ability of classifier to distinguish positive and negative samples and was calculated by Eq. 8. Recall represents the ability of the classifier to recognize all positive samples and was calculated by Eq. 9. F1 score was the harmonic average value of precision and recall and is calculated by Eq. 10. Because there is class imbalance in sample distribution in this study, ROC (receiver operating characteristic) curve and AUC (area under the curve) were also used to evaluate classification performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{F1 score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{10}$$

Functional Annotation

We utilized the Gene Ontology enrichment analysis database (Ye et al., 2006; Waardenberg et al., 2016) to annotate the function of the gene used in the model, shown in **Figure 3**. The R package gogadget and clusterProfiler (Nota, 2016; Yu et al., 2012) were used for gene visualization and clustering.

RESULTS

Workflow

The complete process for predicting the primary site of a tumor is shown in **Figure 1**, which can be divided into three parts. First, we obtained the somatic mutation data from the ICGC database and carried out data preprocessing such as filled null value and filtered invalid data. A matrix of features was generated for follow-up handling. Thereafter, we built a gene selection model using the random forest algorithm. Genes were selected with 10-time cross-validation. Finally, we constructed the classifier by utilizing the logistic regression algorithm, and the final matrix feature was fed into the classifier. The results were obtained with 10-time 10-fold cross-validation, and model performance was analyzed by the evaluation metric.

Data

We obtained the somatic mutation data from ICGC version 28 database for gene selection and tumor classification. Allelic mutations in somatic mutation data can be A/G, C/T, C/A, and so on. Because of limited information and tools, we treated all allele mutations as mutations and counted the number of mutations. And we counted the number of mutations of each sample. The sample distribution of each cancer is shown in **Table 1**. A total of 3,219 primary samples and 155 metastatic samples were used to model training and included 13 types of cancer.

Genes Used to Infer Cancer Tissue of Origin

The role of relative genes was discussed in context of molecular function, biological processes, and cellular components. **Figure 2** shows functional annotation of the top 500 genes selected using the random forest algorithm. Genes were found to enrich cell–cell adhesion, regulation of ion transmembrane transport, modulation of chemical synaptic transmission, forebrain development, and so on. Among these, gene enrichment evidently concentrated on the recognition and adhesion between cells and neurotransmitter conduction. Abnormal proteins that resulted from gene mutations can cause abnormal cell adhesion or differentiation, as well as abnormal neurotransmitter conduction or abnormal neural cell differentiation. Meanwhile, gastric cancer and brain lower-grade glioma account for a high proportion in all samples. Jiang et al. (2004) research the frequency and nature of mutations of the *CDH1* gene in gastric cancer, and proved that the mutation accounts for gastric cancer. The *APC* gene has been found to play an important role in the pathogenesis of soft tissue tumors (Kuhnen et al., 2000). Birnbaum et al. (2012) explored the role of the *APC* gene in colorectal cancer, by investigating 183 cases, and found point mutations in 73% of these cases. Mutation of the *IDH1* gene leads to a reduction in cell survival and proliferation, as well as further invasion of human gliomas

TABLE 3 | Performance metric of test dataset using top 500 genes.

Cancer	Precision	Recall	F1 score	Support	Specificity
BTCA	0.6429	0.6000	0.6207	15.0000	0.9675
CMDI	1.0000	1.0000	1.0000	5.0000	1.0000
COCA	0.7059	0.7500	0.7273	16.0000	0.9673
GACA	0.8148	0.7097	0.7586	31.0000	0.9638
LGG	0.9412	1.0000	0.9697	32.0000	0.9854
LIRI	0.9412	0.8889	0.9143	18.0000	0.9934
LMS	1.0000	1.0000	1.0000	2.0000	1.0000
MALY	1.0000	1.0000	1.0000	9.0000	1.0000
MELA	1.0000	0.8889	0.9412	9.0000	1.0000
NACA	1.0000	1.0000	1.0000	2.0000	1.0000
PAEN	0.3333	1.0000	0.5000	1.0000	0.9881
RECA	0.9583	0.9583	0.9583	24.0000	0.9931
SKCA	0.7143	1.0000	0.8333	5.0000	0.9878
Average	0.8501	0.9074	0.8633	169.0000	0.9890
Accuracy	0.8639	NA	NA	NA	NA

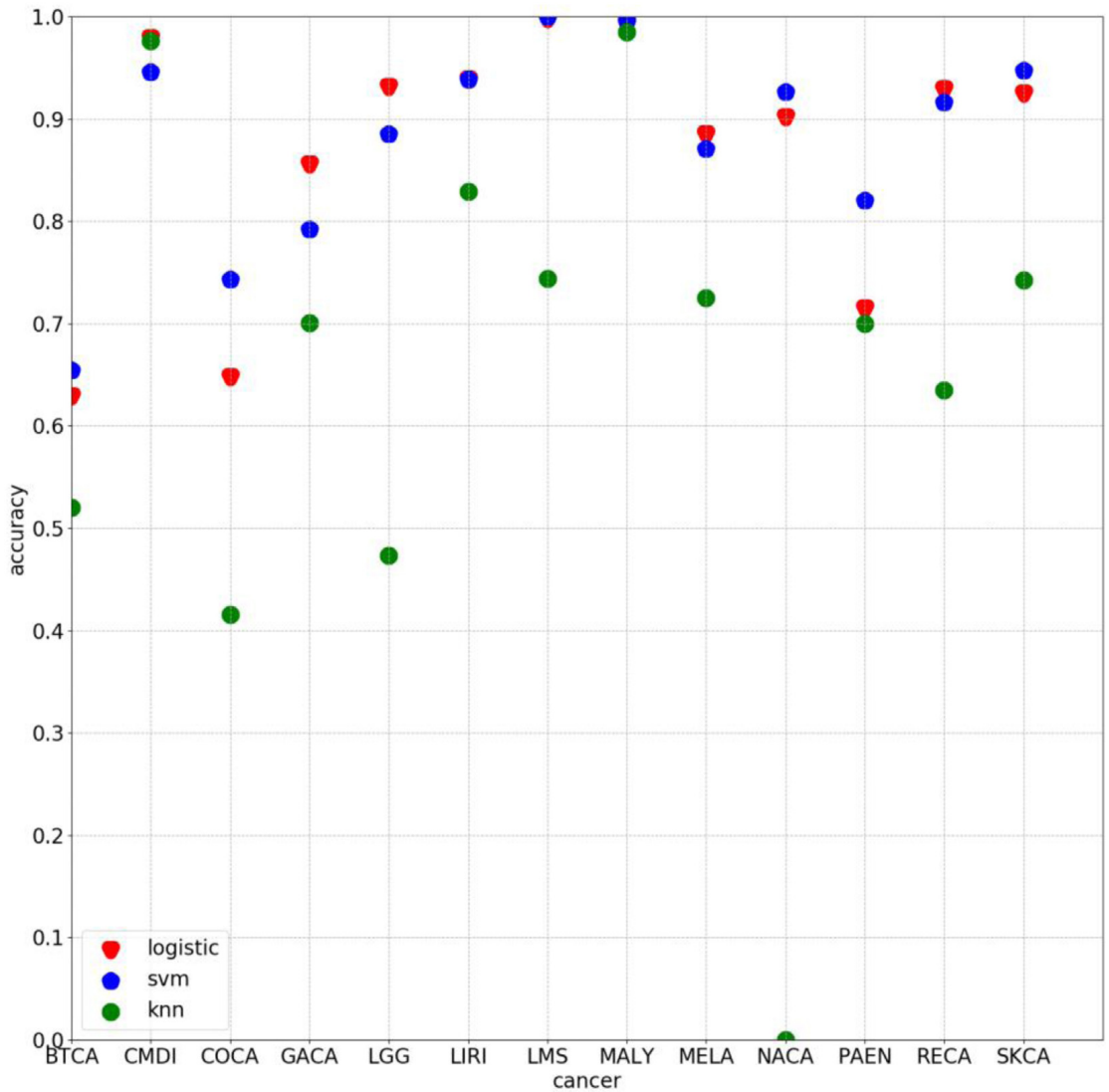


FIGURE 5 | Classification accuracy on each cancer by using 500 chosen genes based on logistic, svm, and knn, respectively.

by malignant tumor cells (Cui et al., 2016). Mutation of the *IDH1* gene has been proved to be the driving oncogenic factor of and has an impact on most brain lower-grade gliomas of different genetic pathways (Ohno et al., 2013; Pieper et al., 2014; Ohka et al., 2017).

According to research carried out on patients with liver cancer from China and southern Africa, a mutational hotspot at codon 249 of the p53 tumor suppressor gene has been identified (Hsu et al., 1993), and HBV and aflatoxin B1 (AFB1) are known synergistic risk factors. Zheng et al. (2005) explored the role of mutation of the DNA polymeraseβ (polβ) gene in human nasopharyngeal cancer and its relationship with EBV. Zhao (2001) carried out investigation on the mutation of the *ras* gene and what role they played in HP infection. They determined the infection of HP through serological examination. The results showed that 28 of 43 cases existed with mutations in codon 12 and a mutation rate of 65.12% (Zhao, 2001). **Supplementary Figure 1** also shows the relationship between gene mutations and cancers. Therefore, we concluded that viral infections could lead to gene mutations and result in cancer. In this study, somatic mutation data were utilized to identify the primary site of a tumor based on machine learning methods, which can contribute to the further diagnosis and treatment of cancer.

Performance Evaluation

Figure 2 compares the accuracy with a different number of genes used in the classifier. Because of gene sequencing and mutation detection being costly and time consuming, we selected 100 and 1,000 as the minimum and maximum number of genes, respectively. And we carried out a large number of experiments, with 100 genes selected as the interval. The highest accuracy

was obtained when using the top 500 genes. These results are shown in **Figure 3** with 10-time 10-fold cross-validation. The average accuracy is 86.71%, and precision, recall, and F1 score are presented in **Table 2**. The ROC curve and AUC of 13 types of cancer are shown in **Figure 4**. Most curves are close to 100%, and the area of each cancer is very close to 1 except BTCA (biliary tract cancer). The micro-average and macro-average are 0.99, which show the prediction value of each dimension and the average of all areas. Combining the metrics of prediction accuracy, ROC, AUC, and so on, our model had the worst overall prediction performance at biliary tract cancer and the best overall prediction performance at malignant lymphoma. Liver cancer, nasopharyngeal cancer, and gastric cancer are caused by HBV, HCV, EBV, and HP, respectively. The performance of our model on nasopharyngeal cancer was comparatively poor. In general, our model can obtain considerable prediction performance with the use of mutation data, which is great help in identification of the primary site of a tumor, follow-up diagnosis, and treatment.

In this study, the metastatic samples were used as test dataset. We carried out experiments by using 500 chosen genes with use of the model trained by training dataset. An average classification accuracy is 86.39%, as shown in **Table 3**. Although the model performed poorly on Pancreatic endocrine neoplasms (PAEN), the overall classification accuracy is satisfying. In this condition, we considered that little error on classification is tolerable.

Some experiments were also conducted by using other algorithm with 500 selected genes. The average classification accuracy values of using k-nearest neighbor (knn) and support vector machine (svm) are 62.66 and 85.27%, respectively, lower than 86.71% obtained by using the method proposed in this study. As **Figure 5** clearly shows, the classification accuracy on each cancer of using logistic algorithm was significantly higher than using knn. The overall performance of logistic is also better than svm. Therefore, the method proposed in this study can provide better prediction performance.

Mean Value of Number of Somatic Mutations on Each Cancer

We mapped the number of somatic mutations in each cancer, as shown in **Supplementary Figure 1**. Columns represent cancers, and rows represent genes. The number of mutations is colored on a logarithmic scale. Also, we used the color bar to show difference in values. The color of rectangles in the heat map represents the relative log number of mutations per gene in each cancer type. Cancers distributed in clusters along the vertical axis had similar values in the number of mutations. Genes also cluster on the horizontal axis, based on the association between cancers.

DISCUSSION

Viruses have been proven an important cause of cancer (Tsai et al., 1996; Lin et al., 2013; Geng and Zhang, 2017). Achieving effective identification of the primary site of a tumor caused by viruses or other factors plays a vital role in the follow-up diagnosis and treatment. Existing research shows that molecular profiling can be used to predict the primary site of a tumor. In this study, somatic mutation data were used to determine cancer tissue origin. Samples of 13 types of cancer were used with 3,374 samples used for feature extraction. The selected top 500 genes with mutation data were selected based on the feature importance score and was trained in the proposed classifier with 10-time 10-fold cross-validation. An average accuracy of 86.71% was obtained with use of machine learning algorithms, random forest algorithm, and logistic regression, utilized for gene selection and cancer classification, respectively.

Our model can achieve considerable performance in prediction of the primary site of common cancers caused by a virus or other factors. However, prediction performances on biliary tract cancer and nasopharyngeal carcinoma are discouraging. According to the sample distribution in **Table 1**, poor performance on nasopharyngeal carcinoma may be attributed to the small quantity of samples tested for this carcinoma. The reason for poor classification of the biliary tract cancer requires further research because of a lack of evidence. Therefore, we infer that there are shortcomings in using mutation data alone to identify the primary site of some cancers, but our model can obtain considerable overall performance. This positively affects the follow-up diagnosis and treatment.

CONCLUSION

As a large number of patients have CUP, tracing the primary site of a tumor has been a long-term challenge. Molecular profiling of tissue-specific genes is available from public database or medical institutions. We conducted experiments using somatic mutation data based on machine learning algorithms. Results showed that the proposed method is beneficial to the diagnosis and treatment of patients with unknown primary sites. However, the model does not perform well on all cancers. This motivates for further research on the identification of tissue origin of more common cancers. And research on performance of combination of somatic mutation data and other molecular profiling will be considered in our future work. Currently, the proposed method can achieve considerable performance and will help in the progress of the follow-up study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/releases/release_28/.

AUTHOR CONTRIBUTIONS

LZ and HZ designed the project. XL and LL analyzed the data, carried out the experiments, and wrote the manuscript.

LP, BW, JL, QL, GT, XZ, and YS modified and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

This research was funded by the National Natural Science Foundation of China (Grant 61803151) and the Project of Scientific Research Fund of Hunan Provincial Education Department (Grant 17A052).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00674/full#supplementary-material

REFERENCES

Ambrosini, V., Nanni, C., Rubello, D., Moretti, A., Battista, G., Castellucci, P., et al. (2006). 18F-FDG PET/CT in the assessment of carcinoma of unknown primary origin. *La Radiol. Med.* 111, 1146–1155. doi: 10.1007/s11547-006-0112-6

Birnbaum, D. J., Laibe, S., Ferrari, A., Lagarde, A., Fabre, A. J., Monges, G., et al. (2012). Expression profiles in stage II colon cancer according to APC gene status. *Transl. Oncol.* 5, 72–76. doi: 10.1593/tlo.11325

Bloom, G., Yang, I. V., Boulware, D., Kwong, K. Y., Coppola, D., Eschrich, S., et al. (2004). Multi-platform, multi-site, microarray-based human tumor classification. *Am. J. Pathol.* 164, 9–16. doi: 10.1016/s0002-9440(10)63090-8

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

Centeno, B. A., Bloom, G., Chen, D.-T., Chen, Z., Gruidl, M., Nasir, A., et al. (2010). Hybrid model integrating immunohistochemistry and expression profiling for the classification of carcinomas of unknown primary site. *J. Mol. Diagn.* 12, 476–486. doi: 10.2353/jmoldx.2010.090197

Chen, Z., Long, L., Wang, K., Cui, F., Zhu, L., Tao, Y., et al. (2016). Identification of nasopharyngeal carcinoma metastasis-related biomarkers by iTRAQ combined with 2D-LC-MS/MS. *Oncotarget* 7, 34022–34037. doi: 10.18632/oncotarget.9067

Cui, D., Ren, J., Shi, J., Feng, L., Wang, K., Zeng, T., et al. (2016). R132H mutation in IDH1 gene reduces proliferation, cell survival and invasion of human glioma by downregulating Wnt/β-catenin signaling. *Int. J. Biochem. Cell Biol.* 73, 72–81. doi: 10.1016/j.biocel.2016.02.007

Dietlein, F., and Eschner, W. (2014). Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. *Hum. Mol. Genet.* 23, 1527–1537. doi: 10.1093/hmg/ddt539

Erlander, M. G., Ma, X.-J., Kesty, N. C., Bao, L., Salunga, R., and Schnabel, C. A. (2011). Performance and clinical evaluation of the 92-Gene Real-Time PCR assay for tumor classification. *J. Mol. Diagn. Jmd* 13, 493–503. doi: 10.1016/j.jmoldx.2011.04.004

Erlander, M. G., Moore, M. W., Cotter, P., Reyes, M., Stahl, R., Hamati, H., et al. (2004). Molecular classification of carcinoma of unknown primary by gene expression profiling from formalin-fixed paraffin-embedded tissues. *J. Clin. Oncol.* 22(Suppl. 14):9545. doi: 10.1200/jco.2004.22.14_suppl.9545

Farinati, F., Valiante, F., Libera, G. D., Baffa, R., Rugge, M., Fanton, M. C., et al. (1993). Prevalence of *Helicobacter* pylori infection (HP) in patients with precancerous changes and gastric cancer. *Eur. J. Cancer Prevent.* 2(Suppl.):9. doi: 10.1097/00008469-199301001-00026

Fencl, P., Belohlavek, O., Skopalova, M., Jaruskova, M., Kantorova, I., and Simonova, K. (2007). Prognostic and diagnostic accuracy of [18F]FDG-PET/CT in 190 patients with carcinoma of unknown primary. *Eur. J. Nucl. Med. Mol. Imag.* 34, 1783–1792. doi: 10.1007/s00259-007-0456-8

Fu, Z., Chen, X., Yang, X., and Li, Q. (2019). Diagnosis of primary clear cell carcinoma of the vagina by 18F-FDG PET/CT. *Clin. Nuc. Med.* 44, 493–494.

Geng, W., and Zhang, H. Y. (2017). Research on the mechanism of HP mediated PI3K/AKT/GSK3β pathways in gastric cancer. *Eur. Rev. Med. Pharmacol. Sci.* 21(Suppl. 3):33.

Gonzaga, L., Coelho, V., Martins, G. M., Passos, M., and Castro, L. P. (2002). Once-daily, low-cost, highly effective H. pylori (HP)treatment to family members of gastric cancer patients. *Alim. Pharmacol. Ther.* 97, S59–S59.

Greco, F. A. (2013). Cancer of unknown primary or unrecognized adnexal skin primary carcinoma? Limitations of gene expression profiling diagnosis. *J. Clin. Oncol.* 31, 1479–1481.

Gross-Goupil, M., Massard, C., Lesimple, T., Merrouche, Y., Blot, E., Loriot, Y., et al. (2012). Identifying the primary site using gene expression profiling in patients with carcinoma of an unknown primary (CUP): a feasibility study from the GEFCAPI. *Onkologie* 35, 54–55. doi: 10.1159/000336300

Hainsworth, J. D., Rubin, M. S., Spigel, D. R., Boccia, R. V., Raby, S., Quinn, R., et al. (2013). Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the sarah cannon research institute. *J. Clin. Oncol.* 31, 217–223. doi: 10.1200/jco.2012.43.3755

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158, 929–944. doi: 10.1016/j.cell.2014.06.049

Hsu, I. C., Tokiwa, T., Bennett, W., Metcalf, R. A., Welsh, J. A., Sun, T., et al. (1993). P53 gene mutation and integrated hepatitis B viral DNA sequences in human liver cancer cell lines. *Carcinogenesis* 14, 987–992. doi: 10.1093/carcin/14.5.987

Hu, J., and Ludgate, L. (2007). HIV–HBV and HIV–HCV coinfection and liver cancer development. *Cancer Treat. Res.* 133, 241–252. doi: 10.1007/978-0-387-46816-7_9

Huebner, G., Morawietz, L., Floore, A., Buettner, R., Folprecht, G., Stork-Sloots, L., et al. (2007). 503 POSTER Comparative analysis of microarray testing and immunohistochemistry in patients with carcinoma of unknown primary. *Syndrome* 5, 90–91. doi: 10.1016/s1359-6349(07)70442-1

Hui, A., Cheung, S., Fong, Y., Lo, K., and Huang, D. (1998). Characterization of a new EBV-associated nasopharyngeal carcinoma cell line. *Cancer Genet. Cytogenet.* 101:83. doi: 10.1016/s0165-4608(97)00231-8

Janick, S., Elodie, L.-M., Marie-Christine, M., Philippe, R., and Marius, I. (2018). Immunohistochemistry for diagnosis of metastatic carcinomas of unknown primary site. *Cancers* 10, 108–110.

Jiang, Y., Wan, Y. L., Wang, Z. J., Zhao, B., and Huang, Y. T. (2004). Germline E-cadherin gene mutation screening in familial gastric cancer kindreds. *Chin. J. Surg.* 42, 914–917.

Kandalaft, P. L., and Gown, A. M. (2015). Practical applications in immunohistochemistry: carcinomas of unknown primary site. *Arch. Pathol. Lab. Med.* 140, 508–526.

Krishna, S. M., James, S., and Balaram, P. (2006). Expression of VEGF as prognosticator in primary nasopharyngeal cancer and its relation to EBV status. *Virus Res.* 115, 0–90.

Kuhnen, C., Herter, P., Monse, H., Kahmann, S., Muehlberger, T., Vogt, P. M., et al. (2000). APC and β-catenin in alveolar soft part sarcoma (ASPS) - immunohistochemical and molecular genetic analysis. *Pathol. Res. Pract.* 196, 0–304.

Küsters-Vandevelde, H. V. N., Kruse, V., Maerken, T. Van, Boterberg, T., Pfundt, R., Creytens, D., et al. (2017). Copy number variation analysis and methylome profiling of a GNAQ-mutant primary meningeal melanocytic tumor and its liver metastasis. *Exp. Mol. Pathol.* 102, 25–31. doi: 10.1016/j.yexmp.2016.12.006

Kwee, T. C., Basu, S., Cheng, G., and Alavi, A. (2010). FDG PET/CT in carcinoma of unknown primary. *Eur. J. Nuc. Med. Mol. Imag.* 37, 635–644.

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Li, M., Hong, G., Cheng, J., Li, J., Cai, H., Li, X., et al. (2016). Identifying reproducible molecular biomarkers for gastric cancer metastasis with the aid of recurrence information. *Sci. Rep.* 6:24869.

Lin, H., Ha, N. B., and Ahmed, A. (2013). Both HCV and HBV are major causes of liver cancer in southeast asians. *J. Immigr. Minor. Health* 15, 1023–1029. doi: 10.1007/s10903-013-9871-z

Malhi, A., and Gao, R. (2005). PCA-based feature selection scheme for machine defect classification. *Instrument. Measur.* 53, 1517–1525. doi: 10.1109/tim.2004.834070

Marquard, A. M., Birkbak, N. J., Thomas, C. E., Favero, F., Krzystanek, M., Lefebvre, C., et al. (2016). TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *Bmc Med. Genom.* 8: 58–59.

Muthukrishnan, R., and Rohini, R. (2016). "LASSO: a feature selection technique in predictive modeling for machine learning," in *Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, Coimbatore.

Nota, B. (2016). Gogadget: an r package for interpretation and visualization of go enrichment results. *Mol. Inform.* 36:1600132. doi: 10.1002/minf.201600132

Ohka, F., Yamamichi, A., Kurimoto, M., Motomura, K., Tanahashi, K., Suzuki, H., et al. (2017). A novel all-in-one intraoperative genotyping system forIDH1-mutant glioma. *Brain Tumor Pathol.* 34, 91–97. doi: 10.1007/s10014-017-0281-0

Ohno, M., Narita, Y., Miyakita, Y., Matsushita, Y., and Shibui, S. (2013). Secondary glioblastomas with IDH1/2 mutations have longer glioma history from preceding lower-grade gliomas. *Brain Tumor Pathol.* 30, 224–232. doi: 10.1007/s10014-013-0140-6

Oien, K. A., and Dennis, J. L. (2012). Diagnostic work-up of carcinoma of unknown primary: from immunohistochemistry to molecular profiling. *Ann. Oncol.* 23(Suppl. 10), 271–277.

Pieper, R. O., Ohba, S., and Mukherjee, J. (2014). Mutant IDH1-driven cellular transformation increases RAD51-mediated homologous recombination and temozolomide (TMZ) resistance. *Cancer Res.* 74, 4836–4844. doi: 10.1158/0008-5472.can-14-0924

Pranoto, H., Gunawan, F. E., and Soewito, B. (2015). Logistic models for classifying online grooming conversation. *Proc. Comp. Sci.* 59, 357–365. doi: 10.1016/j.procs.2015.07.536

Qu, K. Z., Li, H., Whetstone, J. D., Sferruzza, A. D., and Bender, R. A. (2007). Molecular identification of carcinoma of unknown primary (CUP) with gene expression profiling. *J. Clin. Oncol.* 25(Suppl. 18), 21024–21024. doi: 10.1200/jco.2007.25.18_suppl.21024

Rizwan, M., and Zulfiqar, M. (2010). Carcinoma of unknown primary. *J. Pak. Med. Assoc.* 60, 598–599.

Rosenwald, S., Gilad, S., Benjamin, S., Lebanony, D., Dromi, N., Faerman, A., et al. (2010). Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. *Mod. Pathol.* 23, 814–823. doi: 10.1038/modpathol.2010.57

Sandri, M., and Zuccolotto, P. (2006). *Variable Selection Using Random Forests. in Data Analysis, Classification and the Forward Search*. Berlin: Springer.

Shaw, P. H. S., Adams, C. J., and Crosby, T. D. L. (2007). A clinical review of the investigation and management of carcinoma of unknown primary in a single cancer network. *Clin. Oncol.* 19, 87–95. doi: 10.1016/j.clon.2006.09.009

Sheffield, B. S., Tessier-Cloutier, B., Li-Chang, H., Shen, Y., Pleasance, E., Kasaian, K., et al. (2016). Personalized oncogenomics in the management of gastrointestinal carcinomas-early experiences from a pilot study. *Curr. Oncol.* 23, 68–73.

Tsai, S. T., Jin, Y.-T., and Su, I.-J. (1996). Expression of EBER1 in primary and metastatic nasopharyngeal carcinoma tissues using in situ hybridization: a correlation with WHO histologic subtypes. *Cancer* 77, 231–236. doi: 10.1002/(sici)1097-0142(19960115)77:2<231::aid-cncr2>3.0.co;2-p

Voigt, J. J. (2008). Immunohistochemistry: a major progress in the classification of carcinoma of unknown primary. *Oncologie* 10, 693–697.

Waardenberg, A. J., Bassett, S. D., Bouveret, R., and Harvey, R. P. (2016). Erratum to: 'CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments'. *BMC Bioinform.* 17:179–185.

Wang, L. B., Jiang, Z. N., Fan, M. Y., Xu, C. Y., and Shen, J. G. (2008). Changes of histology and expression of MMP-2 and nm23-H1 in primary and metastatic gastric cancer. *World J. Gastroenterol.* 14, 1612–1616.

Xiuping, X., Yan, Q., Wang, Y., and Dong, X. (2016). NTN4 is associated with breast cancer metastasis via regulation of EMT-related biomarkers. *Oncol. Rep.* 37, 449–457. doi: 10.3892/or.2016.5239

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., et al. (2006). WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* 34, 293–312.

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

Zhang, B., Chai, H., Yang, Z., Liang, Y., Chu, G., and Liu, X. (2014). Application of L1/2 regularization logistic method in heart disease diagnosis. *Bio Med. Mater. Eng.* 24, 3447–3454. doi: 10.3233/bme-141169

Zhao, D. (2001). Investigation of the mutation of ras gene in gastric cancer and their relation to *helicobacter* pylori(HP)infection. *Cancer Res. Prevent. Treatm.* 18, 68–83.

Zheng, H., Ming-Shan, L. I., Zhao, G. Q., and Dong, Z. M. (2005). DNA polymerase β gene mutation in human nasopharyngeal cancer and its relationship with EBV infection. *J. Fourth Milit. Med. Univ.* 68, 198–234.

Conflict of Interest: JL, BW, QL, and GT are employed by Genesis Beijing Co., Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 July 2020
doi: 10.3389/fgene.2020.00725

Prognostic Implications of Immune-Related Genes' (IRGs) Signature Models in Cervical Cancer and Endometrial Cancer

Hao Ding[1], Guan-Lan Fan[1], Yue-Xiong Yi[1], Wei Zhang[1], Xiao-Xing Xiong[2] and Omer Kamal Mahgoub[2]*

[1] Department of Gynecology, Zhongnan Hospital of Wuhan University, Wuhan, China, [2] Central Laboratory, Renmin Hospital of Wuhan University, Wuhan, China

Cervical cancer and endometrial cancer remain serious threats to women's health. Even though some patients can be treated with surgery plus chemoradiotherapy as a conventional option, the overall efficacy is deemed unsatisfactory. As such, the development for new treatment approaches is truly necessary. In recent years, immunotherapy has been widely used in clinical practice and it is an area of great interest that researchers are keeping attention on. However, a thorough immune-related genes (IRGs) study for cervical cancer and endometrial cancer is still lacking. We therefore aim to make a comprehensive evaluation of IRGs through bioinformatics and large databases, and also investigate the relationship between the two types of cancer. We reviewed the transcriptome RNAs of IRGs and clinical data based on the TCGA database. Survival-associated IRGs in cervical/endometrial cancer were identified using univariable and multivariable Cox proportional-hazard regression analysis for developing an IRG signature model to evaluate the risk of patients. In the end, this model was validated based on the enrichment analyses through GO, KEGG, and GSEA pathways, Kaplan-Meier survival curve, ROC curves, and immune cell infiltration. Our results showed that out of 25/23 survival-associated IRGs for cervical/endometrial cancer, 13/12 warranted further examination by multivariate Cox proportional-hazard regression analysis and were selected to develop an IRGs signature model. As a result, enrichment analyses for high-risk groups indicated main enriched pathways were associated with tumor development and progression, and statistical differences were found between high-risk and low-risk groups as shown by Kaplan-Meier survival curve. This model could be used as an independent measure for risk assessment and was considered relevant to immune cell infiltration, but it had nothing to do with clinicopathological characteristics. In summary, based on comprehensive analysis, we obtained the IRGs signature model in cervical cancer (*LTA, TFRC, TYK2, DLL4, CSK, JUND, NFATC4, SBDS, FLT1, IL17RD, IL3RA, SDC1, PLAU*) and endometrial cancer (*LTA, PSMC4, KAL1, TNF, SBDS, HDGF, LTB, HTR3E, NR2F1, NR3C1, PGR, CBLC*), which can effectively evaluate the prognosis and risk of patients and provide justification in immunology for further researches.

Keywords: immune-related genes, cervical cancer, endometrial cancer, TCGA, prognostic model

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Shankar Suman,
The Ohio State University, United States
Alfred Grant Schissler,
University of Nevada, Reno, United States

****Correspondence:***
Wei Zhang
zw6676@163.com

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

Received: *27 March 2020*
Accepted: *15 June 2020*
Published: *21 July 2020*

Citation:
Ding H, Fan G-L, Yi Y-X, Zhang W, Xiong X-X and Mahgoub OK (2020) Prognostic Implications of Immune-Related Genes' (IRGs) Signature Models in Cervical Cancer and Endometrial Cancer. Front. Genet. 11:725. doi: 10.3389/fgene.2020.00725

INTRODUCTION

Cervical cancer is the fourth most commonly occurring cancer in women worldwide (Yang et al., 2019), for which a major cause is chronic infection with high-risk HPV types (HPV types 16 and 18) (Cohen et al., 2019). This condition is considered the leading cause of death and disability for women, although progress has been made for diagnostic methods and treatment in recent years in the context of improved test panels that provide detailed screening around the world. In 2018, approximately 570,000 patients were diagnosed with cervical cancer and 31,000 died from it globally (Bray et al., 2018). In Japan, it has been estimated that there are 13,000 new cases and 3,500 deaths associated with cervical cancer each year (Ishikawa et al., 2020). The 5-year survival rate can be encouraging for local cervical cancer, as approximately 75–85% after effective treatments such as surgery. However, the 5-year survival rate for recurrence is approximately 15% (Liu et al., 2019b). Histopathologically, squamous cell carcinoma accounts for about 80–85% and adenocarcinoma about 15–20% (Yang et al., 2019). Traditionally, a patient may be treated with surgical removal of the lesions and adjacent lymph nodes in combination with cycles of radiotherapy and chemotherapy (Cosper et al., 2019). Recently, immunotherapy has been increasingly used in clinical settings (Crusz and Miller, 2020) and has now becomes one of the important areas of cancer research.

Endometrial cancer is another very common gynecological tumor, ranking as the sixth cause of cancer incidence in women following breast cancer, colorectal cancer, lung cancer, cervical cancer, and thyroid cancer (Bray et al., 2018). Statistics show that the incidence of endometrial cancer is second only to cervical cancer (Feng et al., 2019; Zhou and Ling, 2019) among gynecological malignancies in China. The survival rate for endometrial cancer varies with tumor progression; there was a big difference in 5-year survival rate by 83–97% in localized to 43–67% in stage III, and finally only 13–25% in stage IV (Liu, 2019). Traditional treatment options including surgery, radiotherapy, and chemotherapy can be effective for the condition in early stages but advanced diseases are not significantly responsive (Miller et al., 2020). As novel immunotherapies are being used to treat endometrial cancer (Grywalska et al., 2019), a new option is now available for doctors (Lynam et al., 2019).

Immunotherapies for cancer have attracted more and more attention from scientific researchers (Irvine and Dane, 2020). In recent years, traditional modalities have found more limitations to the treatment of cancer. Immunotherapies have provided more opportunities to modern precision medicine and personalized medicine (Martin et al., 2020). In fact, many immunotherapeutic methods have been applied in clinical practice, such as the typical immune checkpoint inhibitor that targets programmed cell death protein 1 (PD-1) for lung cancer (Gainor et al., 2020) and breast cancer (Barroso-Sousa et al., 2020), as well as CD19-specific CART (Shen et al., 2019) immune cell therapy for leukemia (June et al., 2018). In addition, immunotherapies are also widely used to treat gynecological tumors (Rubinstein and Makker, 2020). Existing studies on immunotherapy for cervical cancer focus mainly on human papillomavirus vaccine, immune checkpoint inhibitors, and adoptive cellular therapy. The main biological mechanism of the human papillomavirus vaccine is the viral vectors expressing HPV-16 or -18 (E6 or E7) to stimulate the body's immune response to malignant cells. These vaccines can be divided into two categories – prophylactic and therapeutic. Currently, there are three clinically available prophylactic human papillomavirus vaccines – Gardasil, Cervarix, and Gardasil 9 – that were approved by the U.S. Food and Drug Administration (FDA) in 2006, 2009, and 2016, respectively (Matanes and Gotlieb, 2019). Therapeutic human papillomavirus vaccine is also an important part of vaccine research, including live vector, nucleic acid, protein, whole cell, and combinatorial vaccines. However, although there are some promising vaccine candidates (Vici et al., 2016; Yang et al., 2016; Kim, 2017), there are currently no vaccine products available for human use. For immune checkpoint inhibitors, anti-programmed death 1 (PD-1) and anti-programmed death ligand 1 (PD-L1) immunoglobulin, as an important representative, have been the focus of research, and many drugs, such as pembrolizumab and nivolumab, have achieved encouraging results and were approved by the FDA (Wang and Li, 2019). At present, studies on adoptive cellular therapy in cervical cancer are insufficient. While some scholars have confirmed the efficacy of human papillomavirus-targeted tumor-infiltrating T lymphocytes (TILs) in cervical cancer (Stevanovic et al., 2015), they still face many problems. The main challenge lies in how to effectively identify the tumor-associated antigens (TAAs) from individual patients and how to amplify the TILs while inducing a targeted immune response to these tumor sites, which became the focus in subsequent studies. Compared to immunological studies on cervical cancer, studies on endometrial cancer are relatively few and mainly focus on the immune checkpoint inhibitors. Studies have shown that PD-1 and PD-L1 are expressed in 80% of primary endometrial carcinoma patients and almost 100% of metastatic tumors (Mo et al., 2016), The inhibitor pembrolizumab was FDA-approved for use in microsatellite instability-high (MSI-H) or mismatch repair (MMR)-deficient endometrial carcinoma patients (Le et al., 2017). Studies (Maskey et al., 2019) have shown that the number of CD4+ and CD8+ lymphocytes is not similar between normal cervical cells infected by HPV and cervical cancer cells, and this difference becomes more complicated for epithelial and stromal layers in cervical tissues. Based on a study of endometrial cancer, it was (Zhou and Ling, 2019) found that the survival rate correlated with the number of cytotoxic T lymphocytes. Despite the fact that *in vivo* and *in vitro* experiments are performed during plenty of studies on immune cell changes in gynecologic tumors, a more comprehensive and specific immune mechanism is still unclear.

Abbreviations: AUC, An area under the ROC curve; FDR, false discovery rate; GO, gene ontology; GSEA, Gene Set Enrichment Analysis; HPV, human papillomavirus; IRGs, immune-related genes; KEGG, Kyoto Encyclopedia of Genes and Genomes; OS, overall survival; ROC, Receiver Operating Characteristic curve; TCGA, The Cancer Genome Atlas; TFs, transcription factors; TIMER, Tumor IMmune Estimation Resource.

As modern high-throughput sequencing technology is being improved and rapid growth is achieved in computer science (Ma et al., 2019), more and more free of charge, large-scale, and comprehensive gene transcriptomics as well as relevant clinical databases are available, which makes it possible to provide comprehensive analyses of genetic molecular biomarkers in a more accurate and fast fashion. These molecular biomarkers play an important role in predicting the prognosis of patients and evaluating their risk levels. Therefore, we hope to further explore those data that provide details in immune related genes (IRGs) for patients with cervical cancer and those with endometrial cancer. Beyond that, efforts will also be made to evaluate and predict the prognosis of patients using these molecular biomarkers or other gene signatures. By combining the gene expression profiles and clinical data of IRGs with bioinformatics statistical methods, we obtained and analyzed those IRGs signatures and then verified them in patients with cervical cancer and those with endometrial cancer. These results will provide us a basic idea for follow-up and in-depth studies on these IRGs, thus laying foundation for precise and individualized medical treatment.

MATERIALS AND METHODS

Clinical Samples and Data Acquisition

For cervical and endometrial cancers, transcriptome RNA-sequencing data from FPKM file as well as clinical data were downloaded from The Cancer Genome Atlas (TCGA) database containing 3 non-tumor samples and 304 tumor samples from patients with cervical cancer, and 35 non-tumor samples and 543 tumor samples from those with endometrial cancer. All clinical data and transcriptome data did not correspond exactly because the clinical data were not completely provided, leading to exclusion from the subsequent analyses. Immune-related genes (IRGs) were derived from the Immunology Database and Analysis Portal (ImmPort) system (Bhattacharya et al., 2014) which was continuously updated and maintained to provide immune-related data that had endorsement by scholars. These resulting genes were thought to be involved in human's immune-related activities.

Differential Gene Analysis and Enrichment Analysis

All of these genes, including immune-related genes (IRGs) and all transcriptome RNA-sequencing genes that were differentially expressed in normal and tumor samples, were screened in association with cervical and endometrial cancer, respectively, through R-Limma package (R version 3.6.1), and the screening criteria were met based on false discovery rate (FDR) < 0.05 and $\log2\,|\text{fold change}| > 1$. Functional enrichment analyses through GO and KEGG pathways were conducted for differentially expressed IRGs using the online database webgestalt (Liao et al., 2019)[1].

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

Identification of Survival-Associated IRGs

We extracted the clinical data of overall survival (OS) time and survival state corresponding to cervical cancer and endometrial cancer, respectively, and the transcriptome of IRGs combined with corresponding clinical data to perform survival analysis and thus identify survival-associated IRGs using univariate Cox proportional hazard regression. To meet the screening criteria, $p < 0.05$ and $p < 0.01$ were defined for cervical cancer and endometrial cancer, respectively. Since many different IRGs were found for endometrial cancer, which was not helpful for subsequent analyses, more appropriate screening criteria should be followed.

Screening of Transcription Factors (TFs) and Construction of Networks

Three hundred and eighteen transcription factors (TFs) were downloaded from the cistrome online database[2] to figure out the differential genes in cervical and endometrial cancer, respectively, in a similar way used for IRG selection, using R-limma package (R version 3.6.1). The selection criteria were defined as false discovery rate (FDR) < 0.05 and $\log2\,|\text{fold change}| > 1$. Subsequently, the differentially expressed TFs and selected survival-associated IRGs were used to establish regulatory networks by Pearson correlation analysis with correlation coefficient > 0.4 at $p < 0.001$ after which the regulatory networks were imported into the cytospace software (version 3.7.2) for visual procedures.

Establishment and Evaluation of the IRG Signature Model

The survival-associated IRGs were further screened to establish the IRGs signature model, which was examined by multivariate Cox proportional-hazard regression analysis. This model would be used for subsequent evaluation and analysis of risk measures for the patients' risk values after assigning these patients into high-risk and low-risk groups. The risk score for each patient was computed using the formula as follows:

$$\text{risk score} = \Sigma_{k=1}^{n} \text{Coe f}_k^* \text{X}_k$$

where Coef_k represents the coefficient and X_k represents the expression level of each IRG. Subsequently, the validity of the IRGs signature model was evaluated by analyzing the difference between high- and low-risk groups using the Kaplan-Meier survival curve, Receiver Operating Characteristic (ROC) curve and heatmap. Similarly, Gene Set Enrichment Analysis (GSEA) was applied to compare signaling pathways and biological processes between high and low risk groups by GSEA (version 4.0.3) software. The landscape of genetic alterations across these IRGs in the signature model was examined through the online database cbioportal[3].

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

[3]https://www.cbioportal.org/

Evaluation of IRGs' Signature Model Along With Clinicopathological Characteristics and Tumor-Infiltrating Immune Cells

Whether the patient risk score could be used as an independent prognostic measure was further evaluated by univariate and multivariate Cox proportional-hazard regression analyses. The tumor infiltrating immune cell index were download from the online Tumor IMmune Estimation Resource (TIMER) (Li et al., 2017)[4], which provided detailed information about infiltrating immune cells including B cells, T cells, macrophages, neutrophils, and dendritic cells. Acceptable compatibility between these data and TCGA database is maintained; that is why the information thereof has been widely used in scientific researches in recent years. Therefore, it is helpful for us to further understand the changes of immune cells in tumor tissues. The relevance between risk scores and infiltrating immune cells was investigated herein using Pearson correlation analysis.

Statistical Analysis

All data were processed using the R software (version 3.6.1). The independent samples *t*-test was used to evaluate the relationship between risk scores and clinicopathological characteristics, and $P < 0.05$ was considered statistically significant. For Kaplan-Meier survival curves, the log-rank test was performed to demonstrate if there could be significant difference in OS between groups. Univariate and multivariate Cox proportional-hazard regression analyses were used to access the association between risk scores and OS. The area under the ROC curve (AUC) was measured for indicating the accuracy of prognosis as shown by the IRG signature model. All these analyses were performed at a significance level of $P < 0.05$.

RESULTS

Identification of Differentially Expressed IRGs

Based on the results derived from the R software, we found that there were 3192 differentially expressed genes in cervical cancer, including 1833 upregulated and 1359 downregulated; and 5665 differentially expressed genes in endometrial cancer, including 3316 upregulated and 2349 downregulated. A total of 2498 immune-related genes (IRGs) are described in the Immunology Database and Analysis Portal (ImmPort). We extracted differentially-expressed IRGs common to the TCGA and the ImmPort, which yielded 88 upregulated and 117 downregulated for cervical cancer, along with 226 upregulated and 171 downregulated for endometrial cancer. During enrichment analyses for these differentially expressed IRGs, the cervix-related genes were mainly found to be involved in "response to stimulus," "biological regulation," and "cell communication" for GO enrichment, and "cytokine-cytokine receptor interaction," "Ras signaling pathway," and "MAPK signaling pathway" as shown in Kyoto Encyclopedia of Genes and Genomes (KEGG). In comparison, endometrial cancer related genes showed biological processes in a similar manner, as they were also mainly involved in "biological regulation," "response to stimulus," and "cell communication" for GO enrichment, and "cytokine-cytokine receptor interaction," "chemokine signaling pathway," and "PI3K-Akt signaling pathway" as shown in Kyoto Encyclopedia of Genes and Genomes (KEGG) (**Figure 1**). The above findings suggested that these IRGs were strongly associated with the development, progression and invasion of tumors.

[4]https://cistrome.shinyapps.io/timer/

Identification of Survival-Associated IRGs

From the previous step, we obtained the differentially expressed IRGs. However, in our clinical studies, we paid more attention to the IRGs that were associated with the survival and prognosis of patients because these genes may be the key biomarkers for evaluating patients. During further screening, we obtained 25 survival-associated IRGs for cervical cancer and 23 for endometrial cancer, respectively (**Figures 2A,B**).

Identification of Differentially Expressed TFs and Construction of IRGs-TFs Regulatory Network

Progress has been made for researches on the change of DNA transcription factors (TFs) level in tumors, which is always the important direction for biological processes. By establishing the matrix for these TFs corresponding to the gene expression profiles, we found that there were 47 upregulated and 28 downregulated TFs in cervical cancer, as well as 44 upregulated and 53 downregulated TFs in endometrial cancer (**Figure 3**). The gene expression profiles were then extracted for survival-associated IRGs and differential expression TFs, respectively, to construct the IRGs-TFs regulatory network. As shown in the network diagram, the network for cervical cancer tissues was formed by 13 TFs and 10 IRGs. Similarly, 17 TFs and 9 IRGs formed the regulatory network for endometrial cancer. We found the IRG LTA in the regulatory network for both cervical cancer and endometrial cancer. As shown in the regulatory network, TFs STAT1 and FOXP3 were involved in regulatory relationship with multiple IRGs for cervical cancer, while IRGs *PGR* and *LTA* were associated with multiple TFs for endometrial cancer (**Figures 2C,D**). Both STAT1 and STAT2 are important members of the family of signal transducer and activator of transcription (STAT), but STAT1 plays the more important role (Verhoeven et al., 2020). Current studies have proven that STAT1 is an important activating mediator of type I and type II interferon (IFN), participating in the body's immune defense response against foreign pathogens and other viruses (Zhang et al., 2017). The biological function of STAT1 is still controversial and unclear. Studies of breast (Hou et al., 2018) and ovarian cancer (Tian et al., 2018) found that STAT1 is overexpressed in malignant tumors and plays an oncogenic role. However, studies on colorectal cancer (Crncec et al., 2018) and other breast cancers (Varikuti et al., 2017) found that

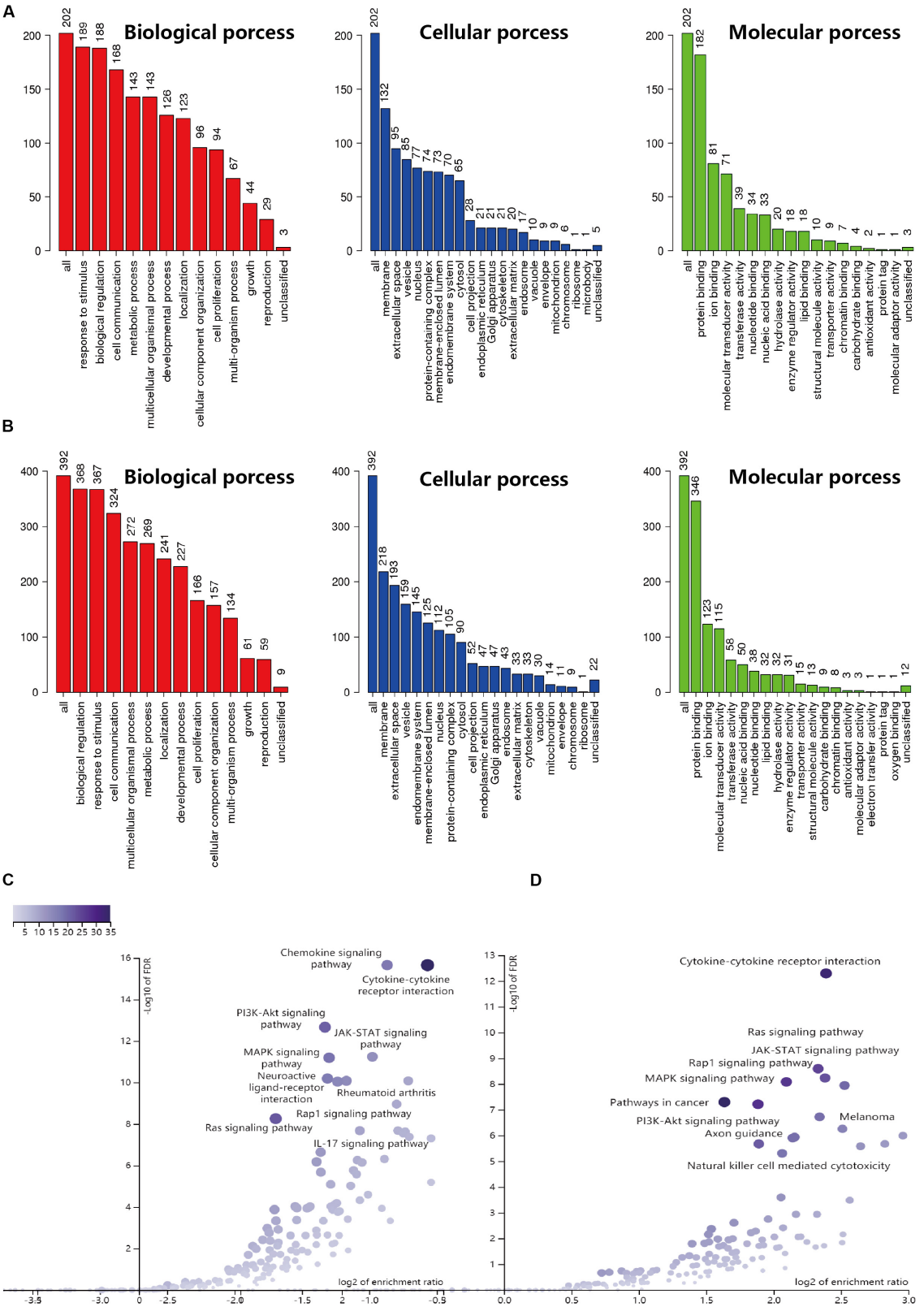


FIGURE 1 | GO and KEGG enrichment result for IRGs in CESC and UCEC. **(A)** GO enrichment analysis for CESC, the vertical axis represents the number of differentially expressed IRGs. **(B)** GO enrichment analysis for UCEC, the vertical axis represents the number of differentially expressed IRGs. **(C)** Volcano of KEGG enrichment result for CESC. **(D)** Volcano of KEGG enrichment result for UCEC.

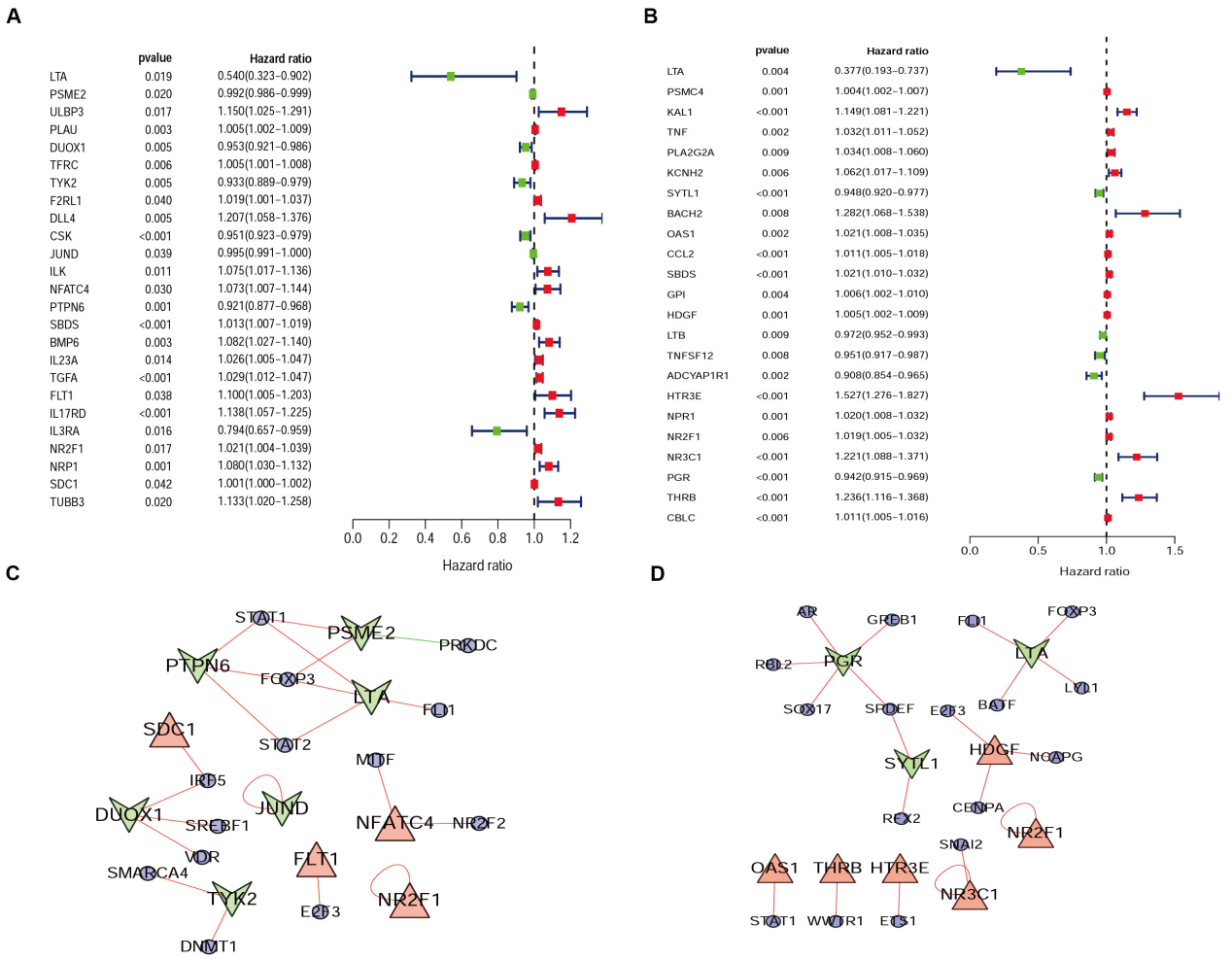


FIGURE 2 | Forest plot of survival-associated IRGs and IRGs-TFs regulatory network. **(A)** Survival-associated IRGs in cervical cancer. **(B)** Survival-associated IRGs in endometrial cancer, red dots represent high-risk genes (HR > 1), and green dots represent low-risk genes (HR < 1). The IRGs-TFs regulatory network in cervical cancer **(C)** and endometrial cancer **(D)**. The blue ellipse represent TFs, the triangle and taper represent IRGs, red/green represent high-Risk IRGs/low-Risk IRGs, red/green lines represent TFs positive/negative regulation IRGs.

TAT1 may act as a tumor suppressor. Even a recent meta-analysis of TAT1 in multiple types of tumors reported that the prognostic factor of STAT1 still depends on cancer type (Zhang et al., 2020). From this IRG–TF regulatory network, we can see that in cervical cancer, STAT1 positively regulates low-risk IRGs (*PSME2*, *LTA*, and *PTPN6*), but STAT1 positively regulates high-risk IRGs (*OAS1*) in endometrial cancer, suggesting that transcription factor STAT1 may play different biological roles in these two types of cancers. The functional role of transcription factor FOXP3 is also unclear in existing studies. On the one hand, FOXP3 can act as a tumor suppressor in breast cancer (Zuo et al., 2007), ovarian cancer (Zhang and Sun, 2010), colon cancer (Li et al., 2013), and gastric cancer (Ma et al., 2013), but it can act as an oncogene in non-small cell lung cancer (Yang et al., 2017), lung adenocarcinoma (Li et al., 2016), and thyroid cancer (Chu et al., 2015). From this IRG–TF regulatory network, we can see that in cervical cancer, FOXP3 positively regulates low-risk IRGs (LTA and PTPN6) and positively regulates low-risk IRGs (LTA) in endometrial cancer, suggesting that FOXP3 may act as a tumor suppressor in these two types of cancers.

Establishment and Evaluation of the IRGs Signature Model

Since different IRGs expression profiles may indicate the differences in disease condition among patients, it is of significance to establish the IRGs' prognosis signature model for patient risk evaluation. In this way, the IRGs' prognosis signature models were established for cervical cancer and endometrial cancer, respectively (**Supplementarys Table S1, S2**). Patients were divided into a high-risk group and a low-risk group according to the median risk score. Calculation based on the IRGs prognosis signature model resulted in 147 and 147 patients assigned to high-risk and low-risk subsets, respectively, for those with cervical cancer, and similarly 255 and 255 patients to two

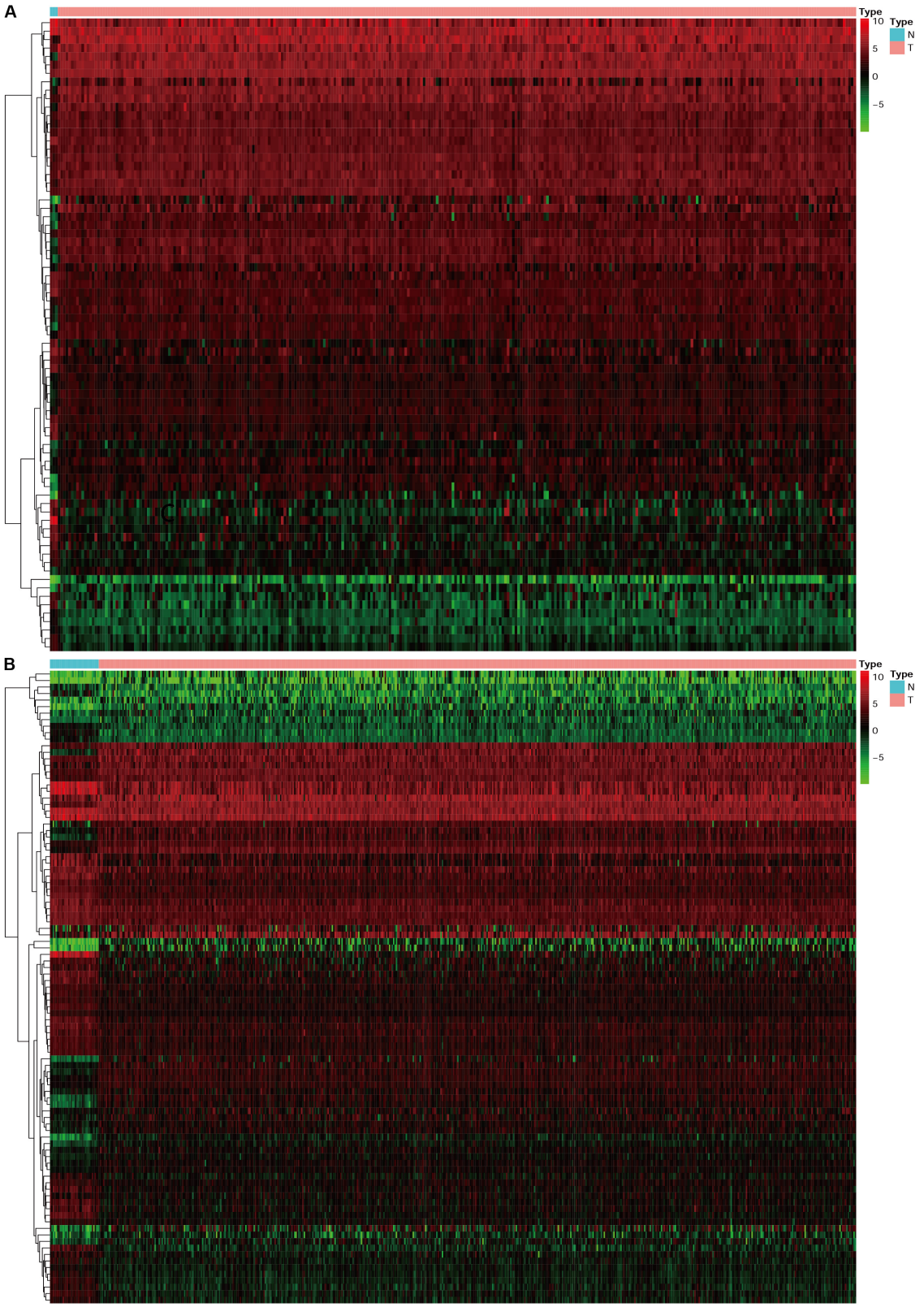


FIGURE 3 | The heatmap for differentially expressed TFs in cervical cancer and endometrial cancer. N represent normal, T represent tumor or cancer. **(A)** Cervical cancer. **(B)** Endometrial cancer.

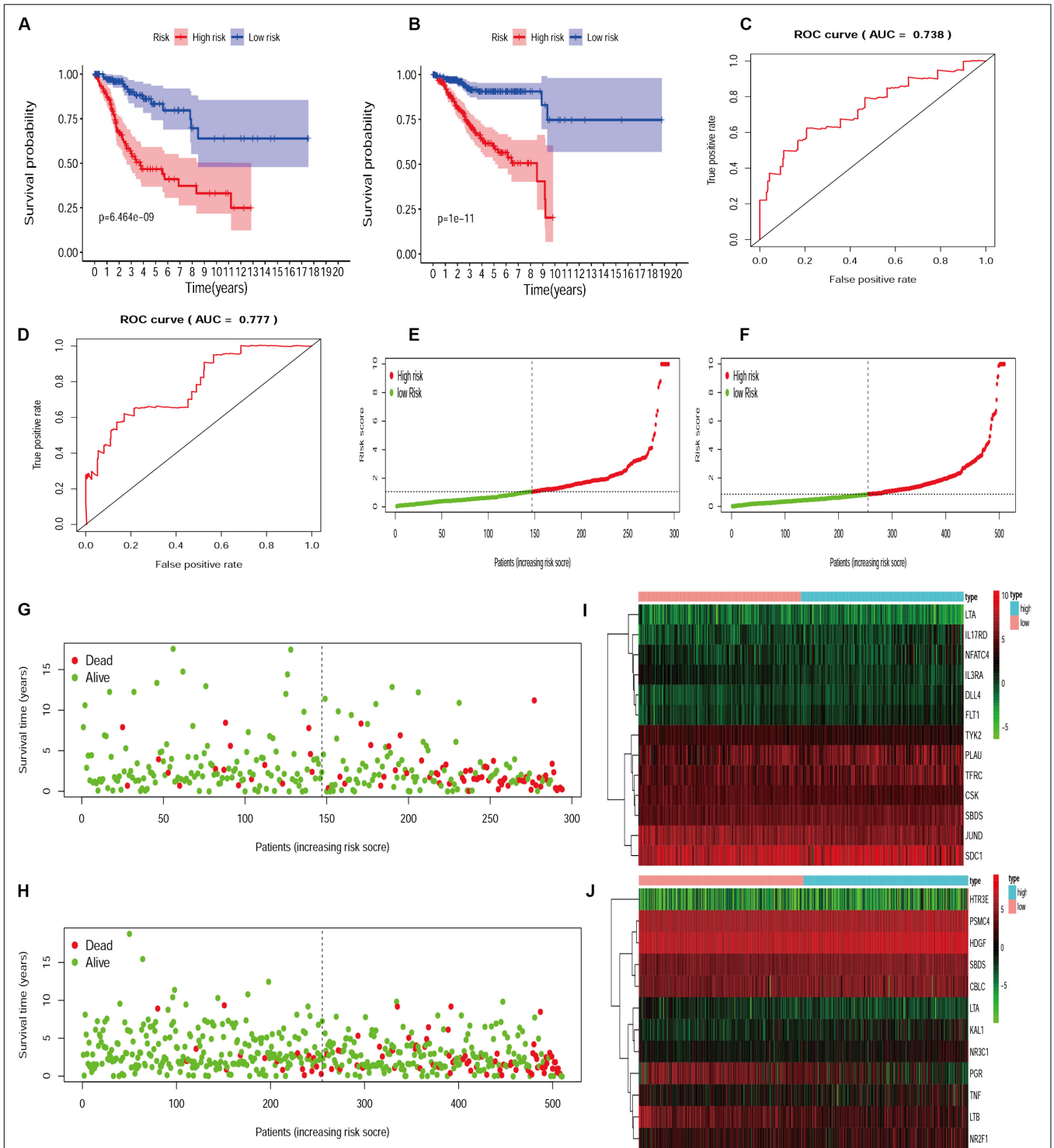


FIGURE 4 | The evaluation of IRGs signature model in cervical cancer and endometrial cancer. Kaplan-Meier survival curves for cervical cancer **(A)** and endometrial cancer **(B)**. AUC for cervical cancer **(C)** and endometrial cancer **(D)**. The distribution and survival status of patients with risk score in the cervical cancer **(E,G)** and endometrial cancer **(F,H)**. The heatmap for IRGs in cervical cancer **(I)** and endometrial cancer **(J)**.

subsets for those with endometrial cancer. Statistical evaluation was subsequently made to analyze this model by performing the comparison of Kaplan-Meier survival curves, evaluation of ROC curves, and drawing of distribution plots of patients at high/low risk. All these suggested that the IRGs signature model could be considered appropriate to evaluate the clinical

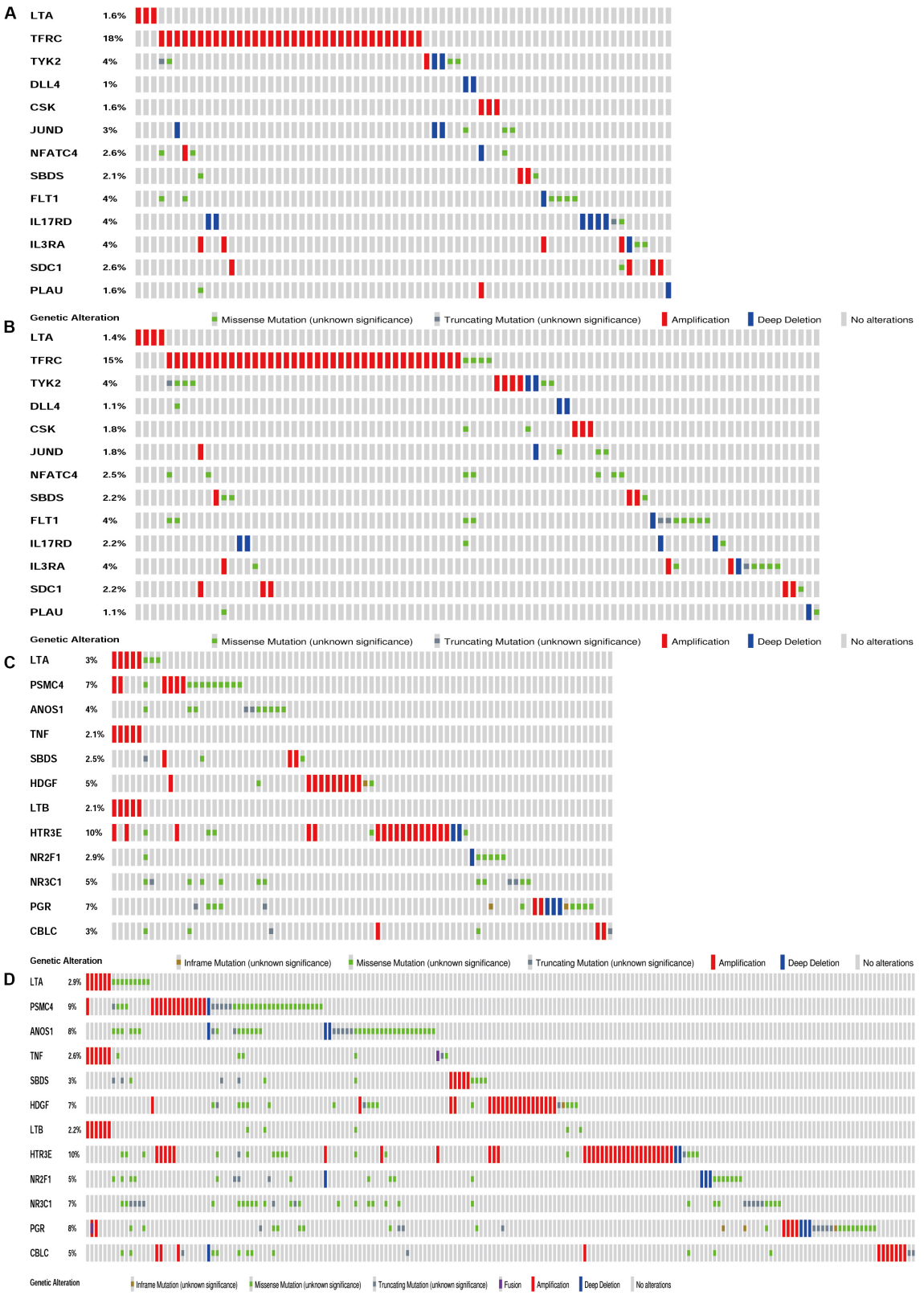


FIGURE 5 | Genetic alteration landscape of IRGs in gene signatures model of CESC and UCEC. **(A)** Genetic alteration in the TCGA-CESC cohort (191 samples). **(B)** Genetic alteration in the TCGA-Pan Cancer Atlas cohort (278 samples). **(C)** Genetic alteration in the TCGA-UCEC cohort (242 samples). **(D)** Genetic alteration in the TCGA-Pan Cancer Atlas cohort (509 samples).

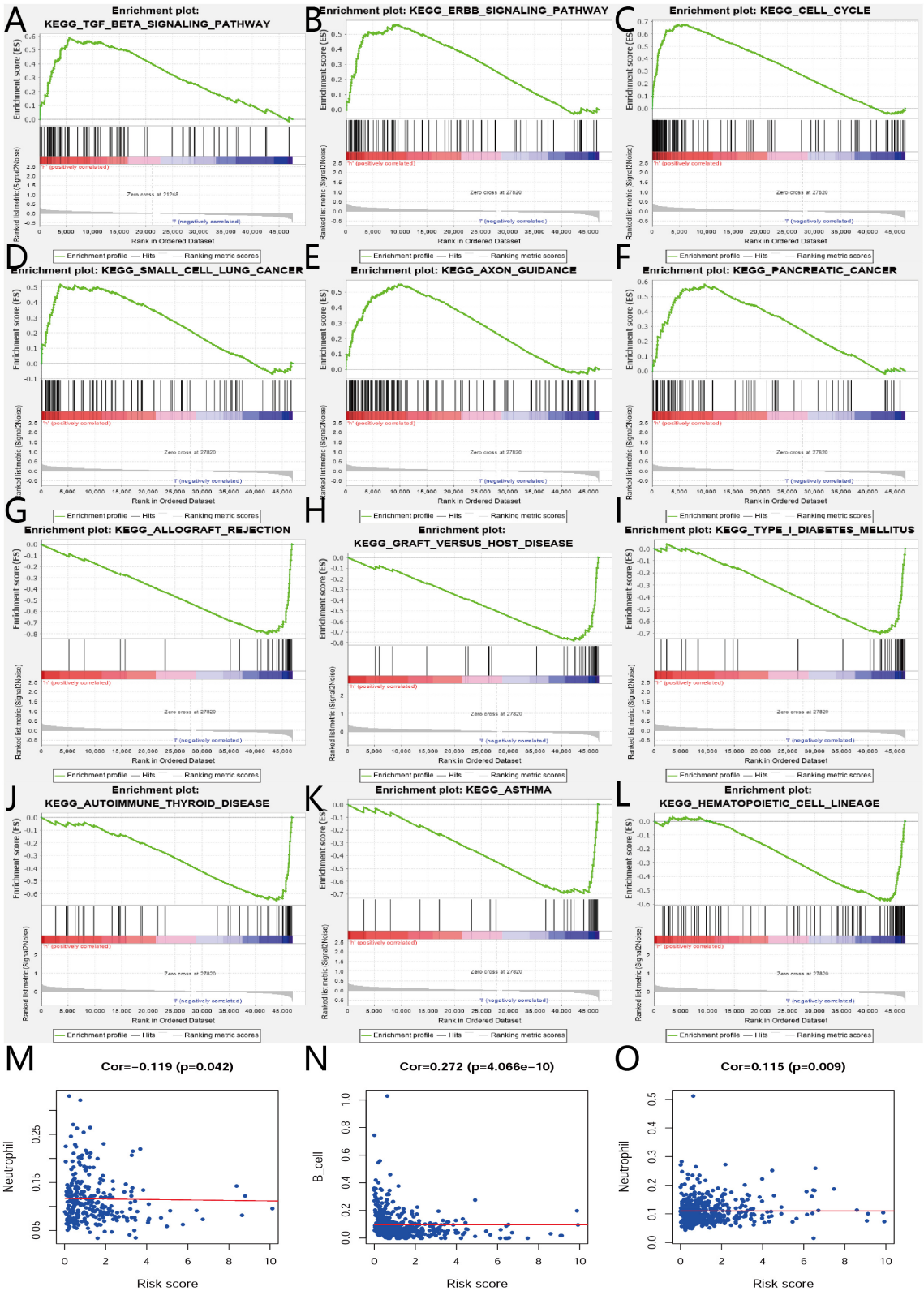


FIGURE 6 | GSEA and relationships between IRGs and tumor infiltrating immune cells index in CESC and UCEC. **(A)** GSEA results of high risk group in cervical cancer. **(B–F)** GSEA results of high risk group in endometrial cancer. **(G–L)** GSEA results of low risk group in endometrial cancer. **(M)** Relationships between risk score of IRGs model in cervical cancer and tumor infiltrating immune cells index. **(N,O)** Relationships between risk score of IRGs model in endometrial cancer and tumor infiltrating immune cells index.

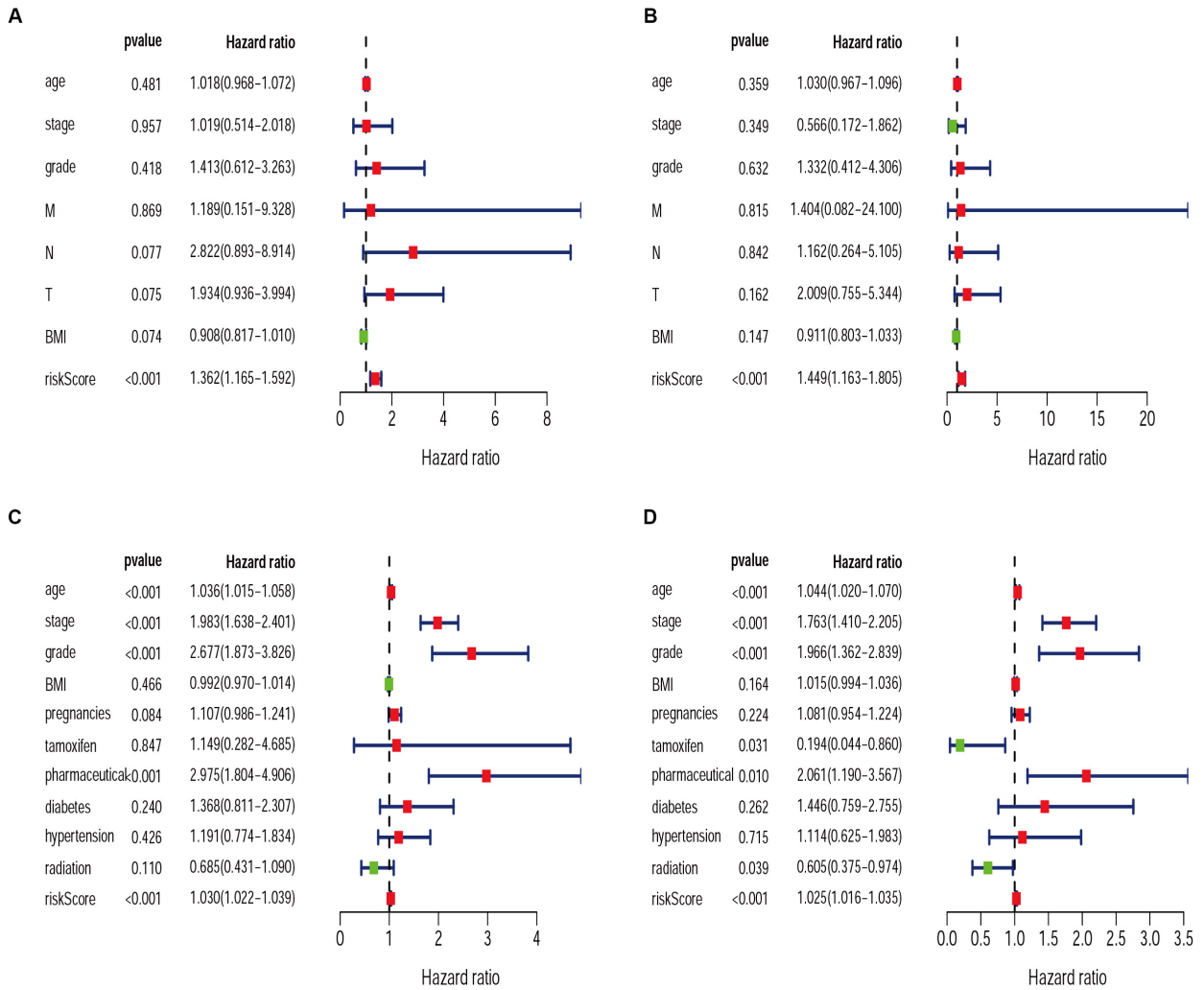


FIGURE 7 | Univariate and Multivariate Cox regression analysis of clinical characteristics for CESC **(A,B)** and UCEC **(C,D)**.

prognosis of patients, with the exception of the moderately AUC (**Figure 4**).

Multiple Evaluation of IRGs' Signature Model Combined With Clinicopathology, Gene Expression Profiles, GSEA, and TIMER

The biological characteristics regarding clinicopathology were also part of our considerations, including age, cancer stage, body mass index (BMI), and (TNM) stage, etc. We found the risk score resulted from the IRGs' signature model could be satisfactory as an independent statistical measure to evaluate the risk levels of patients. As an exception, the IRGs signature model for endometrial cancer developed the following independent clinical measures for risk levels evaluation, including age, cancer stage, and tumor pathological grades. Statistical difference was observed in clinicopathological characteristics among many IRGs expression profiles. Based on the online database cbioportal, datasets of TCGA-CESC/TCGA-UCEC cohort and TCGA-PanCancer Atlas were applied (310 samples in CESC vs. 297 samples in PanCancer Atlas; 548 samples in UCEC vs. 529 samples in PanCancer Atlas). Only samples harboring both mutations and CAN data were included. In terms of CESC, IRGs were altered in 69 (36%) of 191 queried samples (TCGA-CESC) (**Figure 5A**), as compared with those altered IRGs detected in 88 (32%) of 278 queried samples (PanCancer Atlas) (**Figure 5B**). In terms of UCEC, IRGs were changed in 80 (33%) of 242 queried samples (TCGA-UCEC) (**Figure 5C**), compared with 192 (38%) of 509 samples (PanCancer Atlas) (**Figure 5D**). GSEA analysis in cervical cancer revealed that the high-risk group was significantly associated with the TGF-beta signaling pathway (NES = 2.059, FDR = 0.017) (**Figure 6A**), but no pathway was significantly relevant to the low risk group. However, the GSEA analysis in endometrial cancer showed that these pathways including erbb-signaling-pathway (NES = 2.099, FDR = 0.028), cell-cycle (NES = 2.195, FDR = 0.034), axon-guidance (NES = 2.106, FDR = 0.038), pancreatic-cancer (NES = 2.049, FDR = 0.039),

TABLE 1 | Relationships between the expressions of the IRGs and the clinicopathological characteristics in cervical cancer.

Id	Age t(p)	Stage t(p)	Grade t(p)	M t(p)	N t(p)	T t(p)	BMI t(p)
LTA	−1.168 (0.293)	0.527 (0.608)	−1.462 (0.148)	1.984 (0.103)	0.07 (0.945)	0.668 (0.529)	−2.633 (0.010)
TFRC	−0.149 (0.886)	0.811 (0.431)	1.066 (0.290)	1.004 (0.364)	−1.127 (0.271)	−7.596 (1.456e − 04)	0.948 (0.347)
TYK2	0.31 (0.766)	−0.047 (0.963)	−0.252 (0.802)	−0.764 (0.500)	0.503 (0.618)	1.611 (0.162)	−0.068 (0.946)
DLL4	1.257 (0.235)	−0.259 (0.798)	0.038 (0.970)	−0.065 (0.952)	0.846 (0.403)	−0.818 (0.452)	0.249 (0.804)
CSK	0 (1.000)	−1.261 (0.224)	−0.23 (0.819)	1.887 (0.138)	0.102 (0.919)	2.566 (0.050)	1.782 (0.081)
JUND	0.721 (0.492)	1.523 (0.153)	0.764 (0.448)	−0.819 (0.472)	−0.067 (0.947)	−0.473 (0.659)	−1.485 (0.142)
NFATC4	1.557 (0.156)	0.66 (0.520)	0.178 (0.859)	−0.846 (0.458)	−0.016 (0.987)	0.933 (0.396)	−1.993 (0.050)
SBDS	−0.446 (0.669)	−0.019 (0.985)	0.643 (0.522)	0.884 (0.438)	−0.667 (0.511)	−0.676 (0.535)	1.376 (0.176)
FLT1	0.348 (0.736)	−0.347 (0.733)	−0.478 (0.634)	0.6 (0.588)	−0.068 (0.946)	−1.14 (0.298)	−0.061 (0.952)
IL17RD	0.521 (0.617)	−0.709 (0.490)	−0.475 (0.636)	0.46 (0.670)	0.734 (0.467)	3.091 (0.004)	−2.338 (0.022)
IL3RA	4.431 (1.404e − 04)	0.905 (0.378)	−1.234 (0.221)	0.094 (0.930)	−0.918 (0.365)	1.262 (0.261)	−2.366 (0.020)
SDC1	−2.491 (0.034)	0.228 (0.823)	2.122 (0.038)	1.035 (0.370)	−1.724 (0.098)	−2.498 (0.048)	1.131 (0.261)
PLAU	−0.922 (0.394)	−1.368 (0.189)	−0.344 (0.732)	1.174 (0.314)	0.517 (0.609)	−0.892 (0.416)	2.693 (0.010)
riskScore	−1.076 (0.312)	0.154 (0.880)	1.254 (0.215)	1.864 (0.103)	−0.695 (0.495)	−1.819 (0.138)	0.182 (0.856)

t, t-value of student's t-test; P, P-value of student's t-test.

and small-cell-lung-cancer (NES = 1.932, FDR = 0.048) were significantly relevant to the high risk group (**Figures 6B–F**), and graft-versus-host disease (NES = −1.916, FDR = 0.031), type-I-diabetes-mellitus (NES = −1.886, FDR = 0.034), allograft-rejection (NES = −1.989, FDR = 0.038), autoimmune-thyroid-disease (NES = −1.917, FDR = 0.038), hematopoietic-cell-lineage (NES = -1.831, FDR = 0.046), and asthma (NES = −1.928, FDR = 0.048) to the low risk group (**Figures 6G–L**). Finally, we evaluated the relationship between the IRG signature model and immune cell infiltration and thereby found that the infiltration of neutrophils was negatively correlated with the IRGs signature model for cervical cancer. However, the infiltration of B cells and neutrophils was positively correlated with this model for endometrial cancer (**Figures 6M–O**).

DISCUSSION

Many pre-existing scientific researches have demonstrated that the occurrence and progression of tumors are strongly related to immune cells and chemokines in the human body (Han et al., 2019; Rosenthal et al., 2019), which can be verified through the mechanism of immune escape (Luo et al., 2019). In some diseases and tumor biological processes, the changes of these immune biomarkers are clear and even can be predicted in the immune microenvironment (Silva-Santos et al., 2019). However, data on systematic and comprehensive molecular mechanisms in genome-wide profiling are limited for cervical cancer and endometrial cancer. Therefore, our study is designed to explore which types of IRGs show changes or may be going to change in patients with cervical cancer and endometrial cancer. Furthermore, we also investigated whether these differences could properly predict the clinical prognosis of patients and help to demonstrate the relationship between these IRGs and clinicopathological characteristics. This provides relevant information for us to develop better understanding on the biological changes of cervical and endometrial cancer.

Cancer cells have been shown to accumulate in inflammatory microenvironments, usually in the early stage of tumorigenesis (Hanahan and Weinberg, 2011); thus, it is genuinely helpful to identify differentially expressed IRGs in tumor tissues. Unfortunately, such studies on cervical cancer and endometrial cancer are rare. In this context, we extracted and calculated these differentially expressed IRGs for cervical cancer and endometrial cancer, respectively, resulting in a total of 146 differentially expressed IRGs that were shared by these two tumor types. Transcription factors (TFs), which also play a very important role in the human body, have been shown to regulate gene transcription at the nucleic acid level, thus affecting the expression of proteins (Lambert et al., 2018). After extraction and calculation of these differentially expressed TFs, 49 TFs were shared by these two diseases. Co-occurrence of differentially expressed IRGs and TFs is of great significance to guide our subsequent studies and those shared IRGs and TFs may suggest similar biological processes in both tumor types.

After that, enrichment analyses through GO and KEGG pathways were performed for these differentially expressed IRGs. The results showed genes were mainly enriched in the pathways including "cytokine-cytokine receptor interaction," "Ras signaling pathway," and "MAPK signaling pathway" for cervical cancer and the pathways including "cytokine-cytokine receptor interaction," "chemokine signaling pathway," and "PI3K-Akt signaling pathway" for endometrial cancer, indicating a possible relationship of these IRGs with tumor-associated development, progression, and invasion. Subsequently, we made further screening to select survival-associated IRGs. To investigate the relationship between these differentially expressed survival-associated IRGs and differentially expressed TFs, the regulatory network was constructed for these survival-associated IRGs and differentially expressed TFs, respectively. We found that the IRG LTA appeared in the regulatory network for both

TABLE 2 | Relationships between the expressions of the IRGs and the clinicopathological characteristics in endometrial cancer.

Id	Age t(p)	Stage t(p)	Grade t(p)	BMI t(p)	Pregnancies t(p)	Tamoxifen t(p)	Pharmaceutical t(p)	Diabetes t(p)	Hypertension t(p)	Radiation t(p)
LTA	0.908 (0.364)	2.775 (0.006)	−0.665 (0.507)	0.584 (0.561)	0.051 (0.959)	−0.384 (0.714)	0.354 (0.725)	0.351 (0.726)	−0.705 (0.481)	−0.26 (0.795)
PSMC4	−1.329 (0.184)	−2.344 (0.020)	−4.647 (5.004e−06)	−0.321 (0.749)	−1.626 (0.106)	−0.153 (0.882)	0.062 (0.951)	0.022 (0.982)	−1.175 (0.241)	−1.446 (0.150)
KAL1	−2.104 (0.036)	−1.024 (0.307)	−1.559 (0.120)	−2.168 (0.032)	−0.839 (0.403)	−1.101 (0.313)	−0.397 (0.693)	−0.536 (0.593)	0.088 (0.930)	−0.509 (0.611)
TNF	−2.04 (0.042)	−1.555 (0.122)	−1.91 (0.057)	0.44 (0.661)	−0.712 (0.477)	−0.886 (0.410)	−1.937 (0.058)	−1.535 (0.127)	−0.468 (0.640)	0.647 (0.518)
SBDS	−0.653 (0.514)	−3.11 (0.002)	−7.062 (6.468e−12)	1.414 (0.162)	−0.376 (0.707)	1.142 (0.294)	−0.28 (0.780)	2.266 (0.024)	2.271 (0.024)	−1.578 (0.115)
HDGF	−3.582 (3.826e−04)	−3.583 (4.11e−04)	−8.674 (6.483e−17)	1.154 (0.251)	−1.512 (0.132)	−0.69 (0.515)	−1.805 (0.076)	1.174 (0.242)	0.528 (0.597)	−1.28 (0.201)
LTB	−1.397 (0.163)	−0.288 (0.773)	−0.715 (0.475)	0.823 (0.413)	0.72 (0.472)	−0.882 (0.411)	−0.218 (0.828)	0.306 (0.760)	0.951 (0.342)	1.553 (0.121)
HTR3E	0.9 (0.369)	−1.025 (0.307)	−1.06 (0.290)	−1.06 (0.290)	1.053 (0.293)	0.511 (0.611)	0.873 (0.383)	1.017 (0.310)	0.814 (0.416)	−0.98 (0.328)
NR2F1	1.173 (0.242)	0.118 (0.906)	−1.776 (0.077)	0.858 (0.393)	0.953 (0.341)	0.511 (0.626)	1.934 (0.056)	0.24 (0.811)	0.876 (0.382)	0.023 (0.982)
NR3C1	−1.703 (0.090)	−3.071 (0.002)	−5.572 (4.443e−08)	1.248 (0.216)	−0.612 (0.541)	−1.169 (0.286)	−1.606 (0.113)	−0.234 (0.815)	0.705 (0.481)	−1.204 (0.230)
PGR	4.139 (4.121e−05)	5.709 (2.527e−08)	8.551 (3.876e−16)	−4.745 (5.258e−06)	0.449 (0.654)	4.321 (0.003)	4.974 (3.228e−06)	−2.129 (0.035)	−1.698 (0.090)	0.865 (0.387)
CBLC	−1.752 (0.081)	−1.01 (0.314)	−1.915 (0.056)	−0.431 (0.667)	−0.266 (0.790)	−0.331 (0.750)	−0.139 (0.890)	0.031 (0.976)	1.097 (0.273)	0.107 (0.915)
riskScore	1 (0.318)	−1 (0.319)	−1 (0.318)	−1 (0.318)	1 (0.318)	1 (0.318)	1 (0.318)	1 (0.318)	1 (0.318)	−1 (0.319)

t, t-value of student's t-test; P, P−value of student's t-test.

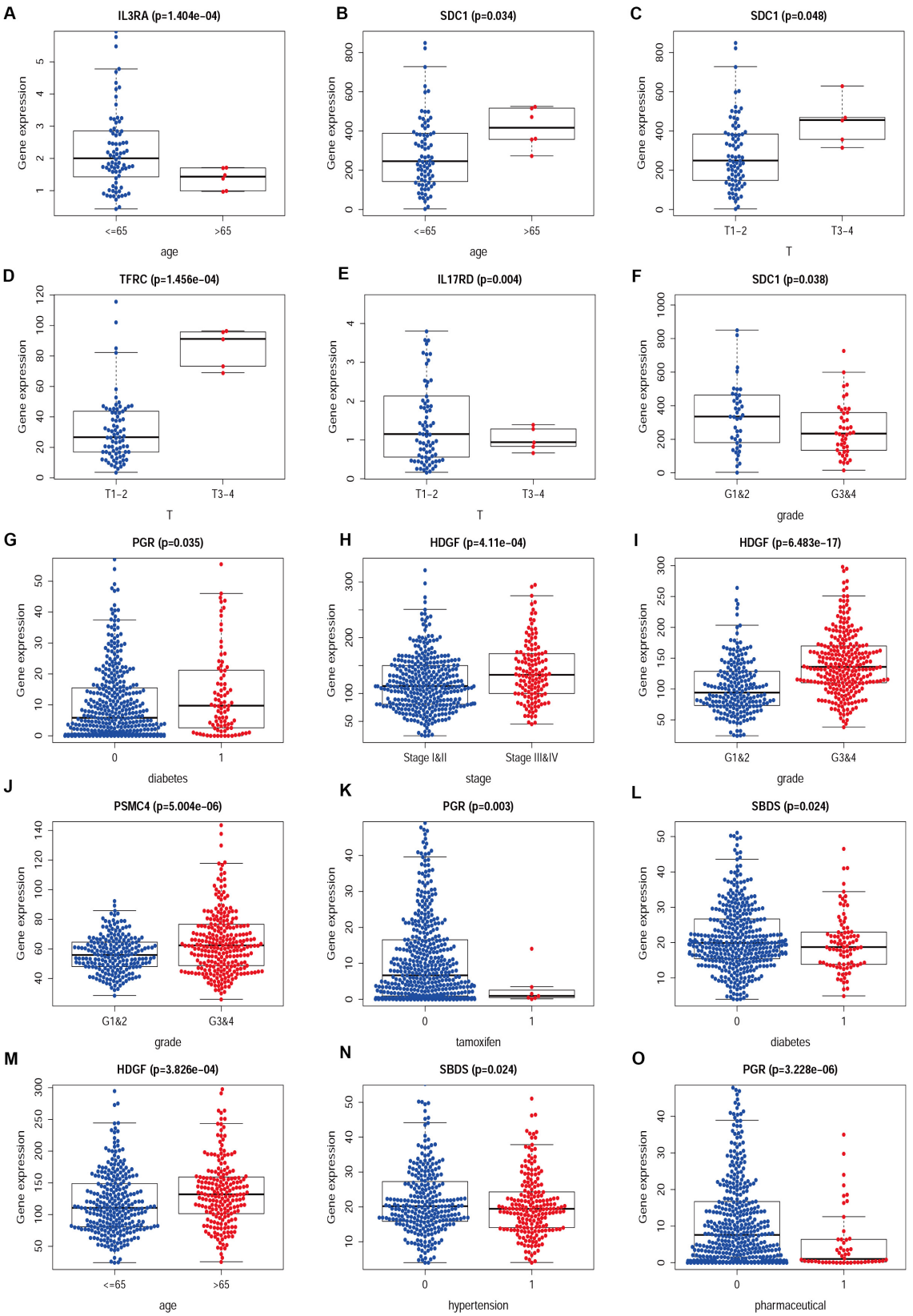


FIGURE 8 | The difference between IRGs expression profile and clinicopathological characteristics in CESC **(A–F)** and UCEC **(G–O)**. Show part only.

cancers. Previous studies have shown that the role of LTA varies with patient's condition. In patients with breast cancer, LTA can be used as a possible tumor marker to evaluate the prognosis (Kohaar et al., 2009). In contrast, in the study of gastric cancer (Mou et al., 2015), the occurrence of gastric cancer is related to the genetic variation of LTA. However, no sufficient data is available to draw a comprehensive picture to characterize this gene, and additional studies are needed to make further demonstration.

In living organisms, biological processes are often characterized by involvement in various genes, multiple courses, and continuous biological responses; therefore, it is difficult to predict and explain condition changes and prognosis with a single or a small number of gene expression profiles. In this case, an overall analysis of "gene signatures" involving different genes provide us with a good method for predicting the prognosis of patients. As early as in 2007, scholars (Chen et al., 2007) used a gene signature model to evaluate the clinical prognosis of non-small cell lung cancer (NSCLC), and the results demonstrated high reliability of this method. Since then, this kind of multigene signature model has been applied more frequently in other diseases, for example, ovary cancer (An et al., 2018), lung cancer (Liu et al., 2019a), colon cancer (Mo et al., 2019), lung adenocarcinoma (Wang et al., 2019), and colorectal cancer (Zhou et al., 2019). Based on these findings, we decided to establish an IRGs signature model for cervical cancer and endometrial cancer to evaluate the prognosis of patients.

For the establishment of gene signature models, 25/23 survival-associated IRGs were selected for cervical cancer and endometrial cancer, respectively. As a result, 13 IRGs in cervical cancer and 12 IRGs in endometrial cancer were found appropriate to establish the model. We used the two IRGs signature models to calculate the risk levels for each patient and found the differences in Kaplan-Meier survival curves were statistically significant between high- and low-risk groups for both cervical cancer (p = 6.464e-9) and endometrial cancer (p = 1e-11). In terms of survival and death, statistics of patients were also significantly different between high- and low-risk groups. These data obtained from evaluation models suggested that the IRGs' signature model may be a good way to assess the risk levels of patients; however, the area under the ROC curve (AUC) moderately only reached to about 0.738 in cervical cancer and 0.777 in endometrial cancer.

Exploring the relationship between clinicopathological characteristic and patient prognosis can provide us with more valuable information. Based on univariate and multivariate Cox regression analyses, we found the risk score resulted from the IRGs signature model could be considered an independent statistical measure to evaluate the overall survival (OS) in patients with cervical cancer ($P < 0.001$) (**Figures 7A,B**). Similar findings were obtained for the risk score resulted from the IRG signature model for endometrial cancer (**Figures 7C,D**), where independent clinical measures included age ($P < 0.001$), cancer stage ($P < 0.001$), tumor pathological grade ($P < 0.001$), the use of estrogen antagonist tamoxifen (Gaber-Wagener and Marth, 2020) ($P < 0.05$), and radiation therapy (Mirza, 2020) ($P < 0.05$). These results were highly consistent with previous studies (Casablanca et al., 2019; Trojano et al., 2019; Wu et al., 2019).

We then examined the correlation between the risk score and clinicopathological characteristics of the patients. However, there was no statistical difference in either the risk score proved by the IRGs' signature model or the clinicopathological characteristics between patients with cervical cancer and those with endometrial cancer, but statistical difference was observed in the IRG expression profiles and clinicopathological characteristics (**Tables 1**, **2**, **Figure 8** and **Supplementary Table S3**) for which further investigation could be considered. GSEA analysis for cervical cancer indicated that the high-risk group was significantly associated with the TGF beta signaling pathway, while in endometrial cancer the results showed relevance to erbb signaling pathway, cell cycle, axon guidance pathway, pancreatic cancer, and small-cell lung cancer. These pathways were associated with tumor development and progression, suggesting that these molecular pathways were likely to be activated in high-risk groups. Thus, the validity of this IRGs signature model for predicting risk scores was well established.

By evaluating the relationship between the risk score provided by the IRGs' signature model and immune cell infiltration, we found that neutrophil infiltration was negatively correlated with risk scores in cervical cancer; however, the infiltration of immune B cells and neutrophils were positively correlated with risk scores in endometrial cancer. Studies (Dong et al., 2019) have shown that the ratio of neutrophils to lymphocytes is an independent measure to determine prognosis and lymph node metastasis in endometrial cancer (Aoyama et al., 2019). It is also reported that (Wisdom et al., 2019) neutrophils can increase the resistance of tumor cells to radiation therapy, and neutrophil-lymphocyte ratio can be considered as a predictor in stage IVB or recurrent cervical cancer patients treated by chemotherapy. Neutrophil-lymphocyte ratio $\geq$ 3.6 has been identified as an independent predictor of poor oncologic outcomes with respect to OS (Farzaneh et al., 2019; Ittiamornlert and Ruengkhachorn, 2019). These data suggest that the relationship between the risk score provided by the IRG signature model and neutrophil infiltration is well-established. The determination for B lymphocytes still requires data due to lack of relevant studies on cervical and endometrial cancers.

CONCLUSION

In conclusion, our analyses for this IRG signature model still leave some limitations for us to improve. First of all, in cervical and endometrial cancer, there are different pathological types, but no different pathological type models have been developed yet. As such, there might be differences for some special pathological types. Second, we just selected immune-related genes to establish the IRG signature model, whereas in the human body, the occurrence and progression of cancer or other diseases is a comprehensive process that involves nucleic acid transcriptome, proteomics, minerals, and other important elements. As the present study focuses on transcriptome of RNA, there may exist a certain selection bias in this model. Third, our model lacks

independent databases for verification which is why we are very careful to draw these conclusions. At last, these models are provided with *in vivo* and *in vitro* experimental data. Although the model developed by us has some shortages, we still hope to provide new ideas and guidance for future researches in the treatment of cervical cancer and endometrial cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov.

AUTHOR CONTRIBUTIONS

HD and G-LF contributed to the conception of the study. Y-XY contributed significantly to analysis and manuscript preparation. HD, X-XX, and OM performed the data analyses and wrote the manuscript. WZ helped perform the analysis with constructive discussions. Final approval of manuscript by all authors.

ACKNOWLEDGMENTS

We would like to thank the openly and friendly free public TCGA databases.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00725/full#supplementary-material

REFERENCES

An, Y., Bi, F., You, Y., Liu, X., and Yang, Q. (2018). Development of a novel autophagy-related prognostic signature for serous ovarian cancer. *J. Cancer* 9:4058. doi: 10.7150/jca.25587

Aoyama, T., Takano, M., Miyamoto, M., Yoshikawa, T., Kato, K., Sakamoto, T., et al. (2019). Pretreatment neutrophil-to-lymphocyte ratio was a predictor of lymph node metastasis in endometrial cancer patients. *Oncology* 96, 259–267. doi: 10.1159/000497184

Barroso-Sousa, R., Keenan, T. E., Pernas, S., Exman, P., Jain, E., Garrido-Castro, A. C., et al. (2020). Tumor mutational burden and PTEN alterations as molecular correlates of response to PD-1/L1 blockade in metastatic triple-negative breast cancer. *Clin. Cancer Res.* 26:clincanres.3507.2019.

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

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

Casablanca, Y., Tian, C., Powell, M., Winterhoff, B., Chan, J., Shriver, C., et al. (2019). 50 Age, histology and stage predict survival following adjuvant chemotherapy and radiation versus radiation alone in high-risk endometrial cancer: a study based on portec-3 criteria. *Intern. J. Gynecol. Cancer* 29(Suppl. 3), A28.3–A29.

Chen, H.-Y., Yu, S.-L., Chen, C.-H., Chang, G.-C., Chen, C.-Y., Yuan, A., et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.* 356, 11–20.

Chu, R., Liu, S. Y., Vlantis, A. C., Van Hasselt, C. A., Ng, E. K. W., Fan, M. D., et al. (2015). Inhibition of Foxp3 in cancer cells induces apoptosis of thyroid cancer cells. *Mol. Cell. Endocrinol.* 399, 228–234. doi: 10.1016/j.mce.2014.10.006

Cohen, P. A., Jhingran, A., Oaknin, A., and Denny, L. (2019). Cervical cancer. *Lancet* 393, 169–182.

Cosper, P. F., Mcnair, C., González, I., Wong, N., Knudsen, K. E., Chen, J. J., et al. (2019). Decreased local immune response and retained HPV gene expression during chemoradiotherapy are associated with treatment resistance and death from cervical cancer. *Intern. J. Cancer* 146, 2047–2058. doi: 10.1002/ijc.32793

Crncec, I., Modak, M., Gordziel, C., Svinka, J., Scharf, I., Moritsch, S., et al. (2018). STAT1 is a sex-specific tumor suppressor in colitis-associated colorectal cancer. *Mol. Oncol.* 12, 514–528. doi: 10.1002/1878-0261.12178

Crusz, S. M., and Miller, R. E. (2020). Targeted therapies in gynaecological cancers. *Histopathology* 76, 157–170. doi: 10.1111/his.14009

Dong, Y., Cheng, Y., and Wang, J. (2019). The ratio of neutrophil to lymphocyte is a predictor in endometrial cancer. *Open Life Sci.* 14, 110–118. doi: 10.1515/biol-2019-0012

Farzaneh, F., Faghih, N., Hosseini, M. S., Arab, M., Ashrafganjoei, T., and Bahman, A. (2019). Evaluation of neutrophil-lymphocyte ratio as a prognostic factor in cervical intraepithelial neoplasia recurrence. *Asian Pacif. J. Cancer Prevent.* 20:2365. doi: 10.31557/apjcp.2019.20.8.2365

Feng, R.-M., Zong, Y.-N., Cao, S.-M., and Xu, R.-H. (2019). Current cancer situation in China: good or bad news from the 2018 global cancer statistics? *Cancer Commun.* 39:22. doi: 10.1186/s40880-019-0368-6

Gaber-Wagener, A., and Marth, C. (2020). "Role of hormonal therapy in advanced stage endometrial cancer," in *Management of Endometrial Cancer*, ed. M. Mirza (Cham: Springer), 243–248. doi: 10.1007/978-3-319-64513-1_17

Gainor, J., Rizvi, H., Aguilar, E. J., Skoulidis, F., Yeap, B., Naidoo, J., et al. (2020). Clinical activity of programmed cell death 1 (PD-1) blockade in never, light, and heavy smokers with non-small cell lung cancer and PD-L1 Expresion= 50%. *Ann. Oncol.* 31, 404–411. doi: 10.1016/j.annonc.2019.11.015

Grywalska, E., Sobstyl, M., Putowski, L., and Roliński, J. (2019). Current possibilities of gynecologic cancer treatment with the use of immune checkpoint inhibitors. *Intern. J. Mol. Sci.* 20:4705. doi: 10.3390/ijms20194705

Han, Q., Zhao, H., Jiang, Y., Yin, C., and Zhang, J. (2019). HCC-derived exosomes: critical player and target for cancer immune escape. *Cells* 8:558. doi: 10.3390/cells8060558

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

Hou, Y., Li, X., Li, Q., Xu, J., Yang, H., Xue, M., et al. (2018). STAT 1 facilitates oestrogen receptor α transcription and stimulates breast cancer cell proliferation. *J. Cell. Mol. Med.* 22, 6077–6086. doi: 10.1111/jcmm.13882

Irvine, D. J., and Dane, E. L. (2020). Enhancing cancer immunotherapy with nanomedicine. *Nat. Rev. Immunol.* 20, 321–334. doi: 10.1038/s41577-019-0269-6

Ishikawa, M., Nakayama, K., Nakamura, K., Yamashita, H., Ishibashi, T., Minamoto, T., et al. (2020). High PD-1 expression level is associated with an unfavorable prognosis in patients with cervical adenocarcinoma. *Arch. Gynecol. Obstet.* 302, 209–218. doi: 10.1007/s00404-020-05589-0

Ittiamornlert, P., and Ruengkhachorn, I. (2019). Neutrophil-lymphocyte ratio as a predictor of oncologic outcomes in stage IVB, persistent, or recurrent cervical cancer patients treated by chemotherapy. *BMC Cancer* 19:51. doi: 10.1186/s12885-019-5269-1

June, C. H., O'connor, R. S., Kawalekar, O. U., Ghassemi, S., and Milone, M. C. (2018). CAR T cell immunotherapy for human cancer. *Science* 359, 1361–1365.

Kim, H. J. (2017). Current status and future prospects for human papillomavirus vaccines. *Archiv. Pharm. Res.* 40, 1050–1063. doi: 10.1007/s12272-017-0952-8

Kohaar, I., Tiwari, P., Kumar, R., Nasare, V., Thakur, N., Das, B. C., et al. (2009). Association of single nucleotide polymorphisms (SNPs) in TNF-LTA locus with breast cancer risk in Indian population. *Breast Cancer Res. Treat.* 114:347. doi: 10.1007/s10549-008-0006-5

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665.

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B., Aulakh, L. K., et al. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 357, 409–413.

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, W., Katoh, H., Wang, L., Yu, X., Du, Z., Yan, X., et al. (2013). FOXP3 regulates sensitivity of cancer cells to irradiation by transcriptional repression of BRCA1. *Cancer Res.* 73, 2170–2180. doi: 10.1158/0008-5472.can-12-2481

Li, Y., Li, D., Yang, W., Fu, H., and Liu, Y. (2016). Overexpression of the transcription factor FOXP3 in lung adenocarcinoma sustains malignant character by promoting G1/S transition gene CCND1. *Tumor Biol.* 37, 7395–7404.

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205.

Liu, Y. (2019). Immune response characterization of endometrial cancer. *Oncotarget* 10:982. doi: 10.18632/oncotarget.26630

Liu, Y., Wu, L., Ao, H., Zhao, M., Leng, X., Liu, M., et al. (2019a). Prognostic implications of autophagy-associated gene signatures in non-small cell lung cancer. *Aging* 11:11440. doi: 10.18632/aging.102544

Liu, Y., Wu, L., Tong, R., Yang, F., Yin, L., Li, M., et al. (2019b). PD-1/PD-L1 inhibitors in cervical cancer. *Front. Pharmacol.* 10:65. doi: 10.3389/fonc.2019.0065

Luo, X., Donnelly, C. R., Gong, W., Heath, B. R., Hao, Y., Donnelly, L. A., et al. (2019). HPV16 drives cancer immune escape via NLRX1-mediated degradation of STING. *J. Clin. Invest.* 130, 1635–1652. doi: 10.1172/jci129497

Lynam, S., Lugade, A. A., and Odunsi, K. (2019). Immunotherapy for gynecologic cancer: current applications and future directions. *Clin. Obstetr. Gynecol.* 63, 48–63. doi: 10.1097/grf.0000000000000513

Ma, G.-F., Chen, S.-Y., Sun, Z.-R., Miao, Q., Liu, Y.-M., Zeng, X.-Q., et al. (2013). FoxP3 inhibits proliferation and induces apoptosis of gastric cancer cells by activating the apoptotic signaling pathway. *Biochem. Biophys. Res. Commun.* 430, 804–809. doi: 10.1016/j.bbrc.2012.11.065

Ma, J., Sun, G., Zhu, P., Liu, S., Ou, M., Chen, Z., et al. (2019). Determination of the complexity and diversity of the TCR β-chain CDR3 repertoire in bladder cancer using high-throughput sequencing. *Oncol. Lett.* 17, 3808–3816.

Martin, J. D., Cabral, H., Stylianopoulos, T., and Jain, R. K. (2020). Improving cancer immunotherapy using nanomedicines: progress, opportunities and challenges. *Natu. Rev. Clin. Oncol.* 17, 251–266. doi: 10.1038/s41571-019-0308-z

Maskey, N., Thapa, N., Maharjan, M., Shrestha, G., Maharjan, N., Cai, H., et al. (2019). Infiltrating CD4 and CD8 lymphocytes in HPV infected uterine cervical milieu. *Cancer Manag. Res.* 11:7647. doi: 10.2147/cmar.s217264

Matanes, E., and Gotlieb, W. H. (2019). Immunotherapy of gynecological cancers. *Best Pract. Res. Clin. Obstetr. Gynaecol.* 60, 97–110. doi: 10.1016/j.bpobgyn.2019.03.005

Miller, D. S., Randall, M. E., and Filiaci, V. (2020). Progress in endometrial cancer: contributions of the former gynecologic oncology group. *Gynecol. Oncol.* 57, 312–322. doi: 10.1016/j.ygyno.2020.01.012

Mirza, M. R. (2020). Role of radiation therapy. *Manag. Endomet. Cancer* 223–229. doi: 10.1007/978-3-319-64513-1_15

Mo, S., Dai, W., Xiang, W., Li, Y., Feng, Y., Zhang, L., et al. (2019). Prognostic and predictive value of an autophagy-related signature for early relapse in stages I-III colon cancer. *Carcinogenesis* 40, 861–870. doi: 10.1093/carcin/bgz031

Mo, Z., Liu, J., Zhang, Q., Chen, Z., Mei, J., Liu, L., et al. (2016). Expression of PD-1, PD-L1 and PD-L2 is associated with differentiation status and histological type of endometrial cancer. *Oncol. Lett.* 12, 944–950. doi: 10.3892/ol.2016.4744

Mou, X., Li, T., Wang, J., Ali, Z., Zhang, Y., Chen, Z., et al. (2015). Genetic variation of BCL2 (rs2279115), NEIL2 (rs804270), LTA (rs909253), PSCA (rs2294008) and PLCE1 (rs3765524, rs10509670) genes and their correlation to gastric cancer risk based on universal tagged arrays and Fe3O4 magnetic nanoparticles. *J. Biomed. Nanotechnol.* 11, 2057–2066. doi: 10.1166/jbn.2015.2113

Rosenthal, R., Cadieux, E. L., Salgado, R., Al Bakir, M., Moore, D. A., Hiley, C. T., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature* 567, 479–485.

Rubinstein, M. M., and Makker, V. (2020). Optimizing immunotherapy for gynecologic cancers. *Curr. Opin. Obstetr. Gynecol.* 32, 1–8. doi: 10.1097/gco.0000000000000603

Shen, R., Pham, C., Wu, M., Munson, D., and Aftab, B. (2019). 8CD19 chimeric antigen receptor (CAR) engineered epstein-barr virus (EBV) specific T cells-an off-the-shelf, allogeneic CAR T-cell immunotherapy platform. *Cytotherapy* 21:S11.

Silva-Santos, B., Mensurado, S., and Coffelt, S. B. (2019). γδ T cells: pleiotropic immune effectors with therapeutic potential in cancer. *Nat. Rev. Cancer* 19, 392–404. doi: 10.1038/s41568-019-0153-5

Stevanovic, S., Draper, L., Langhan, M. M., Campbell, T. E., Kwong, M. L., Wunderlich, J. R., et al. (2015). Complete regression of metastatic cervical cancer after treatment with human papillomavirus-targeted tumor-infiltrating T cells. *J. Clin. Oncol.* 33, 1543–1550. doi: 10.1200/jco.2014.58.9093

Tian, X., Guan, W., Zhang, L., Sun, W., Zhou, D., Lin, Q., et al. (2018). Physical interaction of STAT1 isoforms with TGF-β receptors leads to functional crosstalk between two signaling pathways in epithelial ovarian cancer. *J. Exper. Clin. Cancer Res.* 37:103.

Trojano, G., Olivieri, C., Tinelli, R., Damiani, G. R., Pellegrino, A., and Cicinelli, E. (2019). Conservative treatment in early stage endometrial cancer: a review. *Acta Bio Med. Atenei Parmen.* 90:405.

Varikuti, S., Oghumu, S., Elbaz, M., Volpedo, G., Ahirwar, D. K., Alarcon, P. C., et al. (2017). STAT1 gene deficient mice develop accelerated breast cancer growth and metastasis which is reduced by IL-17 blockade. *Oncoimmunology* 6:e1361088. doi: 10.1080/2162402x.2017.1361088

Verhoeven, Y., Tilborghs, S., Jacobs, J., De Waele, J., Quatannens, D., Deben, C., et al. (2020). The potential and controversy of targeting STAT family members in cancer. *Semin. Cancer Biol.* 60, 41–56. doi: 10.1016/j.semcancer.2019.10.002

Vici, P., Pizzuti, L., Mariani, L., Zampa, G., Santini, D., Lauro, L. D., et al. (2016). Targeting immune response with therapeutic vaccines in premalignant lesions and cervical cancer: hope or reality from clinical studies. *Expert Rev. Vaccines* 15, 1327–1336.

Wang, Y., and Li, G. (2019). PD-1/PD-L1 blockade in cervical cancer: current studies and perspectives. *Front. Med.* 13, 438–450. doi: 10.1007/s11684-018-0674-4

Wang, Y., Zhang, Q., Gao, Z., Xin, S., Zhao, Y., Zhang, K., et al. (2019). A novel 4-gene signature for overall survival prediction in lung adenocarcinoma patients with lymph node metastasis. *Cancer Cell Intern.* 19:100.

Wisdom, A. J., Hong, C. S., Lin, A. J., Xiang, Y., Cooper, D. E., Zhang, J., et al. (2019). Neutrophils promote tumor resistance to radiation therapy. *Proc. Natl. Acad. Sci. U.S.A.* 116, 18584–18589.

Wu, Y., Sun, W., Liu, H., and Zhang, D. (2019). Age at menopause and risk of developing endometrial cancer: a meta-analysis. *Biomed. Res. Intern.* 2019:8584130.

Yang, A., Farmer, E., Wu, T. C., and Hung, C. (2016). Perspectives for therapeutic HPV vaccine development. *J. Biomed. Sci.* 23:75.

Yang, S., Liu, Y., Li, M.-Y., Ng, C. S., Yang, S.-L., Wang, S., et al. (2017). FOXP3 promotes tumor growth and metastasis by activating Wnt/β-catenin signaling pathway and EMT in non-small cell lung cancer. *Mol. Cancer* 16:124.

Yang, S., Wu, Y., Deng, Y., Zhou, L., Yang, P., Zheng, Y., et al. (2019). Identification of a prognostic immune signature for cervical cancer to predict survival and response to immune checkpoint inhibitors. *Oncoimmunology* 8:e1659094.

Zhang, H., and Sun, H. (2010). Up-regulation of Foxp3 inhibits cell proliferation, migration and invasion in epithelial ovarian cancer. *Cancer Lett.* 287, 91–97.

Zhang, J., Wang, F., Liu, F., and Xu, G. (2020). Predicting STAT1 as a prognostic marker in patients with solid cancer. *Therapeut. Adv. Med. Oncol.* 12:1758835920917558.

Zhang, X., Li, X., Tan, F., Yu, N., and Pei, H. (2017). STAT1 Inhibits miR-181a expression to suppress colorectal cancer cell proliferation through PTEN/Akt. *J. Cell. Biochem.* 118, 3435–3443.

Zhou, X., and Ling, Z. (2019). Systematic analysis of tumor-infiltrating immune cells in human endometrial cancer: a retrospective study. *medRxiv* [Preprint], doi: 10.1101/19003707

Zhou, Z., Mo, S., Dai, W., Ying, Z., Zhang, L., Han, L., et al. (2019). Development and validation of an autophagy score signature for the prediction of post-operative survival in colorectal cancer. *Front. Oncol.* 9:878. doi: 10.3389/fonc.2019.00878

Zuo, T., Liu, R., Zhang, H., Chang, X., Liu, Y., Wang, L., et al. (2007). FOXP3 is a novel transcriptional repressor for the breast cancer oncogene SKP2. *J. Clin. Invest.* 117, 3765–3773.

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 Oncology

ORIGINAL RESEARCH
published: 28 July 2020
doi: 10.3389/fonc.2020.01269

The Better Survival of MSI Subtype Is Associated With the Oxidative Stress Related Pathways in Gastric Cancer

Lei Cai†, Yeqi Sun†, Kezhou Wang, Wenbin Guan, Juanqing Yue, Junlei Li, Ruifen Wang and Lifeng Wang**

Department of Pathology, Xinhua Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Xin Wang,
Houston Methodist Hospital, United States
Sing-Wai Wong,
University of North Carolina at Chapel Hill, United States
Jie Yan,
Yale University School of Medicine, United States

***Correspondence:**
Ruifen Wang
wangruifen@xinhuamed.com.cn
Lifeng Wang
wanglifeng@xinhuamed.com.cn

†These authors have contributed equally to this work

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

Received: *20 February 2020*
Accepted: *18 June 2020*
Published: *28 July 2020*

Citation:
Cai L, Sun Y, Wang K, Guan W, Yue J, Li J, Wang R and Wang L (2020) The Better Survival of MSI Subtype Is Associated With the Oxidative Stress Related Pathways in Gastric Cancer. Front. Oncol. 10:1269. doi: 10.3389/fonc.2020.01269

Background: Gastric cancer (GC) is the third leading fatal cancer in the world and its incidence ranked second among all malignant tumors in China. The molecular classification of GC, proposed by the The Cancer Genome Atlas (TCGA), was added to the updated edition (2019) of WHO classification for digestive system tumor. Although MSI and EBV subtypes appeared as ever-increasingly significant roles in immune checkpoint inhibitor therapy, the underlying mechanisms are still unclear.

Methods: We systematically summarized the relationship between EBV, d-MMR/MSI-H subtypes and clinicopathological parameters in 271 GC cases. Furthermore, GSE62254/ACRG and TCGA-STAD datasets, originated from Gene Expression Omnibus (GEO) and TCGA respectively, were analyzed to figure out the prognosis related molecular characteristics by bioinformatics methods.

Results: Patients with MSI subtype had better prognosis than the MSS subtype ($P = 0.013$) and considered as an independent biomarker by the univariate analysis ($P = 0.017$) and multivariate analysis ($P = 0.050$). While there was no significant difference between EBV positive and negative tissues ($P = 0.533$). The positive prognostic value conferred by MSI in different cohorts was revalidated via the clinical analysis of GSE62254/ACRG and TCGA-STAD datasets regardless of race. Then key gene module that tightly associated with better status and longer OS time for MSI cases was obtained from weighted gene co-expression network analysis(WGCNA). *NUBP2* and *ENDOG* were screened from the gene cluster and oxidative phosphorylation, reactive oxygen species(ROS) and glutathione metabolism were analyzed to be the differential pathways in their highly expressed groups.

Conclusions: Our results manifested the significant prognostic value of MSI in Chinese GC cohort and comparisons with other populations. More opportunities to induce apoptosis of cancer cells, led by the unbalance between antioxidant system and ROS accumulation, lay foundations for unveiling the better prognosis in MSI phenotype through the bioinformatics analysis.

Keywords: gastric cancer, microsatellite instability, Epstein-Barr virus, prognosis, bioinformatics analysis

INTRODUCTION

Gastric cancer (GC), a highly heterogeneous disease, is the third most common cause of cancer-related death worldwide with a particular high incidence and mortality in Asia (1). Although the operation, chemotherapy and radiotherapy were widely used, the therapeutical efficacy was still such limited for some patients. The advent and development of next-generation sequencing (NGS) has revolutionized our understanding of its pathogenesis and molecular alterations. TCGA had presented four distinct subtypes-Epsterin-Barr virus (EBV), microsatellite instability (MSI), chromosomal instability(CIN) and genome stable(GS) through comprehensive molecular evaluation of 295 primary gastric cancer (2–4). Recognition of molecular subtypes can indeed help to establish a new paradigm of cancer therapeutics especially as the development of immunotherapy. Nevertheless, each molecular subtype had divergent response and therapeutical effects to immunotherapy. Impressive results from some clinical trials have demonstrated that solid tumors with MSI phenotype had more significant responses to anti-PD1 inhibitors than that with Microsatellite Stable (MSS) in patients who failed conventional therapy and GC was one of them (5–7). Compared with GS and CIN, metastatic GC patients with the MSI and EBV subtype manifested a dramatic response to PD1 inhibitor (8). Furthermore, GC patients with MSS status could benefit from 5-FU-based adjuvant chemotherapy in TNM stage II–III (9). Therefore, correct evolution of EBV infection and MSI status could be served as a potential biomarker for anti-PD1/PD-L1 targeted therapy and 5-FU based traditional chemotherapy in GC.

High-Microsatellite Instability (MSI-H) phenotype has been widely acknowledged to be the predictive factor for immunotherapy as its high PD-L1 expression. Some researchers has represented that MSI is an independent predictive factor while others observed that there are no significant difference of prognosis between divergent MSI status (10–16). The complex interactions that involved in the p53 signal pathways or E2F/DP1 transcription factors may largely contribute to the outcome (17–19). It also has been revealed that the EBV infection may be connected with the GC carcinogenesis at an early stage though the exact mechanism is still unclear. Enhanced understanding of the clinicopathological and prognostic implications of these molecular subtypes will assist to acquire the reasonable evaluation of the biological behavior of tumors. In fact, considerable literatures have investigated the relations between MSI phenotype and their prognosis but the conclusion is still in the air (15). So did the similar condition in EBV subtype. For example, Ahn et al. and Setia et al. separately revealed a significant survival advantage for EBV associated gastric cancer (EBVaGC), whereas Genitsch et al. showed that there was no association between EBV infection and clinical outcome of GC patients (10, 20, 21).

Moreover, results from Shen et al. manifested that EBV+ patients had a poorer OS than EBV- patients (12). The discrepancies may be due to a number of factors, such as different ethnic background of the enrolled patients or multiple methods for detecting the presence of EBV/MSI alteration. Hence, more robust tools for detection of EBV/MSI phenotype and better-tailored investigation should be applied to elucidate the real contributions of them to prognosis in various regions.

Since most aforementioned data about MSI and EBV (+) GC are derived from studies of western population, little investigations have reported for Chinese cohorts. In this study, we adopted the most widely used methods to identify EBV infection (EBV–encoded RNA by *in situ* hybridization) and Mismatch Repair deficiency/High-Microsatellite Instability (d-MMR/MSI-H) status (joint application of immunohistochemical staining & PCR-based MSI testing according to NCI panel) in 279 Chinese GC patients. Then, the clinicopathological characteristics and prognostic significance of EBV+ and MSI were in-depthly explored in present study. In addition, the data derived from TCGA and GEO had been used to compare the clinical differences among diverse cohorts. The associated molecular mechanisms were also analyzed by utilization of the bioinformatics.

MATERIALS AND METHODS

Patients and Samples

A total of 279 consecutive cases with gastric cancer were included at our institution. For each patient, all available archives including clinical data, hematoxylin and eosin (H&E)-stained slides and formalin-fixed paraffin embedded (FFPE) blocks were collected in this study. These patients were treated with surgical resection of primary gastric tumors between April 2010 and December 2015. Those diagnoses were confirmed by routine pathological examination after surgery. Ethics approval was obtained from the Ethics Committee of Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine. None of the patients received preoperative radiotherapy or chemotherapy. Pathologic parameters of all cases were reassessed according to the 4th edition of WHO classification for stomach tumors. The follow-up time was from initial diagnosis to September 2017 (range from 3 to 89 months).

Data Collection

Refer to our data size, the clinical information and expression profiling of GSE62254 about 300 samples was downloaded from Gene Expression Omnibus (GEO) database and about 315 samples from TCGA (The Cancer Genome Atlas). The expression profiling of TCGA was downloaded from the UCSC Xena browser (http://xena.ucsc.edu/) and FPKM normalized. Their corresponding clinical information obtained from the online tool cbioportal (http://www.cbioportal.org/). All cases had determined subtypes but overall survival (OS) time of

Abbreviations: GC, Gastric Cancer; MSI, Microsatellite Instability; TCGA, The Cancer Genome Atlas; EBV, Epstein-Barr virus; EBVaGC, Epstein-Barr virus associated gastric cancer; TCGA-STAD, Stomach adenocarcinoma samples in The Cancer Genome Atlas; GEO, Gene Expression Omnibus; ACRG, Asian Cancer Research Group; MSS, Microsatellite Stable; WGCNA, Weighted Gene Co-expression Network Analysis; ROS, Reactive Oxygen Species; MMR, Mismatch Repair; d-MMR/MSI-H, Mismatch Repair deficiency/High-Microsatellite Instability; p-MMR/MSI-L, MMR-proficient/Low-Microsatellite Instability; KEGG, The Kyoto Encyclopedia of Gene and Genomes; GO, Gene ontology; GSEA, Gene Set Enrichment Analysis; GSVA, Gene Set Variation Analysis; OXPHOS, Oxidative Phosphorylation System; FPKM, Fragments Per Kilobase per Million.

four cases and the AJCC pathological tumor stage of two cases were not available. Thus, 309 cases were analyzed for the clinicopathological characteristics and prognostic significance.

MMR Immunochemistry (IHC) and EBV *in situ* Hybridization (ISH)

One representative FFPE block of the cancer region in each case was chosen for IHC and ISH analysis. Unstained 4-um thick tissue sections were tested by IHC antibodies to MLH1 (Clone M1, ready-to-use; Roche), PMS2 (Clone EPR3947, ready-to-use; Roche), MSH2 (Clone 219-1129, ready-to-use; Roche) and MSH6 (Clone 44, ready-to-use; Roche) for detection of MMR status, as well as chromogenic ISH with EBV-encoded RNA (EBER, Ventana) probe to prove EBV infection, using Benchmark automated staining device (Ventana Medical Systems, Roche, Switzerland) according to the manufacturer' instructions. All IHC and ISH stained sections were reviewed and scored independently by two professional digestive pathologists (WLF and WRF) without knowledge of previous clinical or pathological parameters.

The slides were evaluated as follows: at least one of the MMR proteins (MLH1, MSH2, MSH6, and PMS2) with complete loss of nuclear reactivity in tumor cells but consistently preserved nuclear staining in background non-tumor cells was taken as d-MMR(aberrant expression). When the tumor cells demonstrated intact nuclear immunostaining of all four MMR proteins, the tumor was judged as p-MMR (normal expression). For EBER, tumors with strong blue-black nuclear staining were considered positive.

DNA Extraction and MSI Analysis

Total DNA was isolated from FFPE tumor and paired normal tissue samples though the DNA extraction kit (TIANGEN, Beijing, China) following the manufacturer's recommendation and was used for subsequent multiplex fluorescent PCR. MSI status was assessed with the amplification of six mononucleotide repeat markers (BAT25, BAT26, NR21, NR24, MONO27, and NR 27) described either in NCI (National Cancer Institute) - or Promega- panel. In addition, the final panel also contained one gender loci (Amel) and two pentanucleotide repeat markers (Peta C and Peta D) as internal controls. Co-amplification of these targets was performed on ABI 7500 using a 25 μl reaction volume advised by MSI-testing reagent kit (SINOMDgene, Beijing, China). The PCR conditions were carried out according to the operation protocols. Fluorescent PCR products were analyzed by capillary electrophoresis using an ABI 3500DX Genetic Analyzer (Applied Biosystems) and Genemaker software 2.0 (SINOMDgene, Beijing, China).

Tumors with instability at two or more of these 6 markers were defined as MSI-H, while those without instability or showing instability at only one marker were classified as MSS and Low-Microsatellite Instability(MSI-L) tumors, respectively.

Construction of Weighted Gene Co-expression Network

To identify the key module that most associated with the OS time and status in 51 MSI cases and then investigate the underlying molecular connections, the weighted gene co-expression network analysis(WGCNA) was performed on the TCGA-STAD dataset (22). The variances of all genes were calculated and approximately top 6,000 genes were performed by use of the WGCNA R package.

To identify co-expressed genes, WGCNA use the soft thresholding power to determine the correlations between genes via the Sigmoid or Exponential function. In this study, the soft thresholding procedure was firstly performed to set the cutoff to identify the modules. Secondly, in order to identify the adjacent gene modules, the topological overlap dissimilarity measure (TOM) was used to calculate the correlation among genes. The hierarchical clustering was constructed and the minimum size was appropriately set to meet the different datasets' need. Thirdly, connecting modules to the external clinical traits could show us the key module that most associated with the OS time and status traits. After the key module had been identified, genes were put into the GO and KEGG enrichment analysis.

Identification of the Hub Genes

After acquiring the module-trait relationships, the critical module was emerged. The pink module was obtained from the WGCNA that consisted of 250 nodes and 1,081 edges. This edge file was put into the Cytoscape software and constructed the gene co-expression network. The top five genes were *NUBP2, CTU1, ENDOG, SSNA1,* and *BCL7C* that the GS $>$ 0.14 and MM $>$ 0.75. But in the further validation in the GSE62254 dataset, the *CTU1* was not detected. Therefore, only four genes were considered the hub genes to be manifested and we selected the *NUBP2* and *ENDOG* as the typical hub genes to be in-depth functionally analyzed in TCGA-STAD dataset.

Function Enrichment Analysis

After the key module was identified, genes were analyzed by The Kyoto Encyclopedia of Gene and Genomes (KEGG) and Gene ontology (GO). The clusterprofiler package in R software was used to realize these two gene enrichment analysis and the $P < 0.050$ (23).

Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA)

To probe the function of *NUBP2* and *ENDOG* in the dataset and elucidate their role in the good prognosis of MSI phenotype well, all MSI cases were divided into *NUBP2* or *ENDOG* high and low expression groups according to the median expression. The GSEA software downloaded from http://software.broadinstitute.org/gsea/ and annotated gene set c2.cp.kegg.v7.0.symbols.gmt. The top five significant pathways that derived from GSEA ($P < 0.05$) were shown in one graphic. GSVA was carried out in the high and low expression by the GSVA R package that also annotated gene set c2.cp.kegg.v7.0.symbols.gmt.

Statistical Analysis

Clinicopathological parameters between groups were assessed for differences using the Pearson's X^2 test, Yate's correction or Fisher's exact test. The Kaplan-Meier method (and the log-rank test) as well as Cox's proportional hazards regression model were used

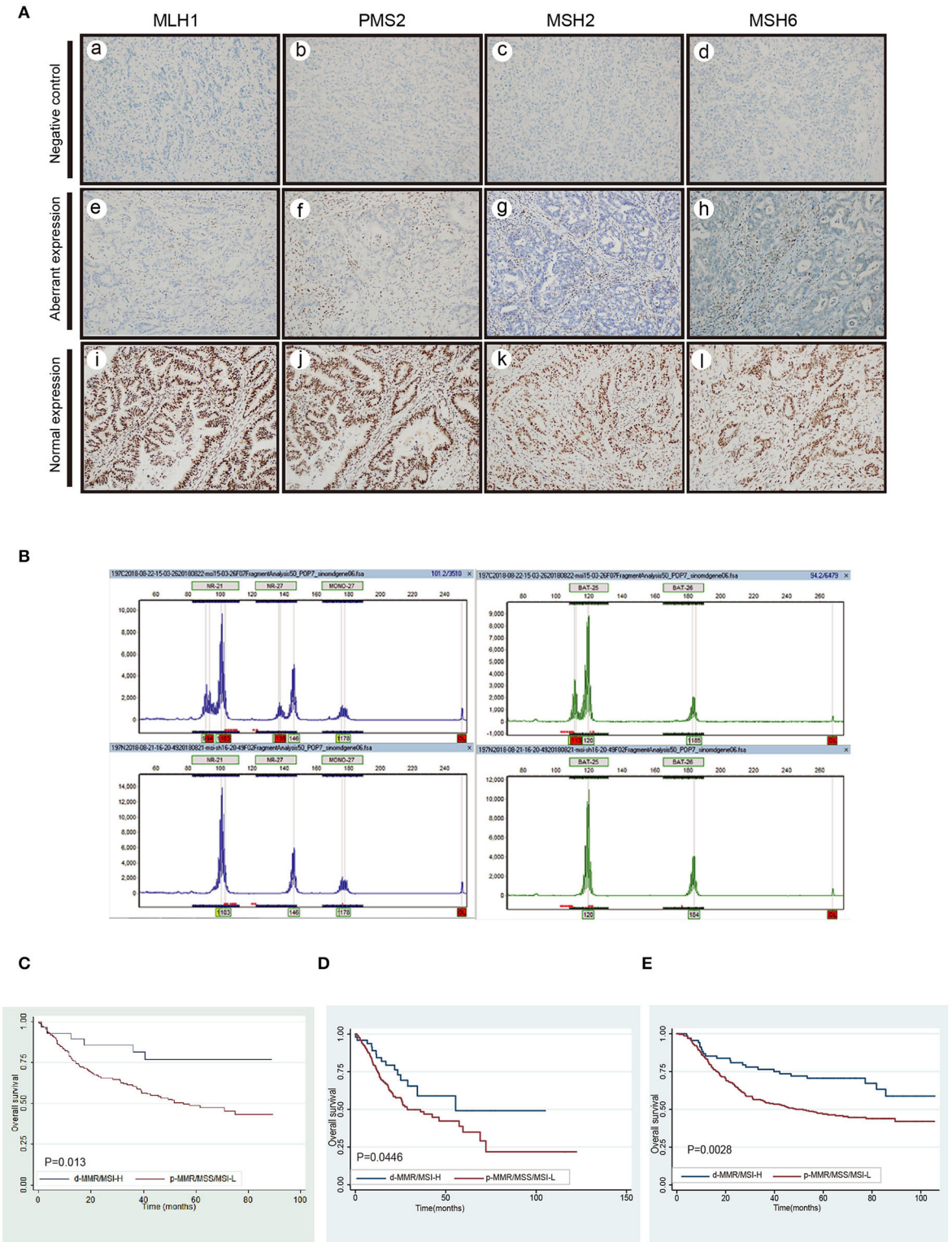


FIGURE 1 | The detection of MMR, MSI and the survival analysis in different cohorts. **(A)** The negative control staining of MLH1 (a), PMS2 (b), MSH2 (c), and MSH6 (d). Complete loss of nuclear expression of MLH1 (e), PMS2 (f), MSH2 (g), and MSH6 (h) in tumor cells but preserved nuclear staining in background non-tumor cells(aberrant expression). The expressions of MLH1 (i), PMS2 (g), MSH2 (k), and MSH6 (l) in tumor cells are intact (normal expression) **(B)** The MSI-PCR testing results of MSI-H. **(C)** Survival analysis of d-MMR/MSI-H on the prognosis of gastric cancer in our study. **(D,E)** Survival analysis of MSI subtype in TCGA-STAD and GSE62254/ACRG cohort.

TABLE 1 | The relationship between EBV and d-MMR/MSI-H subtypes and clinicopathological parameters in 271 gastric cancers.

		EBV			d-MMR/ MSI-H		
		(+)	(–)	*P*	(+)	(–)	*P*
	Median age	**61**	**65**		**70.5**	**64**	
Age (%)	>65	4 (44.4)	122 (46.6)	1.000	19 (67.9)	107 (44)	**0.017**
	≦65	5 (55.6)	140 (53.4)		9 (32.1)	136 (56)	
Gender (%)	Male	6 (66.7)	176 (46.4)	1.000	13 (46.4)	169 (69.5)	**0.014**
	Female	3 (33.3)	86 (53.6)		15 (53.6)	74 (30.5)	
Location (%)	GEJ-cardia	2 (22.2)	27 (10.3)	0.248	1 (3.6)	28 (11.5)	0.197
	Non-GEJ-cardia	7 (77.8)	235 (89.7)		27 (96.4)	215 (88.5)	
Location (%)	Antrum	4 (44.4)	150 (57.3)	0.674	20 (71.4)	134 (55.1)	0.099
	Non- antrum	5 (55.6)	112 (42.7)		8 (28.6)	109 (44.9)	
Size (%)	<5	3 (33.3)	110 (28.6)	0.862	8 (28.6)	105 (43.2)	0.137
	≧5	6 (66.7)	152 (71.4)		20 (71.4)	138 (56.8)	
Differentiation (%)	Well-moderate	1 (11.1)	45 (10.7)	0.980	3 (10.7)	43 (17.7)	0.505
	Poor	8 (88.9)	217 (89.3)		25 (89.3)	200 (82.3)	
Lauren (%)	Intestinal	3 (33.3)	119 (45.4)	0.707	13 (46.4)	109 (44.9)	0.874
	Nonintestinal	6 (66.7)	143 (54.6)		15 (53.6)	134 (55.1)	
T (%)	T1-T3	8 (88.9)	180 (68.7)	0.355	26 (92.9)	162 (47.2)	**0.000**
	T4	1 (11.1)	82 (31.3)		2 (7.1)	81 (52.8)	
N (%)	N0	4 (44.4)	53 (20.2)	0.181	14 (50)	43 (17.7)	**0.000**
	N+	5 (55.6)	209 (79.8)		14 (50)	200 (82.3)	
M (%)	M0	8 (88.9)	259 (98.9)	0.127	28 (100)	239 (98.4)	1.000
	M1	1 (11.1)	3 (1.1)		0 (0)	4 (1.6)	
TNM (%)	I–II	5 (55.6)	97 (37)	0.436	19 (67.9)	83 (34.2)	**0.000**
	III–IV	4 (44.4)	165 (63)		9 (32.1)	160 (65.8)	
WHO (%)	Medullary	3 (33.3)	3 (1.1)	**0.000**	2 (7.1)	4 (1.6)	0.119
	Non-medullary	6 (66.7)	259 (98.9)		26 (92.9)	239 (98.4)	
WHO (%)	Papillary-tubular	3 (33.3)	143 (54.6)	0.359	19 (67.9)	128 (52.7)	0.127
	Non-papillary-tubular	6 (66.7)	119 (45.4)		9 (32.1)	115 (47.3)	

The bold values indicate significant difference and P < 0.05.

for univariate survival analysis. Multivariate survival analysis was performed by Cox's proportional hazards regression model. The performance of the model was evaluated by applying the area under curve of receiver operating characteristic (auROC). Overall survival (OS) was defined as the interval between diagnosis and date of death or last-documented contact with patient. The cut-off value of *NUBP2* and *ENDOG* was determined by the X-tile software (24). A two-sided *P*-value < 0.05 was regarded as statistically significant and all statistical calculations were done using STATA 10.1(stata corp., College Station, TX, USA) or R software(version 3.5.3).

RESULTS

Prognosis and Potential Predictive Value of d-MMR/MSI-H Status in Different Cohorts

Of the 279 GC cases, the definite results of both IHC staining and MSI-testing were made in 275 cases. But four cases were detected to have the inconsistent results. Nuclear negative expression of MLH1, PMS2, MSH2, and MSH6 was seen in 27 (10.0%), 27 (10.0%), 1 (0.3%), and 1 (0.3%) in the rest of 271 cases, respectively. The normal and aberrant expression of MMR proteins were displayed in **Figure 1A**. MSI-PCR analysis revealed 28 cases of MSI-H, 2 MSI-L and 241 MSS. During this experiment, 27 cases showed instability at all six microsatellite loci and one case presented instability at five microsatellite loci except the Bat-6 (**Figure 1B**). Taken together, there were 28 cases with d-MMR/MSI-H and 243 cases with MMR-proficient/Low-Microsatellite Instability/Microsatellite stable (p-MMR/MSI-L/MSS) in 271cases. The detail was summarized in **Table 1**. Kaplan-Meier analysis and univariate analysis indicated that OS of GC patients with the d-MMR/MSI-H phenotype was better than that of patients with the p-MMR/MSI-L or MSS phenotype ($P = 0.013$) in the **Figure 1C** and **Table 2**. It was characterized by elderly age ($P = 0.017$), female ($P = 0.014$), without lymph node involvement ($P < 0.0001$), the lower depth of tumor invasion ($P < 0.0001$) and early TNM stage ($P < 0.0001$). We collected approximately 315 stomach adenocarcinoma from TCGA and described their clinical features as in the **Table 3**. Of the 315 cases, after six cases with undetermined subtype were removed, there were 50 MSI-H samples and the other 259 cases were considered as the MSI-low/MSS. The Kaplan-Meier survival analysis also

TABLE 2 | Univariate and multivariable analysis of overall survival in 271 gastric cancer.

	Univariate analysis				Multivariable analysis		
	P (log-rank test)	*P* (Cox's test)	HR	95% CI	*P* (Cox's test)	HR	95% CI
d-MMR/MSI-H status (Yes vs. No)	0.013	0.017	2.73	1.20–6.23	0.050	2.33	1.00–5.43
EBV (+ vs. –)	0.533	0.534	0.75	0.31–1.85			
Age	0.000	0.003	1.03	1.01–1.05	0.000	1.04	1.02–1.06
Sex (male vs. female)	0.092	0.093	1.39	0.95–2.95			
Location (antrum vs. nonantrum)	0.424	0.422	0.86	0.59–1.25			
Size (<5 vs. ≧5)	0.078	0.080	1.41	0.96–2.07			
Differentiation (well-moderate vs. poor)	0.226	0.228	1.39	0.82–2.35			
Lauren (intestinal vs. nonintestinal)	0.003	0.004	1.78	1.20–2.64	0.302	1.31	0.78–2.21
WHO (poorly cohesive components vs. remant)	0.000	0.000	2.04	1.40–2.97	0.130	1.48	0.89–2.46
pT stage (T1 + T2 vs. T3 + T4)	0.000	0.000	2.61	1.79–3.81	0.065	2.44	0.95–6.32
pN stage (N0 vs. N+)	0.002	0.003	2.35	1.34–4.12	0.360	0.65	0.26–1.63
M stage (M0 vs. M1)	0.467	0.476	0.49	0.07–3.51			
TNM (I + II vs. III + IV)	0.000	0.000	3.86	2.39–6.23	0.003	3.47	1.55–7.77

indicated that patients with MSI phenotype had better prognosis than MSS among all different races which included Asian, White, Black or African American($P = 0.045$) (**Figure 1D**). Then the GSE62254, derived from the ACRG research also illustrated the similar results. Among the 300 samples in dataset, 68 cases were MSI-H which also has a better correlation with prognosis ($P = 0.003$) (**Figure 1E**). The performance of the Cox's proportional hazards regression model was determined by applying the Receiver Operating Characteristic (ROC) curve analysis. The Area Under the Curve (AUC) value was 0.791. Therefore, this model has the predictive value for prognosis and was feasible.

Clinical and Prognostic Features of EBV in Different Cohorts

The incidence of EBV-positive GC in the 271 cases with consistent results was 3% (8/271). It has more frequent presence of EBV positive cases at GEJ/cardia-portion ($P = 0.043$) and medullary carcinoma ($P < 0.0001$) than EBV (–) cases (**Table 1**). Unlike d-MMR/MSI-H status, EBV infection itself by contrast was not prognostic factor in predicting OS of GC patients ($P = 0.533$, **Figure 2A**). In TCGA-STAD and GSE62254 dataset, 29 and 18 samples were detected to be EBV (+), respectively. The survival analysis also observed that it had no significant difference between EBV (–) and EBV(+) cases ($P = 0.795$, $P = 0.867$, **Figures 2B,C**). The EBER positive and negative case were displayed in **Figure 2D**.

Identification of the Key Module That Associated With OS Time and Status and Its Annotation in MSI Sample

In MSI subtype of GC, samples were clustered to detect the outliers while we did not delete any samples by average linkage method. The clinical trait data also could be input and the color representation of traits combined with the sample dendrogram (**Figure 3A**). The determination of soft-threshholding powers is the critical step to process this analysis. It was picked by the specific function in the WGCNA package and $\beta = 9$ was the most appropriate power to construct the adjacency ($R^2 = 0.870$; **Figure 3B**). The ME dissimilarity threshold was set at 0.3 and twelve modules were manifested for this group (**Figure 3C**). Then we got the primary module separation and the dissimilarity of module eigengenes (ME) was calculated to merge the similar modules to form the merged dynamic tree (**Figure 3D**). Through connecting the gene module to clinical traits, pink module was highly negatively correlated with the status and also pink module had the longest survival. It represented these genes were most associated with good prognosis in the heatmap (**Figure 3E**). Then all genes were shown in the heatmap accoring to the MSI and MSS subtype in the pink module (**Figure 4A**). Meanwhile, the heatmap was also drawn for all genes in OS time and status (**Supplementary Figures 2A,B**). GO enrichment indicated that genes cluster to mitochondrial protein formation and ncRNA process (**Figure 4B**). These processes occurred in mitochondria and ncRNA may confer to the mitochondrial circle DNA (**Figure 4C**). There were lots of unknown molecular functions and GO enrichment could not be manifested by the clusterprofiler R package (**Supplementary Figure 1B**). Thus, the activity should be further detected by the hub genes to probe the alterations in the mitochondria.

The Determination of the Hub Genes and Validation

The edge file, acquired from the WGCNA, put into the Cytoscape and genes were analyzed in the pink module (**Figure 4D**). According to the genes with high intra-modular connectivity ranked by the software, *NUBP2, CTU1, ENDOG, SSNA1,* and *BCL7C* could be considered as the hub genes. Also, these *CTU1, ENDOG, SSNA1,* and *BCL7C* had relative high GS and MM.

TABLE 3 | The relationship between MSI subtype and clinicopathological parameters in 309 gastric cancers in TCGA.

		d-MMR/MSI-H		
		(–)	(+)	*P*
Number		259	50	
Median age				
Age (%)	≤65	124 (47.9)	14 (28.0)	0.015
	>65	135 (52.1)	36 (72.0)	
Gender (%)	Male	177 (68.3)	26 (52.0)	0.039
	Female	82 (31.7)	24 (48.0)	
Location (%)	Non-cardia	188 (72.6)	45 (90.0)	0.015
	Cardia	71 (27.4)	5 (10.0)	
Location (%)	Non-antrum	173 (66.8)	20 (40.0)	0.001
	Antrum	86 (33.2)	30 (60.0)	
Race (%)	NA	32 (12.4)	13 (26.0)	0.068
	White	165 (63.7)	25 (50.0)	
	Asian	52 (20.1)	11 (22.0)	
	Black	10 (3.9)	1 (2.0)	
T (%)	T1–T3	191 (73.7)	31 (62.0)	0.129
	T4	68 (26.3)	19 (38.0)	
N (%)	N0	70 (27.0)	19 (38.0)	0.217
	N+	185 (71.4)	31 (62.0)	
	NX	4 (1.5)	0 (0.0)	
M (%)	M0	227 (87.6)	46 (92.0)	0.564
	M1	18 (6.9)	3 (6.0)	
	MX	14 (5.4)	1 (2.0)	
Stage (%)	I–II	108 (41.7)	27 (54.0)	0.147
	III–IV	151 (58.3)	23 (46.0)	
NUBP2 (%)	Low	184 (71.0)	25 (50.0)	0.006
	High	75 (29.0)	25 (50.0)	
ENDOG (%)	Low	151 (58.3)	12 (24.0)	<0.001
	High	108 (41.7)	38 (76.0)	

The GS of *NUBP2* was relative lower but its MM was such high that could not be neglected. As they were selected from the MSI samples in TCGA-STAD, hub genes were validated by the GSE62254/ACRG dataset. In GSE62254 dataset, *CTU1* could not be observed and only the other four hub genes were used to be further analyzed to assist us to uncover the specific activity that tightly associated with the good prognosis in MSI samples. *NUBP2* and *ENDOG* had significant difference between MSI and MSS subtype in TCGA-STAD. But the expression of *BCL7C* and *SSNA1* had no significant difference in these two subtypes (**Figure 4E**). There were 68 MSI samples in GSE62254. Others were considered the MSS phenotype. Then we found that the four hub genes highly expressed in MSI samples and had more significant difference than that was in MSS in GSE62254 (**Figure 4F**). Then all samples were divided into high and low expression group according to the appropriate cutoff value of these hub genes. The expression of *NUBP2* ($P = 0.006$) and *ENDOG* ($P < 0.001$) had significant difference between MSI and MSS subtypes (**Table 3**).

GSEA and GSVA for the Hub Genes

GSEA and GSVA were conducted to further shed light on the function of hub genes by comparing the differential expression group. According to the median expression of *NUBP2, CTU1, ENDOG, SSNA1,* and *BCL7C*, all cases were divided into the high and low expression group. Based on the nominal $P < 0.050$ and the normalized enrichment score (NES), top five KEGG pathways were illustrated in *ENDOG* and *NUBP2* highly expressed group (**Supplementary Figures 3A,B**). It more inclined to enrich in oxidative phosphorylation, glutathione metabolism and DNA repair. The common HALLMARK gene sets were reactive oxygen species pathway, oxidative phosphorylation, MYC targets and DNA repair that characterized by the mitochondrial impairment and oxidant stress (**Figures 5A,B**). The GSVA for *NUBP2* and *ENDOG* made similar conclusions as well (**Figures 5C,D** and **Supplementary Figures 3C,D**). *CTU1*, *BCL7C,* and *SSNA1* were carried out the same analysis and shown in **Supplementary Figures 4**, **5**. It could be concluded that both the cell impairment and anti-impairment associated pathways existed in this group. Nevertheless, the expression of *MYC* and *CASP3*, encoding the caspase3, was higher in MSI than it was in MSS samples in TCGA-STAD and GSE62254 (**Figures 5E–H**). Therefore, oxidative phosphorylation and reactive oxygen species pathways facilitate the apoptosis and had significant difference between MSI and MSS subtype. In addition, we performed the GSEA in GSE62254/ACRG in which the samples derived from Samsung Medical Center (Asian ethnicity) for status and the differential expression of hub genes (**Supplementary Figures 6A–C**). The results were similar with that performed in TCGA-STAD dataset and consistent with the above investigations as well (**Supplementary Figure 6D**).

DISCUSSION

The immune checkpoint therapy has become the most dazzling star in recent years since it has revealed the therapeutic efficacy in melanoma (25). The ever-increasingly comprehensive research also enhanced it to be the hotpoint in different cancer types. Nevertheless, not all types of cancer could actually benefit from the immunotherapy while not all patients had response for the determined effective cancer (26). Under this circumstance, the identification and well understanding of the subtypes greatly assisted to improve the efficacy of the anti-PD1 therapy. The recent proposal of TCGA molecular classification broadened our view of GC molecular characteristics by highlighting four main subtypes that summarized the western population. Then the Asian Cancer Research Group (ACRG) also came up with a molecular classification by analyzing the expression profiling of Asian population. Both of these two classifications involved the MSI and EBV phenotype. Investigations had confirmed that immune checkpoint blockade gained the better efficacy in MSI and EBV associated GC (27–30). The discovery of correlations between this classification and differential therapeutic response reaffirmed that clinical-relevance of each subtype was caused by distinct molecular mechanism of GC. Nonetheless, the clinical course of immunotherapy related EBV (+) or MSI-H GC is

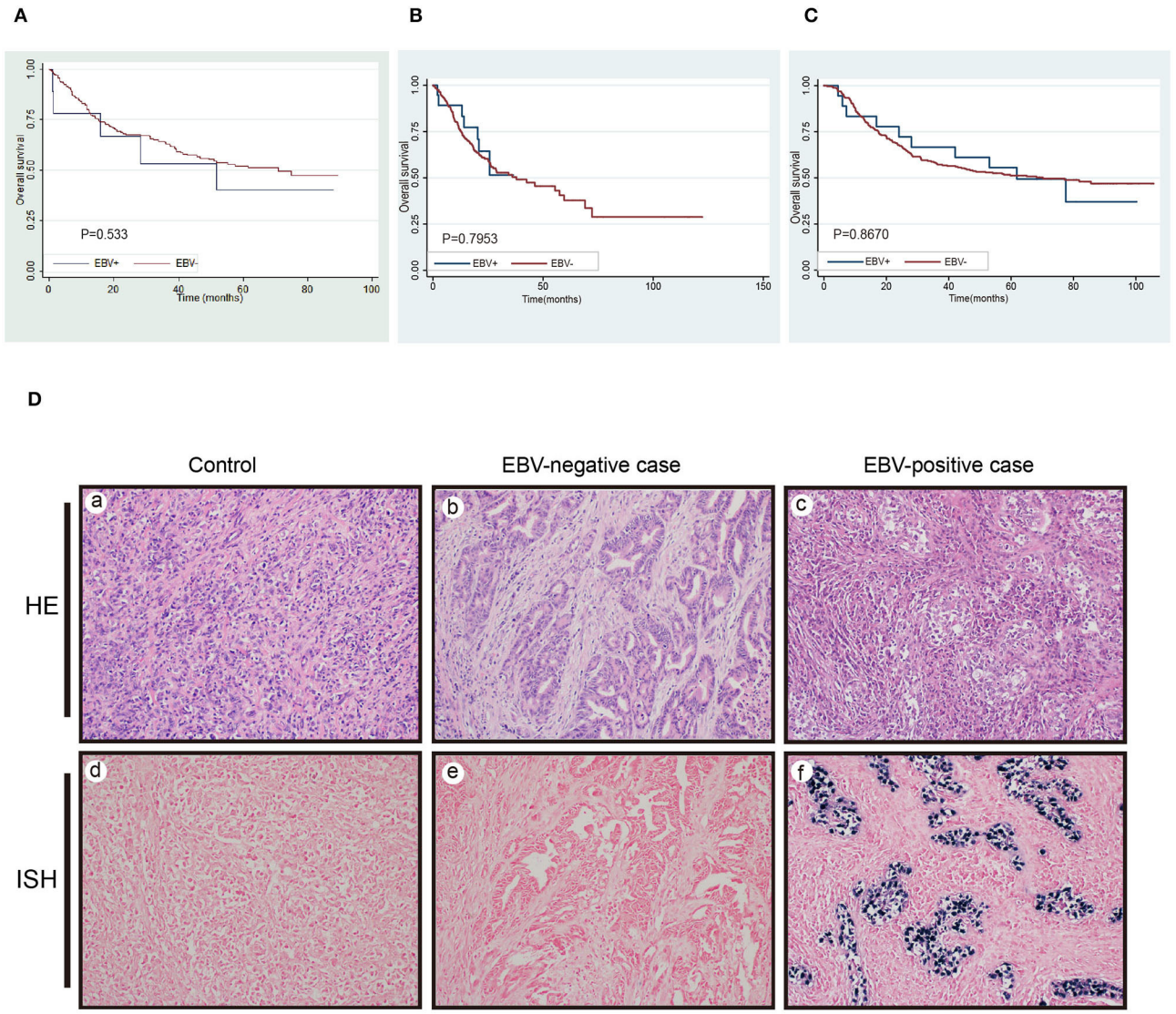


FIGURE 2 | The detection of EBV infection and its survival analysis. **(A)** Survival analysis of EBV(+)gastric cancer in our study. **(B,C)** The survival analysis in TCGA-STAD and GSE62254/ACRG cohort for the same parameters. **(D)** The negative probe control of ISH in EBV+ infection cases (a) HE, and (d) ISH. One case with EBV-negativity (b) HE, and (e) ISH. One case with EBV-positivity (c) HE, and (f) ISH.

not fully understood. As far as we concerned, current data that concentrate on the associations of EBV infection and MSI phenotype with clinical parameters and outcome of GC was relatively scarce in East-Asia. To address it, our retrospectively analysis of the EBV infection and MSI status were chosen based on reliable detection methods in a cohort of Chinese GC patients($n = 279$).

Currently, PCR combined with capillary electrophoresis and IHC was routinely performed to detect the MSI (31, 32). In this study, an integrated testing panel containing the mononucleotides of BAT-25, BAT-26, NR21, NR24 and MONO27 were carried out in these 279 cases. Meanwhile, IHC was also adopted to test the four MMR associated proteins MLH1, MSH2, MSH6, and PMS2. Studies often focused on the consistency rate of these two methods varied a lot (91.2–97.8%) (33, 34). Previously, investigations indicated the proportion of d-MMR/MSI-H was ∼8.2–44.5% in different cohort while the incidence was 10.3% (28/271) in our study. It reconfirmed that D-MMR/MSI-H GC was related to older age, female, lower depth of tumor invasion, without frequency of lymph node metastasis and lower TNM stage, but was not consistent with tumor size, distal location, medullary carcinoma and intestinal subtype that previously reported. Even so, there was still a trend toward higher rates of antrum-located location, large size ($\geqq$ 5), medullary carcinoma and papillary-tubular type seen in our d-MMR/MSI-H GC. Scientists reported that MSI-H was significantly related

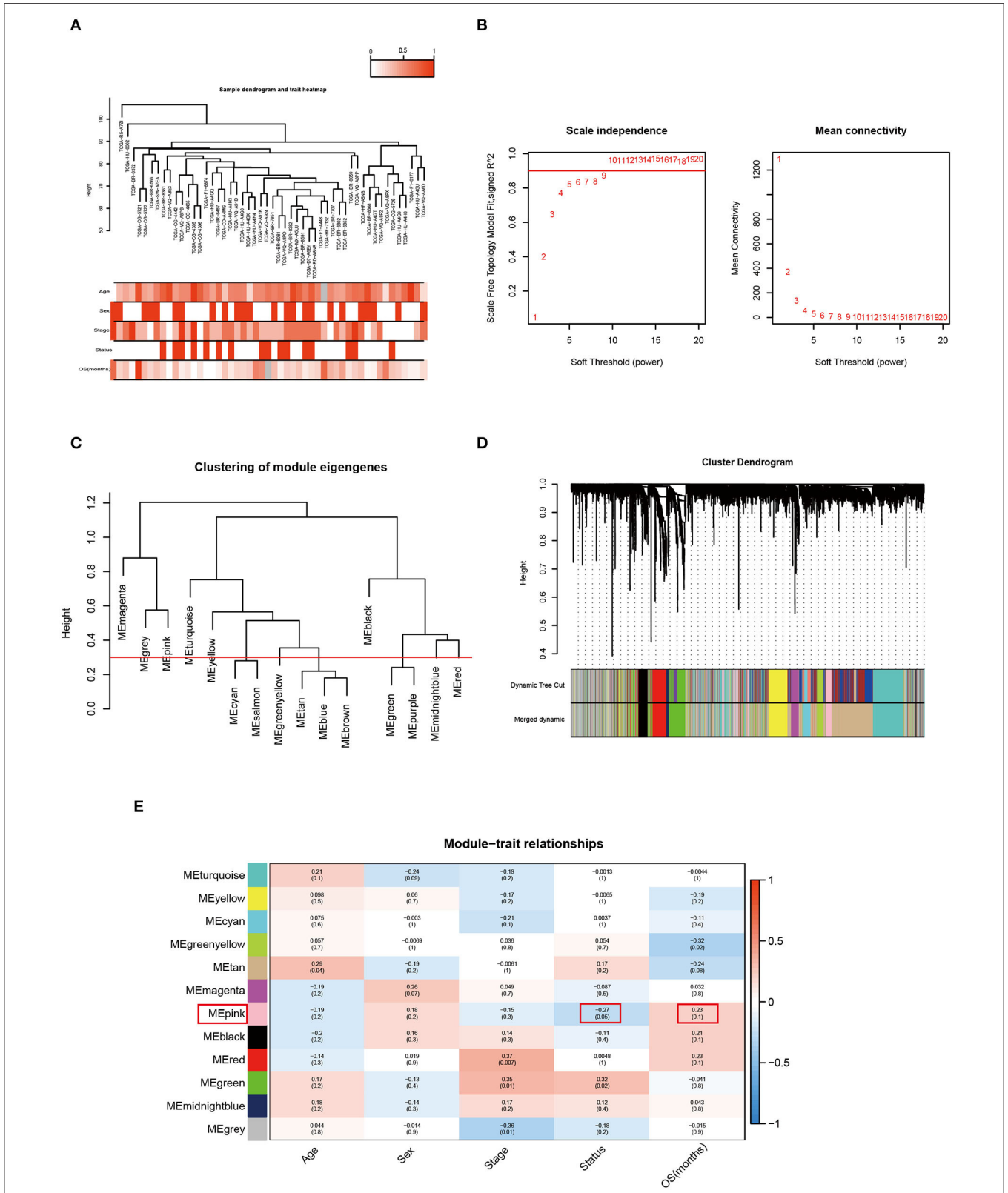


FIGURE 3 | Weighted Correlation Network analysis was performed to construct the correlation between gene module and the clinical traits to find the key module that tightly associated with the prognosis. **(A)** Clustering the dendrogram of 51 MSI samples and the combination with its clinical traits. **(B)** Screening out the soft-thresholding power through scale independence and mean connectivity. **(C)** Clustering of the modules and set the criteria to merge the similar modules. **(D)** The dynamic cut tree after merging the similar modules. **(E)** The heatmap for module-trait relationships in DGC samples. The pink module was the key module with good status and long survival.

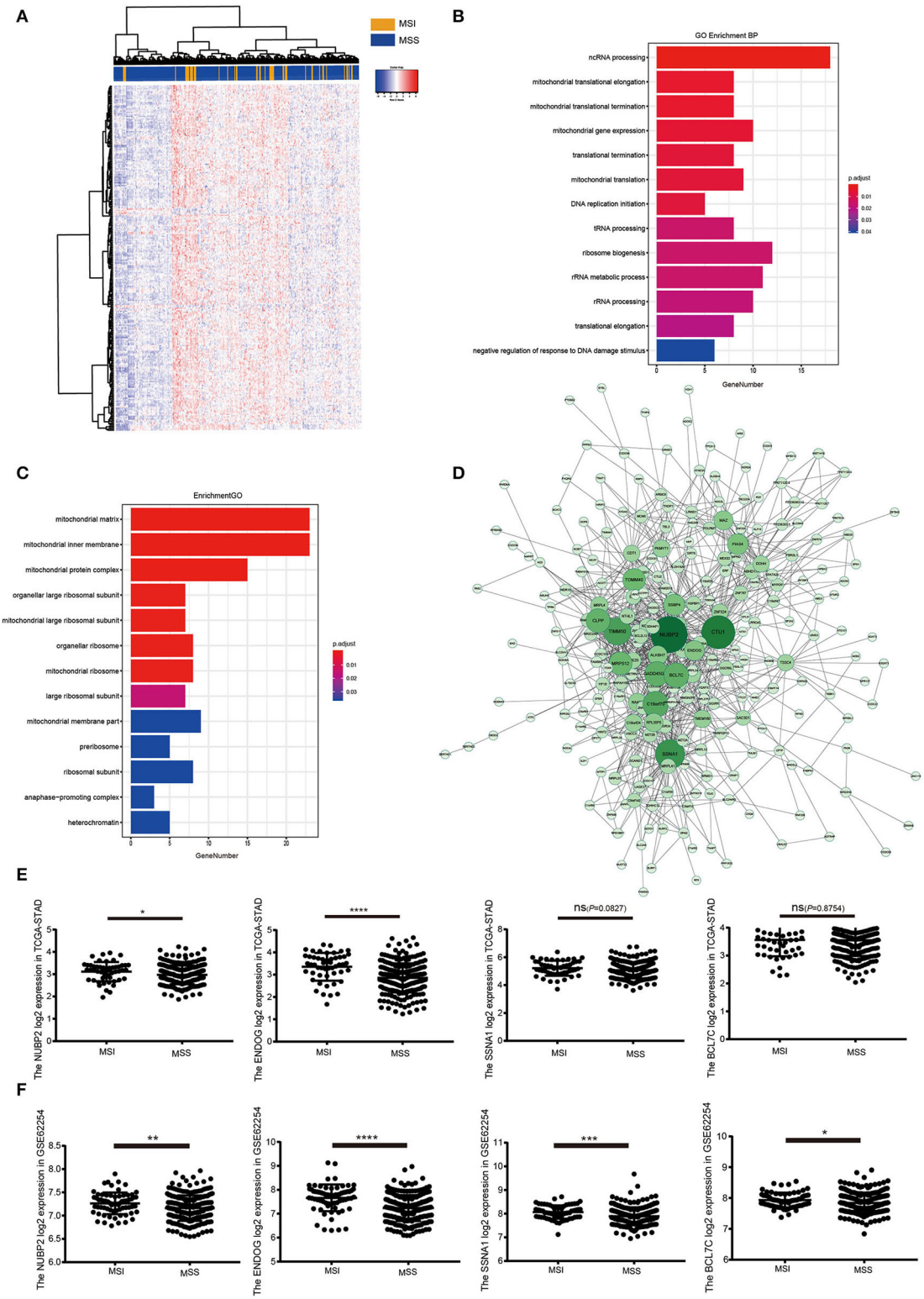


FIGURE 4 | GO enrichment for the pink module and identification of the hub genes. **(A)** Heatmap for the expression pattern of all genes in pink module at MSI and MSS phenotype. **(B,C)** The biological process and cellular component for the pink module in GO annotations. **(D)** The coexpression network for the genes in pink module and identification of the hub genes. **(E)** The expression of *NUBP2*, *ENDOG*, *SSNA1* and *BCL7C* that had been log2 normalized in TCGA-STAD dataset in MSI ($N = 51$) and MSS group ($N = 264$). **(F)** The validation of *NUBP2*, *ENDOG*, *SSNA1* and *BCL7C* expression that had been log2 normalized in GSE62254/ACRG in MSI ($N = 68$) and MSS group ($N = 232$). $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

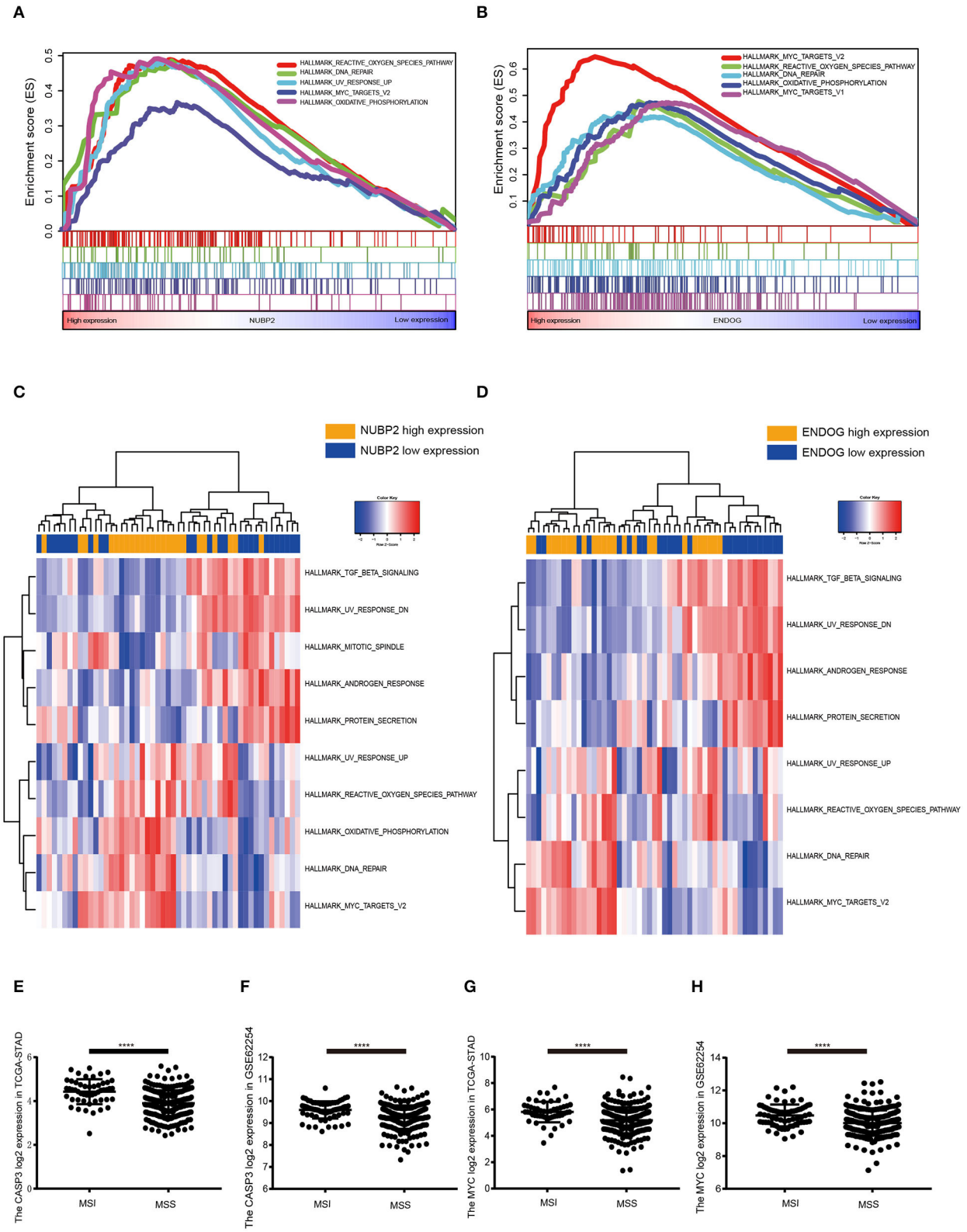


FIGURE 5 | Gene set enrichment analysis (GSEA) and gene set variation analysis(GSVA) for the *NUBP2* and *ENDOG* in TCGA-STAD. **(A,B)** The top five gene sets(according to the enrichment score) enriched in the high expression of single hub gene for HALLMARK gene sets. **(A)** *NUBP2*; **(B)** *ENDOG*. **(C,D)** The heatmaps of differentially expressed pathways for single hub gene through the calculation of GSVA. **(E,F)** The log2 normalized expression of *MYC* and *CASP3* in MSI and MSS samples in TCGA-STAD dataset. **(G,H)** The log2 normalized expression of *MYC* and *CASP3* in GSE62254 dataset. ****$P < 0.0001$.

with higher survival at 15 years of follow-up and an independent prognostic factor that reminded us its predictive role relied on further prolonged follow-up. Besides that, we collected the expression profiling of GC tissues and corresponding clinical traits from TCGA ($n = 315$) and GEO (GSE62254/ACRG). The cohort consisted of Asian, Black or African American, Native Hawaiian or other Pacific island and white in the TCGA-STAD dataset. In 2014, TCGA did not investigate the significant OS differences between MSI and other subtypes. But it had a better prognosis in this TCGA dataset which was downloaded by us. This was reconfirmed in Asian population involved in the GSE62254/ACRG by Kaplan-Meier analysis.

At present, EBV infection could be tested by several methods, such as polymerase chain reaction (PCR), electron microscopy, southern blot hybridization, IHC and ISH that was considered as the gold-standard test. Due to limited sample size (3%), EBVaGC was just associated with proximal location and medullary carcinoma, but not with reported characteristics of male predominance in this study. Additionally, consistent with some previous research, it was also found that this subtype could not reflect a long-term survival (12). It made the identical conclusion from the TCGA-STAD and GSE62254. During this process, we have also observed that these two situations (EBV positive and d-MMR/MSI-H status) are virtually mutually exclusive in line with previous reports though both of these subtypes showed a good response to immunotherapy (2, 3, 10, 12, 21). Obviously, the distinct PD-L1 associated expression profile owed by them need to be further studied.

The different prognostic influence between EBV and MSI subtypes was worthful to be explored as their common efficacy to immune checkpoint blockade. Nevertheless, the MSS subtype could respond to the immunotherapy with EBV infection and there were little cases with both MSI and EBV phenotype (30). It means that at least two distinct molecular mechanisms exist in these two subsets and the immunotherapy can be controlled based on this. Some recent investigations have reported the clinical characteristics for the status of EBV infection in GC. There was approximately average 10% EBV positive cases in GC samples worldwide (35, 36). For Latvia GC population, EBV positivity was a favorable prognostic factor in GC while it had no significant difference in other relative large cohort (37). Not only the intrinsic alterations of MSI and EBV infected cells but also the tumor microenvironment (TME) contributes to these clinical characteristics to some extent. On the one hand, the metabolic differences manifested that lipid metabolism was evident in MSI tissues as the fatty acid synthase (FASN) increased in colorectal cancer is associated with MSI (38, 39). Sirt1, the critical histone deacetylase that cross the mitochondrial metabolism and DNA damage repair, correlated with MSI (40). Meanwhile, some fatty acid biosynthesis related enzymes FASN and PLA2G4A decreased in EBVaGC and could lead to the worse survival. Similarly, EBV infection could make the metabolic reprogramming and it is the foundation of the poor clinical prognosis in GC patients (41). On the other hand, the discrepancy of tumor microenvironment was another controversial topic and may interpret the molecular mechanism. Recently, bioinformatics analysis come up with the TMEscore and Immunoscore which also could be consider as a prognostic and predictive tool for GC by a large scale microarray data (42–44). During the process of accessing the tumor purity, higher TMEscore was associated with a good prognosis and characterized by the response to virus and IFNγ that was consistent with the features of MSI in latest researches (42). Furthermore, the activation of immune response commonly observed in MSI and EBV subtypes and immunomicroenvironment appears complicated and play a role in metabolic reprogramming as well. Of note, T cell metabolism could not be easily ignored as it involved in the IFN-γ and fatty acid synthesis in the TME (45).

As the good prognosis was such evident for MSI, it was appealing to explore the critical factors that associated with the longer survival. The WGCNA provide the pink module that could be considered as the key one with the better status and longest overall survival in MSI samples. Genes predominantly enriched in mitochondria or ribosome and played a role in the process of ncRNA, mitochondrial translation elongation or termination and mitochondrial gene expression. As amount of unknown molecular functions in these modules, the specific function of this module was probed by the hub genes to detect the concrete activity to ensure the OS for MSI cases. Apparently, not only the mitochondria associated proteins but also the genes involved the MSI conditions. Mitochondrial activities need to be in-depth studied and may uncover the origins of MSI.

The fetched five hub genes by WGCNA, *NUBP2, CTU1, ENDOG, SSNA1,* and *BCL7C* were illustrated by Cytoscape. They were revalidated by the GSE62254/ACRG and had more significant difference in MSI cases than MSS. Then *NUBP2* and *ENDOG* was the top two genes close to the centrality could reflect the activity of this module in mitochondria to largely extent. Oxidative phosphorylation, reactive oxygen species pathway, MYC targets, glutathione metabolism and DNA repair were the obvious pathways that tightly associated with the high expression of hub genes and the alteration of mitochondrial translation. On the basis of these parts, we could conclude that the better prognosis associated with the function of mitochondrial proteins that mainly played a great part in oxidant phosphorylation, ROS pathway and MYC targets as the apoptosis was increasing. Actually, the glutathione metabolism, base excision repair and DNA repair reflected the activity of antioxidant response and anti-impairment. But weigh the two factors, high expression of apoptosis associated gene determined that the injury factors play a dominant role in MSI subtype.

Until now, rare investigations reported the relationships between the MSI phenotype and mitochondrial activity. Mitochondrial microsatellites instability (mtMSI) was easier to be ignored than the nuclear MSI. As the alterations emerged in mitochondrial matrix according to the GO annotation, the ncRNA process and DNA replication initiation had a great probability to represent the mitochondrial DNA variations. Considerable investigations had revealed its links with the prognosis of colorectal cancer while rare researches involved in GC (46). Though the function of mitochondrial DNA is less powerful than the nuclear DNA, it is convenient to regulate the oxidative phosphorylation system (OXPHOS) (47).

NUBP2, the nucleotide binding protein 2, encodes adenosine triphosphate (ATP) and metal-binding protein that modulate the iron-metabolism that was essential for ATP production and mitochondrial metabolism (48). It was the essential component that could assemble the iron-sulfur clusters through the process of cytosolic iron-sulfur cluster assembly (CIA) outside of the mitochondria (49, 50). Compared with the hypoxic microenvironment, the upregulation of *NUBP2* indicated the normoxia and ensure the oxidative phosphorylation in the MSI samples in GC. The normal activity of oxidative phosphorylation decreased the tumor cell atypia and its malignancy that lead to the better prognosis. *ENDOG*, the Endonuclease G, was the nuclear encoded gene and its corresponding protein mainly localized in mitochondria. This protein is capable of initiating the mitochondrial DNA replication by generating the RNA primers (51, 52). On the one hand, it was the downstream effector of caspase-3 and facilitated the Myc-induced genetic instability and apoptosis (53, 54). On the other hand, *ENDOG* regulate the mRNA alternative splicing of hTERT (52, 55). As the non-active splice variant hTERT increased, the activity of telomerase is suppressed and lead to the short telomere which acquired the replicated senescence for tumor cells (56). On the basis of these reasons, tumor cells have more opportunities and prone to be induced apoptosis and cell senescence in MSI subtype.

CONCLUSIONS

Taken together, we classified the clinical characteristics of MSI and EBV in Chinese GC cohort to some extent with the limited cases. Combining with the public datasets, we summarized that MSI could serve as a prognostic factor for good survival while it had no significant difference in EBV associated cases. The prognostic value tightly associated with the oxidative phosphorylation system, reactive oxygen species and MYC targets pathways through the modulation of mitochondria. The glutathione metabolism and DNA repair were also active but the antioxidant response could not resist the accumulation of ROS and genetic instability that contribute to more opportunities for cell apoptosis in MSI samples. Based on these discoveries, some attractive strategies of up regulating the *ENDOG* or *NUBP2* could be utilized to increase the oxidative phoshorylation for MSS subtype which could imitate the easily apoptotic effects. Certainly, more experimental and clinical trials should apply to optimize and achieve the potential to acquire the similar prognostic effects like MSI in other subtypes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study, these can be found in The Cancer Genome Atlas via the UCSC Xena browser (http://xena.ucsc.edu/); the NCBI Gene Expression Omnibus (GSE62254).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Xin Hua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine. 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.

AUTHOR CONTRIBUTIONS

LC and YS performed the research study, analyzed the data, and wrote the paper. KW, WG, JY, and JL supported during performing of the experiments, collected patient data, and contributed essential reagents and laboratory equipment. RW and LW conceived and designed the study. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by funds from the Shanghai Municipal Commission of Health and Family Planning (No. 201740031), Joined Medicine-Engineering Fund of Shanghai Jiao Tong University (No. YG2016QN72) and Xinhua Hospital for the Introduction of Talent, School of Medicine, Shanghai Jiao Tong University (No. 005).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.01269/full#supplementary-material

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer.* (2015) 136:E359–86. doi: 10.1002/ijc.29210
2. Cancer Genome Atlas Research N. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* (2014) 513:202–9. doi: 10.1038/nature13480
3. Cristescu R, Lee J, Nebozhyn M, Kim KM, Ting JC, Wong SS, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat Med.* (2015) 21:449–56. doi: 10.1038/nm.3850
4. Li X, Wu WK, Xing R, Wong SH, Liu Y, Fang X, et al. Distinct subtypes of gastric cancer defined by molecular characterization include novel mutational signatures with prognostic capability. *Cancer Res.* (2016) 76:1724–32. doi: 10.1158/0008-5472.CAN-15-2443
5. Brahmer JR, Drake CG, Wollner I, Powderly JD, Picus J, Sharfman WH, et al. Phase I study of single-agent anti-programmed death-1 (MDX-1106) in refractory solid tumors: safety, clinical activity, pharmacodynamics, and immunologic correlates. *J Clin Oncol.* (2010) 28:3167–75. doi: 10.1200/JCO.2009.26.7609
6. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science.* (2017) 357:409–13. doi: 10.1126/science.aan6733

7. Le DT, Uram JN, Wang H, Bartlett BR, Kemberling H, Eyring AD, et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N Engl J Med.* (2015) 372:2509−20. doi: 10.1056/NEJMoa1500596
8. Kim ST, Cristescu R, Bass AJ, Kim KM, Odegaard JI, Kim K, et al. Comprehensive molecular characterization of clinical responses to PD-1 inhibition in metastatic gastric cancer. *Nat Med.* (2018) 24:1449–58. doi: 10.1038/s41591-018-0101-z
9. An JY, Kim H, Cheong JH, Hyung WJ, Kim H, Noh SH. Microsatellite instability in sporadic gastric cancer: its prognostic role and guidance for 5-FU based chemotherapy after R0 resection. *Int J Cancer.* (2012) 131:505–11. doi: 10.1002/ijc.26399
10. Ahn S, Lee SJ, Kim Y, Kim A, Shin N, Choi KU, et al. High-throughput protein and mRNA expression-based classification of gastric cancers can identify clinically distinct subtypes, concordant with recent molecular classifications. *Am J Surg Pathol.* (2017) 41:106–15. doi: 10.1097/PAS.0000000000000756
11. Falchetti M, Saieva C, Lupi R, Masala G, Rizzolo P, Zanna I, et al. Gastric cancer with high-level microsatellite instability: target gene mutations, clinicopathologic features, and long-term survival. *Hum Pathol.* (2008) 39:925–32. doi: 10.1016/j.humpath.2007.10.024
12. Shen H, Zhong M, Wang W, Liao P, Yin X, Rotroff D, et al. EBV infection and MSI status significantly influence the clinical outcomes of gastric cancer patients. *Clin Chim Acta.* (2017) 471:216–21. doi: 10.1016/j.cca.2017.06.006
13. Wirtz HC, Muller W, Noguchi T, Scheven M, Ruschoff J, Hommel G, et al. Prognostic value and clinicopathological profile of microsatellite instability in gastric cancer. *Clin Cancer Res.* (1998) 4:1749–54.
14. Polom K, Marano L, Marrelli D, De Luca R, Roviello G, Savelli V, et al. Meta-analysis of microsatellite instability in relation to clinicopathological characteristics and overall survival in gastric cancer. *Br J Surg.* (2018) 105:159–67. doi: 10.1002/bjs.10663
15. Seo HM, Chang YS, Joo SH, Kim YW, Park Y-K, Hong SW, et al. Clinicopathologic characteristics and outcomes of gastric cancers with the MSI-H phenotype. *J Surg Oncol.* (2009) 99:143–7. doi: 10.1002/jso.21220
16. Zhu L, Li Z, Wang Y, Zhang C, Liu Y, Qu X. Microsatellite instability and survival in gastric cancer: a systematic review and meta-analysis. *Mol Clin Oncol.* (2015) 3:699–705. doi: 10.3892/mco.2015.506
17. Janic A, Valente LJ, Wakefield MJ, Di Stefano L, Milla L, Wilcox S, et al. DNA repair processes are critical mediators of p53-dependent tumor suppression. *Nat Med.* (2018) 24:947–53. doi: 10.1038/s41591-018-0043-5
18. Lario LD, Ramirez-Parra E, Gutierrez C, Casati P, Spampinato CP. Regulation of plant MSH2 and MSH6 genes in the UV-B-induced DNA damage response. *J Exp Bot.* (2011) 62:2925–37. doi: 10.1093/jxb/err001
19. Aasland D, Gotzinger L, Hauck L, Berte N, Meyer J, Effenberger M, et al. Temozolomide induces senescence and repression of DNA repair pathways in glioblastoma cells via activation of ATR-CHK1, p21, and NF-kappaB. *Cancer Res.* (2019) 79:99–113. doi: 10.1158/0008-5472.CAN-18-1733
20. Genitsch V, Novotny A, Seiler CA, Kroll D, Walch A, Langer R. Epstein-Barr virus in gastro-esophageal adenocarcinomas - single center experiences in the context of current literature. *Front Oncol.* (2015) 5:73. doi: 10.3389/fonc.2015.00073
21. Setia N, Agoston AT, Han HS, Mullen JT, Duda DG, Clark JW, et al. A protein and mRNA expression-based classification of gastric cancer. *Mod Pathol.* (2016) 29:772–84. doi: 10.1038/modpathol.2016.55
22. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* (2008) 9:559. doi: 10.1186/1471-2105-9-559
23. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS.* (2012) 16:284–7. doi: 10.1089/omi.2011.0118
24. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res.* (2004) 10:7252–9. doi: 10.1158/1078-0432.CCR-04-0713
25. Drake CG, Lipson EJ, Brahmer JR. Breathing new life into immunotherapy: review of melanoma, lung and kidney cancer. *Nat Rev Clin Oncol.* (2014) 11:24–37. doi: 10.1038/nrclinonc.2013.208
26. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *New Engl J Med.* (2012) 366:2443–54. doi: 10.1056/NEJMoa1200690
27. Goodman AM, Sokol ES, Frampton GM, Lippman SM, Kurzrock R. Microsatellite-stable tumors with high mutational burden benefit from immunotherapy. *Cancer Immunol Res.* (2019) 7:1570–3. doi: 10.1158/2326-6066.CIR-19-0149
28. Duffy MJ, Crown J. Biomarkers for predicting response to immunotherapy with immune checkpoint inhibitors in cancer patients. *Clin Chem.* (2019) 65:1228–38. doi: 10.1373/clinchem.2019.303644
29. Jin Z, Yoon HH. The promise of PD-1 inhibitors in gastro-esophageal cancers: microsatellite instability vs. PD-L1. *J Gastrointest Oncol.* (2016) 7:771–88. doi: 10.21037/jgo.2016.08.06
30. Panda A, Mehnert JM, Hirshfield KM, Riedlinger G, Damare S, Saunders T, et al. Immune activation and benefit from avelumab in EBV-positive gastric cancer. *J Natl Cancer Inst.* (2018) 110:316–20. doi: 10.1093/jnci/djx213
31. Luchini C, Bibeau F, Ligtenberg MJL, Singh N, Nottegar A, Bosse T, et al. ESMO recommendations on microsatellite instability testing for immunotherapy in cancer, and its relationship with PD-1/PD-L1 expression and tumour mutational burden: a systematic review-based approach. *Ann Oncol.* (2019) 30:1232–43. doi: 10.1093/annonc/mdz116
32. Hempelmann JA, Lockwood CM, Konnick EQ, Schweizer MT, Antonarakis ES, Lotan TL, et al. Microsatellite instability in prostate cancer by PCR or next-generation sequencing. *J Immunother Cancer.* (2018) 6:29. doi: 10.1186/s40425-018-0341-y
33. McConechy MK, Talhouk A, Li-Chang HH, Leung S, Huntsman DG, Gilks CB, et al. Detection of DNA mismatch repair (MMR) deficiencies by immunohistochemistry can effectively diagnose the microsatellite instability (MSI) phenotype in endometrial carcinomas. *Gynecol Oncol.* (2015) 137:306–10. doi: 10.1016/j.ygyno.2015.01.541
34. Zheng J, Huang B, Nie X, Zhu Y, Han N, Li Y. The clinicopathological features and prognosis of tumor MSI in East Asian colorectal cancer patients using NCI panel. *Future Oncol.* (2018) 14:1355–64. Epub 2018/01/26. doi: 10.2217/fon-2017-0662
35. Naseem M, Barzi A, Brezden-Masley C, Puccini A, Berger MD, Tokunaga R, et al. Outlooks on Epstein-Barr virus associated gastric cancer. *Cancer Treat Rev.* (2018) 66:15–22. doi: 10.1016/j.ctrv.2018.03.006
36. Akiba S, Koriyama C, Herrera-Goepfert R, Eizuru Y. Epstein-Barr virus associated gastric carcinoma: epidemiological and clinicopathological features. *Cancer Sci.* (2008) 99:195–201. doi: 10.1111/j.1349-7006.2007.00674.x
37. Gasenko E, Isajevs S, Camargo MC, Offerhaus GJA, Polaka I, Gulley ML, et al. Clinicopathological characteristics of Epstein-Barr virus-positive gastric cancer in Latvia. *Eur J Gastroenterol Hepatol.* (2019) 31:1328–33. doi: 10.1097/MEG.0000000000001521
38. Ogino S, Kawasaki T, Ogawa A, Kirkner GJ, Loda M, Fuchs CS. Fatty acid synthase overexpression in colorectal cancer is associated with microsatellite instability, independent of CpG island methylator phenotype. *Human Pathol.* (2007) 38:842–9. doi: 10.1016/j.humpath.2006.11.018
39. Wood SM, Gill AJ, Brodsky AS, Lu S, Friedman K, Karashchuk G, et al. Fatty acid-binding protein 1 is preferentially lost in microsatellite instable colorectal carcinomas and is immune modulated via the interferon gamma pathway. *Mod Pathol.* (2017) 30:123–33. doi: 10.1038/modpathol.2016.170
40. Nosho K, Shima K, Irahara N, Kure S, Firestein R, Baba Y, et al. SIRT1 histone deacetylase expression is associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. *Mod Pathol.* (2009) 22:922–32. doi: 10.1038/modpathol.2009.49
41. Yoon SJ, Kim JY, Long NP, Min JE, Kim HM, Yoon JH, et al. Comprehensive multi-omics analysis reveals aberrant metabolism of Epstein-Barr-virus-associated gastric carcinoma. *Cells.* (2019) 8:1220. doi: 10.3390/cells8101220
42. Zeng D, Li M, Zhou R, Zhang J, Sun H, Shi M, et al. Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. *Cancer Immunol Res.* (2019) 7:737–50. doi: 10.1158/2326-6066.CIR-18-0436
43. Jiang Y, Zhang Q, Hu Y, Li T, Yu J, Zhao L, et al. ImmunoScore signature: a prognostic and predictive tool in gastric cancer. *Ann Surg.* (2018) 267:504–13. doi: 10.1097/SLA.0000000000002116
44. Zeng D, Zhou R, Yu Y, Luo Y, Zhang J, Sun H, et al. Gene expression profiles for a prognostic immunoscore in gastric cancer. *Br J Surg.* (2018) 105:1338–48. doi: 10.1002/bjs.10871

45. Bantug GR, Galluzzi L, Kroemer G, Hess C. The spectrum of T cell metabolism in health and disease. *Nat Rev Immunol.* (2018) 18:19–34. doi: 10.1038/nri.2017.99
46. Ling X-L, Fang D-C, Wang R-Q, Yang S-M, Fang L. Mitochondrial microsatellite instability in gastric cancer and its precancerous lesions. *World J Gastroenterol.* (2004) 10:800. doi: 10.3748/wjg.v10.i6.800
47. Shuwen H, Xi Y, Yuefen P. Can mitochondria DNA provide a novel biomarker for evaluating the risk and prognosis of colorectal cancer? *Dis Mark.* (2017) 2017:5189803. doi: 10.1155/2017/5189803
48. Kypri E, Christodoulou A, Maimaris G, Lethan M, Markaki M, Lysandrou C, et al. The nucleotide-binding proteins Nubp1 and Nubp2 are negative regulators of ciliogenesis. *Cell Mol Life Sci.* (2014) 71:517–38. doi: 10.1007/s00018-013-1401-6
49. Grossman JD, Gay KA, Camire EJ, Walden WE, Perlstein DL. Coupling nucleotide binding and hydrolysis to iron-sulfur cluster acquisition and transfer revealed through genetic dissection of the Nbp35 ATPase site. *Biochemistry.* (2019) 58:2017–27. doi: 10.1021/acs.biochem.8b00737
50. Anwar S, Dikhit MR, Singh KP, Kar RK, Zaidi A, Sahoo GC, et al. Interaction between Nbp35 and Cfd1 proteins of cytosolic Fe-S cluster assembly reveals a stable complex formation in Entamoeba histolytica. *PloS ONE.* (2014) 9:e108971. doi: 10.1371/journal.pone.0108971
51. Wiehe RS, Gole B, Chatre L, Walther P, Calzia E, Ricchetti M, et al. Endonuclease G promotes mitochondrial genome cleavage and replication. *Oncotarget.* (2018) 9:18309. doi: 10.18632/oncotarget.24822
52. Vasina DA, Zhdanov DD, Orlova EV, Orlova VS, Pokrovskaya MV, Aleksandrova SS, et al. Apoptotic endonuclease endog inhibits telomerase activity and induces malignant transformation of human CD4+ T cells. *Biochem Biokhim.* (2017) 82:24–37. doi: 10.1134/S0006297917010035
53. Liu X, He Y, Li F, Huang Q, Kato TA, Hall RP, et al. Caspase-3 promotes genetic instability and carcinogenesis. *Mol Cell.* (2015) 58:284–96. doi: 10.1016/j.molcel.2015.03.003
54. Cartwright IM, Liu X, Zhou M, Li F, Li C-Y. Essential roles of Caspase-3 in facilitating Myc-induced genetic instability and carcinogenesis. *Elife.* (2017) 6:e26371. doi: 10.7554/eLife.26371
55. Zhdanov DD, Pokrovsky VS, Orlova EV, Orlova VS, Pokrovskaya MV, Aleksandrova SS, et al. Intracellular localization of apoptotic endonuclease endog and splice-variants of telomerase catalytic subunit hTERT. *Biochem Biokhim.* (2017) 82:894–905. doi: 10.1134/S0006297917080041
56. Zhdanov DD, Vasina DA, Orlova EV, Orlova VS, Pokrovsky VS, Pokrovskaya MV, et al. Cisplatin-induced apoptotic endonuclease EndoG inhibits telomerase activity and causes malignant transformation of human CD4+ T lymphocytes. *Biochemistry.* (2017) 11:251–64. doi: 10.1134/S199075081703012X

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 04 August 2020
doi: 10.3389/fbioe.2020.00766

Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection

Junyi Li[1†], *Qingzhe Xu*[1†], *Mingxiao Wu*[1], *Tao Huang*[2*] *and Yadong Wang*[1*]

[1] Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, [2] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

Cancer is a one of the severest diseases and cancer classification plays an important role in cancer diagnosis and treatment. Some different cancers even have similar molecular features such as DNA copy number variant. Pan-cancer classification is still non-trivial at molecular level. Herein, we propose a computational method to classify cancer types by using the self-normalizing neural network (SNN) for analyzing pan-cancer copy number variation data. Since the dimension of the copy number variation features is high, the Monte Carlo feature selection method was used to rank these features. Then a classifier was built by SNN and feature selection method to select features. Three thousand six hundred ninety-four features were chosen for the prediction model, which yields the accuracy value is 0.798 and macro F1 is 0.789. We compared our model to random forest method. Results show the accuracy and macro F1 obtained by our classifier are higher than those obtained by random forest classifier, indicating the good predictive power of our method in distinguishing four different cancer types. This method is also extendable to pan-cancer classification for other molecular features.

Keywords: cancer classification, pan-cancer, self-normalizing neural network, copy number variation, feature selection

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Guimin Qin,
Xidian University, China
Jie Yan,
Yale University, United States
Mao Peng,
Westerdijk Fungal Biodiversity Institute, Netherlands

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yadong Wang
ydwang@hit.edu.cn

[†]*These authors have contributed equally to this work*

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

Received: *05 May 2020*
Accepted: *17 June 2020*
Published: *04 August 2020*

Citation:
Li J, Xu Q, Wu M, Huang T and Wang Y (2020) Pan-Cancer Classification Based on Self-Normalizing Neural Networks and Feature Selection. Front. Bioeng. Biotechnol. 8:766. doi: 10.3389/fbioe.2020.00766

BACKGROUND

Cancer is a one of the severest diseases which cause abnormal cell growths or tumors that metastasize to other parts of human body (Mayer et al., 2017). There are around 8 million human deaths related to cancer each year (Wild et al., 2014). Cancer classification is important for cancer diagnosis and drug discovery and can help improving treatment of patients and their life quality (Lu and Han, 2003). To decrease the effect of cancer to human health, tremendous research has been done to the cancer diagnosis and treatment, among which molecular-feature-based cancer classification is an important perspective. Due to the drop in the cost of sequencing technology in recent years, the output of sequencing data has increased dramatically. This provides adequate data for cancer analysis. Copy number variance (CNV) has also been shown to be associated with different cancers (Greenman et al., 2007; Wang et al., 2013). Some different cancers even have similar CNV patterns and mechanisms (Hoadley et al., 2018). We focus on CNV data analysis in this study. We aim to find out an applicable computational method to classify different cancer types. At present, some machine learning models are widely used in data analysis. Some models have been used to analyze the CNV data for cancer analysis (Ostrovnaya et al., 2010; Ding et al., 2014). The utility of machine learning in revealing

relationships between recurrent constitutional CNVs and cancers shows CNV data analysis is applicable to multi-type of cancers with a significant molecular component.

Deep learning has recently been widely used in computational scientific areas such as computer vision, natural language processing, computational biology (LeCun et al., 2015; Najafabadi et al., 2015; Angermueller et al., 2016; Sultana et al., 2020). The essence of deep learning algorithms is the domain independent idea of using hierarchical layers of learned abstraction to efficiently accomplish a complicated task. It uses many layers of convolutional or recurrent neural networks. The feed-forward neural network (FNN) is suitable for data without sequential features. However, there are some drawbacks of the FNNs. For instance, internal covariate shift (Ioffe and Szegedy, 2015) might causes the low training speed and poor generalization (Bengio et al., 1994; Pascanu et al., 2012, 2013). FNN might leads to invalid gradient too (Klambauer et al., 2017). Therefore, normalization is used and the self-normalizing neural network (SNN) (Klambauer et al., 2017) is proposed to overcome these short backs. SNNs make it possible for deep network applications on general data such as sequencing CNV data and SNNs have yielded the best results on some drug discovery and astronomy tasks.

In this study, we use a SNN-based prediction model to classify and analyze cancer patients with four cancers (LUAD, OV, LIHC, and BRCA). The data we used come from CNV data of The Cancer Genome Atlas (TCGA) (Grossman et al., 2016). We integrate a method which was used by Pan et al. (2018) to identify atrioventricular septal defect in Down syndrome patients to build our prediction model. Since the CNV data has a very high dimension, feature selection method is applied to identify important CNV features. Then a deep SNN model is trained based on these CNV features to perform pan-cancer classification. The normally used classification algorithm random forest (Cutler et al., 2012) is also used to compare with our model for its predictive ability in four different types of patient samples.

METHODS

Data Retrieval and Preprocessing

We download and collate the copy number variation data of 518 Lung adenocarcinoma (LUAD) patients, 597 Ovarian serous cystadenocarcinoma (OV) patients, 372 Hepatocellular carcinoma (LIHC) patients and 597 Breast cancer (BRCA) patients from TCGA database (Grossman et al., 2016), including the information of the copy number variation of probes. We use GISTIC2.0 (Mermel et al., 2011) to analyze the data. GISTIC2.0 can identify the key drivers of somatic copy number alterations (SCNAs) by the frequency and magnitude of mutation events. By using GISTIC2.0, we can select more important copy number variant genes, and then model the molecular information data of cancer patients more precisely. From the result generated form GISTIC2.0, we get a table which has 23,109 features. A series of discrete values is used to represent the specific type of copy number variation.

Approach for Cancer Classification

Feature Analysis

Since the dimensions of the CNV features are high, in order to avoid over-fitting, we need to select some features that can effectively classify patients. Therefore, we employed Monte Carlo Feature Selection (MCFS) (Draminski et al., 2008) and Incremental Feature Selection (IFS) methods as we used these two method before (Pan et al., 2018).

Monte Carlo Feature Selection method is proposed to improve a feature ranking obtained from an ensemble of decision trees. The general idea is to select s subset of the original d features, each with m features randomly selected. We repeat the selection process for s times, so that s feature subsets and a total of $t \times s$ tree classifier was obtained. Each feature f is assigned a score called relative importance (RI_f) which is assigned greater to feature f if it contributes more in the classification using the tree classifiers. RI of f is estimated by the Equation (1):

$$RI_f = \sum_{\tau=1}^{s^*t} (wAcc)^u \sum_{n_f(\tau)} IG(n_f(\tau)) \left(\frac{no.\ in\ n_f(\tau)}{no.\ in\ \tau} \right)^v \quad (1)$$

$\boldsymbol{wAcc}$ is the weighted accuracy and $\boldsymbol{IG(n_f(\tau))}$ is the information gain of node $\boldsymbol{n_f(\tau)}$. no.in $\boldsymbol{n_f(\tau)}$ is the number of patients in $\boldsymbol{n_f(\tau)}$ and no.in $\boldsymbol{\tau}$ is the number of patients in tree $\boldsymbol{\tau}$. u and v are a fixed real number.

The $\boldsymbol{wAcc}$ is defined by Draminski as Equation (2):

$$wAcc = \frac{1}{c} \sum_{i=1}^{c} \frac{n_{ii}}{n_{i1}+n_{i2}+\ldots+n_{ic}} \quad (2)$$

In Equation (2), c is the number of classes and $\boldsymbol{n_{ij}}$ is the number of patients from class i that are classified as class j. The $\boldsymbol{IG(n_f(\tau))}$ is defined by Equation (3):

$$IG\left(n_f(\tau)\right) = Entropy(T) - Entropy\left(T, f\right) \quad (3)$$

In Equation (3), T is the class label of node $\boldsymbol{n_f(\tau)}$, Entropy(T) is the entropy of the frequency table of T and Entropy (T, f) is the entropy of the frequency table of the two variables T and f.

We used the MCFS method of Draminski and obtained a ranked feature list according to their RI values evaluate by the algorithm, which can be defined as Equation (4).

$$F = \left[f_1, f_2, \ldots, f_M\right] \quad (4)$$

And in Equation (4) M means the 23,109 CNV features.

Then we aimed to select a subgroup of CNV features to build a classification model. Therefore, in order to avoid training all CNV feature sets, we used Incremental Feature Selection method on previous obtained feature list. We first determine the approximate feature interval from which we can find optimal features. We defined CNV feature subsets as $S_1^1, S_2^1, \ldots, S_I^1$, where $S_i^1 = f_1, f_2, \ldots, f_{i^*k}$, i.e., and the ith feature subset had the first i

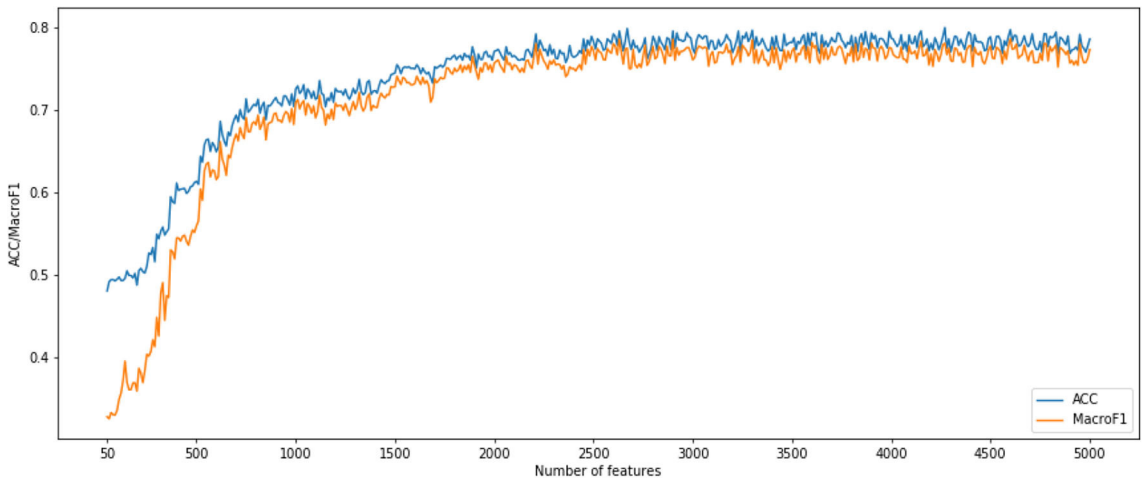


FIGURE 1 | Incremental feature selection (IFS) curves derived from the IFS method and SNN algorithm. IFS curve with *X*-values from 50 to 5,000.

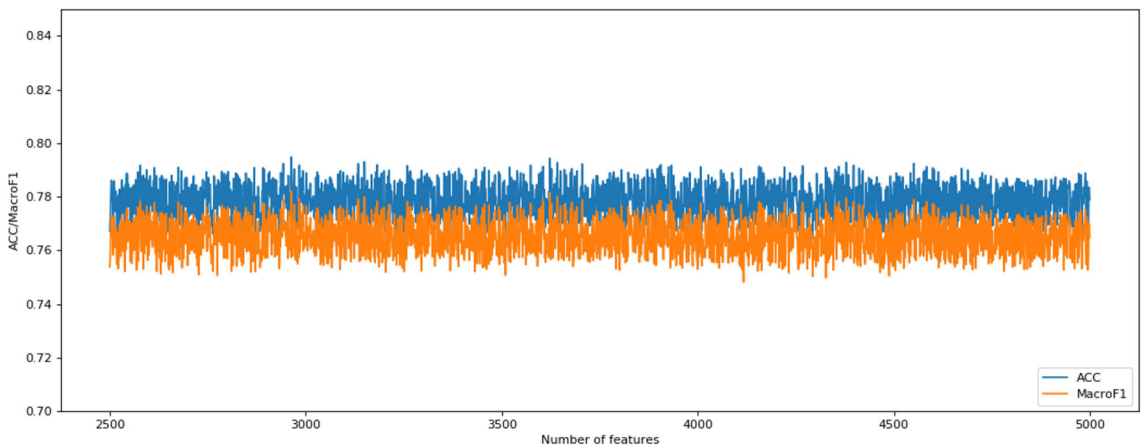


FIGURE 2 | Incremental feature selection (IFS) curves derived from the IFS method and SNN algorithm. IFS curve with X-values of 2501–4,999 for SNN algorithm.

times k features in the original M CNV feature list. Classification model was built by using features in each feature subset of corresponding patient samples in dataset. To estimate the CNV feature interval, we tested performances of different classification model based on different subsets. The feature subset was selected when it had the best performance.

Classification Methods

We need an algorithm to classify pan-cancer patients based on the selected subset of CNV features. Here, neural network SNN was used and RF method was applied for comparison.

(a) Self-Normalizing Neural Network Algorithm

SNN is proposed to enable high-level abstract representations through keeping neuron activations converge toward zero mean and unit variance (Klambauer et al., 2019). Klambauer et al. proposed a Scale ELU (SELU) function as activation function.

$$selu\,(x) = \lambda \begin{cases} x, & x > 0 \\ \alpha e^{x} - \alpha, & x \leq 0 \end{cases} \tag{5}$$

where scale $\lambda = 1.0507$ and $\alpha = 1.6733$ (see Klambauer et al., 2017 for details on the derivation of these two parameters).

By using the Banach fixed-point theorem, Klambauer et al. prove that activations close to zero mean and unit variance that are propagated through many network layers will converge toward zero mean and unit variance. A specific method to initialize SNNs and alpha dropout (Klambauer et al., 2017) are also proposed to make SNNs have a fixed point at zero mean and unit variance. In this study, the SNN classifiers those we constructed have three hidden layers with 200 hidden nodes of each layer.

(b) Random Forest Algorithm

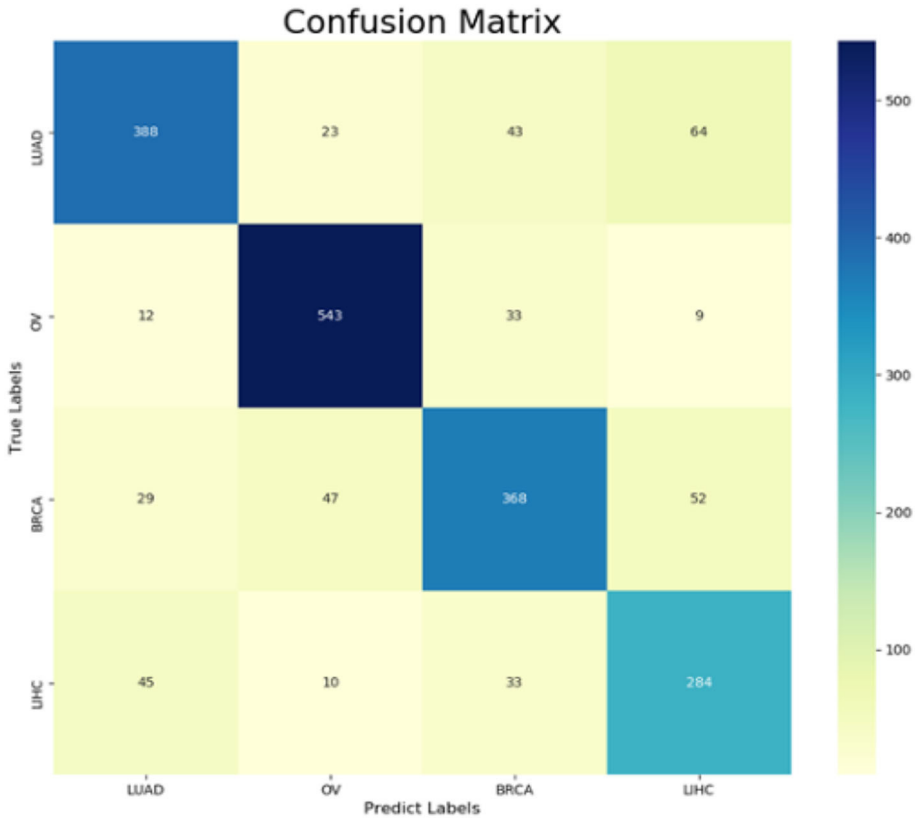


FIGURE 3 | Confusion matrix from pan-cancer classification by using SNN and features selection.

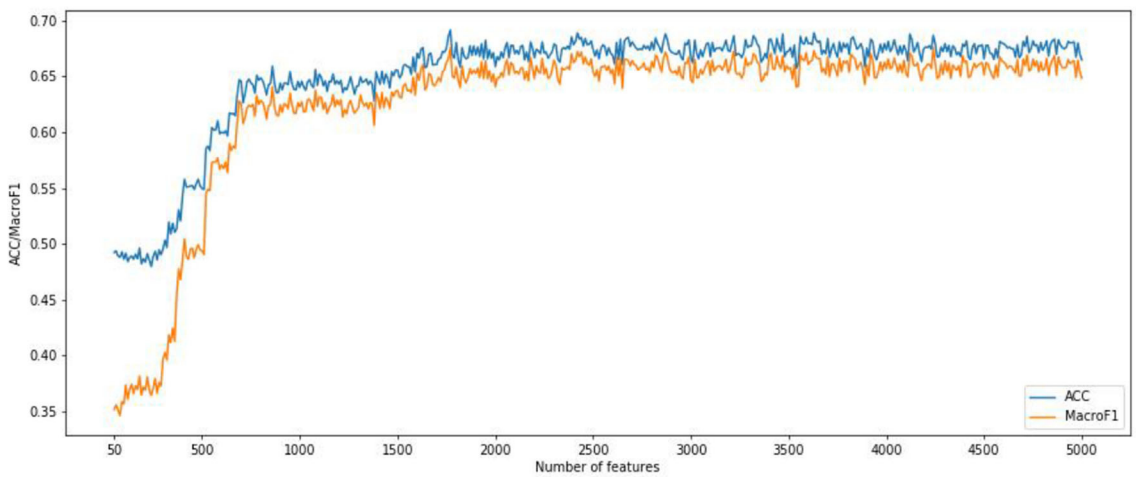


FIGURE 4 | Incremental feature selection (IFS) curves derived from the IFS method and RF algorithm. IFS curve with *X*-values from 50 to 5,000.

The random forest (RF) method is a supervised classification and regression algorithm (Cutler et al., 2012). The RF method builds multiple decision trees and merges them together to get a more accurate prediction. It adds additional randomness to the model when it growing the trees. Instead of searching for the most important feature when splitting a node, it searches for the best feature among a random subset of features. This generally results in a better model. The RF method has been widely used in machine learning area and is applied here to compare our model.

Performance Evaluation

Since pan-cancer classification is a multi-classification problem, we use accuracy (ACC) to measure the performance. There are also precision and recall to measure performance in a binary classification problem. One measurement closely related to these two values is F-score, which is a comprehensive indicator of precision and recall. That means, F-score is a parameter used to adjust the ratio of these two parts. When this parameter is 1, it degenerates into a harmonic average called F1-score. The multi-classification evaluation was split into multiple binary classification problems, and each F1-score was calculated. The average of the F1 scores was defined as Macro F1. To evaluate prediction of SNN classifier, we performed a 10-fold cross-validation (Kohavi, 1995; Chen et al., 2017, 2018).

RESULTS

To evaluate the best features for discriminating four types of cancer samples, a MCFS method was used to rank all features according to their RI values by using Monte Carlo method and decision trees. We selected the top 5,000 CNV features and applied IFS method.

After using MCFS for CNV feature sorting, we obtained two feature subset series. For the first CNV feature subsets, the parameter k is set to 10. That means, the i-th feature subset contains the first 10 times i features in the original CNV feature list. We constructed an SNN-based classification model on each feature subset, performed a 10-fold cross-validation and calculated its accuracy and macro F1 values. To show the changes of accuracy and macro F1 values, an IFS curve was generated as **Figure 1**. In **Figure 1**, the accuracy and macro f1 values are the Y axis and the number of features is the X axis. Both curves become stable after number of features >2,500 and them reached acceptable values. Therefore, we selected the number interval as [2,500, 4,999] for classifier to select the best number of features.

The following CNV feature subset is constructed by using the number of features in the number interval [2,500, 4,999]. By testing all of these subsets, we obtained the corresponding accuracy and macro F1 values. We also plotted the IFS curves to show these values in **Figure 2**. The best accuracy and macro

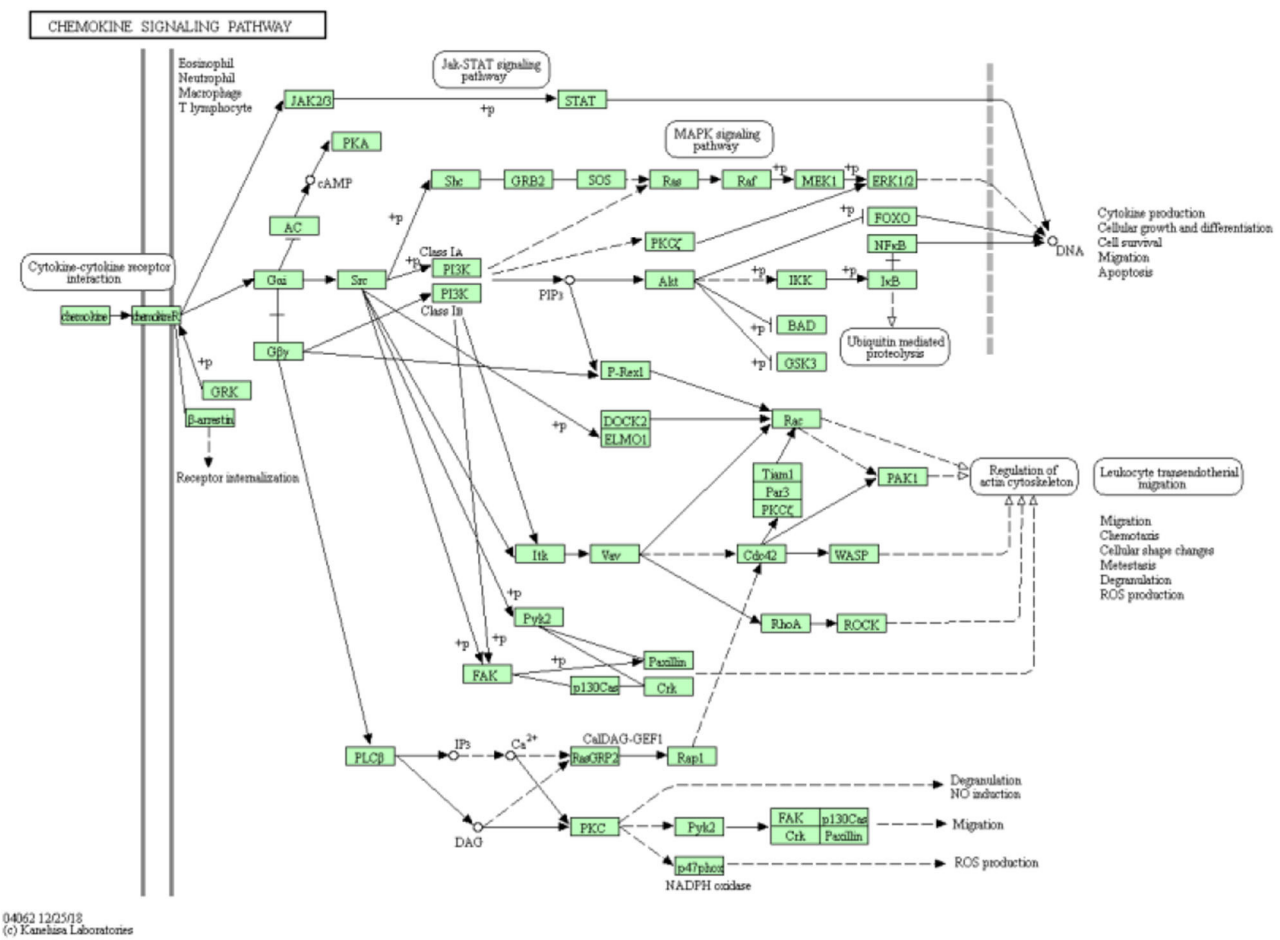


FIGURE 5 | Chemokine signaling pathway from KEGG has the highest counts for selected feature genes.

F1 values were generated when using the first 3,694 features to construct the SNN-based classification model. Thus, these first 3,694 genes were select for the final model. In the meantime, we used RF method as a comparison. The RF generated accuracy and macro F1 are much lower than the SNN one, which proves the efficiency of the deep SNN classifier. Therefore, we obtained the best feature subset and the optimal SNN-based model. Its ACC is 0.798 and the corresponding macro F1 is 0.789. **Figure 3** is confusion matrix and shows the good classification result from our model.

We also implemented the RF algorithm to construct a classifier on the CNV features subset obtained from the IFS method and evaluate each classifier through a 10-fold cross-validation test. Since the fast speed of RF method, which promised all CNV feature sets were tested. In order to compare the classification feature selection results, the IFS curves of accuracy and macro F1 were plotted in **Figure 4**. It can be seen that the optimal accuracy value is 0.689 and the macro F1 is 0.667 when using the first 1,693 features in the CNV feature list. Therefore, the first 1,693 features and RF algorithms can construct the best RF classification model. It can be seen that the accuracy and macro F1 obtained by the best RF classifier are much lower than those obtained by the best SNN-based classification model. That means our SNN-based model is effective in pan-cancer classification analysis.

DISCUSSION

DNA copy number variation is a straight-forward mechanism, which provides insight into genomic instability and structural dynamism in cancer researches. We applied Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis to the first 200 selected features and checked whether these were significant pathway information as shown as **Figure 5**. The highest counts are on the Chemokine signaling pathway, where chemoattractant proteins play an important role in controlling leukocyte migration during development, homeostasis, and inflammation. These processes are closely related to the occurrence and development of various cancers.

CONCLUSIONS

In this study, we use machine learning method for CNV-based pan-cancer classification. Considering the high dimension of data, MCFS and IFS are used to classify four different cancer patients effectively. And the feature subsets generated from IFS method are classified by integrating SNN method. Comparison experiments show that our SNN-based classification method has significant advantages over random forest in cancer classification. We demonstrate the advantages and potential of this method for copy number variant data. We suggest that this model can be extended and transferred to other pan-cancer classification fields. For future research, we will improve the models of other complex and large-scale data and expand our training data sets to further improve classification results.

DATA AVAILABILITY STATEMENT

The data and code are available at https://github.com/KohTseh/CancerClassification.

AUTHOR CONTRIBUTIONS

JL and QX leaded the method application, experiment conduction, the result analysis, and drafted the manuscript. QX and MW participated in the data extraction and preprocessing. TH and YW provided theoretical guidance and the revision of this paper. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the grants from the National 863 Key Basic Research Development Program (2014AA021505) and the startup grant of Harbin Institute of Technology (Shenzhen). National Natural Science Foundation of China (31701151), National Key R&D Program of China (2018YFC0910403), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), Shanghai Sailing Program (16YF1413800) and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

REFERENCES

Angermueller, C., Pärnamaa T, Parts, L., and Stegle, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.* 12:878. doi: 10.15252/msb.20156651

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural. Netw.* 5, 157–166. doi: 10.1109/72.279181

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

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., Huang, T., and Cai, Y.-D. (2017). Identify key sequence features to improve CRISPR sgRNA efficacy. *Biochim. Biophys. Acta* 5, 26582–26590. doi: 10.1109/ACCESS.2017.2775703

Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). "Random forests," in *Ensemble Machine Learning*, eds C. Zhang and Y. Ma (Golden Valley, MN: Springer), 157–175. doi: 10.1007/978-1-4419-9326-7_5

Ding, X., Tsang, S. Y., Ng, S. K., and Xue, H. (2014). Application of machine learning to development of copy number variation-based prediction of cancer risk. *Genomics Insights*. 7, 1–11. doi: 10.4137/GEI.S15002

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

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375, 1109–1112. doi: 10.1056/NEJMp1607591

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e296. doi: 10.1016/j.cell.2018.03.022

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv*.

Klambauer, G., Hochreiter, S., and Rarey, M. (2019). Machine learning in drug discovery. *J. Chem. Inf. Model.* 59, 945–946. doi: 10.1021/acs.jcim.9b00136

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. (2017). "Self-normalizing neural networks" in *Advances in Neural Information Processing Systems* (Long Beach, CA), 972–81.

Kohavi, R. (1995). *A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. Montreal, CA: IJCAI.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lu, Y., and Han, J. J. (2003). Cancer classification using gene expression data. *Inf. Syst*. 28, 243–268. doi: 10.1016/S0306-4379(02)00072-8

Mayer, D. K., Nasso, S. F., and Earp, J. A. (2017). Defining cancer survivors, their needs, and perspectives on survivorship health care in the USA. *Lancet Oncol*. 18, e11–e18. doi: 10.1016/S1470-2045(16)30573-3

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *J. Big Data* 2:1. doi: 10.1186/s40537-014-0007-7

Ostrovnaya, I., Nanjangud, G., and Olshen, A. B. (2010). A classification model for distinguishing copy number variants from cancer-related alterations. *BMC Bioinformatics* 11:297. doi: 10.1186/1471-2105-11-297

Pan, X., Hu, X., Zhang, Y. H., Feng, K., Wang, S. P., Chen, L., Huang, T., and Cai, Y. D. (2018). Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* 9:208. doi: 10.3390/genes9040208

Pascanu, R., Mikolov, T., and Bengio, Y. J. C. (2012). Understanding the exploding gradient problem. *Arvix.* 2:417.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). "On the difficulty of training recurrent neural networks," in *International Conference on Machine Learning*, (Atlanta, GA), 1310–1318.

Sultana, J., Rani, M. U., and Farquad, M. (2020). "An extensive survey on some deep-learning applications," in *Emerging Research in Data Engineering Systems and Computer Communications*, eds P. V. Krishna and M. S. Obaidat (Springer), 511–519. doi: 10.1007/978-981-15-0135-7_47

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., et al. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* 5: 91. doi: 10.1186/gm495

Wild, C. P., Stewart, B. W., and Wild, C. (2014). *World Cancer Report 2014*. Geneva: World Health Organization.

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 Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 August 2020
doi: 10.3389/fbioe.2020.00897

A New Method for CTC Images Recognition Based on Machine Learning

Binsheng He[1†], Qingqing Lu[2,3†], Jidong Lang[2,3], Hai Yu[2], Chao Peng[2], Pingping Bing[1], Shijun Li[4], Qiliang Zhou[1*], Yuebin Liang[2,3*] and Geng Tian[2,3*]*

[1] Academician Workstation, Changsha Medical University, Changsha, China, [2] Geneis (Beijing) Co., Ltd., Beijing, China, [3] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [4] Department of Pathology, Chifeng Municipal Hospital, Chifeng, China

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Juanying Xie,
Shaanxi Normal University, China
Khanh N. Q. Le,
Taipei Medical University, Taiwan

***Correspondence:**
Binsheng He
hbscsmu@163.com
Qiliang Zhou
13974942986@163.com
Yuebin Liang
liangyb@geneis.cn
Geng Tian
tiang@geneis.cn

† These authors have contributed equally to this work

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

Received: 09 March 2020
Accepted: 13 July 2020
Published: 06 August 2020

Citation:
He B, Lu Q, Lang J, Yu H, Peng C, Bing P, Li S, Zhou Q, Liang Y and Tian G (2020) A New Method for CTC Images Recognition Based on Machine Learning. Front. Bioeng. Biotechnol. 8:897. doi: 10.3389/fbioe.2020.00897

Circulating tumor cells (CTCs) derived from primary tumors and/or metastatic tumors are markers for tumor prognosis, and can also be used to monitor therapeutic efficacy and tumor recurrence. Circulating tumor cells enrichment and screening can be automated, but the final counting of CTCs currently requires manual intervention. This not only requires the participation of experienced pathologists, but also easily causes artificial misjudgment. Medical image recognition based on machine learning can effectively reduce the workload and improve the level of automation. So, we use machine learning to identify CTCs. First, we collected the CTC test results of 600 patients. After immunofluorescence staining, each picture presented a positive CTC cell nucleus and several negative controls. The images of CTCs were then segmented by image denoising, image filtering, edge detection, image expansion and contraction techniques using python's openCV scheme. Subsequently, traditional image recognition methods and machine learning were used to identify CTCs. Machine learning algorithms are implemented using convolutional neural network deep learning networks for training. We took 2300 cells from 600 patients for training and testing. About 1300 cells were used for training and the others were used for testing. The sensitivity and specificity of recognition reached 90.3 and 91.3%, respectively. We will further revise our models, hoping to achieve a higher sensitivity and specificity.

Keywords: circulating tumor cells (CTCs), imFISH, machine learning, image segmentation, CNN network

INTRODUCTION

The metastasis of cancers is a complex and multistage process. The circulating tumor cells (CTCs) are the "seeds" shed from the primary tumor and/or metastatic lesions and rooted in a new "soil" transferred by the circulatory system (Paget, 1989). Circulating tumor cell is an intermediate stage of cancer metastasis, correlated with cancer aggressiveness and the likelihood of metastasis, and therefore can be used to predict disease progression and survival on a real-time basis by liquid biopsy (Lindsay et al., 2017; Praharaj et al., 2018; Anand and Roszik, 2019; Baek et al., 2019; Maly et al., 2019; Marcuello et al., 2019; Pan et al., 2019; Riebensahm et al., 2019). The molecular subtypes of CTCs, not only the CTCs count, are interrelated with the prognosis (Banys-Paluchowski et al., 2015; Cristofanilli et al., 2019; Dong et al., 2019; Stefanovic et al., 2019). What's more, the PD-L1

expression in CTCs is correlated with the response to immunity inhibitors (Kloten et al., 2019). PD-L1$^+$/EMT$^+$ CTCs were associated with significantly poorer survival after curative surgery, showing that PD-L1 expression and Epithelial Mesenchymal Transition (EMT) of CTCs are negative survival predictors for Non-small cell lung cancer (NSCLC) patients (Janning et al., 2019; Manjunath et al., 2019). Pre-treatment PD-L1$^+$ CTCs are usually associated with a bad prognosis in patients treated with PD-1 inhibitors in NSCLC, such as nivolumab (Guibert et al., 2018).

The liquid biopsies worked as an ongoing monitoring system to assess tumor heterogeneity, and make it possible to detect a single CTC or clusters of cells (Wan et al., 2017; Merker et al., 2018; Praharaj et al., 2018; Asante et al., 2020). The breakthrough for CTC-detection is the application of immunomagnetic CTC enrichment combined with flow cytometry, which is still the "gold" standard of CTC-detection (Racila et al., 1998). However, this method that lack of the cancer specific markers still remains lots of limitation (Grover et al., 2014; Ferreira et al., 2016; Gabriel et al., 2016; Keller et al., 2019). Thus, the multi-marker immunofluorescence staining is required for recognize CTCs. Antibodies against chromosome 8 centromere duplication (CEP8)/chromosome 17 centromere duplication (CEP17) are used to mark the rapidly dividing tumor cells; antibodies against CD45 as typical leukocytes filaments, as well as 4′,6-diamidino-2-phenylindole (DAPI) for labeling nuclears (Koudelakova et al., 2016; Lu et al., 2017; Liu et al., 2018; Lee et al., 2019). Although there are great advantages in enrichment technology, the automatic recognition of CTCs still remains problems. Manual identification is very time-consuming and unreliable. With the continuous deepening of the application of CTCs recognition in various cancer diseases, the demand for rapid and automatic identification and counting methods of CTCs is increasing. Several studies have reported the automated screening process (Nagrath et al., 2007; Yang et al., 2018). Kraeft et al. (2004) performed a fluorescence-based automated microscope system, REIS, for cell detection. This scanning can quantify the number of cells reliably and reproducibly and categorize positive cells based on the marker expression profile. Ligthart et al. (2011) redefined the CTCs by computer algorithms after the manual counting. The stricter definition, with the standard deviation of the signal in the CK-PE channel, the peak signal value in both the DNA-DAPI and CD45-APC channels and the size of the objects used as classifier, was well validated CTC by clinical outcome using a perfectly reproducing automated algorithm. Mingxing et al. reported an automated CTC enumeration (Zhou et al., 2017). All images with different colors were transferred to a grayscale image and the grayscale images were used to identify the position and outline of cells. However, despite the widely accepted, these classification methods still remain subjective, as the rules are set artificially. The fixed conditions may not identify the morphologically heterogeneous CTCs integrally. What's more, different technologies usually use different antibodies, making comparison and standardization across different platforms challenging (Marcuello et al., 2019).

With the maturity of artificial intelligence (AI) recent years, machine learning become an exciting field for research. The U.S. Food and Drug Administration (FDA) has approved several commercial products using machine-learning algorithms in the medical diagnosis and research. The cardiovascular MRI analysis software of Arterys was the world's first internet platform for medical imaging, AI powered and FDA cleared. This software is able to analyze multiple, multi-period MR images to determine blood flow in heart and main vessels. The cloud platform will enable software to collect and analyze the vast amount of cardiovascular data from MR scanners in real time, which will speed up doctors' diagnosis. This artificial machine is consistent and tireless and is able to identify characters beyond human perception, which provided a substantial interest in the field of medical research, specifically medical images (Dominguez et al., 2017; Erickson et al., 2017; Lundervold and Lundervold, 2019; Maier et al., 2019). Many algorithms are developed for selecting the best weights for features, involving neural networks (Hornik et al., 1989), decision trees (Quinlan, 1986), support vector machines (Cristianini and Shawe-Taylor, 2000), the naïve Bayes (Lowd and Domingos, 2005), k-nearest neighbors (Zhou and Chen, 2006), and deep learning (McBee et al., 2018; Wainberg et al., 2018; Zou et al., 2019). Deep learning, as well as deep neural network learning, refers to the use of neural networks with more than 20 layers, able to integrate vast datasets, learn arbitrarily complex relationships and incorporate existing knowledge. Convolutional neural networks (CNNs) is a powerful algorithm for advancing biomedical image analysis as it assumes that the input layer has a geometric relationship, such as the rows and columns of images (Anthimopoulos et al., 2016; Poplin et al., 2018). It has been successfully applied in the cancer diagnosis and nuclei or tissue identification (Le et al., 2017, 2018; Le et al., 2019). Xing et al. (2015) present a novel method for automated nucleus segmentation powered by CNNs. The features involved in the images are considered as a part of the search process, and there is no need to limit the features compared to the traditional machine learning methods, which will eliminate the bias created subjective. Here, we apply deep learning to the recognition of CTCs in order to reduce the artificial errors and improve accuracy.

MATERIALS AND METHODS

Patients and Samples Preparation

A cohort of 600 patients with cancers were enrolled in this study during 2018–2019, which was approved by the ethics committee of Chifeng Municipal Hospital. The clinical pathological characteristics of patients including age, gender, CTC number, and cancer type are summarized in **Table 1**. Four milliliter of peripheral venous blood was routinely collected for every patient. The first 2 ml blood samples obtained after puncture was discarded in order to avoid the skin epithelial cells contamination. Then the blood was placed in anticoagulation tubes and store at room temperature. The test was completed within 24 h.

All the 600 patients were divided into two parts according to the collecting date. The earlier 300 patients we collected were used as the training data, the others were used as the independent

TABLE 1 | Clinical pathological characteristics.

Clinicopathologic variable	Category	Clinical level
Age	Mean	65(11–90)
Gender	Male	256
	Female	141
	Unknown	203
Samples type	Peripheral blood	100%
CTC number	Mean	7.8(0–185)
Cancer type	Lung cancer	158(26.3%)
	Liver cancer	12(2.0%)
	Gastrointestinal cancer	45(7.5%)
	Breast cancer	70(11.7%)
	Carcinoma of thyroid	1(0.2%)
	NPC	9(1.5%)
	Other	305(50.8%)

testing data. Thousand three hundred cells images in the earlier received 300 patients were selected to build the CTC recognition model, which will be further tested by the 1000 cells images of the test dataset. There was no cross part between the two datasets in order to avoiding the over-fitting.

Enrichment and imFISH Identification of CTCs

The Cyttel method was used to isolate and enumerate CTCs. The peripheral blood was first centrifuged at 600 g for 5 min to get the precipitation and then washed by CS1 buffer (Cyttel Biosciences Co. Ltd., Beijing, China). Then the red blood cells were lysed by CS2 buffer (Cyttel). After centrifuged at 600 g for 5 min, the precipitate was washed by CS1 buffer. Then the cells were incubated completely with anti-CD45 monoclonal antibody-conjugated beads (Cyttel) for 20 min. Three milliliter separation medium was used to separate the beads and the CTCs by gradient centrifugation at 300 g for 5 min. Then the upper rare cell layer was centrifuged at 600 g for 5 min and re-suspended by CS1. The tube was put on a magnetic stand for 2 min. After smeared, fixed and dried, cells were used to perform the imFISH.

The slides were fixed, dehydrated and then dried at room temperature. 10 μl CEP-8/CEP-17 antibody was added to the cells and the slides were placed in a hybridization and denatured for 1.5 h at 37°C. The probe was eluted and the slides were washed twice in 2 × SSC. Then the CD45 fluorescent antibody was added to the sample area and the slides were put in a wet box and incubate for 1 h at 33°C. After incubation, CD45 fluorescent antibody was aspirated and 10 μl mounting media containing DAPI was added to the sample area. After mounted, the cells can be observed and counted under a fluorescence microscope.

The Manual Interpretation Standard of CTCs Counting

After imFISH, lots of images were acquired with different fluorescent colors. Usually, manual counting is the "gold standard," but it's a time consuming and exhausted procession. The Manual interpretation standard of CTCs counting is: (1) Eliminates the aggregation, superposition and interference of nucleus or impurity, (2) DAPI positive, (3) CD45 negative, and (4) Three or more than three CEP-8^+/CEP-17^+ signal points. It will be regarded as one signal point if the distance between two signal points is smaller than the diameter of one point.

The Image Segmentation Method Was Used to Segment Single Nucleus and Give Labels of Cells Instead of Manual

Since the obtained microscopic image is very huge, the algorithm will be limited by the memory and cannot be executed normally on a conventional computer. We first selected part of the image containing one CTC cell and several non-CTC cells around to perform the following test. The chosen resolution is 2728 × 2192.

The openCV package of python was used to process the CTCs images, including conversion of color and morphological transformations.

(1) The RGB image was converted to the gray image;
(2) The derivatives were calculated using the OpenCV function Sobel from an image;
(3) Morphological transformations operations based on the image shape.

The Morphological package of python was used to segment the images of CTCs by image denoising, image filtering, edge detection, image expansion and contraction.

Nuclei were segmented in the blue channel (DAPI), and the proportion of red in the red channel was detected based on the position of the nucleus. The nucleus with proportion of red higher than 30% was defined as having a common leukocyte antigen. The orange channel was used to detect the number of $CEP8^+$ chromosomes and the green channel was used to detect the number of centromere probes extracted by $CEP17^+$. Different cell types were distinguished by different colors (**Figure 1**).

The CNN Deep Learning Method Was Used for CTCs Identification

With the development of AI, machine learning has been wildly used in the procession of medical images. Deep learning is a big improvement on artificial neural networks, allowing higher-level feature extraction and better data prediction with more layers. After segmentation, CNN network were used to identify CTC cells in single nucleus. Finally, it enters the output layer and output the result, i.e., CTCs or non-CTCs.

Our CNN model was built based on AlexNet, which was first introduced in 2012 (Krizhevsky et al., 2012). The network consists of eight weighted layers (**Figure 2**); the first five layers are convolution layers, and the remaining three layers are full connection layers. The output of the last full connection layer is the input of the 1000 dimensional softmax values, which will generate the distribution network of two types of labels.

The five-fold cross validation was used to prevent overfitting and select hyper-parameters of the model. The best cross-validation score was obtained by searching the hyper-parameter space round and round. The final hyper-parameters involved in

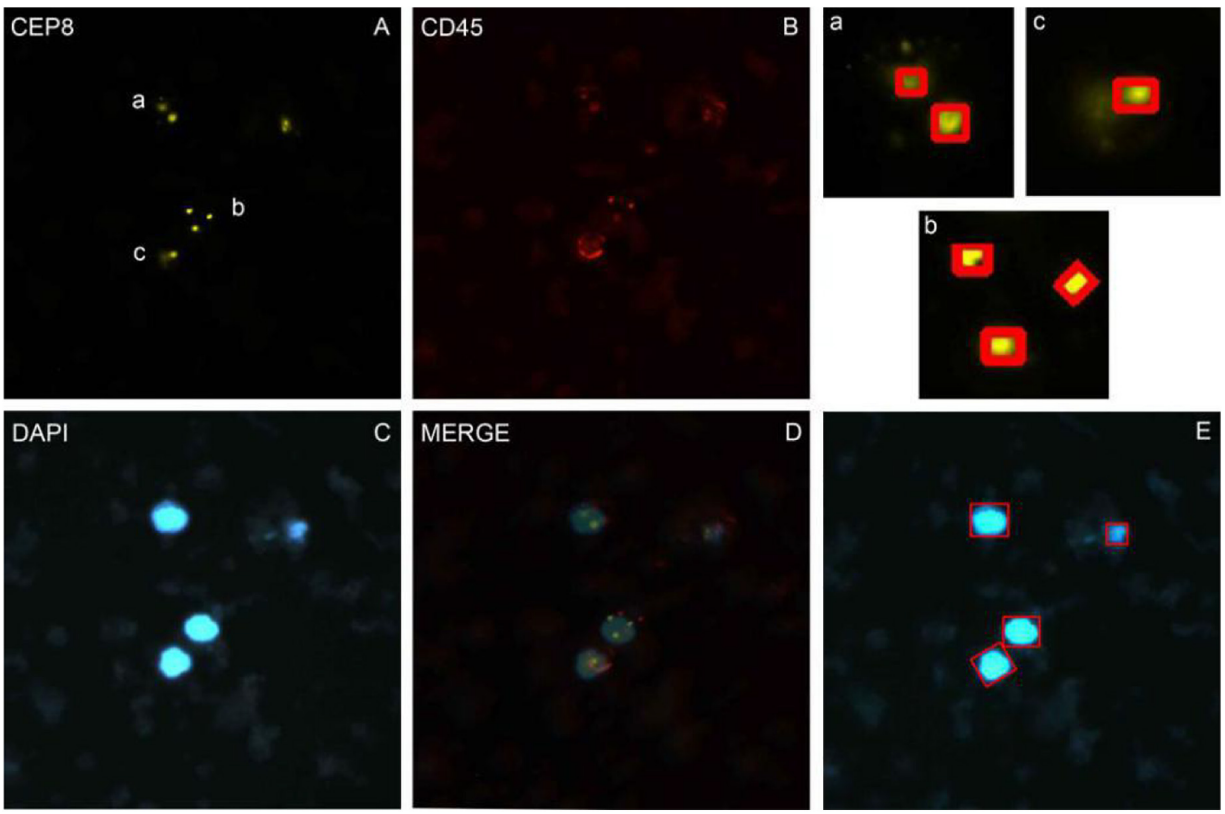


FIGURE 1 | The imFISH result and the segmentation of chromosome and nuclear. **(A–C)** The imFISH result of CEP-8, CD45 and DAPI; **(D)** The merge of panels **(A–C)**; **(E)** The CTCs were identified by openCV segmentation method and marked in red box; **(a–c)** The CEP-8 signal points were identified by openCV segmentation method and marked in red box.

our model are activation function, kernel regularizer type and regularization factor. The workflow is shown below:

(1) The grid was defined on 3-dimensions with each of these maps for hyper-parameter sets, e.g., hyper-parameters = (activation function, kernel regularizer type, regularization factor); activation function = ("softmax," "ReLU," "tanh"); kernel regularizer type = ("l1," "l2"); regularization factor = ("0.01," "0.02");
(2) The range of possible values were defined of each dimension;
(3) All the possible configurations were searched for establishing the best one.

Evaluation Criteria for Classification Models

After segmentation, some performance evaluation criteria (Xie et al., 2019) were involved in to evaluate the performance of the classification model, such as sensitivity (Se or recall), specificity (Sp), precision, F1 score and area under the receiver operating characteristic curve (AUC).

$$Se(recall) = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$\mathrm{F1} = \frac{2 \times precision \times recall}{precision + recall} \tag{4}$$

In the equations, *TP* stands for the number of positive CTC cells which are correctly recognized as positive CTC cells. *FP* stands for the number of negative CTC cells that are incorrectly recognized as positive CTC cells. *FN* stands for the number of positive CTC cells incorrectly recognized as negative CTC cells. *TN* stands for the number of negative CTC cells correctly recognized as negative CTC cells (**Table 2**).

RESULTS

Patient Characteristics

A total of 600 patients were enrolled in this study from January 2017 to June 2019. The average age is 65 years old. Patients with lung cancer count 26.3% of all patients, and the next is breast cancer and gastrointestinal cancer (**Table 1**).

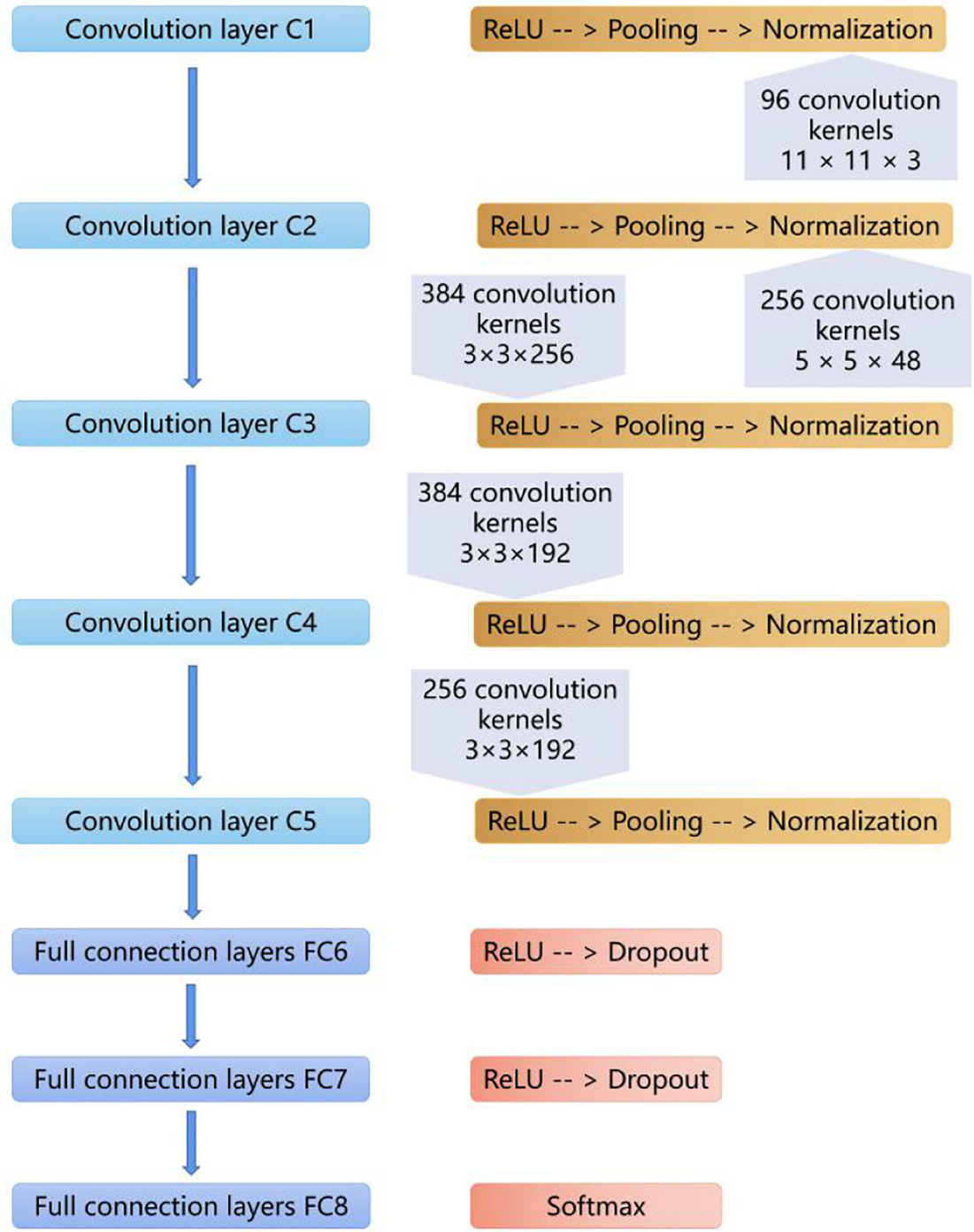


FIGURE 2 | The layers of the CNN model. The first five layers are convolution layers, and the remaining layers are full connection layers.

Three Sub-Images Were Required for Manual Counting

We performed imFISH for all the 600 patients and required 2300 images of CTCs cells. Every image was divided into 3 or 4 channels with different color. The orange channel represented the chromosome 8 with CEP8$^+$ (**Figure 1A**), the green channel represented the centromere of chromosome 17 with CEP17$^+$ (**Supplementary Figure S1**), the red channel represented the white cell with CD45$^+$ (**Figure 1B**), the blue channel represented the nuclei with DAPI$^+$ (**Figure 1C**). The mergence was shown in **Figure 1D**. We then manually labeled all these sub-images according

TABLE 2 | Confusion matrix definitions.

Confusion Matrix		Prediction	
		Positive	Negative
True	Positive	True positive (TP)	False Negative (FN)
	Negative	False positive (FP)	True Negative (TN)

to the standard. Among our results, 316 patients are CTCs positive.

The Segmentation of Nuclear and Identifying CTCs by OpenCV Segmentation Method

In order to avoid the artificial error and save costs, we performed the traditional image identification method for CTCs counting (**Figure 1**). The nucleus was separated in the blue channel (DAPI) (**Figure 1E**), and the red proportion of the red channel was detected according to the location of the cell nucleus. The proportion higher than 30% was defined as the number of the CEP8 chromosome detected by the common antigen orange channel of white blood cells (**Figures 1A–C**), the number of centromeric probes detected by the green channel, such as CEP17 (**Supplementary Figure S1**).

After segmentation of nuclear, we used openCV segmentation method to identify CTC cells from single nucleus regions in 1000 testing dataset by the manual interpretation standard of CTCs counting. After identification and judgment, 645 cells of 700 negative nuclei were recognized as CTC negative. About 278 cells of 300 positive nuclei were recognized as CTC negative. The sensitivity and specificity were 93.7 and 92.1%, while the precision and F1 score reached 83.6 and 88.4%, respectively (**Table 3**).

We also applied the region-based image segmentation algorithm such as watershed algorithm in the segmentation process. The watershed algorithm was implemented the by watershed function in OpenCV (python 3.6 and OpenCV 4.1.1). In this method, optimal threshold value was used respectively in binaryzation process by setting THRESH_OTSU mode. The traditional watershed algorithm was sensitive to noise and the accuracy was lower than our segmentation method on CTC negative data set in size of 100 (**Supplementary Table S3**).

The Hyper-Parameters Selected for Evaluating the CNN Method

We used GridSearchCV class in scikit-learn by providing a dictionary of hyper-parameters to determine the hyper-parameters of the model. After the cross-validation process, activation function was set to ReLU, kernel regularizer type was set to l2 and regularization factor was set to 0.01 as shown in **Table 4** with the best performance. Further, the hyper-parameters we selected were used to construct the model on the whole training dataset.

TABLE 3 | The confusion matrix of the models for test dataset.

Method	Confusion Matrix		Prediction	
			Positive	Negative
openCV	True	Positive	281	19
		Negative	55	645
ALexNet	True	Positive	271	29
		Negative	61	639

TABLE 4 | Tuning of the hyper-parameters of AlexNet.

Activation function	Kernel regularizer type	Regularization factor	
		0.01	0.02
softmax	l1	0.93	0.91
	l2	0.93	0.92
ReLU	l1	0.96	0.94
	l2	<u>0.96</u>	0.94
tanh	l1	0.94	0.93
	l2	0.94	0.93

The underline value shows the best result of AUC value in the tuning process of the hyper-parameters of AlexNet.

The Identification of CTCs by CNN Method

We got 2300 nuclei of 600 patients by segmentation process. **Figure 3** showed the whole flowchart of the experiment. About 1300 nuclei were used for training, the left 1000 were used for testing. We use the same images for testing. 639 cells of 700 negative nuclei were recognized as CTC negative and 271 cells of 300 were recognized as CTC positive. The sensitivity and specificity were 90.3 and 91.3%, while the precision and F1 score reached 81.6 and 85.7%, respectively (**Table 3** and **Figure 4**).

Before that, we also compared the performance of AlexNet model with others, such as ResNet and Xception. All of them have close AUC values (**Figure 4**), but the AlexNet was less time-consuming in the training and test process (**Supplementary Table S1**).

DISCUSSION

This study showed a method for CTC counting powered by machine learning. The use of machine learning for image interpretation can capture important image features, reduce errors caused by manually setting interpretation standards, and save time and labor costs. Although this method shows a higher sensitivity and specificity in CTC counting, it is slightly worse than the first method for the data used in this study. Actually, we have analyzed that the main reason is that there are fewer positive samples for training, and the algorithm cannot extract features of more positive samples. In addition, some pictures in the group were excluded due to quality problems. Unfortunately, the CTC images included in the group doesn't cover the whole film, but a

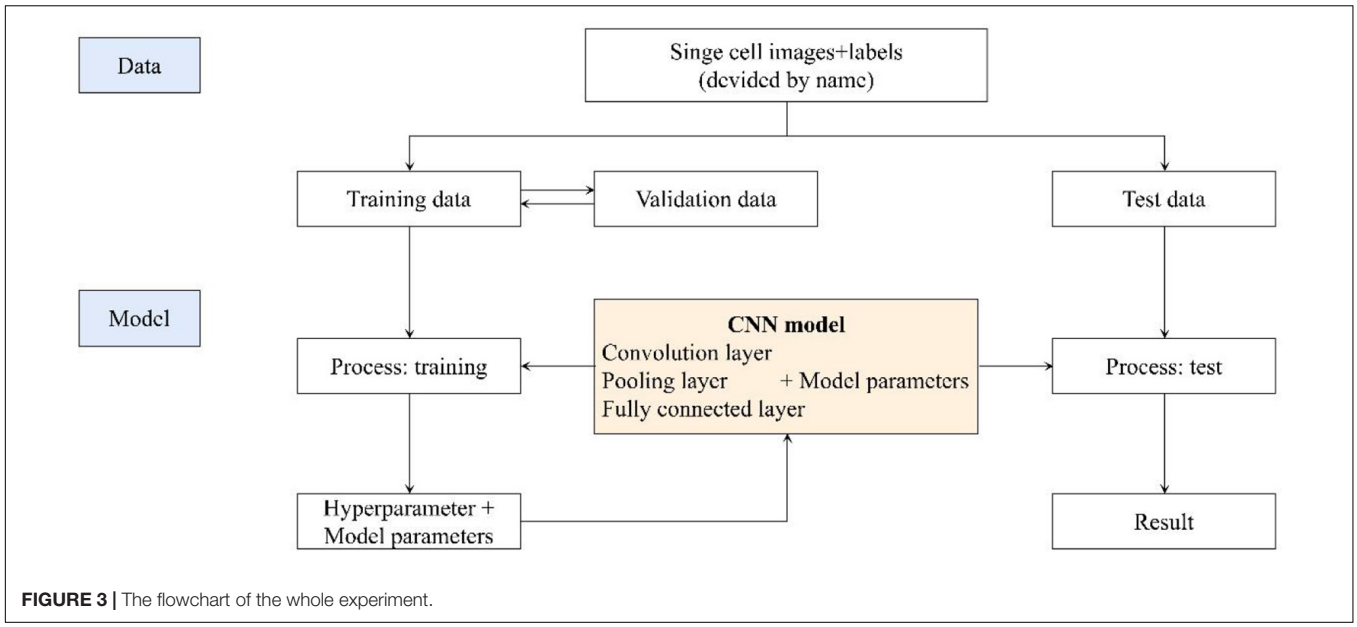


FIGURE 3 | The flowchart of the whole experiment.

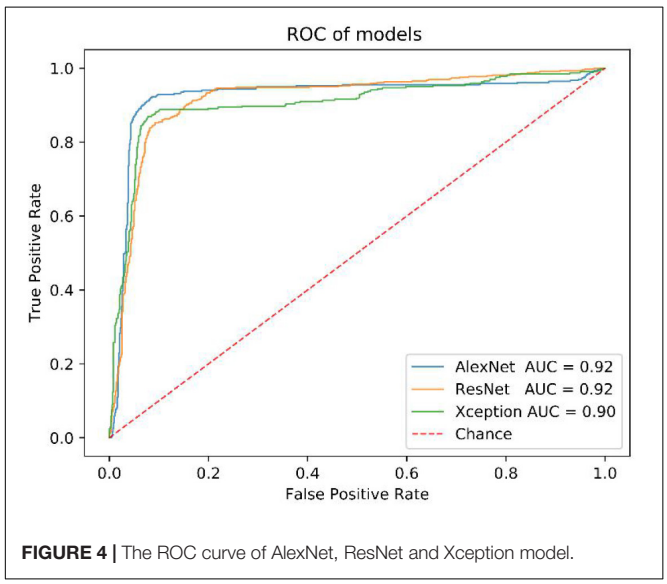


FIGURE 4 | The ROC curve of AlexNet, ResNet and Xception model.

picture just focused on a certain CTC-positive cell under the microscope, which results in that the machine learning method has no advantage in recognition speed compared with the traditional image recognition method. Enlarging the scope of images and collected more samples is also that need to be improved in the future.

Deep learning has already been shown to be suitable for detection of CTCs because of the high sensitivity and specificity in CTC counting. We had changed the filter size and number in all convolution layers in order to find the best CNN parameters. We found different filter size and number will influence the results largely. We changed filter number from range 5 to 128 in our training process. We found that the training result was not convergence when the number was less than 16. It showed that the range of the feature number of the image is about 32–128. We tried to increase the filter size from 5 to 20, but the result was not changed a lot and the convergence speed even became slower when the filter size higher than 10. From this process, we summarized that the feature size in CTCs could not be greater than 10 pixels. Furthermore, there are many appropriately AI models such as VGG, InceptionV1-4. We will apply them on the CTCs dataset to establish a more suitable model in the later testing.

Circulating tumor cell is an important marker for early screening and prognosis of tumors. In addition, CTCs, originating from the primary tumor, may be more effective for tumor tissue tracing and molecular classification. Image recognition can only obtain the characteristics of the cell surface. If strict tissue tracing is required, other molecular biological experimental data such as the isolation of CTC cells and single cell sequencing may be required. Besides, in this study, we also evaluated the performance of AlexNet model in variant types of cancers. **Supplementary Table S2** and **Figure S2** showed that our model presents a better performance in Lung cancer than Gastrointestinal cancer and Breast cancer. One of the reasons may be that the training data size of Lung cancer (158) is much larger than those of Gastrointestinal cancer (45) and Breast cancer (70). Further, postoperative recurrence may occur in approximately 45% of patients, even after complete resection of NSCLC (Yano et al., 2014). These proteins, especially epithelial proteins, such as EpCAM, PIK3CA, AKT2, TWIST, and ALDH1, may have more activities (Hanssen et al., 2016), which will lead more influence in the morphology of cells and affecting the recognition performance thereby. Therefore, the multi-image omics, including CT images, HE staining, and immunohistochemical images, as well as the sequencing data, may be urgently needed at this stage.

CONCLUSION

In the present study, we established a CTC cell recognition software based on deep learning. In order to make it more practical, we collected samples from the real world, instead of using the public databases. We performed the CTC enrichment and imFISH experiments and screened the fluorescence images according to the figure's quality. In order to improve the efficiency, we used the machine instead of doing manual screening. First, the python's package was used to do image segmentation. The obtained recognition sensitivity and specificity are 93.7 and 92.1%, respectively. In addition, the recognition sensitivity and specificity can also reach to 90.3 and 91.3%, respectively using CNN instead of manual intervention. In the future studies, we will focus on the improvement of the accuracy and sensitivity with a more suitable deep learning model, promoting this technology to the clinic as soon as possible.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics Committee of Chifeng Municipal Hospital. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

GT, YL, BH, and QZ conceived the concept of the work. BH, QL, JL, PB, HY, and SL performed the experiments. QL and BH wrote the manuscript. CP and HY reviewed the manuscript. All authors approved the final version of this manuscript.

FUNDING

This research was funded by Hunan Provincial Innovation Platform and Talents Program (No. 2018RS3105), the Natural Science Foundation of China (No. 61803151), the Natural Science Foundation of Hunan Province (No. 2018JJ3570), and the Project of Scientific Research Fund of Hunan Provincial Education Department (Nos. 19A060 and 19C0185).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00897/full#supplementary-material

REFERENCES

Anand, K., and Roszik, J. (2019). Pilot study of circulating tumor cells in early-stage and metastatic uveal melanoma. *Cancers (Basel)* 11:856. doi: 10.3390/cancers11060856

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans. Med. Imaging* 35, 1207–1216. doi: 10.1109/tmi.2016.2535865

Asante, D. B., Calapre, L., Ziman, M., Meniawy, T. M., and Gray, E. S. (2020). Liquid biopsy in ovarian cancer using circulating tumor DNA and cells: ready for prime time? *Cancer Lett.* 468, 59–71. doi: 10.1016/j.canlet.2019.10.014

Baek, D. H., Kim, G. H., Song, G. A., Han, I. S., Park, E. Y., Kim, H. S., et al. (2019). Clinical potential of circulating tumor cells in colorectal cancer: a prospective study. *Clin. Transl. Gastroenterol.* 10:e00055. doi: 10.14309/ctg.0000000000000055

Banys-Paluchowski, M., Schneck, H., Blassl, C., Schultz, S., Meier-Stiegen, F., Niederacher, D., et al. (2015). Prognostic relevance of circulating tumor cells in molecular subtypes of breast cancer. *Geburtshilfe Frauenheilkd* 75, 232–237. doi: 10.1055/s-0035-1545788

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge university press.

Cristofanilli, M., Pierga, J. Y., Reuben, J., Rademaker, A., Davis, A. A., Peeters, D. J., et al. (2019). The clinical use of circulating tumor cells (CTCs) enumeration for staging of metastatic breast cancer (MBC): International expert consensus paper. *Crit. Rev. Oncol. Hematol.* 134, 39–45. doi: 10.1016/j.critrevonc.2018.12.004

Dominguez, C., Heras, J., and Pascual, V. (2017). IJ-OpenCV: combining ImageJ and OpenCV for processing images in biomedicine. *Comput. Biol. Med.* 84, 189–194. doi: 10.1016/j.compbiomed.2017.03.027

Dong, J., Zhu, D., Tang, X., Qiu, X., Lu, D., Li, B., et al. (2019). Detection of circulating tumor cell molecular subtype in pulmonary vein predicting prognosis of stage I-III non-small cell lung cancer patients. *Front. Oncol.* 9:1139. doi: 10.3389/fonc.2019.01139

Erickson, B. J., Korfiatis, P., Akkus, Z., and Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics* 37, 505–515. doi: 10.1148/rg.2017160130

Ferreira, M. M., Ramani, V. C., and Jeffrey, S. S. (2016). Circulating tumor cell technologies. *Mol. Oncol.* 10, 374–394. doi: 10.1016/j.molonc.2016.01.007

Gabriel, M. T., Calleja, L. R., Chalopin, A., Ory, B., and Heymann, D. (2016). Circulating tumor cells: a review of non-EpCAM-based approaches for cell enrichment and isolation. *Clin. Chem.* 62, 571–581. doi: 10.1373/clinchem.2015.249706

Grover, P. K., Cummins, A. G., Price, T. J., Roberts-Thomson, I. C., and Hardingham, J. E. (2014). Circulating tumour cells: the evolving concept and the inadequacy of their enrichment by EpCAM-based methodology for basic and clinical cancer research. *Ann. Oncol.* 25, 1506–1516. doi: 10.1093/annonc/mdu018

Guibert, N., Delaunay, M., Lusque, A., Boubekeur, N., Rouquette, I., Clermont, E., et al. (2018). PD-L1 expression in circulating tumor cells of advanced non-small cell lung cancer patients treated with nivolumab. *Lung Cancer* 120, 108–112. doi: 10.1016/j.lungcan.2018.04.001

Hanssen, A., Wagner, J., Gorges, T. M., Taenzer, A., Uzunoglu, F. G., Driemel, C., et al. (2016). Characterization of different CTC subpopulations in non-small cell lung cancer. *Sci. Rep.* 6:28010. doi: 10.1038/srep28010

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi: 10.1016/0893-6080(89)90020-8

Janning, M., Kobus, F., Babayan, A., and Wikman, H. (2019). Determination of PD-L1 expression in circulating tumor cells of NSCLC patients and correlation with response to PD-1/PD-L1 inhibitors. *Cancers (Basel)* 11:835. doi: 10.3390/cancers11060835

Keller, L., Werner, S., and Pantel, K. (2019). Biology and clinical relevance of EpCAM. *Cell Stress* 3, 165–180. doi: 10.15698/cst2019.06.188

Kloten, V., Lampignano, R., Krahn, T., and Schlange, T. (2019). Circulating tumor Cell PD-L1 expression as biomarker for therapeutic efficacy of immune checkpoint inhibition in NSCLC. *Cells* 8:809. doi: 10.3390/cells8080809

Koudelakova, V., Trojanec, R., Vrbkova, J., Donevska, S., Bouchalova, K., Kolar, Z., et al. (2016). Frequency of chromosome 17 polysomy in relation to CEP17 copy number in a large breast cancer cohort. *Genes Chromosomes Cancer* 55, 409–417. doi: 10.1002/gcc.22337

Kraeft, S. K., Ladanyi, A., Galiger, K., Herlitz, A., Sher, A. C., Bergsrud, D. E., et al. (2004). Reliable and sensitive identification of occult tumor cells using the improved rare event imaging system. *Clin. Cancer Res.* 10, 3020–3028. doi: 10.1158/1078-0432.ccr-03-0361

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). "ImageNet classification with deep convolutional neural networks," in *Paper Presented at the NIPS*, (Lake Tahoe: Harrahs and Harveys).

Le, N. Q., Ho, Q. T., and Ou, Y. Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Comput. Chem.* 38, 2000–2006. doi: 10.1002/jcc.24842

Le, N. Q., Ho, Q. T., and Ou, Y. Y. (2018). Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. *Anal. Biochem.* 555, 33–41. doi: 10.1016/j.ab.2018.06.011

Le, N. Q. K., Huynh, T. T., Yapp, E. K. Y., and Yeh, H. Y. (2019). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. *Comput. Methods Programs Biomed.* 177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Lee, K., Kim, H. J., Jang, M. H., Lee, S., Ahn, S., and Park, S. Y. (2019). Centromere 17 copy number gain reflects chromosomal instability in breast cancer. *Sci. Rep.* 9:17968. doi: 10.1038/s41598-019-54471-w

Ligthart, S. T., Coumans, F. A. W., Attard, G., Cassidy, A. M., de Bono, J. S., and Terstappen, L. W. M. M. (2011). Unbiased and automated identification of a circulating tumour cell definition that associates with overall survival. *PLoS One* 6:e27419. doi: 10.1371/journal.pone.0027419

Lindsay, C. R., Faugeroux, V., Michiels, S., Pailler, E., Facchinetti, F., Ou, D., et al. (2017). A prospective examination of circulating tumor cell profiles in non-small-cell lung cancer molecular subgroups. *Ann. Oncol.* 28, 1523–1531. doi: 10.1093/annonc/mdx156

Liu, X., Zhang, Z., Zhang, B., Zheng, Y., Zheng, C., Liu, B., et al. (2018). Circulating tumor cells detection in neuroblastoma patients by EpCAM-independent enrichment and immunostaining-fluorescence in situ hybridization. *EBioMedicine* 35, 244–250. doi: 10.1016/j.ebiom.2018.08.005

Lowd, D., and Domingos, P. (2005). "Naive Bayes models for probability estimation," in *Proceedings of the 22nd International Conference on Machine Learning*, (New York, NY: Association for Computing Machinery).

Lu, S. S., Pan, Q. J., Cao, J., Xu, X., Zhao, H., and Shen, D. H. (2017). Fluorescence in situ hybridization combined with cytomorphology for the detection of lung cancer in bronchial brushing specimens. *Zhonghua Zhong Liu Za Zhi* 39, 595–599. doi: 10.3760/cma.j.issn.0253-3766.2017.08.007

Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Z Med. Phys.* 29, 102–127. doi: 10.1016/j.zemedi.2018.11.002

Maier, A., Syben, C., Lasser, T., and Riess, C. (2019). A gentle introduction to deep learning in medical image processing. *Z Med. Phys.* 29, 86–101. doi: 10.1016/j.zemedi.2018.12.003

Maly, V., Maly, O., Kolostova, K., and Bobek, V. (2019). Circulating tumor cells in diagnosis and treatment of lung cancer. *In Vivo* 33, 1027–1037. doi: 10.21873/invivo.11571

Manjunath, Y., Upparahalli, S. V., Avella, D. M., and Deroche, C. B. (2019). PD-L1 expression with epithelial mesenchymal transition of circulating tumor cells is associated with poor survival in curatively resected non-small cell lung cancer. *Cancers (Basel)* 11:806. doi: 10.3390/cancers11060806

Marcuello, M., Vymetalkova, V., Neves, R. P. L., Duran-Sanchon, S., Vedeld, H. M., Tham, E., et al. (2019). Circulating biomarkers for early detection and clinical management of colorectal cancer. *Mol. Aspects Med.* 69, 107–122. doi: 10.1016/j.mam.2019.06.002

McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., et al. (2018). Deep learning in radiology. *Acad. Radiol.* 25, 1472–1480. doi: 10.1016/j.acra.2018.02.018

Merker, J. D., Oxnard, G. R., Compton, C., Diehn, M., Hurley, P., Lazar, A. J., et al. (2018). Circulating tumor DNA analysis in patients with cancer: american society of clinical oncology and college of American pathologists joint review. *J. Clin. Oncol.* 36, 1631–1641. doi: 10.1200/jco.2017.76.8671

Nagrath, S., Sequist, L. V., Maheswaran, S., Bell, D. W., Irimia, D., Ulkus, L., et al. (2007). Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* 450, 1235–1239. doi: 10.1038/nature06385

Paget, S. (1989). The distribution of secondary growths in cancer of the breast. 1889. *Cancer Metastasis Rev.* 8, 98–101.

Pan, L., Yan, G., Chen, W., Sun, L., Wang, J., and Yang, J. (2019). Distribution of circulating tumor cell phenotype in early cervical cancer. *Cancer Manag. Res.* 11, 5531–5536. doi: 10.2147/cmar.s198391

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164. doi: 10.1038/s41551-018-0195-0

Praharaj, P. P., Bhutia, S. K., Nagrath, S., Bitting, R. L., and Deep, G. (2018). Circulating tumor cell-derived organoids: current challenges and promises in medical research and precision medicine. *Biochim. Biophys. Acta Rev. Cancer* 1869, 117–127. doi: 10.1016/j.bbcan.2017.12.005

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106.

Racila, E., Euhus, D., Weiss, A. J., Rao, C., McConnell, J., Terstappen, L. W., et al. (1998). Detection and characterization of carcinoma cells in the blood. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4589–4594. doi: 10.1073/pnas.95.8.4589

Riebensahm, C., Joosse, S. A., Mohme, M., Hanssen, A., Matschke, J., Goy, Y., et al. (2019). Clonality of circulating tumor cells in breast cancer brain metastasis patients. *Breast Cancer Res.* 21:101. doi: 10.1186/s13058-019-1184-2

Stefanovic, S., Deutsch, T. M., Wirtz, R., Hartkopf, A., and Sinn, P. (2019). Molecular subtype conversion between primary and metastatic breast cancer corresponding to the dynamics of apoptotic and intact circulating tumor cells. *Cancers (Basel)* 11:342. doi: 10.3390/cancers11030342

Wainberg, M., Merico, D., Delong, A., and Frey, B. J. (2018). Deep learning in biomedicine. *Nat. Biotechnol.* 36, 829–838. doi: 10.1038/nbt.4233

Wan, J. C. M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J. D., Caldas, C., et al. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* 17, 223–238. doi: 10.1038/nrc.2017.7

Xie, J., Liu, R., Luttrell, J., and Zhang, C. (2019). Deep learning based analysis of histopathological images of breast cancer. *Front. Genet.* 10:80. doi: 10.3389/fgene.2019.00080

Xing, F., Xie, Y., and Yang, L. (2015). An automatic learning-based framework for robust nucleus segmentation. *IEEE Trans. Med. Imaging* 35, 550–566. doi: 10.1109/tmi.2015.2481436

Yang, C., Zhang, N., Wang, S., Shi, D., Zhang, C., Liu, K., et al. (2018). Wedge-shaped microfluidic chip for circulating tumor cells isolation and its clinical significance in gastric cancer. *J. Transl. Med.* 16:139. doi: 10.1186/s12967-018-1521-8

Yano, T., Okamoto, T., Fukuyama, S., and Maehara, Y. (2014). Therapeutic strategy for postoperative recurrence in patients with non-small cell lung cancer. *World J. Clin. Oncol.* 5, 1048–1054. doi: 10.5306/wjco.v5.i5.1048

Zhou, C. Y., and Chen, Y. Q. (2006). Improving nearest neighbor classification with cam weighted distance. *Pattern Recognition* 39, 635–645. doi: 10.1016/j.patcog.2005.09.004

Zhou, M., Zheng, H., Wang, Z., Li, R., Liu, X., Zhang, W., et al. (2017). Precisely enumerating circulating tumor cells utilizing a multi-functional microfluidic chip and unique image interpretation algorithm. *Theranostics* 7, 4710–4721. doi: 10.7150/thno.20440

Zou, J., Huss, M., Abid, A., Mohammadi, P., and Torkamani, A. (2019). A primer on deep learning in genomics. *Nat. Genet.* 51, 12–18. doi: 10.1038/s41588-018-0295-5

Conflict of Interest: QL, JL, HY, CP, YL, and GT were employed by the company Geneis (Beijing) Co., Ltd.

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 August 2020
doi: 10.3389/fgene.2020.00768

DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images

Zhijun Wu[1†], Lin Wang[2†], Churong Li[3], Yongcong Cai[4], Yuebin Liang[5], Xiaofei Mo[5], Qingqing Lu[5], Lixin Dong[6] and Yonggang Liu[7*]*

[1] Department of Oncology, The First People's Hospital of Changde City, Changde, China, [2] Department of Oncology, Hainan General Hospital, Haikou, China, [3] Sichuan Cancer Hospital and Institute, The Affiliated Cancer Hospital, School of Medicine, UESTC, Chengdu, China, [4] Sichuan Cancer Hospital, Chengdu, China, [5] Geneis (Beijing) Co., Ltd., Beijing, China, [6] The First Hospital of Qinhuangdao, Qinhuangdao, China, [7] Baotou Cancer Hospital, Baotou, China

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Tao Huang,
Shanghai Institute for Biological Sciences (CAS), China
Khanh N. Q. Le,
Taipei Medical University, Taiwan

***Correspondence:**
Lixin Dong
Donglixin6447@sina.com
Yonggang Liu
Zlnkldf8000@sina.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: *27 March 2020*
Accepted: *29 June 2020*
Published: *25 August 2020*

Citation:
Wu Z, Wang L, Li C, Cai Y, Liang Y, Mo X, Lu Q, Dong L and Liu Y (2020) DeepLRHE: A Deep Convolutional Neural Network Framework to Evaluate the Risk of Lung Cancer Recurrence and Metastasis From Histopathology Images. Front. Genet. 11:768. doi: 10.3389/fgene.2020.00768

It is critical for patients who cannot undergo eradicable surgery to predict the risk of lung cancer recurrence and metastasis; therefore, the physicians can design the appropriate adjuvant therapy plan. However, traditional circulating tumor cell (CTC) detection or next-generation sequencing (NGS)-based methods are usually expensive and time-inefficient, which urge the need for more efficient computational models. In this study, we have established a convolutional neural network (CNN) framework called DeepLRHE to predict the recurrence risk of lung cancer by analyzing histopathological images of patients. The steps for using DeepLRHE include automatic tumor region identification, image normalization, biomarker identification, and sample classification. In practice, we used 110 lung cancer samples downloaded from The Cancer Genome Atlas (TCGA) database to train and validate our CNN model and 101 samples as independent test dataset. The area under the receiver operating characteristic (ROC) curve (AUC) for test dataset was 0.79, suggesting a relatively good prediction performance. Our study demonstrates that the features extracted from histopathological images could be well used to predict lung cancer recurrence after surgical resection and help classify patients who should receive additional adjuvant therapy.

Keywords: lung cancer, recurrence, hematoxylin and eosin staining, histopathological image, convolutional neural network

INTRODUCTION

Lung cancer accounts for 13% of newly diagnosed cancer incidences worldwide, resulting in 1.4 million deaths annually (Travis et al., 2011). According to the American Joint Committee on Cancer (AJCC), the TNM staging system is widely used for describing the anatomical extent of the disease on the basis of the assessment of three components: the extent of the primary tumor (T), presence and extent of regional lymph node metastasis (N), or presence of distant metastasis (M). The current TNM staging system is relatively accurate in defining the tumor stage. The recurrence rates of lung cancer patients in TNM stages I, II, and III are 34, 55, and 74%, respectively.

As is known to us all, the first-line treatment plan for a cancer patient is surgical removal of the primary tumor if there is no metastasis. However, the 5-year survival rate of postsurgical patients with early-stage lung cancer is only 54%, which is significantly worse than that of patients with breast cancer (∼90%) (Kaplan et al., 2016; Meng et al., 2018; Sun et al., 2019). One key factor leading to the poor postsurgical outcome for lung cancer patients is the loss of pulmonary function. Lobectomy leads to the loss or compromise of limited pulmonary function. On the other hand, wedge resections, which largely depend on the surgical resection margin, can save lung parenchyma but are associated with a nearly twofold increase in local cancer recurrence. It is crucial to decide the type of surgery to be performed because the 2-year survival rate in patients with local recurrence will drop to about 20% (Hung et al., 2009).

To alleviate the risk of surgical-related recurrence risk and increase the survival rate of postsurgical lung cancer patients, some invasive or non-invasive techniques have been used in clinical practice. First, the detection of circulating tumor cells (CTCs) at the time of surgery may represent an approach for identifying patients at a high risk of recurrence. A recent study indicated that the detection of pulmonary venous CTCs (PV-CTCs) at surgical resection could be used to evaluate future relapse (Chemi et al., 2019). Second, a few types of genomic alterations could be utilized to evaluate the risk of lung cancer recurrence owing to the strong association between genetic instability and tumorigenesis (Chan and Hughes, 2015). Next-generation sequencing (NGS) has a better testing performance with compatibility of low-input DNA. The National Comprehensive Cancer Network guideline of non-small-cell lung cancer recommended biomarkers favorable for target therapies such as epidermal growth factor receptor (EGFR) mutation (Couraud et al., 2014; Xu et al., 2016). Plasma and urine EGFR mutation levels could be used to predict the response of chemotherapy (Reckamp et al., 2016). NGS-based liquid biopsy is complemented with traditional tissue biopsy, which might be a promising strategy in the molecular profiling of lung cancer in the future. Furthermore, circulating tumor DNA and tissue assay might be combined to better predict lung cancer recurrence (Reckamp et al., 2016).

Since the rapid rise in the incidence and mortality of lung cancer, many researchers have shifted their focus on advanced discovery of novel diagnostic approach and predictive markers of metastasis, therefore, to assist clinical professionals to design individualized therapy for patients. Cancer recurrence following surgery or chemotherapy for lung cancer is a significant failure of local treatment as well as reduces the patient outcomes. Currently, cancer immunotherapy has been applied to cancer therapy. It has been recognized as adjuvant therapy for patients do not qualify for surgical intervention. The novel approach can be used to identify driver genes and predictive genes. For example, we have explained some lung cancer-specific gene mutation, gene sequencing, and biomarkers. PD-1 is an antibody against program death receptor and has been approved for second-line therapy of squamous cell carcinomas (Beer et al., 2002). Moreover, many preclinical trials demonstrated that combined traditional strategy and novel gene therapy may have improved patient overcome (Weiner et al., 2012).

Compared with other techniques, visual inspection of histologically stained slices is considered standard and used by pathologists to evaluate tumor stage, subtype, metastatic location, and prognosis (Fischer et al., 2008). With the absence of definitive pathological features, microscopic assessment requires experienced pathologists to evaluate stained slices. This process could be quite challenging and time-consuming for pathologists, and the results also depend on the quality of hematoxylin and eosin (H&E)-stained slices. Furthermore, accurate interpretation of an H&E image could be difficult because the distinction among different types of lung cancer is relatively unclear (MacConaill, 2013). To assist pathologists, deep learning tools have been developed to interpret the whole-slide image (WSI), which is helpful for developing an appropriate treatment plan and predicting survival outcomes. Yu et al. combined conventional image processing techniques with machine learning algorithms such as random forest, support vector machine, and naïve Bayes classifier to achieve acceptable prediction accuracy for lung cancer subtypes (Yu et al., 2016). The area under the receiver operating characteristic (ROC) curve (AUC) was approximately 0.75 in distinguishing two subtypes of lung cancer (Blumenthal et al., 2018). Furthermore, deep learning has also been successfully applied to the subtype classification of multiple cancers such as breast cancer, bladder cancer, and lung cancer (Zachara-Szzakowki et al., 2015; Araujo et al., 2017). The AUC reached approximately 0.83 by using The Cancer Genome Atlas (TCGA) dataset (Zachara-Szzakowki et al., 2015). Convolutional neural network (CNN) approach is not only used in cancer field, but it has been used in biochemical field as well. CNN has also served as a powerful approach to identify specific proteins located in electron transport chain, achieving good sensitivity (0.83%), specificity (94.4%), and accuracy (92.3%). This study demonstrated that the CNN approach can also be used in understanding the biochemical mechanism of important proteins such as electronic (Le et al., 2017, 2019). The same study team also used CNN to identify fertility related protein, which also received good sensitivity, specificity, and accuracy. Fertility-related proteins have critical function in reproductive organs and hormone-related fertility (Le, 2019). In fact, deep learning-based annotations of medical images are now close to, if not better than, those of pathologists for many types of cancers at present. With the development of image segmentation techniques (Simon et al., 2018), the WSI has been widely used for nuclei identification, tissue segmentation, and epithelial tissue identification in several cancers such as renal cancer, bladder cancer, and breast cancer (de Bel et al., 2018).

In this study, we established a novel machine learning framework to predict lung cancer recurrence by using the H&E-stained histopathological images. We first patched the H&E WSI into images of the size 512 × 512 pixels, which were then subject to a few image preprocessing steps such as image quality control and normalization. We then established a lung cancer tumor region prediction model and a cancer recurrence prediction model on the basis of the patched images. The prediction results based on patched images of a WSI were then combined to

evaluate the recurrence risk of a lung cancer patient. Our model is cost-effective and could meet large clinical demands.

MATERIALS AND METHODS

Data Preparation

Hematoxylin and eosin images and clinical data of lung cancer were downloaded from TCGA database[1], which is a landmark cancer genomics program that characterized thousands of primary cancers and matched normal samples spanning many cancer types. The labels that matched H&E images downloaded from TCGA contained information about metastasis and recurrence, and H&E image with SVS format was analyzed by the Python package OpenSlide. H&E images from those patients with the risk of metastasis and recurrence were labeled as "1" and "0" for those without metastasis and recurrence (**Figure 1A**).

[1]https://portal.gdc.cancer.gov/repository/

Image Preprocessing

To predict cancer recurrence and metastasis, tumor regions were annotated with the help of an expert pathologist by visual assessment. The morphology, color, and size of the nucleus of tumor cells are shown inside of tumor region, with the blue solid dotted lines representing the boundary of tumor (**Figure 1B**). For image preprocessing, each WSI was divided into computationally memory-affordable tiles of 512 × 512 pixels as input dataset. For noise reduction, Python's OpenCV (version 4.1.1) package was applied to remove blank or blurred spaces in tumor region and to help reduce non-association interference in model training process. The non-association region was calculated as the ratio of the blank area or blurred spaces to the total area. The defined threshold of ratio was used to remove false-positive structures by definitive cutoff threshold. Further analysis of segmentation of H&E slice was performed by image de-noising, filtering, edge detection, expansion, and contraction techniques with OpenCV package (**Figure 2A**).

The performance of the computational technique for H&E-stained tissue image analysis is compromised by variable

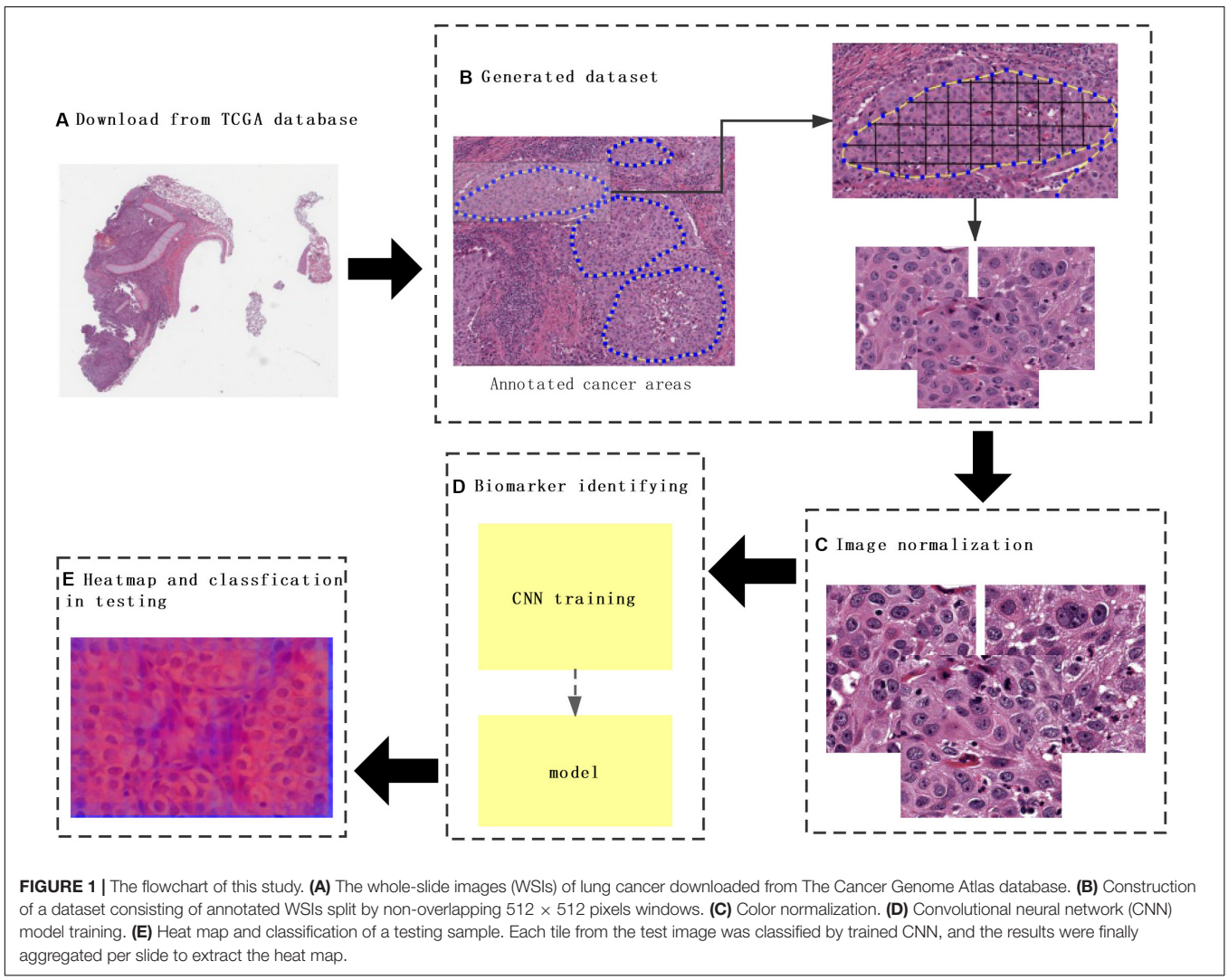


FIGURE 1 | The flowchart of this study. **(A)** The whole-slide images (WSIs) of lung cancer downloaded from The Cancer Genome Atlas database. **(B)** Construction of a dataset consisting of annotated WSIs split by non-overlapping 512 × 512 pixels windows. **(C)** Color normalization. **(D)** Convolutional neural network (CNN) model training. **(E)** Heat map and classification of a testing sample. Each tile from the test image was classified by trained CNN, and the results were finally aggregated per slide to extract the heat map.

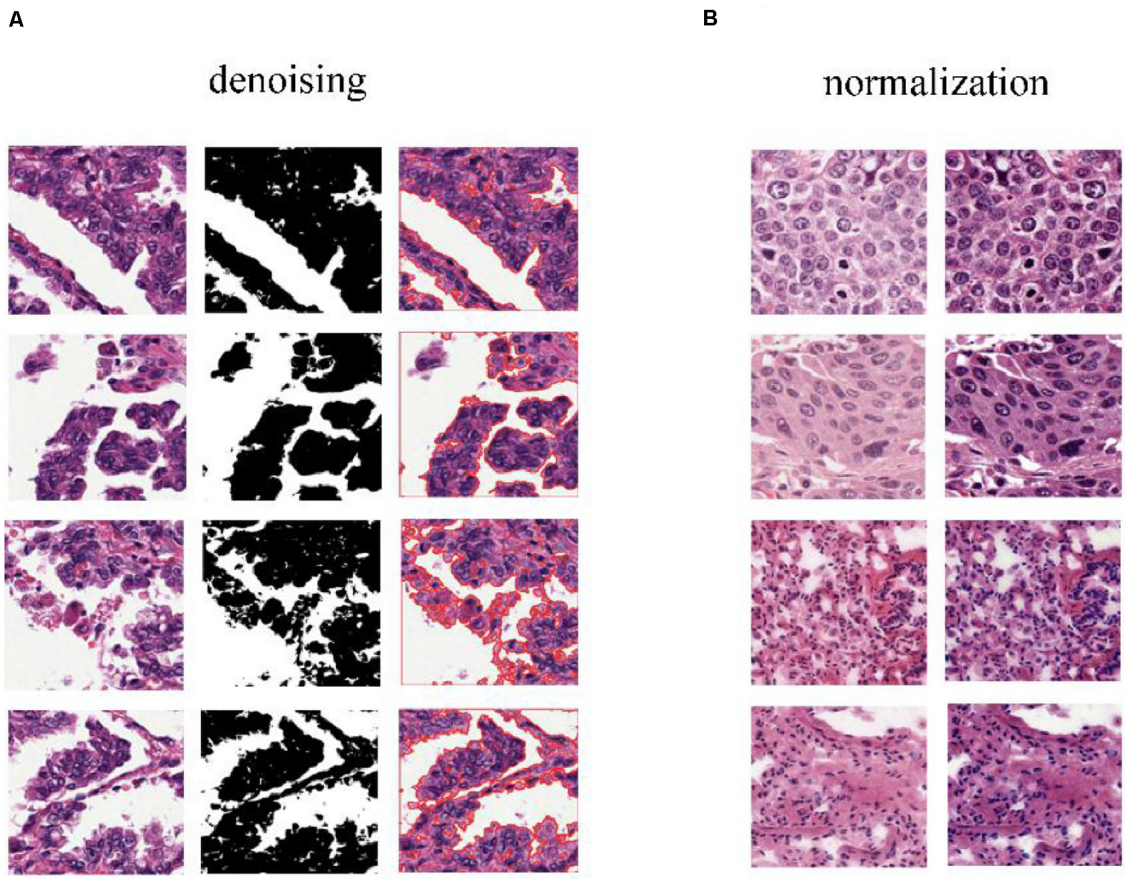


FIGURE 2 | Color normalization of H&E slices. **(A)** The de-noising process applied to regions that have large blank spaces in the tumor regions. **(B)** The deep convolutional Gaussian mixture model (DCGMM) used for color normalization. The left column represents original images, and the right column represents imaging after color normalization.

image colors due to H&E reagent concentration, staining process, and absorption caused by tissue fixation and staining method. To remove potential influenced variables, multiple color normalization (CN) approaches have been established (Vahadane et al., 2016), and unsupervised generative neural networks were applied in our study for performing stain-CN based on deep convolutional Gaussian mixture models (DCGMMs) in the stained H&E images (Shen et al., 2017; Wang et al., 2017; Qaiser et al., 2018; Simon et al., 2018). The DCGMM represents parameters of a fully CNN that are combined with the GMM parameters to optimize CN (**Figure 2B**).

The Convolutional Neural Network + ResNet Model

In our study, Tensorflow 2.0.0 package was applied to conduct our model. To be specific, CNN was used effectively to identify tumor diagnoses by analyzing H&E-stained slices. CNNs are the most popular deep learning models for processing color images. The CNN deep learning network includes the input layer, intermediate hidden layer, and output layer. The intermediate hidden layer consists of multiple convolutional layers and pooling layers followed by more fully connected layers. The CNN could adapt and extract the feature hierarchy and classify images by error back propagation, which is a relatively effective gradient descent algorithm to update the weights connecting its inputs to the outputs during the training process.

After being transformed from the input layer, the image data were trained sequentially into the convolution layer composed of 32 $n \times n$ convolution kernels (e.g., $n = 5$) and the pooling layer for dimensional reduction through the ReLU excitation layer. The data were output to complete the entire feature extraction process afterward. Then, the data entered the second and third intermediate hidden layers, respectively. After the entire process was completed, all the features were extracted completely.

Batch normalization layer was then applied with the CNN to improve the generalization ability of the network and to expedite the training for higher learning rate. Increasing the number of layers of a deep CNN after reaching a certain depth could not improve the classification performance further, resulting in slower network convergence and worse classification accuracy due to the disappearance gradient problem.

ResNet was introduced to deal with this problem. The difference between residual and ordinary networks is the

introduction of jump connection that can make the information of the previous residual block flow into the next one unimpeded, improve the information flow, and also avoid the disappearance gradient problem and the degradation caused by over depth of the network. Suppose there is a large neural network called big NN and its input is x and its output activation value is A[l]. After increasing the depth of the network, adding two additional layers to the network, and receiving the final output as A[l + 2], these two layers could be regarded as a residual block with a shortcut connection, and the activation function used in the whole network is relu. The function of relu is written by g(x). Linear function is written by W * A + B. We can get A[l + 2] = g(Z[L + 2] + A[l]), where Z[L + 2] = W[l + 2] * A[l + 1] + B[l + 1]. If W[l + 2] = 0, B[l + 1] = 0, we can know A[l + 2] = g(A[I]) then. When A[l] ≥ 0, A[l + 2] = A[l]. This is equivalent to establishing the linear relationship between A[l] and A[l + 2], when W and B is 0. It is equivalent to neglecting the two neural layers behind A[l] and realizing the linear transfer of the interlayer. The model itself can tolerate the deeper network, and this extra residual block will not affect its performance, and the relations are shown in **Figure 3**.

In fact, the residual network is composed of several shallow networks, and a shallow network could avoid the appearance of the vanishing gradient problem during training, thus accelerating the convergence of the network.

Heat Map Generation

The probability maps were generated from the tumor region for high metastasis score detection (**Figure 1E**). The color in the probability map as shown in **Figure 4B** indicates the predicted metastasis score by pixels in the tumor region. The red color represents a high score, and blue color indicates a low score. H&E images were scanned by a 512 × 512 window in a step-wise manner, and results were obtained by the CNN model at each window. We applied the results on the pixels that were included in the window. We summed up all the values that pass the pixel and determined their average value, which was the predicted metastasis score of the pixel. The probability to recurrence and metastasis of every pixel was turned into color value with clear probability visualization. The probability value was mapped in the range of (0, 1) to RGB color from pure blue color (0, 0, 255) to pure red color (255, 0, 0) linearly. As a result, the red pixel image represents as a lower risk of metastasis; meanwhile, the light blue pixels represent no risk of metastasis, as shown in **Figure 4B**.

The WSI was divided into tiles, and each tile gets a probability result by model prediction during window sliding. Results of all tiles were integrated by fusion algorithm and computed as the final probability results for a specific slide. The average probability of top *n* windows was defined as identification score. Identification score predicted the risk of recurrence and metastasis with specific cutoff threshold. Scores higher than threshold were interpreted as positive results, whereas the top number value served as a hyper-parameter and is decided by cross-validation.

Hyper-Parameter Tuning by Cross-Validation

A fivefold cross-validation was applied to prevent overfitting and to select hyper-parameters of the model for selecting the hyper-parameter space with best cross-validation score. The hyper-parameters that we tried to use in our model are activation function, patch, and top number. Our workflow is shown in the following three steps:

(1) Defining a grid on three dimensions with each of these maps for a hyper-parameter; for example, *n* = (activation function, patch, top number).
(2) For each dimension, defining the range of possible values.

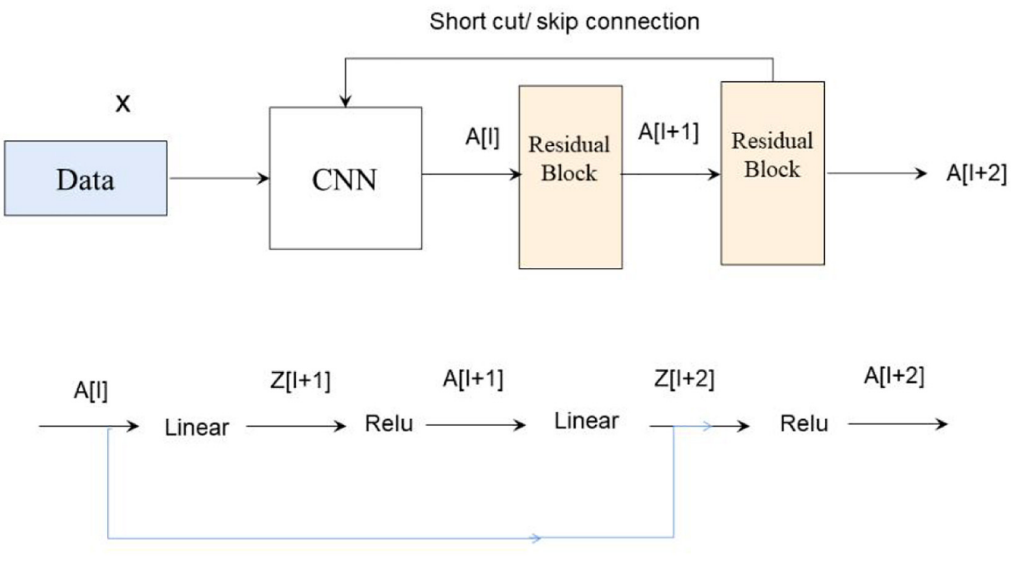


FIGURE 3 | The ResNet network workflow.

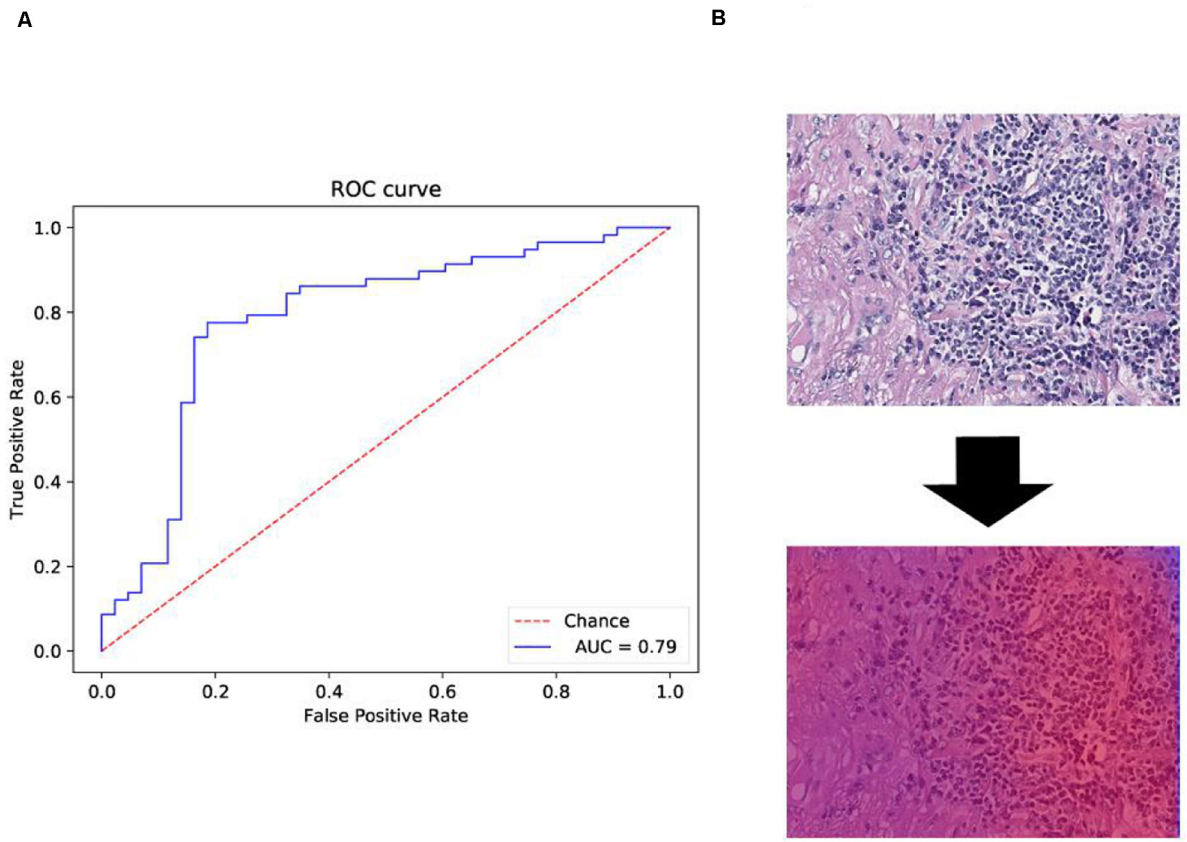


FIGURE 4 | Receiver operating characteristic (ROC) and heat map on The Cancer Genome Atlas (TCGA) training data. **(A)** ROC curve of test data with the 512 × 512 pixel image. **(B)** Heat map of the tumor region applied in the convolutional neural network (CNN) model by using TCGA dataset. We also obtained the heat map given by the model shown in **B**. From the heat map, we found that the color of suspected tumor area was red and that the color of normal area was partial blue. The results were consistent as we have considered.

(3) Searching for all the possible configurations and waiting for the results to establish the best one.

Performance Evaluation Criteria

Several well-established performance evaluation criteria were employed to evaluate the performance of the classification model, including sensitivity (Se) or recall, specificity (Sp), precision, and the AUC.

$$\mathrm{Se} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

In the equations, TP stands for the number of images correctly recognized as positive samples. FP stands for the number of images that were incorrectly recognized as positive samples. FN stands for the number of images incorrectly recognized as negative samples. TN stands for the number of images correctly recognized as negative samples. We indicate TP, FP, TN, and TP by confusion matrix as shown in **Table 1**.

RESULTS

Clinical Characteristics of Training Dataset

A total of 110 H&E images of lung cancer patients with metastasis or recurrence information were downloaded from TCGA, and the available datasets were selected with required condition with data type of slide image, data format of SVS, primary site for

TABLE 1 | Confusion matrix definitions.

Confusion matrix		Prediction	
		Positive	Negative
True	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

bronchus and lung, and white ethnicity (**Table 2**). The average age of the selected patient cohort was 54 years, and 68% of the patients have metastasis or recurrence. We labeled data to positive with new_tumor_event_type of Distant Metastasis and Locoregional Recurrence. At the same time, we labeled data to negative with tumor_status of Tumor Free.

Data Pre-treatment

The 110 H&E images with corresponding clinical data downloaded from TCGA are all in SVS format. Whole images could not be used as the input data for the network. Hence, we segmented them into tiles with a 512 × 512 pixel size from the 110 H&E images in which tumor regions were annotated in the WSIs by the expert pathologist. The tiles with a low amount of information (e.g., more than 70% of the surface was covered by background) were removed. Thereafter, a template image was selected by an expert pathologist. Then we trained the DCGMM by using this template image. After training, we applied the model on the H&E image on the upper row of the compared color normalized image down the row (**Figure 1C**). The results are shown in **Figure 2B**.

Model Construction and Hyper-Parameter Selection

DeepLRHE model was constructed with ResNet network and top five selection algorithm of WSI (**Figure 1D**). We used GridSearchCV class in scikit-learn by providing a dictionary of hyper-parameters to determine the hyper-parameters of the model. The hyper-parameters we selected are shown as follows:

$$\text{Top number} = [5, 3]$$

$$\text{Patch} = [100, 150, 200]$$

$$\text{Activation function} = [\text{softmax, relu, tanh}].$$

After the cross-validation process, activation function is set to relu, patch number is set to 150, and top number is set to 5 as shown in **Table 3**, in which performance is the best (the AUC value reached a maximum value of 0.84). We used selected hyper-parameters to construct the DeepLRHE model on whole 110 training dataset.

TABLE 2 | Clinical characteristics.

Clinical variable	Category	Clinical level
Age	Mean	54 (31–83)
Gender	Male	62
	Female	47
	Unknown	1
Samples type	H&E	1
Metastasis and recurrence period	Tumor Free	35
	Loco regional recurrence	15
	Distant metastasis	60
Cancer subtype	Adenocarcinoma	58
	Squamous carcinoma	52

TABLE 3 | Tuning of the hyper-parameters.

Activation function	Patch	Top number	
		5	3
softmax	100	0.79	0.73
	150	0.82	0.75
	200	0.82	0.78
relu	100	0.81	0.74
	150	0.84	0.78
	200	0.83	0.80
tanh	100	0.73	0.6
	150	0.76	0.67
	200	0.77	0.69

TABLE 4 | The confusion matrix of the model for test dataset.

True	Prediction		Total
	High risk	Low risk	
High risk	49	9	58
Low risk	14	29	43
Total	63	38	101

Performance Evaluation on Test Dataset

Another 101 H&E images were downloaded with their clinical information in different project from train dataset in TCGA. The datasets were available on TCGA in condition with data type of slide image, data format is SVS, and primary site of bronchus and lung and ethnicity is not reported (this condition is different from that of the training dataset). The trained model was applied on those data and obtained the confusion matrix below and ROC curve in **Figure 4A**.

The performance evaluation results were calculated from the confusion matrix in **Table 4**. The results showed that the sensitivity and specificity of the model were 0.84 and 0.67, respectively. The precision and F1 score reached 0.78 and 0.81, respectively, in the independent test dataset. In the meantime, the model achieved 0.79 AUC score. The AUC value on independent test dataset was lower compared with AUC value (0.79 vs. 0.84) of fivefold cross-validation method on train dataset. The performance evaluation from independent test dataset was more convincing.

DISCUSSION

Machine learning algorithms have been widely used in clinical practice. They can map unstructured information into a structured form as well as enable automatic identification and extraction of relevant information. Such an automated system enables us to significantly reduce time-consuming diagnostic procedures. With a dramatic improvement in the affordability of the testing, it has also brought challenges pertaining to the evaluation of effectiveness and accuracy of gene testing, which could affect diagnosis and subsequent therapy. Therefore, machine learning algorithms have been a hot topic and a

dynamically changing area in the recent years. Therefore, these models require human experts to encode the domain knowledge through feature engineering. However, the results of such models are still controversial and time dependent.

Recently, multilayer NNs or deep learning has been applied to gain insights from heterogeneous clinical data. The major difference between deep learning and conventional NN is the number of hidden layers as well as their capability to learn meaningful abstractions of the input. Deep learning has been applied to process aggregated clinical documents (imaging, pathological slices from biopsy, and other reports). Several studies have used deep learning to predict disease prognosis from medical documentation; for example, one study used a four-layer CNN to predict congestive heart failure and chronic obstructive pulmonary disease that showed promising performance. CNN is a powerful algorithm for advancing biomedical images and analysis (Makowski and Hayes, 2008; Shackelford et al., 2013). It can be applied for pathological image analysis tasks such as tumor detection and quantification of cellular features by using either general staining slices or in combination with immunohistological markers (Mogi and Kuwano, 2011; Morris et al., 2013). Computerized image processing histopathological analysis system has been impressive in the prognostic determination of various tumors and even precancerous lesions in the esophagus (Chiang et al., 2016). Recent studies showed that many histological features are associated with survival outcomes. Deep learning tumor detection allows for tumor size calculation and shape estimation. Tumor size and shape are a well-established prognostic marker for lung cancer, and the boundary of the tumor region has been reported to be associated with a poor local prognosis marker as well (Esteva et al., 2017). Furthermore, most tumor-related features including the tumor area, perimeter, convex area, and filled area of the tumor region were associated with poor survival outcome (Popin et al., 2018). Extracting tumor features from H&E were usually conducted by experienced experts; however, the extraction process is subject to human bias and is time-consuming. CNNs, as the most popular deep learning model for imaging processing, could directly handle multidimensional color image and extract the regional boundary of the pixels. Moreover, CNNs can retain parameters during imaging processing as well as effectively identify similar images.

In this study, to identify tumor regions, the pathological images were divided into 512 × 512 pixel patches to classify as tumor, non-malignant, or white categories using the CNN model. The CNN model was trained on image patches that were downloaded from TCGA database for lung squamous cell carcinoma (LUSC). Moreover, we compared the performance of our model on the test set with the performance of experienced pathologists. Our results reached an 81% AUC score. Moreover, our model has strong generalizability for learning comprehensive tissue and cell morphological changes that could be used as an auxiliary approach to make a pathological diagnosis for different types of cancers. Also, our results suggest that deep learning of histopathological imaging features can predict the prognosis of lung cancer patients, thereby assisting health professionals to make precision treatment plans.

Our study has several limitations. TCGA images exclusively composed of lung adenocarcinoma (LUAD) cells, LUSC cells, or normal lung tissues. However, several images contain features that the model has not been trained to recognize, making the classification task more challenging. For example, we observed several non-specific features including blood vessels, inframammary cell infiltration, and necrotic regions in the lung tissue as well as bronchial cartilage and fibrous scars. Moreover, this study did not include an independent set to validate our model, which may have compromised the accuracy of the results.

Overall, here, we established a novel deification model for pathological diagnoses. This model interpreted predictions through convolutional natural language and visual attention that could help pathologists to analyze histological slices. Our model could allow diagnostic consistency and establish cost-effective systems to meet large clinical demands with less manual intervention and time efficiency by analyzing precise pixels objectively. Future studies are necessary to testify its performance for other types of cancers such as gastrointestinal cancers.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

YBL and LD designed the project. ZW, LW, CL, YC, YBL, XM, and QL analyzed the data, performed the experiments, and wrote the manuscript. ZW and YGL modified and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Araujo, T., Guilherme, A., Eduardo, C., José, R., Paulo, A., Catarina, E., et al. (2017). Classification of breast cancer histology images using convolutional neural networks. *PLoS One* 12:e017754.

Beer, D. G., Sharon, L. R. K., Chiang-Ching, H., Thomas, J. G., Albert, M. L., David, E. M., et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.* 8, 816–824.

Blumenthal, G. M., Bunn, P. A. Jr., Chaft, J. E., Caroline, E. M., Edith, A. P., Giorgio, V. S., et al. (2018). Current status and future perspectives on neoadjuvant therapy in lung cancer. *J. Thorac. Oncol.* 13, 1818–1831.

Chan, B. A., and Hughes, B. G. (2015). Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. *Transl. Lung Cancer Res.* 4, 36–54.

Chemi, F., Rothwell, D. G., McGranahan, N., Sakshi, G., Chris, A., Simon, P. P., et al. (2019). Pulmonary venous circulating tumor cell dissemination before tumor resection and disease relapse. *Nat. Med.* 25, 1534–1539. doi: 10.1038/s41591-019-0593-1

Chiang, S., Weigelt, B., Wen, H. C., Fresia, P., Ashwini, R., Luciano, G. M., et al. (2016). IDH2 mutations define a unique subtype of breast cancer with altered nuclear polarity. *Cancer Res.* 76, 7118–7129. doi: 10.1158/0008-5472.can-16-0298

Couraud, S., Vaca-Paniagua, F., Villar, S., Oliver, J., Schuster, T., Blanche, H., et al. (2014). Noninvasive diagnosis of actionable mutations by deep sequencing of circulating free DNA in lung cancer from never-smokers: a proof-of-concept study from BioCAST/IFCT-1002. *Clin. Cancer Res.* 2014, 4613–4624. doi: 10.1158/1078-0432.ccr-13-3063

de Bel, T., Hermsen, M., Smeets, B., Hilbrands, L., van der Laak, J., and Litjens, G. (2018). "Automatic segmentation of histopathological slides of renal tissue using deep learning," in *Medical Imaging 2018: Digital Pathology*, 10581, 1058112 (Houston, TX: SPIE Medical Imaging). doi: 10.1117/12.2293717

Esteva, A., Brett, K., Roberto, A. N., Justin, K., Susan, M. S., Helen, M. B., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 115–118. doi: 10.1038/nature21056

Fischer, A. H., Jacobson, K. A., Rose, J., and Zeller, R. (2008). Hematoxylin and eosin staining of tissue and cell sections. *CSH Protoc.* 2008:pdb.prot4986. doi: 10.1101/pdb.prot4986

Hung, J. J., Hsu, W. H., Hsieh, C. C., Huang, B. S., Huang, M. H., Liu, J. S., et al. (2009). Post-recurrence survival in completely resected stage I non-small cell lung cancer with local recurrence. *Thorax* 2009, 192–196. doi: 10.1136/thx.2007.094912

Kaplan, J. A., Liu, R., Freedman, J. D., Robert, P., John, S., Yolonda, L. C., et al. (2016). Prevention of lung cancer recurrence using cisplatin-loaded superhydrophobic nanofiber meshes. *Biomaterials* 2016, 273–281. doi: 10.1016/j.biomaterials.2015.10.060

Le, N., Ho, Q. T., and Ou, Y. Y. (2017). Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *J. Chmput. Chem.* 28, 2000–2006. doi: 10.1002/jcc.24842

Le, N. Q. K. (2019). Fertility –GRU: indentifying fertility-related proteins by incorporating deep-gated recurrent units and original position-specific scoring matrix profiles. *J. Proteome Res.* 18, 3503–3511. doi: 10.1021/acs.jproteome.9b00411

Le, N. Q. K., Edward, K. Y. Y., and Yeh, H. Y. (2019). ET-GRU: using multi-layer gated recurrent units to identify electron transport proteins. *BMC Bioinformatic* 20:377.

MacConaill, L. E. (2013). Existing and emerging technologies for tumor genomic profiling. *J. Clin. Oncol.* 2013, 1815–1824. doi: 10.1200/jco.2012.46.5948

Makowski, L., and Hayes, D. N. (2008). Role of LKB1 in lung cancer development. *Br. J. Cancer* 99, 683–688. doi: 10.1038/sj.bjc.6604515

Meng, Y., Chi-Wei, C., Mingo, M. H. Y., Wei, S., Jing, S., Zhuqing, L., et al. (2018). DUOXA1-mediated ROS production promotes cisplatin resistance by activating ATR-Chk1 pathway in ovarian cancer. *Cancer Lett.* 428, 104–116. doi: 10.1016/j.canlet.2018.04.029

Mogi, A., and Kuwano, H. (2011). TP53 mutations in nonsmall cell lung cancer. *J. Biomed. Biotechnol.* 2011:583929. doi: 10.1155/2011/583929

Morris, L. G., Andrew, M. K., Yongxing, G., Deepa, R., Logan, A. W., Sevin, T., et al. (2013). Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.* 45, 253–261. doi: 10.1038/ng.2538

Popin, R., Avinash, V. V., Katy, M., Yun, L., Michael, V. M., Greg, S. C., et al. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* 2, 158–164. doi: 10.1038/s41551-018-0195-0

Qaiser, T., Tsang, Y. W., Epstein, D., and RajpootEma, N. (2018). *Medical Image Understanding and Analysis: 21st Annual Conference on Medical Image Understanding and Analysis*. New York, NY: Springer International Publishing.

Reckamp, K. L., Melnikova, V. O., Karlovich, C., Sequist, L. V., Camidge, D. R., Wakelee, H., et al. (2016). A highly sensitive and quantitative test platform for detection of NSCLC EGFR mutations in urine and plasma. *J. Thorac. Oncol.* 2016, 1690–1700. doi: 10.1016/j.jtho.2016.05.035

Shackelford, D. B., Evan, A., Laurie, G., Debbie, S. V., Atsuko, S., Wei, L., et al. (2013). LKB1 inactivation dictates therapeutic response of non-small cell lung cancer to the metabolism drug phenformin. *Cancer Cell* 23, 143–158. doi: 10.1016/j.ccr.2012.12.008

Shen, D., Wu, G., and Suk, H. I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248.

Simon, O., Yacoub, R., Jain, S., Tomaszewski, J. E., and Sarder, P. (2018). Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. *Sci. Rep.* 8:2032.

Sun, J., Xin, C., Mingo, M. H. Y., Wei, Z., Jing, L., Yi, Z., et al. (2019). miR-137 mediates the functional link between c-Myc and EZH2 that regulates cisplatin resistance in ovarian cancer. *Oncogene* 38, 564–580. doi: 10.1038/s41388-018-0459-x

Travis, W. D., Elisabeth, B., Masayuki, N., Andrew, G. N., Kim, G., Yasushi, Y., et al. (2011). International association for the study of lung cancer/american thoracic society/european respirator society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Onco.* 6, 244–285.

Vahadane, A., Tingying, P., Amit, S., Shadi, A., Lichao, W., Maximilian, B., et al. (2016). Structure-preserving color normalization methods and sparse stain separation for histological images. *IEEE Trans.* 35, 1962–1971. doi: 10.1109/tmi.2016.2529665

Wang, X., Andrew, J., Yu, Z., Rajat, T., Pingfu, F., Kurt, S., et al. (2017). Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images. *Sci. Rep.* 7:13543.

Weiner, L. M., Murray, J. C., and Casey, W. S. (2012). Antibody based immunotherapy of cancer. *Cell* 148, 1081–1084.

Xu, S., Lou, F., Wu, Y., Sun, D. Q., Zhang, J. B., Chen, W., et al. (2016). Circulating tumor DNA identified by targeted sequencing in advanced-stage non-small cell lung cancer patients. *Cancer Lett.* 2016, 324–331. doi: 10.1016/j.canlet.2015.11.005

Yu, K. H., Zhang, C., Gerald, J. B., Russ, B. A., Christopher, R., Daniel, L. R., et al. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nat. Commun.* 7:12474.

Zachara-Szzakowki, S., Verdun, T., and Churg, A. (2015). Accuracy of classifying poorly differentiated non-small cell lung carcinoma biopsies with commonly used lung carcinoma markers. *Hum. Pathol.* 46, 766–782.

Conflict of Interest: YBL, XM, and QL are employed by the company Geneis (Beijing) Co., Ltd.

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

frontiers in Oncology

SYSTEMATIC REVIEW
published: 02 September 2020
doi: 10.3389/fonc.2020.01214

The Significance of the *CLDN18-ARHGAP* Fusion Gene in Gastric Cancer: A Systematic Review and Meta-Analysis

Wei-Han Zhang[1†], *Shou-Yue Zhang*[2†], *Qian-Qian Hou*[2†], *Yun Qin*[3], *Xin-Zu Chen*[1], *Zong-Guang Zhou*[4], *Yang Shu*[2*], *Heng Xu*[2*] and *Jian-Kun Hu*[1*‡]

[1] State Key Laboratory of Biotherapy, Department of Gastrointestinal Surgery and Laboratory of Gastric Cancer, West China Hospital, Sichuan University and Collaborative Innovation Center of Biotherapy, Chengdu, China, [2] State Key Laboratory of Biotherapy, West China Hospital, Sichuan University and Collaborative Innovation Center, Chengdu, China, [3] Department of Radiology, West China Hospital, Sichuan University, Chengdu, China, [4] State Key Laboratory of Biotherapy, West China Hospital, Department of Gastrointestinal Surgery and Laboratory of Digestive Surgery, Sichuan University, and Collaborative Innovation Center for Biotherapy, Chengdu, China

Objective: The objective of this study was to summarize the clinicopathological characteristics of the *CLDN18-ARHGAP* fusion gene in gastric cancer patients.

Background: The *CLDN18-ARHGAP26* fusion gene is one of the most frequent somatic genomic rearrangements in gastric cancer, especially in the genomically stable (GS) subtype. However, the clinical and prognostic meaning of the *CLDN18-ARHGAP* fusion in gastric cancer patients is unclear.

Methods: Studies that investigated *CLDN18-ARHGAP* fusion gastric cancer patients were identified systematically from the PubMed, Cochrane, and Embase databases through the 28th of February 2020. A systematic review and meta-analysis were performed to estimate the clinical significance of *CLDN18-ARHGAP* fusion in patients.

Results: A total of five eligible studies covering 1908 patients were selected for inclusion in the meta-analysis based on specified inclusion and exclusion criteria. Several fusion patterns were observed linking *CLDN18* and *ARHGAP26* or *ARHGAP6*, with the most common type being *CLDN18/exon5-ARHGAP26/exon12*. The survival outcome meta-analysis of the *CLDN18-ARHGAP* fusion gene showed that it was associated with overall survival outcomes in gastric cancer (HR, 2.03, 95% CI 1.26–3.26, $P < 0.01$, random-effects). In addition, diffuse gastric cancer had a greater proportion of CLDN18-ARHGAP fusions than intestinal gastric cancer (13.3%, 151/1,138 vs. 1.8%, 8/442; $p < 0.001$). Moreover, gastric cancer patients with the *CLDN18-ARHGAP* fusion gene are more likely to be female or have a younger age, lymph node metastasis and advanced TNM stages.

Conclusion: The *CLDN18-ARHGAP* fusion is one of the molecular characteristics of diffuse gastric cancer and is also an independent prognostic risk factor for gastric cancer. In addition, it is also related to multiple clinical characteristics, including age, sex, lymph node metastasis and tumor stage. However, the mechanism of the *CLDN18-ARHGAP* fusion gene and potential targeted therapeutic strategies need further exploration.

Keywords: gastric cancer, fusion gene, *CLDN18-ARHGAP*, survival, therapy

OPEN ACCESS

Edited by:
Ling Kui,
Harvard Medical School,
United States

Reviewed by:
Hong Zheng,
Stanford University, United States
Bogang Wu,
George Washington University,
United States
Xinglei Liu,
Albert Einstein College of Medicine,
United States

***Correspondence:**
Yang Shu
shuyang1986@gmail.com
Heng Xu
xuheng81916@scu.edu.cn
Jian-Kun Hu
hujkwch@126.com

[†]These authors have contributed equally to this work

[‡]ORCID:
Jian-Kun Hu
orcid.org/0000-0002-3294-3471

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

Received: *03 March 2020*
Accepted: *15 June 2020*
Published: *02 September 2020*

Citation:
Zhang W-H, Zhang S-Y, Hou Q-Q, Qin Y, Chen X-Z, Zhou Z-G, Shu Y, Xu H and Hu J-K (2020) The Significance of the CLDN18-ARHGAP Fusion Gene in Gastric Cancer: A Systematic Review and Meta-Analysis. Front. Oncol. 10:1214. doi: 10.3389/fonc.2020.01214

INTRODUCTION

Recurrent chromosomal translocations have been implicated in multiple tumor types. It has been demonstrated that frequent fusion genes are involved in oncogenesis and progression as driver events. The Philadelphia chromosome is well reported as the first cancer-associated chromosomal rearrangement, resulting in the *BCR-ABL* fusion, which was also identified as a diagnostic feature and therapeutic target of chronic myeloid leukemia patients. Thereafter, oncologic fusion genes were frequently identified in leukemia, lymphoma and sarcoma, but with a relatively low incidence rate in epithelial tumors (1). For epithelial tumors, the most well-known fusion gene is the *EML4-ALK* gene, which was detected in ∼5–10% of non-small cell lung cancer patients (2, 3).

Regarding gastric cancer, The Cancer Genome Atlas (TCGA) project proposes the molecular classification of gastric cancers and divides them into four separate subtypes, and the *CLDN18-ARHGAP26* fusion gene is highly enriched in the genomically stable (GS) subtype (4). Histologically, the majority of GS subtype cancers were the diffuse type according to the Lauren classification, with common somatic mutations located in the *CDH1* and *RHOA* genes. It is notable that *CLDN18-ARHGAP* fusions were mutually exclusive with *RHOA* mutations in the classification of TCGA (4). A subsequent functional study indicated that the introduction of the *CLDN18-ARHGAP26* fusion to tumor cells can direct the loss of the epithelial phenotype, epithelial-mesenchymal transition, and inhibition of the RHOA signaling pathway and contribute to tumor invasiveness in cancer cell lines (5). Our study group previously used a whole-genome sequencing approach to characterize the genomic features of signet ring cell gastric cancer and identified frequent *CLDN18-ARHGAP* fusions (6). More importantly, we linked multiple clinical characteristics with *CLDN18-ARHGAP28/6* fusions, including the proportion of signet ring cell content, TNM stage, and poor prognosis with the current chemotherapy strategy (6). These findings were quickly validated by an independent Japanese study (7, 8). Meanwhile, a further Korean study found that the *CLDN18-ARHGAP* fusion gene can promote the invasion and migration capacity of gastric cancer cells (9). However, there is a lack of studies systematically evaluating the clinicopathological characteristics and prognostic meaning of *CLDN18-ARHGAP* fusions in gastric cancers.

Therefore, in this meta-analysis and systematic review, we will systematically summarize and assess the clinical significance and advances of the *CLDN18-ARHGAP* fusion gene in gastric cancer. The primary endpoint of the present study is the survival outcomes of patients with the *CLDN18-ARHGAP* fusion gene, and other endpoints are the relationship of the *CLDN18-ARHGAP* fusion gene with tumor-related clinicopathological characteristics, such as age, sex, tumor location and tumor stage.

Abbreviations: TCGA, The Cancer Genome Atlas; NOS, Newcastle–Ottawa Scale; SD, standard deviation; OR, odds ratio; MD, mean difference; HR, hazard ratio; CI, confidence intervals; GRAF, GTPase Regulator Associated with Focal Adhesion Kinase; EMT, Epithelial-Mesenchymal Transition; PDX, Patient-Derived Xenograft; RT-PCR, reverse transcription-polymerase chain reaction; FISH, Fluorescence *in situ* Hybridization.

METHODS

Search Strategy

We searched the Web of Knowledge, PubMed, Embase and Cochrane Collaborative Center Register of Controlled Trials databases on the 28th of February 2020 by using the terms "gastric cancer," "gastric carcinoma," "gastric neoplasm," "stomach cancer," "stomach carcinoma," "stomach neoplasm," "CLDN," "claudin," "ARHGAP," "Rho GTPase-activating protein," "oligophrenin-1-like" and "OPHN1L" and strictly restricted search results to titles, abstracts and keywords. We also searched previously published meta-analyses and systematic reviews. All of those articles were independently screened by two authors (WH Zhang and SYZ) based on the inclusion and exclusion criteria of the study. Because the studies included in this meta-analysis have been published, ethical approval was not needed from ethics committees. The results of this study were reported according to the PRISMA statement (10).

Study Selection

Those studies that reported the relationship between the *CLDN18-ARHGAP* fusion gene and the clinicopathological characteristics or survival outcomes of gastric cancer patients were included. The exclusion criteria included the following: (1) mixed benign disease of the stomach; (2) articles in languages other than English; and (3) incomplete data or duplicated data. For studies with more than one article and with duplicated data, only the article with the most complete data was included for analysis in this study.

Data Extraction and Quality Assessment

Data from the included studies were independently extracted by two authors (WH Zhang and QQ Hou). For each study, we recorded the following information: name of the first author, year of publication, country of the study, study design, time period of the study and examined method for the *CLDN18-ARHGAP* fusion gene. Furthermore, the following clinicopathological characteristics were also extracted and included in the present study: fusion types of the *CLDN18* and *ARHGAP* genes, age (years), sex (male or female), tumor location (upper third of stomach), tumor stage (T stage, N stage and TNM stage) and survival outcomes between *CLDN18-ARHGAP* fusion-positive and CLDN18-ARHGAP fusion-negative patients. The patients were divided into a *CLDN18-ARHGAP* fusion-positive group and a *CLDN18-ARHGAP* fusion-negative group according to the status of the expression of *CLDN18-ARHGAP26*/6 fusions.

The quality assessment of the included studies was evaluated by two authors (WH Zhang and QQ Hou) independently. Retrospective studies were assessed by the Newcastle–Ottawa Scale (NOS), which is a 9-point scale (11). Studies with NOS scores lower than 6 were deemed moderate or low-quality studies. Any disagreements regarding the quality assessment were resolved by discussion with supervisors (H Xu and JK Hu).

Statistical Analysis

This study was performed according to the Cochrane guidelines (12). For studies that only reported the medians and ranges for continuous variables, the data were converted to means and

standard deviations (SDs) with the method reported by Hozo et al. (13). Categorical variables are presented as ratios and were analyzed by the Mantel-Haenszel method, and continuous variables are presented as the mean ± SD and were analyzed by the inverse variance method. The odds ratio (OR) and mean difference (MD) were used to evaluate dichotomous and continuous data, respectively. The hazard ratio (HR) was used to evaluate survival outcomes. The OR, HR and MD were reported with 95% confidence intervals (CIs). Heterogeneity among studies was assessed by the I^2 value. According to the I^2 value, the studies were determined to have low ($I^2 < 30\%$), moderate (30–50%) or considerable ($I^2 \geq 50\%$) heterogeneity. Begg's test was used to assess publication bias. For the survival analysis, we updated the survival information to Jan 2019 of our previous study (all 829 patients, Shu et al.) (6). In addition, individual survival information from the TCGA cohort was also updated according to a recent report from the TCGA research network (14). Survival information from other studies was extracted with the method reported by Tierney et al. by Engauge Digitizer software (Version 11.2) (15). A P-value < 0.05 was considered statistically significant for the present study. All of the statistical analyses were performed by R software (http://

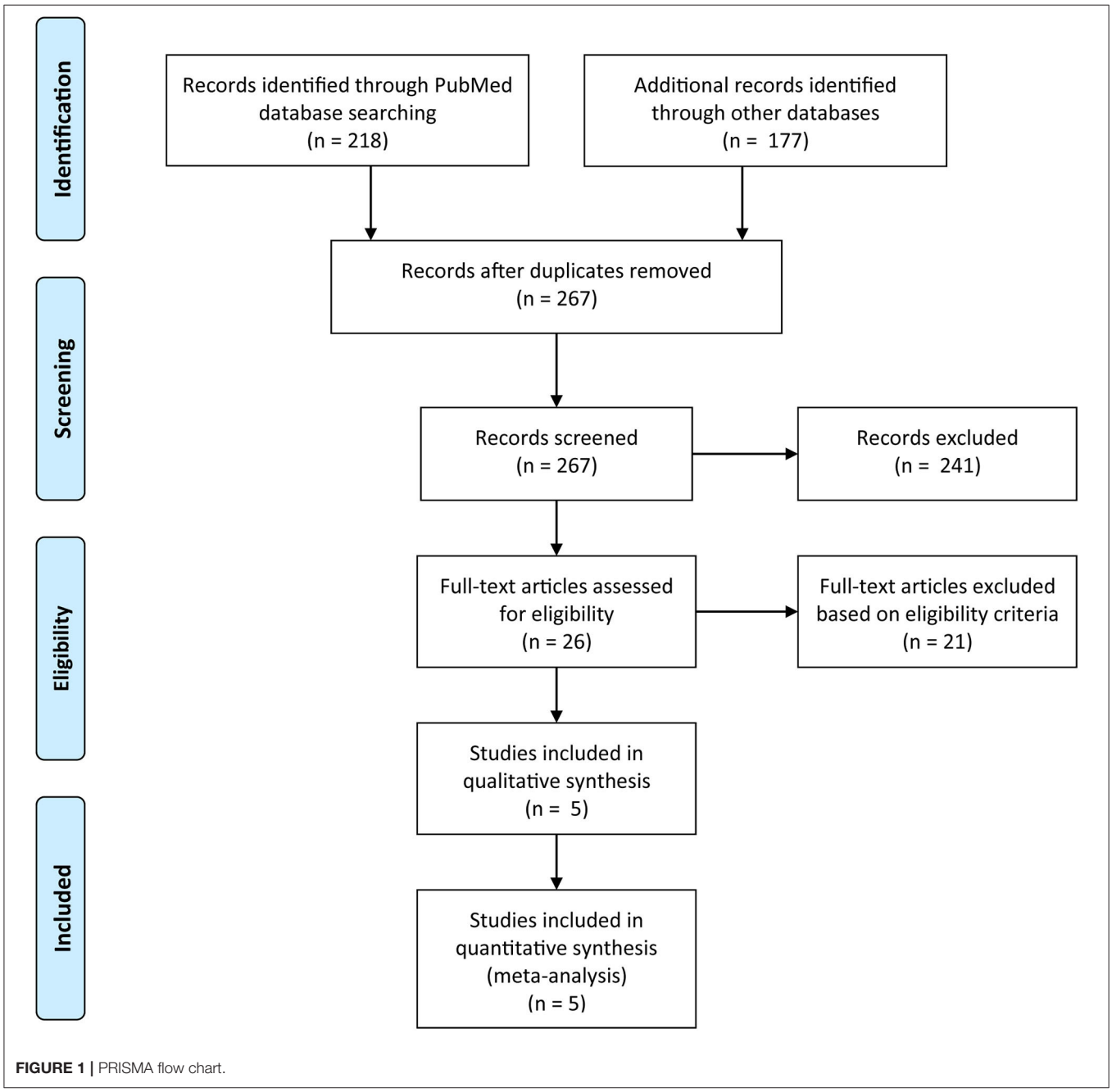


FIGURE 1 | PRISMA flow chart.

www.R-project.org/) with the "survival," "survminer," "ggplot2," "meta," and "metafor" packages.

RESULTS

General Characteristics

We retrieved 395 records with 128 duplicates. After reading the titles and abstracts, 26 articles remained for reassessment according to their full texts. After reading the full texts of these articles, we included five studies that presented the relationship of clinicopathological characteristics and survival outcomes with *CLDN18-ARHGAP* fusion gene status (**Figure 1**, PRISMA Flow Diagram). We also evaluated the quality of all included studies with the Newcastle-Ottawa Scale, and the results showed that all studies had a score ≥ 6.

All 5 studies were from four countries (United States, Japan, Korea and China) and were published between 2014 and 2019. Finally, a total of 1,908 patients were included in the present study: 151 (7.9%) patients in the *CLDN18-ARHGAP* fusion-positive group and 1,757 (92.1%) patients in the *CLDN18-ARHGAP* fusion-negative group. General characteristics of those included five studies were summarized in **Table 1**. There were several fusion types in *CLDN18-ARHGAP* fusion gene gastric cancer patients. Also, the fusion types from the five reported studies between *CLDN18 and ARHGAP* genes are presented in **Table 2**.

The CLDN18/exon5-ARHGAP26/exon12 fusion was the most common fusion pattern among present reports and is clearly understood. In addition, the mutation counts between the CLDN18-ARHGAP fusion-positive and CLDN18-ARHGAP fusion-negative patients in cohorts of Shu et al. and TCGA were analyzed (**Supplementary Tables 1, 2**). However, there was no overlap of the significant mutation counts gene between CLDN18-ARHGAP fusion-positive and -negative patients in the Shu and TCGA cohorts. Different pathological subtypes and the sample size difference of the two cohorts may be the reasons for the results. However, the study of Shu et al. was based on the whole genome sequencing, so we cannot analyze the RNA expression level between fusion positive and negative patients. Therefore, we analyzed copy number variation in the cohort of Shu et al. (**Supplementary Table 3**) and the RNA expression level (**Supplementary Table 4**) in the TCGA cohort between CLDN18-ARHGAP fusion-positive and CLDN18-ARHGAP fusion-negative gastric cancer patients.

In the meta-analysis of clinicopathological characteristics, the *CLDN18-ARHGAP* fusion gene-positive group had more patients with a younger age (MD: −5.85, 95% CI: −11.22 to −0.48, $p = 0.03$), a lower proportion of male patients (OR: 0.40, 95% CI 0.23–0.70, $p = 0.001$), patients with a more advanced N stage (OR: 3.41, 95% CI 2.00–5.82, $p < 0.001$) and patients with a more advanced TNM stage (OR: 3.07, 95% CI 1.56–6.05, $p = 0.001$) than the *CLDN18-ARHGAP* fusion gene-negative group (**Table 3**). However, tumor location ($p = 0.43$) and T stage ($p = 0.07$) were not significantly different between the two groups. Moreover, we found that diffuse gastric cancers had a greater proportion of *CLDN18-ARHGAP* fusion genes than intestinal gastric cancers (13.3%, 151/1,138 vs. 1.8%, 8/442; $p < 0.001$).

Survival Analysis

CLDN18-ARHGAP fusion-positive gastric cancer patients had significantly poorer overall survival outcomes than *CLDN18-ARHGAP* fusion-negative patients in the meta-analysis (HR:

TABLE 1 | Characteristics of studies reported clinicopathological characteristics between *CLDN18-ARHGAP* fusion and gastric cancers.

Study	Country	Period	Samples	Lauren Type			*CLDN18-ARHGAP Fusion positive*			Tumor stage	Examine methods	NOS#	Special Characteristics
				N = (%)			*N* = (%)						
				DGC*	IGC	NA	All	DGC	IGC				
Nakayama et al. (7)	Japan	2006–2015	146	136 (93.2)	10 (6.8)	NA	22 (15.1)	22 (16.2)	0 (0.0)	I–IV	Fusion-FISH, RNA-seq	7	Young age Patients (≤40)
Shu et al. (6)	China	2009–2014	829	358 (43.2)	154 (18.6)	317 (38.2)	73 (8.8)	55 (15.4)	2 (1.3)	I–IV	WGS, RT-PCR	8	Fusion related to the proportion of SRCC
Tanake et al. (8)	Japan	2000–2013	254	172 (67.7)	82 (32.3)	0	26 (10.2)	22 (12.8)	4 (4.9)	I–IV	RT-PCR, FISH	6	Fusion-positive DGCs E-cad expression
Yang et al. (9)	Korea	2003–2017	384	384 (100.0)	0	0	17 (4.4)	17 (4.4)	0	I–IV	RT-PCR, RNA seq	6	Fusion related to H. pylori infections
TCGA (4)	United State	NM	295	88 (29.8)	196 (66.4)	11 (3.7)	13 (4.4)	10 (11.4)	2 (1.0)	I–IV	WGS or RNA-seq	8	Fusion related to GS tumors

**Included mixed type.*

#NOS, The Newcastle-Ottwa Scale (11).

DGC, diffuse gastric cancer; IGC, intestinal gastric cancer; TCGA, The Cancer Genome Atlas; WGS, whole-genome sequence; RNA-seq, RNA sequence; RT-PCR, reverse transcription-polymerase chain reaction; FISH, fluorescence in situ hybridization, SRCC, signet ring cell cancer; GS, genomically stable; NA, not applicable; NM, not mentioned.

TABLE 2 | The *CLDN18-ARHGAP* fusion models in reported studies.

Study	Number of *CLDN18-ARHGAP* fusion	Fusion Mode		Numbers (%)
Nakayama et al. (7)	22	*CLDN18/exon5*	*ARHGAP26/exon10*	18 (81.8)
			ARHGAP26/exon 12	
		CLDN18/exon5	*ARHGAP6/exon2*	2 (9.1)
		CLDN18/exon5	*ARHGAP42/exon7*	1 (4.5)
		CLDN18/exon5	*ARHGAP10/exon8*	1 (4.5)
Shu et al. (6)	73	*CLDN18/exon5*	*ARHGAP26/exon12*	58 (79.5)
		CLDN18/exon5	*ARHGAP26/exon10*	7 (9.6)
		CLDN18/exon4	*ARHGAP26/exon11*	1 (1.4)
		CLDN18/exon5	*ARHGAP6/exon2*	7 (9.6)
Tanake et al. (8)	26	*CLDN18/exon5*	*ARHGAP26/exon12*	24 (92.3)
		CLDN18/exon5	*ARHGAP26/exon10*	1 (3.8)
		CLDN18/exon5	*ARHGAP6/exon2*	1 (3.8)
Yang et al. (9)	17	*CLDN18*	*ARHGAP26*	13 (76.5)
		CTNND1	*ARHGAP26**	2 (11.8)
		ANXA2	*MYO9A**	2 (11.8)
TCGA (4)	13	*CLDN18/exon5*	*ARHGAP26/exon12*	10 (76.9)
		CLDN18/exon5	*ARHGAP26/exon10*	1 (7.7)
		CLDN18/exon5	*ARHGAP6/exon2*	2 (15.4)

TCGA, The Cancer Genome Atlas.
**Also known as RhoGAP domain-containing fusions.*

TABLE 3 | The Meta-analysis of clinicopathological characteristics between patients with *CLDN18-ARHGAP* fusion positive and negative patients.

Characteristics	Included Study	*CLDN18-ARHGAP* Fusion (+) *N* = (%)	*CLDN18-ARHGAP* fusion (–) *N*= (%)	Test of Heterogeneity			Meta-analysis		
				χ^2	I^2 (%)	*p*-value	OR or MD	95% CI	*p*-value
Age (years)	(4, 6, 8)	112	1177	8.76	77	0.01	−5.85	−11.22 to −0.48	0.03
Gender (Male)	(4, 6–9)	61/151 (40.4)	1,107/1,668 (66.4)	8.51	53	0.07	0.40	0.23–0.70	0.001
Tumor Location (Upper)	(4, 6–9)	31/151 (20.5)	621/1,668 (37.2)	14.19	72	0.007	0.68	0.27–1.75	0.43
T stage (T2–T4)	(4, 6–8)	116/134 (86.6)	1,131/1,301 (86.9)	0.63	0	0.89	1.76	0.96–3.22	0.07
N stage (*N*+)	(4, 6–8)	110/134 (82.1)	902/1,301 (69.3)	3.33	10	0.34	3.41	2.00–5.82	<0.001
TNM stage (III–IV)	(4, 6, 7, 9)	120/151 (79.5)	1,046/1,668 (62.7)	7.66	48	0.10	3.07	1.56–6.05	<0.001

95% CI, 95% confidence interval; MD, mean difference; OR, odds ratio.

2.03, 95% CI 1.26–3.26, $p < 0.01$, random effects) (**Figure 2A**), and the survival results in the meta-analysis were relatively stable in the sensitivity analysis (**Figure 2B**). Because there were only 4 studies included in the survival analysis, publication bias was only evaluated by Begg's test. The results demonstrated that there was no publication bias according to Begg's test with continuity correction ($p = 0.555$).

In addition, we acquired updated individual survival information from the cohort of Shu et al. and the TCGA cohort. Therefore, the survival difference between *CLDN18-ARHGAP* fusion-positive and *CLDN18-ARHGAP* fusion-negative patients was evaluated in these two cohorts (**Figures 3A,B**). A significant survival difference was found between the *CLDN-ARHGAP* fusion gene-positive and *CLDN-ARHGAP* fusion gene-negative groups with the combination of the data from the two cohorts ($p < 0.001$) (**Figure 3C**). In addition, the Cox proportional hazards model was used to present the independent prognostic risk factors in the merged data of the Shu and TCGA cohorts (**Table 4**). In multivariate survival analysis, positive *CLDN18-ARHGAP* fusion (HR: 1.365, 95% CI 1.031–1.809, $p = 0.030$) and TNM stage (stage III vs. stage I, HR: 3.018, 95% CI 1.763–5.164, $p < 0.001$; stage IV vs. stage I, HR: 7.155, 95% CI 4.083–12.538, $p < 0.001$) were independent prognostic risk factors.

DISCUSSION

In the present study, a total of 5 cohort studies reported the presence of *CLDN18-ARHGAP* fusions and its relationship with clinicopathological characteristics and survival outcomes

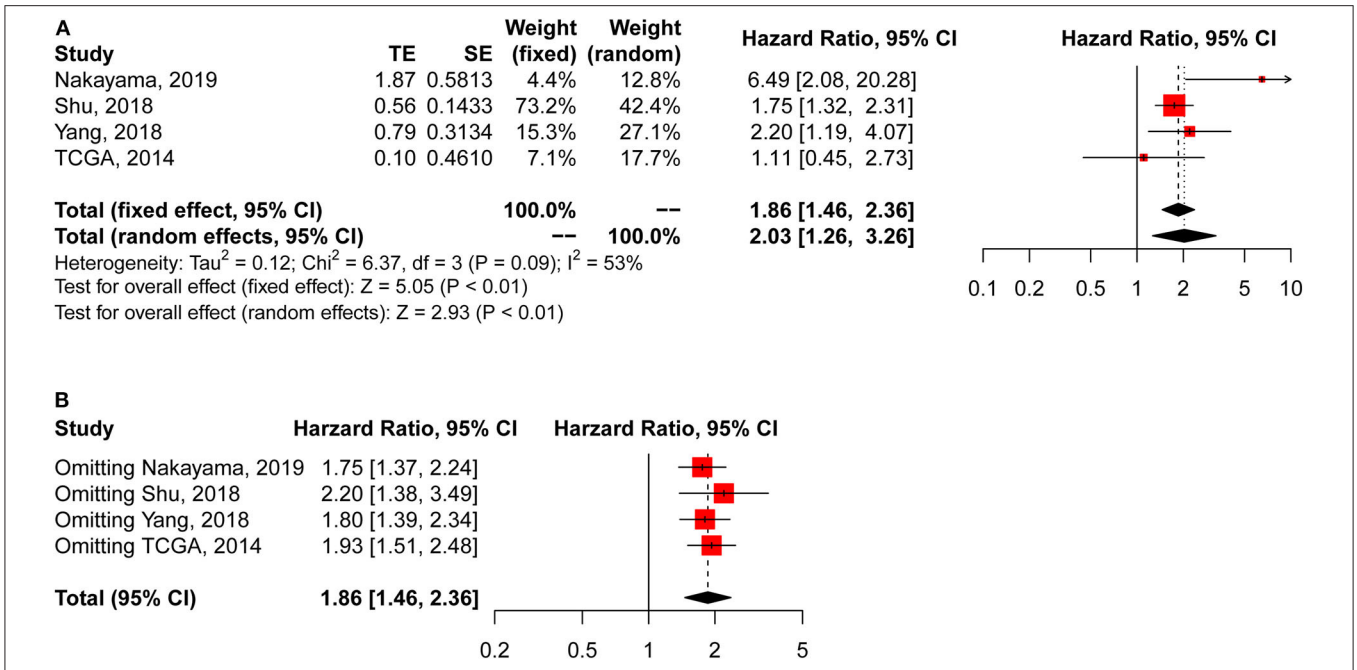


FIGURE 2 | Meta-analysis of survival outcomes between CLDN18-ARHGAP fusion-negative and CLDN18-ARHGAP fusion-positive gastric cancer patients. **(A)** Survival outcomes of the CLDN18-ARHGAP26 fusion gene (positive vs. negative); **(B)** sensitivity analysis of the included studies.

in gastric cancer patients. Multiple clinical characteristics were observed to be correlated with the frequency of *CLDN18-ARHGAP* fusions. Finally, significant enrichment of *CLDN18-ARHGAP* fusions was observed in female patients and patients with a younger age, diffuse gastric cancer by Lauren classification and more advanced tumor stages (N stage and TNM stage), but *CLDN18-ARHGAP* fusions were not related to the primary tumor location. Most importantly, the *CLDN18-ARHGAP* fusion was significantly related to poor survival outcomes in the meta-analysis (HR: 2.03, 95% CI 1.26–3.26, $p < 0.01$, random effects). Meanwhile, in the survival analysis with the combination of individual data from the Shu et al. and TCGA cohorts, the CLDN18-ARHGAP fusion was an independent prognostic risk factor for overall survival outcomes (HR: 1.365, 95% CI 1.031–1.809, $p = 0.030$).

The *CLDN18-ARHGAP* fusion gene is formed by chromosomal rearrangements of the *CLDN18* and *ARHGAP* genes, mainly *CLDN18-ARHGAP26* and *CLDN18-ARHGAP6* fusions. *CLDN18-ARHGAP26/6* contains a nearly full coding region of *CLDN18* and the conserved domain of *ARHGAP26/6*. Functionally, the *CLDN18* gene encodes the claudin-18 protein, which forms tight junctions in epithelial cells. The CLDN18-ARHGAP fusion protein may disrupt the structure of the wild-type CLDN18 protein, which may impact the cellular adhesion of cancer cells. The *ARHGAP26* gene encodes the ARHGAP26 protein, a member of the Rho GTPase activating the protein, which is also known as the GTPase regulator associated with focal adhesion kinase (GRAF). GRAF not only regulates the activity of RHOGAP family proteins (16) but also coordinates membrane remodeling, which is necessary for the CLIC/GEEC endocytic pathway (17). Regev et al. suggested that the GRAF protein may play a role in the maintenance of the normal epithelial phenotype, the depletion of which can induce a neoplastic transformation-related epithelial-mesenchymal transition (EMT)-like process (18). For the fusion proteins of CLDN18-ARHGAP, the large segment of ARHGAP fuses to the carboxy terminus of CLDN18 and retains the carboxy-terminal GAP domain, which may affect ARHGAP's regulation of the RHOA pathway and/or the epithelial phenotype of gastric cancer cells. Several studies have indicated that the introduction of the *CLDN18-ARHGAP26* fusion in cancer cells can increase their migration and invasion ability (5, 6), which can partially explain the advanced tumor stages in *CLDN18-ARHGAP* fusion-positive patients.

In the TCGA gastric cancer cohort, the *CLDN18-ARHGAP26/6* fusion was enriched in patients with the genomically stable subtype, which has higher frequency of lower third tumors, patients with a younger age and more diffuse histological subtype tumors (4). In our previous study, we found that the *CLDN18-ARHGAP26/6* fusion was significantly associated with the proportion of signet ring cancer cells and tumor stage (6). Tanaka et al. also observed a higher frequency of *CLDN18-ARHGAP26/6* fusions in diffuse gastric cancers than in intestinal gastric cancers (22/172, 12.8% vs. 4/82, 4.8%) (8). In the present study, we found a significant difference in the frequency of the *CLDN18-ARHGAP* fusion gene between diffuse gastric cancer and intestinal gastric cancer (13.3%, 151/1,138 vs. 1.8%, 8/442; $p < 0.001$). Therefore, the *CLDN18-ARHGAP* fusion may be an important molecular characteristic of diffuse gastric cancer.

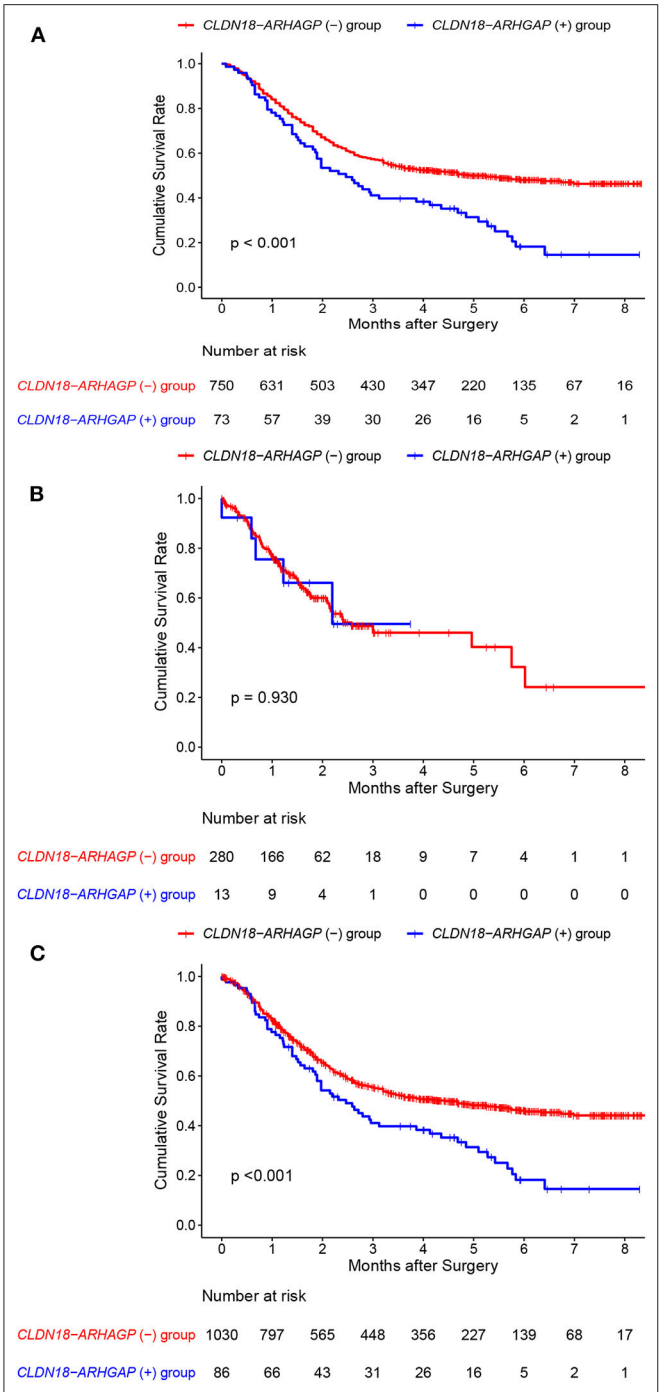


FIGURE 3 | Individual overall survival outcomes of CLDN18-ARHGAP26/6 fusion gastric cancer patients in the Shu et al. and TCGA cohorts. **(A)** Overall survival outcomes in the Shu et al. cohort; **(B)** overall survival outcomes in the TCGA cohort; **(C)** overall survival outcomes in the two cohorts.

It is well reported that the diffuse subtype has a significantly poorer prognosis than the intestinal subtype of gastric cancer according to the Lauren classification, probably because of the more advanced tumor stages and potential resistance to traditional chemotherapy regimens of diffuse gastric cancers (19).

We previously reported the prognostic value of the *CLDN18-ARHGAP26/6* fusion, which is a risk factor for overall survival and confers postoperative chemotherapy resistance (6). The present study summarized the survival outcomes of previously reported studies focused on the C*LDN18-ARHGAP* fusion gene. Thereafter, the survival outcome meta-analysis showed that patients with *CLDN18-ARHGAP* fusion have a significantly poorer prognosis than patients without *CLDN18-ARHGAP* fusion. Due to the enrichment of the *CLDN18-ARHGAP* fusion in patients with more advanced stages, it is important to assess the factors independently associated with the *CLDN18-ARHGAP* fusion. Multivariate analyses of individual data from the Shu and TCGA cohorts presented a significant association of the *CLDN18-ARHGAP* fusion status with poor treatment outcomes after adjusting for tumor stage, indicating that the *CLDN18-ARHGAP* fusion is an independent prognostic factor for gastric cancers. In addition, it is necessary to mention that some of the included studies were not traditional clinical studies, and limited follow-up durations (such as that in the TCGA cohort) may increase the bias risk in the survival analysis above.

According to a previous study (6), gastric cancer patients with the *CLDN18-ARHGAP26/6* fusion gene cannot obtain survival benefits from 5-FU/oxaliplatin-based chemotherapy, which may partially explain the poor prognosis of *CLDN18-ARHGAP* fusion patients. However, no other study analyzed the relationship between the *CLDN18-ARHGAP26* fusion gene and chemotherapy drug therapeutic sensitivity. Mechanistically, resistance to these chemotherapy drugs was observed after the introduction of the *CLDN18-ARHGAP26/6* fusion into cell lines (6). Because no reported gastric cancer cell lines carry *CLDN18-ARHGAP26/6* fusions according to the Cancer Cell Line Encyclopedia database (20, 21), patient-derived xenograft (PDX) and organoid models may be the breakthrough point for future research to help validate drug resistance, screen fusion-targeted drugs, and guide personalized therapy (22–24). Yan et al. described a gastric cancer organoid model that can be used to assess the efficacy of chemotherapy (25). Unfortunately, no patient with *CLDN18-ARHGAP26* fusion was captured in their gastric cancer organoid bank. Nakayama et al. established two *CLDN18-ARHGAP26* fusion-positive cell lines from 125 gastric cancer PDXs (26). Collectively, these results suggest that the establishment of PDX and organoid models can help researchers conduct drug sensitivity screening and explore personalized medicine applications for therapy response testing in the future.

Specifically, the aberrant activation of claudin-18 splice variant 2 (claudin-18.2) was detected in multiple types of cancer compared with its limited expression in normal tissues. The rate of claudin-18.2-positive patients was more than 80% according to a Japanese study on gastric cancer, and more than 40% of patients had moderate-to-strong expression ($\geq$ 2+ membrane staining intensity in $\geq$ 40% of tumor cells) in both the primary tumor and metastatic lymph nodes (27). Claudin-18.2 has been considered a novel druggable target for some epithelial tumors (28). Indeed, a chimeric monoclonal antibody drug has been recently developed (i.e., zolbetuximab, formerly known as IMAB362), which induces the immune-mediated lysis of CLDN18.2-positive cancer cells by activating immune effector mechanisms (29).

TABLE 4 | Univariate and Multivariate survival analysis of *CLDN18-ARHGAP26/6* fusion gene in TCGA and Shu cohort.

Characteristics		Univariate Analysis			Multivariate Analysis		
		OR	95% CI	*P*-value	OR	95% CI	*P*-value
Age (years)	<65 vs. ≥65	1.14	0.877–0.959	0.136	1.343	1.125–1.603	0.090
Gender	Male vs. Female	1.042	0.868–1.252	0.658	0.988	0.817–1.195	0.900
Tumor location	Upper vs. Other	0.973	0.810–1.168	0.769	0.868	0.722–1.044	0.134
T stage	T2 vs. T1	3.119	0.960–10.130	0.058			
	T3 vs. T1	3.953	1.253–12.470	0.019			
	T4 vs. T1	6.455	2.073–20.100	0.001			
N stage	N1 vs. N0	1.999	1.379–2.897	<0.001			
	N2 vs. N0	2.453	1.733–3.471	<0.001			
	N3 vs. N0	3.94	2.888–5.376	<0.001			
TNM stage	II vs. I	1.29	0.725–2.294	0.386	1.339	0.753–2.385	0.32
	III vs. I	2.864	1.668–4.889	<0.001	3.018	1.763–5.164	<0.001
	IV vs. I	6.669	3.825–11.628	<0.001	7.155	4.083–12.538	<0.001
CLDN18-ARHGAP26/6 fusion	Positive vs. Negative	1.629	1.247–2.127	<0.001	1.365	1.031–1.809	0.03

OR, odds ratio; 95% CI, 95% confidence interval.

**Only TNM stage entered into the Cox regression model due to the potentially confounding effect.*

Clinical trials evaluating the safety and efficacy of zolbetuximab for claudin-18.2-positive cancer patients are ongoing. The up-to-date evidence is promising but remains to be validated by high-quality clinical trials (30, 31). We also noticed that there is an ongoing clinical trial, which is focused on safety and efficacy of anti-claudin18.2 chimeric antigen receptor t-cell (CAR-T) immunotherapy in patients with advanced gastric cancer or pancreatic cancer (ClinicalTrials.gov Identifier: NCT03159819). The final results of the anti-claudin-18.2 CAR-T immunotherapy as a new anti-tumor targeted immunotherapy study are highly anticipated. Considering that a higher claudin-18.2 positive rate was observed in patients with the diffuse subtype of gastric cancer than in those with the intestinal type (57.5 vs. 39.0%) (27), an unsolved and important question is whether the *CLDN18-ARHGAP* fusion gene is correlated with claudin-18.2 protein expression and thus is suitable for zolbetuximab treatment, which requires solid clinical evidence.

According to previous studies, whole-genome sequencing, RNA sequencing, reverse transcription-polymerase chain reaction (RT-PCR) and fluorescence *in situ* hybridization (FISH) are all effective methods for the detection of *CLDN18-ARHGAP* fusions. Once the clinical significance of the *CLDN18-ARHGAP* fusion has been proven and potential target treatment regimens have been determined, establishing a stable and effective detection method for this fusion gene is particularly important. Considering the preservation of tumor tissue as well as the stability and economic cost of the examination, FISH may be the first choice for promotion in clinical practice. In addition, RNA *in situ* hybridization techniques may be useful in the detection of fusion genes. However, the sensitivity and specificity of the detection of fusion genes by such methods should be validated by clinical studies. In addition, oncogenic fusion circRNAs (f-circRNAs) derived from cancer-associated chromosomal translocations exhibit properties of tumor-promoting cellular transformation, cell viability and resistance to treatment (32, 33). Consistently, f-circRNAs derived from *SLC32A2-ROS1* and *EML4-ALK* fusion genes, which have been determined as biomarkers for the use of targeted drugs in lung cancer, were also demonstrated to impact cell migration, invasion and cell proliferation in lung cancer cells (34, 35). More importantly, the f-circRNAs of *EML4-ALK* can be detected in the plasma of *EML4-ALK*-positive NSCLC patients (36). These results suggest the following: (1) f-circRNAs are involved in the mechanism of tumorigenesis, progression and therapy resistance; and (2) cell-free f-circRNAs could be a novel "liquid biopsy" biomarker to monitor the status of fusion genes in a noninvasive way. Therefore, we speculate and propose the following hypothesis: f-circRNAs of *the CLDN18-ARHGAP* fusion gene exist and can affect tumor function or act as a potential "liquid biopsy" biomarker for targeted drugs (e.g., zolbetuximab).

The present study also has some limitations. (1) This study only included five retrospective studies. Therefore, selection bias and quality deviation are likely among these studies, which may have an influence on the results of the meta-analysis. (2) The detection methods varied among the included studies. The potential false positive and false negative rates of *CLDN18-ARHGAP* fusion in the included studies may have influenced the results of the meta-analysis. (3) In addition, the limited follow-up duration of the included studies was another limitation of the presented studies. (4) Although previous studies successfully demonstrated that the *CLDN18-ARHGAP26* fusion gene can induce EMT, the loss of the epithelial phenotype, and cell-cell and cell-extracellular matrix adhesion, as well as increase the invasion ability and resistance to chemotherapy drugs in cancer cell lines (5, 6), the specific molecular mechanisms by which *CLDN18-ARHGAP26* regulates downstream molecules and pathways remain unclear. As different fusion modes generate various fusion proteins, it is difficult to design and develop antibodies to specifically target these fusion proteins, which may hinder mechanistic investigation of the *CLDN18-ARHGAP* fusion gene. Furthermore, large sample size multicenter studies are expected to validate the clinical significance and prognostic meaning of the *CLDN18-ARHGAP26* fusion gene in gastric cancer patients.

CONCLUSIONS

The CLDN18-ARHGAP fusion gene is characterized as one of the features of diffuse gastric cancer. The *CLDN18-ARHGAP* fusion gene is correlated with advanced tumor stages in gastric cancer, as well as poor survival outcomes. Although *CLDN18-ARHGAP* fusion can increase the invasion and migration ability of gastric cancer cells *in vitro*, the molecular mechanism remains to be elucidated. Furthermore, the early detection of the *CLDN18-ARHGAP* fusion and targeted drugs for this fusion may potentially improve the survival outcomes of gastric cancer patients.

DATA AVAILABILITY STATEMENT

The raw data supporting the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

W-HZ, HX, and J-KH designed the study. W-HZ, S-YZ, Q-QH, X-ZC, YQ, and YS collected information and analyzed and interpreted the data. Z-GZ, YS, HX, and J-KH supervised this study. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the National Natural Science Foundation of China (grant numbers: 81902437, 81673452, and 81973408); the 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (grant number: ZY2017304); the Post-Doctoral Research Project, West China Hospital, Sichuan University (grant numbers: 2018HXBH010 and 2018HXBH024); the China Postdoctoral Science Foundation (grant numbers: 2019M653418 and 2019M653428); and the Fostering Academic and Technical Leaders of Sichuan Province (No. [2016] 183-19).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2020.01214/full#supplementary-material

REFERENCES

1. Brien GL, Stegmaier K, Armstrong SA. Targeting chromatin complexes in fusion protein-driven malignancies. *Nat Rev Cancer.* (2019) 19:255–255 doi: 10.1038/s41568-019-0132-x
2. Koivunen JP, Mermel C, Zejnullahu K, Murphy C, Lifshits E, Holmes AJ, et al. EML4-ALK fusion gene and efficacy of an ALK kinase inhibitor in lung cancer. *Clin Cancer Res.* (2008) 14:4275–427 doi: 10.1158/1078-0432.CCR-08-0168
3. Stransky N, Cerami E, Schalm S, Kim JL, Lengauer C. The landscape of kinase fusions in cancer. *Nat Commun.* (2014) 5:4846. doi: 10.1038/ncomms5846
4. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* (2014) 513:202–20 doi: 10.1038/nature13480
5. Yao F, Kausalya JP, Sia YY, Teo AS, Lee WH, Ong AG, et al. Recurrent fusion genes in gastric cancer: CLDN18-ARHGAP26 induces loss of epithelial integrity. *Cell Rep.* (2015) 12:272–272 doi: 10.1016/j.celrep.2015.06.020
6. Shu Y, Zhang W, Hou Q, Zhao L, Zhang S, Zhou J, et al. Prognostic significance of frequent CLDN18-ARHGAP26/6 fusion in gastric signet-ring cell cancer. *Nat Commun.* (2018) 9:2447. doi: 10.1038/s41467-018-04907-0
7. Nakayama I, Shinozaki E, Sakata S, Yamamoto N, Fujisaki J, Muramatsu Y, et al. Enrichment of CLDN18-ARHGAP fusion gene in gastric cancers in young adults. *Cancer Sci.* (2019) 110:1352–135 doi: 10.1111/cas.13967
8. Tanaka A, Ishikawa S, Ushiku T, Yamazawa S, Katoh H, Hayashi A, et al. Frequent CLDN18-ARHGAP fusion in highly metastatic diffuse-type gastric cancer with relatively early onset. *Oncotarget.* (2018) 9:29336–293 doi: 10.18632/oncotarget.25464
9. Yang H, Hong D, Cho SY, Park YS, Ko WR, Kim JH, et al. RhoGAP domain-containing fusions and PPAPDC1A fusions are recurrent and prognostic in diffuse gastric cancer. *Nat Commun.* (2018) 9:4439. doi: 10.1038/s41467-018-06747-4
10. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* (2009) 339:b2535. doi: 10.1136/bmj.b2535
11. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol.* (2010) 25:603–60 doi: 10.1007/s10654-010-9491-z
12. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. *Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019).* Cochrane (2019). Available online at: https://training.cochrane.org/handbook
13. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. *BMC Med Res Methodol.* (2005) 5:13. doi: 10.1186/1471-2288-5-13
14. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell.* (2018) 173:400–400tics doi: 10.1016/j.cell.2018.02.052
15. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials.* (2007) 8:16. doi: 10.1186/1745-6215-8-16
16. Amin E, Jaiswal M, Derewenda U, Reis K, Nouri K, Koessmeier KT, et al. Deciphering the molecular and functional basis of RHOGAP family proteins: a systematic approach toward selective inactivation of RHO family proteins. *J Biol Chem.* (2016) 291:20353–203 doi: 10.1074/jbc.M116.736967
17. Lundmark R, Doherty GJ, Howes MT, Cortese K, Vallis Y, Parton RG, et al. The GTPase-activating protein GRAF1 regulates the CLIC/GEEC endocytic pathway. *Curr Biol.* (2008) 18:1802–18 doi: 10.1016/j.cub.2008.10.044
18. Regev M, Sabanay H, Kartvelishvily E, Kam Z, Bershadsky AD. Involvement of Rho GAP GRAF1 in maintenance of epithelial phenotype. *Cell Adh Migr.* (2017) 11:367–367 doi: 10.1080/19336918.2016.1227910
19. Cheng X, Yu S, Wang Y, Cui Y, Li W, Yu Y, et al. The role of oxaliplatin in the adjuvant setting of different Lauren's type of gastric adenocarcinoma after D2 gastrectomy: a real-world study. *Gastric Cancer.* (2019) 22:587–587 doi: 10.1007/s10120-018-0895-x
20. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature.* (2019) 569:503–50 doi: 10.1038/s41586-019-1186-3
21. Li H, Ning S, Ghandi M, Kryukov GV, Gopal S, Deik A, et al. The landscape of cancer cell line metabolism. *Nat Med.* (2019) 25:850–850 doi: 10.1038/s41591-019-0404-8
22. Hidalgo M, Amant F, Biankin AV, Budinska E, Byrne AT, Caldas C, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.* (2014) 4:998–998er doi: 10.1158/2159-8290.CD-14-0001
23. Tuveson D, Clevers H. Cancer modeling meets human organoid technology. *Science.* (2019) 364:952–95 doi: 10.1126/science.aaw6985

24. Kondo J, Inoue M. Application of cancer organoid model for drug screening and personalized therapy. *Cells.* (2019) 8:470. doi: 10.3390/cells8050470
25. Yan HHN, Siu HC, Law S, Ho SL, Yue SSK, Tsui WY, et al. A comprehensive human gastric cancer organoid biobank captures tumor subtype heterogeneity and enables therapeutic screening. *Cell Stem Cell.* (2018) 23:882–882 Ste doi: 10.1016/j.stem.2018.09.016
26. Nakayama I, Shinozaki E, Mashima T, Kawaguchi T, Sakata S, Yamamoto N, et al. Abstract 1954: functional analyses of CLDN18-ARHGAP26 fusion gene in gastric cancers. *Cancer Res.* (2018) 78(13 Suppl.):1954. doi: 10.1158/1538-7445.AM2018-1954
27. Rohde C, Yamaguchi R, Mukhina S, Sahin U, Itoh K, Tureci O. Comparison of Claudin 18.2 expression in primary tumors and lymph node metastases in Japanese patients with gastric adenocarcinoma. *Jpn J Clin Oncol.* (2019) 49:870–87 doi: 10.1093/jjco/hyz068
28. Sahin U, Koslowski M, Dhaene K, Usener D, Brandenburg G, Seitz G, et al. Claudin-18 splice variant 2 is a pan-cancer target suitable for therapeutic antibody development. *Clin Cancer Res.* (2008) 14:7624–762 doi: 10.1158/1078-0432.CCR-08-1547
29. Singh P, Toom S, Huang Y. Anti-claudin 18.2 antibody as new targeted therapy for advanced gastric cancer. *J Hematol Oncol.* (2017) 10:105. doi: 10.1186/s13045-017-0473-4
30. Tureci O, Sahin U, Schulze-Bergkamen H, Zvirbule Z, Lordick F, Koeberle D, et al. A multicentre, phase IIa study of zolbetuximab as a single agent in patients with recurrent or refractory advanced adenocarcinoma of the stomach or lower oesophagus: the MONO study. *Ann Oncol.* (2019) 30:1487–148 doi: 10.1093/annonc/mdz199
31. Sahin U, Schuler M, Richly H, Bauer S, Krilova A, Dechow T, et al. A phase I dose-escalation study of IMAB362 (Zolbetuximab) in patients with advanced gastric and gastro-oesophageal junction cancer. *Eur J Cancer.* (2018) 100:17–170 doi: 10.1016/j.ejca.2018.05.007
32. Guarnerio J, Bezzi M, Jeong JC, Paffenholz SV, Berry K, Naldini MM, et al. Oncogenic role of fusion-circRNAs derived from cancer-associated chromosomal translocations. *Cell.* (2016) 165:289–289l doi: 10.1016/j.cell.2016.03.020
33. Oncogenic circular RNAs arise from chromosomal translocations. *Cancer Discov*. (2016) 6:OF20. doi: 10.1158/2159-8290.CD-RW2016-068
34. Tan S, Sun D, Pu W, Gou Q, Guo C, Gong Y, et al. Circular RNA F-circEA-2a derived from EML4-ALK fusion gene promotes cell migration and invasion in non-small cell lung cancer. *Mol Cancer.* (2018) 17:138. doi: 10.1186/s12943-018-0887-9
35. Wu K, Liao X, Gong Y, He J, Zhou JK, Tan S, et al. Circular RNA F-circSR derived from SLC34A2-ROS1 fusion gene promotes cell migration in non-small cell lung cancer. *Mol Cancer.* (2019) 18:98. doi: 10.1186/s12943-019-1028-9
36. Tan S, Gou Q, Pu W, Guo C, Yang Y, Wu K, et al. Circular RNA F-circEA produced from EML4-ALK fusion gene as a novel liquid biopsy biomarker for non-small cell lung cancer. *Cell Res.* (2018) 28:693–5. doi: 10.1038/s41422-018-0033-7

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

frontiers in Genetics

METHODS
published: 25 November 2020
doi: 10.3389/fgene.2020.553587

Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach

Narjes Rohani[1] *and Changiz Eslahchi*[1,2]*

[1] Department of Computer and Data Sciences, Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran, [2] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Cancer is a complex disease with a high rate of mortality. The characteristics of tumor masses are very heterogeneous; thus, the appropriate classification of tumors is a critical point in the effective treatment. A high level of heterogeneity has also been observed in breast cancer. Therefore, detecting the molecular subtypes of this disease is an essential issue for medicine that could be facilitated using bioinformatics. This study aims to discover the molecular subtypes of breast cancer using somatic mutation profiles of tumors. Nonetheless, the somatic mutation profiles are very sparse. Therefore, a network propagation method is used in the gene interaction network to make the mutation profiles dense. Afterward, the deep embedded clustering (DEC) method is used to classify the breast tumors into four subtypes. In the next step, gene signature of each subtype is obtained using Fisher's exact test. Besides the enrichment of gene signatures in numerous biological databases, clinical and molecular analyses verify that the proposed method using mutation profiles can efficiently detect the molecular subtypes of breast cancer. Finally, a supervised classifier is trained based on the discovered subtypes to predict the molecular subtype of a new patient. The code and material of the method are available at: https://github.com/nrohani/MolecularSubtypes.

Keywords: cancer molecular subtypes, breast cancer, machine learning, somatic mutations, clustering, tumor classification

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong, Hong Kong

Reviewed by:
Rodrigo Gularte Mérida,
Cornell University, United States
Wenji Ma,
Columbia University, United States

***Correspondence:**
Changiz Eslahchi
ch-eslahchi@sbu.ac.ir

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

Received: *19 April 2020*
Accepted: *25 August 2020*
Published: *25 November 2020*

Citation:
Rohani N and Eslahchi C (2020) Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. Front. Genet. 11:553587. doi: 10.3389/fgene.2020.553587

1. INTRODUCTION

Breast cancer is a heterogeneous disease at the molecular and clinical levels; thus, the effectiveness of a treatment is hugely different based on the tumor characteristics. This heterogeneity is a challenge for tumor classification to reach an appropriate clinical outcome. To solve this problem, many researchers have developed numerous methods to classify tumor masses, such as histopathological classification based on the morphological characteristics or immunohistochemical (IHC) markers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Elston, 1999; Perou et al., 2000; Sørlie et al., 2001; Hu et al., 2006; Hofree et al., 2013; Ali et al., 2014; List et al., 2014). Moreover, Sorlie et al. have used hierarchical clustering on the gene expression data that led to the identification of significant breast cancer subtypes (Perou et al., 2000). The high cost of gene expression analysis for many genes was a significant obstacle in applying this method. To overcome this issue, the researchers have reduced the gene list to a relevant gene signature for breast cancer subtypes detection. Parker et al. (2009) have presented biomarker genes that can efficiently detect molecular subtypes. These genes could be an excellent

alternative to whole transcriptome microarray analysis. The subtypes found by these genes are known as *PAM*50 subtypes. Diversity of gene expression data in the subtypes is an indicator for the clinical prognosis of the patients, such as survival outcome (Sørlie et al., 2003).

In some studies, the microarray-based breast cancer classification has been considered as the gold standard (Peppercorn et al., 2007). However, the microarray-based methods cannot classify tumors consistently, due to the dynamic nature of gene expression data (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010).

Some studies have recently identified cancer subtypes based on somatic mutation profiles of tumors (Vural et al., 2016; Zhang et al., 2018b; Kuijjer et al., 2018). Somatic mutations are more stable and have critical functions in cancer development and progression (Vural et al., 2016; Kuijjer et al., 2018). Moreover, investigating somatic mutation profiles can aid in cancer diagnosis and treatment due to the vast number of clinical guidelines based on single gene mutation (Kuijjer et al., 2018). Therefore, the classification of cancers based on the mutation profiles can help identify subtypes of patients and their treatments (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010; Kuijjer et al., 2018). On the other side, with the development of new sequencing technologies, genome sequencing has become an appropriate tool for diagnostic purposes. Therefore, tumor classification based on somatic mutation profiles and application of the results in the clinical decisions can be crucial in the personalized medicine (Kuijjer et al., 2018).

Some studies have merged different kinds of the molecular data for breast cancer classification. Curtis et al. (2012) have developed a method to classify breast cancer by integrating genome and transcriptome data of 2,000 breast cancer patients. Based on the impact of somatic copy number alterations (CNAs) on the transcriptome, they have introduced new subtypes for breast cancer. Furthermore, Ali et al. (2014) have classified breast cancer into ten subtypes based on the combination of CNAs and gene expression data. In another study, List et al. (2014) have proposed a machine learning-based method that merges the gene expression and DNA methylation data for breast cancer classification. In a novel study, Hofree et al. (2013) have proposed a network stratification algorithm to classify tumors by fusing somatic mutation profiles with gene interaction network and have identified four subtypes for breast cancer. As somatic mutations are often sparse, it is sometimes challenging to predict cancer subtypes using somatic mutations. Therefore, previous studies have used other molecular information beside the somatic mutation data to detect cancer subtypes (Hofree et al., 2013).

In the most previous works, conventional clustering methods have been used to classify tumors; however, numerous innovative clustering methods have been proposed recently with various capabilities, which may help identify cancer subtypes. Moreover, the number of clusters typically has been determined using the silhouette criterion, which may lead to biologically meaningless clusters. In addition to the mentioned issues, the discovered clusters using somatic mutations are not analyzed extensively in previous works. In this study, the novel subtypes are presented using analysis of the somatic mutations and CNAs data from 861 breast tumors in the cancer genome atlas (TCGA) database (The International Cancer Genome Consortium, 2010). We used the network propagation method for smoothing somatic mutation profiles besides the gene interaction network; then, we used deep embedded clustering (DEC) (Xie et al., 2016) to find new breast cancer subtypes. Moreover, we used novel metrics such as AUMF (Maddi et al., 2019) and MMR (Brohee and Van Helden, 2006) for finding the best number of clusters. Afterward, the biological features of discovered subtypes were analyzed. Finally, a supervised model was trained to predict the breast cancer subtype of new patients. Also, the random forest (RF) was used to find the most important genes for classification.

2. MATERIALS AND METHODS

2.1. Extracting and Smoothing Data

We used somatic mutation profiles collected by Zhang et al. (2018b). They have obtained somatic mutation data of 861 breast tumors from TCGA. A gene is recognized altered if at least one of the following conditions satisfies:

- It has a non-silent somatic mutation.
- It is a well-defined oncogene or tumor suppressor.
- It happens within a CNA.

The somatic mutation profiles are sparse, that is, in each tumor, the number of mutated genes is relatively small compared to the total number of genes (Hofree et al., 2013; Zhang et al., 2018a). In most machine learning techniques, sparse data cannot train the model well (Zhang et al., 2018a), so data need to be smoothed. One of the most effective solutions for smoothing data is the network propagation (Hofree et al., 2013). By combining somatic mutation profiles and gene interaction networks, we can obtain profiles that are not sparse. Here, the protein–protein interaction (PPI) information in the STRING database (Szklarczyk et al., 2016) was used to create a gene interaction network. For this purpose, the *Homosapiens* PPI network was obtained from the STRING database. Then, the gene interaction network was created from the PPI network by mapping proteins to genes. The mutation profile of each tumor was integrated with the gene interaction network. In fact, the entire vertices of the network were labeled based on the mutation profile of each tumor. If a gene is mutated, the corresponding vertex is labeled one, and zero otherwise.

Then, in the network propagation process, a random walk with restart was applied on the networks as Equation (1).

$$D_{i+1} = \alpha D_i A + (1-\alpha) D_0, \quad i = 0, 1, 2, \ldots \tag{1}$$

The adjustment parameter α controls the amount of distance that a mutation can be propagated in the network. The optimal value of α varies for each network (in this study, it is subjectively set to 0.4). The network propagation process iterates until D_{i+1} is converged (i.e., $||D_{i+1} - D_i|| < 1 \times 10^{-6}$). D_0 is the original profile of tumor mutations, which is a $k \times n$ matrix (k is the number of tumors and n is the number of genes). D_i is the modified profile of mutations in the ith iteration. Matrix A is computed

by $A = H \times D$, where $H = [h_{ij}]$ is the adjacent matrix of the network and $D = [d_{ij}]$ is a diagonal matrix, such that:

$$d_{ij} = \begin{cases} \frac{1}{\sum_j h_{ij}} & \text{If } i = j \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

After the convergence, D_{i+1} was considered as the propagated mutation profile that has values between zero and one.

2.2. Clustering Method

To cluster propagated mutation profiles, we used DEC method (Xie et al., 2016). Suppose we have n tumors with the feature vectors x_i in space X with m dimension that should be grouped to k clusters with centers $\mu_j, j = 1, \ldots, k$. Instead of clustering the data in the initial space X, the data are mapped to the latent feature space Z by a nonlinear function $f_\theta : X \rightarrow Z$, where θ is a set of trainable parameters. Usually, in order to avoid the curse of dimensionality, the dimension of Z is less than m. A deep neural network can be used to implement f_θ, because of its theoretical function approximation characteristics (Hornik, 1991), and the capabilities in learning features (Bengio et al., 2013).

DEC is an iterative method, which learns cluster assignments and feature embedding simultaneously. In each iteration, the cluster centers $\{\mu_j \in Z\}_{j=1}^{k}$ as well as parameters θ are updated. This algorithm consists of two parts:

1. Parameter initialization using a stacked auto-encoder (SAE) (for θ) (Suk et al., 2015) and k-means algorithm (for centroids).
2. Parameter optimization that contains the alternative iteration of two steps: calculation of the auxiliary target distribution function, and updating the parameters using minimization of the Kullback–Leibler divergence (KLD).

In the initialization phase, the SAE is used to learn the feature embedding in an unsupervised manner. The SAE in this paper consists of two auto-encoders. Every auto-encoder has two layers as follows:

$$u = f(w_1(Dropout(x)) + b_1) y = g(w_2(Dropout(u)) + b_2) \quad (3)$$

where Dropout function (Baldi and Sadowski, 2013) randomly sets some of input elements to zero, f is the encoder function, g is the decoder function, w_i is the weight of ith layer, and b_i is the bias of ith layer. The parameter set $\theta = \{w_1, w_2, b_1, b_2\}$ is learned in order to minimize the loss function $||y - x||_2^2$. After learning the first auto-encoder, the output of encoder (u) is regarded as the input of the second auto-encoder. When the SAE was trained, the feature vector x_i could be embedded to the latent feature z_i by applying the first and second encoders on it.

Next, a clustering layer is added after the encoder layers to cluster the latent features. The cluster centers (μ_j) are initialized by running k-means on the latent features. The weights of the clustering layer were initialized by cluster centers.

In the optimization part, the latent features and clustering assignments are improved using alternating two following steps. In the first step, the latent feature (z_i) is softly assigned to cluster center (μ_j) with probability q_{ij}:

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2)^{-1}}{\sum_{j'} (1 + ||z_i - \mu_{j'}||^2)^{-1}} \quad (4)$$

In the second step, the KLD between soft assignment distribution (q_{ij}) and an auxiliary distribution (p_{ij}) is calculated.

$$KLD(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (5)$$

The auxiliary distribution is defined as:

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \quad (6)$$

where $f_j = \sum_i q_{ij}$ are the soft cluster frequencies. Then, the cluster center (μ_j) and latent feature (z_i) are updated in order to minimize the KLD using the stochastic gradient descent (Bottou, 2012).

These two steps are iterated until the convergence. The convergence criterion is satisfied when the assigned clusters to samples in two subsequent iterations are changed in <0.001 portion of data.

We tuned hyperparameters of the model, and the best number of neurons in the stacked auto-encoder layers was 514, 500, 200, 500, and 514, respectively. Moreover, the best number of neurons for clustering layer was found to be 4. The scheme of the method is presented in **Figure 1**. Also, the code and material of the method are available at: https://github.com/nrohani/MolecularSubtypes.

2.3. Finding the Best Number of Clusters

The clustering method requires the number of clusters (k) as the input. For selecting the best number of clusters, the clustering algorithm was implemented with different values of k. There are some appropriate criteria to compare results and choose the best number of clusters.

An approach to find the number of clusters is to evaluate the clustering based on microarray-based classes (PAM50) (Parker et al., 2009) as the prior information. For this purpose, a weighted bipartite graph G was formed, where the nodes of one part were the clusters of PAM50, represented by p_i symbols, and the nodes of another part were the discovered clusters, represented by c_j symbols. We weighted the edge (p_i, c_j), represented by v_{ij}, which shows the number of tumors shared between the clusters p_i and c_j. Moreover, the vertices p_i and c_j were labeled by the their sizes, represented by l_i and k_j, respectively. **Figure 2** shows the general scheme of such graph. After creating the graph, the following metrics were calculated in order to find the best number of clusters:

$$PPV = \frac{\sum_{j=1}^{K} \max_i v_{ij}}{\sum_{i=1}^{L} \sum_{j=1}^{K} v_{ij}} \quad (7)$$

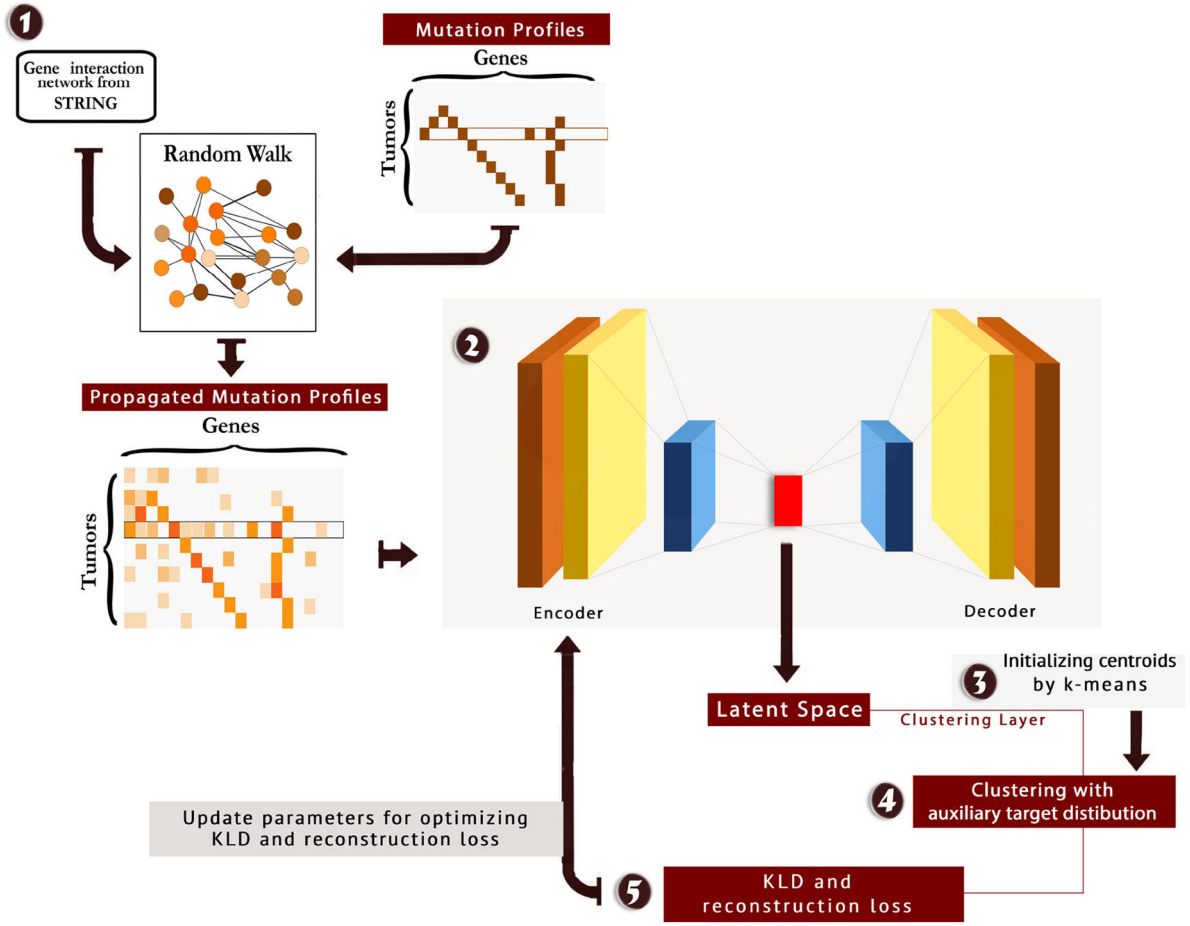


FIGURE 1 | The scheme of MSDEC (discovering Molecular Subtypes by using Deep Embedded Clustering). **1**: The gene interaction network is obtained from STRING, and the nodes in the network are labeled based on the mutation profile of tumors. Applying random walk on this network yields propagated mutation profiles. **2**: The propagated mutation profiles are mapped to a latent space. **3**: A clustering layer is appended after encoder layers. The cluster centroids are initialized using k-means. **4**: The tumors are clustered using the auxiliary target distribution. **5**: The Kullback–Leibler divergence (KLD) and reconstruction loss are calculated, and the parameters are updated to minimize KLD and reconstruction loss.

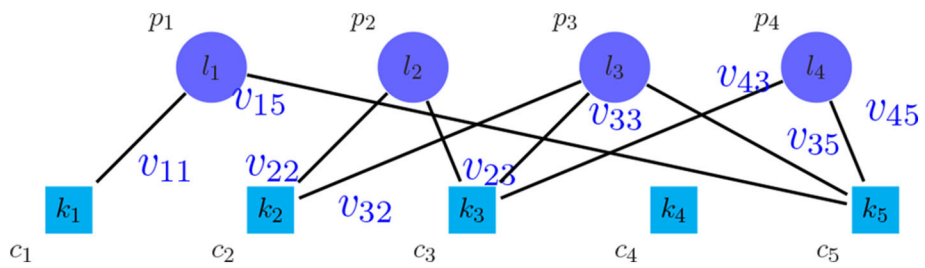


FIGURE 2 | Bipartite graph between the method to be evaluated and PAM50.

$$SN = \frac{\sum_{i=1}^{L} \max_j v_{ij}}{\sum_{i=1}^{L} l_i} \tag{8}$$

$$ACC = \sqrt{SN \times PPV} \tag{9}$$

Brohee and Van Helden (2006) have introduced these criteria. *ACC* is the geometric mean of PPV and SN; thus, it is more comprehensive than PPV and SN.

Another important criterion is the MMR (Brohee and Van Helden, 2006). For calculating this criterion, graph G was made, and the weights on the edges (v_{ij}) were calculated based on the threshold θ and the affinity score $NA(p_i, c_j)$ as follows:

$$v_{ij} = \begin{cases} NA(p_i, c_j) & NA(p_i, c_j) \geq \theta \\ 0 & (p_i, c_j) < \theta \end{cases} \tag{10}$$

$$NA(p_i, c_j) = \frac{|p_i \cap c_j|^2}{|p_i||c_j|} \tag{11}$$

MMR was defined as follows:

$$MMR = \frac{\sum_{v_{ij} \in Match_w(\mathcal{P},\mathcal{C},\theta)} v_{ij}}{|\mathcal{P}|} \tag{12}$$

where $Match_w(\mathcal{P}, \mathcal{C}, \theta)$ is the maximum weighted matching of G.

The discussed criteria compare the methods qualitatively. Another approach for comparison is the quantitative evaluation. We constructed a graph similar to the graph made for computing MMR. Then, we ignored the weight of the edges. Let $Match(\mathcal{P}, \mathcal{C}, \theta)$ to be the maximum non-weighted matching of this graph. Maddi et al. (2019) have introduced the following set of criteria:

$$N_p^+ = |\{p_i \mid \exists c_j,\ NA(p_i, c_j) \geq \theta,\ (p_i, c_j) \in Match(\mathcal{P}, \mathcal{C}, \theta)\}| \tag{13}$$

$$N_c^+ = |\{c_j \mid \exists p_i,\ NA(p_i, c_j) \geq \theta,\ (p_i, c_j) \in Match(\mathcal{P}, \mathcal{C}, \theta)\}| \tag{14}$$

$$Precision^+ = \frac{N_p^+}{|\mathcal{P}|} \tag{15}$$

$$Recall^+ = \frac{N_c^+}{|\mathcal{C}|} \tag{16}$$

$$F - measure^+ = \frac{2 \times Precision^+ \times Recall^+}{Precision^+ + Recall^+} \tag{17}$$

$F - measure^+$ is the harmonic mean of $Precision^+$ and $Recall^+$; thus, $F - measure^+$ is more meaningful than $Precision^+$ and $Recall^+$. All the mentioned criteria are in the $[0, 1]$ range.

One of the most comprehensive criteria in this issue is the AUMF (Maddi et al., 2019), which combines qualitative and quantitative attitudes. In fact, in this criterion the area under the curve $(MMR + Fmeasure^+, \theta)$ is considered as a clustering measure called AUMF, which is in the $[0, 2]$ range.

We executed DEC with the different numbers of clusters, and the results show that the best number of clusters is four (see **Supplementary Figures 1, 2**). Also, to evaluate the performance of the DEC method, this method was compared with other popular and common clustering methods such as hierarchical clustering (*HC*), k-means clustering, and spectral clustering (*SPC*) (Von Luxburg, 2007). DEC achieved better performance in comparison with other clustering methods.

2.4. Supervised Classification for New Tumors

Using the discovered breast cancer subtypes, we labeled each tumor with its discovered subtype and proposed a supervised classifier to understand how accurate the subtypes of new breast tumors can be predicted based on their somatic mutations. With this classifier, one can predict the subtype of a new patient using the somatic mutation profile as input. Five common machine learning classifiers were executed, namely, RF, support vector machine (SVM), multi-layer perceptron (MLP), naïve bayes (NB), and k-nearest neighbors (KNN) to classify the tumors into k subtypes $\{C_i\}_{i=1}^k$.

Due to the best results of RF (see section 3.6) in the supervised classification of tumors as well as its efficient application in feature selection, the RF was used to find important genes for classification. After training the RF, the importance of features can be calculated by considering the effect of using the features in reducing loss function (in this study, we used the Gini index as the loss function). In other words, the feature importance is the average reduction in loss function that induced by that feature. Then, the features with the importance of more than 0.01 were selected. The selected genes have the highest importance in detecting breast cancer subtypes.

3. RESULTS

After clustering tumors using MSDEC method, four clusters were obtained with the following sizes:

- Subtype 1 (*Primary* subtype): 182 tumors,
- Subtype 2 (*Progressive* subtype): 82 tumors,
- Subtype 3 (*Proliferous* subtype): 499 tumors,
- Subtype 4 (*Perilous* subtype): 98 tumors.

Figure 3 shows the illustration of the MSDEC subtypes. To visualize the tumors based on their mutation profile in a 2D space, we used principal component analysis (PCA) and obtained the first two principal components. Therefore, each tumor with a vector of length n representing the mutation status of the genes can be mapped to a 2D space using the first and second principal components. In **Figure 3**, the tumors are colored based on their assigned subtypes using MSDEC. It can be seen that the subtypes assigned by MSDEC are highly separable in this space. Precisely, all the tumors belonging to *Proliferous* subtype (green circles) are located at left, then *Primary* tumors (purple circles) are located at the right of them. The *Perilous* tumors are placed at the left side of *Primary* tumors. Moreover, *Progressive* tumors are settled at the right of the figure. The location of each subtype is specified and can be separated easily from the other subtypes. This figure shows that MSDEC subtypes have high separability.

To further investigate the discovered subtypes, we conducted the following evaluations.

3.1. Finding the Gene Signature for Each Subtype

One of the efficient evaluations is finding influential genes in each subtype. This evaluation is essential in two ways. First, it is possible to examine the biological significance of the clustering method; second, these genes can be considered as the candidates for the therapeutic purposes in each subtype's patients. For this purpose, the Fisher's exact test was used to find each subtype's gene signature. In the gene signature list, the top 50 genes with the p-value lower than 0.05 were considered and shown in **Supplementary Figures 3–6**. By investigating the top genes, one can conclude that the subtypes' key genes are different; thus, these genes can be suitable clues for choosing the treatment for

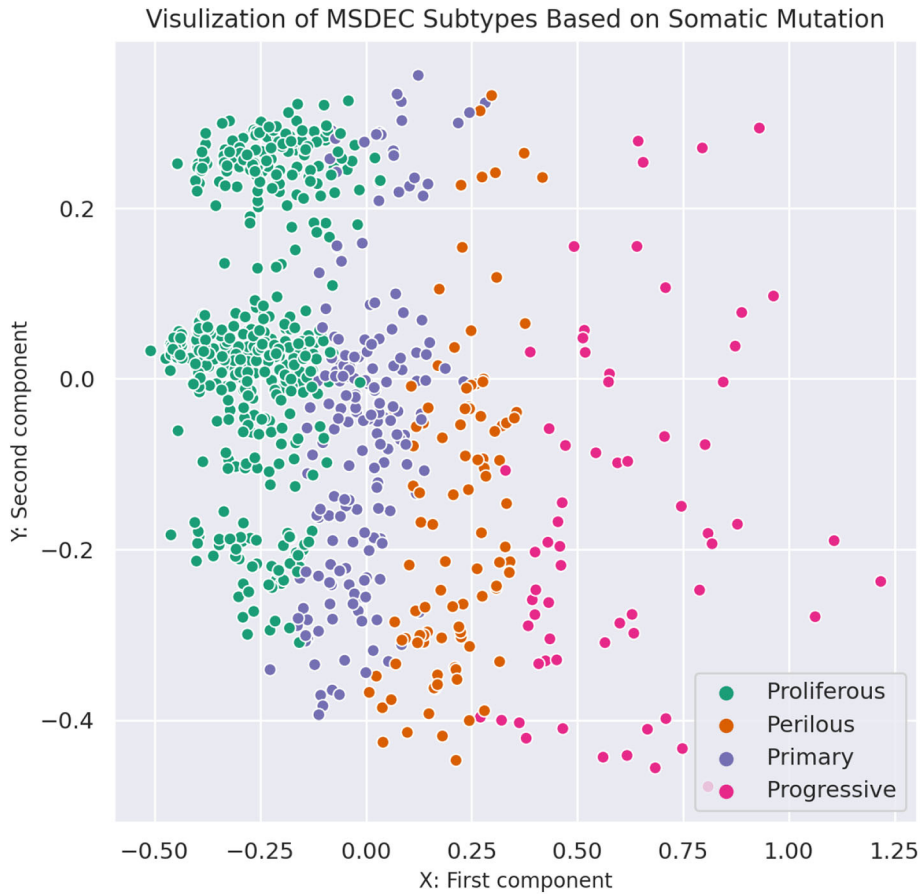


FIGURE 3 | Visualization of MSDEC subtypes based on the somatic mutation profile of tumors. The axes are the first two principal components of propagated mutation profiles.

the patients in each subtype. The gene interaction subnetwork of each subtype is obtained by enriching the subtype's gene signature into STRING database. The subnetwork of each subtype is illustrated in **Supplementary Figure 7**.

Many vital genes were found in the gene signature of the *Primary* subtype. One of them is *CDH*1, which produces E-cadherin protein. This protein is responsible for cell adhesion. Lacking E-cadherin allows the cancer cells to detach quickly and spread over the body and metastasize[1]. *CBFB* is another significant gene for *Primary* subtype. It encodes a transcription factor, which makes a complex by attaching to *RUNX*1[2]. This complex can transcriptionally repress the oncogenic *NOTCH* signaling pathway (Malik et al., 2019). *TBX*3 is a substantial gene in *Primary* subtype, which is needed for normal breast development[3]. Previous studies have shown that *TBX*3 leads to cell proliferation and suppresses apoptosis. *TBX*3 is regarded as a biomarker for breast cancer and has high importance in breast cancer diagnosis and treatment (Yarosh et al., 2008; Krstic et al., 2016). Another important gene in *Primary* subtype is *CTCF*, which encodes a transcription factor called zinc-finger. Studies have indicated that the mutation in *CTCF* is associated with the onset of breast cancer, prostate cancer, and Wilms' tumors (Oh et al., 2017), suggesting that this subtype mainly contains the tumors in early stages.

Many important genes such as *ERBB*2, *TP*53, *BRAF*, and *GNAS* are presented in the gene signature of the *Progressive* subtype. One of the driver genes in breast cancer is *ERBB*2, which is an indicator of tumor invasion (Revillion et al., 1998). Mutations and overexpression of this oncogene show the tendency of a tumor mass to become invasive, which may lead to the poor prognosis. The *BRAF* gene encodes a protein that helps transmit chemical signals from outside the cell to the cell's nucleus. This protein is responsible for regulating cell growth, proliferation, differentiation, migration, and apoptosis. Somatic mutations in this oncogene are prevalent in numerous cancers such as breast cancer, leading to the growth of cancerous cells[4]. The *TP*53 gene also is mutated in about 20 − −40% of breast cancer patients. It is useful to note that the mutation

[1]Genetics Home References, CDH1 gene, URL: https://ghr.nlm.nih.gov/gene/CDH1#normalfunction (accessed March 7, 2020).

[2]Genetics Home References, CBFB gene, URL: https://ghr.nlm.nih.gov/gene/CBFB#synonyms (accessed March 7, 2020).

[3]Genetics Home References, TBX3 gene, (Yarosh et al., 2008).

[4]Targeted Cancer Care, BRAF gene, URL: http://targetedcancercare.massgeneral.org/My-Trial-Guide/Diseases/Breast-Cancer/BRAF.aspx (accessed March 7, 2020).

frequency is higher in patients with recurrent breast cancer (Norberg et al., 2001). Another essential gene for *Progressive* subtype is *GNAS*. The *GNAS* gene encodes the stimulatory alpha subunit of the *G* protein complex, which triggers a complicated network of signaling pathways that affect multiple cell functions by regulating the activity of hormones. This gene is known to be mutated in 0.74% of all cancers such as breast invasive ductal carcinoma, colon adenocarcinoma, lung adenocarcinoma, and rectal adenocarcinoma, in which invasive breast carcinoma has the highest frequency of mutations[5]. Therefore, the *Progressive* subtype is more invasive because its significant genes are mostly mutated in invasive cancers. The probability of the poor prognosis and metastasis may be high in this subtype.

The *Proliferous* subtype contains many important genes, such as *NOTCH*, *KRAS*, *PTEN*, and *WHSC*1*L*1. The *NOTCH* family genes, including *NOTCH*1, *NOTCH*2, *NOTCH*3, and *NOTCH*4, are highly expressed in breast cancer patients. These genes play an important role in the differentiation, proliferation, and cell cycle (Wang et al., 2011). About 80% of cancers have estrogen receptors, which are treated with anti-estrogen drugs. One of the leading causes of death in such patients is their resistance to anti-estrogen drugs. Estrogen pathways have a positive association with anti-estrogen drug resistance in ER-positive breast cancers by suppressing *NOTCH*1 (Hao et al., 2010). The *KRAS* gene produces the $K-Ras$ protein, which affects cell proliferation, differentiation, and apoptosis[6]. The mutations of *KRAS* cause the production of abnormal $K-Ras$ protein that leads to uncontrolled cell proliferation. Somatic mutations in this oncogene are substantial in different cancers, including breast cancer, papillary thyroid carcinoma (PTC), oral squamous cell carcinoma (OSCC), and gastric cancer (Sanaei et al., 2017). *WHSC*1*L*1 provides instructions for making $histone-lysineN-methyltransferase$ NSD3 enzyme. It may involve in carcinogenesis, which is amplified in several cancers such as lung cancer and head and neck cancer[7]. Previous studies have suggested a close relation between *WHSC*1*L*1 mutation and breast cancer initiation and progression. The mutated *WHSC*1*L*1 is regarded as a candidate target for the treatment of breast cancer (Liu et al., 2015). *PTEN* gene encodes a tumor suppressor, which suppresses rapid and uncontrolled cell division. It also controls cell migration and adhesion. Somatic mutations of *PTEN* lead to the uncontrolled growth and division of cancerous cells. These mutations are involved in breast cancer (Zhang et al., 2013). Previous studies have shown that mutation in *PTEN* is a factor of resistance to trastuzumab (Herceptin) drug, which is used for the treatment of breast cancer[8].

Many essential genes are found among the gene signature of *Perlious* subtype such as *MYC*, *ITSN*1, *KDM*5*C*, and *TEP*1. One of the critical regulators of cell growth, proliferation, metabolism, differentiation, and apoptosis is *MYC*. Mutations of this gene have many roles in the development and progression of breast cancer, activation of oncogenes, and inactivation of tumor suppressors (Xu et al., 2010). *TEP*1 is one of the telomeres length genes that is linked with cancer (Pellatt et al., 2013). Previous studies have provided evidence for the relation of mutations in *TEP*1 and breast cancer (Savage et al., 2007). *ITSN*1 provides instructions for making a cytoplasmic membrane-associated protein. It is associated with the actin cytoskeleton reconstruction in breast cancer (Xie et al., 2019). *KDM*5*C* controls the transcription and chromatin remodeling regulation. TCGA has identified *KDM*5*C* mutation as a cancer driver mutation in the genes encoding the histone demethylases. Studies on oncometabolite have shown that the *KDM*5*C* is involved in cancer-related metabolic reprogramming and the tumor suppression (Chang et al., 2019). Thus, mutations of this oncogene are associated with tumor progression. It is mutated in 0.22% of all cancers, such as breast invasive ductal carcinoma, lung adenocarcinoma, prostate adenocarcinoma, and high-grade ovarian serous adenocarcinoma. Among these cancers, mutations of *KDM*5*C* are the most prevalent in invasive breast carcinoma[9].

3.2. Survival Analysis

We used Kaplan–Meier estimator (Kleinbaum and Klein, 2012) for survival analysis in each subtype, which is shown in **Figure 4**. The horizontal axis is the time after diagnosis, and the vertical axis represents the percentage of patients. The percentage of patients that are survived after specific days are plotted, and colored lines link the patients with the same subtype. The lower plot of survival demonstrates the more hazardous subgroup of people.

It was mentioned in section 3.1, that *Progressive* subtype is invasive, due to the set of significant genes in this subtype. This issue is consistent with survival analysis. It can be seen that the *Progressive* subtype has the lowest survival.

Moreover, the cox hazard regression was computed for further survival analysis. The diagram of cox hazard regression is presented in **Supplementary Figure 8**. To examine the significance of subtypes in predicting the patient's survival, chi-squared test was used, which shows that subtype is an essential feature in cox hazard regression ($p = 0.00475$). This analysis indicates that MSDEC subtypes have a significant correlation with the hazard rate.

3.3. Protein Complexes Analysis

We investigated the essential protein complexes in each subtype because most of the cell activities are carried out by protein complexes. The gene signature of each subtype was entered to the *iRefWeb* (Turner et al., 2010) website; then, the sorted complexes of each subtype were obtained (see **Supplementary Tables 1–4**). More information on these complexes is available in the *CORUM* database (Ruepp et al., 2009). **Figures 5A–D** visualizes five protein complexes in the *Primary*, *Progressive*, *Proliferous*,

[5]My Cancer Genome, GNAS gene, URL: https://www.mycancergenome.org/content/gene/gnas (accessed March 7, 2020).

[6]Genetics Home References, KRAS gene, URL: https://ghr.nlm.nih.gov/gene/KRAS (accessed March 7, 2020).

[7]Cancer Genetics Web, NSD3 gene, URL: http://www.cancerindex.org/geneweb/WHSC1L1.htm (accessed March 7, 2020).

[8]Genetics Home References, PTEN gene, URL: https://ghr.nlm.nih.gov/gene/PTEN#conditions (accessed March 7, 2020).

[9]My Cancer Genome, KDM5C gene, URL: https://www.mycancergenome.org/content/gene/kdm5c (accessed March 7, 2020).

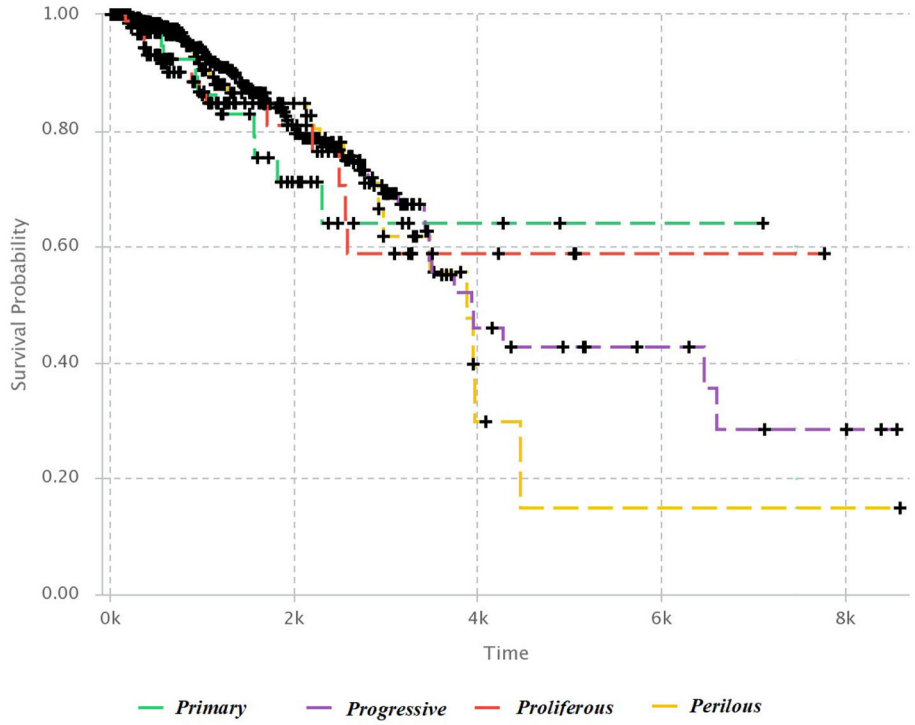


FIGURE 4 | Kaplan–Meier survival diagram. Each line shows the percentage of survived patients with a subtype of breast cancer after a specific time.

and *Perilous* subtypes, respectively. The nodes of these graphs represent the proteins that are involved in five complexes, which are obtained from *CORUM* database (Ruepp et al., 2009). The interactions between proteins were obtained from *STRING* database (Szklarczyk et al., 2016) and were shown by the edges in these graphs. The numbers beside the nodes represent the complexes that the protein are cooperating in them. Moreover, the nodes are colored based on their complexes.

One of the notable complexes in the *Primary* subtype is the $p27 - cyclinE - CDK2$ complex, which contains two *CDK*2 and *CDKN*1*B* genes. This complex is involved in cell cycle regulation, cell cycle control, and *DNA* processing. One of the crucial regulators of the cell cycle is *CDKN*1*B*, which inhibits $G1/S$ by clinging to *CDK*2 and suppressing it. Overexpression of *CDKN*1*B* gene in specific cancer cells prevents DNA replication and tumorigenesis, whereas its deficiency plays an inhibitory role in human cancers and decreases the chance for developing breast, prostate, colon, lung, and esophagus cancers (Xu et al., 2007).

BRCC complex includes the genes *BRCA*1, *BRCA*2, *BRCC*3, *RAD*51, and *BRE*, which is among the influential complexes in the *Progressive* subtype. The function of the *BRCA*1 gene in *DNA* repair and cell cycle control in response to *DNA* damage is regulated by other complexes. Interaction of *BRCA*1 with *RAD*51 has a direct impact on the double-strand breaks of *DNA* (Christou and Kyriacou, 2013). Not only has *ERCC* complex a direct interaction with *TP*53 in the destruction of *DNA*, but also it causes the displacement of *DNA*. Recently, the expressions of two new members of the complex, namely *BRCC*36 and *BRCC*45, have been discovered in breast cancer cells (Dong et al., 2003).

The set of *TBL*1*X*, *HDAC*3, and *NCOR*2 genes together make the *SMRT* complex, which plays a vital role in *Proliferous* tumors. The *SMRT* complex is both an activator and a suppressor of the estrogen receptor-α ($ER - \alpha$), which its overexpression in breast cancer can make therapeutic outcomes more complicated. The activity of this complex inhibits the regulated cell death using the genes involved in apoptosis. This complex activates the anti-apoptotic genes and suppresses the pro-apoptotic genes. Thus, by activating multiple pathways, this complex leads to the progression and proliferation of breast cancer with declining apoptosis (Blackmore et al., 2014).

*ESR*1 − *MDM*4 complex that is consisted of two genes *ESR*1 and *MDM*4 proteins is essential in the *Perlious* subtype. The estrogen hormone receptor *ESR*1 is a nuclear hormone receptor that is expressed in approximately 70% of patients with breast cancer (Stanford et al., 1986). The expression of *MDM*4 gene is positively correlated with the expression of *ER*α in primary breast tumors. Also, *ER*α enhances the expression of *MDM*2 (Baunoch et al., 1996).

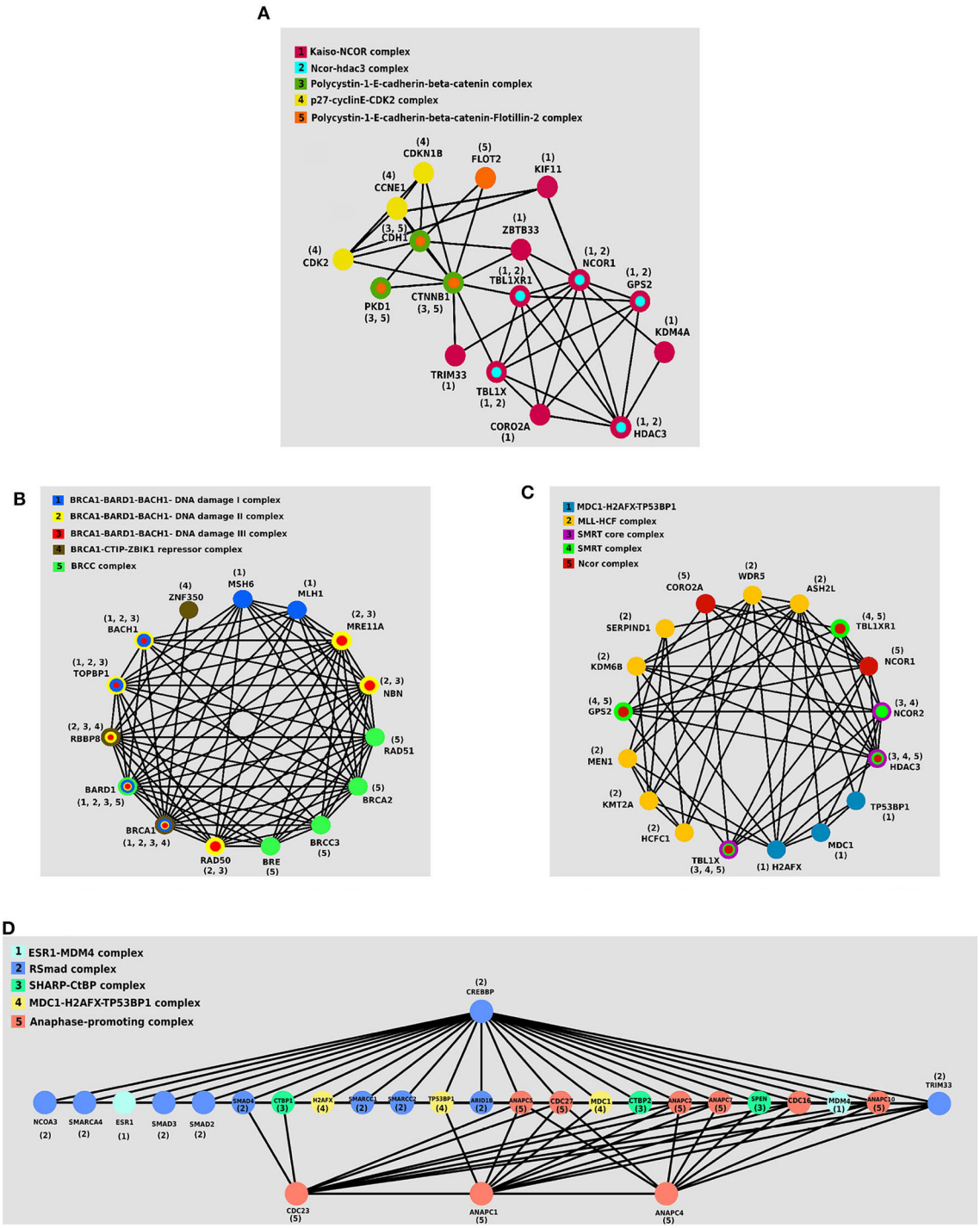


FIGURE 5 | The protein–protein interaction (PPI) networks of protein complexes in discovered subtypes. The proteins assigned to the same complex are shown with the same color and labeled with the same number. **(A)** Five protein complexes in *Primary* subtype. **(B)** Five protein complexes in *Progressive* subtype. **(C)** Five protein complexes in *Proliferous* subtype. **(D)** Five protein complexes in *Perilous* subtype.

3.4. Clinical Examination

We investigated the relationship between each subtype and the clinical features such as *ER* status, *PR* status, *HER2* status, TP53 status, and histopathological subtypes using the chi-squared test. The contingency tables of these analyses are shown in **Supplementary Figures 9–13**. The MSDEC

subtypes have a significant correlation with the mentioned clinical features.

Supplementary Figure 9 shows the relation of the *ER* status with the MSDEC subtypes ($p < 2.2E-16$ by chi-squared test and $p = 1E-06$ by Fisher's exact test). By considering the results of two tests, it can be concluded that the *ER* status of tumors is not significantly independent of the MSDEC subtypes. Thus, MSDEC subtypes are related to this clinical factor. Moreover, it can be seen that the majority of tumors in *Primary* and *Proliferous* subtypes are mostly ER-positive.

The contingency table in **Supplementary Figure 10**, shows the relationship of the *PR* status with MSDEC subtypes. The *p*-values of the chi-squared test and Fisher's exact test on this table were $2.2E-16$ and $1E-06$, respectively. Therefore, the MSDEC subtypes are not significantly independent of the *PR* status of patients. The rate of *PR* positive is higher than *PR* negative in the *Primary* and *Proliferous* subtypes, while most tumors in the *Progressive* and *Perlious* subtypes are *PR* negative.

The contingency table in **Supplementary Figure 11**, was constructed to examine the association of *HER*2 status with the MSDEC subtypes. The *p*-values of the chi-squared test and Fisher's exact test in this table were $1.445E-07$ and $1E-06$, respectively, which indicate a significant relationship between the clinical status of *HER*2 and the MSDEC subtypes. It can also be carefully deduced from this table that the *Primary* and *Proliferous* subtypes are significant *HER*2 negative.

The contingency table that indicates the relation of the *TP*53 status with MSDEC subtypes is shown in **Supplementary Figure 12**. The *p*-values of the chi-squared test and Fisher's exact test on this table were $2.2E-16$. Therefore, the MSDEC subtypes are not significantly independent of the *TP*53 mutations in patients. One of the interesting points in this table is the low rate of *TP*53 mutations in *Proliferous* and *Primary* subtypes, which indicates a noninvasive and better diagnostic status for *Primary* and *Proliferous* tumors. Thus, the *Primary* and *Proliferous* subtypes include tumors that have a better prognosis. In the *Progressive* and *Perilous* subtypes, the mutations pattern of *TP*53 is reversed, and its mutated state is more prevalent than its wild type.

We examined the association of the MSDEC subtypes with the histopathological subtypes. The distribution of these two variables in relation to each other is shown in **Supplementary Figure 13**, which has $p = 0.0001615$ by the chi-squared test and $p = 5.4E-05$ by the Fisher's exact test. As a result, there is strong evidence for the significant correlation between the two types of classification.

On the whole, the characteristics of the MSDEC subtypes can be summarized as follows.

Primary and *Proliferous* subtypes are consisted of tumors that are *ER*+ and *PR*+. The higher rate of *PR* positive than *PR* negative in the *Primary* and *Proliferous* subtypes indicate that most tumors in these two subtypes are *luminal* tumors. It can also be carefully deduced from the **Supplementary Figure 11** that the *Primary* and *Proliferous* subtypes are significantly negative for *HER*2. These tumors have wild-type *TP*53, and one of their most significant genes is *CDH*1.

Moreover, *Progressive* and *Perilous* subtypes mostly contain tumors that are *PR*−. *TP*53, *ERBB*2, *BRCA*1, and *MYC* are the significant genes in *Progressive* and *Perilous* subtypes. Mutations of the *BRCA*1 and *MYC* genes exacerbate breast cancer (Xu et al., 2010). Additionally, high rate of *TP*53 mutations in these subtypes suggest that the *Progressive* and *Perilous* subtypes may have poor diagnostic status.

3.5. Comparison Between MSDEC and PAM50 Subtypes

We compared the MSDEC subtypes from somatic mutation with PAM50 subtypes obtained from micro-array data; thus, the following evaluations were conducted to investigate their similarities and differences.

The contingency table in **Supplementary Figure 14** shows the intersection of tumors between the MSDEC subtypes and PAM50 subtypes. It is noteworthy that this table is not static since the assignment of tumors to PAM50 subtypes changes dynamically (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010; Vural et al., 2016). The dependency of these two clusterings was evaluated by using chi-squared test, which yielded $p < 2.2E-16$, and Fisher's exact test, which led to $p = 1E-06$. Moreover, the composition for each subtype with ER+/–, PR+/–, HER2+/–, and TP53 (mutated/wild type), and the PAM50 is visualized in **Figures 6A,B**, respectively.

Among the PAM50 subtypes, *luminal A* and *luminal B* are *HER*2 negative and *ER* positive. These tumors have a good prognosis and long survival. These subtypes are most similar to *Primary* and *Proliferous* subtypes due to the status of *ER*, *HER*2, and based on their prognosis and survival. Moreover, *Primary* and *Proliferous* tumors have wild-type *TP*53. One of their most significant genes is *CDH*1, which is highly expressed in the *luminal A* and *luminal B* subtypes, while it has low activity in *HER*2 − *positive* and *basal* − *like* subtypes (Zaha et al., 2019). However, the higher rate of *PR* positive than *PR* negative in the *Primary* and *Proliferous* subtypes may differ from *LuminalB* tumors.

Moreover, *basal* − *like* and *HER*2 subtypes mostly contains tumors that are *PR*−, which suggest that these two subtypes are more similar to *Progressive* and *Perilous* tumors. *TP*53, *ERBB*2, *BRCA*1, and *MYC* are the significant genes in *Progressive* and *Perilous* subtypes. Mutations of the *BRCA*1 and *MYC* genes exacerbate breast cancer (Xu et al., 2010). The *MYC* gene is highly expressed in the *basal* − *like* subtype of breast cancer, which is being targeted for treatment in these patients. Given the poor diagnostic status and high rate of *TP*53 mutations in the *basal* − *like* and *HER*2 subtypes, one can conclude that the *Progressive* and *Perilous* subtypes are related to the *basal* − *like* and *HER*2 subtypes (Xu et al., 2010).

To sum up, the *Primary* and *Proliferous* mostly contain *luminal A* and *luminal B* tumors, while the majority of tumors in *Progressive* and *Perilous* subtypes are *HER*2 − *positive* and *basal* − *like*. It is noteworthy that although the majority of tumors in *Primary* and *Proliferous* are *luminal A* and *luminal B*, numerous *HER*2 − *positive* and *basal* − *like* tumors are

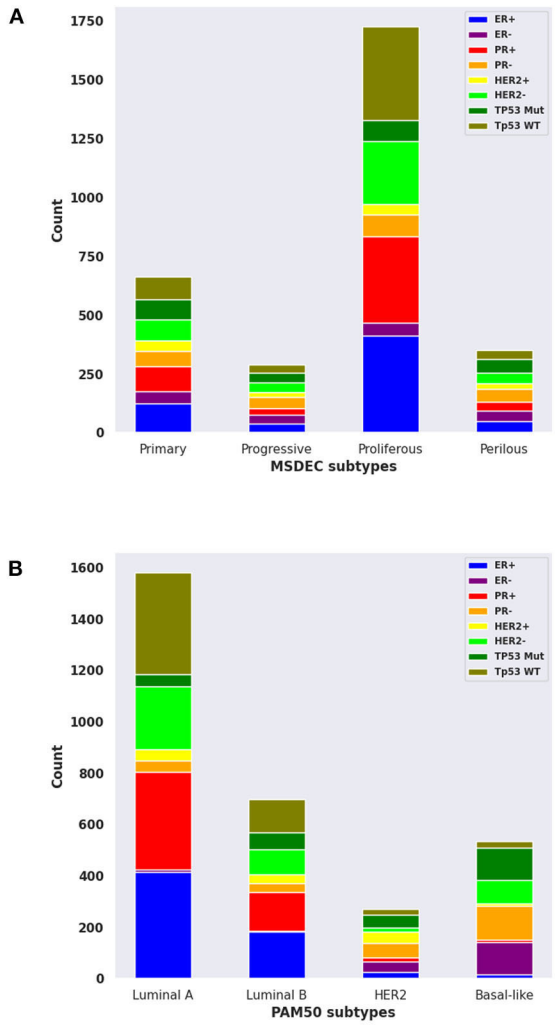


FIGURE 6 | The composition of ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in MSDEC and PAM50 subtypes. **(A)** The number of tumors with ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in each MSDEC subtypes. **(B)** The number of tumors with ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in each PAM50 subtypes.

included in these two subtypes. A similar issue is true for *Progressive* and *Perilous* subtypes. Thus, the MSDEC subtypes are not fully matched with PAM50 subtypes. It is worth mentioning that PAM50 subtypes were obtained by clustering microarray data, whereas the MSDEC subtypes are the results of clustering the mutation profiles. Since applying different unsupervised methods on different features yield different results, it is obvious that the MSDEC and PAM50 subtypes are not the same.

To compare the separability of subtypes identified by MSDEC and PAM50, we visualized the *PAM*50 subtypes in 2D space. To this aim, we used PCA to reduce the dimension of data and colored the tumors based on their subtypes. For the sake of simplicity in comparing subtypes identified by MSDEC and PAM50, we first applied PCA on the mutation profile of tumors, used the first two principal components to visualize the tumors, and colored them based on the PAM50 subtypes. **Figure 7A** shows the illustration of the PAM50 subtypes based on somatic mutation. One can figure out by the comparison of **Figures 3A**, **7** that the location of tumors are the same in these figures, while having different color scheme, one based on MSDEC and another based on PAM50 subtypes. In spite of **Figure 3** that shows high separation in the MSDEC subtypes, the PAM50 subtypes in **Figure 7A** do not have favorable separation and all the subtypes seems to be mixed up in 2D space. Moreover, since PAM50 is clustering tumors based on gene expression, we plotted the tumors on the 2D space based on the first two principal components of the gene expression profiles to have a fair notion of the visualization of PAM50 subtypes. **Figure 7B** shows the illustration of *PAM*50 clusters based on gene expression. Same as in **Figure 7A**, the other illustrations of PAM50 subtypes in **Figure 7B** does not demonstrate high separability.

Moreover, we computed the silhouette criterion for assessing MSDEC and PAM50 clustering quantitatively. The silhouette criterion measures the difference between the similarity of a tumor to its own cluster (cohesion) compared to its similarity to other clusters (separation). The value of this criterion ranges from -1 to $+1$. The higher the silhouette, the better tumors are matched to their own clusters rather than other clusters. For a tumor i in cluster C_k, the silhouette value is computed as formula 18.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{18}$$

where $a(i)$ and $b(i)$ are the cohesion and separation values for tumor i, which are calculated as follows:

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \neq i, j \in C_k} d(i,j) \tag{19}$$

$$b(i) = \min_{l \neq k} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j) \tag{20}$$

$d(i,j)$ is the Euclidean distance between tumors i and j. The silhouette criterion for a clustering method is computed by averaging the $s(i)$ values over all tumors. This criterion demonstrates that how tightly are the tumors in a cluster and how far are the tumors in diverse clusters. Therefore, this can be a measure for assessing the appropriateness of clustering methods. The computed silhouette criterion for MSDEC was 0.07011, while the computed silhouette criterion for PAM50 clusters based on gene expression and mutation profiles was 0.00956 and -0.00577, respectively. Comparison of the silhouette for MSDEC and PAM50 shows that MSDEC yields more appropriate subtypes.

3.6. Evaluation of Supervised Methods

Five classifiers, namely, RF, SVM, MLP, KNN, and NB, were compared using tenfold cross-validation. In tenfold cross-validation, the whole set of tumors was randomly divided into ten subsets with almost the same size. Then, one subset was

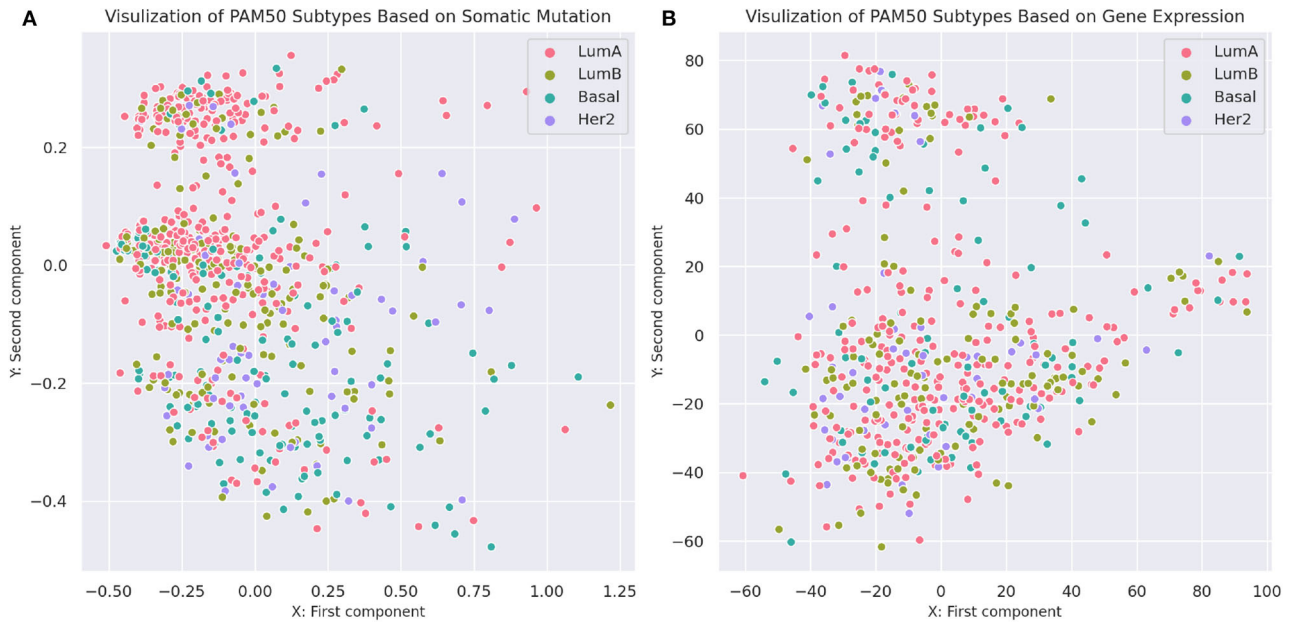


FIGURE 7 | Visualization of PAM50 subtypes based on somatic mutation and gene expression profiles of the tumors. **(A)** Visualization based on somatic mutation profiles. The axes are the first two principal components of the propagated mutation profiles. **(B)** Visualization based on gene expression profiles. The axes are the first two principal components of gene expression profiles.

put aside, and the model was trained with nine other subsets and evaluated with the remaining subsets. This process was repeated, such that each of the ten subsets was considered as the test data once. In this study, the tenfold cross-validation was repeated 100 times, and the average performance of the model was reported. The performance of the model was measured by standard evaluation criteria such as Accuracy, Sensitivity, Precision, F-measure, and AUC.

$$Accuracy = \frac{\sum_{i=1}^{k} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{k} \tag{21}$$

$$Precision = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} (TP_i + FP_i)} \tag{22}$$

$$Recall = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} (TP_i + FN_i)} \tag{23}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{24}$$

where TP_i, TN_i, FP_i, and FN_i stand for the number of True Positives, True Negatives, False Positives, and False Negatives of class $\{C_i\}_{i=1}^{k}$. Since the values of Accuracy, Precision, Recall, and F-measure are dependent on the value of a threshold, we also evaluated methods using AUC, which is the area under the receiver operating characteristic (ROC) curve. The ROC curve plots True Positive Rate (TPR) vs. False Positive Rate (FPR). For each class i, AUC_i is the area under the curve plotting TPR_i vs. FPR_i. Moreover, AUC for all classes is the area under the ROC curve of all classes, which is plotted with two approaches, namely, micro_average and macro_average. In micro_average, the ROC curve plots TPR_{micro} vs. FPR_{micro}, while in macro_average, the ROC curve plots TPR_{macro} vs. FPR_{macro}. AUC criterion indicates the efficiency of methods independent of the threshold value.

$$TPR_i = \frac{TP_i}{TP_i + FN_i} \tag{25}$$

$$FPR_i = \frac{FP_i}{FP_i + TN_i} \tag{26}$$

$$TPR_{macro} = \frac{\sum_{i=1}^{k} TPR_i}{k} \tag{27}$$

$$FPR_{macro} = \frac{\sum_{i=1}^{k} FPR_i}{k} \tag{28}$$

$$TPR_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FN_i} \tag{29}$$

$$FPR_{micro} = \frac{\sum_{i=1}^{k} FP_i}{\sum_{i=1}^{k} FP_i + TN_i} \tag{30}$$

According to **Supplementary Figure 15**, NB method has the worst performance, and SVM, KNN, and MLP have average performances. The best method with regard to all criteria is the RF with AUC of 99%, Accuracy of 86%, Precision of 90%, Recall of 85%, and F-measure of 87%, which has achieved great results. It can be concluded that the discovered subtypes by MSDEC method are separable; also, these subtypes can be predicted only by receiving mutations of 16 important genes for new tumors that were obtained using RF. The 16 important genes is as follows: *AKT2*, *CARD11*, *EIF4A2*, *FLNA*, *HNF1A*, *IDH2*, *LAMA1*, *LTBP1*, *MAP2K1*, *NCOR2*, *NOS2*, *PPP1R12A*, *PTPRU*, *SMC1A*, *TPR*, and

UPF3B. The mutational frequency of 16 important genes in each subtype is shown in **Supplementary Figure 16**. **Figure 8** shows the ROC curves of the RF classifier for each subtype. The value of *AUC* is excellent for each subtype and very close to one. However, the value of AUC for the *Proliferous* subtype is equal to one, which indicates that the model fits well on the tumors of the *Proliferous* subtype.

3.7. GSEA Enrichment

To find a family of genes that are related to cancer, we enriched the gene signature of each subtype (see **Supplementary Material**) by Gene Set Enrichment Analysis (GSEA) tool (Subramanian et al., 2005). We recognized that the most of these genes belong to transcription factor and protein kinase gene families, which are known to be associated with the progression of breast cancer. The results are described in **Supplementary Figures 17–20**. Besides, **Figure 9** shows the GSEA enrichment of 16 important genes, obtained using RF. It verifies that many of these genes are the most important genes in cancer.

4. DISCUSSION

Cancer is a heterogeneous disease; so, accurate classification of cancer is crucial to find the appropriate treatment. Recent

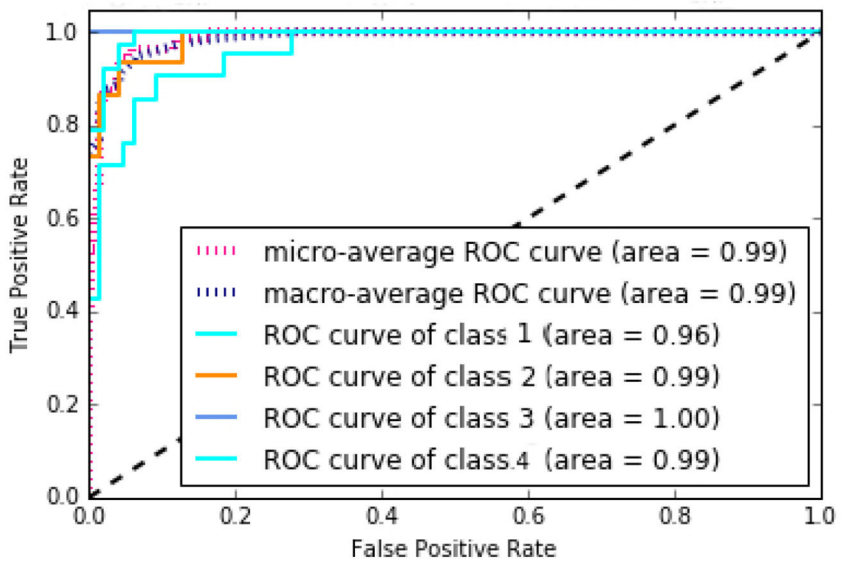


FIGURE 8 | Area under the ROC curves of the random forest (RF).

	cytokines and growth factors	transcription factors	homeodomain proteins	cell differentiation markers	protein kinases	translocated cancer genes	oncogenes	tumor suppressors
tumor suppressors	0	1	1	0	0	0	0	1
oncogenes	0	0	0	0	1	2	5	
translocated cancer genes	0	0	0	0	0	2		
protein kinases	0	0	0	0	2			
cell differentiation markers	0	0	0	0				
homeodomain proteins	0	1	1					
transcription factors	0	2						
cytokines and growth factors	1							

FIGURE 9 | GSEA enrichment of 16 important genes. The numbers show how many of important genes are incorporated in each family.

advances in molecular biology have provided high-quality and diverse data for the researchers. Recently, somatic mutation has attracted much attention in molecular cancer subtypes detection because it is more stable than other types of data and is commonly used for cancer treatment due to a large number of guidelines for single-gene mutations. In this study, the novel breast cancer molecular subtypes were presented using the profile of somatic mutations. Four discovered subtypes were obtained using network propagation with DEC. To analyze the characteristics of tumors in each subtype, we conducted numerous experiments, including finding gene signatures, protein complexes, gene families, and clinical features.

The results show that the *Primary* and *Proliferous* subtypes are mainly *ER*+, *PR*+, *HER*2−, and wild-type *TP*53; however, they have different important gene signature and protein complexes. Also, both of these subtypes contain the early stage and noninvasive tumors; the tumors in *Primary* have a higher probability of survival. Moreover, *Progressive* and *Perlious* subtypes are mainly *PR*− and have mutated *TP*53 gene. Numerous tumor suppressors and oncogenes were found in the gene signature of these two subtypes suggesting that these subtypes contain invasive tumors. It is noteworthy that these subtypes are different in terms of crucial protein complexes and gene signature. Moreover, the *Perlious* tumors have a lower probability of survival.

The RF classification algorithm was used for supervised classification to detect subtypes for new breast cancer patients. Also, 16 critical genes were identified using RF that can be used for detecting breast cancer subtypes of new tumors. Consequently, the MSDEC subtypes obtained from somatic mutations were clinically meaningful and provide an informative insight into molecular subtype diagnosis and suggesting efficient clues for cancer treatment.

For future research, we intend to use the proposed method to detect subtypes of other cancers, such as glioblastoma. Moreover, we aim to use other data such as gene expression and methylation features of tumors for finding more appropriate subtypes. Furthermore, we propose to examine the importance of each data in detecting cancer subtypes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/nrohani/MolecularSubtypes.

AUTHOR CONTRIBUTIONS

NR and CE conceived the analysis. NR implemented the method, calculated the results, and wrote the manuscript. CE helped to improve the paper. Both authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

All authors thank Farzaneh Rami and Fatemeh Ahmadi Moughari for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.553587/full#supplementary-material

REFERENCES

Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., et al. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* 15:431. doi: 10.1186/s13059-014-0431-1

Baldi, P., and Sadowski, P. J. (2013). "Understanding dropout," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 2814–2822.

Baunoch, D., Watkins, L., Tewari, A., Reece, M., Adams, L., Stack, R., et al. (1996). MDM2 overexpression in benign and malignant lesions of the human breast. *Int. J. Oncol.* 8, 895–899. doi: 10.3892/ijo.8.5.895

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Blackmore, J. K., Karmakar, S., Gu, G., Chaubal, V., Wang, L., Li, W., et al. (2014). The smrt coregulator enhances growth of estrogen receptor-α-positive breast cancer cells by promotion of cell cycle progression and inhibition of apoptosis. *Endocrinology* 155, 3251–3261. doi: 10.1210/en.2014-1002

Bottou, L. (2012). "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, eds G. Montavon, G. B. Orr and K. R. Müller (Berlin; Heidelberg: Springer), 421–436.

Brohee, S., and Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7:488. doi: 10.1186/1471-2105-7-488

Chang, S., Yim, S., and Park, H. (2019). The cancer driver genes IDH1/2, JARID1C/KDM5c, and UTX/KDM6A: crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism. *Exp. Mol. Med.* 51, 1–17. doi: 10.1038/s12276-019-0230-6

Christou, C., and Kyriacou, K. (2013). BRCA1 and its network of interacting partners. *Biology* 2, 40–63. doi: 10.3390/biology2010040

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Dong, Y., Hakimi, M.-A., Chen, X., Kumaraswamy, E., Cooch, N. S., Godwin, A. K., et al. (2003). Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in dna repair. *Mol. Cell* 12, 1087–1099. doi: 10.1016/S1097-2765(03)00424-6

Elston, C. W. (1999). Pathological prognostic factors in breast cancer. *Crit. Rev. Oncol. Hematol.* 31, 209–223. doi: 10.1016/S1040-8428(99)00034-7

Gusterson, B. (2009). Do'basal-like'breast cancers really exist? *Nat. Rev. Cancer* 9, 128–134. doi: 10.1038/nrc2571

Hao, L., Rizzo, P., Osipo, C., Pannuti, A., Wyatt, D., Cheung, L. W., et al. (2010). Notch-1 activates estrogen receptor-α-dependent transcription via ikkα in breast cancer cells. *Oncogene* 29, 201–213. doi: 10.1038/onc.2009.323

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96. doi: 10.1186/1471-2164-7-96

Kleinbaum, D. G., and Klein, M. (2012). "Kaplan-meier survival curves and the log-rank test," in *Survival Analysis*, eds M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis and W. Wong (New York, NY: Springer), 55–96.

Krstic, M., MacMillan, C. D., Leong, H. S., Clifford, A. G., Souter, L. H., Dales, D. W., et al. (2016). The transcriptional regulator TBX3 promotes progression from non-invasive to invasive breast cancer. *BMC Cancer* 16:671. doi: 10.1186/s12885-016-2697-z

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *Br. J. Cancer* 118, 1492–1501. doi: 10.1038/s41416-018-0109-7

List, M., Hauschild, A.-C., Tan, Q., Kruse, T. A., Baumbach, J., and Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinformatics* 11, 1–14. doi: 10.1515/jib-2014-236

Liu, L., Kimball, S., Liu, H., Holowatyj, A., and Yang, Z.-Q. (2015). Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. *Oncotarget* 6, 2466–2482. doi: 10.18632/oncotarget.2967

Maddi, A. M., Moughari, F. A., Balouchi, M. M., and Eslahchi, C. (2019). CDAP: An online package for evaluation of complex detection methods. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-49225-7

Malik, N., Yan, H., Moshkovich, N., Palangat, M., Yang, H., Sanchez, V., et al. (2019). The transcription factor CBFB suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* 10, 1–15. doi: 10.1038/s41467-019-10102-6

Norberg, T., Klaar, S., Lindqvist, L., Lindahl, T., Ahlgren, J., and Bergh, J. (2001). Enzymatic mutation detection method evaluated for detection of P53 mutations in cdna from breast cancers. *Clin. Chem.* 47, 821–828. doi: 10.1093/clinchem/47.5.821

Oh, S., Oh, C., and Yoo, K. H. (2017). Functional roles of CTCF in breast cancer. *BMB Rep.* 50, 445–453. doi: 10.5483/BMBRep.2017.50.9.108

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, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Pellatt, A. J., Wolff, R. K., Torres-Mejia, G., John, E. M., Herrick, J. S., Lundgreen, A., et al. (2013). Telomere length, telomere-related genes, and breast cancer risk: the breast cancer health disparities study. *Genes Chromos. Cancer* 52, 595–609. doi: 10.1002/gcc.22056

Peppercorn, J., Perou, C. M., and Carey, L. A. (2007). Molecular subtypes in breast cancer evaluation and management: divide and conquer. *Cancer Invest.* 26, 1–10. doi: 10.1080/07357900701784238

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W. F. (2006). Molecular classification of breast cancer: limitations and potential. *Oncologist* 11, 868–877. doi: 10.1634/theoncologist.11-8-868

Revillion, F., Bonneterre, J., and Peyrat, J. (1998). ERBB2 oncogene in human breast cancer and its clinical significance. *Eur. J. Cancer* 34, 791–808. doi: 10.1016/S0959-8049(97)10157-5

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., et al. (2009). Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* 38, D497–D501. doi: 10.1093/nar/gkp914

Sanaei, S., Hashemi, M., Eskandari, E., Hashemi, S. M., and Bahari, G. (2017). KRAS gene polymorphisms and their impact on breast cancer risk in an iranian population. *Asian Pac. J. Cancer Prevent.* 18, 1301–1305. doi: 10.22034/APJCP.2017.18.5.1301

Savage, S., Chanock, S., Lissowska, J., Brinton, L., Richesson, D., Peplonska, B., et al. (2007). Genetic variation in five genes important in telomere biology and risk for breast cancer. *Br. J. Cancer* 97, 832–836. doi: 10.1038/sj.bjc.6603934

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.191367098

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100

Stanford, J. L., Szklo, M., and Brinton, L. A. (1986). Estrogen receptors and breast cancer. *Epidemiol. Rev.* 8, 42–59. doi: 10.1093/oxfordjournals.epirev.a036295

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

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* 220, 841–859. doi: 10.1007/s00429-013-0687-3

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

The International Cancer Genome Consortium (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010:baq023. doi: 10.1093/database/baq023

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. doi: 10.1007/s11222-007-9033-z

Vural, S., Wang, X., and Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst. Biol.* 10:62. doi: 10.1186/s12918-016-0306-z

Wang, J., Fu, L., Gu, F., and Ma, Y. (2011). Notch1 is involved in migration and invasion of human breast cancer cells. *Oncol. Rep.* 26, 1295–1303. doi: 10.3892/or.2011.1399

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol.* 220, 263–280. doi: 10.1002/path.2648

Xie, C., Xiong, W., Li, J., Wang, X., Xu, C., and Yang, L. (2019). Intersectin 1 (ITSN1) identified by comprehensive bioinformatic analysis and experimental validation as a key candidate biological target in breast cancer. *OncoTargets Ther.* 12, 7079–7093. doi: 10.2147/OTT.S216286

Xie, J., Girshick, R., and Farhadi, A. (2016). "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning (Vienna)*, 478–487.

Xu, J., Chen, Y., and Olopade, O. I. (2010). MYC and breast cancer. *Genes Cancer* 1, 629–640. doi: 10.1177/1947601910378691

Xu, S., Abbasian, M., Patel, P., Jensen-Pergakes, K., Lombardo, C. R., Cathers, B. E., et al. (2007). Substrate recognition and ubiquitination of SCFSKP2/CKS1 ubiquitin-protein isopeptide ligase. *J. Biol. Chem.* 282, 15462–15470. doi: 10.1074/jbc.M610758200

Yarosh, W., Barrientos, T., Esmailpour, T., Lin, L., Carpenter, P. M., Osann, K., et al. (2008). TBX3 is overexpressed in breast cancer and represses P14ARF by interacting with histone deacetylases. *Cancer Res.* 68, 693–699. doi: 10.1158/0008-5472.CAN-07-5012

Zaha, D. C., Jurca, C. M., Bungau, S., Cioca, G., Popa, A., Sava, C., et al. (2019). Luminal versus non-luminal breast cancer CDH1 immunohistochemical expression. *Rev. Chim.* 70, 465–469. doi: 10.37358/RC.19.2.6936

Zhang, H.-Y., Liang, F., Jia, Z.-L., Song, S.-T., and Jiang, Z.-F. (2013). PTEN mutation, methylation and expression in breast cancer patients. *Oncol. Lett.* 6, 161–168. doi: 10.3892/ol.2013.1331

Zhang, W., Flemington, E. K., and Zhang, K. (2018a). Driver gene mutations based clustering of tumors: methods and applications. *Bioinformatics* 34, i404–i411. doi: 10.1093/bioinformatics/bty232

Zhang, W., Ma, J., and Ideker, T. (2018b). 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.

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