COMPUTATIONAL TOOLS IN INFERRING CANCER TISSUE-OF-ORIGIN AND MOLECULAR CLASSIFICATION TOWARDS PERSONALIZED CANCER THERAPY, VOLUME II

EDITED BY: Min Tang, Cheng Guo, Ling Kui and Jialiang Yang

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-408-7
DOI 10.3389/978-2-88971-408-7

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 TOOLS IN INFERRING CANCER TISSUE-OF-ORIGIN AND MOLECULAR CLASSIFICATION TOWARDS PERSONALIZED CANCER THERAPY, VOLUME II

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

Citation: Tang, M., Guo, C., Kui, L., Yang, J., eds. (2021). Computational Tools in Inferring Cancer Tissue-of-Origin and Molecular Classification Towards Personalized Cancer Therapy, Volume II. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-408-7

Table of Contents

frontiers in Genetics

EDITORIAL
published: 04 August 2021
doi: 10.3389/fgene.2021.735103

Editorial: Computational Tools in Inferring Cancer Tissue-of-Origin and Molecular Classification Towards Personalized Cancer Therapy, Volume II

Ling Kui[1,2,3†], *Wenzhe Mao*[1†], *Aasim Majeed*[4*] *and Jian Chen*[5*]

[1] *Shenzhen Qianhai Shekou Free Trade Zone Hospital, Shenzhen, China,* [2] *Harvard Medical School, Dana-Farber Cancer Institute, Brookline, MA, United States,* [3] *School of Pharmacy, Jiangsu University, Zhenjiang, China,* [4] *Molecular Genetics Laboratory, Department of Botany, Central University of Punjab, Bathinda, India,* [5] *International Genome Center, Jiangsu University, Zhenjiang, China*

Keywords: cancer, molecular classification, computational tools, machine learning, cancer therapeutics

Editorial on the Research Topic

Computational Tools in Inferring Cancer Tissue-of-Origin and Molecular Classification Towards Personalized Cancer Therapy, Volume II

With the advancement of sequencing technologies, there has been a rapid accumulation of data making it difficult to decipher the key genetic elements, their nature, and alterations for progression or regression of cancers. Further, enlightening an optimum gene set for optimal classification, diagnosis, and prognosis of cancer types is difficult without reliable and efficient computational tools. Progress in bioinformatics tools is parallel with sequencing data accumulation. Computational tools are being continuously refined and developed to meet specific challenges in cancer biology, the primary problem with them is that their accuracy is often insufficient for clinical use (Kui et al., 2021). As the molecular classification, appropriate diagnosis and prognosis, and markers for cancer improve better, more strategic novel drugs and efficient cancer treatments can be developed. Computation tools would play a great role in this direction.

In this editorial, we presented an account of how computational tools have greatly facilitated in unearthing the classification, prognosis, and therapeutic treatments of different cancer types. This editorial is based on 12 research articles, which sheds light on the power of computational tools to reveal the novel targets for cancer therapy and enhance the survival of patients with but not limited to glioblastoma, ciliated muconodular papillary tumors, adenosquamous carcinoma (ASC), breast cancer, esophageal cancer (EC), in colorectal cancer (CRC), metastatic melanoma, and multiple myeloma. The computational pipelines developed in these studies have a great potential to be extended to uncover novel therapeutic targets of other cancer types.

Four studies focused on the usage of specific computational approaches to address diverse problems. Chen et al. proposed to use machine learning algorithms to identify primary lesions for primary metastatic tumors. The fundamental idea behind their model is that different tumor types exhibit specific expression profiles for certain genes, which could be captured through machine learning models to classify the primary lesions. In essence, they used gene expression data from TCGA and GEO, analyzed and processed it to obtain a relatively suitable machine learning model followed by evaluation of the efficiency of diagnosis of primary lesions. They used XGBoost for classification and their results revealed that by combining tumor data with machine learning

OPEN ACCESS

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

***Correspondence:**
Aasim Majeed
majeedaasim@gmail.com
Jian Chen
jianchen0722@163.com

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

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

Received: *02 July 2021*
Accepted: *12 July 2021*
Published: *04 August 2021*

Citation:
Kui L, Mao W, Majeed A and Chen J (2021) Editorial: Computational Tools in Inferring Cancer Tissue-of-Origin and Molecular Classification Towards Personalized Cancer Therapy, Volume II. Front. Genet. 12:735103. doi: 10.3389/fgene.2021.735103

methods, the classification of different cancers can be achieved with specific accuracy, which can be used to predict the location of primary metastatic tumors. Cui et al. developed a novel pipeline, which not only compares two single-cell clusters but also calls for differential gene expression, coexpression network modules, etc. They used two single-cell data sets; Usoskin from the GEO database and Xin dataset of the human pancreas. Different types of analysis were then performed sequentially through a variety of computational tools to create a smooth pipeline. The pipeline implements DEsingle and SigEMD for differential gene expression analysis, DGCA for differential correlation analysis, WGCNA for network analysis, and DNA for differential network analysis. This pipeline is very effective to unravel the key differences between cell clusters and cell types and provides one place for easy computational analysis of single-cell data sets. Zhao et al. designed an Autoencoder-based computational framework, which could capture both intrinsic and extrinsic features of melanoma. They used the expression data of the TCGA metastatic melanoma gene RNA-seq dataset from Firehose and decomposed it into a small number of representative nodes. Further, microarray datasets from GEO for melanoma were used for prognosis analysis. They identified many nodes that were significantly associated with the prognosis of melanoma patients using Cox proportional hazard models. A tumor-intrinsic (TI) signature and a tumor-extrinsic (TE) signature were established from the two most prognostic nodes. Both these signatures highly predicted the patient's overall survival. In addition, the TE signature successfully predicted the response of patients to immunotherapy techniques. Using an integrative approach of somatic mutations and gene expression data, Jiang and Jin proposed a novel method for the identification of breast cancer-associated mutated genes. The fundamental theme behind their analysis is to first create a mutation matrix data and evaluating mutation frequency for each gene, then to create a gene expression matrix with expression values for each gene. Finally, both data sets are mapped to identify the co-expression profile. Their results indicated that this integrative approach is effective in breast cancer classification.

Two studies focused on the classification and prognosis of glioblastoma multiform (GBM), which lacks accurate prognostic markers and drug targets. Yuan et al. aimed to create a new molecular classification and to provide new therapeutic targets for GBM. They performed an integrated analysis based on the SNPs, DNA copy, DNA methylation, and mRNA expression profile data of 117 patients. The data was obtained from the TCGA database and Genomic Data Commons database (GDC). MutSigCV and GISTIC modules from GenePattern were used in the analysis of driver genes and landmark CNV events in GBM, respectively. Using the cluster of cluster analysis (CoCA), they found two novel subtypes, HX-1 and HX-2 depicting three variable methylation positions and fifteen gene mutations. These subtypes may act as potential prognostic biomarkers for patients with glioblastoma. Zhang et al. tried to classify glioblastoma subtypes on the basis of different degrees of gene methylation. They used the methylation datasets from Gene Expression Omnibus (GEO), identified the methylation loci, which served as potential biomarkers to classify and annotate the different GBM subgroups. They used powerful machine learning algorithms to achieve their goals. Monte Carlo feature selection (MCFS) and incremental feature selection (IFS) methods were used to extract 4,100 essential methylation sites and support vector machine (SVM), random forest (RF) metaclassifier, and repeated incremental pruning to produce error reduction (RIPPER) were used during classification. Functional enrichment analysis of these dysmethylated genes using GO and KEGG databases revealed several biological functions related to GBM classification.

Three studies aimed to reveal the novel prognosis-related signatures in different cancers. Esophageal cancer (EC) is a global fatal disease with a poor prognosis. Huang et al. aimed to evaluate the significance of genetic alteration (*CDK4* amplification) in the prognosis of esophageal squamous cell cancer (ESCC). Through tissue microarray and fluorescence *in situ* hybridization they found that among the investigated 520 patients with ESCC, 8.5% exhibiting *CDK4* amplification showed a negative correlation with disease progression and significantly better survival. Thus, they declared *CDK4* amplification as an independent prognostic factor for the survival of patients with ESCC. With the aim of identifying the novel DNA damage and repair-related prognostic genes in colorectal cancer (CRC), Wang et al. identified 1,545 genes related to DNA damage and repair. They used gene expression data of 471 COAD (Colon adenocarcinoma) and 41 normal samples from The Cancer Genome Atlas-Colon adenocarcinoma (TCGA-COAD) and 4 datasets of colon cancer from the GEO database. Following the gene set enrichment analysis (GSEA), the prognostic relevance of the individual genes was evaluated through Cox regression analysis on the TCGA-COAD dataset. A set of 12 genes related to DNA damage-and-repair were identified, which classified COAD patients into high and low-risk groups. Genes co-expressing with these 12 genes were identified through Pearson's correlation method. WGCNA with Topological Overlap Matrix (TOM) was used to construct the gene-coexpression network. Functional annotation of the functional gene modules was carried out using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). The gene set identified in this study has great potential in the prognosis, and treatment of CRC. Further, using The Cancer Genome Atlas (TCGA)-Multiple Myeloma Research Foundation (MMRF) dataset as a training dataset, Wang et al. analyzed the expression profiles through R package limma and evaluated the prognostic relevance of each gene through univariate Cox regression. Risk predictions were established through Lasso and stepwise Cox regressions followed by validation using GEO datasets. A set of eight RBP hub genes were identified, which classified multiple myeloma patients into high- and low-score groups. Functional analysis through Gene Ontology, KEGG Enrichment Analysis, and Gene Set Enrichment Analysis (GSEA) revealed that the major pathway through which RBP's could lead the development of myeloma may be the spliceosome pathway.

Two studies identified mutations in driving cancers. Yang et al. used the whole exon sequencing and immune checkpoint analysis of five patients with ciliated muconodular papillary tumors (CMPT) to elucidate the molecular details and histogenesis in CMPT. They observed 77 gene mutations in the patient's

tumor tissue and 31 mutations in the border tissue. Interestingly, CMPT shared the same phylogeny with cancer tissue. These results suggest the CMPT indeed are neoplastic processes with immune escape and have malignant potential. Recent studies have revealed that the clinical outcome of multiple cancers could be predicted through tumor mutational burden (TMB). To ascertain the relationship between TMB level and clinical features and outcomes of lung Adenosquamous carcinoma (ASC), Cheng et al. used NGS and immunohistochemistry approach and identified 95 unique genes with somatic variations from a total of 475 genes evaluated. TMB was found to be associated with pathological stages, invasion of lymph node, and overall survival but not with age, sex, smoking history, and tumor size in lung ASC. Moreover, no correlation between TMB and mutations in *TP53* and *EGFR* was observed. This study, therefore, provided an evidence that higher TMP correspond to lesser survival and higher lymph node invasion.

One study focused its interest on improving the existing medical imaging technology, which is a commonly useful approach in disease diagnosis and progression. With rapid advancements in deep learning, medical imaging technology has been revolutionized. Most medical imaging techniques involve encoder-decoder system, the classical architecture of which is implemented in U-Net. Several modified versions of U-Net have been introduced till now, all of which have two major limitations; loss of diversity features caused by fixed receptive field of the convolution kernel and loss of information when a single convolutional sequence is used in extracting features at each scale. With the aim of overcoming these limitations, Su et al. developed a new version of U-Net called multiscale U-Net (MSU-Net), which employed a new image segmentation architecture. It is based on Multi-scale blocks composed of convolution sequences with different receptive fields, which facilitates extraction of more information with diversified features. Their results showed that MSU-Net enabled significant improvement of semantic segmentation. MSU-Net integrates multiple convolution sequences having receptive fields of different sizes, which produces more conspicuous object features during forward propagation. Besides, MSU-Net is flexible enough to be integrated with other network structures. MSU-Net showed improved results, with 5-fold cross validation when applied on five biomedical image segmentation datasets; (1) 30 serial section Transmission Electron Microscopy (ssTEM) images (512 × 512 pixels) of the first instar larva ventral nerve cord (VNC) of the *Drosophila,* (2) The Breast Ultrasound Dataset B (BUL) comprising of 163 ultrasound images (760 × 570 pixels) of breast lesions, (3) 800 Chest X-ray (CXR) images (4,456 × 4,456 pixels) from the standard digital image database for Tuberculosis, (4) 2,594 RGB images of skin lesions (2,166 × 3,188 pixels), and (5) Nuclei Segmentation (NS) dataset from The Cancer Genome Atlas (TCGA) comprising of 30 digitized Hematoxylin and Eosin-stained frozen sections (512 × 512 pixels). This imaging technology may perform better in tracing, diagnosis, prognosis, and possible treatment of different cancer types.

AUTHOR CONTRIBUTIONS

This editorial was designed by LK and written by LK and AM. WM and JC revised the editorial. All authors made a direct and intellectual contribution to this topic and approved the article for publication.

FUNDING

This work was supported by the grant from Jiangsu University (No. 20JDG47) and Jiangsu University High-Level Talent Funding (No. 20JDG34).

REFERENCES

Kui, L., Guo, C., Li, S. C., Yang, J., and Tang, M. (2021). Computational methods in inferring cancer tissue-of-origin and cancer molecular classification. *Front. Genet.* 12:644542. doi: 10.3389/fgene.2021.644542

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 29 September 2020
doi: 10.3389/fgene.2020.579737

Ciliated Muconodular Papillary Tumors of the Lung: Distinct Molecular Features of an Insidious Tumor

*Xinxin Yang[1], Yunjing Hou[1], Jiashi Geng[2], Jingshu Geng[1] and Hongxue Meng[1]**

[1] Department of Pathology, Harbin Medical University Cancer Hospital, Harbin, China, [2] Department of Radiology, Harbin Medical University Cancer Hospital, Harbin, China

Introduction: Ciliated muconodular papillary tumors (CMPTs) are rare special peripheral pulmonary nodule composed of different cell proportions, characterized by papillary structures and significant alveolar mucus. Because of their rarity, underrecognized processes, the full range clinical course and histogenesis of CMPTs remains uncertain.

Methods: Molecular features of 5 CMPTs cases (one case with mucinous adenocarcinoma simultaneously) were observed by whole exon gene detection. The histological features of CMPTs and the development trends of three major constituent cells were studied by immunohistochemistry and PCR.

Results: NGS revealed 77 gene mutations in the patient's tumor tissue and 31 mutations in the border tissue. TMB of CMPT tends to TMB of cancer tissues, and both are higher than normal tissues, CMPT share the same phylogenetic tree with cancer tissues. Moreover, PDL1, B7H3, and B7H4 were overexpressed in high columnar cells and eosinophilic ciliated cells of CMPT, tends to cancer tissues, while LAG3 and siglec15 were not found in CMPT.

Conclusion: The high prevalence of driver gene mutations in CMPTs, similar TMB and phylogenetic tree with cancer tissues indicate their malignant potential. Distinct molecular and immune check point features of each component support the notion that ciliated columnar cells in CMPT are insidious with immune escape.

Keywords: ciliated muconodular papillary tumors, molecular analysis, histogenesis, immune escape, whole exon gene detection

OPEN ACCESS

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

Reviewed by:
Yuyan Cheng,
University of California, Los Angeles,
United States
Sha Li,
Sanford Burnham Prebys Medical Discovery Institute, United States
Meng Xu,
National Institutes of Health (NIH),
United States

****Correspondence:***
Hongxue Meng
menghongxue15@163.com

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

Received: *03 July 2020*
Accepted: *09 September 2020*
Published: *29 September 2020*

Citation:
Yang X, Hou Y, Geng J, Geng J and Meng H (2020) Ciliated Muconodular Papillary Tumors of the Lung: Distinct Molecular Features of an Insidious Tumor. Front. Genet. 11:579737. doi: 10.3389/fgene.2020.579737

INTRODUCTION

Ciliated muconodular papillary tumors (CMPTs) are rare peripheral pulmonary nodules characterized by papillary structures and significant alveolar mucus in different proportions. They are composed of a mixture of proliferating ciliated columnar cells, goblet cells, and basal cells surrounded by intra-alveolar mucin pools in the peripheral lung (Kamata et al., 2015;

Abbreviations: CMPTs, ciliated muconodular papillary tumors; FFPE, formalin-fixed paraffin-embedded; GGO, ground glass opaque; TMB, Tumor Mutational Burden; TTF-1, thyroid transcription factor-1.

Liu et al., 2016; Taguchi et al., 2017; Chang et al., 2018; Kataoka et al., 2018). Only about 70 cases have been reported worldwide, and the clinicopathological characteristics and histogenesis have not yet been defined in detail. One case of CMPT coexisting with mucinous adenocarcinoma was reported in our cases, it may be a basis for the malignant potential of a CMPT. Through the detection of immune checkpoints, perhaps we can find out the similarities between CMPT and immune escape of malignant tumors.

Recent genetic studies revealed mutations in some driver oncogenes (*BRAF, EGFR, KRAS, AKT1, or ALK*), and they supported the notion that the lesion tends to be a neoplastic lesion with malignant potential (Chuang et al., 2014; Kamata et al., 2015, 2016; Jin et al., 2017; Kim et al., 2017; Taguchi et al., 2017; Udo et al., 2017; Chang et al., 2018; Kataoka et al., 2018). In particular, mutations in *BRAF* (40%) and *EGFR* (30%), as identified by Kamata, support the development of a CMPT as a true tumor process rather than a response or metaplastic disease (Kamata et al., 2016; **Table 2**). Here, we performed whole exon gene sequencing and immune check point analysis on five CMPT patients to clarify the molecular features and histogenesis of each cell component in CMPT.

MATERIALS AND METHODS

Patients

This study was approved by the institutional review board of Harbin Medical University Cancer Hospital (Harbin, China). Five cases with characteristic features of CMPT were identified between 2016 and 2019. Their clinical and pathologic information were reviewed (**Table 1**). Tumor tissues and tissues adjacent to cancer were obtained by pathological sampling after surgery. In addition, border tissues beside tumor were enucleated by macrodissection under a stereo microscope.

Immunohistochemistry

Immunohistochemical analysis was performed on formalin-fixed paraffin-embedded (FFPE) sections (4 μm thick) using a fully automated system (Ventana Medical Systems, Tucson, AZ, United States). The slides were stained with antibodies against CK5/6 (clone CK5/6.007, ZSJ-bio, China), thyroid transcription factor-1 (TTF-1) (clone SPT24, Maixin, China), p40 (clone ZR8, Maixin, China), PD-L1 (clone sp22C3, Dako, Japan), B7H3 (Abcam, United States) and B7H4 (Abcam, United States).

Next-Generation Sequencing

Genomic DNA was sheared into fragments with the size of ∼200 bp. The adapters were added to both ends then were purified with Agencourt AMPure SPRI beads (Beckman Coulter, Inc., Brea, CA, United States). Ligation-mediated PCR was performed to amplify the extracted DNA. For enrichment the PCR products was hybridized to the SureSelect biotinylated RNA library (Agilent Technologies, Santa Clara, CA, United States) according to the manufacturer's instructions. Paired-end multiplex samples were sequenced with the Illumina HiSeq 2000 System. Sequencing depth was ∼100 × per sample.

PCR

Total RNA was extracted using an RNeasy Micro kit (Qiagen, Hilden, Germany), then treated according to the manufacturer's instructions. Complementary DNA (cDNA) was synthesized using a QuantiTect Reverse Transcription Kit (Qiagen). The PCR was performed using cDNA as a template. PCR products were analyzed by 4% agarose gel electrophoresis and stained by ethidium bromide. We used PCR to detect the expression of LAG3 and siglec15 in tissues.

RESULTS

Clinical Findings

The patients had a ground glass opacity (GGO) nodule in the lung by chest CT examination, and the size of the nodule in the five patients was generally less than 1 cm. The clinical staging (one case with mucinous adenocarcinoma simultaneously) of the five patients are both T1M0N0. One of them underwent pulmonary lobectomy, and the remaining four cases received wedge excision (**Table 1** and **Supplementary Figure S1**).

Histologic Findings

Microscopic observation of the tumor reveals a hyperplastic zone with unclear boundaries, and a mucus lake could be seen in the alveolar cavity. Three main components could be seen under a high powered microscope: basal cells, high columnar cells, and eosinophilic ciliated cells (**Figure 1**).

Immunohistochemical Findings

Transcription factor-1 of basal cells and columnar cells was stained, and these cells were stained more than eosinophilic ciliated cells. CK7 staining coloration was continuous in basal cells, and the Ki67 index was less than 5%. Five patients had a positive expression of B7H3, B7H4, PDL1, and OFD1, mainly in high columnar cells and eosinophilic ciliated cells. Among them, adenocarcinoma and CMPT coexisted in the tissue of patient 2, which showed high expression of PDL1 in the CMPT compared to the other four patients, which prompted the activity of immune escape **Figure 1** and **Table 2**).

Molecular Findings

Molecular analysis of NGS revealed 77 gene mutations in the patient's tumor tissue and 31 mutations in the border tissue (**Supplementary Figure S2**). We performed a functional enrichment analysis of the tumor variant gene. According to the functional enrichment analysis (**Supplementary Figure S3**), the first three enriched signal pathways were (1) negative regulation of apoptosis, (2) processing of O-glycogen, and (3) positive regulation of GTPase activity. Both (1) and (3) are associated with excessive proliferation of cells. Furthermore, there were six genes (*EGR1, MUC20, MUC3A, NBPF19, NOL4L,* and *OR4L1*) that were simultaneously mutated in the tumor tissues and junction tissues (**Supplementary Figure S4** and **Table 3**). By analyzing the evolutionary relationship of the taxa, it can be seen that there are three pairs of genes in

TABLE 1 | Characteristics of CMPT patients.

Number	Age	Smoking	Family history	Location	CT finding	Size (mm)	Treatment
1	40–45	−	+	RUL	Ground glass opaque (GGO) nodule	10	Wedge excision
2	60–65	−	−	RLL	Ground glass opaque (GGO) nodule	10	pulmonary lobectomy
3	60–65	+	−	RLL	Ground glass opaque (GGO) nodule	7	Wedge excision
4	60–65	−	−	LLL	Ground glass opaque (GGO) nodule	8	Wedge excision
5	55–60	−	−	RLL	Ground glass opaque (GGO) nodule	10	Wedge excision

RUL, Right upper lobe; RLL, Right lower lobe; and LLL, Left lower lobe.

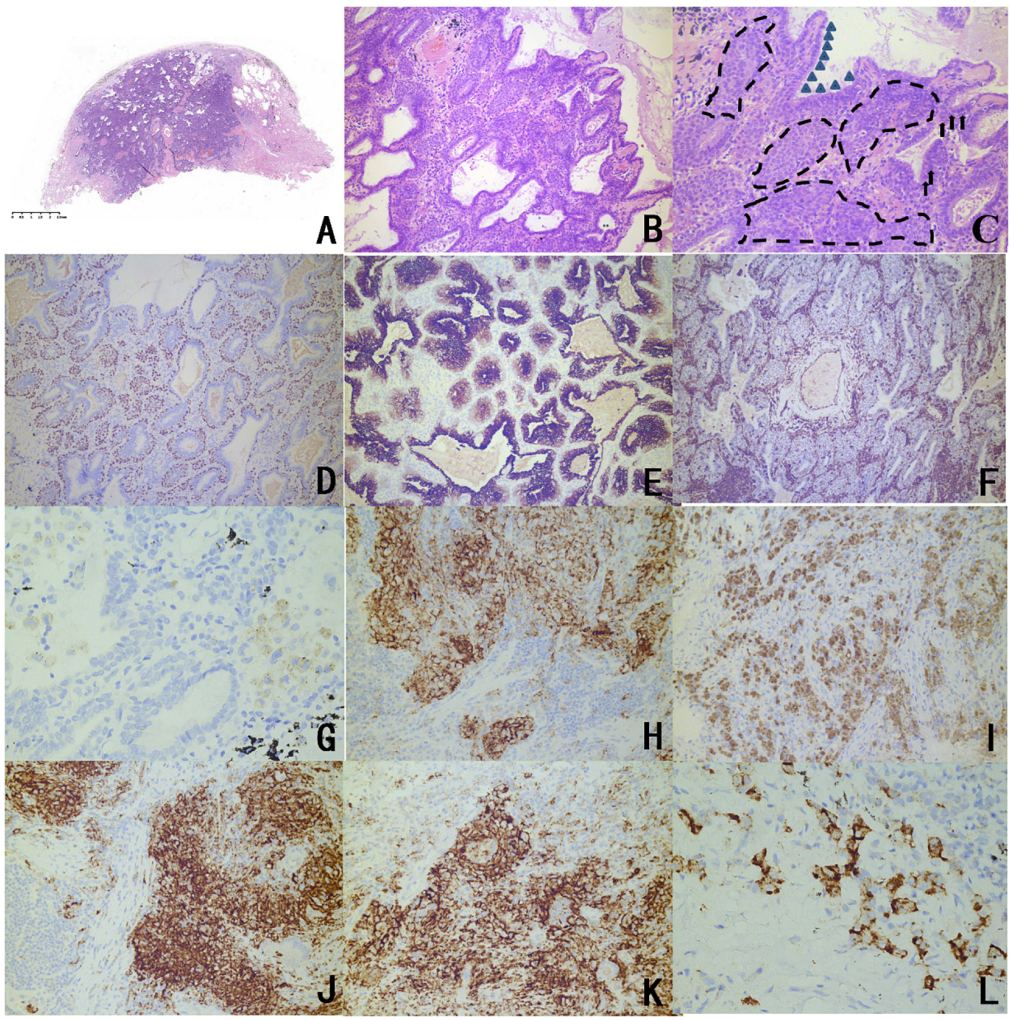

FIGURE 1 | Representative histopathological findings and immunophenotype of CMPTs. **(A)** Microscopic observation revealed a hyperplastic zone with unclear boundaries, and a mucus lake in the alveolar cavity (H&E staining). **(B)** A tubular papillary growth was observed under a medium magnification microscope with chronic inflammatory cell infiltration. **(C)** The microscope was highly magnified, and three main components can be seen in the lesion: basal cells, high columnar cells, and eosinophilic ciliated cells. The triangles are marked as high columnar cells, the arrows are marked mucous cells, and the dashed lines are marked as basal cells. **(D)** TTF-1 of basal cells and columnar cells was stained, and stained stronger than eosinophilic ciliated cells according to immunohistochemistry. **(E)** CK7 staining showed continuous coloring in the basal cells surrounding the adenoid structure and the papillary structure. **(F)** The Ki67 index was less than 5%. Immunohistochemistry for panel **(G)** PDL1 (negative in normal issue); **(H)** PDL1 (positive mainly on high columnar cells and eosinophilic ciliated cells in CMPT); and **(I)** PDL1 (positive in carcinoma). Immunohistochemical analyses of panel **(J)** B7H3, **(K)** B7H4, and **(L)** OFD1 in the CMPTs.

the same branch of the phylogenetic tree in the CMPT and adenocarcinoma tissues for patient 2. By comparing the TMB in normal tissues and CMPT and adenocarcinoma tissues, it can be seen that the TMB of CMPT is similar to the TMB of cancer tissues, and both are higher than the TMB of normal tissues (**Figure 2**).

TABLE 2 | Summary of immunohistochemical findings from previous reports and the present cases.

Authors	CK5/6	TTF-1	Ki-67	CK7	CK20	MUC5AC	MUC5B	MUC6	MUC2	P53	HNF4α	P63/P40	CDX2	CEA	MUC1	CA125	PDL1	B7H3	B7H4	OFD1
Taguchi et al., 2017	–	0/1	3.7%	1/1	–	1/1	–	1/1	1/1	1/1	–	–	–	1/1	–	–	–	–	–	–
Kataoka et al., 2018	–	4/4	<5%	–	–	3/4	–	2/4	0/4	–	0/4	–	–	–	4/4	4/4	–	–	–	–
Chang et al., 2018	+[a]	19/25	–	–	–	–	–	–	–	–	–	+[a]	–	–	–	–	–	–	–	–
Kamata et al., 2015)	–	4/4	–	–	–	–	–	–	–	–	–	4/4	–	–	–	–	–	–	–	–
Jin et al., 2017	1/1	1/1	–	1/1	–	–	–	–	–	–	–	1/1	–	–	–	–	–	–	–	–
Udo et al., 2017	–	4/4	<5% 3<10% 1	4/4	0/4	4/4	4/4	–	–	1/4	4/4	4/4	0/4	–	–	–	–	–	–	–
Ishikawa et al., 2016	2/5	3/5	<5% 3<10% 1	5/5	0/5	–	–	–	–	3/3	–	–	–	5/5	–	–	–	–	–	–
Kim et al., 2017	–	1/1	0/1	1/1	–	–	–	–	–	1/1	–	1/1	–	–	–	–	–	–	–	–
Kon et al., 2016	–	5/5	<1% 5	5/5	0/5	0/5	–	0/5	0/5	0/5	–	5/5	–	5/5	5/5	–	–	–	–	–
Chuang et al., 2014	1/1	1/1	<1%	1/1	0/1	–	–	–	–	0/1	–	1/1	–	–	–	–	–	–	–	–
Sato et al., 2010	–	2/2	3%;10%	2/2	0/2	1/2	–	–	–	–	–	–	–	–	–	–	–	–	–	
Our series	5/5	0/5	<5% 5	–	–	–	–	–	–	–	–	5/5	–	–	–	–	5/5	5/5	5/5	5/5

–, not available. [a] Number of positive/negative not reported.

TABLE 3 | Summary of the detected gene mutations from past reports and the present cases.

Authors	EGFR	BRAF	KRAS	Others
Taguchi et al., 2017	0[a]/1[b]	0/1	0/1	Alk 1/1
Kataoka et al., 2018	2/4	1/4	1/4	–
Chang et al., 2018	5/21	6/21	4/21	HRAS 1/21
Kamata et al., 2015	–	1/1	–	AKT1 1/1
Jin et al., 2017	–	–	–	ALK 1/1
Udo et al., 2017	–	1/4	1/4	AKT1 1/4
Kim et al., 2017	–	1/1	–	–
Chuang et al., 2014	0/1	–	0/1	–
Kamata et al., 2016	3/10	5/10	–	–
Our series	1/5	–	–	EGR1, MUC20, MUC3A, NBPF19, NOL4L, OR4L1

–, not available. [a] Positive number. [b] Number of people surveyed.

DISCUSSION

In this study, we identified a high prevalence of driver gene mutations in the CMPTs; a similar TMB and phylogenetic tree with cancer tissues and an adenocarcinoma coexisted in one case. Distinct molecular and immune check point features of each component provided evidence that these enigmatic lesions are indeed neoplastic processes with immune escape.

Five cases had characteristic features of CMPT, and one of them had mucinous adenocarcinoma simultaneously. Chang et al. (2018) reported one case of CMPT coexisting with adenocarcinoma, and they showed that CMPT may be malignant. The CMPT population has features similar to malignant features, including alveolar structural damage and elastic fiber aggregation, tumor cells proliferating along the alveolar wall, jumping lesions, no capsules, and CEA positivity (Kamata et al., 2015; Kon et al., 2016; Taguchi et al., 2017). Because histology is invasive, a CMPT is easily misdiagnosed as adenocarcinoma with a diagnosis based on frozen pathology. Therefore, we should conduct in-depth research on the cellular components, composition, and developmental trend of CMPT to provide more accurate guidance for clinical work. Chuang et al. (2014) suggested that although CMPT does not meet the criteria for ciliated adenocarcinoma, it has the characteristics of pre-mutation, including goblet (mucus) cell metaplasia and goblet cell TTF-1 staining loss. According to the immunohistochemistry results reported in previous studies, in many cases, CK7/CEA/TTF-1 expressions were positive, and most CK20 expressions were negative. These findings are very similar to those for adenocarcinoma (Sato et al., 2010; Chuang et al., 2014; Ishikawa et al., 2016; Kamata et al., 2016; Kon et al., 2016; Lau et al., 2016; Jin et al., 2017; Kim et al., 2017; Taguchi et al., 2017; Udo et al., 2017; Miyai et al., 2018) and indicate that CMPTs are potential malignant tumors.

Nonetheless, the long-term biological behavior of CMPTs could not be established by the present study, which had a limited follow-up, and larger studies with longer follow-ups are necessary to accurately determine the course of CPMTs

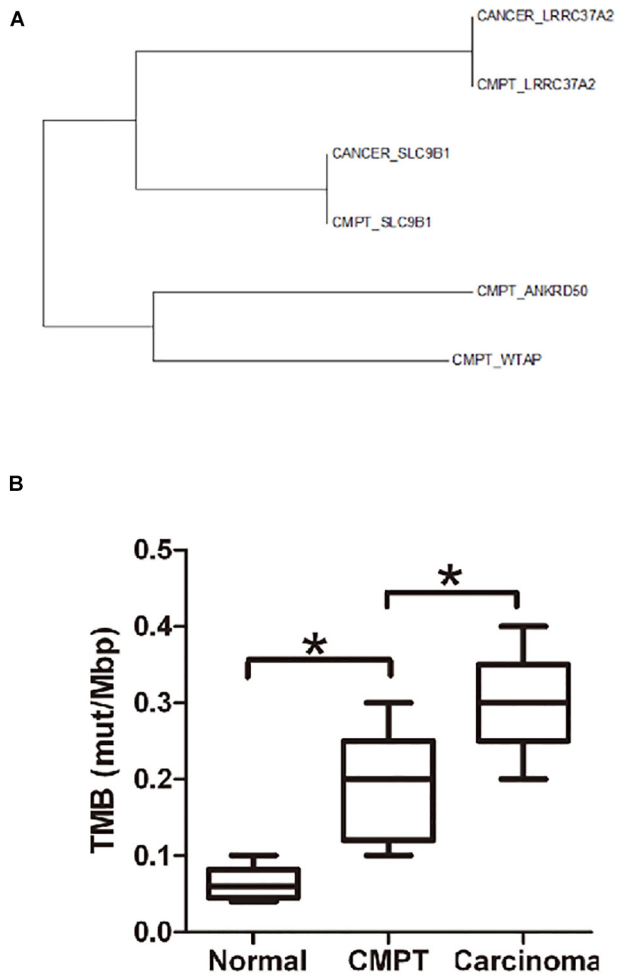

FIGURE 2 | Evolutionary relationships of the taxa and the TMB in different tissues. **(A)** Evolutionary relationships of the taxa. The evolutionary history was inferred using the neighbor-joining method. The optimal tree with the sum of branch length = 2.65875810 is shown. The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved six amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 117 positions in the final dataset. Evolutionary analyses were conducted with MEGA7. **(B)** The TMB (mut/Mbp) in normal tissue, CMPTs, and lung carcinoma. The TMB of CMPT tends to be similar to the TMB of cancer tissues, and both are higher than that in normal tissues. *$p < 0.05$.

(Kamata et al., 2015). Among the five cases, one had a family history of lung cancer, which was her mother (Patient 1), and one case coexisted with lung cancer (Patient 2). This allows us to question whether CMPT really has a malignant potential and whether its subsequent process is lung cancer. Through genetic testing, we identified a high prevalence of driver gene mutations in all CMPTs by whole exon sequencing, and we also found a non-frame shift insertion mutation in exon 20 of *EGFR* in the tumor tissues, which has been considered to be a key driving gene for lung cancer (**Supplementary Figure S5**). According to the functional enrichment analysis of the tumor variant gene, enriched signal pathways are associated with the excessive proliferation of cells. *MUC20* and *MUC3A* co-mutated at the junction of the tumor and tumor tissues are mucin family genes that are involved in the development of various adenocarcinomas, including lung cancer. Many studies have shown that mucins can be misexpressed in malignant tumors (Zheng et al., 2018). Is this related to the formation of mucus lakes in CMPT? Exploration of more cases is necessary. These results provide a good basis for the tumor properties of CMPT.

Similarly, inconclusive here is whether CMPTs have any potential for malignant transformation with immune escape. We observed the specific influence structure of a CMPT malignant potential and the mechanism of a CMPT malignant potential. We tested the CD28 family of immune escape targets on CMPTs. The expression of PD-L1 (B7H1/CD274), B7H3 (CD276), and B7H4 in tissues was observed in all five patients (**Table 2**). It is well known that PDL1, B7H3, and B7H4 are highly expressed in tumor tissues to achieve immune escape and promote tumorigenesis (Wiegering et al., 2019). It is notable that PDL1, B7H3, B7H4, and other indicators are mostly expressed in mucus cells of CMPTs. Moreover, we found PDL1 overexpression in CMPTs with adenocarcinoma coexisting compared with other CMPT cases, prompting the presence of immune escape. There is a growing consensus on the importance of PDL1 as a diagnostic biomarker or favorable prognostic factor in CMPTs.

In addition, we also found the high expression of OFD1 in CMPTs by immunohistochemistry, which is an important inhibitor of primary cilia in cancer cells (Wiegering et al., 2019). The elevation of OFD1 indicates a decrease in autophagy and the disappearance of cilia, and studies have shown a close relationship between the disappearance of cilia and tumorigenesis (Tang et al., 2013). The test results for these indicators support CMPTs having a certain malignant potential.

In addition to the CD28 family, we also observed the expression of LAG3 and siglec15 on CMPT tissue, both of which were negative. LAG3 and Siglec15 are novel immunomodulatory targets that inhibit antigen-specific T cell responses, and siglec15 is a major immunosuppressive molecule of PDL1-negative tumors (Nguyen and Ohashi, 2015; Wang et al., 2019). Combined with the CD28 family of immune escape target results, siglec15 negativity coincided with our expectations, and these results support the view that CMPT has malignant potential (Janakiram et al., 2017). Moreover, through biological tree evolution analysis, we found that CMPT and mucinous adenocarcinoma genes share a common evolutionary direction. At the same time, CMPT has the same TMB as adenocarcinoma, and it is higher than that in normal tissue. Among them, high TMB may have a relationship with the gene mutations we detected. This may suggest that mucus cells in CMPT may become cancerous and develop into mucinous adenocarcinoma.

We identified a high prevalence of driver gene mutations in all the CMPTs, a similar TMB and phylogenetic tree as with cancer tissues, and adenocarcinoma coexisted in one case. Distinct molecular and immune check point features of each component provided evidence that these enigmatic lesions are indeed neoplastic processes with immune escape.

CONCLUSION

The high prevalence of driver gene mutations in CMPTs, similar TMB and phylogenetic tree with cancer tissues indicate their malignant potential. Distinct molecular and immune check point features of each component support the notion that ciliated columnar cells in CMPT are insidious with immune escape.

DATA AVAILABILITY STATEMENT

We have uploaded the genetic sequencing data in our manuscript to Dryad Digital Repository.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Ethics Committee of Harbin Medical University Cancer Hospital (KY2017-27). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XY performed the sequence alignment and drafted the manuscript. YH and JiaG carried out the immunoassays. JinG and HM participated in the design of the study and performed the statistical analysis. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Nature Science Foundation of China (81600539), the Postdoctoral Scientific Research Developmental Fund of Heilongjiang Province (LBH-Q18076), the N10 Found project of Harbin Medical University Cancer Hospital (2017-03), the Youth Elite Training Foundation of Harbin Medical University Cancer Hospital (JY2016-06), the Outstanding Youth Foundation of Harbin Medical University Cancer Hospital (JCQN-2018-05), National Nature Science Foundation of Heilongjiang Province (YQ2020H036), National Nature Science Foundation of china (2021, HM), Special funds of central finance to support the development of local University (2019, HM), and Wu-Jieping Medical Foundation (320.6750.19089-22, 320.6750.19089-48).

ACKNOWLEDGMENTS

We wish to acknowledged the Harbin Medical University Cancer Hospital, specifically Department of Pathology for assistance in tissue preparation and immunohistochemistry.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.579737/full#supplementary-material

FIGURE S1 | Clinical findings. **(A)** A representative image of gray-whitish tumor was observed in the peripheral lung (arrows) (Patient 1). **(B)** Chest CT shows the right lower lobe ground glass-like nodules (1 × 1 × 1 cm).

FIGURE S2 | Mutant genes of tumor and junctional tissues (SNV and Indel).

FIGURE S3 | Functional enrichment analysis of mutant genes in tumors.

FIGURE S4 | Gene mutation map. Gene mutation map **(A)** EGR1, **(B)** MUC20, MUC3A, **(C)** NOL4L, and **(D)** OR4L1.

FIGURE S5 | Mutation site map. Mutation site map of the **(A)** U2AF1.p34F mutation and **(B)** EGFR 20 exon non-frame shift insertion mutation.

FIGURE S6 | The mRNA content of PDL1, B7H3 and B7H4 in CMPT and normal tissues. mRNA content **(A)** PDL1, **(B)** B7H3, **(C)** B7H4. $^{*}p < 0.05$.

FIGURE S7 | Negative control. Negative control **(A)** PDL1, **(B)** B7H3, **(C)** B7H4.

REFERENCES

Chang, J. C., Montecalvo, J., Borsu, L., Lu, S., Larsen, B. T., Wallace, W. D., et al. (2018). Bronchiolar adenoma expansion of the concept of ciliated muconodular papillary tumors with proposal for revised terminology based on morphologic, immunophenotypic, and genomic analysis of 25 cases. *Am. J. Surg. Pathol.* 42, 1010–1026. doi: 10.1097/pas.0000000000001086

Chuang, H. W., Liao, J. B., Chang, H. C., Wang, J. S., Lin, S. L., Hsieh, P. P., et al. (2014). Ciliated muconodular papillary tumor of the lung: a newly defined peripheral pulmonary tumor with conspicuous mucin pool mimicking colloid adenocarcinoma: a case report and review of literature. *Pathol. Int.* 64, 352–357. doi: 10.1111/pin.12179

Ishikawa, M., Sumitomo, S., Imamura, N., Nishida, T., Mineura, K., Ono, K., et al. (2016). Ciliated muconodular papillary tumor of the lung: report of five cases. *J. Surg. Case Rep.* 8, 1–4.

Janakiram, M., Shah, U. A., Liu, W., Zhao, A., Schoenberg, M. P., and Zang, X. (2017). The third group of the B7-CD28 immune checkpoint family: HHLA2, TMIGD2, B7x, and B7-H3. *Immunol. Rev.* 276, 26–39. doi: 10.1111/imr.12521

Jin, Y., Shen, X., Shen, L., Sun, Y., Chen, H., and Li, Y. (2017). Ciliated muconodular papillary tumor of the lung harboring ALK gene rearrangement: case report and review of the literature. *Pathol. Int.* 67, 171–175. doi: 10.1111/pin.12512

Kamata, T., Sunami, K., Yoshida, A., Shiraishi, K., Furuta, K., Shimada, Y., et al. (2016). Frequent BRAF or EGFR mutations in ciliated muconodular papillary tumors of the lung. *J. Thorac. Oncol.* 11, 261–265. doi: 10.1016/j.jtho.2015.10.021

Kamata, T., Yoshida, A., Kosuge, T., Watanabe, S., Asamura, H., and Tsuta, K. (2015). Ciliated muconodular. (papillary)tumors of the lung: a clinicopathologic analysis of 10 cases. *Am. J. Surg. Pathol.* 39, 753–760. doi: 10.1097/pas.0000000000000414

Kataoka, T., Okudela, K., Matsumura, M., Mitsui, H., Suzuki, T., Koike, C., et al. (2018). A molecular pathological study of four cases of ciliated muconodular papillary tumors of the lung. *Pathol. Int.* 68, 353–358. doi: 10.1111/pin.12664

Kim, L., Kim, Y. S., Lee, J. S., Choi, S. J., Park, I. S., Han, J. Y., et al. (2017). Ciliated muconodular papillary tumor of the lung harboring BRAF V600E mutation and p16INK4a overexpression without proliferative. *J. Thorac. Dis.* 9, E1039–E1044.

Kon, T., Baba, Y., Fukai, I., Watanabe, G., Uchiyama, T., Murata, T., et al. (2016). Ciliated muconodular papillary tumor of the lung: a report of five cases. *Pathol. Int.* 66, 633–639. doi: 10.1111/pin.12460

Lau, K. W., Aubry, M. C., Tan, G. S., Lim, C. H., and Takano, A. M. (2016). Ciliated muconodular papillary tumor: a solitary peripheral lung nodule in a teenage girl. *Hum. Pathol.* 49, 22–26.

Liu, L., Aesif, S. W., Kipp, B. R., Voss, J. S., Daniel, S., Aubry, M. C., et al. (2016). Ciliated muconodular papillary tumors of the lung can occur in western patients and show mutations in BRAF and AKT1. *Am. J. Surg. Pathol.* 40, 1631–1636. doi: 10.1097/pas.0000000000000707

Miyai, K., Takeo, H., Nakayama, T., Obara, K., Aida, S., Sato, K., et al. (2018). Invasive form of ciliated muconodular papillary tumor of the lung: a case report and review of the literature. *Pathol. Int.* 68, 530–535. doi: 10.1111/pin.12708

Nguyen, L. T., and Ohashi, P. S. (2015). Clinical blockade of PD1 and LAG3–potential mechanisms of action. *Nat. Rev. Immunol.* 15, 45–56. doi: 10.1038/nri3790

Sato, S., Koike, T., Homma, K., and Yokoyama, A. (2010). Ciliated muconodular papillary tumour of the lung: a newly defined low-grade malignant tumour. *Interact. Cardiovasc. Thorac. Surg.* 11, 685–687. doi: 10.1510/icvts.2009.229989

Taguchi, R., Higuchi, K., Sudo, M., Voss, J. S., Daniel, S., Aubry, M. C., et al. (2017). A case of anaplastic lymphoma kinase (ALK)-positive ciliated muconodular papillary tumor (CMPT) of the lung. *Pathol. Int.* 67, 99–104. doi: 10.1111/pin.12504

Tang, Z., Lin, M. G., Stowe, T. R., Chen, S., Zhu, M., Stearns, T., et al. (2013). Autophagy promotes primary ciliogenesis by removing OFD1 from centriolarsatellites. *Nature* 502, 254–257. doi: 10.1038/nature12606

Udo, E., Furusato, B., Sakai, K., Prentice, L. M., Tanaka, T., Kitamura, Y., et al. (2017). Ciliated muconodular papillary tumors of the lung with KRAS/BRAF/AKT1 mutation. *Diagn. Pathol.* 12:62.

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

Wiegering, A., Rüther, U., and Gerhardt, C. (2019). the role of primary cilia in the crosstalk between the ubiquitin?proteasome system and autophagy. *Cells* 8:241. doi: 10.3390/cells8030241

Zheng, Q., Luo, R., Jin, Y., Shen, X., Shan, L., Shen, L., et al. (2018). So-called "non-classic" ciliated muconodular papillary tumors: a comprehensive comparison of the clinicopathological and molecular features with classic ciliated muconodular papillary tumors. *Hum. Pathol.* 82, 193–201. doi: 10.1016/j.humpath.2018.07.029

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fgene.2020.604336

Distinguishing Glioblastoma Subtypes by Methylation Signatures

Yu-Hang Zhang[1,2†], Zhandong Li[3†], Tao Zeng[4], Xiaoyong Pan[5], Lei Chen[6], Dejing Liu[7], Hao Li[3], Tao Huang[7] and Yu-Dong Cai[1]**

[1] School of Life Sciences, Shanghai University, Shanghai, China, [2] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [3] College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [4] Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China, [5] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, China, [6] College of Information Engineering, Shanghai Maritime University, Shanghai, China, [7] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Yanding Zhao,
Baylor College of Medicine, United States
Xuefeng Gu,
Shanghai University of Medicine and Health Sciences, China

***Correspondence:**
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.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: *09 September 2020*
Accepted: *02 November 2020*
Published: *24 November 2020*

Citation:
Zhang Y-H, Li Z, Zeng T, Pan X, Chen L, Liu D, Li H, Huang T and Cai Y-D (2020) Distinguishing Glioblastoma Subtypes by Methylation Signatures. Front. Genet. 11:604336. doi: 10.3389/fgene.2020.604336

Glioblastoma, also called glioblastoma multiform (GBM), is the most aggressive cancer that initiates within the brain. GBM is produced in the central nervous system. Cancer cells in GBM are similar to stem cells. Several different schemes for GBM stratification exist. These schemes are based on intertumoral molecular heterogeneity, preoperative images, and integrated tumor characteristics. Although the formation of glioblastoma is remarkably related to gene methylation, GBM has been poorly classified by epigenetics. To classify glioblastoma subtypes on the basis of different degrees of genes' methylation, we adopted several powerful machine learning algorithms to identify numerous methylation features (sites) associated with the classification of GBM. The features were first analyzed by an excellent feature selection method, Monte Carlo feature selection (MCFS), resulting in a feature list. Then, such list was fed into the incremental feature selection (IFS), incorporating one classification algorithm, to extract essential sites. These sites can be annotated onto coding genes, such as *CXCR4, TBX18, SP5*, and *TMEM22*, and enriched in relevant biological functions related to GBM classification (e.g., subtype-specific functions). Representative functions, such as nervous system development, intrinsic plasma membrane component, calcium ion binding, systemic lupus erythematosus, and alcoholism, are potential pathogenic functions that participate in the initiation and progression of glioblastoma and its subtypes. With these sites, an efficient model can be built to classify the subtypes of glioblastoma.

Keywords: glioblastoma, methylation, signature, subtype, classification

INTRODUCTION

Glioblastoma, also called as glioblastoma multiform (GBM), is the most aggressive cancer that initiates within the brain. The cause of this disease is unclear. The risk factors of GBM include genetic factors and environmental factors, such as smoking and exposure to pesticides. Similar to other brain cancers, GBM can cause epilepsy, nausea, vomiting, headaches, and mild hemiplegia. The typical symptoms of glioblastoma are deteriorating memory and personality or decline in

neurological function. Most symptoms are caused by the destruction of the temporal lobes and the frontal lobes. Different subspecies of glioblastomas are produced in the central nervous system, and cancer cells in GBM are similar to stem cells.

Several different schemes for glioblastoma stratification exist. One is based on intertumoral molecular heterogeneity in GBM. This scheme identities the subtypes of procedural and mesenchymal glioblastoma on the basis of the biomarker genes *VEGF-A*, *VEGF-B*, *ANG1*, and *ANG2* (Sharma et al., 2017). The second technique involves the use of preoperative images as predictive markers of GBM subtypes; in this approach, the distinctive imaging phenotypes and imaging patterns of glioblastoma subtypes are detected by employing machine-learning techniques (Macyszyn et al., 2016). The third technique is based on integrated tumor subtypes, which have been discovered through an integrative subtype analysis of the GBM dataset from the cancer genome atlas (TCGA) (Shen et al., 2012).

The promoter region is a functional part of the genome that is regulated by methylation and contributes to the regulation of gene expression during the pathogenesis of glioblastoma. Such genomic modification affects the expression of a group of important proteins, including *MGMT*, *GATA6*, and *CASP8*; the dysmethylation of these genes is remarkable in glioblastoma (Skiriute et al., 2012). For example, through whole-genome wide methylation screening, a study found that 5 m-dC level is the best discriminant among methylation classes, and the upregulation of *LINE1* methylation is an independent prognostic factor in GBM diseases (Lai et al., 2014). Although the formation of glioblastoma is related to gene methylation, glioblastoma has been poorly classified on the basis of epigenetics.

Preliminary attempts on clustering GBMs using epigenetic biomarkers have already started. According to a systematic analysis on the DNA methylation-based classification of central nervous system tumors (Guardiola Bagán et al., 2017; Capper et al., 2018), central nerve system (CNS) tumors can be further classified into multiple subgroups based on the whole-genome wide methylation status. As one important part of the CNS tumors, GBM can be further classified into eight classes, which is DMG K27, GBM G34, GBM MES, GBM RTK I, GBM RTK II, GBM RTK III, GBM MID, and GBM MYCN. Researchers tried to use unsupervised clustering of reference samples using t-SNE dimensionality reduction. According to the original publications, group DMG K27 can be easily distinguished from other seven groups based on the results of t-SNE based separation. However, the differences between the other seven subgroups cannot be clarified clearly and the specific methylation locus that contribute to the separation have not been identified. Therefore, in this study, we used methylation datasets downloaded from Gene Expression Omnibus (GEO) database to identify specific methylation locus/biomarkers that contribute to the classification and annotation of different GBM subgroups (Capper et al., 2018).

We aimed to identify essential methylation sites (features) in this study, on which the subtypes of glioblastoma can be efficiently classified. To this end, we employed two datasets collected in GEO. One dataset was termed as the training dataset, whereas the other was treated as the independent test dataset. A powerful feature selection method, Monte Carlo feature selection (MCFS) (Dramiński et al., 2007), was applied on the training dataset. A feature list, indicating the importance of features, was produced. After that, incremental feature selection (IFS) (Liu and Setiono, 1998) was executed on this list, which incorporated one classification algorithm, to extract essential methylation sites. As a result, we found 4100 methylation sites (features) associated with the classification of GBM. These sites can be annotated onto coding genes, such as *CXCR4, TBX18, SP5*, and *TMEM22*. Through the further functional enrichment analysis of these dysmethylated genes using GO and KEGG databases, we identified several biological functions related to GBM classification (e.g., subtype-specific functions). Also, with these methylation sites, an efficient model with support vector machine (SVM) (Cortes and Vapnik, 1995) as the prediction engine can be built to classify subtypes of glioblastoma. In summary, on the basis of the powerful computational approaches, we identified various novel potential pathogenic genes at the epigenetics level and revealed several potential pathogenic functions that participate in the initiation and progression of glioblastoma and its subtypes with wide support from recent reports.

MATERIALS AND METHODS

Dataset

Two sets of methylation profiles of patients with GBM were downloaded from GEO with the accession numbers GSE90496 and GSE109379 (Capper et al., 2018). The first dataset included 347 GBM cases and the second dataset contained 324 GBM cases. These two datasets were used as the training dataset and independent test dataset, respectively. All GBM cases are classified into seven categories. The distribution of GBM cases on seven categories is listed in **Table 1**. The methylation levels of 42,383 probes were used to represent each patient. The goal was to identify discriminative methylation features (e.g., dysmethylated sites or genes) corresponding to different GBM subtypes.

Feature Selection

In this study, we first used MCFS (Dramiński et al., 2007) to identify the general interpretable information of features (methylation sites) in tumor samples from the central nervous system. Then, we applied IFS (Liu and Setiono, 1998) to

TABLE 1 | Breakdown of the GBM samples in the training and independent datasets.

Category	Training dataset	Independent dataset
G34	41	13
MES	56	104
MID	14	19
MYCN	16	17
RTK	64	44
RTK II	143	118
RTK III	13	9

improve classification performance by obtaining a group of optimal features with the strong recognition ability of central nervous system tumors.

MCFS

Monte Carlo feature selection is a classical and powerful feature selection method wherein decision trees are used to find distinguishable features for classification (Dramiński et al., 2007). It is quite suitable to analyze datasets with features much more than samples. The datasets described in section "Dataset" are in such type. Thus, we adopted MCFS to analyze the training dataset, aiming to extract essential features. Furthermore, such feature selection method can deeply investigate complicated relationship between features or class labels, extracting essential features in deep levels.

The MCFS method evaluates the importance of features by constructing lots of decision trees. Given a dataset with M features, randomly construct s feature subsets consisting of m features, where m is much smaller than M. For each feature dataset, t bootstrap sample sets are constructed from the original dataset, in which samples are represented by features in such feature subset. Accordingly, t decision trees are built. After all feature subsets are processed by the above procedures, $s{\cdot}t$ decision trees are constructed. Based on these trees, a feature g is assigned a relative importance (RI) value, which can be calculated by

$$RI_g = \sum_{\tau=1}^{st} (wAcc)^u \sum_{n_g(\tau)} IG(n_g(\tau)) \left(\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau} \right)^v, \quad (1)$$

where $IG(n_g(\tau))$ stands for the gain information of node $n_g(\tau)$, (no. in $n_g(\tau)$) represents the number of samples in node $n_g(\tau)$, (no. in τ) denotes the number of samples in tree τ, $wAcc$ indicates the weighted accuracy of the tree. u and v are the regular factors, which were suggested to set to one (Dramiński et al., 2007). All investigated features are ranked in a list with the decreasing order of their RI values. Clearly, features with high ranks are more important than those with low ranks.

In present study, we used the MCFS program retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html. Default parameters were adopted.

IFS

Incremental feature selection is a feature selection method used to distinguish between samples from different classes (e.g., normal and diseased) (Liu and Setiono, 1998). In this study, different classes of samples were discerned by a set of optimal features screened by IFS performed in a rank-descending feature list. We set candidate high-performance feature subsets as feature subsets with large interval sizes (e.g., 10 features) from the ranked feature list. Suppose N candidate feature subsets $\mathrm{F} = [F^1, F^2, \ldots, F^N]$ exist. The i-th feature subset includes $10 * i$ features yielding $F^i = [f_1, f_2, \ldots, f_{i*N}]$. We construct and evaluate the classifier on each candidate feature subset. The candidate feature subset with the maximal prediction performance is the optimal feature subset, and the classifier constructed from these optimal features is the optimal classifier.

Classification Algorithm

Support Vector Machine

The classifier acts as a classification model that maps data samples to a given category for data class prediction. We use support vector machine (SVM) (Cortes and Vapnik, 1995) based on statistical learning theory for supervised data classification. It has wide application for tackling different biological problems (Muthukrishnan et al., 2014; Chen et al., 2017, 2020; Liu et al., 2020; Sang et al., 2020; Zhou et al., 2020a,b). The basic principle is to use a given kernel function (e.g., Gaussian kernel) to transform data from a low-dimensional space to a high-dimensional space. The SVM model can separate the samples of each class/category by maximizing the data interval and also predicts (new) sample categories on the basis of the interval where this sample falls in. For two-class classification, the largest margin between the two categories of samples can be inferred by SVM, where large margins are associated with small generalization error. For multiclass classification, SVM uses the "One Versus the Rest" strategy. In this study, we solved the optimization problem of SVM by using the sequence minimization optimization (SMO) algorithm (Platt, 1998; Keerthi et al., 2001) implemented by the tool "SMO" in Weka software (Frank et al., 2004; Witten and Frank, 2005), which can be downloaded at https://www.cs.waikato.ac.nz/ml/weka/. For convenience, the default parameters were adopted, where the kernel was a polynomial function and the regularization parameter C was set to one.

Random Forest

A random forest (RF) (Breiman, 2001) is a metaclassifier that contains a large number of tree classifiers for establishing final joint classification, which determines the output categories/classes by summarizing votes from different decision trees (Breiman, 2001). The RF is a commonly used method in machine learning and is widely applied in computational biology (Pan et al., 2010; Zhao et al., 2018; Jia et al., 2020; Liang et al., 2020; Yuan et al., 2020). Notably, a slight difference exists between each decision tree and other decision trees in a RF. Thus, the predictions of all decision trees are averaged to obtain the final decision of RF. This approach can avoid over-fitting and improve the performance of the integrated model. However, it slightly increases the bias of the overall model and causes the loss of some interpretability. In this study, we used the tool "RandomForest" in Weka (Frank et al., 2004; Witten and Frank, 2005), which implemented the above RF. The number of decision trees was set to ten.

Rule Learning

In this study, we used the rule learner known as repeated incremental pruning to produce error reduction (RIPPER) to generate classification rules for classifying samples from different GBM subtypes (Cohen, 1995). RIPPER learns interpretable classification rules consisting of IF–ELSE rules. Briefly, RIPPER learns the rules of one class and then moves to learn the next class in a given order, e.g., it learns from the first minority class to the next until the dominant class. To quickly implement the RIPPER algorithm, we directly employed the tool "JRip"

in Weka (Frank et al., 2004; Witten and Frank, 2005). Default parameters were used.

Functional Enrichment Analysis

The selected optimal methylation probes (features) were mapped onto genes on the basis of the annotation files of GPL13534 downloaded from GEO. The enrichments of these genes on GO terms and KEGG pathways were evaluated with hypergeometric tests measured by phyper function in R[1]. The cutoff of the adjusted hypergeometric test *p*-values, i.e., FDR (false discovery rate), was set to 0.05. In other words, only the GO terms and KEGG pathways with FDR < 0.05 were considered to be statistically significant.

Performance Measurement

We employed Matthew Correlation Coefficients (MCC) (Matthews, 1975; Gorodkin, 2004) to evaluate the performance metrics of different kinds of classifiers. The MCC accounts for true and false positives and true and false negatives, and this measurement has values ranging from -1 and $+1$. It is a common method for calculating the correlation between target and prediction classes. Applying 10-fold cross-validation (Kohavi, 1995), we used MCC to evaluate the performance of different training models for glioblastoma classification.

RESULTS

In this study, we investigated the methylation profiles of GBM patients. The entire procedures are illustrated in **Figure 1**.

Results of MCFS Method on the Training Dataset

We first used MCFS to analyze the training dataset. Each feature was evaluated by a RI value. Accordingly, all features were ranked in the decreasing order of their RI values. Obtained feature list is provided in **Supplementary Table 1**.

IFS Results

Next, we generated a series of feature subsets from the MCFS feature list and then subjected them to IFS with SVM, RF, and RIPPER to obtain the best features for classifying different categories of GBM samples. The complete results of the three classifiers using different number of features are given in **Supplementary Table 2**. For an easy observation, an IFS curve was plotted with number of used features as X-axis and MCC as the Y-axis for each classification algorithm, as shown in **Figure 2**, in which the highest MCC of each classification is marked. It can be observed that the highest value of MCC generated by SVM was 0.939 when using the top-ranked 4100 features. Accordingly, we constructed the optimal SVM classifier with these 4100 features. For RF, when using the top-ranked 1690 features, the largest MCC value of 0.882 was achieved. These 1690 features were used to build the optimal RF classifier. When using the top-ranked 1180 features, the highest MCC value of 0.737 was obtained by RIPPER. The optimal RIPPER classifier was built based on these 1180 features. The overall accuracies of above-mentioned classifiers are listed in **Table 2** and the accuracies on seven categories are shown in **Figure 3**. As shown by these results, the optimal classifier was SVM, which was superior to RIPPER and RF although it used additional features.

Performance of Optimal Classifiers on the Test Dataset

To show the generalizability of our pipeline, we also evaluated above-constructed classifiers on a completely independent test dataset. The MCCs generated by the optimal SVM, RF, and RIPPER classifiers were 0.798, 0.832, and 0.937. These results are summarized in **Table 3**, in which the corresponding overall accuracies are also listed. The detailed performance on each category is shown in **Figure 4**. The results indicated that the RIPPER classifier had better generalizability than other two algorithms, and SVM shown the worst generalizability performance in this study.

Results of Enrichment Analysis

On the training dataset, the optimal SVM classifier gave the best performance, which adopted 4100 top-ranked features (methylation sites). These sites were mapped onto genes based on the annotation file of Illumina HumanMethylation450 BeadChip from GEO with platform number of GPL13534[2], resulting

[1]http://finzi.psych.upenn.edu/R/library/stats/html/Hypergeometric.html

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534

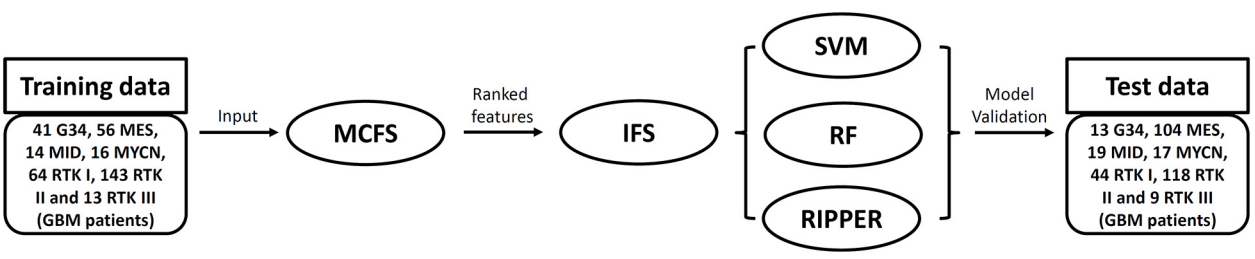

FIGURE 1 | Flowchart of the analysis performed in this study. The training dataset is first analyzed by the Monte Carlo feature selection (MCFS) method. Features are ranked in a list, which is fed into the incremental feature selection (IFS) with one of three classification algorithms. The optimal classifiers based on different classification algorithms are built and further evaluated their performance on a test dataset.

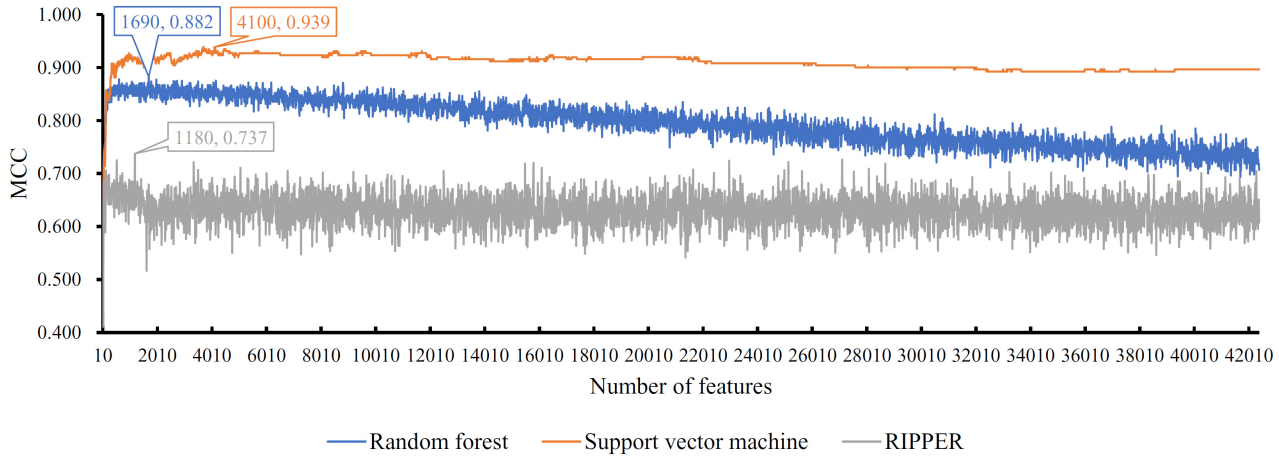

FIGURE 2 | IFS curves with support vector machine, random forest, and RIPPER on the training set. The support vector machine can yield the highest MCC (0.939) when top 4100 features are used, while the highest MCCs of random forest and RIPPER are 0.882 and 0.737, respectively, when top 1690 and 1180, features respectively, are adopted.

in 1813 coding genes, which are provided in **Supplementary Table 3**. For consistency, these genes were called optimal genes in the following text.

The enrichment analysis was done on the above 1813 genes. The results are listed in **Supplementary Table 4**. Several GO terms and KEGG pathways with FDR < 0.05 were obtained. In detail, we obtained 167 biological process (BP) GO terms, 28 cellular component (CC) GO terms, 26 molecular function (MF) GO terms and four KEGG pathways. Some of them would be analyzed in section "Biological Functions Relevant to GBM Based on Optimal Genes" and "Biological Pathways Relevant to GBM Based on Optimal Genes."

DISCUSSION

Optimal Genes Relevant to GBM

As mentioned in section "Results of Enrichment Analysis," 1813 optimal genes were obtained. We selected some of them for analysis in this section. These genes are targeted by probes with high RI values.

The first gene is ***CXCR4*** (targeted by probe **cg02902079** and **cg10824187**), which is a lymphocyte activity regulation molecule and acts as an alpha-chemokine receptor specific for stromal-derived-factor-1. Chemokines play important autocrine and paracrine roles during tumor initiation and progression. Generally, the *in vivo* secretion of chemokines regulates the biological effects of various components in the microenvironment of *CXCR4* (Würth et al., 2014). In cancer stem cells, *CXCR4* is upregulated and plays an irreplaceable role in perivascular invasion, a specific tumor behavior in GBM (Yadav et al., 2016). In addition, *CXCR4* is an effective target for improving tumor sensitivity in GBM in conjunction with radiation therapy (Yadav et al., 2016). Moreover, *CXCR4* is suppressed by *PATZ1*, which is enriched in the proneural subtype and colocalizes with stemness markers of GBMs (Guadagno et al., 2017).

TABLE 2 | 10-fold cross-validation performance of the optimal SVM, RF, and RIPPER classifiers on the training set.

Classification algorithm	Number of features	Overall accuracy	MCC
SVM	4100	0.954	0.939
RF	1690	0.911	0.882
RIPPER	1180	0.804	0.737

The next identified probe turns out to be **cg26558485**, targeting the 5′UTR of CYP4X1. As a member of the cytochrome P450 superfamily of enzyme, such gene has been generally reported to participate in neurovascular function in the brain (Bylund et al., 2002). As for its correlations with GBM, recently, two successive related publications (Wang et al., 2018, 2019) confirmed that CYP4X1 contributes to the inhibition of glioma angiogenesis. Glioma vasculature is quite significant for the initiation and progression of such disease (Hardee and Zagzag, 2012). The methylation of related functional regions of such gene definitely affect its biological functions, which further plays an irreplaceable role for GBM pathogenesis. Therefore, such target gene can be an effective GBM associated gene.

Apart from probe **cg26558485**, another probe named as **cg07028914** targets a transcription factor named as ***TBX18***. According to recent publications, an independent study in 2015 confirmed that microRNA miR-205 prevent the invasion of glioma by targeting *TBX18* (Zheng et al., 2015), reflecting the potential regulatory role of *TBX18* during glioma pathogenesis. Most of microRNAs' biological effects on glioma pathogenesis relied on the regulation on gene expression, which is similar with methylation mediated biological processes. Therefore, the methylation status of such gene may also play a potential regulatory role for the invasion of glioma.

The next gene, ***SP5*** is targeted by multiple probes including **cg26766005** and **cg14768335**. According to recent publications,

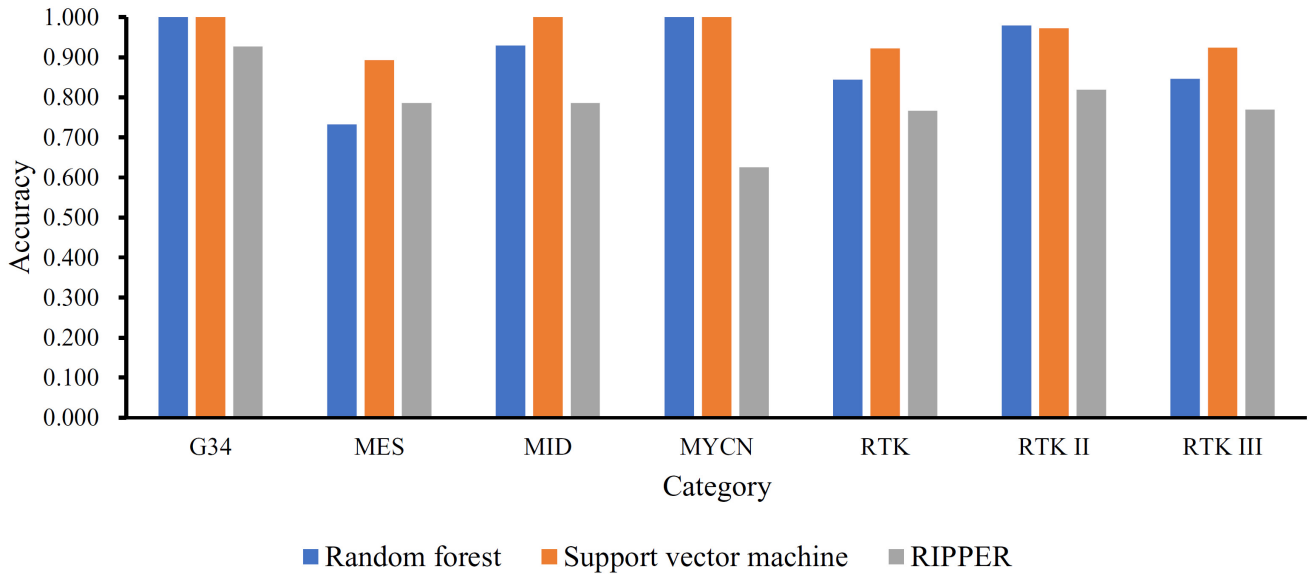

FIGURE 3 | Performance of the optimal SVM, RF, and RIPPER classifiers on different categories in the training dataset. The optimal SVM and RF classifiers are much superior to the optimal RIPPER classifier, and the optimal SVM classifier is slightly superior to the optimal RF classifier.

SP5 has been shown to be therapeutic target and a prognostic biomarker for multiple cancer subtypes, including glioma (Safe and Abdelrahim, 2005; Safe et al., 2014). Considering that methylation can regulate the expression level and biological effects of a target gene, the methylation status of the regulatory region of such gene may also probably affect the pathogenesis of glioma and have different pathological effects in different glioma subgroups indirectly.

As for ***TMEM22***, also known as *SLC35G2*, which is targeted by the optimal features **cg25836094, cg13383019**, and **cg22304507**, it has been generally reported to participate in cell proliferation and tumorigenesis with few publications (Dobashi et al., 2009). Although such gene has not been directly reported to be functionally correlated with glioma, it has been widely reported to be associated with renal cell carcinoma and its homolog which shared similar biological functions, *TMEM97* has been directly confirmed to be correlated with glioma at transcriptomics level. Considering that methylation at gene body is correlated with gene transcription, it is reasonable for us to regard *TMEM22* associated probes as potential glioma associated probes.

The next identified probe turns out to be **cg11823511**, targeting gene ***BARHL2***. According to two independent studies reported by researchers from University of Birmingham (Dunwell et al., 2010) and Memorial Sloan-Kettering Cancer Center (Shen et al., 2012), respectively, the methylation of *BARHL2* is not only related to hematological and epithelial cancers, but nerve system malignancies including glioma and may play a specific role for the integrative subgrouping of glioma (Shen et al., 2012).

RASGRF2 targeted by probe **cg06829830** has also been predicted to be contribute to the pathogenesis of glioma at methylation level. According to recent publications, in 2019, a systematic review (Wu et al., 2014) on the cancer methylation biomarkers confirmed that such gene is a specific biomarker for aggressive gliomas at methylation level using liquid biopsy.

Apart from such gene, the next identified biomarker is ***TLX3***, targeted by probe **cg26844246**. The methylation alteration of such gene has been identified in multiple tumor subtypes, like thyroid cancer (Kikuchi et al., 2013), bladder cancer and lung adenocarcinoma (Pradhan et al., 2013). In a systematic study on the whole-genome wide glioma methylation status, *TLX3* has been shown with specific methylation status in level II and III gliomas (Suzuki et al., 2015).

As for gene ***ANKRD34A*** (correlated with probes **cg10178263, cg18280463**, and **cg13947666**), according to related methylation studies (Giri and Aittokallio, 2018; Ding et al., 2020), such gene has shown to have methylation changes during the initiation and progression of multiple tumor subtypes, including lung, colon, bladder, lymphoma, breast and ovarian cancer. Therefore, it is reasonable for us to connect the methylation status of *ANKRD34A* with glioma. Apart from that, a recent publication (Ding et al., 2020) in 2020 also indicated that the transcript of such gene, which is regulated by methylation status, may participate in the RNA regulatory network in low grade glioma. Therefore, the methylation of such gene may be correlated with glioma and performed differentially in different subgroups.

TABLE 3 | Performance of the optimal SVM, RF, and RIPPER classifiers on the independent test dataset.

Classification algorithm	Overall accuracy	MCC
SVM	0.852	0.798
RF	0.877	0.832
RIPPER	0.954	0.937

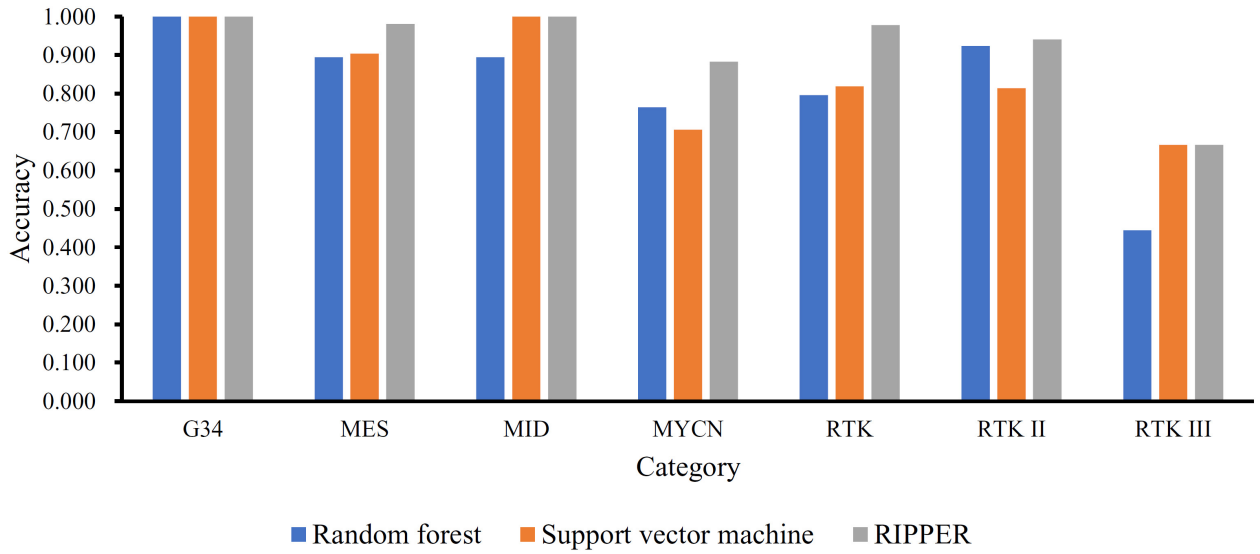

FIGURE 4 | Performance of the optimal SVM, RF, and RIPPER classifiers on different categories in the independent test dataset. The optimal RIPPER classifier gives the best generalizability on the independent test dataset, followed by the optimal RF and SVM classifier.

The last target of the optimal probes is ***MARCH11*** (targeted by probe **cg09017434**), regulating the intracellular transport of lysines. As for its correlations with GBM, according to recent publications, such gene has shown to be correlated with the carcinogenic transformation of cells with different expression levels (Yang et al., 2020). Considering the correlations between gene region methylation and gene expression, it is reasonable for us to speculate that the methylation status of such gene may be correlated with potential malignant alterations, supporting its correlations with GBM.

Biological Functions Relevant to GBM Based on Optimal Genes

Here, to summarize the specific biological functions that may contribute to revealing the differences between different GBM subgroups at methylation level, we performed GO enrichment analyses and pathway analyses on the optimal genes associated with GBM related probes (see **Supplementary Table 4**).

For the GO enrichment analyses results, firstly nervous system development has been screened out. Nervous system development is a biological process related to GBM. The malignant transformation and invasive migration of glioma cells rely on basic cellular components and physical anatomical structure. Therefore, the nervous system may contain proteins that are crucial for GBM. A recent publication confirmed that MT1-MMP, a major component of nervous system development, plays an important role during the pathogenesis of GBM (Beliën et al., 1999). Nervous system development is also associated with DNA methylation. Specific patterns have been seen at the DNA methylation level in the nervous system during the development and pathogenesis of GBM. Some patterns are even shared by two groups (Numata et al., 2012). Therefore, nervous system development, as an effective biological process, can be predicted to contribute to the description of GBM, validating the efficacy and accuracy of our prediction.

Apart from that, the next enriched term calcium ion binding has also been shown to be related to GBM. Various important cells in the central nervous system and the pathogenesis of GBM-like astrocytes participate in complicated metabolite transportation from the blood to the brain. Under pathogenic conditions, glioma cells seize control of the regulation of vascular tone through the Ca^{+}-dependent release of K^{+}, suggesting that calcium ion binding and blood stream in the brain in pathogenic status have important clinical implications (Watkins et al., 2014). Calcium ion binding is also related to methylation. An increase in the ionic strength and a decrease in the methylation reduce the amount of calcium required for the gelation of pectin–calcium systems (Garnier et al., 1993).

Biological Pathways Relevant to GBM Based on Optimal Genes

Apart from GO enrichment analyses, we also performed KEGG pathway analyses on such optimal genes (see **Supplementary Table 4**). The results of this study indicated that alcoholism is related to glioblastoma. Repurposing disulfiram (DSF) is a drug that has been widely used over the past several years to control alcoholism. DSF can inhibit the growth of GBM cells with TMZ resistance without affecting normal cells in the human central nervous system. DSF can suppress the growth and self-renewal of primary cells from GBM tumors, suggesting that an association exists between alcoholism and GBM (Triscott et al., 2012). Alcoholism is also related to the methylation alteration of transporter genes. Methylation status is further affected by alcoholism. The methylation of DAT in peripheral blood has also been validated to be a biomarker for alcohol-dependent patients (Wiers et al., 2015).

CONCLUSION

We found several methylation features (sites) associated with the classification of GBM using our newly presented computational method for classifying glioblastoma subtypes on the basis of gene methylation level. Through the further functional enrichment analysis of dysmethylated genes, such as *CXCR4*, *TBX18*, *SP5*, and *TMEM22*, several potential pathogenic functions are found to participate in the initiation and progression of glioblastoma. These functions include nervous system development, intrinsic plasma membrane component, systemic lupus erythematosus, and alcoholism.

DATA AVAILABILITY STATEMENT

Two sets of methylation profiles of patients with GBM were downloaded from Gene Expression Omnibus (GEO) with the accession numbers GSE90496 and GSE109379.

AUTHOR CONTRIBUTIONS

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

FUNDING

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

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.604336/full#supplementary-material

Supplementary Table 1 | Features ranked by MCFS.

Supplementary Table 2 | 10-fold cross-validation performance of IFS with SVM, RF, and RIPPER when using different number of features ranked by MCFS.

Supplementary Table 3 | Genes corresponding to the top-ranked features by MCFS.

Supplementary Table 4 | Enrichment analysis results on 1813 coding genes.

REFERENCES

Beliën, A. T. J., Paganetti, P. A., and Schwab, M. E. (1999). Membrane-type 1 Matrix Metalloprotease (MT1-MMP) Enables Invasive Migration of Glioma Cells in Central Nervous System White Matter. *J. Cell Biol.* 144, 373–384. doi: 10.1083/jcb.144.2.373

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32.

Bylund, J., Zhang, C., and Harder, D. R. (2002). Identification of a novel cytochrome P450. CYP4X1, with unique localization specific to the brain. *Biochem. Biophys. Res. Commun.* 296, 677–684. doi: 10.1016/s0006-291x(02)00918-x

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.

Chen, L., Pan, X. Y., Guo, W., Gan, Z. J., Zhang, Y. H., Niu, Z. B., et al. (2020). Investigating the gene expression profiles of cells in seven embryonic stages with machine learning algorithms. *Genomics* 112, 2524–2534. doi: 10.1016/j.ygeno.2020.02.004

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

Cohen, W. W. (1995). "Fast effective rule induction," in *The Twelfth International Conference on Machine Learning* (San Francisco, SF: Morgan Kaufmann Publishers Inc), 115–123. doi: 10.1016/b978-1-55860-377-6.50023-2

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

Ding, Y., Liu, H., Zhang, C., Bao, Z., and Yu, S. (2020). Comprehensive Analysis of Prognostic lncRNAs, miRNAs, and mRNAs Forming a Competing Endogenous RNA Network in LGG. Preprint.

Dobashi, S., Katagiri, T., Hirota, E., Ashida, S., Daigo, Y., Shuin, T., et al. (2009). Involvement of TMEM22 overexpression in the growth of renal cell carcinoma cells. *Oncol. Rep.* 21, 305–312.

Dramiński, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., and Komorowski, J. (2007). Monte Carlo feature selection for supervised classification. *Bioinformatics* 24, 110–117. doi: 10.1093/bioinformatics/btm486

Dunwell, T., Hesson, L., Rauch, T. A., Wang, L., Clark, R. E., Dallol, A., et al. (2010). A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Mol. Cancer* 9:44. doi: 10.1186/1476-4598-9-44

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

Garnier, C., Axelos, M. A. V., and Thibault, J.-F. (1993). Phase diagrams of pectin-calcium systems: Influence of pH, ionic strength, and temperature on the gelation of pectins with different degrees of methylation. *Carbohydrate Res.* 240, 219–232. doi: 10.1016/0008-6215(93)84185-9

Giri, A. K., and Aittokallio, T. (2018). DNMT inhibitors increase methylation at subset of CpGs in colon, bladder, lymphoma, breast, and ovarian, cancer genome. *bioRxiv*:395467. Preprint.

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

Guadagno, E., Vitiello, M., Francesca, P., Calì, G., Caponnetto, F., Cesselli, D., et al. (2017). PATZ1 is a new prognostic marker of glioblastoma associated with the stem-like phenotype and enriched in the proneural subtype. *Oncotarget* 8, 59282–59300. doi: 10.18632/oncotarget.19546

Guardiola Bagán, S., Seco, J., Varese, M., Díaz Lobo, M., García Arroyo, J., Teixidó Turà, M., et al. (2017). Toward a novel drug to target the EGF-EGFR interaction: design of metabolically stable bicyclic peptides. *ChemBioChem* 19, 76–84. doi: 10.1002/cbic.201700519

Hardee, M. E., and Zagzag, D. (2012). Mechanisms of glioma-associated neovascularization. *Am. J. Pathol.* 181, 1126–1141. doi: 10.1016/j.ajpath.2012.06.030

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi: 10.1109/access.2020.3009439

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* 13, 637–649. doi: 10.1162/089976601300014493

Kikuchi, Y., Tsuji, E., Yagi, K., Matsusaka, K., Tsuji, S., Kurebayashi, J., et al. (2013). Aberrantly methylated genes in human papillary thyroid cancer and their association with BRAF/RAS mutation. *Front. Genet.* 4:271. doi: 10.3389/fgene.2013.00271

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

Lai, R. K., Chen, Y., Guan, X., Nousome, D., Sharma, C., Canoll, P., et al. (2014). Genome-Wide Methylation Analyses in Glioblastoma Multiforme. *PLoS One* 9:e89376. doi: 10.1371/journal.pone.0089376

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

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

Liu, H., Hu, B., Chen, L., and Lu, L. (2020). *Identifying protein subcellular location with embedding features learned from networks. Current Proteomics*. Sharjah: Bentham Science Publishers.

Macyszyn, L., Akbari, H., Pisapia, J. M., Da, X., Attiah, M., Pigrish, V., et al. (2016). Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol.* 18, 417–425. doi: 10.1093/neuonc/nov127

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

Muthukrishnan, S., Puri, M., and Lefevre, C. (2014). Support vector machine (SVM) based multiclass prediction with basic statistical analysis of plasminogen activators. *BMC Res. Notes* 7:63. doi: 10.1186/1756-0500-7-63

Numata, S., Ye, T., Hyde, T. M., Guitart-Navarro, X., Tao, R., Wininger, M., et al. (2012). DNA Methylation Signatures in Development and Aging of the Human Prefrontal Cortex. *Am. J. Hum. Genet.* 90, 260–272. doi: 10.1016/j.ajhg.2011.12.020

Pan, X. Y., Zhang, Y. N., and Shen, H. B. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9, 4992–5001. doi: 10.1021/pr100618t

Platt, J. (ed.) (1998). *Fast training of support vector machines using sequential minimal optimization*. Cambridge, MA: MIT Press.

Pradhan, M. P., Desai, A., and Palakal, M. J. (2013). Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst. Biol.* 7:141. doi: 10.1186/1752-0509-7-141

Safe, S., and Abdelrahim, M. (2005). Sp transcription factor family and its role in cancer. *Eur. J. Cancer* 41, 2438–2448. doi: 10.1016/j.ejca.2005.08.006

Safe, S., Imanirad, P., Sreevalsan, S., Nair, V., and Jutooru, I. (2014). Transcription factor Sp1, also known as specificity protein 1 as a therapeutic target. *Expert Opin. Therapeut. Targets* 18, 759–769. doi: 10.1517/14728222.2014.914173

Sang, X., Xiao, W., Zheng, H., Yang, Y., and Liu, T. (2020). HMMPred: Accurate Prediction of DNA-Binding Proteins Based on HMM Profiles and XGBoost Feature Selection. *Comput. Math. Methods Med.* 2020:1384749.

Sharma, A., Bendre, A., Mondal, A., Muzumdar, D., Goel, N., and Shiras, A. (2017). Angiogenic Gene Signature Derived from Subtype Specific Cell Models Segregate Proneural and Mesenchymal Glioblastoma. *Front. Oncol.* 7:146. doi: 10.3389/fonc.2017.00146

Shen, R., Mo, Q., Schultz, N., Seshan, V. E., Olshen, A. B., Huse, J., et al. (2012). Integrative subtype discovery in glioblastoma using iCluster. *PLoS One* 7:e35236. doi: 10.1371/journal.pone.0035236

Skiriute, D., Vaitkiene, P., Saferis, V., Asmoniene, V., Skauminas, K., Deltuva, V. P., et al. (2012). MGMT, GATA6, CD81, DR4, and CASP8 gene promoter methylation in glioblastoma. *BMC Cancer* 12, 218–218. doi: 10.1186/1471-2407-12-218

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., et al. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* 47, 458–468. doi: 10.1038/ng.3273

Triscott, J., Lee, C., Hu, K., Fotovati, A., Berns, R., Pambid, M., et al. (2012). Disulfiram, a drug widely used to control alcoholism, suppresses self-renewal of glioblastoma and overrides resistance to temozolomide. *Oncotarget* 3, 1112–1123. doi: 10.18632/oncotarget.604

Wang, C., Chen, Y., Wang, Y., Liu, X., Liu, Y., Li, Y., et al. (2019). Inhibition of COX-2, mPGES-1 and CYP4A by isoliquiritigenin blocks the angiogenic Akt signaling in glioma through ceRNA effect of miR-194-5p and lncRNA NEAT1. *J. Experimen. Clin. Cancer Res.* 38, 1–14.

Wang, C., Li, Y., Chen, H., Huang, K., Liu, X., Qiu, M., et al. (2018). CYP4X1 inhibition by flavonoid CH625 normalizes glioma vasculature through reprogramming TAMs via CB2 and EGFR-STAT3 Axis. *J. Pharmacol. Experimen. Therapeut.* 365, 72–83. doi: 10.1124/jpet.117.247130

Watkins, S., Robel, S., Kimbrough, I. F., Robert, S. M., Ellis-Davies, G., and Sontheimer, H. (2014). Disruption of astrocyte-vascular coupling and the blood-brain barrier by invading glioma cells. *Nat. Commun.* 5:4196.

Wiers, C. E., Shumay, E., Volkow, N. D., Frieling, H., Kotsiari, A., Lindenmeyer, J., et al. (2015). Effects of depressive symptoms and peripheral DAT methylation on neural reactivity to alcohol cues in alcoholism. *Transl. Psychiatry* 5:e648. doi: 10.1038/tp.2015.141

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

Wu, G., Diaz, A. K., Paugh, B. S., Rankin, S. L., Ju, B., Li, Y., et al. (2014). The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat. Genet.* 46:444. doi: 10.1038/ng.2938

Würth, R., Bajetto, A., Harrison, J. K., Barbieri, F., and Florio, T. (2014). CXCL12 modulation of CXCR4 and CXCR7 activity in human glioblastoma stem-like cells and regulation of the tumor microenvironment. *Front. Cell. Neurosci.* 8:144. doi: 10.3389/fncel.2014.00144

Yadav, V. N., Zamler, D., Baker, G. J., Kadiyala, P., Erdreich-Epstein, A., Decarvalho, A. C., et al. (2016). CXCR4 increases in-vivo glioma perivascular invasion, and reduces radiation induced apoptosis: A genetic knockdown study. *Oncotarget* 7, 83701–83719. doi: 10.18632/oncotarget.13295

Yang, Q., Li, K., Li, X., and Liu, J. (2020). Identification of Key Genes and Pathways in Myeloma side population cells by Bioinformatics Analysis. *Int. J. Med. Sci.* 17:2063. doi: 10.7150/ijms.48244

Yuan, F., Pan, X. Y., Zeng, T., Zhang, Y. H., Chen, L., Gan, Z. J., et al. (2020). Identifying Cell-Type Specific Genes and Expression Rules Based on Single-Cell Transcriptomic Atlas Data. *Front. Bioengine. Biotechnol.* 8:350. doi: 10.3389/fbioe.2020.00350

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, G., Jia, X., Peng, C., and Zhimin, H. (2015). miR-205 inhibits invasion of glioma cells via targeting TBX18. *Chin. J. Pathophysiol.* 31, 1219–1224.

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

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

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: 26 November 2020
doi: 10.3389/fgene.2020.565341

Multi-Omics Analysis Reveals Novel Subtypes and Driver Genes in Glioblastoma

Yang Yuan[1†], Pan Qi[2†], Wang Xiang[1], Liu Yanhui[1], Li Yu[3] and Mao Qing[1]**

[1] Department of Neurosurgery, West China Hospital, Sichuan University, Chengdu, China, [2] Department of Dermatology, Chongqing Traditional Chinese Medicine Hospital, Chongqing, China, [3] Department of Anesthesia, West China Hospital, Sichuan University, Chengdu, China

OPEN ACCESS

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

Reviewed by:
Yanding Zhao,
Baylor College of Medicine,
United States
Zhengwang Sun,
Harvard Medical School,
United States
Sha Li,
Sanford Burnham Prebys Medical Discovery Institute, United States
Jianfeng Shen,
Shanghai Jiao Tong University, China

***Correspondence:**
Li Yu
liyu@wchscu.cn
Mao Qing
qingmao2000@163.com

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

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

Received: 08 June 2020
Accepted: 26 October 2020
Published: 26 November 2020

Citation:
Yuan Y, Qi P, Xiang W, Yanhui L, Yu L and Qing M (2020) Multi-Omics Analysis Reveals Novel Subtypes and Driver Genes in Glioblastoma. Front. Genet. 11:565341. doi: 10.3389/fgene.2020.565341

Glioblastoma is the most lethal malignant primary brain tumor; nevertheless, there remains a lack of accurate prognostic markers and drug targets. In this study, we analyzed 117 primary glioblastoma patients' data that contained SNP, DNA copy, DNA methylation, mRNA expression, and clinical information. After the quality of control examination, we conducted the single nucleotide polymorphism (SNP) analysis, copy number variation (CNV) analysis, and infiltrated immune cells estimate. And moreover, by using the cluster of cluster analysis (CoCA) methods, we finally divided these GBM patients into two novel subtypes, HX-1 (Cluster 1) and HX-2 (Cluster 2), which could be co-characterized by 3 methylation variable positions [cg16957313(DUSP1), cg17783509(PHOX2B), cg23432345(HOXA7)] and 15 (PCDH1, CYP27B1, LPIN3, GPR32, BCL6, OR4Q3, MAGI3, SKIV2L, PCSK5, AKAP12, UBE3B, MAP4, TP53BP1, F5, RHOBTB1) gene mutations pattern. Compared to HX-1 subtype, the HX-2 subtype was identified with higher gene co-occurring events, tumor mutation burden (TBM), and poor median overall survival [231.5 days (HX-2) vs. 445 days (HX-1), *P*-value = 0.00053]. We believe that HX-1 and HX-2 subtypes may make sense as the potential prognostic biomarkers for patients with glioblastoma.

Keywords: multi-omics analysis, copy number variation, DNA methylation, mRNA expression, glioblastoma

INTRODUCTION

Gliomas are most common malignant brain tumors which derive from neuroepithelial cells (Rivera et al., 2008). Most patients underwent tumor resection surgery with standard follow-up chemotherapy/radiotherapy, and based on molecular neuropathology diagnosis, they may survive from months to decades (median survival from 1 year to 15 years) (Marton et al., 2019). High-grade gliomas' recurrence was due to their invasive nature. Recent studies on molecular pathology of glioma has outlined some valuable prognosis biomarkers such as IDH1, 1q-19p co-deletion, h3k27, TERT (Killela et al., 2014; Marton et al., 2019; Zhang Z. Y. et al., 2019), but the existed biomarkers still cannot fully predict the overall survival for all glioblastoma patients, such as IDH1 wild-type in WHO grade 2 gliomas or in recurrent gliomas; moreover, we know a little about of the MGMT demethylation status in glioma patients. Unlike many other types of malignant tumor, glioblastoma lacks of effective treatment measures and drug targets (Snape and Warr, 2015; Higashijima and Kanki, 2019; Ruta et al., 2019). Recent phase II/III clinical trials on glioblastoma were all failed,

including immune checkpoint inhibitor PD-1 or PD-L1 (Berghoff and Preusser, 2016; Charoentong et al., 2017; Kurz et al., 2018) or anti-angiogenic drugs like bevacizumab (Kurz et al., 2018; Moriya et al., 2018). Life is composed of complicated regulator control system, the cancer happened normally involved in gene mutation, change of epigenetics and gain of fusion-gene (Liang et al., 2019). Thus, through integrating analysis of multi-omics data on glioblastoma is meaningful, which could systematically study the negative molecular event like genomic instability and somatic mutation (Song et al., 2019; Zhang Z. Y. et al., 2019). In this study, we performed integrated analysis via TCGA database of glioblastoma [(NIH), Genomic Data Commons database (GDC)[1]], aimed to complete a new molecular classification and provide some new treatment targets for GBM. As a result, we enrolled 117patients that all contained SNP, DNA copy, DNA methylation and mRNA expression profile data. After combined the multidimensional data with clinical information and cluster of clusters analysis steps, we divided theses GBM samples into two novel subtypes (HX-1 and HX-2), among the two subtypes, we identified 15 genes and 3 methylation variable position which are associated with overall survival, and the subtype HX-2 has an obvious higher mutation frequency than subtype HX-1, moreover, the NK cells activated rate in HX-2 is also higher than HX-1 group.

RESULTS

Mutation Analysis Reveals

As the first step, we performed statistical analysis for the enrolled 117GBM samples, annotated the mutation types, depicted proportion of different types of base changes and the top 10 mutation genes. Among these patients, the median age at initial diagnosis was 62(from 21 to 89), and 44 of them are female, more details of each patients could find in **Supplementary Table 1**. The overall description of the results is revealed in **Figure 1A**. In glioblastoma, the most common mutation type is C > T. **Figures 1A,B** have displayed the 20 most mutated genes and metadata information such as molecular subtypes information. **Figure 1C** disclosed the frequency distribution of the top20 gene mutations in GBM, the gene with the highest mutation rate is PTEN, 56% of samples had gene mutation on PTEN.

We separately counted the number of somatic mutations in each GBM sample and matched the clinical characteristics of these samples, the clinical features including survival status, tumor recurrence, etc. The analysis results indicated that the somatic mutations between tumor recurrence and progression of disease existed huge difference, and the recurrence samples has a higher number of mutations (**Figure 1D**).

Somatic mutations are widespread events in tumorigenesis, a few of gene mutations could directly cause tumor happening, and those genes are called driver genes (Higashijima and Kanki, 2019). We used MutSigCV to predict driver gene of the samples based on mutation data. When the significance threshold was $q < 0.01$, a total of 925 candidate genes were obtained.

Considering the mutation site of each sample and the bases at 1 bp position upstream and downstream of the mutation site, we divided the mutation into 96 types according to the upstream and downstream mutation site, calculated the frequency distribution of the 96 mutation types of the 117 sample (**Figure 1E**). Moreover, somatic mutations are present in all cells of the human body and occur throughout life. They are the consequence of multiple mutational processes, including the intrinsic slight infidelity of the DNA replication machinery, exogenous or endogenous mutagen exposures, enzymatic modification of DNA and defective DNA repair. Different mutational processes generate unique combinations of mutation types, termed "Mutational Signatures"[2]. In this study, to figure out the relationship between the mutation spectrum distribution of GBM samples and mutational signatures in *cosmic*, we subsequently conducted non-negative matrix factorization analysis based on 96 mutation types of the 117 sample, and extracted three somatic point mutations (**Figure 1F**). We found that the glioblastoma mutation spectrums are mainly related to signature_27 like, signature_1 like and signature_10 like.

Copy Number Variation Analysis

A total of 117 samples were conducted by GISTIC analysis. The results suggested that 7q,7p,19p amplification and 10q,10p, 22q deletion are most notable, and **Figure 2A** revealed the chromosome arms when GISTIC test significant (*Q*-value $< 10^{-5}$). In all tumor samples, there were 10 amplifications and 21 copy number deletions in minimal common regions (MCRs), these MCRs are showed in **Figures 2B,C**, among them, the most significant amplification position are 7p11.2, 12q14.1, the most significant deletion position are 9p21.3,10q23.31. **Figure 2D** revealed the minimal common regions (MCRs) and the genes within the MRCs (the deletion genes in the region is represented by a negative value).

We then used ABSOLUTE software to evaluate the tumor purity and ploidy based on copy number variation (CNV), as showed in **Figures 2E,F**, the tumor purity ranged from 0.16–1, and the tumor cell genome ploidy was ranged from 1.82–10.13, which suggested that genomic disorder is a common event in tumorigenesis.

Clustering by Integrated Platforms

We utilized four single platform data (SNP,DNA copy,DNA methylation,mRNA expression profile) to integrate with clinical information. When the significance threshold is set to 0.01($q < 0.01$), 333 gene mutations, 60 DNA methylation sites, and 123 mRNAs are associated with prognosis of GBM patients; however, there were no significant CNV position with prognosis in our array. According to the expression of 123 mRNAs, the samples can be divided into 3 categories (**Figure 3A**). According to the information of 60 methylation sites, the samples can be divided into 2 subtypes (**Figure 3B**). According to the mutation information of 333 genes the samples can be divided into 2 subtypes (**Figure 3C**).

[1]https://gdc.cancer.gov/

[2]https://cancer.sanger.ac.uk/cosmic/signatures

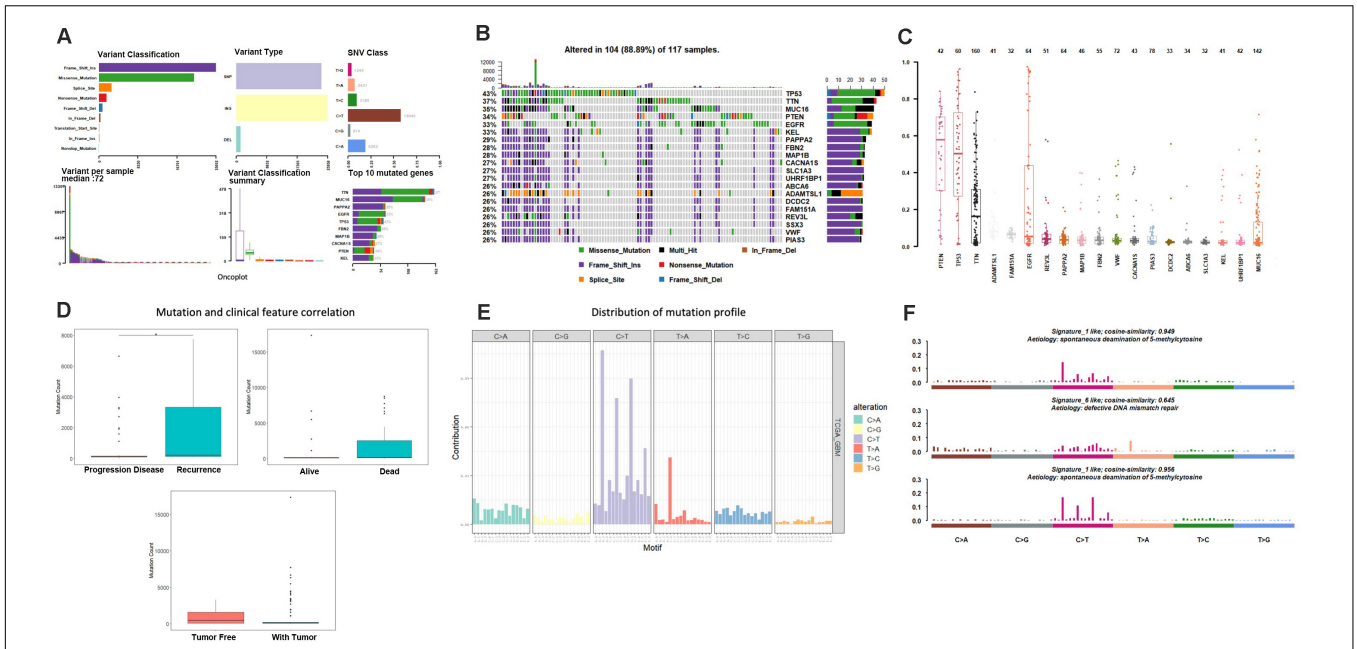

FIGURE 1 | Mutation analysis for enrolled samples. **(A)** Tumor mutation profile of glioblastoma samples. **(B)** Oncoplot with the top 20 most mutated genes. **(C)** Frequency distribution of the top 20 gene mutations in glioblastoma. **(D)** Mutation and clinical feature correlation. **(E)** Distribution of mutation profile. **(F)** Mutation correlation of character and cosmic mutation signature.

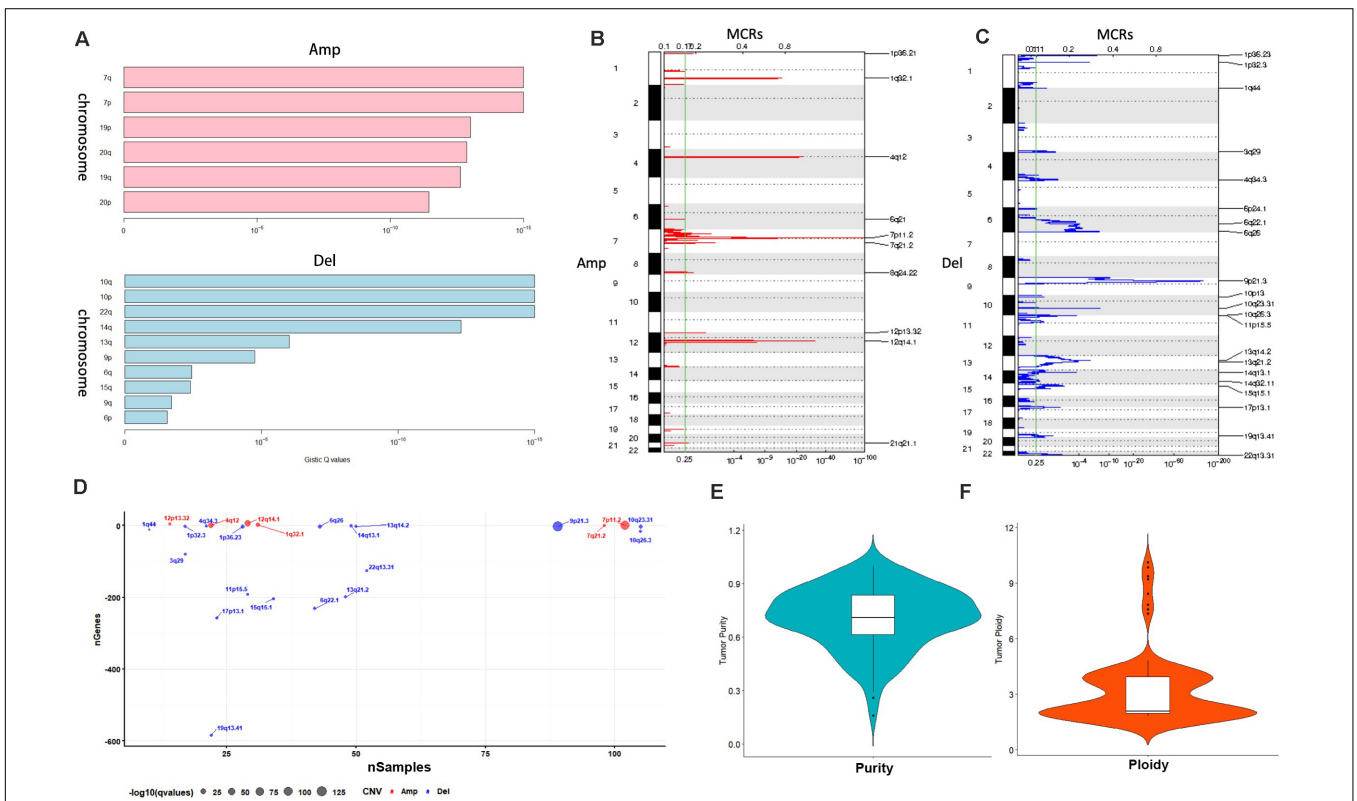

FIGURE 2 | Copy number variation analysis. **(A)** GISTIC analysis for enrolled samples, upper panel: amplification of chromosome arm, lower panel: deletion of chromosome arm. **(B,C)** Distribution of minimal common regions (MCRs). **(D)** The number of genes in minimal common regions (the number of genes in the missing region is represented by a negative value). **(E,F)** Purity and ploidy analysis.

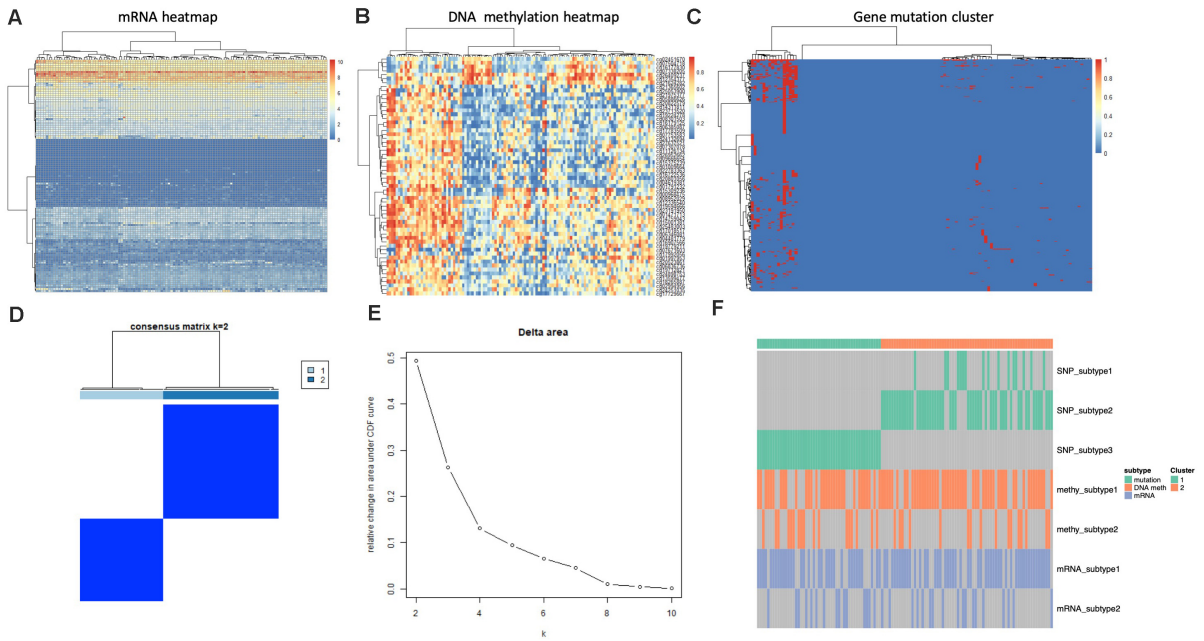

FIGURE 3 | Identification of subtype with cluster of different platforms. **(A)** Clustering mRNA expression profile. **(B)** Clustering of DNA methylation. **(C)** Clustering of gene mutation. **(D,E)** CoCA cluster analysis, all samples can be divided into 2 subtypes, $K = 2$, HX-1(Cluster 1) and HX-2 (Cluster 2) subtype. **(F)** Subtype classification with single platform clustering results display.

We next used CoCA cluster analysis method to conduct cluster analysis again, the data was derived from SNP, DNA methylation and mRNA platform, finally, we obtained two novel subtypes from all GBM samples, and we named these subtypes as HX-1 and HX-2 (**Figure 3D**). **Figure 3E** represents the delta area curve of consensus clustering, indicating the relative change in area under the cumulative distribution function (CDF) curve for each category number k compared with k-1. When the subtypes were classified into two groups ($K = 2$), the area under the cure is biggest. We also plot the information of the subtypes with each platform (**Figure 3F**). It suggests that HX1 and HX2 are more correlated to SNP1, SNP2, and SNP3, but not correlated to methylation subgroups or mRNA subgroups.

Subtype Analysis

Firstly, we analyzed the clinical features for each subtype, such as gender, tumor status, survival status and the median survival time etc. (**Figure 4A** and **Table 1**), the median survival time between each group (HX-1 and HX-2) has significant differences ($P = 0.00053$), the data indicated that the HX-2 had an obviously poor OS (**Figure 4B**), the median OS for HX-1 is 445 days, and the median OS for HX-2 is 231.5 days, P-value is 0.00053.

We further want to identify whether each subtype differs in the type of mutation, as shown in **Figures 4C,D, HX-1 and HX-2** were mainly happened as C > T mutation, and Ti (transition) frequency was higher than TV (transversion) frequency (**Figures 4C,D**).

Many genes that cause cancers often with mutually exclusive or co-occurring events, in order to determine which genes will happen with mutually exclusive or co-occurring events, we conduct Fisher's exact test for any two gene mutations, and we found a plenty of gene co-occurring events in HX-2 subtype instead of HX-1 (**Figures 4E,F**).

APOBEC ("apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like") is a family of evolutionarily conserved cytidine deaminases. In humans, they help protect from viral infections. These enzymes, when misregulated, are a major source of mutation in numerous cancer types (Rebhandl et al., 2015). We used R package maftools to proceed APOBEC analysis. As shown in **Figure 4G**, only subtype HX-2 had APOBEC cluster samples, the genes with mutation rate which significantly high were revealed in **Figure 4G**, the box plot shows differences in mutation load between APOBEC-enriched and non-enriched samples, donut plots display the proportion of mutations in tCw context, bar plots show the top 10 differentially mutated genes between APOBEC-enriched and non-APOBEC- enriched samples.

We also compared the 96 signatures collected in cosmic with each subtype; as a result, the mutational signatures in each subtype were both associated with signature1 (**Figure 4H**), but the HX-1 had high similarity with signature6 subtype independently; the HX-2 subtype also had high similarity with signature10 and signature 27 (**Figure 4I**).

In order to identify the gene mutations for each subtype, we counted the total amount of mutations in each subgroup of each gene, and then conduct chi-square test. Finally, we identified 727 different mutations in reach subtypes, the subtype HX-2 had an obvious higher mutation rate than HX-1 (**Figure 4J**). We subsequent counted the tumor mutation burden (TBM) for each subtype, the result confirmed the TBM in HX-2 (TBM = 55.4) is significantly higher than HX-1

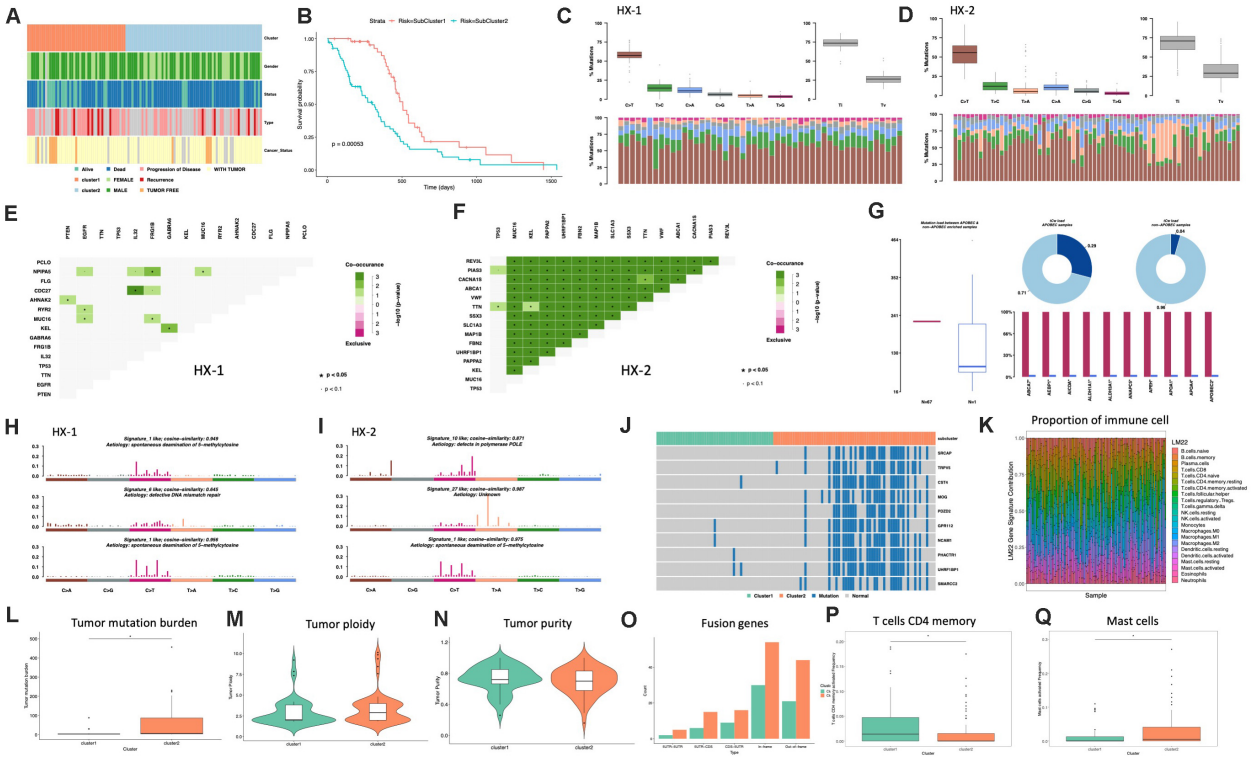

FIGURE 4 | Analysis of subtypes. **(A)** Clinical characters of subtypes. **(B)** Survival curve between subtypes. **(C)** Distribution of mutation types of HX-1. **(D)** Distribution of mutation types of HX-2. **(E)** Mutually exclusive or co-occurring events in HX-1. **(F)** Mutually exclusive or co-occurring events in HX-2. **(G)** Cluster of APOBEC ("apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like") analysis of HX-2. **(H)** Mutation signatures of HX-1. **(I)** Mutation signatures of HX-2. **(J)** Distribution of significant difference genes between HX-1 and HX-2. **(K)** Distribution of immune cells. **(L)** Tumor mutation burden comparison. **(M,N)** Purity and ploidy analysis between HX-1 and HX-2. **(O)** Fusion genes events in HX-1 and HX-2. **(P)** Distribution of T cells CD4 memory activated between HX-1 and HX-2. **(Q)** Distribution of mast cells activated between HX-1 and HX-2.

(TMB = 5.7, P = 3.881e-06, **Figure 4L**). There was no significant difference of each subtype on tumor ploidy (**Figure 4M**) and purity (**Figure 4N**).

We download fusion gene baseline from http://54.84.12.177/PanCanFusV2/database. In total, we identified 144 fusion genes in HX-1 cluster and 284 fusion gene in HX-2 (**Figure 4O**, **Supplementary Table 2**). We uploaded the expression data of 117 GBM samples to cibersort website, calculated the proportion of 22 immune cells in these samples (**Figure 4K**). Then, the distribution of the proportion of each immune cell between the two subgroups was calculated separately, we determined that proportion of T cells CD4 memory activated (**Figure 4P**) and mast cells activated (**Figure 4Q**) was significant different between HX-1 and HX-2.

TABLE 1 | Clinical features for each subtype.

	HX-1	HX-2	*P*-value
Female	16	28	0.5072
Male	32	40	
Not available	1	0	
Alive	18	12	0.0285
Dead	30	56	
Not available	1	0	
Progression of disease	24	25	0.6217
Recurrence	8	5	
Not available	17	38	
Tumor free	6	4	0.4655
With tumor	39	53	
Not available	4	11	

Prognostic Marker Identification and Validation

In order to further identify of the prognostic markers for the subtypes, we conjointly analyzed the 19 DE genes, 27 DE methylation position and 727 DE genes between HX-1 and HX-2. The analysis results show that when the significance threshold is set to 0.05, there had three methylation positions [cg16957313(DUSP1), cg17783509(PHOX2B),cg23432345(HOXA7)] and 21 genes were associated with prognosis, in which 15 genes were same as in Mut2SigCV analysis. In addition, the distribution of all GBM cases based on TCGA is displayed in **Supplementary Figure 1** according to the mutation signature of these 15 genes. The survival curve of these 21 associated prognosis factors

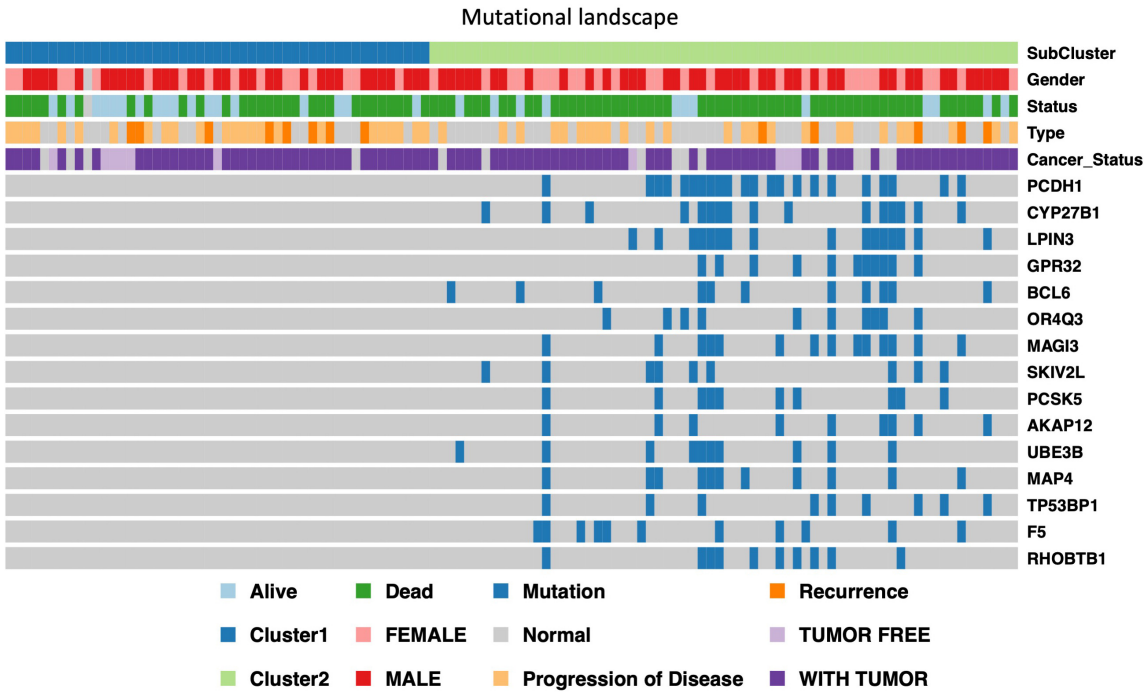

FIGURE 5 | Display of landscape between HX-1 and HX-2, the clinical features and mutation status of 15 driver gene were integrated into subtypes, it can be clearly seen from the figure that the frequency of gene mutations in HX-2 subtype is much higher than that of HX-1.

were showed in **Supplementary Figure 2** We also described the landscape of the 15 genes between the two subtypes (**Figure 5**).

Moreover, to validate the outcome of our analysis, the 15 mutant genes mutation signature genes used to develop a cancer-related risk signature. Samples from the Chinese Glioma Genome Atlas (CGGA) dataset were divided into high risk group and low risk group. These samples carrying mutations within 15 genes were defined as high-risk group ($n = 11$) in CGGA primary GBM cohort; while the others were defined as low-risk group ($n = 42$). According to the Kaplan-Meier survival analysis, the prognosis of high-risk group was strikingly worse than that of low-risk group (**Supplementary Figure 3A**, $P = 0.032$). Moreover, the 15 gene signature in the CGGA primary GBM cohort showed a high area under the receiver operating characteristic curve (AUC = 0.632) (**Supplementary Figure 3B**), close to that in the TCGA GBM cohort (AUC = 0.756) (**Supplementary Figure3C**).

DISCUSSION

Glioblastomas (GBM) is the most invasive and prevalent types of glioma with extremely poor prognosis and limited treatment options (Rebhandl et al., 2015). In recent years, tremendous articles reported the molecular characterization of GBM, make us better understanding of how to use the key molecules to predict the OS for glioma patients (Colaprico et al., 2016; Holdhoff, 2018; Higashijima and Kanki, 2019; Marton et al., 2019). However, most of the published articles were based on single platform analysis, which is hard to explain why the similar molecular pattern may induce diverse prognosis in GBM patients sometimes. In order to make a comprehensive understanding on molecular characteristic of GBM, we used the unsupervised clustering method to cluster the data from four different platforms (SNP,DNA copy,DNA methylation,mRNA expression) and subsequently used the cluster of clusters analysis (CoCA) method to further identify the subtypes of GBM. Therefore, through systematic studying of the integrated multi-omics analysis, genomic instability, somatic mutation and the molecular characteristics of each GBM subgroup, we hope we can provide new ideas and novel theoretical basis for early diagnosis and individualized treatment for GBM patients.

We conducted the SNP associated analysis in the first step, our result showed the glioblastoma was characterized by prominence of C > T. The signatures of mutational processes in human cancer was firstly reported by Michael R. Stratton and his colleagues, they concluded more than 20 distinct mutational signatures from 4,938,362 mutations from 7,042 cancers (Wu et al., 2019). We extracted the mutation characteristics of somatic point mutations, the result showed that the mutation spectrum of glioblastoma is similar to signature27,signature1 and signature10 which collected in cosmic. As reported, the Signature 1A/B is probably related to the relatively elevated rate of spontaneous deamination of 5-methyl-cytosine which results in C > T transitions and which predominantly occurs at NpCpG trinucleotides, and signature10 was the associated with altered activity of the error-prone polymerase Pol ε consequent on

mutations in the gene. However, the reason for signature 27 is still unknown.

We next use clusters analysis (CoCA) method to classified the subtype of enrolled glioblastoma samples as HX-1 and HX-2, the main mutation signature of the two subtypes are the same as C > T, however, there were a plenty of gene co-mutation events in HX-2 but not shown in HX-1, the Tumor mutation burden in HX-2 was significant higher that HX-1, and the median survival forHX-2 is 231.5 days, much shorter than HX-1 445 days, suggested that the HX-2 subtype is more aggressive than HX-1 subtype, and HX-2 occurred from high frequency of gene mutation.

The proportion of *T cells CD4 memory activated* and *mast cells activated* were determined significant difference between HX-1 and HX-2 in our result. Dongrui Wang et al. found that maintenance of the CD4 + subset was positively correlated with the recursive killing ability of CAR T cell products derived from GBM patients (Alexandrov et al., 2013). His finding identified CD4 + CAR T cells as a highly potent and clinically important T cell subset for effective CAR therapy. This may probably explain why the HX-1 had the better prognosis. Moreover, recent research indicated that mast cells (MCs) upon activation by glioma cells produce soluble factors including IL-6, which are documented to be involved in cancer-related activities and promoted glioma cell differentiation and growth (Wang et al., 2018). It was also figured out that MCs exert their effect via inactivation of STAT3 through GSK3β downregulation. This could probably explain why the HX-2 cluster had the shorter OS.

We further analyzed the negatively regulative biomarkers which may distinguish the OS of HX-2 from HX-1, and we identified 3 methylation variable positions [cg16957313(DUSP1), cg17783509(PHOX2B), cg23432345(HOXA7)] and 15 genes (PCDH1, CYP27B1, LPIN3, GPR32, BCL6, OR4Q3, MAGI3, SKIV2L, PCSK5, AKAP12, UBE3B, MAP4, TP53BP1, F5, RHOBTB1) that may induce poor overall survival for HX-2. Some of these genes have been reported to be associated with the malignant behavior of glioblastoma. For example, studies have shown that BCL6 is essential for the survival of GBM cells (Attarha et al., 2017), the overexpression of BCL6 is associated with poor prognosis for glioma patients, BCL6 gene could inhibits the expression of wild-type p53 and its target genes in GBM cells. In gliomas, the expression levels of MAGI3 and PTEN were reported significantly down-regulated, and for glioma C6 cell line, overexpressed MAGI3 will inhibits Akt phosphorylation, and inhibits cell proliferation (Xu et al., 2017). We also identified some novel genes which are still not been reported, such as PCDH1, LPIN3, GPR32, SKIV2L, PCSK5.

In this study, we used a comprehensive bioinformatics method to integrate 4 platform data of glioblastoma, and further identified two novel subtypes of glioblastoma which could be characterized by the cluster of 3 methylation variable position and 15 gene mutation, the multi-omic signatures for the prognosis of glioblastoma developed by us were also be validate in CGGA independent dataset. We hope that our research could provide potential stratification marker for clinical outcome and new theoretical basis for glioblastoma.

MATERIALS AND METHODS

TCGA Data Acquisition

The TCGAbiolinks R package was used to help us obtain patients data from the National Institutes of Health (NIH), Genomic Data Commons database (GDC)[3] (Holdhoff, 2018). Briefly, we get 577 SNP6 Copy Number segment GBM samples data and 411 samples methylation microarrays data from the website http://firebrowse.org/, and we also downloaded 154 GBM samples mRNA expression data from https://portal.gdc.cancer.gov/. After filtrate these data and link sample information, there are 117 sample contained multi-omics data, which means all the filtered samples contained gene mutation data, CNV data, methylation data and mRNA expression data. Our subsequent analysis was based on these data. The fusion gene result subsequently used was acquired from TUMOR FUSION GENE DATA PORTAL database[4].

Single Nucleotide Polymorphism (SNP) Analysis

MutSigCV module in GenePattern was used to analysis the driver gene in GBM[5] (Ma et al., 2015). There are strong correlations between somatic mutation frequencies in cancers and both gene expression level and replication time of a DNA region during the cell cycle, MutsigCV analysis could substantially reduce the number of false positives, especially when applied to tumor samples that have high mutation rates.

We use the maftools R package[6] (Lawrence et al., 2013) and SomaticSignatures[7] (Mayakonda et al., 2018) to conduct mutation analysis and plot the mutation spectrum and characteristics.

Copy Number Variation Analysis

GISTIC module in GenePattern was also used to extract the landmark CNV events in GBM, the parameters in GISTIC algorithm were set as follows: Q-value < 0.05 as statistics significance, confidence levels were set as 95% to confirm peak region. Chromosome arm length was set as 0.98 when analyzed the chromosome arm mutation. Tumor purity and ploidy character were analyzed by the R package R ABSOLUTE[8].

Subtype Identification of Glioblastoma

Unsupervised clustering was proceeded based on the three different platforms (SNP, DNA methylation, and mRNA expression profile) and molecules associated of overall survival, the we conduct the clustering again based on a method called

[3]https://gdc.cancer.gov/

[4]https://tumorfusions.org/PanCanFusV2/

[5]https://cloud.genepattern.org/gp/pages/index.jsf

[6]https://bioconductor.org/packages/release/bioc/html/maftools.html

[7]https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html

[8]https://software.broadinstitute.org/cancer/cga/absolute_download

cluster of clusters analysis (CoCA) (Hoadley et al., 2014; Gehring et al., 2015). Briefly, Subtype calls from each of the 4 platforms analyzed for subtypes within each data type were used to identify relationships between the different classifications. Subtypes defined from each platform were coded into a series of indicator variables for each subtype. The matrix of 1 and 0s was used in ConsensusClusterPlus R package (Gehring et al., 2015) to identify structure and relationship of the samples. Parameters for Consensus cluster were 80% sample resampling with 1000 iterations of hierarchical clustering based on a Pearson correlation distance metric. and ultimately, we acquired the two subtypes result from glioblastoma that integrated the data of different platforms. We named these two subtypes as subtype HX-1 and subtype HX-2.

Characteristic Analysis of Subtypes

Chi-square test was used to the characteristic analysis of GBM subtypes, including survival state and progression of disease.

R package limma[9] (Smaglo et al., 2015) was conduct to screen the valuable biomarkers within the subgroups, we tried to filter the difference expressed (DE) mRNA and methylation variable positions (MVPs), and finally proceed KEEG and GO analysis for those DE mRNA and MVPs.

We also utilized the maftools to map the gene mutation characteristic in GBM subtypes, including C>T, T>C, C>A, T>G, C>G, T>A, Ti(transition) and TV (transversion). Moreover, mutation signature analysis and APOBEC enrichment analysis (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) were also conduct between the subtypes.

Infiltrated Immune Cells Estimate

Tumor immune cell infiltration refers to the migration of immune cells from the blood to the tumor tissue and begins to exert its effects. The infiltration of immune cells in tumor directly affects the overall survival in GBM patients. Thus, to quantify the proportion of immune cells in the enrolled samples, we used CIBERSORT algorithm (Ritchie et al., 2015; Chen et al., 2018; Zhang L. et al., 2019) and LM22 algorithm (Charoentong et al., 2017), and calculated the percentage of 22 types of human immune cells in GBM, concluding the B cells, T cells, natural killer cells, macrophages and dendritic cells.

[9]https://bioconductor.org/packages/release/bioc/html/limma.html

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material.**

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was sponsored by the National Natural Science Young Foundation of China (Grant no. 81904218 to PQ and Grant no. 81902532 to YY) and Innovation and Sparkle Project of Sichuan University (Grant No. 2082604401004/060 to YY).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.565341/full#supplementary-material

Supplementary Figure 1 | Mutation signature of these 15 DE genes.

Supplementary Figure 2 | Survival curves of the three methylation variable position [cg16957313(DUSP1), cg17783509(PHOX2B), cg23432345(HOXA7)] and 15 genes (PCDH1, CYP27B1, LPIN3, GPR32, BCL6, OR4Q3, MAGI3, SKIV2L, PCSK5, AKAP12, UBE3B, MAP4, TP53BP1, F5, RHOBTB1).

Supplementary Figure 3 | Prognostic marker validation. **(A)** Kaplan–Meier survival analysis of Chinese Glioma Genome Atlas (CGGA) dataset. **(B)** Operating characteristic curve (AUC = 0.632) of CGGA primary GBM cohort. **(C)** Operating characteristic curve (AUC = 0.756) of TCGA GBM cohort.

REFERENCES

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

Attarha, S., Roy, A., Westermark, B., and Tchougounova, E. (2017). Mast cells modulate proliferation, migration and stemness of glioma cells through downregulation of GSK3beta expression and inhibition of STAT3 activation. *Cell Signal.* 37, 81–92. doi: 10.1016/j.cellsig.2017.06.004

Berghoff, A. S., and Preusser, M. (2016). In search of a target: PD-1 and PD-L1 profiling across glioma types. *Neuro. Oncol.* 18, 1331–1332. doi: 10.1093/neuonc/now162

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M., and Alizadeh, A. A. (2018). Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol. Biol.* 1711, 243–259.

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

Gehring, J. S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* 31, 3673–3675.

Higashijima, Y., and Kanki, Y. (2019). Molecular mechanistic insights: The emerging role of SOXF transcription factors in tumorigenesis and development. *Semin. Cancer Biol.* S1044-579X, 30146–30149.

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

Holdhoff, M. (2018). Role of Molecular Pathology in the Treatment of Anaplastic Gliomas and Glioblastomas. *J. Natl. Compr. Canc. Netw.* 16, 642–645. doi: 10.6004/jnccn.2018.0045

Killela, P. J., Pirozzi, C. J., Healy, P., Reitman, Z. J., Lipp, E., Rasheed, B. A., et al. (2014). Mutations in IDH1, IDH2, and in the TERT promoter define clinically distinct subgroups of adult malignant gliomas. *Oncotarget* 5, 1515–1525. doi: 10.18632/oncotarget.1765

Kurz, S. C., Cabrera, L. P., Hastie, D., Huang, R., Unadkat, P., Rinne, M., et al. (2018). PD-1 inhibition has only limited clinical benefit in patients with recurrent high-grade glioma. *Neurology* 91:e1355–e1359.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218.

Liang, Q., Li, X., Guan, G., Xu, X., Chen, C., Cheng, P., et al. (2019). Long non-coding RNA, HOTAIRM1, promotes glioma malignancy by forming a ceRNA network. *Aging* 11, 6805–6838. doi: 10.18632/aging.102205

Ma, Q., Zhang, Y., Meng, R., Xie, K. M., Xiong, Y., Lin, S., et al. (2015). MAGI3 Suppresses Glioma Cell Proliferation via Upregulation of PTEN Expression. *Biomed. Environ. Sci.* 28, 502–509.

Marton, E., Giordan, E., Siddi, F., Curzi, C., Canova, G., Scarpa, B., et al. (2019). Over ten years overall survival in glioblastoma: A different disease? *J. Neurol. Sci.* 408:116518. doi: 10.1016/j.jns.2019.116518

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

Moriya, S., Ohba, S., Adachi, K., Nishiyama, Y., Hayashi, T., Nagahisa, S., et al. (2018). A retrospective study of bevacizumab for treatment of brainstem glioma with malignant features. *J. Clin. Neurosci.* 47, 228–233. doi: 10.1016/j.jocn.2017.10.002

Rebhandl, S., Huemer, M., Greil, R., and Geisberger, R. (2015). AID/APOBEC deaminases and cancer. *Oncoscience* 2, 320–333. doi: 10.18632/oncoscience.155

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

Rivera, A. L., Pelloski, C. E., Sulman, E., and Aldape, K. (2008). Prognostic and predictive markers in glioma and other neuroepithelial tumors. *Curr. Probl. Cancer* 32, 97–123. doi: 10.1016/j.currproblcancer.2008.02.003

Ruta, V., Longo, C., Boccaccini, A., Madia, V. N., Saccoliti, F., Tudino, V., et al. (2019). Inhibition of Polycomb Repressive Complex 2 activity reduces trimethylation of H3K27 and affects development in Arabidopsis seedlings. *BMC Plant Biol.* 19:429. doi: 10.1186/s12870-019-2057-7

Smaglo, B. G., Tesfaye, A., Halfdanarson, T. R., Meyer, J. E., Wang, J., Gatalica, Z., et al. (2015). Comprehensive multiplatform biomarker analysis of 199 anal squamous cell carcinomas. *Oncotarget* 6, 43594–43604. doi: 10.18632/oncotarget.6202

Snape, T. J., and Warr, T. (2015). Approaches toward improving the prognosis of pediatric patients with glioma: pursuing mutant drug targets with emerging small molecules. *Semin. Pediatr. Neurol.* 22, 28–34. doi: 10.1016/j.spen.2014.12.003

Song, J., Song, F., Liu, K., Zhang, W., Luo, R., Tang, Y., et al. (2019). Multi-omics analysis reveals epithelial-mesenchymal transition-related gene FOXM1 as a novel prognostic biomarker in clear cell renal carcinoma. *Aging* 11, 10316–10337. doi: 10.18632/aging.102459

Wang, D., Aguilar, B., Starr, R., Alizadeh, D., Brito, A., Sarkissian, A., et al. (2018). Glioblastoma-targeted CD4+ CAR T cells mediate superior antitumor activity. *JCI Insight* 3:e99048.

Wu, F., Chai, R. C., Wang, Z., Liu, Y. Q., Zhao, Z., Li, G. Z., et al. (2019). Molecular classification of IDH-mutant glioblastomas based on gene expression profiles. *Carcinogenesis* 40, 853–860. doi: 10.1093/carcin/bgz032

Xu, L., Chen, Y., Dutra-Clarke, M., Mayakonda, A., Hazawa, M., Savinoff, S. E., et al. (2017). BCL6 promotes glioma and serves as a therapeutic target. *Proc. Natl. Acad. Sci. U S A.* 114, 3981–3986. doi: 10.1073/pnas.1609758114

Zhang, L., Liu, Z., Li, J., Huang, T., Wang, Y., Chang, L., et al. (2019). Genomic analysis of primary and recurrent gliomas reveals clinical outcome related molecular features. *Sci. Rep.* 9, 1–8.

Zhang, Z. Y., Zhan, Y. B., Zhang, F. J., Yu, B., Ji, Y. C., Zhou, J. Q., et al. (2019). Prognostic value of preoperative hematological markers combined with molecular pathology in patients with diffuse gliomas. *Aging* 11, 6252–6272. doi: 10.18632/aging.102186

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: 03 February 2021
doi: 10.3389/fgene.2021.632761

A Novel XGBoost Method to Infer the Primary Lesion of 20 Solid Tumor Types From Gene Expression Data

Sijie Chen[1], Wenjing Zhou[2], Jinghui Tu[1], Jian Li[1], Bo Wang[3,4], Xiaofei Mo[3,4], Geng Tian[3,4], Kebo Lv[1] and Zhijian Huang[5]**

[1] Department of Mathematics, Ocean University of China, Qingdao, China, [2] Department of Oncology, Hiser Medical Center of Qingdao, Qingdao, China, [3] Qingdao Geneis Institute of Big Data Mining and Precision Medicine, Qingdao, China, [4] Geneis Beijing Co., Ltd., Beijing, China, [5] Department of Breast Surgical Oncology, Fujian Cancer Hospital & Fujian Medical University Cancer Hospital, Fuzhou, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Xue Wang,
University of Texas Southwestern Medical Center, United States
Yunpeng Xu,
Rutgers, The State University of New Jersey, United States
Liuyi Hao,
University of North Carolina at Greensboro, United States

***Correspondence:**
Kebo Lv
kewave@ouc.edu.cn
Zhijian Huang
ehuang27@fjmu.edu.cn

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

Received: 24 November 2020
Accepted: 06 January 2021
Published: 03 February 2021

Citation:
Chen S, Zhou W, Tu J, Li J, Wang B, Mo X, Tian G, Lv K and Huang Z (2021) A Novel XGBoost Method to Infer the Primary Lesion of 20 Solid Tumor Types From Gene Expression Data. Front. Genet. 12:632761. doi: 10.3389/fgene.2021.632761

Purpose: Establish a suitable machine learning model to identify its primary lesions for primary metastatic tumors in an integrated learning approach, making it more accurate to improve primary lesions' diagnostic efficiency.

Methods: After deleting the features whose expression level is lower than the threshold, we use two methods to perform feature selection and use XGBoost for classification. After the optimal model is selected through 10-fold cross-validation, it is verified on an independent test set.

Results: Selecting features with around 800 genes for training, the R^2-score of a 10-fold CV of training data can reach 96.38%, and the R^2-score of test data can reach 83.3%.

Conclusion: These findings suggest that by combining tumor data with machine learning methods, each cancer has its corresponding classification accuracy, which can be used to predict primary metastatic tumors' location. The machine-learning-based method can be used as an orthogonal diagnostic method to judge the machine learning model processing and clinical actual pathological conditions.

Keywords: tumor tissue-of-origin, gene expression, XGBoost, feature selection, CUP

INTRODUCTION

Metastatic cancer is a metastatic malignant tumor that has been confirmed by biopsy, but the primary site cannot be found. The cancer cells from the primary site are brought into other organs by invading the lymph, blood, or other means (Pavlidis and Pentheroudakis, 2012). The cause of the tumor is that the focus is small, the position is hidden, or the site of the disease is in the lower part of the mucous membrane and the like, the focus is not easy to find, and the biological behavior of the tumor is worse, leading to the early metastasis of the tumor (Smith et al., 1967).

It is particularly important to find the primary focus in the clinical stage of cancer treatment. Only by finding the primary focus can the clinical cure rate of the patient be improved. Because the biological features often vary with the type of tumor tissue, we can make a pathological diagnosis based on the existing biological knowledge and established pathological methods. Due to the limited tissue and diagnostic staining of tumors and the influence of doctors' professional level,

there are still some loopholes and shortcomings in the thorough search at this stage (Medeiros et al., 2010; Eti et al., 2012; Angela et al., 2017).

The transfer of cancer means that the tumor cells are taken to it from the primary site into the lymphatic vessel, the blood vessel, or other means to continue to grow to form the same type of tumor as the primary site. Common methods of transfer include lymphatic metastasis, vascular metastasis, and the like. About 50% of the lung cancer will have multiple bone metastasis sites, 28–33% of the liver metastasis, and 17–20% of the transfer of the kidney and the epinephrine. The auxiliary imaging examination is usually diagnosed by a biochemical indicator. In the liver metastases, the biochemical biopsy of the liver micro metastases may cause confusion due to the stability of the biochemical indicators; and in the imaging ultrasound examination, the lesions of 1–2 cm could be detected in random tests. The error of uncertain factors in a practical application will accumulate and magnify, resulting in diagnostic confusion.

We aim to establish an automatic processing method to solve this problem. We selected data from gene expression profiles. By analyzing and processing the existing data, a relatively suitable machine learning model is obtained (Fei et al., 2020), and the efficiency of diagnosis of primary lesions can be improved to be more accurate. Different tumorous types have distinct expression profiles on specific genes, and the difference could be captured by the machine learning models and used to classify the primary lesions.

In essence, machine learning trains computers to simulate or realize human learning behavior to acquire new knowledge and skills and reorganize the existing knowledge structure to improve its own performance continuously. The application of medical treatment is also a process of comprehensive doctor diagnosis experience to treat patients. Many machine learning algorithms have been developed for classification problems. It can judge the unknown information by learning from the known information. By studying the existing tumor samples' features, the computer has a certain decision-making ability to judge and evaluate the unknown cancer pathology directly.

XGBoost based on tree boosting is a scalable end-to-end tree boosting system, which was first proposed by Chen and Guestrin (2016). This system is an open-source system available at https://github.com/dmlc/xgboost and is widely used in bioinformatics. Mendik et al. (2018) use XGBoost for analyzing protein translocation between cellular organelles; Li et al. (2019) use XGBoost for predicting gene expression values; Danciu et al. (2020) use XGBoost for predicting early-stage prostate cancer in veterans. We describe the algorithm mechanism in detail in the methods section.

MATERIALS AND METHODS

Data Preparation

Training Set and Oversampling

Data of 5,759 samples, each containing 20,501 gene characteristics, were downloaded from TCGA. After extracting effective information, we normalized the gene expression by the sum of all the sample gene expressions. We use oversampling with stable results to solve the problem of data imbalance, then we select and train the optimal model 10-fold cross-validation on TCGA data.

Test Set

We conduct retrospective testing on a GEO test set containing 42 samples covering five cancers. The trained model predicts the test data, and the results were compared with the true labels of the samples. The specific number of samples per cancer is shown in **Table 1**.

Feature Selection Method

In the training set and the independent verification set, a part of the gene expression level was very low. We set the expression level threshold value as 0.00005, 0.00001, and 0.000001, respectively, for screening. After the intersection of the training set's gene characteristics and the independent verification set, the following feature selection was conducted.

We choose the Chi-Square test and Random Forest in the filtering method for feature selection. The Chi-Square calculates the correlation of qualitative independent variables to qualitative dependent variables. First, we take each gene as an independent hypothesis and then calculate the degree of deviation D between

TABLE 1 | Data size and proportion.

Training data from TCGA		
Cancer type	**Amount**	**Percent**
BRCA	1,056	0.13687622
KIRC	526	0.06817887
UCEC	516	0.0668827
THCA	500	0.06480881
LUAD	486	0.06299417
HNSC	480	0.06221646
COAD	451	0.05845755
LGG	439	0.05690214
STAD	415	0.05379132
PRAD	379	0.04912508
BLCA	301	0.03901491
LIHC	294	0.03810758
OV	261	0.0338302
CESC	258	0.03344135
KIRP	222	0.02877511
LAML	173	0.02242385
GBM	153	0.0198315
READ	153	0.0198315
PAAD	142	0.0184057
SKCM	80	0.01036941
Unknown	430	0.05573558
Testing data from GEO		
BRCA	13	0.27659574
COADREAD	2	0.04255319
LIHC	5	0.10638298
LUAD	15	0.31914894
OV	12	0.25531915

TABLE 2 | Parameters of model evaluation and parameters in the results.

R^2 score	1-MSE(ŷ,y)/Var(y)	
Precision	TP /(TP+FP)	
Recall rate	TP /(TP+FN)	
F1score	·(Precision· Recall)/(Precision+Recall)	
	Relevant	**No relevant**
Retrieved	True positives (TP)	False positives (FP)
Not retrieved	False negatives (FN)	True negatives (TN)

Precision TP / (TP + FP)
Recall rate
TP / (TP + FN) F-Score=(1+β ^{2})· (Precision· Recall)/(β ^{2} · Precision+Recall)

the observed value and the theoretical value. If the deviation is small enough, accept the null hypothesis; otherwise, reject the null hypothesis, and accept the alternative hypothesis. Therefore, the larger the deviation value D, the greater the deviation from the original hypothesis. That is, the more relevant it is, the better the selection process becomes at calculating the deviation value D of each gene and the type of cancer, and to order them from large to small, and to take the first *k* genes.

The application of random forest in feature selection needs to calculate the feature importance. The specific steps are as follows: First, we calculate each feature's importance and sort it in descending order. After that, we determine the proportion to be eliminated and get a new feature set by eliminating the corresponding proportion of features according to their importance. Repeat the process with the new feature set until there are m features left, which is the preset value. Finally, we select the feature set with the lowest out-of-bag error rate according to each feature set obtained in the above process and the corresponding out-of-bag error rate of the feature set.

Training Method

XGBoost is based on gradient tree boosting. Unlike traditional trees, which only do the first-order Taylor expansion, XGBoost performs the second-order Taylor expansion, which realizes the parallel computation (Li et al., 2019). It can use the combination of weak learners to create a single strong learner to reach a fast execution speed and a good model performance. Its main idea is to continuously add a tree and continuously perform feature splitting to grow a tree. Each time a tree is added, it is learning a new function to fit the last prediction residuals. If we get *k*-trees after training, we need to predict the score of a sample. In fact, according to the characteristics of this sample, each tree will fall

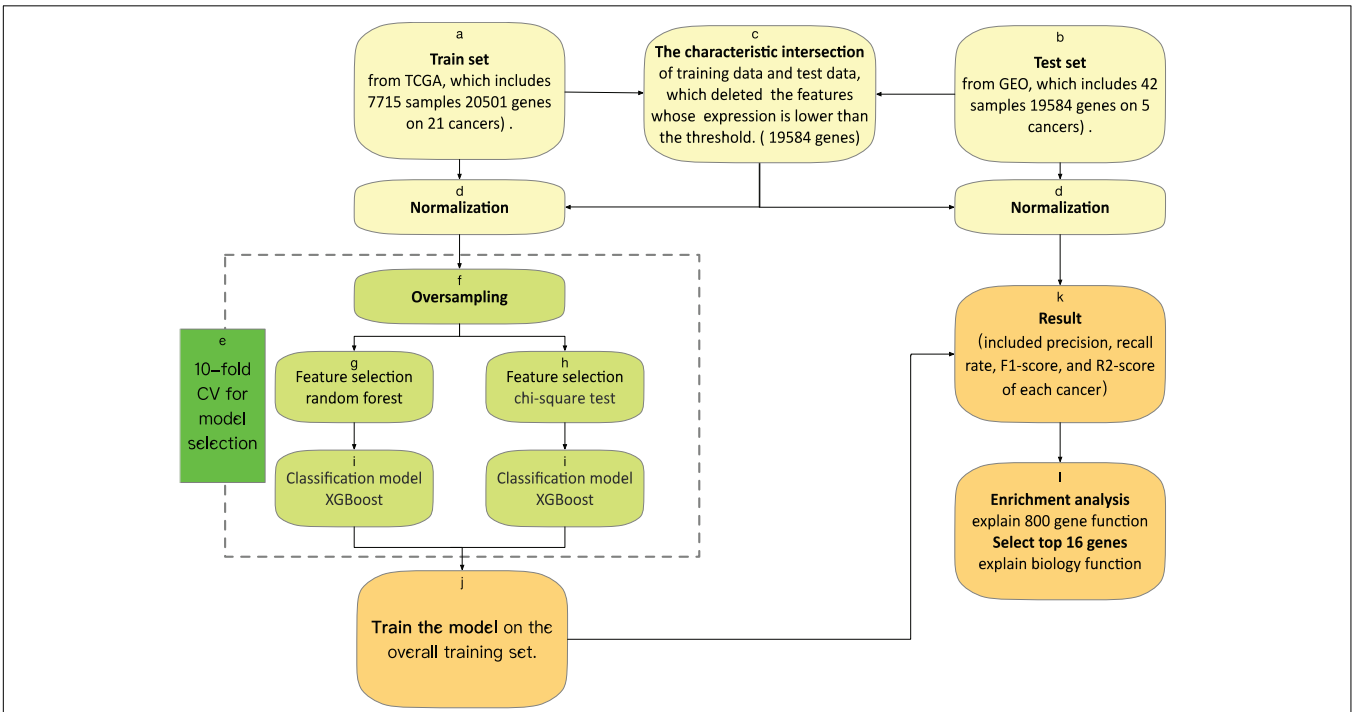

FIGURE 1 | (a) The training set was downloaded from TCGA, which obtains 5,759 samples, each containing 20,501 gene characteristics on 21 cancers. (b) The test set was downloaded from GEO, which obtains 42 samples, each containing 19,584 gene characteristics on five cancers. (c) The characteristic intersection of training data and test data deletes the features whose expression is lower than the threshold. (d) We normalize the gene expression by the sum of all the gene expressions in each sample. (e) Model selection by the result of 10-fold cross-validation. (f) We oversample the train set, and then we have data for 854 samples of each tumor. (j) We train the model on the overall training set, with optimal features and an optimal model. (k) We test the model and output the result included precision, recall rate, F1-score, and R^2-score of each cancer. (l) We do an enrichment analysis of the top 800 genes with KEGG (Kyoto Encyclopedia of Genes and Genomes) and GO (Gene Ontology) by metascape, and explain the top 16 genes from feature selection, which include biological function.

TABLE 3 | 10-fold CV results of variety with the number of features in Chi-Square and Random Forest.

Feature number	10-fold CV result of using the Chi-Square in feature selection	10-fold CV result of using Random Forest in feature selection
100	0.929750576	0.936357298
200	0.947377573	0.951911924
300	0.957487752	0.956577824
400	0.956709878	0.961505816
500	0.960339005	0.960726262
600	0.961894081	0.960854956
700	0.961894081	0.962541414
800	0.961890889	**0.963838431**
900	0.962538726	0.963707385
1,000	**0.962278986**	0.963448150

The bold values in each column are the optimal results for this method.

to a corresponding leaf node, and each leaf node corresponds to a score. It is necessary to add up the scores corresponding to each tree to be the predicted value of the sample. Chen and Guestrin (2016) descript the mathematical formula of gradient tree boost and XGBoost with scientific rigor. And Li et al. (2019) described the parameters of XGBoost.

We fine-tuned three hyperparameters within the 10-fold cross-validation. The parameter "n estimators" is the number of trees to be used in the forest. The parameter "max depth" is the deepest depth of all trees. The parameter "min child weight parameter" in XGBoost is the minimum sum of instance weight (hessian) needed in a child. If the tree partition step results in a leaf node with the sum of instances weighing less than the min child weight, the building process will give up further partitioning. This parameter is used to avoid overfitting. When its value is large, the model can be prevented from learning from outliers. But if this value is too high, it will cause under-fitting. The max depth is also used to avoid overfitting. The greater the max depth, the more outliers the model will learn.

Parameters of Model Evaluation and Parameters in the Results

Use the R^2 score as an indicator of the evaluation model. At the same time, the test results are output, which included the R^2 score, precision, recall rate, and the F1 score of each cancer calculation result shown in **Table 2**.

The predicted value is $\hat{y}$ and the true value is y. R^2 score the problem that MSE (Mean Absolute Error), RMSE (Root Mean Squared Error), and MAE (Mean Absolute Error) cannot solve when dimensions are different, and it is difficult to measure the effectiveness of the model. R^2 score = 1, reaches the maximum value, and then MSE as the molecule is 0, which means that the predicted value and the true value in the sample are the same, without any error. In other words, the model that has been established perfectly fits all the real data, which is the model with the best effect and where the R^2 score value reaches the maximum. The model is usually not so perfect; there are

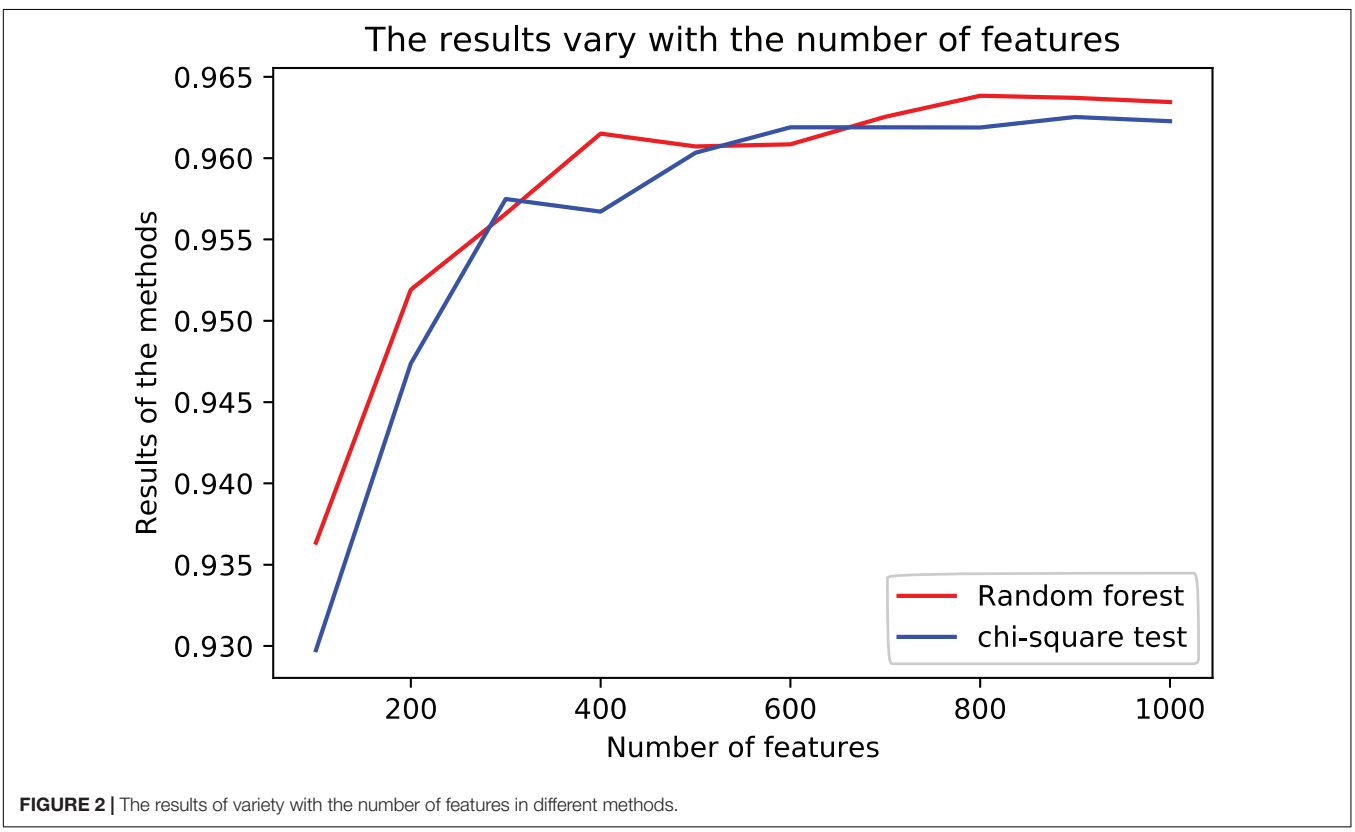

FIGURE 2 | The results of variety with the number of features in different methods.

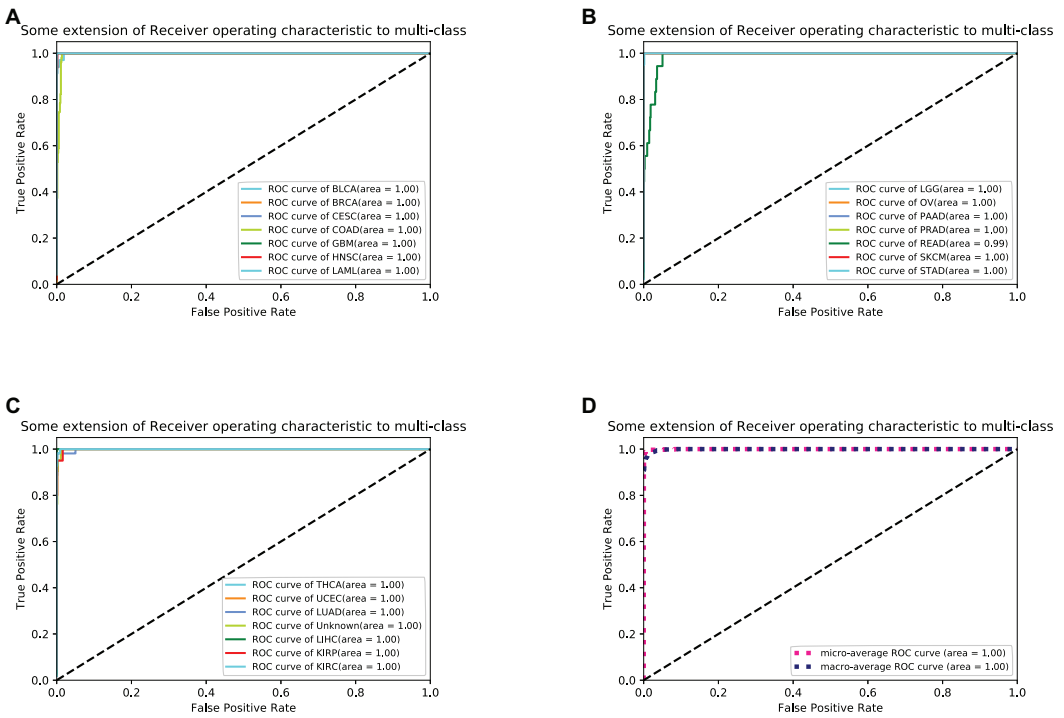

FIGURE 3 | (A–C) shows 21 cancers' ROC curve of the optimal 10-fold CV's results. **(D)** shows the average ROC curve.

always errors; when the error is small, the numerator is less than the denominator; when the model tends to 1, it is still a good model. Precision is defined as (true-positives)/(true positives + false- positives). Recall rate is defined as (true-positives)/(true-positives + false-negatives), which intuitively represents the classifier's ability to identify all positive cases correctly. F1 score is the harmonic mean of precision and recall. Precision and Recall do not have much of a relationship with the formula, but they are mutually restricted in practice. We all hope that the model is accurate, and the recall rate is high, but when the precision rate is high, the recall rate is often low. When $\beta = 1$, it becomes the F1-score, in which case both recall, and accuracy are important and have the same weight. In some cases, if we think accuracy is more important, we adjust the β value to be less than 1, and if we think the recall is more important, we adjust the β value to be greater than 1, such as the F2-score.

We determined the data list as the first 800 genes from the feature selection list. We used software: Cytoscape and metascape for GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Gene and Genomes) Enrichment Analysis.

RESULTS

Genes Selected by Random Forest Were More Informative Than Chi-Square

We used 10-fold cross-validation in the training set to evaluate the performance of the feature selection methods.

TABLE 4 | The model test result (precision, recall, F1-score, and R^2-score) on 9 cancers on the GEO dataset.

Abbreviation	Precision	Recall	F1-score	R^2-score	Support
BRCA	1	0.75	0.86	0.75	12
COADREAD	1	1	1	1	1
LIHC	1	1	1	1	5
LUAD	0.85	0.92	0.88	0.92	12
OV	1	0.82	0.9	0.82	11
Avg/total	0.93	0.83	0.87	0.83	42

With leave-one-out cross-validation, the algorithm is repeatedly retrained, which included oversampling, feature selection, and classification model, leaving out one sample in each round and testing each sample on a classifier that was trained without this sample. The framework of the 10-fold CV is shown in **Figure 1**.

The results are shown in **Table 3**. The average R^2-score of 10-fold cross-validation of the two feature selection methods is very high. The average R^2-score was 96.23 and 96.38% (95% confidence interval) for the chi-square test as feature selection and random forest as feature selection. Although these two results are very close, the R^2-score of Random Forest is slightly higher than the Chi-Square within the same feature number range, and the Rise of R^2-score of random forest is more stable, as shown in **Figure 2**. Considering all the results of the average R^2-score, the Random Forest is used for feature selection in the next flow.

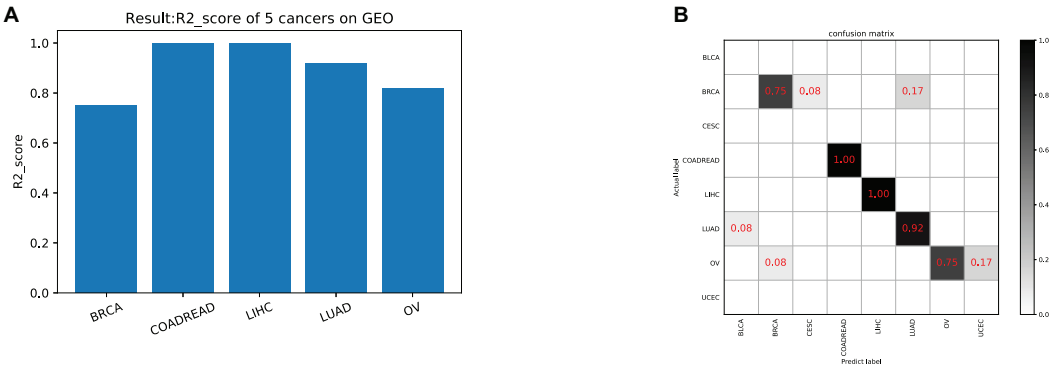

FIGURE 4 | (A) The model test result (R^2-score) on five cancers on GEO. **(B)** The confusion matrix on testing data. Our trainer contained 21 cancer tags, but only five cancers in the test set. There was a partial error in the classifier's prediction outside of the five cancers.

TABLE 5 | The basic information of top 16 genes on feature selection.

Mark rank	Gene symbol	Gene name	RefSeq DNA sequence	UniProtKB/Swiss-Prot
1	AFAP1L2	Actin filament associated protein 1 like 2	NC_000010.11	Q8N4 × 5-AF1L2_HUMAN
2	CREB3L4	CAMP responsive element binding protein 3 like 4	NC_000001.11	Q8TEY5-CR3L4_HUMAN
3	HOXB13	Homeobox B13	NC_000017.11	Q92826-HXB13_HUMAN
4	KLK3	Kallikrein related peptidase 3	NC_000019.10	P07288-KLK3_HUMAN
5	PLCB2	Phospholipase C beta 2	NC_000015.10	Q00722-PLCB2_HUMAN
6	RC3H1	Ring finger and CCCH-type domains 1	NC_000001.11	Q5TC82-RC3H1_HUMAN
7	TMEM176A	Transmembrane protein 176A	NC_000007.14	Q96HP8-T176A_HUMAN
8	TMPRSS2	Transmembrane serine protease 2	NC_000021.9	O15393-TMPS2_HUMAN
9	WT1	WT1 transcription factor	NC_000011.10	P19544-WT1_HUMAN
10	CCL16	C-C motif chemokine ligand 16	NC_000017.11 NT_187614.1	O15467-CCL16_HUMAN
11	CDH17	Cadherin 17	NC_000008.11	Q12864-CAD17_HUMAN
12	H3F3C	Histone variant H3.5	NC_000012.12	Q6NXT2-H3C_HUMAN
13	HNF1A	HNF1 homeobox A	NC_000012.12	P20823-HNF1A_HUMAN
14	KLK2	Kallikrein related peptidase 2	NC_000019.10	P20151-KLK2_HUMAN
15	SLC45A3	Solute carrier family 45 member 3	NC_000001.11	Q96JT2-S45A3_HUMAN
16	STEAP2	STEAP2 metalloreductase	NC_000007.14	Q8NFT2-STEA2_HUMAN

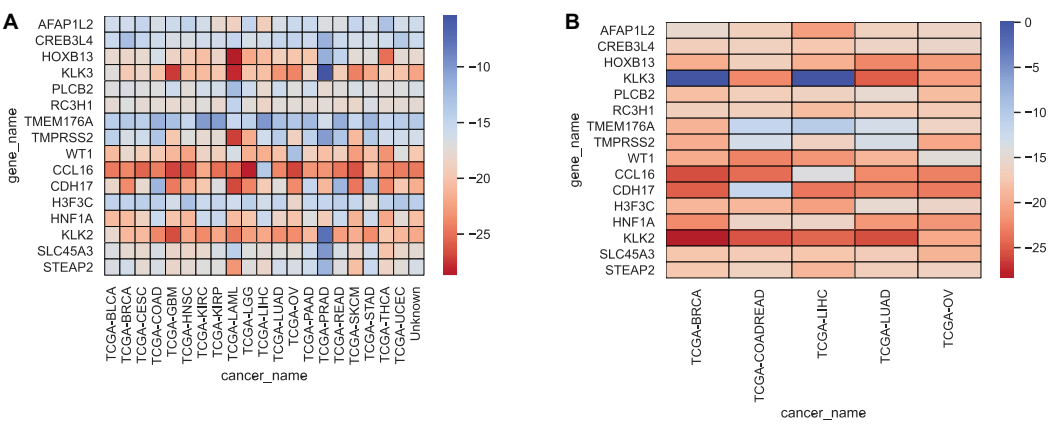

FIGURE 5 | Heatmap representing the expressions of 16 genes for each cancer sample in the training set and test set, averaged, and then logarithmic. Cool colors represent a higher expression level, and warm colors a lower expression level. **(A,B)** Represent the expression levels of 16 genes in the training set and the test set respectively.

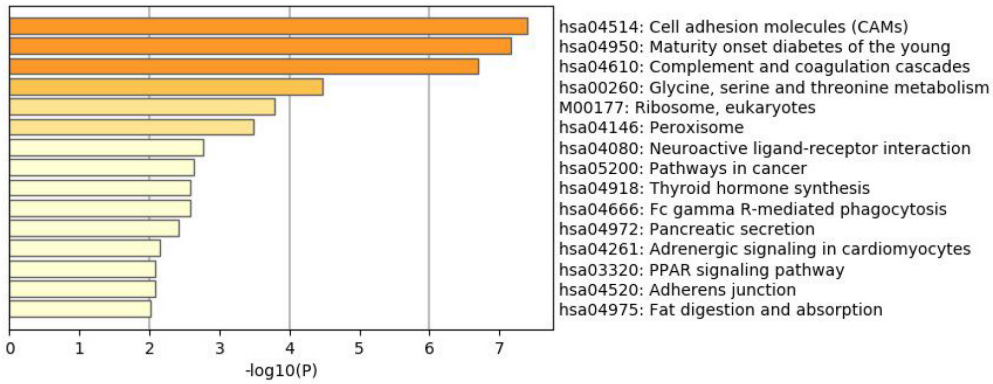

FIGURE 6 | KEGG enrichment analysis of the 800 selected genes.

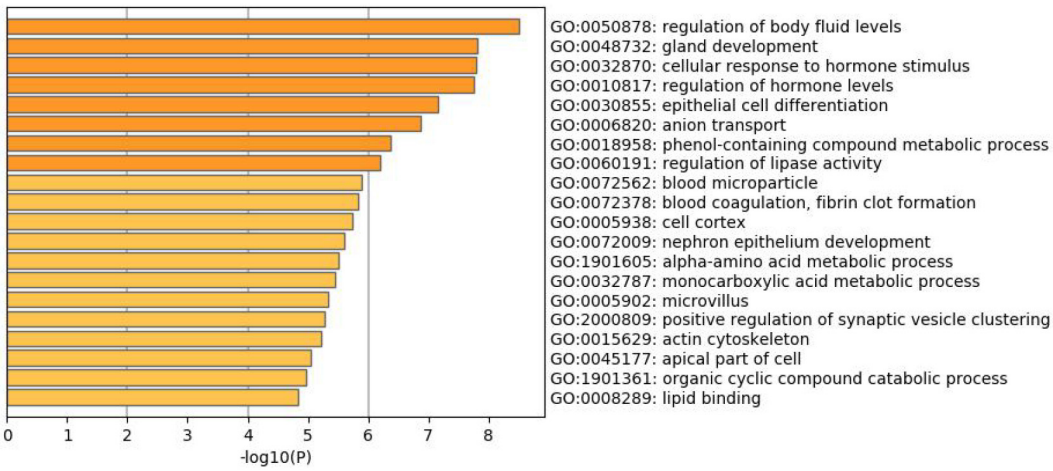

FIGURE 7 | GO enrichment analysis of the 800 selected genes.

The XGBoost Algorithm Showed Good Generalization Performance on the GEO Dataset

We selected 800 genes with Random Forest characteristics, using XGBoost as a classifier. Taking the R^2-score as the model evaluation index, 10-fold CV was carried out in the training data, and finally, the parameters, n estimators = 250, max depth = 7, min child weight = 1, in the optimal model of XGBoost were obtained. The results of this model in leaving out one data are shown in **Figure 2**.

For each sample, the type of tumor predicted was compared with the type diagnosed. When the predicted tumor type matches the reference diagnosis, it is a true positive. When the predicted tumor type does not match the diagnosis, the sample is considered a false-positive. For each cancer, sensitivity was defined as the ratio of true positive results to the total positive samples analyzed, and specificity was defined as the ratio of (1- false positive) to (total test results - total positive). To better measure the classification results, we took sensitivity and specificity as the horizontal axis and the vertical axis, respectively, and drew the ROC (Receiver Operating Characteristic) curve to the results as shown in **Figure 3**.

The model was trained according to N estimators = 250, Max depth = 7, and min child weight = 1 in the whole training data for independent testing. The R^2-score average of independent testing results is 83.3%, which obtained 42 samples cover five cancers. The trainer had good generalization for COADREAD (Colon Adenocarcinoma and Rectum Adenocarcinoma), LIHC (Liver Hepatocellular Carcinoma), LUAD (Lung Adenocarcinoma), and OV (Ovarian Serous Cystadenocarcinoma), and the R^2-score respectively was 1, 1, 0.92 and 0.82, shown in **Table 4** and **Figure 4A**. For BRCA (Breast Invasive Carcinoma), we can see from **Figure 4B** that it is often incorrectly predicted for CESC (Csquamous Cell Carcinoma and Endocervical Adenocarcinoma) and LUAD.

Top 16 Genes on Feature Selection

We often use molecular experiments to distinguish the origin of metastatic cancer. Our supporting results combined with the literature review found that the accuracy of cancer classification

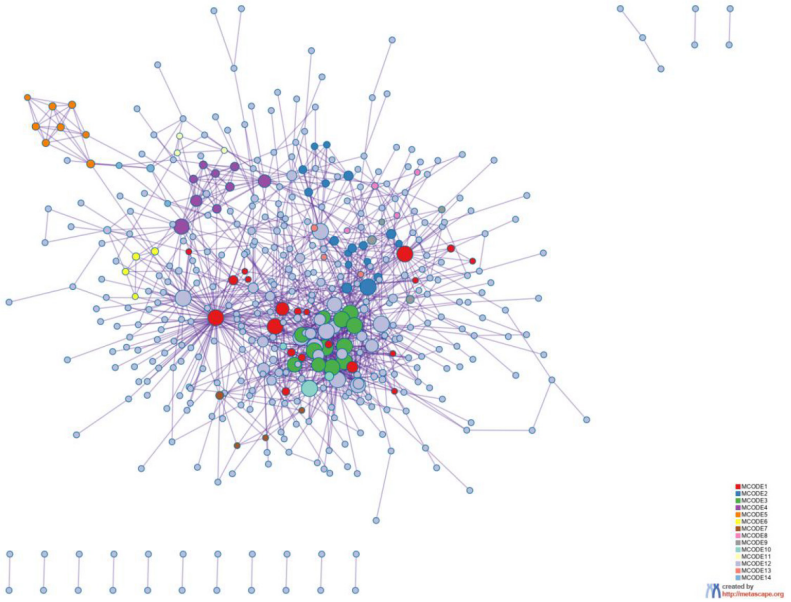

FIGURE 8 | Protein-protein interaction network. The MCODE algorithm was then applied to this network to identify neighborhoods where proteins are densely connected. Each MCODE network is assigned a unique color. The GO enrichment analysis was applied to each MCODE network to assign "meanings" to the network component.

was low for fixed cancer types, which was similar to other data methods. We selected 16 genes, shown in **Table 5**, with high expression levels, to analyze the potential relationship between these genes and cancer. The heat maps of the expressions of 16 genes in the training set and the test set are shown in **Figure 5**.

Genes control protein expression. A gene contains introns and exons, in which the coding region of the protein is encoded. Gene coding of a protein is a DNA-mRNA- protein process. The genes we analyzed are all protein-coding genes.

WT1 is a tumor suppressor gene associated with the development of a Wilms' Tumor, for which it was named. This gene encodes a transcription factor that contains four zinc-finger motifs at the C-terminus and a proline/glutamine-rich DNA-binding domain at the N-terminus. CCL16 is one of several cytokine genes clustered on the q-arm of chromosome 17. Cytokines are a family of secreted proteins involved in immunoregulatory and inflammatory processes. The CC cytokines are proteins characterized by two adjacent cysteines. The cytokine encoded by this gene displays chemotactic activity for lymphocytes and monocytes but not for neutrophils. This cytokine also shows a potent myelosuppressive activity and suppresses the proliferation of myeloid progenitor cells. The expression of this gene is upregulated by IL-10. The CDH17 gene is a member of the cadherin superfamily, genes encoding calcium-dependent, membrane-associated glycoproteins. Diseases associated with CDH17 include Metanephric Adenoma and Cleft Lip/Palate-Ectodermal Dysplasia Syndrome, which is provided by RefSeq et al. Histones are basic nuclear proteins that are responsible for the nucleosome structure of the chromosomal fiber in eukaryotes. Nucleosomes consist of approximately 146 bp of DNA wrapped around a histone octamer composed of pairs of each of the four core histones (H2A, H2B, H3, and H4). Among its related pathways are Transcriptional misregulation in cancer and Activated PKN1, which stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3. HNF1A encodes a transcription factor required for the expression of several liver-specific genes. Diseases associated with HNF1A include Maturity-Onset Diabetes of the Young, Type 3 and Diabetes Mellitus, and Insulin-Dependent 20.

Enrichment Analysis

To better understand why those genes could tell the origin of the primary lesion, we performed the enrichment analysis using the 800 selected genes. The results of KEGG (Kyoto Encyclopedia of Gene and Genomes) (**Figure 6**) and GO (Gene Ontology) (**Figure 7**) are shown in **Figures 8**, **9**.

The 800 selected genes were significantly enriched in some cancer-related pathways. Cell adhesion molecules (CAM) (Okegawa et al., 2004) played important roles in invasive and metastasis and cancer progression. Loss of the tumor cells' intercellular adhesion might result in cells escaping from the primary lesion and metastasizing. CAM is also involved in various functions such as cell growth, differentiation, site-specific gene expression, and morphogenesis, which could explain why the different tissues have different expression profiles among those genes.

The 800 genes were also significantly enriched in some organ-specific pathways. The selected genes were representative in thyroid hormone synthesis, pancreatic secretion, and fat digestion—absorption pathways. Since those pathways were organ-specific, we could show that the random forest algorithm found the differentially expressed genes among different organs.

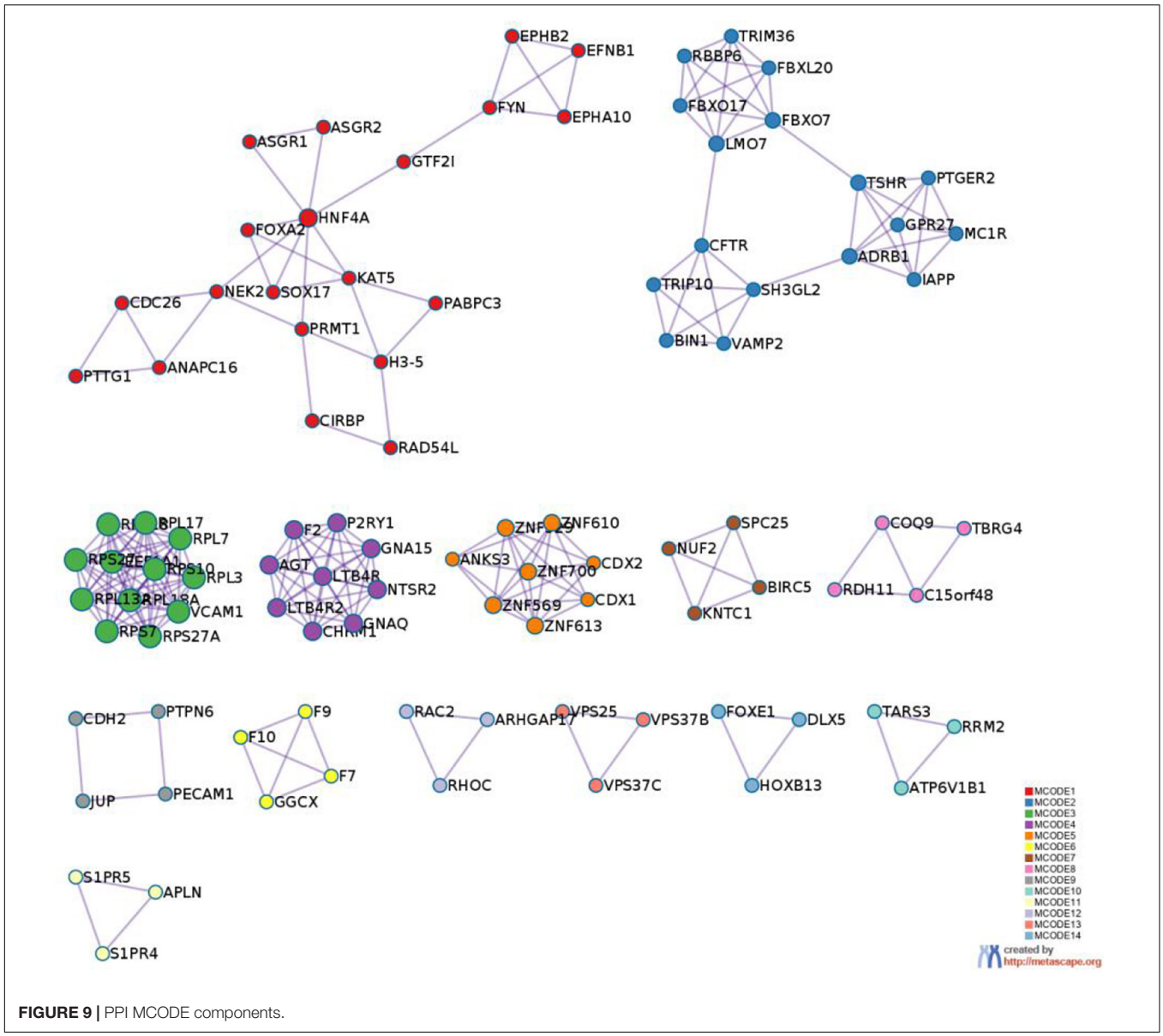

FIGURE 9 | PPI MCODE components.

DISCUSSION

Nowadays, CUP cases are characterized by small primary tumors (difficult to be detected by existing technologies) (Hainsworth and Greco, 2018), primary tumors being eliminated by the body's autoimmune system, and primary tumors being excised during surgery (without histological examination), which makes it difficult to find the primary tumors, leading to generally poor prognosis of patients treated with chemotherapy. Our study hopes to help doctors clinically identify the primary of CUP and to use more effective targeted therapies for CUP patients according to these identification results.

In this paper, we show that our result is better than in recent studies. Our average R^2-score of the classification based on XGBoost can reach 96.38%, while the average accuracy of the support vector machine (SVM) classifier is 82–89% (Tothill et al., 2005; Ma et al., 2006). We train a classifier, selected feature by random forests, classified by XGBoost, on data containing 7,715 samples and 19,854 genes from TCGA, and test it on data including 42 samples and five cancers. Currently, the prediction for CUP cancer is between 80%–95% (Sarah, 2010; Greco et al., 2012; Meiri et al., 2012; Conway et al., 2019), and this data fluctuation is related to the different evaluation indicators and sample types of each model. In the test R^2-score of 83.3% in particular, our classifier was relatively accurate in predicting LIHC (liver hepatocellular carcinoma) which is, LUAD (lung adenocarcinoma), OV (ovarian serous cystadenocarcinoma).

Although we have made progress in these studies, there are also limitations. Our test data are collected from 8 series, and there was some detection method between each series. This may be due to the fact that our test results are not as high as the cross-validation results.

Further studies could be done in several main aspects. First, the SNP (single nucleotide polymorphism) or methylation data may be combined with expression profiles to further improve the prediction utilities to infer primary lesions for metastatic tumors. Second, the eQTL (expression Quantitative Trait Loci), which supplies us with new insights between expression profile and mutation profile, might also help determine the primary lesions.

CONCLUSION

These findings suggest that by combining multiple tumor data with machine learning methods, each cancer has its corresponding classification accuracy, which can be used to predict primary metastatic tumors' location. At the same time, it can also be used as an orthogonal diagnostic method to utilize the machine learning model processing for auxiliary diagnosis methods.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

KL and ZH designed the study. SC, WZ, JT, and BW collected the data, analyzed the data, interpreted the data. SC wrote the manuscript. JL, XM, and GT reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Natural Science Foundation of Fujian Province (No. 2020J011112).

REFERENCES

Angela, H., Wordsworth, S., Fermont, J. M., Page, S., Kaur, K., Camps, C., et al. (2017). Clinical applicability and cost of a 46-gene panel for genomic analysis of solid tumours: retrospective validation and prospective audit in the UK national health service. *J. PLoS Med.* 14:e1002230. doi: 10.1371/journal.pmed.1002230

Chen, T., and Guestrin, C. (2016). "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: Association for Computing Machinery, 785–794.

Conway, A. M., Mitchell, C., Kilgour, E., Brady, G., Dive, C., and Cook, N. (2019). Molecular characterisation and liquid biomarkers in carcinoma of unknown primary (CUP): taking the 'U' out of 'CUP'. *Br. J. Cancer* 120, 141–153. doi: 10.1038/s41416-018-0332-2

Danciu, I., Erwin, S., Agasthya, G., Janet, T., McMahon, B., Tourassi, G., et al. (2020). Using longitudinal PSA values and machine learning for predicting progression of early stage prostate cancer in veterans. *J. Clin. Oncol.* 38:e17554. doi: 10.1200/jco.2020.38.15_suppl.e17554

Eti, M., Mueller, W. C., Rosenwald, S., Zepeniuk, M., Klinke, E., Edmonston, T. B., et al. (2012). A second-generation microRNA-based assay for diagnosing tumor tissue origin. *J. Oncologist* 17, 801–812. doi: 10.1634/theoncologist.2011-0466

Fei, Y., Lin, L., and Quan, Z. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *J. Biochim. Biophys. Acta. Mol. Basis Dis.* 1866:165822. doi: 10.1016/j.bbadis.2020.165822

Greco, F. A., Lennington, W. J., Spigel, D. R., Varadhachary, G. R., and Hainsworth, J. D. (2012). Carcinoma of unknown primary site: outcomes in patients with a colorectal molecular profile treated with site specific chemotherapy. *J. Cancer Therapy* 3, 37–43. doi: 10.4236/jct.2012.31005

Hainsworth, J. D., and Greco, F. A. (2018). Cancer of unknown primary site: new treatment paradigms in the era of precision medicine. *Am. Soc. Clin. Oncol. Educ. Book* 38, 20–25. doi: 10.1200/edbk_100014

Li, W., Yin, Y., Quan, X., and Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Front. Genet.* 10:1077.

Ma, X. J., Patel, R., Wang, X., Salunga, R., Murage, J., and Desai, R. (2006). Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. *Arch. Pathol. Lab Med.* 130, 465–473.

Medeiros, F., Lyons-Weiler, M., and Henner, W. D. (2010). Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. *J. Diagn. Pathol.* 5:3.

Meiri, E., Mueller, W. C., Rosenwald, S., Zepeniuk, M., Klinke, E., and Edmonston, T. B. (2012). A second-generation microRNA-based assay for diagnosing tumor tissue origin. *Oncologist* 17, 801–812. doi: 10.1634/theoncologist.2011-0466

Mendik, P., Dobronyi, L., Hári, F., Kerepesi, C., Maia-Moço, L., Buszlai, D., et al. (2018). Translocatome: a novel resource for the analysis of protein translocation between cellular organelles. *Nucleic Acids Res.* 47, D495–D505.

Okegawa, T., Pong, R. C., Li, Y., and Hsieh, J. T. (2004). The role of cell adhesion molecule in cancer progression and its application in cancer therapy. *Acta Biochim. Pol.* 51, 445–457. doi: 10.18388/abp.2004_3583

Pavlidis, N., and Pentheroudakis, G. (2012). Cancer of unknown primary site. *Lancet* 379, 1428–1435. doi: 10.1016/S0140-6736(11)61178-1

Sarah, E. K. (2010). Multisite validation study to determine performance characteristics of a 92-gene molecular cancer classifier. *J. Clin. Cancer Res.* 18, 3952–3960. doi: 10.1158/1078-0432.ccr-12-0920

Smith, P. E., Krementz, E. T., and William, C. (1967). Metastatic cancer without a detectable primary site. *J. Elsevier* 113, 633–637. doi: 10.1016/0002-9610(67)90309-1

Tothill, R. W., Kowalczyk, A., Rischin, D., Bousioutas, A., Haviv, I., and van Laar, R. K. (2005). An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res.* 65, 4031–4040. doi: 10.1158/0008-5472.can-04-3617

Conflict of Interest: BW, XM, 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: 11 February 2021
doi: 10.3389/fgene.2021.639930

MSU-Net: Multi-Scale U-Net for 2D Medical Image Segmentation

Run Su[1,2], *Deyun Zhang*[3], *Jinhuai Liu*[1,2*] *and Chuandong Cheng*[4,5,6]

[1] *Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China,* [2] *Science Island Branch of Graduate School, University of Science and Technology of China, Hefei, China,* [3] *School of Engineering, Anhui Agricultural University, Hefei, China,* [4] *Department of Neurosurgery, The First Affiliated Hospital of University of Science and Technology of China (USTC), Hefei, China,* [5] *Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China,* [6] *Anhui Province Key Laboratory of Brain Function and Brain Disease, Hefei, China*

Aiming at the limitation of the convolution kernel with a fixed receptive field and unknown prior to optimal network width in U-Net, multi-scale U-Net (MSU-Net) is proposed by us for medical image segmentation. First, multiple convolution sequence is used to extract more semantic features from the images. Second, the convolution kernel with different receptive fields is used to make features more diverse. The problem of unknown network width is alleviated by efficient integration of convolution kernel with different receptive fields. In addition, the multi-scale block is extended to other variants of the original U-Net to verify its universality. Five different medical image segmentation datasets are used to evaluate MSU-Net. A variety of imaging modalities are included in these datasets, such as electron microscopy, dermoscope, ultrasound, etc. Intersection over Union (IoU) of MSU-Net on each dataset are 0.771, 0.867, 0.708, 0.900, and 0.702, respectively. Experimental results show that MSU-Net achieves the best performance on different datasets. Our implementation is available at https://github.com/CN-zdy/MSU_Net.

Keywords: multi-scale block, U-net, medical image segmentation, convolution kernel, receptive field

OPEN ACCESS

Edited by:
Jialiang Yang,
Geneis Co. Ltd, China

Reviewed by:
Khanh N. Q. Le,
Taipei Medical University, Taiwan
Bing Wang,
Anhui University of Technology, China

****Correspondence:***
Jinhuai Liu
jhliu@iim.ac.cn

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

Received: *10 December 2020*
Accepted: *20 January 2021*
Published: *11 February 2021*

Citation:
Su R, Zhang D, Liu J and Cheng C (2021) MSU-Net: Multi-Scale U-Net for 2D Medical Image Segmentation. Front. Genet. 12:639930. doi: 10.3389/fgene.2021.639930

1. INTRODUCTION

Medical imaging analysis has made a significant breakthrough with the rapid progress of deep learning (Long et al., 2015; Chen et al., 2018a; Salehi et al., 2018; Wang et al., 2019b). Among these techniques, encoder-decoder architecture has been widely used in the medical image segmentation task (Salehi et al., 2017; Xiao et al., 2018; Guan et al., 2019). U-Net (Ronneberger et al., 2015) is the most classic encoder-decoder structure for medical image segmentation. In recent years, the original U-Net has been modified by many researchers. As a result, many variants of the original U-Net have been proposed (Poudel et al., 2016; Oktay et al., 2018; Roth et al., 2018).

However, the variants of the original U-Net come with two limitations. First, the diversity of features is lost due to the fixed receptive field of the convolution kernel. The same scale feature maps extracted from the convolution kernel with different receptive fields are semantically different. As a result, the performance of the network may vary with the size of the receptive field, and the performance depends on the size of the receptive field in the convolution kernel. Redundant features will be extracted when the receptive field of the convolution kernel is too small. Smaller targets are ignored when the receptive field of the convolution kernel is too large. For example, in the pulmonary lesion or multi-organ segmentation task, the edge detail of the smaller lesion/organ is not fine by the large receptor field and the structure of the lesion/organ is not obvious by the small receptor field. Therefore, it is very important to use the convolution kernel with different

receptive fields to process the image (Luo et al., 2016; Peng et al., 2017; Shen et al., 2019). In the natural image processing task, satisfactory results are obtained by combining the convolution of different receptive fields (Seif and Androutsos, 2018). To the best of our knowledge, there are few reports based on different receptive fields in medical image segmentation tasks. Second, some information may be lost using a single convolutional sequence to extract features at each scale. More feature information can be obtained by multiple convolutional sequences. The loss of feature information can be reduced by the structure of multiple convolutional sequences in the process of down-sampling and up-sampling. Therefore, the learning capacity of the network is aided by multiple convolutional sequences (He et al., 2015).

In this paper, a new image segmentation architecture (multi-scale U-Net) is proposed by us to overcome the above limitations. This architecture is a generalization segmentation architecture. Multi-scale U-Net (MSU-Net) consists of blocks of multi-scale whose multi-scale blocks are composed of convolution sequences with different receptive fields. The multi-scale block introduced in MSU-Net achieves the following advantages. First, more feature information can be obtained because of the multiple convolutional sequences structure embedded in the network. The input of the convolution sequence is all the same, while their convolution kernel is not shared. This design not only improves the performance of segmentation but also facilitates the learning of network in the training process. Second, the features extracted from the multi-scale block are diversified. This is caused by the multiple convolution sequences with different receptive fields in multi-scale block. This is helpful for intensive forecasting tasks that require detailed spatial information. The semantics extracted from the convolution sequence with different receptive fields are different on the same scale feature map. This structure enables the encoder of the network to extract features better and the decoder to restore features better. We construct different types of multi-scale blocks with several commonly used convolution kernels. An extensive evaluation of different types of multi-scale blocks is performed on three segmentation datasets. Our results demonstrate that MSU-Net built by integrated multiple convolution sequences with different receptive fields enables significant improvement of semantic segmentation. Compared with the traditional U-Net architecture, the main improvement of MSU-Net is the integration of multiple convolution sequences with different sizes of receptive fields. This improvement enables the object features to become more conspicuous with forward propagation. In addition, the proposed multi-scale block can be easily integrated into other network structures.

In summary, the main contributions of this paper are summarized as follows:

(1) Multi-scale blocks are proposed by us based on several commonly used convolution kernel. More diverse feature information and better feature maps are captured from the images through multi-scale block.

(2) MSU-Net, a new segmentation architecture for medical image, is proposed for medical image segmentation. This is an improvement on the basic structure of U-Net. Compared to the existing algorithms, the proposed method has a stronger ability to overcome the problems of class-imbalance and overwhelmed.

(3) Different receptive fields are crucial for dense prediction tasks requiring detailed spatial information. It can stimulates learning capacity of network and make the network more robust. Experimental results demonstrate that the proposed method is outperforms the state-of-the-art methods in medical image segmentation task under different imaging modalities.

2. RELATED WORKS

With the development of convolutional neural network (CNN) in the field of natural image processing and medical image analysis, automatic feature learning algorithm using deep learning has become a feasible method for biomedical image segmentation (Le et al., 2019, 2020; Sua et al., 2020). Segmentation method based on deep learning is a learning method with pixel-classification, which is different from the traditional pixel or superpixel classification method (Abramoff et al., 2007; Kitrungrotsakul et al., 2015; Tian et al., 2015) using hand-made features. The limitations of hand-made features are overcome when deep learning approaches are used to learn features. The limitations of hand-made features are overcome when deep learning approaches are used to learn features. Early deep learning methods for medical image segmentation are mostly based on patch. The strategy based on plaque and sliding window was proposed by Ciresan et al. (2012) to segment neuronal membranes from microscopic images. Kamnitsas et al. (2017) adopted a multi-scale 3D CNN architecture with fully connected conditional random field (CRF) to enhance patch based brain leasion segmentation. Pereira et al. (2016) proposed an automatic segmentation method based on CNN to segment brain tumors. Obviously, two main drawbacks are introduced by this solution: the redundant computation caused by sliding window and the global feature cannot be learned.

With the emerging of end-to-end FCN (Long et al., 2015), Ronneberger et al. (2015) proposed U-Net for biomedical image segmentation. U-Net has shown good performance in fields of medical image segmentation. It has become a popular neural network architecture for biomedical image segmentation tasks (LaLonde and Bagci, 2018; Fan et al., 2019; Song et al., 2019). Li et al. (2019) proposed a new dual-U-Net architecture to solve the problem of nuclei segmentation. Milletari et al. (2016) proposed a 3D image segmentation method based on U-Net to perform end-to-end training on prostate MRI. Guan et al. (2019) proposed an improved CNN structure for removing artifact from 2D PAT images reconstructed. Many variants of U-Net has been appeared for different medical image segmentation tasks. In order to improve the learning ability of feature, some new modules are proposed to replace the original modules. Seo et al. (2019) proposed an up-sampling method based on an object and redesigned the remaining paths and skip-connection. The limitation of the traditional U-Net algorithm was overcome in this way. Ge et al. (2019) proposed a k-shaped network of end-to-end deep neural network. The network was used for multi-view segmentation and multi-dimensional quantification of LV

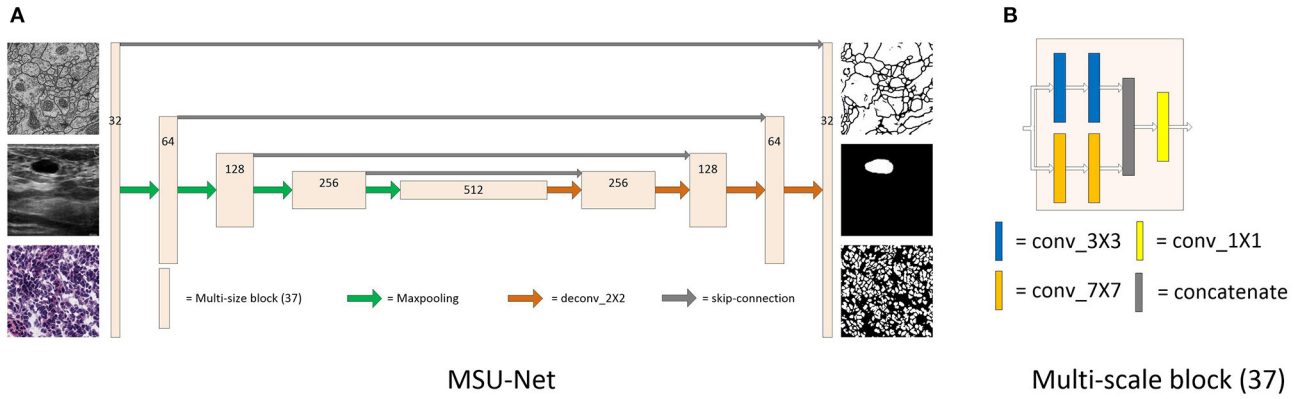

FIGURE 1 | Detailed description of MSU-Net and multi-scale block (37). Panel **(A)** is the architecture of MSU-Net. The overall architecture is similar to the original U-Net. The dimensions of the network are represented by numbers on the block structure. Panel **(B)** is the architecture of multi-scale block (37). This module is embedded in the original U-Net to get MSU-Net.

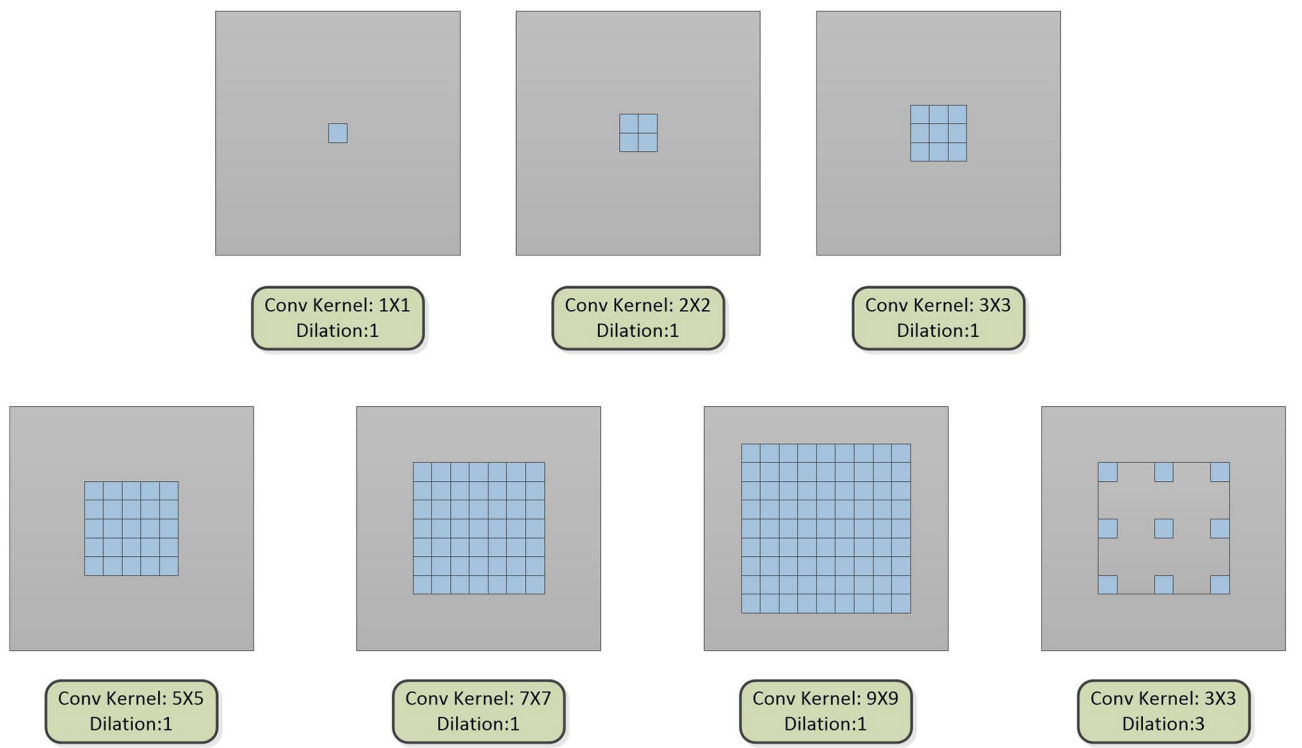

FIGURE 2 | The type of convolution kernel used in this article. Combining the above seven convolution kernels, different types of multi-scale blocks are proposed.

in PEAV sequences. Myronenko (2018) proposed a semantic segmentation method for 3D brain tumor segmentation from multimodal 3D MRIs. An asymmetric encoder was used to extract features, and then two decoders segment the brain tumor and reconstruct the input image, respectively. Oktay et al. (2018) proposed AttU-Net in combination with attention gate. Alom et al. (2018) integrated the structure of Recurrent Neural Network (RNN) and ResNet into the original U-Net. RNN could make the network extract better features. ResNet enables the training of deeper networks. Liu et al. (2020) proposed a ψ-shaped depth neural network (ψ-Net). In the deep stage, semantic information was featured by selective aggregation. In the shallow stage, the semantic information obtained in the deep stage was used to improve the detailed information. Therefore, discriminative features were obtained to provide the basis for accurate subcortical segmentation of brain structures. In addition

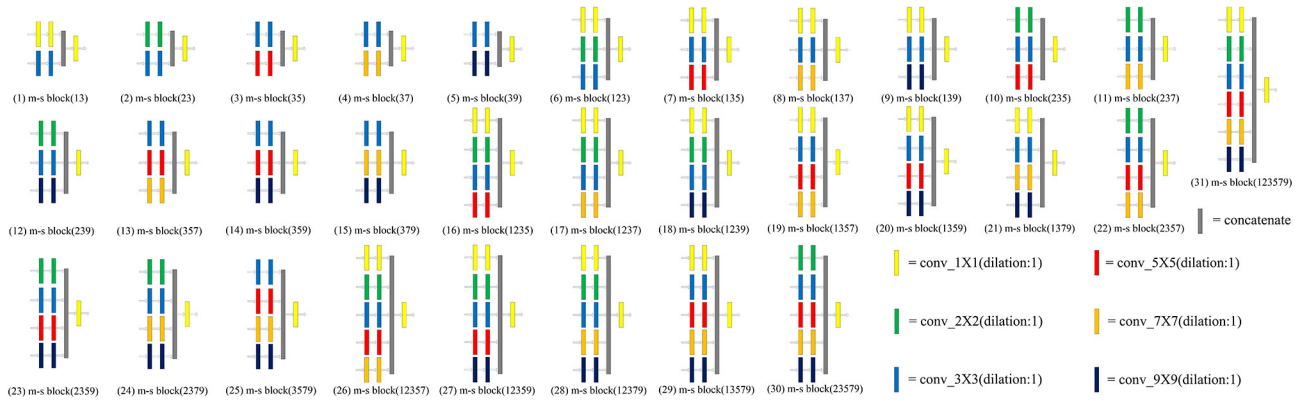

FIGURE 3 | An overview of 31 multi-scale blocks. m-s block represents the multi-scale block. Different multi-scale blocks are designed according to several commonly used convolution kernels. More richer and diverse features can be extracted through this design. Meanwhile, the problem of unknown network width can be alleviated by this design. Conducive to intensive prediction tasks that require detailed spatial information.

to the above achievements in medical image segmentation based on U-Net, some researchers have also improved U-Net to apply in general image segmentation. Zhang et al. (2018) proposed a semantic segmentation neural network based on residual learning and U-Net for road area extraction. Kohl et al. (2018) proposed a generative segmentation model based on a combination of a U-Net with a conditional variational auto-encoder. A new Recurrent U-Net had been proposed by Wang et al. (2019a). This model not only retained the compactness of U-Net, but also achieved a good performance improvement in some benchmarks. TernausNet was proposed by Iglovikov and Shvets (2018). The network replaces the encoder in U-Net with VGG11 and conducts pre-training on ImageNet. TernausNet achieved the best results in the Kaggle Carvana Image Masking Challenge.

Although the architecture of U-Net has been widely used, the most basic architecture has not changed. The convolution blocks of the original U-Net network are adjusted by us to improve the efficiency of the segmentation algorithm. The convolution blocks are arranged in parallel to form a multiple convolution sequence. Richer semantic information is provided by this design. In addition, the convolution kernel of the multiple convolution sequence is adjusted to have different receptive fields. The convolution kernel with different receptive fields enables the network to better extract and restore features.

3. METHOD

The proposed MSU-Net consists of major part: multi-scale block (37), as shown in **Figure 1**. In the following, we first trace the types of multi-scale block and then explain the structure of MSU-Net and extended work of multi-scale block.

3.1. Multi-Scale Block

The multi-scale block is proposed by us, which is composed of multiple convolution sequences with different receptive fields. More diverse semantic information is extracted by this module and more detailed feature maps are generated. The widely used convolution kernel is shown in **Figure 2**.

The convolution kernel with different receptive fields is matched to obtain a multi-scale block. We designed 31 kinds of multi-scale blocks according to the above several convolution kernels. The multi-scale block evolved from the different convolution kernels is shown in **Figure 3**.

The 3×3 convolution kernel has been used in all experiments. The features of the input multi-scale block are processed by the convolution kernel with different receptive fields, and then the obtained features are output after 1×1 convolution. A comprehensive ablation experiment is used to verify the performance of different types of multi-scale blocks. In the experiment, three datasets are used by us. The datasets are EM, BUL, and CXR, respectively (detailed in section 4.1). The experiments are carried out after integrated each multi-scale block into the original U-Net. The experimental results are illustrated in **Table 1**. The performance of multi-scale block (37) is the best. The details of multi-scale block (37) are shown in **Figure 4**.

x represents the characteristics of the input. x_1 and x_2 represent the characteristics obtained by the convolution kernel of different sizes. F is the output result of multi-scale block. F is computed as follows:

$$x_1 = w_{32}(w_{31}x + b_{31}) + b_{32} \tag{1}$$

$$x_2 = w_{72}(w_{71}x + b_{71}) + b_{72} \tag{2}$$

$$X = Cat[x_1, x_2] \tag{3}$$

$$F = w_f X + b_f \tag{4}$$

Feature fusion needs to be used in multi-scale block before 1X1 convolution. Therefore, different fusion methods are validated by us (results in **Table 2**). MSU-Net (37+sum) uses element

TABLE 1 | Ablation study on MSU-Nets of the convolution kernel with different receptive fields.

Applications	BUL	EM	NS
	M ± SD	M ± SD	M ± SD
MSU-Net (13)	0.548 ± 0.076	0.871 ± 0.002	0.678 ± 0.017
MSU-Net (23)	0.610 ± 0.029	0.840 ± 0.035	0.661 ± 0.028
MSU-Net (35)	0.690 ± 0.047	0.884 ± 0.017	0.670 ± 0.036
MSU-Net (37)	**0.708 ± 0.011**	**0.900 ± 0.001**	**0.702 ± 0.010**
MSU-Net (39)	0.699 ± 0.016	0.895 ± 0.009	0.660 ± 0.011
MSU-Net (123)	0.547 ± 0.067	0.862 ± 0.012	0.672 ± 0.015
MSU-Net (135)	0.679 ± 0.005	0.883 ± 0.010	0.676 ± 0.021
MSU-Net (137)	0.696 ± 0.018	0.890 ± 0.015	0.684 ± 0.025
MSU-Net (139)	0.682 ± 0.037	0.880 ± 0.015	0.674 ± 0.020
MSU-Net (235)	0.673 ± 0.036	0.873 ± 0.023	0.684 ± 0.025
MSU-Net (237)	0.703 ± 0.042	0.888 ± 0.017	0.687 ± 0.019
MSU-Net (239)	0.664 ± 0.029	0.893 ± 0.011	0.672 ± 0.023
MSU-Net (357)	0.679 ± 0.018	0.888 ± 0.016	0.682 ± 0.015
MSU-Net (359)	0.693 ± 0.007	0.894 ± 0.006	0.686 ± 0.020
MSU-Net (379)	0.705 ± 0.008	0.894 ± 0.011	0.671 ± 0.023
MSU-Net (1,235)	0.652 ± 0.015	0.877 ± 0.015	0.662 ± 0.038
MSU-Net (1,237)	0.655 ± 0.008	0.886 ± 0.009	0.693 ± 0.025
MSU-Net (1,239)	0.699 ± 0.017	0.885 ± 0.014	0.687 ± 0.031
MSU-Net (1,357)	0.689 ± 0.033	0.895 ± 0.005	0.673 ± 0.023
MSU-Net (1,359)	0.700 ± 0.028	0.898 ± 0.002	0.689 ± 0.015
MSU-Net (1,379)	0.702 ± 0.025	0.898 ± 0.003	0.692 ± 0.017
MSU-Net (2,357)	0.694 ± 0.040	0.894 ± 0.004	0.687 ± 0.023
MSU-Net (2,359)	0.681 ± 0.023	0.884 ± 0.014	0.702 ± 0.018
MSU-Net (2,379)	0.694 ± 0.036	0.882 ± 0.014	0.675 ± 0.013
MSU-Net (3,579)	0.696 ± 0.338	0.893 ± 0.010	0.695 ± 0.011
MSU-Net (12,357)	0.680 ± 0.017	0.893 ± 0.005	0.696 ± 0.027
MSU-Net (12,359)	0.705 ± 0.014	0.892 ± 0.006	0.687 ± 0.040
MSU-Net (12,379)	0.667 ± 0.023	0.893 ± 0.002	0.695 ± 0.021
MSU-Net (13,579)	0.697 ± 0.032	0.899 ± 0.001	0.685 ± 0.025
MSU-Net (23,579)	0.705 ± 0.020	0.889 ± 0.014	0.697 ± 0.008
MSU-Net (123,579)	0.693 ± 0.028	0.896 ± 0.002	0.696 ± 0.017

*The numbers in brackets represent the size of receptive field in MSU-Net. This is corresponds to the different multi-scale blocks in **Figure 3**. Intersection over Union (IoU) is used as the evaluation metric for comparative. Bold values represent the best results.*

summation for feature fusion. MSU-Net (37) uses concatenation for feature fusion.

The dilated convolution is introduced into the multi-scale block after the optimal convolution kernel is obtained. The dilated convolution used in the experiment is described in **Figure 2**. Convolution kernels with different receptive fields are concatenated to verify the effectiveness of the multiple convolution sequence. The details are shown in **Figure 5**. The experimental results are shown in **Table 2**.

3.2. Network Architecture

The architecture of MSU-Net is illustrated in **Figure 1**. MSU-Net has a contraction path and an expansion path. The network architecture follows encoder-decoder. In original U-Net, each block consists of two convolutional layers. However, there is still a drawback in this block. Due to the limitation of the receptive field, the network does not achieve better performance in feature extraction and feature restoration. The convolution blocks in encoder of the original U-Net are replaced with multi-scale blocks to obtain MSU-Net (encoder). The convolution blocks in decoder of the original U-Net are replaced with multi-scale blocks to obtain MSU-Net (decoder). The experimental results are illustrated in **Table 2**. In MSU-Net, the multi-scale block (37) is used to replace the all convolution block in the original U-Net. Multi-scale block enables encoder to extract more detailed information. Multi-scale block makes the features of decoder restoration more complete.

3.3. Extension of Model

Residual (He et al., 2016) is expanded into our model. The residual multi-scale block is shown in **Figure 6**. In addition, multi-scale blocks are also extended to variants of U-Net.

3.3.1. Residual Multi-Scale Block

The idea of residual is introduced with multi-scale blocks to obtain residual multi-scale block (0) and residual multi-scale block (1). Residual multi-scale block (0) and residual multi-scale block (1) are shown in **Figures 6A,B**, respectively. The original convolution block in U-Net was replaced by residual multi-scale block (0) and residual multi-scale block (1) to get Res MSU-Net (0) and Res MSU-Net (1). The experimental results are described in **Table 4**. In **Table 4**, the performance of residual multi-scale block (0) is better than residual multi-scale block (1).

The structure of residual multi-scale block (1) is described below. x_r represents the characteristics of the input. x_{r1} and x_{r2} represent the characteristics obtained by the convolution kernel of different receptive fields. F_R is the output result of the multi-scale block. F_R is computed as follows:

$$x_{r1} = w_{r32}(w_{r31}x_r + b_{r31}) + b_{r32} \tag{5}$$

$$x_{r2} = w_{r72}(w_{r71}x_r + b_{r71}) + b_{r72} \tag{6}$$

$$X_R = Cat[x_r, x_{r1}, x_{r2}] \tag{7}$$

$$F_R = w_{rf}X_R + b_{rf} \tag{8}$$

Residual connection can make the forward and backward propagation of multi-scale block smoother. In forward propagation, the input signal can be propagated directly from the bottom to the top. The problem of network degradation can be alleviated. In back propagation, the error signal can be propagated directly to the lower layer without any intermediate weight matrix transformation. The problem of gradient dispersion can be alleviated. In addition, the generalization capacity of the network can be enhanced by the structure.

3.3.2. Other Structures

In addition to combining the structure with our proposed multi-scale block, we also extend our multi-scale block on the variants of original U-Net. The convolution blocks in AttU-Net (Oktay et al., 2018) and U-Net++ (Zhou et al., 2020)

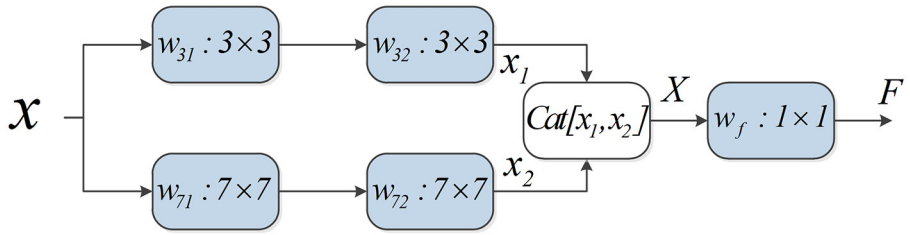

FIGURE 4 | Detailed description of multi-scale block. First, two 3X3 and 7X7 convolution kernels are used to extract features. Second, the extracted features are merged by the feature by cat. Finally, the fused features are output after dimensionality reduction by 1X1 convolution.

TABLE 2 | Ablation study for MSU-Net and its variants.

Architecture	BUL M ± SD	EM M ± SD	NS M ± SD
MSU-Net	**0.708 ± 0.011**	**0.900 ± 0.001**	**0.702 ± 0.010**
MSU-Net(37+sum)	0.694 ± 0.020	0.894 ± 0.013	0.683 ± 0.017
MSU-Net(encoder)	0.646 ± 0.061	0.889 ± 0.013	0.679 ± 0.021
MSU-Net(decoder)	0.656 ± 0.027	0.883 ± 0.018	0.661 ± 0.024
MSU-Net(37+concatenated)	0.642 ± 0.036	0.899 ± 0.004	0.674 ± 0.024
MSU-Net(73+concatenated)	0.707 ± 0.061	**0.900 ± 0.001**	0.667 ± 0.022
MSU-Net(37+dilated)	0.640 ± 0.033	0.877 ± 0.005	0.662 ± 0.013

*MSU-Net is MSU-Net (37) in **Table 1**. MSU-Net (37+ sum) is an MSU-Net with feature fusion by adding. MSU-Net (encoder) and MSU-Net (decoder) are obtained by using multi-scale block to replace the convolution block between encoder and decoder in U-Net. MSU-Net (73+concatenated) and MSU-Net (37+concatenated) are obtained after concatenated the convolution kernel with different receptive fields. MSU-Net (37+dilated) is obtained by dilated convolution. Intersection over Union (IoU) is used as the evaluation metric for comparison. Bold values represent the best results.*

are replaced with multi-scale block, namely MSAttU-Net and MSU-Net++, respectively.

4. EXPERIMENT

4.1. Dataset

Table 3 summarizes the five biomedical image segmentation datasets used in this study. These lesions/organs are derived from the most common medical imaging modalities, such as microscopy, X-ray, B-mode ultrasound, etc. The dataset was randomly divided into six subsets. Five of six are used as a training-validation dataset, and the remaining data as a test dataset. Five-fold cross validation is applied by randomly dividing training-validation into five subsets. The training process alternates with a fixed ratio of 4:1 between the training dataset and the validation dataset.

(1) Electron Microscopy (EM): The dataset is provided by the EM segmentation challenge (Cardona et al., 2010), which is a part of ISBI 2012. The dataset contains 30 images (512 × 512 pixels) from a serial section Transmission Electron Microscopy (ssTEM) dataset of the Drosophila first instar larva ventral nerve cord (VNC). The images has not been resized. The images size of the input network is 512 × 512. An example of dataset is shown in **Figure 7**. Each image has a completely annotated ground truth segmentation map of the corresponding cell (white) and membranes (black).

(2) Breast Ultrasound Lesions (BUL): The Breast Ultrasound Dataset B (BUL) open-sourced in (Yap et al., 2017) is used in this study. This dataset includes 163 ultrasound images of breast lesions from different women. The image size of average is 760 × 570 pixels where each of the images presented one or more lesions. For our experiments, the data is resampled to 128 × 128 pixels. The ground truths provided in the BUL are in the form of binary masks of the lesions, as illustrated in **Figure 7**.

(3) Chest X-ray (CXR): The standard digital image database for Tuberculosis (Candemir et al., 2013; Jaeger et al., 2013) is created by the National Library of Medicine, Maryland, USA in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. The Chest X-rays are from out-patient clinics. There are 800 images in the Chest X-rays dataset. However, the ground truth of 96 images is unknown. Seven hundred and four images of corresponding GT in the dataset were used by us. The image size of average is 4456 × 4456 pixels. The images are rescaled to 128 × 128 for this implementation. Referring to the example in **Figure 7**.

(4) Skin Lesions (SL): The dataset is provided by the ISIC 2018: Skin Lesion Analysis Toward Melanoma Detection grand challenge dataset (Tschandl et al., 2018; Codella et al., 2019). This dataset consists of 2594 RGB images of skin lesions with an average image size of 2166 × 3188 pixels. For our experiments, the dataset is resampled to 256 × 256 pixels with cross validation. The training samples include the original image and the binary image containing the lesion. Pixels outside the target lesion are represented by 0.

(5) Nuclei Segmentation (NS): This dataset is provided by The Cancer Genome Atlas (TCGA). This dataset can be downloaded from Kaggle. The dataset comprising 30 digitized Hematoxylin and Eosin (H&E)-stained frozen sections (512 × 512 pixels) derived from 10 different human organs. The dataset were selected from different laboratories to maximize the staining variability in the data set. Image tiles (3 per tissue) were extracted from adrenal gland, larynx, lymph nodes, mediastinum, pancreas, pleura, skin, testes, thymus, and thyroid gland. Like the EM dataset, this dataset was not sampled prior to input. The image size of the input is 512 × 512.

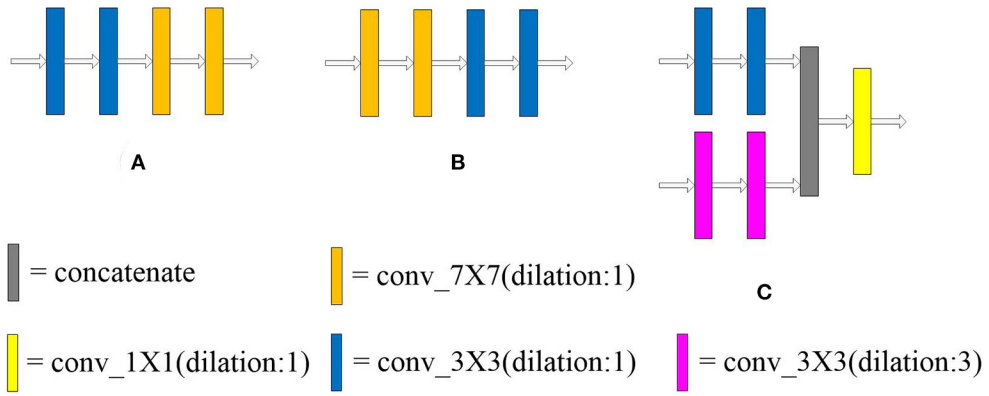

FIGURE 5 | Arrangement of convolution kernel with different receptive fields and dilated convolution. Panels **(A,B)** lay out the convolution kernel in different order, respectively. In **(C)**, the large convolution kernel in the multi-scale block (37) is replaced by dilated convolution. The receptive field of the convolution kernel is enlarged without increasing the number of parameters.

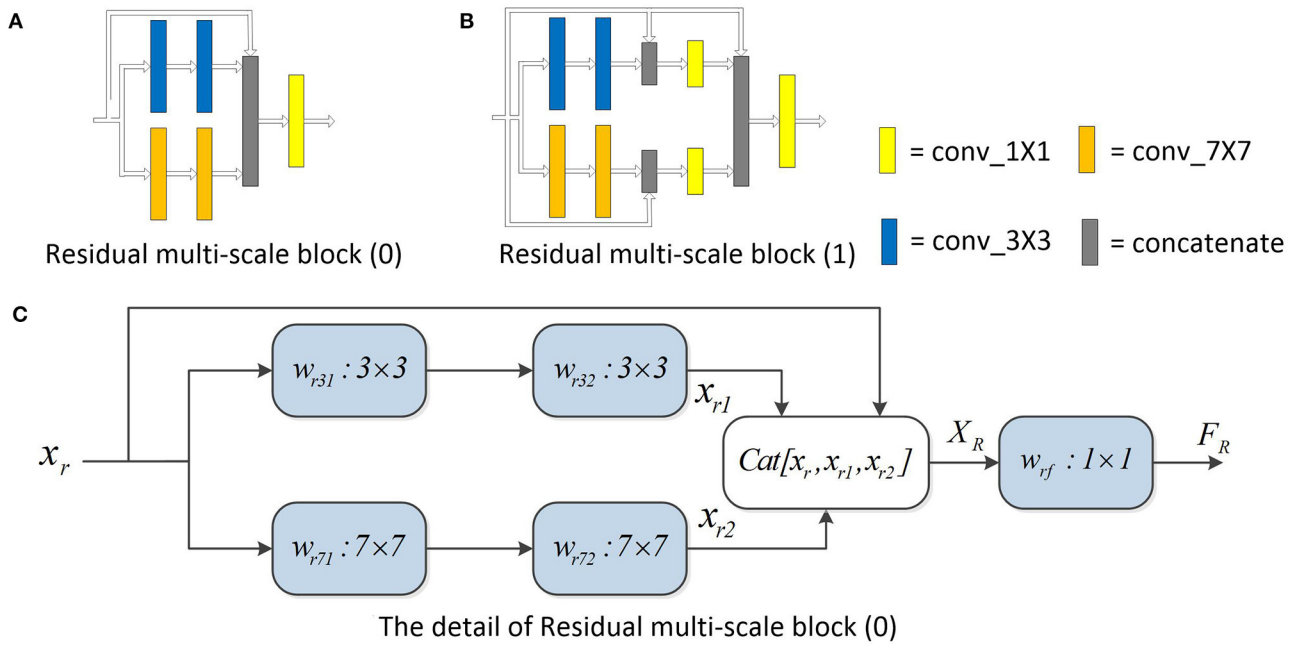

FIGURE 6 | Residual multi-scale block. Panel **(A)** is the first structure designed to incorporate residual thinking. Panel **(B)** is the second. Experimental results show that the performance of **(A)** structure is better than **(B)**. Panel **(C)** describes **(A)** in detail. The different multi-scale blocks are described in **Figure 3**. The information of the input features is directly transmitted to the deep layer of the network through the residual connection.

4.2. Baselines and Implementation

For comparison, the original U-Net is used to implement the segmentation task. U-Net is a common performance baseline for medical image segmentation. In addition, a wide U-Net with a similar number of parameters to our proposed architecture was designed. This is to ensure that the performance gain yielded by our architecture is not simply due to the increased number of parameters.

In this experiment, the program was based on the Pytorch (Paszke et al., 2019) framework. SGD (Robbins and Monro, 1951) was used as the optimizer with the learning rate of 1e-2. Both networks were constructed from the original U-Net. All the experiments are performed using an NVIDIA GeForce RTX 2080 Ti GPUs with 11 GB memory.

4.3. Evaluation Measures

In this paper, the Intersection over Union (IoU) is used as the main evaluation indicator to evaluate the results. Alternative measurement metrics could be found in **Table 6**, such as dice coefficient, precision, area Under Curve (AUC), and statistical analysis. These metrics were calculated as follows:

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

TABLE 3 | Summary of biomedical image segmentation datasets used in our experiments.

Applications	Images	Input size	Modality	Provider
EM	30	512 × 512	Microscopy	ISBI 2012 (Cardona et al., 2010)
BUL	163	128 × 128	Ultrasound	Breast Ultrasound Lesions Dataset (Yap et al., 2017)
CXR	704	128 × 128	X-ray	Chest X-ray Database (Candemir et al., 2013; Jaeger et al., 2013)
SL	2594	256 × 256	Demoscopy	ISIC 2018 (Tschandl et al., 2018; Codella et al., 2019)
NS	30	512 × 512	Digitize	Kaggle

TABLE 4 | Ablation study for U-Net, wide U-Net, MSU-Net, Res MSU-Net(0), and Res MSU-Net(1).

Architecture	BUL	EM	NS
	M ± SD	M ± SD	M ± SD
U-Net (Ronneberger et al., 2015)	0.608 ± 0.037	0.884 ± 0.007	0.675 ± 0.018
wide U-Net (Ours)	0.643 ± 0.025	0.889 ± 0.016	0.677 ± 0.012
MSU-Net (Ours)	0.708 ± 0.011	**0.900 ± 0.001**	0.702 ± 0.010
Res MSU-Net (0) (Ours)	**0.713 ± 0.032**	**0.900 ± 0.001**	**0.704 ± 0.010**
Res MSU-Net (1) (Ours)	0.628 ± 0.025	0.848 ± 0.056	0.675 ± 0.022

Wide U-Net is obtained by extending the width of the U-Net network. The wide U-Net has the same number of parameters as the MSU-Net. Res MSU-Net (0)/Res MSU-Net (1) are proposed based on Residual multi-block. Intersection over Union (IoU) is used as the evaluation metric for comparison. Bold values represent the best results.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

where TP, FP, and FN represent the number of true positive, false positive, and false negative, respectively. In addition, the area under receiver operation characteristic curve (AUC) is used to measure the segmentation performance. The closer the AUC is to 1.0, the higher authenticity of the segmentation method. When it is equal to 0.5, it has the lowest authenticity and no application value.

5. RESULTS

5.1. Selection of Multi-Scale Block

31 kinds of multi-scale blocks were designed by combining the convolution kernel with different receptive fields. The different multi-scale blocks are shown in **Figure 3**. All multi-scale blocks were embedded into the original U-Net respectively. Subsequently, an ablation analysis of multi-scale block is made on three datasets. The experimental results of different multi-scale blocks on the dataset are illustrated in **Table 1**. Two key findings are illustrated in our results: (1) The wider network structure is not always better, (2) The optimal width of the network depends on the difficulty and size of the dataset. Although these findings may facilitate the automatic search of neural structures, this approach is hampered by limited computational resources (Elsken et al., 2018; Liu et al., 2018, 2019; Zoph et al., 2018).

The influence of the difference receptive field on the network performance is shown in **Table 1**. Among them, multi-scale block (37) achieves the best performance on datasets.

Different arrangements of convolution blocks and different convolution kernels are verified in **Table 2**. The robustness of the multiple convolution sequence is demonstrated by experimental results.

5.2. Results of the Extended Model

The multi-scale block was extended by us. First, the idea of residuals was introduced into the proposed module. Two multi-scale blocks based on residuals were constructed. The structure is shown in **Figure 6**. Second, the proposed multi-scale block was extended to the existing U-Net variants. Convolution kernel in AttU-Net and U-Net++ was replaced by multi-scale block. The experimental results are shown in **Tables 4**, **5**. Experimental results show that the proposed method has good scalability and compatibility.

It can be seen from the experimental results that the performance of wide U-Net is better than U-Net. The main reason is that there are more parameters in wide U-Net. When the residual idea is not introduced, MSU-Net achieves very robust performance on all three data sets. Compared with U-Net, MSU-Net is higher than 0.1, 0.016, and 0.027 on the three datasets. The performance of the network is improved by introducing residual ideas. In addition, the extended experiment on U-Net variants also confirmed the effectiveness and universality of multi-scale block. By comparing the performance of MSU-Net (37+encoder) and U-Net, we found that the ability of network to extract features was enhanced by combining multi-scale blocks.

5.3. Semantic Segmentation Results

In order to verify the performance of the network, MSU-Net was compared with the current more advanced segmentation network (Ronneberger et al., 2015; Badrinarayanan et al., 2017; Chen et al., 2018b; Zhou et al., 2020). In addition, chest X-ray and skin lesion segmentation datasets were added to the experiment. These two datasets are larger than the three previously mentioned datasets. **Figure 7** depicts a qualitative comparison of the results between the different split schemas. Compared with other architectures, the segmentation results of MSU-Net are more detailed. SegNet cannot be trained on EM datasets. Therefore, SegNet has not experimented on the EM dataset.

Table 6 shows the segmentation performance of the architectures on different datasets. A statistical analysis based on independent two-sample *t*-tests is performed by us for each

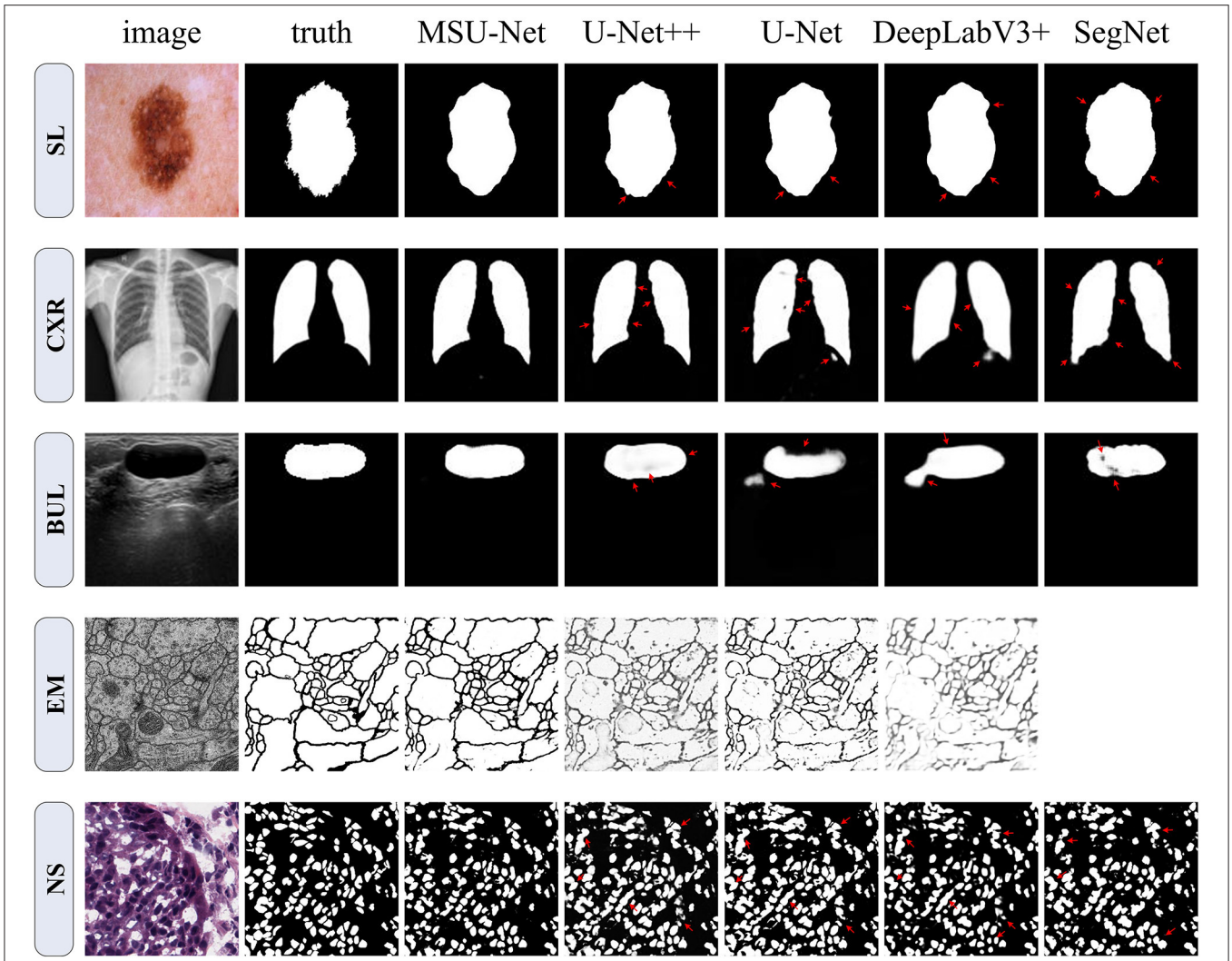

FIGURE 7 | Qualitative comparison among SegNet, DeepLabV3+, U-Net, U-Net++, and MSU-Net. It shows the segmentation application of the architectures on five different biomedical image datasets. The red arrows indicate areas of incorrect segmentation. SegNet can not be trained on EM datasets. Therefore, the result of SegNet on the EM dataset is vacant. The ground truth is illustrated in the second column (from **left** to **right**).

TABLE 5 | Ablation study for AttU-Net, MSAttU-Net, U-Net++, and MSU-Net++.

Architecture	BUL	EM	NS
	M ± SD	M ± SD	M ± SD
AttU-Net (Oktay et al., 2018)	0.607 ± 0.039	0.853 ± 0.043	0.655 ± 0.020
MSAttU-Net (Ours)	0.674 ± 0.005	0.895 ± 0.004	0.677 ± 0.010
U-Net++ (Zhou et al., 2020)	0.670 ± 0.020	0.885 ± 0.013	0.665 ± 0.012
MSU-Net++ (Ours)	0.687 ± 0.009	0.895 ± 0.002	0.691 ± 0.022

MSAttU-Net and MSU-Net ++ are extended versions of AttU-Net and U-Net ++. Intersection over Union (IoU) is used as the evaluation metric for comparison.

pair of data between different structures. Our results show that MSU-Net is an effective network structure.

The results in **Table 5** suggest that our proposed MSU-Net is more robust in semantic segmentation. Compared with the U-Net, MSU-Net achieves a significant IoU gain over both architectures for all the five tasks of SL (↑0.01), CXR (↑0.01), BUL (↑0.1), EM (↑0.016), NS (↑0.027) segmentation. AUC of different architectures on the data set is illustrated in **Figure 8**. **Figure 8** shows the ROC curve of different architectures on the datasets. Our model achieves the best performance in all datasets. Fine Precision is not captured by our model on the SL dataset. However, the high sensitivity of our model is shown in **Figure 8**. This allows false positives and false negatives in the data to be better balanced by our model. It is mainly due to the multiple convolution sequence with different receptive fields. This design makes the features in the network richer and more diverse.

6. DISCUSSION

Medical image segmentation plays an important role in diagnosis, treatment and prognosis evaluation. In the process of diagnosis, the main applications include morphological

TABLE 6 | Semantic segmentation results measured by different metrics for different network architectures.

Metric	Architecture	SL		CXR		BUL		EM		NS	
		M ± SD	*p*-value	M ± SD	*p*-value	M ± SD	*p*-value	M ± SD	*p*-value	M ± SD	*p*-value
IoU	SegNet (Badrinarayanan et al., 2017)	0.752 ± 0.007	9.824e-4	0.832 ± 0.008	6.179e-5	0.630 ± 0.033	0.001	—	—	0.586 ± 0.021	4.084e-6
	DeepLabV3+ (Chen et al., 2018b)	0.762 ± 0.002	2.202e-3	0.847 ± 0.005	3.261e-4	0.558 ± 0.034	1.761e-5	0.837 ± 0.015	1.582e-5	0.582 ± 0.019	1.717e-6
	U-Net (Ronneberger et al., 2015)	0.751 ± 0.005	1.872e-4	0.857 ± 0.005	0.020	0.608 ± 0.037	4.789e-4	0.884 ± 0.007	6.873e-4	0.675 ± 0.018	0.020
	U-Ne++ (Zhou et al., 2020)	0.746 ± 0.008	2.725e-4	0.863 ± 0.004	0.232	0.670 ± 0.020	0.013	0.885 ± 0.013	0.031	0.665 ± 0.012	8.243e-4
	MSU-Net(Ours)	**0.771 ± 0.004**	—	**0.867 ± 0.006**	—	**0.708 ± 0.011**	—	**0.900 ± 0.001**	—	**0.702 ± 0.011**	—
Dice	SegNet (Badrinarayanan et al., 2017)	0.852 ± 0.006	0.002	0.908 ± 0.005	6.393e-5	0.770 ± 0.026	0.002	—	—	0.738 ± 0.017	5.941e-6
	DeepLabV3+ (Chen et al., 2018b)	0.857 ± 0.003	0.002	0.917 ± 0.003	3.123e-4	0.713 ± 0.029	3.215e-5	0.911 ± 0.009	2.104e-5	0.734 ± 0.016	2.830e-6
	U-Net (Ronneberger et al., 2015)	0.850 ± 0.004	1.696e-4	0.923 ± 0.003	0.020	0.753 ± 0.029	6.919e-4	0.938 ± 0.004	7.314e-4	0.805 ± 0.013	0.022
	U-Ne++ (Zhou et al., 2020)	0.847 ± 0.006	2.892e-4	0.926 ± 0.002	0.230	0.800 ± 0.014	0.015	0.939 ± 0.007	0.032	0.797 ± 0.008	5.129e-4
	MSU-Net(Ours)	**0.865 ± 0.003**	—	**0.929 ± 0.004**	—	**0.827 ± 0.008**	—	**0.947 ± 0.001**	—	**0.824 ± 0.007**	—
Precision	SegNet (Badrinarayanan et al., 2017)	0.886 ± 0.010	0.161	0.856 ± 0.009	4.465e-4	0.725 ± 0.040	0.115	—	—	0.873 ± 0.008	0.203
	DeepLabV3+ (Chen et al., 2018b)	0.892 ± 0.008	0.037	0.875 ± 0.005	0.029	0.798 ± 0.054	2.227e-4	0.864 ± 0.029	6.076e-4	0.860 ± 0.019	0.065
	U-Net (Ronneberger et al., 2015)	**0.899 ± 0.014**	0.024	0.878 ± 0.006	0.079	0.760 ± 0.061	0.018	0.913 ± 0.014	0.007	**0.888 ± 0.019**	0.917
	U-Net++ (Zhou et al., 2020)	0.895 ± 0.010	0.030	0.882 ± 0.005	0.274	0.786 ± 0.043	0.011	0.919 ± 0.025	0.196	0.853 ± 0.059	0.267
	MSU-Net(Ours)	0.873 ± 0.015	—	**0.887 ± 0.009**	—	**0.842 ± 0.006**	—	**0.935 ± 0.003**	—	0.887 ± 0.021	—

We have performed independent two sample t*-test between and highlighted boxes in red when the differences are statistically significant (p<0.05). Bold values represent the best results.*

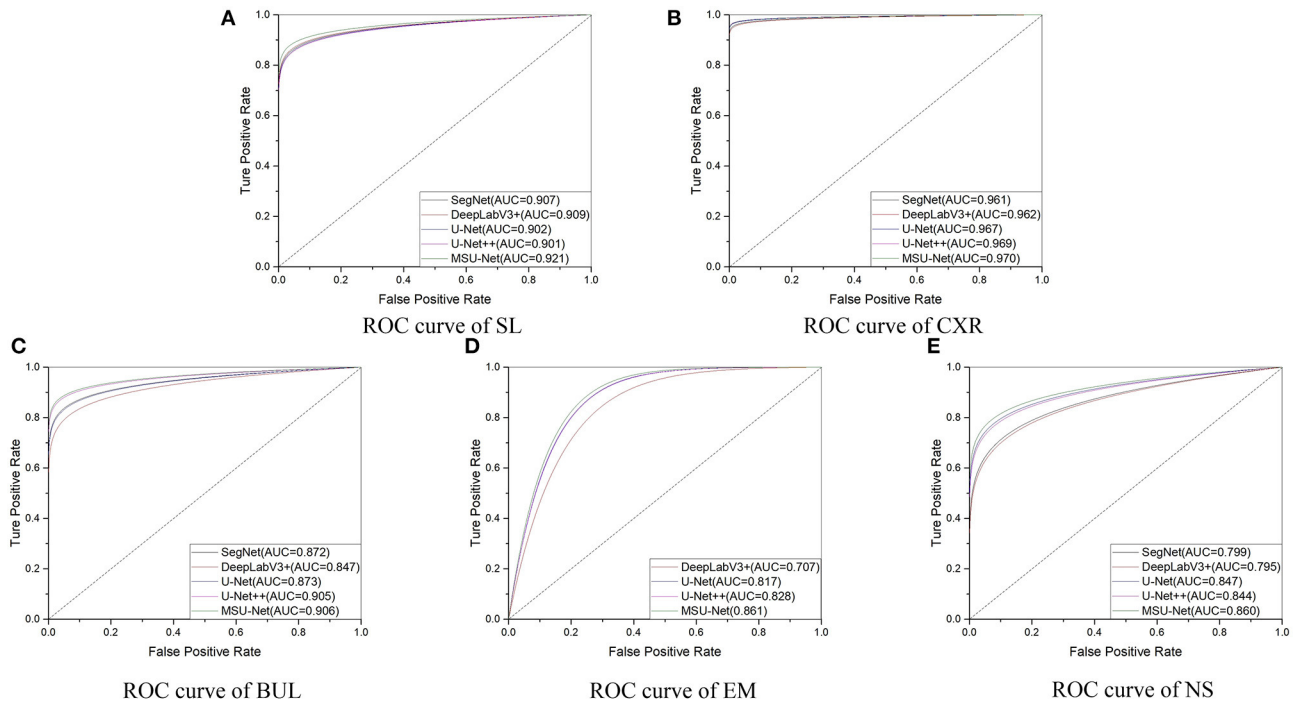

FIGURE 8 | ROC curves for different architectures. AUC is the area under the curve.

analysis, volume calculation, anatomical structure analysis, etc. In surgical treatment planning, the commonly used methods include preoperative biopsy guidance, target area planning of radiotherapy, image registration fusion and path planning, and target tracking in medical robot, etc. In the prognostic assessment, the most important segmentation is the analysis of lesion volume change and the analysis of lesion histological characteristics. In addition, medical image segmentation can be applied to three-dimensional reconstruction visualization, which can provide clinicians with more intuitive pathological morphology and spatial anatomy. In recent years, the method based on deep learning has been widely used in medical image segmentation. However, the performance of segmentation is greatly affected by the network architecture and the ability to acquire features in learning process.

U-Net is a very classical network architecture in the field of medical image segmentation. At present, U-Net is widely used in medical image segmentation. However, the basic architecture of U-Net has not been significantly modified by the researchers. Large receptive fields play an important role when we need to make dense per-pixel predictions. In order to improve the existing segmentation model, multi-scale blocks are constructed by convolution sequence and multiple convolution kernel with different receptive fields. The different types of multi-scale blocks are illustrated in **Figure 3**. In addition, MSU-Net is proposed after all the convolution blocks in the original U-Net are replaced by multi-scale block. The details of the MSU-Net are illustrated in **Figure 1**. Multiple convolution sequences are used to extract more semantic features from images. In addition, convolution kernels with different receptive fields are used to make features more diverse. The problem of unknown network width is alleviated by effective integration of multiple convolution sequences with different receptive fields.

The most important innovation described in this paper is the combination of multiple convolution sequences and convolution kernel with different receptive fields to improve the segmentation performance. It can be seen from the **Table 1** that the performance of the network is affected by different receptive fields. Good performance was achieved by combining advanced ideas with multi-scale blocks. In addition, multi-scale blocks are extended to the variants of original U-Net. The results in **Tables 4**, **5** describes that the segmentation performance of the network is improved by combining the multiple convolution sequence and the convolution kernel with the different receptive fields. The strategies of our proposed strategy has the following advantages: (1) More diverse features are extracted through the convolution kernel of different receptive fields. This is useful for intensive forecasting tasks that require detailed spatial information. At the same time, the problem of unknown network width can be alleviated. (2) More feature information is extracted by multi-convolution sequence, which is helpful to the segmentation task. Our method has obtained the best performance compared with the advanced models through the demonstration of multiple medical image segmentation datasets (see in **Table 6**). The highest AUC is obtained by our architecture (see in **Figure 8**). This suggests that our model has a stronger ability to balance false positives and false negatives in the data. In general, the proposed method is useful for intensive

forecasting tasks requiring detailed spatial information. Different receptive fields can provide diverse semantic information for tasks, which is beneficial to the segmentation of lesions. More detailed segmentation results can provide doctors with more detailed lesion areas, which is helpful for the diagnosis of disease and the formulation of treatment plan.

Although we have widely evaluated the performance of the network on different datasets, there are still some deficiencies in our network. First, the convolution kernels with a larger receptive field are not attempted due to objective factors. The performance of the network may be improved through greater receptive field. Second, the dilated convolution can increase the receptive field of the convolution kernel without increasing the number of parameters. Unfortunately, dilated convolution was not attempted in our experiment. Third, our network has not been validated against the 3D medical image segmentation dataset. The above work may be completed by us in the future.

7. CONCLUSION

In order to obtain more accurate segmentation image, a new structure called multi-scale block was proposed by us. The convolution blocks in the original U-Net are replaced by multi-scale blocks to obtain MSU-Net. The improvement of MSU-Net performance is attributed to multiple convolution sequence and convolution kernels with different receptive fields. Two key issues are addressed by this design: (1) The diversity of features is lost due to the fixed size of the convolution kernel. (2) Feature information may be lost at each scale using a single convolutional sequence to extract features. Five different public datasets were used to conduct an extensive evaluation of MSU-Net. The experimental results show that MSU-Net achieves the best performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: all datasets can be found in **Table 3**.

AUTHOR CONTRIBUTIONS

RS, DZ, and JL: conceptualization and writing (review and editing). RS, DZ, and CC: data curation. RS and DZ: methodology, validation, and writing (original draft). RS: project administration and visualization. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was supported by two research grants: (1) National Natural Science Foundation of China (62033002). (2) Science and Technology Project grant from Anhui Province (Grant Nos. 1508085QHl84, 201904a07020098). (3) Fundamental Research Fund for the Central Universities (Grant No. WK 9110000032).

ACKNOWLEDGMENTS

The authors express their sincere gratitude to the creator of the public dataset for many valuable discussions and educational help in the growing field of medical image analysis.

REFERENCES

Abramoff, M. D., Alward, W. L., Greenlee, E. C., Shuba, L., Kim, C. Y., Fingert, J. H., et al. (2007). Automated segmentation of the optic disc from stereo color photographs using physiologically plausible features. *Investig. Ophthalmol. Vis. Sci.* 48, 1665–1673. doi: 10.1167/iovs.06-1081

Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., and Asari, V. K. (2018). Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation. *arXiv [preprint]. arXiv:1802.06955*. doi: 10.1109/NAECON.2018.8556686

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 2481–2495. doi: 10.1109/TPAMI.2016.2644615

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., et al. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Trans. Med. Imaging* 33, 577–590. doi: 10.1109/TMI.2013.2290491

Cardona, A., Saalfeld, S., Preibisch, S., Schmid, B., Cheng, A., Pulokas, J., et al. (2010). An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. *PLoS Biol.* 8:e1000502. doi: 10.1371/journal.pbio.1000502

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D. (2018a). Drinet for medical image segmentation. *IEEE Trans. Med. Imaging* 37, 2453–2462. doi: 10.1109/TMI.2018.2835303

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 11211, 833–851. doi: 10.1007/978-3-030-01234-2_49

Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*. 2843–2851. Available online at: https://dl.acm.org/doi/10.5555/2999325.2999452

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., et al. (2019). Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). *arXiv [preprint].*. Available online at: https://arxiv.org/abs/1902.03368

Elsken, T., Metzen, J. H., and Hutter, F. (2018). Neural architecture search: a survey. *arXiv [preprint]. arXiv:1808.05377*. doi: 10.1007/978-3-030-05318-5_11

Fan, F., Huang, Y., Wang, L., Xiong, X., Jiang, Z., Zhang, Z., et al. (2019). A semantic-based medical image fusion approach. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1906.00225

Ge, R., Yang, G., Chen, Y., Luo, L., Feng, C., Ma, H., et al. (2019). K-Net: Integrate left ventricle segmentation and direct quantification of paired echo sequence. *IEEE Trans. Med. Imaging* 39, 1690–1702. doi: 10.1109/TMI.2019.2955436

Guan, S., Khan, A. A., Sikdar, S., and Chitnis, P. V. (2019). Fully dense unet for 2-D sparse photoacoustic tomography artifact removal. *IEEE J. Biomed. Health Inform.* 24, 568–576. doi: 10.1109/JBHI.2019.2912935

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. doi: 10.1109/CVPR.2016.90

He, P., Huang, W., Qiao, Y., Loy, C. C., and Tang, X. (2015). Reading scene text in deep convolutional sequences. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1506.04395

Iglovikov, V., and Shvets, A. (2018). Ternausnet: U-Net with VGG11 encoder pre-trained on imagenet for image segmentation. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1801.05746

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* 33, 233–245. doi: 10.1109/TMI.2013.2284099

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Kitrungrotsakul, T., Han, X.-H., and Chen, Y.-W. (2015). "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," in *2015 IEEE International Conference on Image Processing (ICIP)* (IEEE), 3368–3371. doi: 10.1109/ICIP.2015.7351428

Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J. R., Maier-Hein, K., et al. (2018). "A probabilistic U-Net for segmentation of ambiguous images," in *Advances in Neural Information Processing Systems*, 6965–6975. Available online at: https://arxiv.org/abs/1806.05034v4

LaLonde, R., and Bagci, U. (2018). Capsules for object segmentation. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1804.04241

Le, N. Q. K., Ho, Q.-T., Yapp, E. K. Y., Ou, Y.-Y., and Yeh, H.-Y. (2020). Deepetc: A deep convolutional neural network architecture for investigating and classifying electron transport chain's complexes. *Neurocomputing* 375, 71–79. doi: 10.1016/j.neucom.2019.09.070

Le, N. Q. K., Yapp, E. K. Y., Nagasundaram, N., and Yeh, H.-Y. (2019). Classifying promoters by interpreting the hidden information of dna sequences via deep learning and combination of continuous fasttext n-grams. *Front. Bioeng. Biotechnol.* 7:305. doi: 10.3389/fbioe.2019.00305

Li, X., Wang, Y., Tang, Q., Fan, Z., and Yu, J. (2019). Dual U-Net for the segmentation of overlapping glioma nuclei. *IEEE Access* 7, 84040–84052. doi: 10.1109/ACCESS.2019.2924744

Liu, C., Chen, L.-C., Schroff, F., Adam, H., Hua, W., Yuille, A. L., et al. (2019). "Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 82–92. doi: 10.1109/CVPR.2019.00017

Liu, H., Simonyan, K., and Yang, Y. (2018). DARTS: differentiable architecture search. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1806.09055

Liu, L., Hu, X., Zhu, L., Fu, C.-W., Qin, J., and Heng, P.-A. (2020). ψ-net: Stacking densely convolutional lstms for sub-cortical brain structure segmentation. *IEEE Trans. Med. Imaging*. 39:2806–2817 doi: 10.1109/TMI.2020.2975642

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440. doi: 10.1109/CVPR.2015.7298965

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). "Understanding the effective receptive field in deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 4898–4906. Available online at: https://arxiv.org/abs/1701.04128

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-Net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (IEEE), 565–571. doi: 10.1109/3DV.2016.79

Myronenko, A. (2018). "3D MRI brain tumor segmentation using autoencoder regularization," in *International MICCAI Brainlesion Workshop* (Springer), 11384, 311–320. doi: 10.1007/978-3-030-11726-9_28

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention U-Net: Learning where to look for the pancreas. *arXiv [preprint]*. Available online at: https://arxiv.org/abs/1804.03999

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 8026–8037. Available online at: https://arxiv.org/abs/1912.01703

Peng, C., Zhang, X., Yu, G., Luo, G., and Sun, J. (2017). "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4353–4361. doi: 10.1109/CVPR.2017.189

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/TMI.2016.2538465

Poudel, R. P., Lamata, P., and Montana, G. (2016). "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *Reconstruction, Segmentation, and Analysis of Medical Images*, eds M. A. Zuluaga, K. Bhatia, B. Kainz, M. H. Moghari, and D. F. Pace (Athens: Springer), 83–94. doi: 10.1007/978-3-319-52280-7_8

Robbins, H., and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* 22, 400–407. doi: 10.1214/aoms/1177729586

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 9351, 234–241. doi: 10.1007/978-3-319-24574-4_28

Roth, H. R., Shen, C., Oda, H., Sugino, T., Oda, M., Hayashi, Y., et al. (2018). "A multi-scale pyramid of 3D fully convolutional networks for abdominal multi-organ segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 11073, 417–425. doi: 10.1007/978-3-030-00937-3_48

Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017). Auto-context convolutional neural network (auto-net) for brain extraction in magnetic resonance imaging. *IEEE Trans. Med. Imaging* 36, 2319–2330. doi: 10.1109/TMI.2017.2721362

Salehi, S. S. M., Khan, S., Erdogmus, D., and Gholipour, A. (2018). Real-time deep pose estimation with geodesic loss for image-to-template rigid registration. *IEEE Trans. Med. Imaging* 38, 470–481. doi: 10.1109/TMI.2018.2866442

Seif, G., and Androutsos, D. (2018). "Large receptive field networks for high-scale image super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 763–772. doi: 10.1109/CVPRW.2018.00120

Seo, H., Huang, C., Bassenne, M., Xiao, R., and Xing, L. (2019). Modified U-Net (MU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in ct images. *IEEE Trans. Med. Imaging* 39, 1316–1325. doi: 10.1109/TMI.2019.2948320

Shen, X., Wang, C., Li, X., Yu, Z., Li, J., Wen, C., et al. (2019). "RF-Net: an end-to-end image matching network based on receptive field," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8132–8140. doi: 10.1109/CVPR.2019.00832

Song, T., Meng, F., Rodriguez-Paton, A., Li, P., Zheng, P., and Wang, X. (2019). U-next: a novel convolution neural network with an aggregation u-net architecture for gallstone segmentation in CT images. *IEEE Access* 7, 166823–166832. doi: 10.1109/ACCESS.2019.2953934

Sua, J. N., Lim, S. Y., Yulius, M. H., Su, X., Yapp, E. K. Y., Le, N. Q. K., et al. (2020). Incorporating convolutional neural networks and sequence graph transform for identifying multilabel protein lysine ptm sites. *Chemometr. Intell. Lab. Syst.* 206:104171. doi: 10.1016/j.chemolab.2020.104171

Tian, Z., Liu, L., Zhang, Z., and Fei, B. (2015). Superpixel-based segmentation for 3d prostate mr images. *IEEE Trans. Med. Imaging* 35, 791–801. doi: 10.1109/TMI.2015.2496296

Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* 5:180161. doi: 10.1038/sdata.2018.161

Wang, W., Yu, K., Hugonot, J., Fua, P., and Salzmann, M. (2019a). "Recurrent U-Net for resource-constrained segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2142–2151. doi: 10.1109/ICCV.2019.00223

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E. K., and Yuille, A. L. (2019b). Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Med. Image Anal.* 55, 88–102. doi: 10.1016/j.media.2019.04.005

Xiao, X., Lian, S., Luo, Z., and Li, S. (2018). "Weighted RES-UNet for high-quality retina vessel segmentation," in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)* (IEEE), 327–331. doi: 10.1109/ITME.2018.00080

Yap, M. H., Pons, G., Marti, J., Ganau, S., Sentis, M., Zwiggelaar, R., et al. (2017). Automated breast ultrasound lesions detection using convolutional neural networks. *IEEE J. Biomed. Health Inform.* 22, 1218–1226. doi: 10.1109/JBHI.2017.2731873

Zhang, Z., Liu, Q., and Wang, Y. (2018). Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett*. 15, 749–753. doi: 10.1109/LGRS.2018.2802944

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2020). UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* 39, 1856–1867. doi: 10.1109/TMI.2019.2959609

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. (2018). "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8697–8710. doi: 10.1109/CVPR.2018.00907

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 February 2021
doi: 10.3389/fgene.2020.609405

The Comprehensive Analyses of Genomic Variations and Assessment of TMB and PD-L1 Expression in Chinese Lung Adenosquamous Carcinoma

Yong Cheng[1†], *Yanxiang Zhang*[2†], *Yuwei Yuan*[2], *Jiao Wang*[1], *Ke Liu*[2], *Bin Yu*[1], *Li Xie*[1], *Chao Ou-Yang*[1], *Lin Wu*[2*] *and Xiaoqun Ye*[1*]

[1] *Department of Respiratory Diseases, The Second Affiliated Hospital of Nanchang University, Nanchang, China,* [2] *Berry Oncology Corporation, Beijing, China*

OPEN ACCESS

Edited by:
Cheng Guo,
Columbia University, United States

Reviewed by:
Tianbao Li,
Geneis (Beijing) Co. Ltd, China
Junlin Xu,
Hunan University, China

***Correspondence:**
Xiaoqun Ye
511201663@qq.com
Lin Wu
wulin@berryoncology.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: *23 September 2020*
Accepted: *18 December 2020*
Published: *17 February 2021*

Citation:
Cheng Y, Zhang YX, Yuan YW, Wang J, Liu K, Yu B, Xie L, Ou-Yang C, Wu L and Ye XQ (2021) The Comprehensive Analyses of Genomic Variations and Assessment of TMB and PD-L1 Expression in Chinese Lung Adenosquamous Carcinoma. Front. Genet. 11:609405. doi: 10.3389/fgene.2020.609405

The poor prognosis and fewer treatment option is a current clinical challenge for patients with lung adenosquamous carcinoma (ASC). The previous studies reported that tumor mutational burden (TMB, numbers of mutation per Megabase) is a predictor of clinical response in trials of multiple cancer types, while fewer studies assessed the relationship between TMB level and clinical features and outcomes of lung ASC. Herein, the present study enrolled Chinese patients with lung ASC. DNA was extracted from formalin-fixed paraffin-embedded tumor samples and subjected to next generation sequencing (NGS), and the 457 cancer related genes were evaluated. The results demonstrated that 95 unique genes with somatic variations were identified in the enrolled patients. The top three of high frequency gene mutations were *TP53*, *EGFR*, *PIK3CA* with rates of 62% (13 cases), 48% (10 cases), and 14% (3 cases), respectively. We identified TMB value was significantly correlated with pathological stages ($p < 0.05$) and invasion of lymph node ($p < 0.05$). However, TMB value was not significantly correlated to other clinicopathologic indexes, for examples, age, sex, smoking history, tumor size, as well as *TP53* and *EGFR* mutations in lung ASC. Moreover, TMB value was associated with the overall survival ($p < 0.01$), but not with the relapse-free survival ($p = 0.23$). In conclusion, this study indicated that lung ASC with high TMB might be associated with the invasion of lymph node and short overall survival. Immunotherapy might be a promising treatment option for lung ASC patients with high TMB.

Keywords: adenosquamous carcinoma, EGFR, lung, next generation sequencing, PD-L1, somatic variations, TMB

INTRODUCTION

Worldwide, lung cancer is the most prevalent cause of cancer related death (Siegel et al., 2018). Adenosquamous carcinoma (ASC) is a small subtype of non-small-cell lung cancer (NSCLC), accounting for <4% of all patients with NSCLC (Uramoto et al., 2010; Li and Lu, 2018). It is defined that lung ASC is a mixed-type tumor, comprised of adenocarcinoma and squamous cell carcinoma. Each component accounts for at least 10% of the total tumor cells, according to the tumor classification by the fifth edition of world health organization (WHO) (Travis et al., 2015).

It is reported that lung ASC displays the worse prognosis than other types of NSCLC. Lung ASC is resistant to the treatment of adjuvant chemotherapy, and more probably to occur local recurrence or distant metastasis in comparison with other histologic types of NSCLC (Hsia et al., 1999; Nakagawa et al., 2003; Maeda et al., 2012).

In recent years, the important advancements have been achieved in NSCLC treatments (Herbst et al., 2018; Testa et al., 2018). For example, the small molecule tyrosine kinase inhibitors (TKIs) were effective for patients with advanced lung adenocarcinoma with the somatic mutation of *epidermal growth factor receptor* (*EGFR*) and the rearrangement of *echinoderm microtubule-associated protein-like 4* (*EML4*) with *anaplastic lymphoma kinase* (*ALK*) (Paez et al., 2004; Soda et al., 2007; Robichaux et al., 2018; Ramalingam et al., 2020). Interestingly, a few case reports and retrospective studies have demonstrated that EGFR-TKIs therapies were effective for the selected patients with advanced ASC of the lung (Song et al., 2013; Kurishima et al., 2014; Fan et al., 2017; Zhang et al., 2018; Lin et al., 2020). Therefore, besides of *EGFR* mutation, the continued research is required to identify more cancer-related gene mutations and the corresponding targeted agents or combined therapies to improve outcomes for lung ASC.

Immune checkpoint inhibitor (ICI) therapies have shown significant benefit in treatment of patients with NSCLC (Herbst et al., 2018), for example, pembrolizumab treatment achieved better clinical outcomes compared to platinum-based chemotherapy in advanced NSCLC patients with high expression of programmed death ligand 1 (PD-L1) in tumor cells (Herbst et al., 2016; Reck et al., 2016). Besides of programmed death 1 (PD-1) and its ligand PD-L1, the resent studies indicated that tumor mutational burden (TMB) could predict clinical outcomes in multiple cancer types, including lung cancer patients receiving immunotherapy (Rizvi et al., 2015, 2018; Samstein et al., 2019). However, there is still lack of prospective data and retrospective study to comprehensively depict the genomic landscape and immune biomarkers, as well as their association with the clinicopathologic features in patients with lung ASC.

To address the limited knowledge, we performed this study in patients with surgically resected lung ASC to evaluate (1) the genomic variations and its correlation with TMB and PD-L1 expression and (2) the clinical relevance of TMB and PD-L1 expression, including clinicopathologic features, relapse-free survival (RFS), and overall survival (OS). Meanwhile, we compared the data of lung ASC to other ethnicities as well as other subtypes such as adenocarcinoma and squamous cell carcinoma.

MATERIALS AND METHODS

Patients and Samples

All the enrolled patients with lung adenosquamous carcinoma (ASC) underwent surgical resection from the Second Affiliated Hospital of Nanchang University between April, 2014 and May, 2019. The criteria of the enrolled patients were as follows: (1) pathological diagnosis of lung ASC according to the tumor classification in the fifth edition of WHO, each component of adenocarcinoma and squamous cell carcinoma at least 10% of the tumor cells; (2) patients without anticancer treatment before surgery; (3) availability of complete medical records, including patient's age, gender, smoking history, immunohistochemistry results, pathological reports, operation time and surgical approach, medication records, tumor response assessment. All the enrolled patients accepted and signed the informed consent, the protocol was approved by the Ethics Committee of medical research, the Second Affiliated Hospital of Nanchang University.

FFPE Preparation and Genomic DNA Extraction

After surgical resection, tumor tissues and normal tissues (incision margin 5 cm away from the tumor) were fixed with formalin, subsequently embedded in paraffin (FFPE). Genomic DNA was extracted from each FFPE sample using the GeneRead DNA FFPE Kit (Qiagen, USA) according to the manufacturer's protocol, respectively.

NGS Based Large-Gene Panel Test

To construct the pre-library, genomic DNA was digested into ∼200 bp fragments by enzymatic method, then subjected to end repairing, A-tailing, adapter ligation and universal amplification. Purified pre-library was hybridized with a customized biotin probe pool (the 457 genes panel, Berry Oncology, Peking, China) to capture target fragments (**Supplementary Table 1**). Captured fragments were amplified with universal primers and purified to acquire the final library. The library of paired-end multiplex samples were sequenced with the NovaSeq 6000 System. Sequencing depth was ∼2,000 x per sample.

The generated sequences were trimmed, low-quality-filtered, and subjected for variant calling. Variants was filtered for nonsynonymous SNPs, indels and spliced variants. Somatic variations were identified with variant allele frequency (cutoff $\geq$ 3%) and cancer hotspots were screened with variant allele frequency (cutoff $\geq$ 1%) and at least 20 high-quality reads.

The tumor mutation burden (TMB) was determined by the number of all the nonsynonymous mutation and indel variants per magabase of coding regions. The 457 gene panels cover the coding region of 1,141,951 bp. Hence, TMB was calculated with the number of all the nonsynonymous mutations and indel variants/1.14 Mb. The threshold of high TMB was set to 10 according to the previous studies (Hellmann et al., 2018; Barroso-Sousa et al., 2020).

Immunohistochemistry Assays

Immunohistochemistry assays were performed on FFPE sections using a primary anti-PD-L1 (SP263) rabbit monoclonal antibody (Roche) and a secondary anti-rabbit-IgG antibody (ZSGB-BIO, Beijing, China), then detected with DBA detection kit (ZSGB-BIO, Beijing, China). PD-L1-positive was determined if membrane staining exhibited in $\geq$ 25% of tumor cells in the tumor sample, as described in the previous study (Shi et al., 2017).

Statistics

R Foundation for Statistics Computing, R script (v3.6.0) was used to perform the statistics analysis. Fisher's exact test was

TABLE 1 | The overview of clinical information of enrolled patients.

Patients ($n = 21$)	
Age	
Median	63
Range	49-75
Sex (%)	
Male	11 (52.4%)
Female	10 (47.6%)
Smoking history (%)	
Current	7 (33.3%)
Former	0
Never	14 (66.7%)
Pathological stage (%)	
I	9 (42.9%)
II	3 (14.3)
III	9 (42.9%)
Histologic type	
Adeno cells accounted for 10–50%	10
Adeno cells accounted for 51–90%	11

used to analyze the relationship between TMB and clinical indexes. Kaplane-Meier method was used to estimate progress-free survival and overall survival. $p < 0.05$ was defined as statistically significant.

RESULTS

Clinicopathological Characteristics of the Enrolled Patients

In total, 21 patients with lung adenosquamous carcinoma (ASC) were finally enrolled in this study. The median age at diagnosis was 63 years (range: 49–75), eleven were male and ten were female, seven were smokers and 14 were non-smokers. The tumors were stage I, II, and III in nine (42.9%), three (14.3%), and nine (42.9%) cases, respectively. Ten patients were with adeno cells accounted for 10%-50% and eleven patients were with adeno cells accounted for 51–90%. The clinical information of patients was overviewed in **Table 1**. The characteristics of each patient were listed in **Supplementary Table 2**.

Somatic Variation Detection

To discover somatic variation in ASC, DNA was extracted from formalin-fixed paraffin-embedded samples and subjected to NGS based large-gene panel test. The somatic mutations of each tumor sample were analyzed and summarized in **Figure 1** and **Supplemental Table 3**. Our study identified 95 unique genes with somatic variations. Among those, the top three of high frequency gene mutations were *TP53*, *EGFR*, *PIK3CA* with rates of 61.9% (13 cases), 47.6% (10 cases), and 14.3% (3 cases), respectively. Coexisting mutations were detected in *TP53* and *EGFR* with rate of 33.3% (7 cases), *EGFR,* and *PIK3CA* with rate of 4.8% (1 case).

Mutations of the *TP53* gene are universal in lung cancer, with mutation rate of about 50% in NSCLC (Mogi and Kuwano, 2011). In lung ASC, our study indicated *TP53* was a highest frequency mutation gene, with a mutation rate of 62%. The common mutation was detected in exon 8 (5 cases), exon 7 (4 cases), exon 6 (2 cases), and exon 4 (1 case). The hotspot mutations of TP53 p.R248Q and TP53 p.R248W were detected in p11 and p19, respectively, which are the target of APR-246 drug. At present, FDA has approved APR-246 in combination therapies to treat myelodysplastic syndromes with TP53 mutation, while more clinical trials are required to prove the efficacy of the drug in treating patients with lung cancer.

For *EGFR* gene, deletion in exon 19 was the most common mutation (7 cases), and the single nucleotide variation in exon 21 resulting in EGFR p.L858R variant was observed only in one case. These mutations are related to the increase of sensitivity to tyrosine kinase inhibitors. The drug resistant mutation was detected in one case harboring insertion mutation in exon 20, while none of T790M variant was observed in this study. Two variants of in-frame deletion in exon 19 and single nucleotide variation resulting in EGFR p.E758D were coexisted in one case, as shown in **Figure 1** and **Supplemental Table 3**.

In addition to *EGFR* mutations, a set of genes involved in the *PI3K* signaling pathway were observed in seven lung ACS cases in our study. Two cases (p2 and p16) harbored a single nucleotide variation in exon 9 of *PIK3CA*, resulting in a p.E545K variant in the helical region, and one case (p14) harbored a mutation in exon 21 of *PIK3CA* which generated a p.H1047L variant in p110α catalytic subunit. The variant of PTEN p.H123Y was present in one case (p13). These variants were sensitive to class I PI3K inhibitor. We also found PIK3C2B p.E545K variant in one case (p9), and PIK3C2G p.M1047I variant in another case (p3), both variants belong to class II PI3K. In addition, gene mutations were observed in *CDKN2A* in two cases (p18 and p21), *NF1* in two cases (p6 and p20), *DDR2* in two cases (p8 and p13), *PBRM1* in two cases (p9 and p13), *WHSC1L1* in one case (p10), *IRF4* in one case (p16), and *HRAS* in one case (p21).

The rearrangement *of anaplastic lymphoma kinase (ALK)*, *c-ros oncogene 1, receptor tyrosine kinase* (*ROS1*), and *ret proto-oncogene* (*RET*) play a role on driving the occurrence and development of NSCLC (Takeuchi et al., 2012; Cancer Genome Atlas Research, 2014). These gene translocations were detected by NGS based large-gene panel test and RNA amplification assays in the present study. However, none of *ALK*, *ROS1*, and *RET* rearrangements were detected in the enrolled patients with lung ASC (data not shown).

Detection of Somatic Copy Number Alterations in Lung ASC

The somatic copy number alterations (CNA) play an important role in the development of lung cancer. In this study, CNA was detected in the enrolled patients with lung ASC. The results indicated that the amplification of six genes in tumor tissues were at least twice than in the normal tissues (**Figure 2** and **Table 2**). Among those, the amplification of *cyclin dependent kinase 4* (*CDK4*) was observed in two cases (p7 and p12). Meanwhile, the coexisting amplification of *cyclin D1* (*CCND1*) and *CDK4* were detected in one case (p12). CDK4 and CCND1 are very important

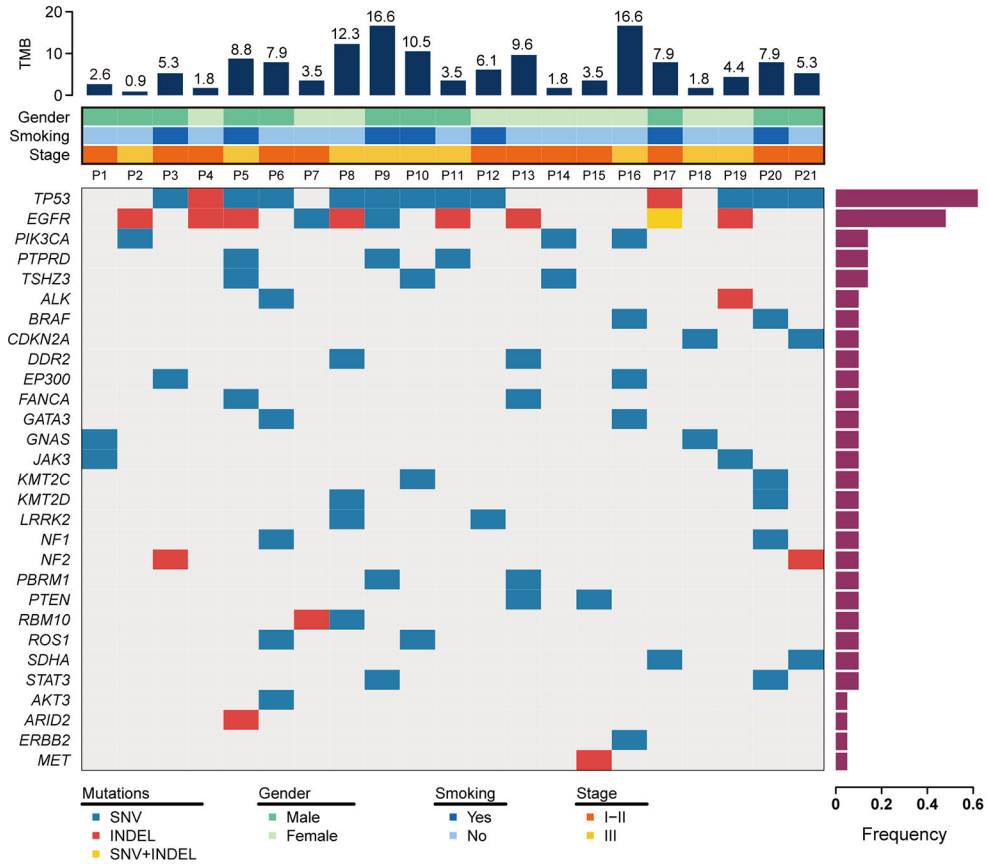

FIGURE 1 | Somatic mutation landscape for 21 patients with lung ASC. The top 30 mutant genes are displayed. The frequency of mutated genes are shown in a histogram on the right. Tumor mutation burden (TMB) and clinical features are illustrated in the upper two panels. SNV+INDEL represents the case carrying both of mutation types of SNV and INDEL.

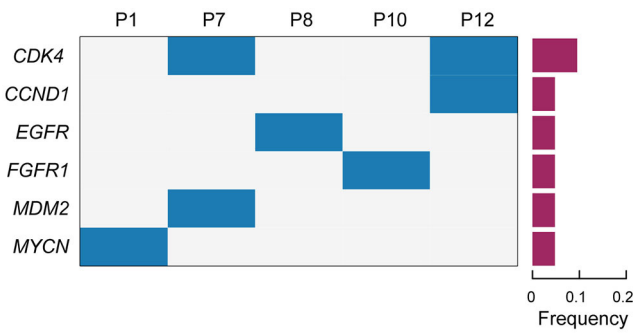

FIGURE 2 | The profiles of somatic copy number alterations for the enrolled patients. Cases with somatic copy number alterations (CNA) are shown. Patient numbers are displayed in the upper lane. Gene names are listed on the left. The frequency of CNA are shown in a histogram on the right.

components involved in regulating the progress of G1/S phase in cell cycle. The complex of CKD4 and CCND1 has been studied as a therapeutic target for cancer (Malumbres and Barbacid, 2009; Musgrove et al., 2011). In addition, the tyrosine kinase receptors *EGFR* and *FGFR1* were amplified for eleven times in p8 and p10

TABLE 2 | Copy number alterations of enrolled patients detected by targeted NGS panel.

Patient ID	CCND1	CDK4	EGFR	FGFR1	MDM2	MYCN
P1						6.1
P7		6.9			6.1	
P8			11.6			
P10				11.9		
P12	6.9	4				

cases, respectively. The present study also indicated p1 and p7 cases harboring *MDM2* and *MYCN* with copy number gains, respectively. Overexpression or amplification of *MDM2* occurs in a variety of cancer types.

Association of Tumor Mutations Burden With Clinicopathological Characteristics

Immune checkpoint inhibitor (ICI) therapies have earned its spurs in the treatment of malignant tumors in recent years (Gandhi et al., 2018). Tumor mutation burden (TMB) is a promising marker to predict survival after immunotherapy

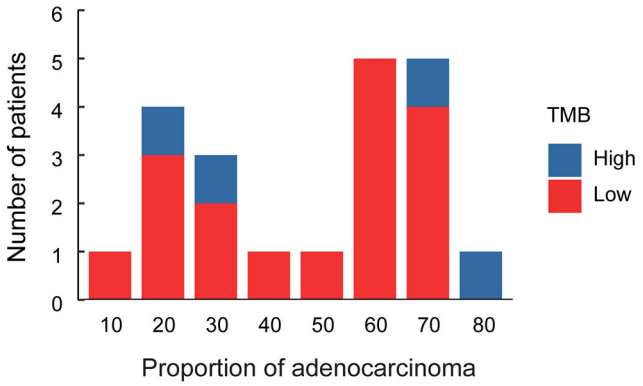

FIGURE 3 | Histogram of proportion of adenocarcinoma cells in lung ASC. Red column represents patients with the low level of TMB, blue column represents patients with the high level of TMB.

TABLE 3 | Relationship between TMB index and clinicopathological characteristics of lung ASC.

Characteristics	Numbers of patients (percentage)	TMB (cutoff = 10)		*p* value
		Low	High	
Age				0.58
< 70	15 (71.4%)	11	4	
≥ 70	6 (28.6%)	4	2	
Sex				0.63
Male	11 (52.4%)	8	3	
Female	10 (47.6%)	7	3	
Smoking history				0.30
Smokers	7 (33.3%)	4	3	
Non-smokers	14 (66.7%)	11	3	
Pathological stage				0.03
I-II	12 (57.1%)	11	1	
III	9 (42.9%)	4	5	
Tumor size (cm)				0.31
≤ 3	7 (33.3%)	6	1	
> 3	14 (66.7%)	9	5	
Lymph node				0.03
N0	12 (57.1%)	11	1	
N1-N2	9 (42.9%)	4	5	
TP53 mutation				0.59
w/t	8 (38.1%)	6	2	
With	13 (61.9%)	9	4	
EGFR mutation				0.27
w/t	11 (52.4%)	9	2	
With	10 (47.6%)	6	4	

across multiple cancer types (Samstein et al., 2019). In our study, TMB value was determined in the enrolled patients. The results indicated that the median TMB was 5.25 mutations per megabases, with a range from 0.88 to 16.64 (**Figure 1**). We analyzed the association between TMB value and the proportion of adeno and squamous cells carcinoma of ASC (**Figure 3**). The results indicated that the high level of TMB was not significantly related to the high proportion of squamous cells in ASC of the lung. Meanwhile, the relationships between TMB level and the clinicopathologic features of adenosquamous cell carcinoma of the lung were analyzed. The results demonstrated that TMB value correlated significantly with pathological stages ($p = 0.03$) and lymph node ($p = 0.03$). The higher TMB value (cutoff ≥ 10 mut/Mb) was related to invasion of lymph node, while the lower TMB value (cutoff < 10 mut/Mb) was related to none of invasion of lymph node. However, there was no significant relationship between TMB and other clinicopathologic indexes, for instances, age, sex, smoking history, as well as *TP53* and *EGFR* mutations in lung ASC (**Table 3**). It was no distant metastasis in the enrolled patients. Therefore, we did not analyze the relationship between TMB level and the index of distant metastasis.

Relationships Between TMB and Clinical Outcomes

In this study, the enrolled patients underwent operation to completely resect the primary tumor tissues and accomplish the lymphadenectomy. In total, five (23.8 %) of 21 patients did not occur disease recurrence, 14 (66.7%) of 21 patients had local recurrence or distant metastasis, two (9.5) of 21 patients were out of contact. The Kaplan–Meier survival curves of relapse-free survival, and overall survival were displayed in **Figure 4**, with the median of 6 and 10 months, respectively. Among the enrolled patients, one case (p8) harboring EGFR exon 19del and amplification of EGFR, received the EGFR-TKI therapy. The patient developed brain metastases after 14 months of treatment with gefitinib. None of the enrolled patients treated with immune checkpoint inhibitor therapies.

We analyzed the relationship between TMB and survival time. The results demonstrated that TMB value was significantly correlated with the overall survival ($p = 0.0078$, **Figure 4A**), but not with the relapse-free survival ($p = 0.23$, **Figure 4B**). As shown in **Figure 4**, the high level of TMB was related to the short survival time. Therefore, immunotherapy might be a promising treatment option to improve the outcomes in lung ASC patients with high TMB.

DISCUSSION

The poor prognosis and fewer treatment option is still a clinical challenge for lung ASC. So far, a few studies introduced the mutational profile of lung ASC (Sasaki et al., 2007; Tochigi et al., 2011; Morodomi et al., 2015; Vassella et al., 2015; Shi et al., 2016; Lin et al., 2020), while most analyses were restricted to small gene panels. The continued studies are required to investigate genetic alterations and explore the potential therapies for lung ASC. The present study displayed the comprehensive analyses of somatic variations in lung ASC. In addition, it is the first study to reveal the clinical relevance of TMB level and PD-L1 expression in lung ASC.

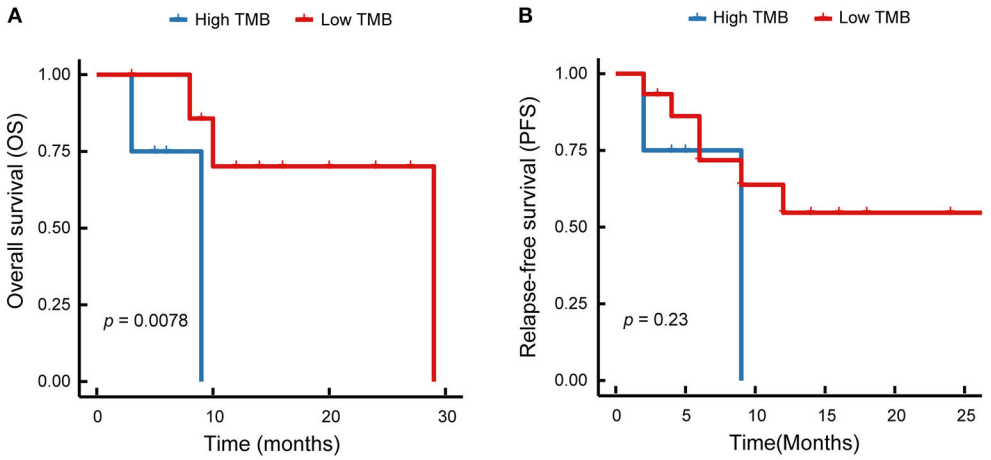

FIGURE 4 | Kaplan–Meier survival curves of patients with lung ASC. **(A)** Overall survival of lung ASCs with high TMB and low TMB are shown. **(B)** Relapse-free survival of lung ASCs with high TMB and low TMB are displayed. Red solid line represents patients with low TMB, blue solid line represents patients with high TMB.

Our study showed a high frequency of *EGFR* mutations in lung ASC, the mutation rate was 48%. However, in contrast to our observation, a lower prevalence of *EGFR* mutations was reported in lung ASC of Caucasian ethnic group, with a mutation rate of 13% (Tochigi et al., 2011). That might be due to the ethnicity differences between Asians and Caucasians. Moreover, consistent with the incidence of *EGFR* mutation in lung adenocarcinoma, the mutation rate was 46.7% in the Asian population and 15% in the white population (Liu et al., 2017). Furthermore, the current study revealed that the landscape of somatic variations of ASC was similar to that of lung adenocarcinomas, and supported the hypothesis that adenocarcinoma components and squamous cell carcinoma components of ASC shared a monoclonal origin (Lin et al., 2020). Therefore, considering the similar profile of somatic variations in ASC and lung adenocarcinoma, TKI might be an effective targeted agents for lung ASC with EGFR mutations. In our study, one resectable patient (pT2aN2M0) received EGFR-TKI therapy after four cycles of adjuvant chemotherapy, and had a clinical benefit from the treatment of gefitinib, with progress-free survival (PFS) of 14 months. In line with the observation, a current multicenter retrospective study also indicated that EGFR-TKIs were effective for patients with advanced ASC of lung, with the median PSF being 10.1 months (Lin et al., 2020). None of *ALK*, *ROS1*, and *RET* rearrangements were detected in our study. In line with our results, the previous study also did not find gene rearrangements in Caucasian patients with lung ASC (Vassella et al., 2015). That might be due to the lower prevalence of gene translocations in lung cancers and the small size of enrolled patients with lung ASC.

In addition, ICI therapies have been applied in treatment of malignant tumors in recent years. We valuated PD-L1 expression in tumor cells of the enrolled patients, while there were no significant associations between PD-L1 expression and the clinicopathologic features of lung ASC (**Supplemental Table 4**). Besides of PD-L1, TMB is a promising marker to predict clinical outcomes of patients with NSCLC to immunotherapy (Carbone et al., 2017; Hellmann et al., 2018; Samstein et al., 2019). The previous studies indicated lung squamous cell carcinoma harboring higher TMB than other solid cancer types (Vogelstein et al., 2013; Zhang et al., 2019). However, our results indicated that the high level of TMB was not significantly related to the high proportion of squamous cells in lung ASC. The current study displayed that TMB was lower in adenocarcinoma component than in squamous cell carcinoma component, with the median of 6.5 and 7.2 mutations/Mb, respectively (Lin et al., 2020). However, it is difficult to distinguish such small differences of TMB value in adenocarcinoma component and squamous cell carcinoma component.

In the present study, we also evaluated the relationships between TMB level and the clinicopathologic features and outcomes of lung ASC, though none of enrolled patients received ICI therapy. Our results indicated that the high level of TMB was related to the invasion of lymph node and the short survival time. Patients with the short survival time might be due to the invasion of lymph node. In line with the results, the previous study indicated that high TMB is a poor prognostic factor for the advanced NSCLC, as well as patient in early stage (Owada-Ozaki et al., 2018). However, as the limited numbers of enrolled patients, we did not obtain lung ASC with high TMB in early stage.

In conclusion, the lung ASC with high TMB might be associated with invasion of lymph node and short overall survival. Therefore, immunotherapy might be a potential treatment option for lung ASC patients with the high level of TMB.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The accession number is CNP0001586,

the link is followed as: https://db.cngb.org/search/project/CNP0001586/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Medical Research, the Second Affiliated Hospital of Nanchang University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

XY and LW designed the project. YZ, YY, KL and YC performed the work. YZ wrote the paper. JW, BY, LX and CO-Y reviewed the paper. All authors accepted the final version of this manuscript for publication.

FUNDING

XY is supported by National Nature Science Foundation of China (No. 81660493) and Natural Science Foundation of Jiang Xi Province (No. 20171BAB205053).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.609405/full#supplementary-material

REFERENCES

Barroso-Sousa, R., Jain, E., Cohen, O., Kim, D., Buendia-Buendia, J., Winer, E., et al. (2020). Prevalence and mutational determinants of high tumor mutation burden in breast cancer. *Ann. Oncol* 31, 387–394. doi: 10.1016/j.annonc.2019.11.010

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

Carbone, D. P., Reck, M., Paz-Ares, L., Creelan, B., Horn, L., Steins, M., et al. (2017). First-line nivolumab in stage IV or recurrent non-small-cell lung cancer. *N. Engl. J. Med* 376, 2415–2426. doi: 10.1056/NEJMoa1613493

Fan, L., Yang, H., Yao, F., Zhao, Y., Gu, H., Han, K., et al. (2017). Clinical outcomes of epidermal growth factor receptor tyrosine kinase inhibitors in recurrent adenosquamous carcinoma of the lung after resection. *Onco. Targets. Ther.* 10, 239–245. doi: 10.2147/OTT.S114451

Gandhi, L., Rodriguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., et al. (2018) Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N. Engl. J. Med* 378, 2078–2092. doi: 10.1056/NEJMoa1801005

Hellmann, M. D., Ciuleanu, T. E., Pluzanski, A., Lee, J. S., Otterson, G. A., Audigier-Valette, C., et al. (2018). Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden. *N. Engl. J. Med.* 378, 2093–2104. doi: 10.1056/NEJMoa1801946

Herbst, R. S., Baas, P., Kim, D. W., Felip, E., Perez-Gracia, J. L., Han, J. Y., et al. (2016). Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* 387, 1540–1550. doi: 10.1016/S0140-6736(15)01281-7

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature* 553, 446–454. doi: 10.1038/nature25183

Hsia, J. Y., Chen, C. Y., Hsu, C. P., Shai, S. E., and Wang, P. Y. (1999). Adenosquamous carcinoma of the lung. Surgical results compared with squamous cell and adenocarcinoma. *Scand. Cardiovasc. J.* 33, 29–32. doi: 10.1080/14017439950142000

Kurishima, K., Ohara, G., Kagohashi, K., Watanabe, H., Takayashiki, N., Ishibashi, A., et al. (2014). Adenosquamous cell lung cancer successfully treated with gefitinib: a case report. *Mol. Clin. Oncol.* 2, 282–284. doi: 10.3892/mco.2013.221

Li, C., and Lu, H. (2018). Adenosquamous carcinoma of the lung. *Onco. Targets. Ther.* 11, 4829–4835. doi: 10.2147/OTT.S164574

Lin, G., Li, C., Li, P. S., Fang, W. Z., Xu, H. P., Gong, Y. H., et al. (2020). Genomic origin and EGFR-TKI treatments of pulmonary adenosquamous carcinoma. *Ann. Oncol.* 31, 517–524. doi: 10.1016/j.annonc.2020.01.014

Liu, L., Liu, J., Shao, D., Deng, Q., Tang, H., Liu, Z., et al. (2017). Comprehensive genomic profiling of lung cancer using a validated panel to explore therapeutic targets in East Asian patients. *Cancer Sci.* 108, 2487–2494. doi: 10.1111/cas.13410

Maeda, H., Matsumura, A., Kawabata, T., Suito, T., Kawashima, O., Watanabe, T., et al. (2012). Adenosquamous carcinoma of the lung: surgical results as compared with squamous cell and adenocarcinoma cases. *Eur. J. Cardiothorac. Surg.* 41, 357–361. doi: 10.1016/j.ejcts.2011.05.050

Malumbres, M., and Barbacid, M. (2009). Cell cycle, CDKs and cancer: a changing paradigm. *Nat. Rev. Cancer* 9, 153–166. doi: 10.1038/nrc2602

Mogi, A., and Kuwano, H. (2011). TP53 mutations in nonsmall cell lung cancer. *J. Biomed. Biotechnol.* 2011:583929. doi: 10.1155/2011/583929

Morodomi, Y., Okamoto, T., Takenoyama, M., Takada, K., Katsura, M., Suzuki, Y., et al. (2015). Clinical significance of detecting somatic gene mutations in surgically resected adenosquamous cell carcinoma of the lung in Japanese patients. *Ann. Surg. Oncol.* 22, 2593–2598. doi: 10.1245/s10434-014-4218-0

Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A., and Sutherland, R. L. (2011). Cyclin D as a therapeutic target in cancer. *Nat. Rev. Cancer* 11, 558–572. doi: 10.1038/nrc3090

Nakagawa, K., Yasumitu, T., Fukuhara, K., Shiono, H., and Kikui, M. (2003). Poor prognosis after lung resection for patients with adenosquamous carcinoma of the lung. *Ann. Thorac. Surg.* 75, 1740–1744. doi: 10.1016/s0003-4975(03)00022-5

Owada-Ozaki, Y., Muto, S., Takagi, H., Inoue, T., Watanabe, Y., Fukuhara, M., et al. (2018). Prognostic impact of tumor mutation burden in patients with completely resected non-small cell lung cancer: brief report. *J. Thorac. Oncol.* 13, 1217–1221. doi: 10.1016/j.jtho.2018.04.003

Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., et al. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* 304, 1497–1500. doi: 10.1126/science.1099314

Ramalingam, S. S., Vansteenkiste, J., Planchard, D., Cho, B. C., Gray, J. E., Ohe, Y., et al. (2020). Overall survival with osimertinib in untreated, EGFR-mutated advanced NSCLC. *N. Engl. J. Med* 382, 41–50. doi: 10.1056/NEJMoa1913662

Reck, M., Rodriguez-Abreu, D., Robinson, A. G., Hui, R., Csoszi, T., Fulop, A., et al. (2016). Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N. Engl. J. Med* 375, 1823–1833. doi: 10.1056/NEJMoa1606774

Rizvi, H., Sanchez-Vega, F., La, K., Chatila, W., Jonsson, P., Halpenny, D., et al. (2018). Molecular determinants of response to anti-programmed cell death (PD)-1 and anti-programmed death-ligand 1 (PD-L1) blockade in patients with non-small-cell lung cancer profiled with targeted next-generation sequencing. *J. Clin. Oncol.* 36, 633–641. doi: 10.1200/JCO.2017.75.3384

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

Robichaux, J. P., Elamin, Y. Y., Tan, Z., Carter, B. W., Zhang, S., Liu, S., et al. (2018). Mechanisms and clinical activity of an EGFR and HER2 exon 20-selective kinase inhibitor in non-small cell lung cancer. *Nat. Med.* 24, 638–646. doi: 10.1038/s41591-018-0007-9

Samstein, R. M., Lee, C. H., Shoushtari, A. N., Hellmann, M. D., Shen, R., Janjigian, Y. Y., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* 51, 202–206. doi: 10.1038/s41588-018-0312-8

Sasaki, H., Endo, K., Yukiue, H., Kobayashi, Y., Yano, M., and Fujii, Y. (2007). Mutation of epidermal growth factor receptor gene in adenosquamous carcinoma of the lung. *Lung Cancer* 55, 129–130. doi: 10.1016/j.lungcan.2006.09.003

Shi, X., Wu, H., Lu, J., Duan, H., Liu, X., and Liang, Z. (2016). Screening for major driver oncogene alterations in adenosquamous lung carcinoma using PCR coupled with next-generation and Sanger sequencing methods. *Sci. Rep.* 6:22297. doi: 10.1038/srep22297

Shi, X., Wu, S., Sun, J., Liu, Y., Zeng, X., and Liang, Z. (2017). PD-L1 expression in lung adenosquamous carcinomas compared with the more common variants of non-small cell lung cancer. *Sci. Rep.* 7:46209. doi: 10.1038/srep46209

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

Soda, M., Choi, Y. L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448, 561–566. doi: 10.1038/nature05945

Song, Z., Lin, B., Shao, L., and Zhang, Y. (2013). Therapeutic efficacy of gefitinib and erlotinib in patients with advanced lung adenosquamous carcinoma. *J. Chin. Med. Assoc.* 76, 481–485. doi: 10.1016/j.jcma.2013.05.007

Takeuchi, K., Soda, M., Togashi, Y., Suzuki, R., Sakata, S., Hatano, S., et al. (2012). RET, ROS1 and ALK fusions in lung cancer. *Nat. Med.* 18, 378–381. doi: 10.1038/nm.2658

Testa, U., Castelli, G., and Pelosi, E. (2018). Lung cancers: molecular characterization, clonal heterogeneity and evolution, and cancer stem cells. *Cancers* 10:248. doi: 10.3390/cancers10080248

Tochigi, N., Dacic, S., Nikiforova, M., Cieply, K. M., and Yousem, S. A. (2011). Adenosquamous carcinoma of the lung: a microdissection study of KRAS and EGFR mutational and amplification status in a western patient population. *Am. J. Clin. Pathol.* 135, 783–789. doi: 10.1309/AJCP08IQZAOGYLFL

Travis, W. D., Brambilla, E., Burke, A. P., Marx, A., and Nicholson, A. G. (2015). Introduction to the 2015 world health organization classification of tumors of the lung, pleura, thymus, and heart. *J. Thorac. Oncol.* 10, 1240–1242. doi: 10.1097/JTO.0000000000000663

Uramoto, H., Yamada, S., and Hanagiri, T. (2010). Clinicopathological characteristics of resected adenosquamous cell carcinoma of the lung: risk of coexistent double cancer. *J. Cardiothorac. Surg.* 5:92. doi: 10.1186/1749-8090-5-92

Vassella, E., Langsch, S., Dettmer, M. S., Schlup, C., Neuenschwander, M., Frattini, M., et al. (2015). Molecular profiling of lung adenosquamous carcinoma: hybrid or genuine type? *Oncotarget* 6, 23905–23916. doi: 10.18632/oncotarget.4163

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122

Zhang, C., Yang, H., Lang, B., Yu, X., Xiao, P., Zhang, D., et al. (2018). Surgical significance and efficacy of epidermal growth factor receptor tyrosine kinase inhibitors in patients with primary lung adenosquamous carcinoma. *Cancer Manag. Res.* 10, 2401–2407. doi: 10.2147/CMAR.S165660

Zhang, X. C., Wang, J., Shao, G. G., Wang, Q., Qu, X., Wang, B., et al. (2019). Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients. *Nat. Commun.* 10:1772. doi: 10.1038/s41467-019-09762-1

Conflict of Interest: YY, LW, and YZ are currently employed by Berry Oncology Corporation.

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: 24 February 2021
doi: 10.3389/fgene.2021.635863

Identification and Validation of a Novel DNA Damage and DNA Repair Related Genes Based Signature for Colon Cancer Prognosis

Xue-quan Wang[1,2], Shi-wen Xu[3], Wei Wang[3], Song-zhe Piao[4], Xin-li Mao[2,5], Xian-bin Zhou[2,5], Yi Wang[2,5], Wei-dan Wu[2,5], Li-ping Ye[3,5] and Shao-wei Li[2,5]**

[1]Laboratory of Cellular and Molecular Radiation Oncology, Department of Radiation Oncology, Radiation Oncology Institute of Enze Medical Health Academy, Affiliated Taizhou Hospital of Wenzhou Medical University, Taizhou, China, [2]Key Laboratory of Minimally Invasive Techniques & Rapid Rehabilitation of Digestive System Tumor of Zhejiang Province, Linhai, China, [3]Wenzhou Medical University, Wenzhou, China, [4]Department of Urology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China, [5]Department of Gastroenterology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Zhi Huang,
Purdue University, United States
Haoyun Lei,
Carnegie Mellon University, United States
Liuyi Hao,
University of North Carolina at Greensboro, United States

***Correspondence:**
Li-ping Ye
yelp@enzemed.com
Shao-wei Li
li_shaowei81@hotmail.com

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

Received: 30 November 2020
Accepted: 01 February 2021
Published: 24 February 2021

Citation:
Wang X-q, Xu S-w, Wang W, Piao S-z, Mao X-l, Zhou X-b, Wang Y, Wu W-d, Ye L-p and Li S-w (2021) Identification and Validation of a Novel DNA Damage and DNA Repair Related Genes Based Signature for Colon Cancer Prognosis. Front. Genet. 12:635863. doi: 10.3389/fgene.2021.635863

Backgrounds: Colorectal cancer (CRC) with high incidence, has the third highest mortality of tumors. DNA damage and repair influence a variety of tumors. However, the role of these genes in colon cancer prognosis has been less systematically investigated. Here, we aim to establish a corresponding prognostic signature providing new therapeutic opportunities for CRC.

Method: After related genes were collected from GSEA, univariate Cox regression was performed to evaluate each gene's prognostic relevance through the TCGA-COAD dataset. Stepwise COX regression was used to establish a risk prediction model through the training sets randomly separated from the TCGA cohort and validated in the remaining testing sets and two GEO datasets (GSE17538 and GSE38832). A 12-DNA-damage-and-repair-related gene-based signature able to classify COAD patients into high and low-risk groups was developed. The predictive ability of the risk model or nomogram were evaluated by different bioinformatics- methods. Gene functional enrichment analysis was performed to analyze the co-expressed genes of the risk-based genes.

Result: A 12-gene based prognostic signature established within 160 significant survival-related genes from DNA damage and repair related gene sets performed well with an AUC of ROC 0.80 for 5 years in the TCGA-CODA dataset. The signature includes CCNB3, ISY1, CDC25C, SMC1B, MC1R, LSP1P4, RIN2, TPM1, ELL3, POLG, CD36, and NEK4. Kaplan-Meier survival curves showed that the prognosis of the risk status owns more significant differences than T, M, N, and stage prognostic parameters. A nomogram was constructed by LASSO regression analysis with T, M, N, age, and risk as prognostic parameters. ROC curve, C-index, Calibration analysis, and Decision Curve Analysis showed the risk module and nomogram performed best in years 1, 3, and 5. KEGG, GO, and GSEA enrichment analyses suggest the risk involved in a variety of important biological

processes and well-known cancer-related pathways. These differences may be the key factors affecting the final prognosis.

Conclusion: The established gene signature for CRC prognosis provides a new molecular tool for clinical evaluation of prognosis, individualized diagnosis, and treatment. Therapies based on targeted DNA damage and repair mechanisms may formulate more sensitive and potential chemotherapy regimens, thereby expanding treatment options and potentially improving the clinical outcome of CRC patients.

Keywords: mRNA signature, DNA damage, DNA repair, prediction, prognosis, colon cancer

INTRODUCTION

Colon cancer is a malignant intestinal disease with the highest incidence among gastrointestinal diseases. Colorectal cancer is the third most common cancer and one of the major cancers for mortality all over the world (Bray et al., 2018). The application of combined drugs, including adjuvant chemotherapy and radiotherapy (Dekker and Rex, 2018), is currently a worldwide accepted standard treatment for colon cancer. Besides, early diagnosis of primary or recurrent colon cancer is one of the key factors for the prognosis. Unfortunately, how to diagnose colon cancer early remains one of the most difficult issues in cancer treatment. The study reported in-depth research on the diagnosis and treatment of colon cancer, such as endoscopic diagnosis (Dekker and Rex, 2018), tumor markers (Sveen et al., 2020), and molecular targeted therapy (Ganesh et al., 2019). The American Joint Committee on Cancer divided the patients into stages I, IIa, IIb, IIIa, IIIb, IIIc, and IV according to the tumor-node-metastasis (TNM). The TNM staging can distinguish patients with different prognoses (O'Connell et al., 2004). There is still a possibility of recurrence in stage I to III patients who underwent curative resection, and the likelihood of recurrence increases with time and stage. However, due to complex pathogenesis and high metastasis rate, the diagnosis is still unsatisfactory, and the prognosis is poor (Kobayashi et al., 2007). Therefore, there is an urgent need to identify new diagnostic and prognostic biomarkers, therapeutic targets, and look into the potential molecular mechanisms of CRC. Today, the revolution helps to identify disease-related biomarkers through more novel bioinformatics analysis and the use of next-generation sequencing technology (Moody et al., 2017), which will help the early identification of colon cancer and the development of personalized treatment plans to benefit more patients.

There is an increasing interest in the search for new genes and the construction of multi-gene prediction models recently. Genome analysis based on the TCGA network project containing 276 patients' CRC samples and corresponding germline DNA samples showed that some genes have been shown to be associated with highly mutated CRC (Ganesh et al., 2019). In hypermutated cancers, APC, TGFBR2, BRAF, MSH3, MSH6, SLC9A9, and TCF7L2 were highly mutated, in particular the frequent mutations of BRAF (V600E). On the contrary, the mutation rate of TP53 and APC was lower. In non-hypermutated cancer, APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, and NRAS were frequently mutated. Based on the mutation status, CRC could be divided into the non-hypermutated group (84%) and the hypermutated group (16%; Moody et al., 2017). Different studies have identified that CDX2, LC3B, ULBP2, SEMA5A, VEGF-D, and SMAD7 are potential biomarkers for the prognosis of colon cancer (Lord and Ashworth, 2017; Gourley et al., 2019; Mauri et al., 2020). However, the prognostic value of a single-gene related clinical prognostic model for CRC patients based on these genes is still not ideal. Yang et al. have constructed a 20 gene signature based on the expression profile of GSE44076 about colon cancer, which were considered as diagnosis targets for colon cancer (Chen et al., 2014).

In recent years, research on new therapeutic targets for different cancer types has gradually focused on genomic changes in the DNA damage response (DDR) pathway (Mauri et al., 2020). The current research on anti-tumor drugs mainly focuses on two main types: Platinum compounds and poly ADP-ribose polymerase inhibitors (Lord and Ashworth, 2017; Gourley et al., 2019). DDR changes were originally found in breast cancer and ovarian cancer, while it has now expanded to prostate and pancreatic cancer (Mauri et al., 2020). The role of DDR alterations in colorectal cancer is still not fully studied. There are only a few studies on its clinical impact and no orderly study system has been established (Chen et al., 2014; Lei et al., 2019; Sun et al., 2019; Karpov et al., 2020; Mauri et al., 2020; Scagliarini et al., 2020; Yu et al., 2020).

In our study, we aimed to construct a DNA damage and repair related gene-based signature and nomogram to make an improvement on the prognostic value of CRC through comprehensive bioinformatics methods.

MATERIALS AND METHODS

Data Collection

The DNA damage and DNA repair related genes list were collected from GSEA gene sets[1] by the keyword "DNA AND damage" or "DNA AND repair." At last, 1545 genes related to DNA damage and repair were included in the analysis (**Supplementary Table 1**).

The gene expression data of HTseq RNA profiles FPKM (fragments per kilobase of exon per million reads mapped)

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

of 471 COAD and 41 compared normal samples were extracted from The Cancer Genome Atlas-Colon adenocarcinoma (TCGA-COAD).[2] Survival endpoint (vital status, days to the last follow-up, and days to death), age, stage, and histological type of primary of each patient were also retrieved.

The public expression profiles data of colon cancer were extracted from the GEO database[3] by the keywords ["Colonic Neoplasms" (MeSH)]. The selected data must meet the following inclusion criteria: human gene expression profiles data of solid tissues of colon cancer, the datasets contained prognosis survival information, and enough samples for analysis. Four eligible data (GSE17538 and GSE38832, GSE44861 and GSE44076), based on the platform of Affymetrix-GPL570, Affymetrix-GPL570, Affymetrix-GPL, and Affymetrix-GPL13667 respectively, that met the above criteria were annotated based on the annotation platform and enrolled in this study, each GEO data set was checked the gene expression distribution was through the histogram and normalization. Furthermore, the related clinical data of the four datasets were retrieved.

Construction of the DNA Damage and DNA Repair Related Gene Signature

All analyses in this study conducted in R language used R version 4.03. Univariate Cox regression analysis (Cox, 1972) was first performed with DNA damage and DNA repair related genes, and genes with a p value of less than 0.05 were considered a statistically significant difference. After randomly separating samples into the training set and testing set, genes that were strongly associated with OS of COAD patients were used for multivariate Cox hazards regression base based on the training set with the stepwise method in My.stepwise package (Hu, 2017). The process and results are shown in the **Supplementary Material**. Then a multivariate cox hazards regression model was built to assess the prognostic value for COAD.

The hazards model was established by the selected final gene signature, and the risk score was generated according to the following formula:

$$Risk\ score = \sum_{i=1}^{N} \beta i * Ei$$

(N represents the total number of signature genes, βi and Ei represent the coefficient index, and the gene expression level of each gene, respectively)

Based on the risk score of each patient, samples were grouped into high risk and low-risk groups based on the risk score of each patient, and the relationship between risk and clinical data was then investigated.

The Nomogram Establishing

All clinical prognostic factors T, M, N, age, and stage together with risk group were used for the selection of the prognostic parameters by Least Absolute Shrinkage and Selection Operator (LASSO; Friedman et al., 2010) regression analysis. And a related prognostic nomogram to assess the probability of 0.5-, 1-, and 3-year OS for COAD patients were built by "rms" R package. Calibration plots were used to evaluate the discriminative ability of the nomogram.

Validation of the Multi-Gene Prognostic Signature

Firstly, survival analysis between high and low groups combined with clinical stage and the histological type was evaluated by the Kaplan-Meier curve (Ranstam and Cook, 2017) and log-rank test (Kleinbaum, 1998). The ROC curve (Kamarudin et al., 2017) and the AUC, C-index, Calibration analysis, and Decision Curve Analysis (Vickers and Elkin, 2006) were performed by "timeROC," "rmda," and "survcomp" packages to evaluate the risk model and the nomogram. Similarly, we evaluated the prediction efficiencies of the risk score system in the testing sets and GEO validation sets too.

The Cutoff Value of the Km Curve

To better evaluate the validation model and the whole cohort model, we obtained a relatively fixed cutoff value by "Surv_cutpoint" function through the training cohort. This can ensure that the corresponding cutoff value will not be biased after different groups, and the verification of the model will be relatively more accurate. This cutoff value is only the best cutoff value obtained by the training group. This cutoff value will vary with the sample changes. Each cohort was divided into high-risk groups and low-risk groups according to their respective cutoff value.

Gene Co-expression Network and Gene Functional Enrichment Analysis

Genes which co-expressed with the 12 risk-related genes were selected by the Pearson correlation method in TCGA-COAD high-risk group, low-risk group, and normal samples, and $p < 0.05$ were considered as significant. The co-expressed genes with Pearson correlation coefficient $|R| > 0.6$ were converted into a Topological Overlap Matrix (TOM) by "plotNetworkHeatmap" in the "WGCNA" package (Friedman et al., 2010), and the co-expressed genes with Pearson correlation coefficient $|R| > 0.7$ were converted into gene co-expression network by "network_plot" in the "correlate" package.

Gene ontology (GO) term analysis, Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) pathway enrichment analyses were then performed with the "clusterprofiler" package to investigate the biological functions and pathway of the genes list used in the TOM heatmap. Gene set enrichment analysis (GSEA, https://software.broadinstitute.org/gsea/index.jsp) was used to analyze signaling pathway enrichment in high- and low-risk groups. The result of the enrichment analysis of biological functions and pathways were displayed by visual graphics. The top 10 most significant results of BP (biological process), CC (cellular components), MF (molecular function), and KEGG were selected, respectively.

[2]https://cancergenome.nih.gov/
[3]https://www.ncbi.nlm.nih.gov/geo/

The GSEA analysis was performed with the following settings: FDR < 0.25, NOM value of $p < 0.05$, and $|NES| > 1$.

RESULT

Characteristics of COAD Patients in the TCGA Dataset and GEO Dataset

We enrolled 439 patients with follow-up time >30 days in total as the discovery set for construction and validation of the model. 263 and 176 patients were separated by random into two groups: the training group and the testing group. The patient characteristics of the training set and test set were in balance ($p > 0.1$). The average age in years was 66.8 ± 12.2, and 119 females (45.2%) in the training set; while the average age in years was 65.4 ± 13.4, 85 females (48.2%) in the testing set (**Table 1**).

Meanwhile, we downloaded four eligible datasets (GSE17538, GSE38832, GSE44861, and GSE44076) from GEO. However, two datasets (GSE44861 and GSE44076) were discarded for containing only 8 of 12 related genes we screened out, and the other two datasets (GSE17538 and GSE38832) containing 11 of 12 related genes are kept as validation datasets. Using the same exclusion criteria of the training group, 232 colon cancer patients out of a total of 238 samples were selected from GSE17583 datasets [average age in years was 64.7 ± 13.4, 110 females (47.4%)]. GSE38832 contains 122 colon cancer patients with disease-free survival and disease-specific survival information, but not overall survival information.

TABLE 1 | TGCA patient characteristics.

Variable		Number			
		Total set	Training set	Testing set	*p* value
Case		439	263	176	/
Gender	Female	204	119	85	0.396
	Male	235	144	91	
Survival status	Alive	346	210	136	0.811
	Dead	93	53	40	
Endpoint time		2.4 ± 2.0	2.5 ± 2.2	2.3 ± 1.8	0.943
Age		66.3 ± 12.7	66.8 ± 12.2	65.5 ± 13.4	0.649
M	M0	324	194	130	
	M1	61	39	22	0.994
	MX	49	27	22	
N	N0	258	149	109	
	N1	103	65	38	0.574
	N2	78	49	29	
T	T1	11	6	4	
	T2	78	42	36	
	T3	299	174	125	0.313
	T4	51	40	11	
	NA	11	6	5	
Stage	STAGE I	75	41	34	
	STAGE II	167	100	67	0.499
	STAGE III	125	77	48	
	STAGE IV	61	39	22	

NA, not reported.

Characteristics of patients in the training set, testing set of TCGA, GSE17583, and GSE38832 are summarized in **Table 2**.

Selection of DNA Damage and DNA Repair Related Genes and Construction of the Signature

In the training set, univariate Cox regression analysis was performed for all the DNA damage and repair related genes selected from GSEA. As shown in **Figure 1A**, 27 DNA damage and repair related genes play a favorable role for COAD patients' survival (blue, Hazard Ratio (HR) < 1, $p < 0.05$), and 133 genes were in risk roles (red, HR > 1, $p < 0.05$), while 1,385 gene showed no significance. Twelve genes were selected by stepwise multivariate regression analysis as reliable predictors, including CCNB3, ISY1, CDC25C, SMC1B, MC1R, LSP1P4, RIN4, TPM1, ELL3, POLG, CD36, and NEK4 (**Figure 1B**). All the above genes except CDC25C show an independent prognostic manner ($p < 0.05$). Among them, CCNB3, ISY1, SMC1B, MC1R, LSP1P4, RIN2, ELL3, POLG, and CD36 may be considered as oncogenes, whereas CDC25C, TPM1, and NEK4 may be tumor suppressor genes. The coefficients of these DNAs indicated their impact on survival prediction. Subsequently, the risk score system for TCGA-CAOD samples based on the expression level and the corresponding beta value of each gene was constructed by the following formula:

$$RS = (3.5) \times ExpCCNB3 + (0.27) \times ExpISY1 + (-0.081) \times ExpCDC25C + (0.48) \times ExpSMC1B + (0.26) \times ExpMC1R + (0.34) \times ExpLSP1P4 + (0.11) \times ExpRIN4 + (-0.039) \times ExpTPM1 + (0.3) \times ExpELL3 + (0.11) \times ExpPOLG + (0.19) \times ExpCD36 + (-0.46) \times ExpNEK4.$$

According to the optimal cutoff value of 2.95 simulated by "Surv_cutpoint" function in "survminer" package, the TCGA-COAD patients were classified into high- and low-risk sets (**Figure 2A**). The patients' status, survival time, and DNA expression levels of the test TCGA set, total TCGA set, and training TCGA set are shown in **Figures 2B–G**.

The survival analysis presented that the OS of the low-risk set was better than that of the high-risk set in the training set of TCGA (hazard ratio, HR = 0.16, 95% confidence interval, 95% CI (0.1–0.24; **Figure 2H**). The results were consistent in the TCGA total set (HR = 0.138, 95% CI (0.079–0.24); $p < 0.001$; **Figure 2I**) and testing set (HR = 0.234, 95% CI (0.12–0.44); $p < 0.001$; **Figure 2J**). The 5-year survival rate for high and low risk is 11 and 79%, respectively, (**Figure 2I**). The area under the ROC curve (AUC) for 1-, 3-, 5-, and 10-year OS were all above 0.8 in the TCGA training set (**Figure 2K**), and in the TGCA total set (**Figure 2L**) and TCGA testing set (**Figure 2M**), they were all above 0.75. Meanwhile, we investigated the relationship between risk score and clinicopathologic features including T, N, M, and stage in the TCGA total cohort. As shown in **Figures 3A–D**, respectively comparing the clinical data of patients of the same T, N, M, and stage in the high-risk and low-risk groups, the prognosis of patients was significantly different. Under the same T, N, M, or stage, the

TABLE 2 | GEO patient characteristics.

GSE17583			GSE38832		
Case	232		Case	122	
Gender	Female	110	dfs time (year)	3.84 ± 2.77	
	Male	122	dfs status	no recurrence	83
Survival status	Alive	139		recurrence	9
	Dead	93		NA	30
Endpoint time (year)		3.95 ± 2.56	dss status	no death	94
Age		64.73 ± 13.43		death from cancer	28
Ajcc stage	1	28	Ajcc atage	1	18
	2	72		2	35
	3	76		3	39
	4	56		4	30
Tumor differentiation	WD	17			
	MD	235		/	
	PD	30			

NA, not reported; WD, well differentiated; MD, moderately differentiated; PD, poorly differentiated.

survival time of patients in the low-risk group was longer than that of the high-risk group.

Validation of the Genes Signature in GEO Dataset

GSE17538 and GSE38832 datasets both based on the platform of Affymetrix-GPL570 included the 11 above risk-related genes except LSP1P4 were used for the following analysis. The results showed that though the new gene signature missing a significant gene, the 11-gene based signature still had a significant performance for OS, DFS, and DSS prediction in the two GEO validation datasets (**Figures 4A,D,G,I,J,L**). The relationship between risk score of "ajcc_stage" and tumor differentiated grade was also investigated in the two sets, which showed that in the same stage or differentiated level, the survival time of patients in the low-risk group was apparently longer than that of the high-risk group (**Figures 4B,C,H,K**), similar to the results in the training set. Together, we considered that the 11-gene signature had a prominent prognostic ability not only for OS prediction but also DFS and DSS prediction.

Comparison of the Prognostic Performance of Genes Signature With Clinical Predictive Factors

Given the fact that T, N, M, and stage have been thought to be predictive factors of the prognosis of COAD in the past, we managed to compare the prognostic performance of these clinicopathologic features with our 12-gene signature. Survival analysis of the above clinical indicators was completed, respectively, in the high-risk and low-risk groups (**Figures 2L**, **3E-H**). The survival analysis presented that these clinicopathologic features showed less satisfactory performance for OS prediction than that of 12-gene signature. The area under the ROC curve (AUC) for 1-, 3-, 5-, and 10-year OS of T (the size of the tumor) were 0.67, 0.634, 0.576, and 0.543 in the total TCGA set, comparing with the AUC of 12-gene signature in the total TCGA set (0.832, 0.797, 0.843, and 0.797). The results were consistent in the GEO colon cancer validation sets containing not only COAD patients (**Figure 4**). Combining the above results, a 12-gene signature can be used as a satisfactory indicator to predict the prognosis of COAD patients or the whole colon cancer types.

Establishment and Validation of the Nomogram Survival Model

By the usage of multivariable Cox regression analyses, pathologic M, pathologic T, pathologic N, stage, age, gender, and risk score status were selected to assess the independent prognostic manner in the COAD samples. Based on the result shown in **Figure 1C**, the risk score can be used as an independent prognostic factor without being affected by clinicopathologic features. And the HR of the high-risk group is 4.56 (2.87–7.25) times danger than that of the low-risk group (**Figure 1C**) The result of the multivariable Cox regression analysis of 12 genes along with clinicopathologic features was revealed in **Figure 1D**, indicating that most of these genes except MC1R can also act as independent prognostic factors, and may have an excellent suggestive effect on predicting the survival of COAD patients. Among these genes, CCNB3, ELL3, LSP1P4, and SMC1B showed a significant harmful effect on COAD OS (HR > 1.5, $p < 0.05$).

To establish a clinical method to predict the survival probability of COAD patients, we created a nomogram by LASSO regression analysis based on the TCGA cohort to estimate the probability of the 1-, 3-, and 5-year OS with T, N, M, age, gender, stage, and risk group status (**Figure 5A**). LASSO regression analysis established that the nomogram contained 5 prognostic factors including age, T, M, N, and risk (**Figures 5C,D**). The AUC of 1-, 3-, and 5-year OS predictions all above 0.8 (**Figure 5G**).

Calibration curves were used to evaluate the consistency between actual and predicted survival rates. As shown in **Figure 5B**, the accuracy of this model in predicting a 5-year survival rate is low, but in predicting a 1- and 3-year survival

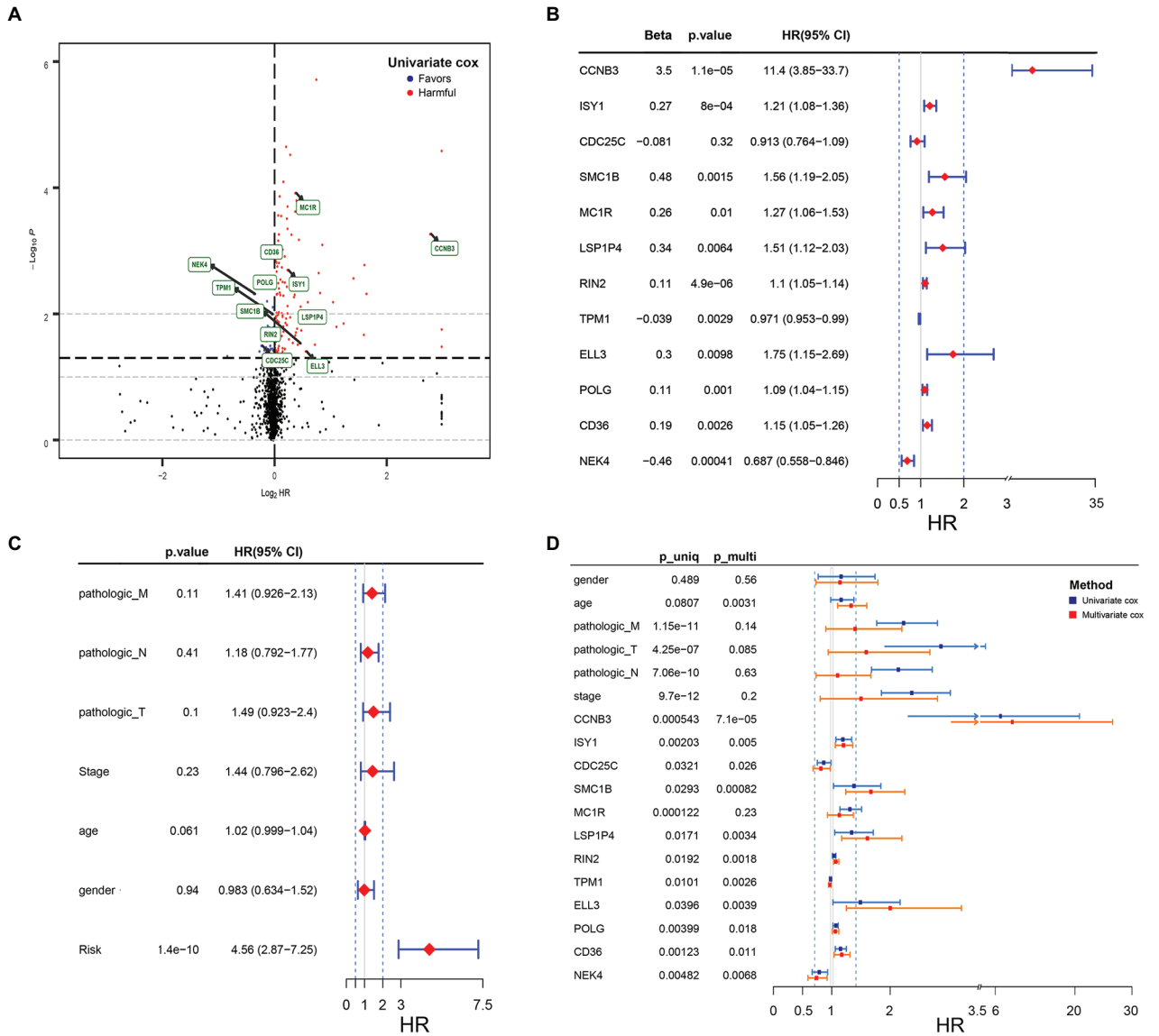

FIGURE 1 | Volcano plot of DNA damage and repair related genes and forest plot of the multivariate Cox regression analysis in TCGA cohorts. **(A)** Volcano plot of DNA damage and repair related genes: blue indicates protective genes, red indicates harmful genes and black indicates no significance genes. **(B)** Forrest plot of the multivariate Cox regression analysis OS of 12 genes. **(C)** Forrest plot of the multivariate Cox regression analysis OS of clinical factors and risk score. **(D)** Forrest plot of the multivariate Cox regression analysis OS of clinical factors and 12 genes. Beta values represent the coefficient index β for each gene and clinical factors.

rate it is high, showing that the nomogram was best for predicting 1-, 3-year OS in COAD patients. The concordance index (C-index) was calculated to evaluate the model prognosis capability. The values of 0.5 and 1.0 represent a random probability and an excellent performance for predicting survival with the model. The C-index of the risk score and nomogram were all above 0.75 between the 1–5 years OS prediction, which was much better than any other independent predictor (**Figure 5E**). We used DCA analysis to confirm a range of threshold probabilities for a prediction mode, as shown in **Figure 5F**, the nomogram threshold probability based on 12-gene combinations was significantly better than the default strategies of treating all or none at a threshold probability more than 0.1, and the results come better than any other predictor used in this study.

Function and Signaling Pathways Analysis of Genes in the Prognosis Module

The model constructed by 12 genes can effectively distinguish patients with different prognoses, which suggests that patients with different risk scores may be involved in different important pathways that cause differences in the final prognosis. Based on the above conjectures, we performed GSEA analysis in high- and low-risk patients, respectively, to confirm the significant

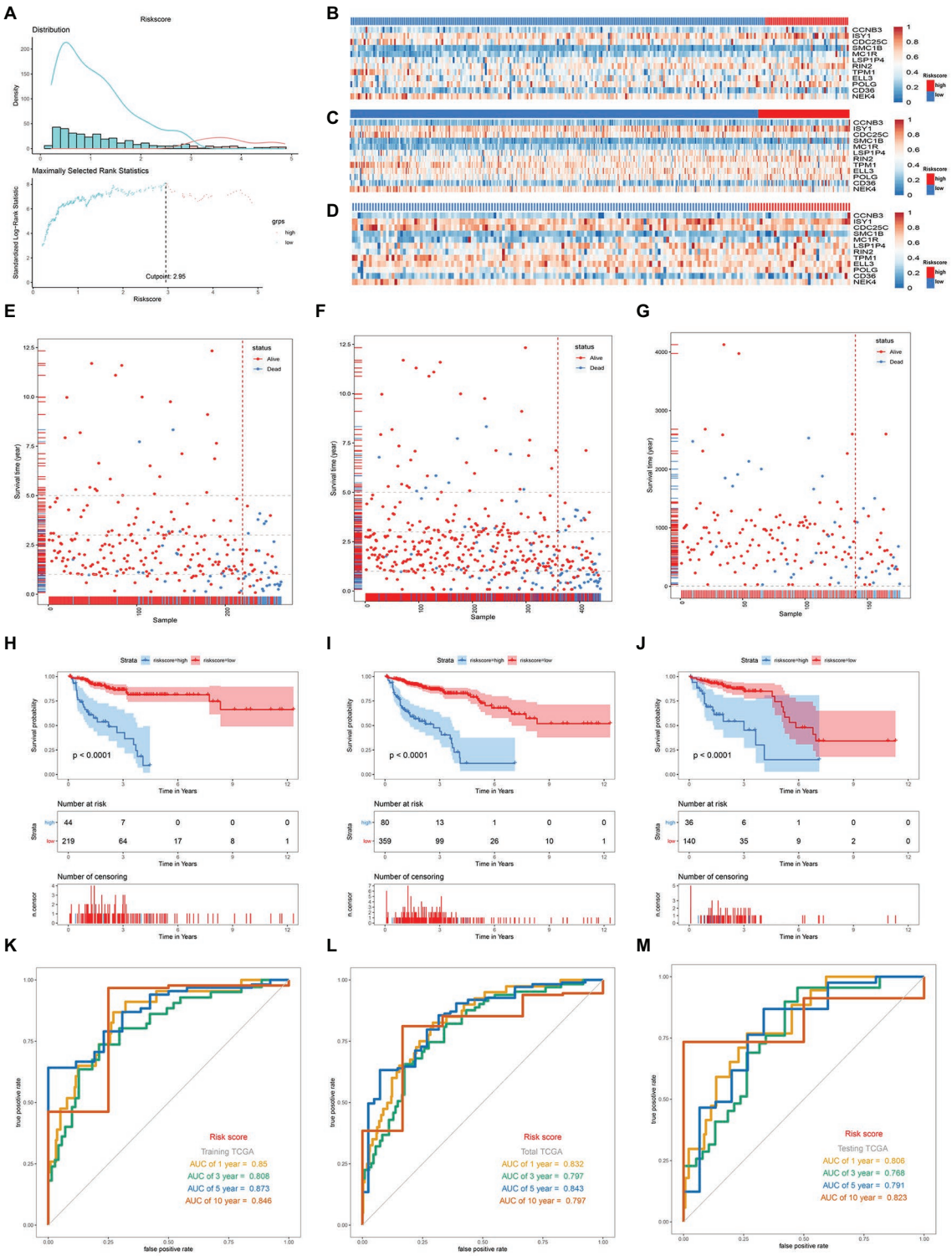

FIGURE 2 | Distribution of risk score, Gene expression heatmaps, Kaplan-Meieranalysis and ROC analysis of 12-gene signature in the training TGCA set, total TCGA set, and testing set. **(A)** Distribution of risk score and the cutoff point. **(B–D)** Gene expression heatmaps in the training TGCA cohort **(B)**, total TCGA cohort **(C)**, and testing TCGA (**D**; The blue color is the low-risk group and the red color is the high-risk group). **(E,F)** Correlation between the prognostic signature and the OS of patients in the training TGCA cohort **(E)**, total TCGA cohort **(F)**, and testing TCGA **(G)**. **(H–J)** Kaplan-Meier survival analysis of the low- and high-risk group patients in the training TGCA cohort **(H)**, total TCGA cohort **(I)**, and testing TCGA **(J)**. **(K–M)** ROC curve analysis according to the 1, 3, 5, 10-year survival of the area under the AUC value in the training TGCA cohort **(K)**, total TCGA cohort **(L)**, and testing TCGA **(M)**.

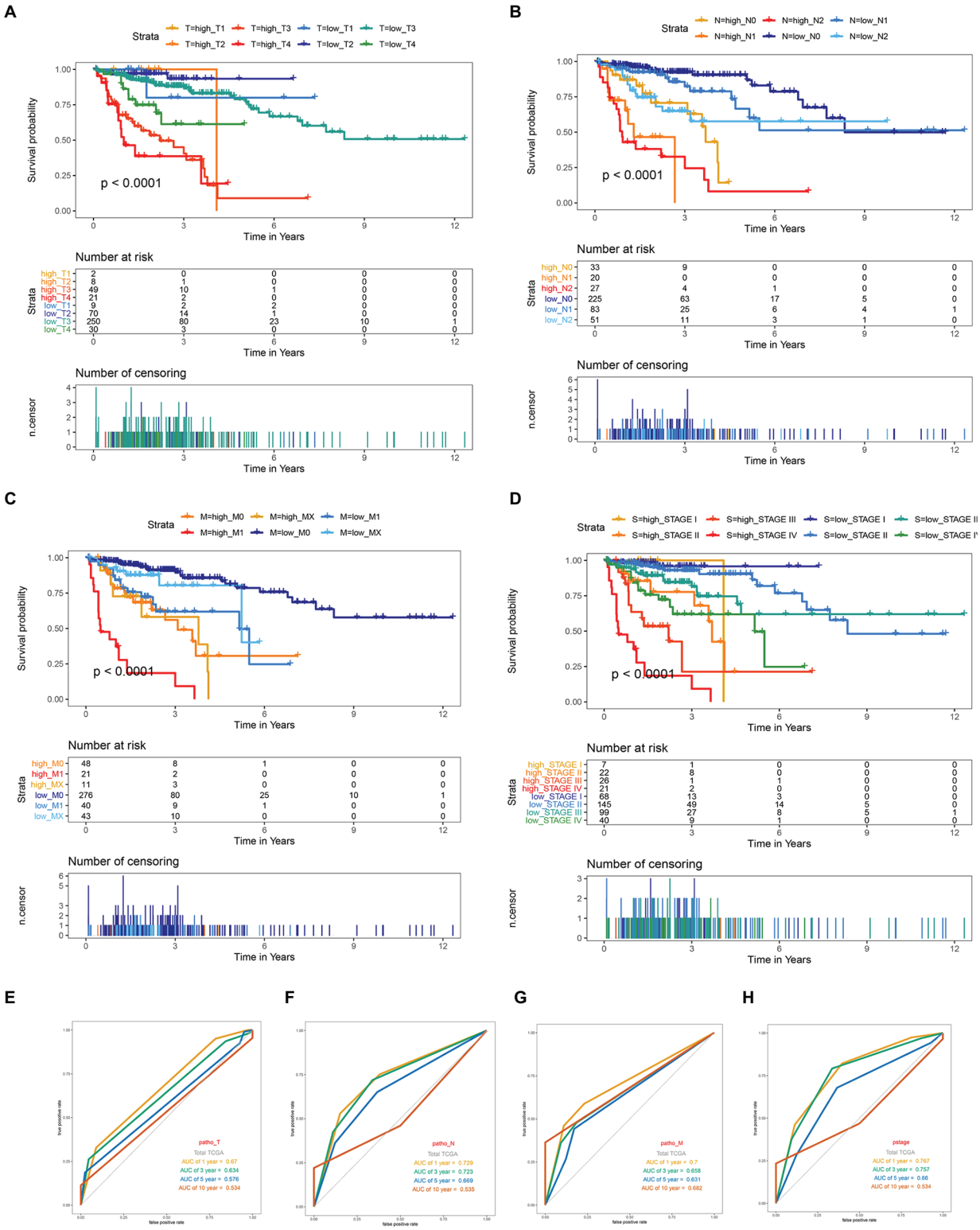

FIGURE 3 | Kaplan-Meier survival for OS in high-risk and low-risk group of different subgroup and ROC curve analysis of T, N, M and stage in the total TCGA cohort. **(A)** In subgroups stratified by T1, T2, T3, and T4. **(B)** In subgroups stratified by N0, N1, and N2. **(C)** In subgroups stratified by M0, M1, and MX. **(D)** In subgroups stratified by stage I, stage II, stage III, and stage IV. **(E–H)** ROC curve analysis of T, N, M and stage according to the 1, 3, 5, and 10-year survival of the area under the AUC value in the total TCGA cohort.

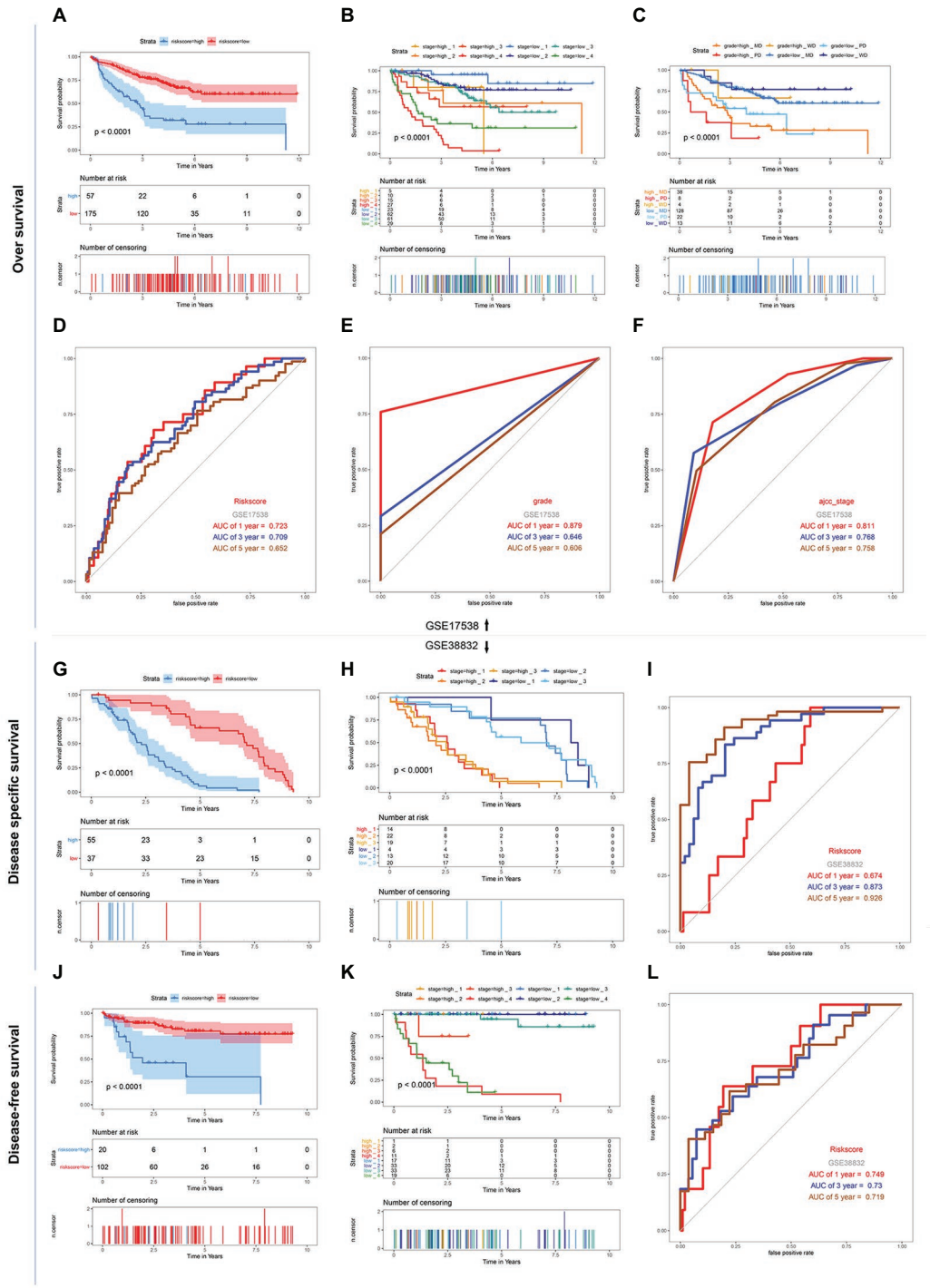

FIGURE 4 | Kaplan-Meier survival and ROC curves of the 12-DNA signature, grade and stage in the two GEO sets. **(A)** Correlation between the 12-DNA signature and the overall survival of patients in the GSE 17538 set. **(B,C)** Kaplan-Meier survival for OS in high-risk and low-risk group of different subgroup in the GSE 17538 set: in subgroups stratified by stage I, stage II, stage III, and stage IV, in subgroups stratified by grade MD, grade PD, and grade WD. **(D–F)** ROC curve analysis of risk score, stage and grade according to the 1, 3, 5, and 10-year survival of the area under the AUC value in the GSE 17538 set. **(G)** Correlation between the 12-DNA signature and the disease specific survival of patients in the GSE 38832 set. **(H)** Kaplan-Meier survival for disease specific survival in stage 1, 2, and 3 subgroups of high-risk and low-risk group in the GSE 38832 set. **(I)** ROC curve analysis of risk score according to the 1, 3, and 5-year disease specific survival of the area under the AUC value in the GSE 38832 set. **(J)** Correlation between the 12-DNA signature and the disease-free survival of patients in the GSE 38832 set. **(K)** Kaplan-Meier survival for disease-free survival in stage 1, 2, 3, and 4 subgroups of high-risk and low-risk group in the GSE 38832 set. **(L)** ROC curve analysis of risk score according to the 1, 3, and 5-year disease-free survival of the area under the AUC value in the GSE 38832 set.

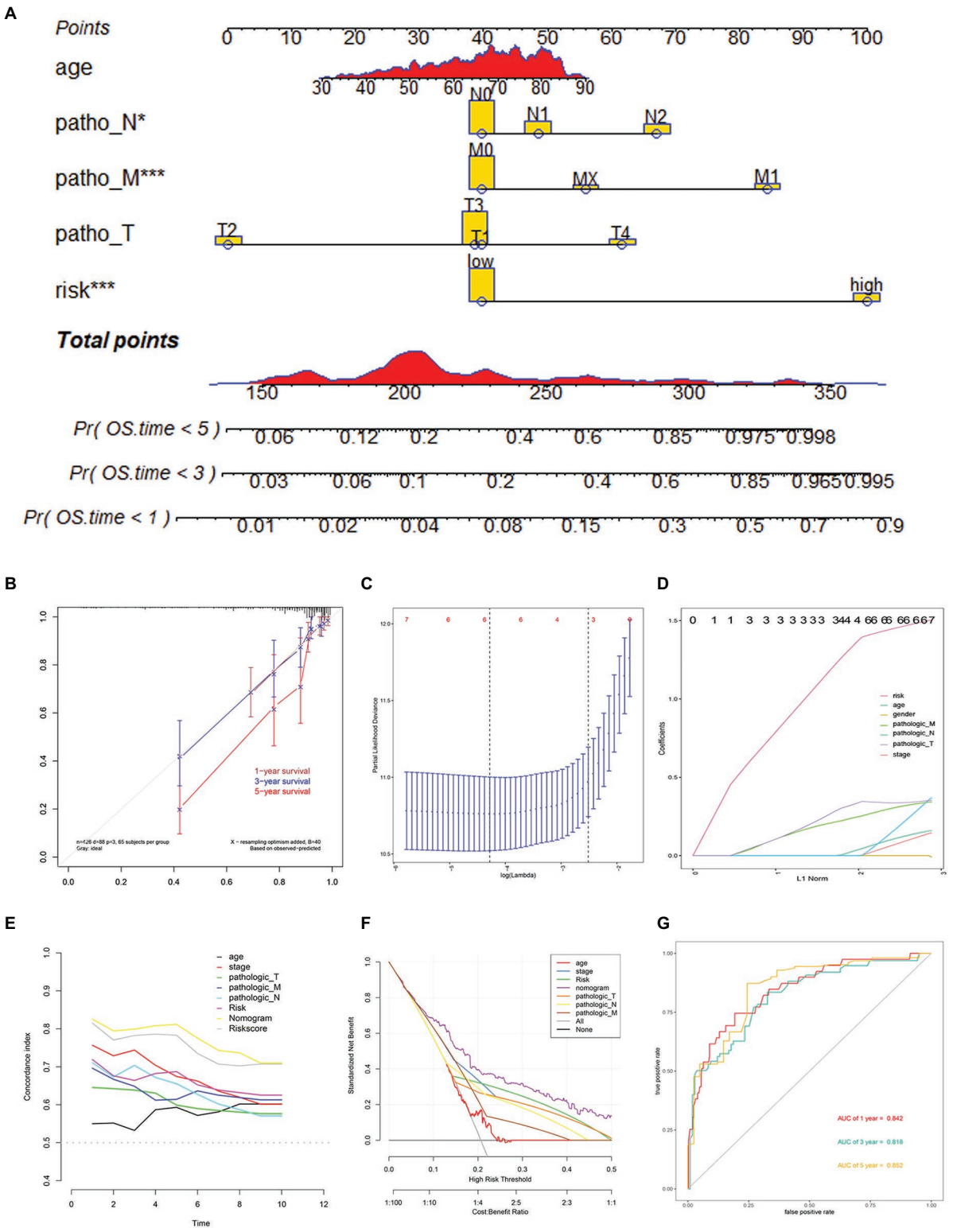

FIGURE 5 | Nomogram construction based on 12-gene signature and prognostic value of 12 genes. **(A)** The nomogram for predicting the proportion of patients with 1-, 3-, or 5-year OS. **(B)** Calibration plots of the nomogram. **(C,D)** LASSO regression analysis used 10-fold cross-validation *via* the maximum criteria. **(E)** C-index of the nomogram **(F)** Decision curve analysis of nomogram predicting 1-, 3-, and 5-year OS of COAD comparing age, stage, the risk score, Pathologic T, Pathologic N, and Pathologic M. **(G)** Time-dependent ROC analysis of nomogram predicting 1-, 3-, and 5-year OS of COAD.

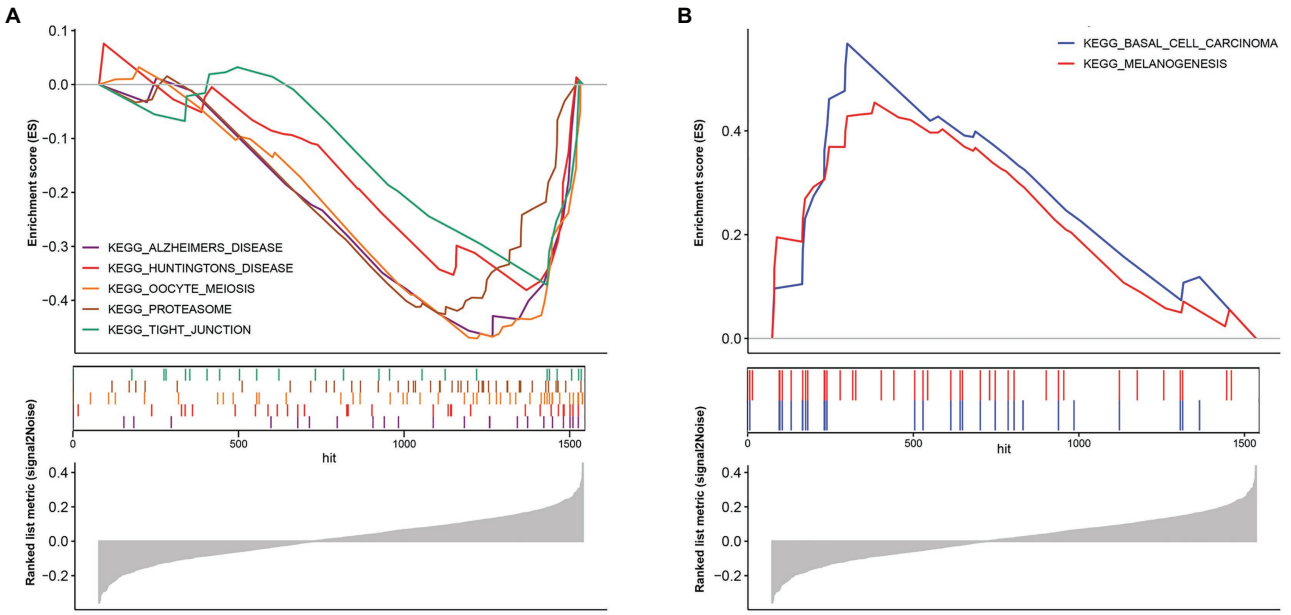

FIGURE 6 | Biological pathways in two different risk groups by GSEA analysis. **(A)** Enriched pathways in the low-risk group. **(B)** Enriched pathways in the high-risk group.

pathways in each group. According to the enrichment results, two different groups have their characteristic pathways. Multiple pathways such as Alzheimers disease, Huntingtons disease, Oocye meiosis, Proteasome, and Tight junction are downregulated in patients with a low-risk score (**Figure 6A**). On the other hand, in the high-risk group, two pathways, including Basal cell carcinoma and Melanogenesis, were up-regulated (**Figure 6B**).

Biological processes are often not the result of the action of a single gene but are often realized through the interaction between genes. Considering that gene expression varies in different individuals and different statuses, we searched for genes related to 12 genes in the normal group, low-risk group, and high-risk group and took the intersection of the three as the gene group of 12 genes co-expression. We used R = 0.6 and $p < 0.01$ as the cutoff value and the correlation with any one of the 12 genes met the condition that they were included in the statistics. Finally, 16,505, 9,561, and 5,260 (including 12 genes) were found in the normal group, low-risk group, and high-risk group, respectively (**Figure 7A**). The number of genes related to 12 genes is the largest in the normal group and the least in the high-risk group, which is related to tumor heterogeneity. The lowest number of genes in high-risk patients suggests more significant heterogeneity, which is consistent with the final poor prognosis. We used WGCNA to build the Topological Overlap Matrix (TOM), which proved that the selected gene group has a good correlation (**Figure 7C**). Next, we further screened the related genes with a cutoff value > 0.7, resulting in a total of 42 genes including the genes of the module. These genes are roughly classified into three clusters, most of the 12 genes (10/12) are located in the upper left corner, and there is a clear correlation between the other two clusters of genes, which further proves the relative independence of the genes of the module and the reliability of the co-expressed genes (**Figure 7B**).

GO enrichment analysis and KEGG pathway enrichment analysis are performed to investigate the biological functions and pathways of the Co-expressed genes. The results of KEGG enrichment analysis showed that the co-expressed genes were significantly enriched in important biological pathways, such as RNA transport, Cell cycle, Spliceosome, and so on (**Figure 7D**). The cellular components (CC) analysis indicated that proteins encoded by genes were mostly located in the chromosomal region, nuclear speck, condensed chromosome, chromosome centromeric region, and spindle (**Figure 7E**). Those molecular function (MF) were significantly associated with ATPase activity, helicase activity, ATPase activity coupled, catalytic activity acting on DNA, and so on (**Figure 7F**). For the biological process (BP), genes were mainly enriched in chromosomal segregation, organelle fission, nuclear division, DNA replication (**Figure 7G**).

DISCUSSION

COAD has one of the highest fatality rates of tumors in the digestive system. It is more common in men over the age of 40. However, early diagnosis of COAD was extremely difficult, and many patients have progressed to advanced cancer when they are diagnosed with COAD, leading to a bad prognosis. Early diagnosis and treatment of COAD can greatly improve the prognosis of COAD patients, which will not only reduce the economic burden of patients but also improve the quality of life. TNM staging is the one that is currently widely used, but this staging has certain drawbacks, and differences in treatment

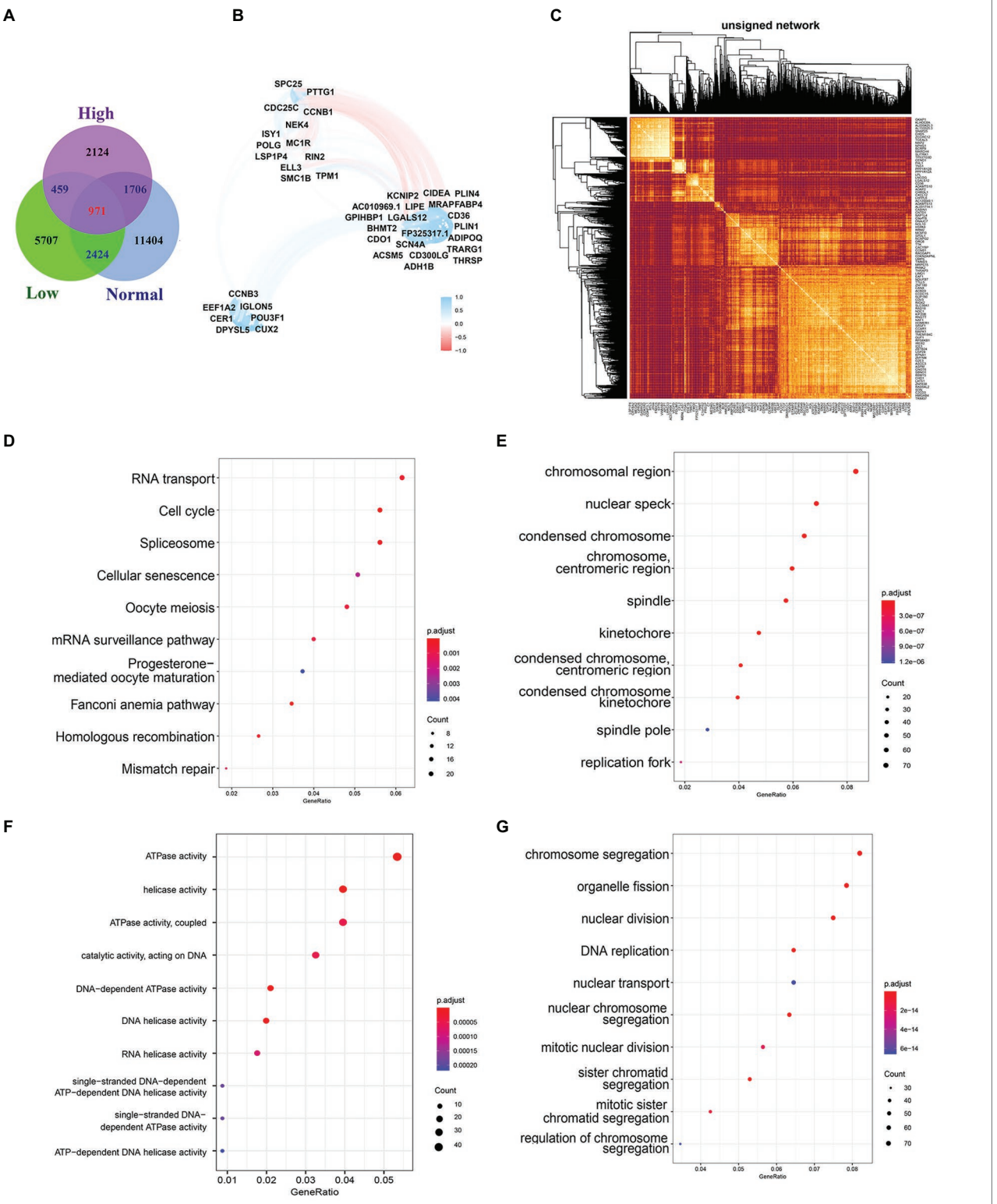

FIGURE 7 | Biological functions and pathways of co-expressed genes. **(A)** Venn diagram of overlapping genes among Normal group, Low-risk group, and High-risk group. **(B)** Topological overlap heatmap of gene co-expression network. Dark colors mean high topological overlap, while Light colors mean low topological overlap. **(C)** Co-expressed genes selected by R2 > 0.7. **(D)** The top 10 most significant results of KEGG. **(E–G)** The GO enrichment analysis of co-expressed genes, including the CC **(E)**, the MF **(F)**, and the BP **(G)**.

options may have caused unexpected differences in survival outcomes. For example, patients with stage IIIA disease receiving chemotherapy have better survival than those with stage IIB disease, where the survival difference is based on the benefit of chemotherapy or whether the stage IIA tumor itself is unknown according to aggressiveness (O'Connell et al., 2004). Meanwhile, with the intensive research on the molecular mechanism of tumors, the advantage of prognosis prediction based on gene-level is gradually exhibited. For example, colorectal cancers (CRCs) are classified into MMR and MMR-d based on whether they have normal DNA mismatch repair (MMR) function, a phenotype that is also an important prognostic indicator. It has been controversial whether the MMR-d/MSI-H phenotype benefits from 5-fluorouracil – based chemotherapy (Stadler, 2015). Therefore, the discovery, identification, and evaluation of new biomarkers are greatly important for COAD patients.

By consulting the previous literature, DNA damage and repair have been proved to be related to the proliferation and metastasis of CRC, but there is no research to clarify its direct relationship with the prognosis or consider DNA damage and repair related genes as prognosis predictors, which serves as the breakthrough point of our research. DNA is constantly on the exposure to endogenous and exogenous sources of damage, destroying genomic integrity (Hoeijmakers, 2009). Unable to repair DNA damage in a precise and well-timed way will lead to various genomic aberrations, including point mutations, chromosomal translocations, and the acquisition or loss of chromosomes. The accumulation of these aberrations will further cause changes in the cells, thus driving the tumorigenesis (Burrell et al., 2013; Khanna, 2015; Jeggo et al., 2016). The contrasting activity of multiple DNA repair pathways plays a key role in interrupting this accumulation and maintaining genomic integrity (Mouw et al., 2017). DNA repair and damage have been described as being related to the occurrence and development of various cancers, such as breast cancer and ovarian cancer. So, we legitimately speculated that DNA damage and repair were closely related to the development of CRC. We used DNA damage and repair related gene sets collected from GSEA gene sets and TCGA-COAD cohort to assess their diagnostic value.

The fast development of sequencing technology produces massive data, which facilitates tumor biomarker identification and a lot of resources have been invested in corresponding research. For example, Yang et al. (2019) construct a prognosis model based on the methylation profiles of 18 CpG that can help to identify new biomarkers, precise drug targets, and molecular subtype classification of COAD patients. Ma et al. (2019) constructed a 10 differentially expressed microRNA prediction model that has high accuracy for OS. In this study, we constructed 12 DNA damage and repair related genes which showed a significant performance for OS prediction in the TCGA cohort and two GSE validation cohort. ROC, DCA, KM, and C-index all proved the 12-gene signature could be an excellent predictor for OS prediction. Meanwhile, we built a nomogram survival model to predict 1/3/5 years survival rate by combining Pathologic M, pathologic T, pathologic N, age, and stage.

There is a point worth making, all the samples included in the TCGA database were COAD, however, the samples in the GEO database include all types of colon cancer and the model constructed by TCGA has 12 genes, while the GEO database only contains 11 of them, which leads to the result that the model has an ideal prediction effect in the train and test groups of TCGA, while the validation effect in GEO is not as good as that in TCGA. We also note that one of the GEO databases only have DFS and DSS information to illustrate our model established by COAD samples. Relapse or tumor-induced death also has a good predictive function, but there is no other corresponding data to verify. In our research, we also refer to a novel web analysis tool suite, TSUNAMI, which can be used for data download, preprocessing and enrichment analysis (Huang et al., 2019).

After reviewing the existing literature, we found that the 12 genes are more or less related to tumors. The cyclin B1-Cdk1 complex is a key regulator of a large number of phosphorylated proteins mitotic entry. Regulation of the mitotic events is linked to activity control of the cyclin B1-Cdk1 complex to make cells enter mitosis, arrest at G2-phase, or skip mitosis (Nakayama and Yamaguchi, 2013). Base excision DNA repair (BER) is the most vital pathway to remove oxidized or mono-alkylated DNA, and APE1 is an important multifunctional enzyme in BER. Oxidative damage induces ISY1 expression. This gene promotes the 5'-3' endonuclease activity of APE1, thereby enhancing the repairability of DNA damage in the cell genome (Jaiswal et al., 2020). Cell Division Cycle 25C (CDC25C) plays an important role in the regulation of G2/M processes and mediates DNA damage repair by checkpoint protein regulation in case of DNA damage. The abnormal expression of cdc25c is related to tumorigenesis and development, and it is a promising therapeutic target (Liu et al., 2020). A large number of mitochondrial DNA (mtDNA) deletion is related to many human diseases and aging. DSB (Double-Strand Breaks) is one of the causes of mtDNA deletion. The exonuclease function of POLG can quickly degrade mtDNA fragments, which minimizes the effect of DSB on mtDNA deletion. The abnormality of POLG will eventually increase the deletion of mtDNA, which has been confirmed in mutant and aging individuals (Nissanka et al., 2018). SMC1B exists in mammalian somatic cells and is related to mitotic cohesion proteins, which help to maintain genome stability and the normal process of gene transcription (Mannini et al., 2015). SMC1B is found to be mutated in UBC and plays an important role in it (van der Lelij et al., 2017). Ras and Rab interactor 2 (RIN2) can associate with GTP-bound Rab5 and take part in early endocytosis (Syx et al., 2010). This gene and SLC22A18, PIGR, and GJA12 can effectively divide Barrett's Esophagus into three groups with different risks and can detect dysplasia/early-stage neoplasia (Alvi et al., 2013).TPM1, as a tumor suppressor gene, was found to be significantly downregulated in colorectal cancer, mainly because of epigenetic and genetic events, which are closely related to the occurrence of colorectal cancer (Mlakar et al., 2009). ELL3 is encoded by an androgen-response gene in the prostate, and it is homologous with ELL and ELL2 (Miller et al., 2000). It was found that the lack of ELL significantly hindered the transcription resumption of RNA Pol II (RNA polymerase II) after DNA repair and increased the RNA Pol II retention to the chromatin, which proved to

be an important member of RNA Pol II restart and participated in the transcription recovery after DNA repair (Mourgues et al., 2013). Through bioinformatics methods, CD36 was found to be associated with lipid metabolism and immune response (Hao et al., 2019), and its high expression was associated with poor prognosis of COAD, and it was found that CD36 was the target of quercetin on COAD (Pang et al., 2019). MC1R is a G-protein-coupled receptor, can cause increased pigmentation, G 1-like cell cycle arrest induced by ultraviolet B, and control senescence and melanoma *in vivo* and *in vitro*, which plays a central role in the prevention of melanoma (Chen et al., 2017). The expression of CCNB3 is usually limited to the testis and encodes a protein with premeiotic function, CyclinB3. CCNB3 can form a fusion gene with BCOR, BCOR-CCNB3, which defines a new subtype of bone sarcoma (Astolfi et al., 2019). NEK4 encodes NIMA-related kinase 4. Inhibition of NEK4 can lead to decreased response to DNA damage and damage the anti-tumor activity of p53. NEK4 is expressed in different stages of CRC, with the highest expression in stage I patients and the lowest expression in stage IV patients. It indicates that a low level of NEK4 is an adverse prognostic factor in CRC patients (Huo et al., 2017). Collectively, we suggested our 12-DNA signature and nomogram could be practical and reliable prognostic tools for COAD. In terms of COAD's overall survival prediction, they can provide higher clinical value than traditional prediction systems and utilize treatment decisions.

Through the gene functional enrichment analysis of 12 genes and their co-expressed genes, we can find that 12 genes are involved in the occurrence and development of COAD by participating in a variety of important biological pathways, meanwhile, through GSEA analysis, we found that there were different pathways in the high- and low-risk group. For example, in the low-risk patient group, it is mainly concentrated in Alzheimers disease, Huntingtons disease, Oocye meiosis, Proteasome, and Tight junction, in which Tight junction is closely related to intestinal inflammation and the occurrence of intestinal tumor (Sharma et al., 2018). The proteasome pathway is widely studied, thanks to the proteasome's ability to control cellular protein quality by degrading misfolded or damaged proteins, which is also key to tumor cell survival (Konstantinopoulos and Papavassiliou, 2006). UPP (The ubiquitin-proteasome pathway) abnormalities play an important role in the occurrence and development of colon cancer. For example, APC (Adenomatous Polyposis Coli) gene mutations in patients with familial adenomatous polyposis syndrome can promote the occurrence of final colon cancer (Konstantinopoulos and Papavassiliou, 2006).

Although the 12-gene signature and nomogram showed excellent performance in the training set and test sets, it had the following defects. First, the gene signature was built with 12-genes but validated by 11-genes in the GEO cohort for the GEO database only contains 11 of them. A relative NRI analysis showed that the 12-gene model performed better than the latter model (**Supplementary Figures 1A–C**). The NRI > 0 for the difference between the two model predictions of the 1, 3, and 5 year survival. This means that the 12-gene model has improved predictive ability compared to the 11-gene model. Meanwhile, though missing a significant gene, the predictive ability for OS, DFS, and DSS of the risk model was significant in the two GEO validation datasets, as we have shown in the results. Second, although the 12-gene signature performed well in predicting the survival of COAD patients, it lacked the verification of large-scale prospective trials. Third, all the samples included in the TCGA database were COAD, while the samples in the GEO database include all types of colon cancer. The TCGA data is gene sequencing data while the GEO data is gene chip data, these differences may mean that the results to come from the validation data may not fully reflect the real prognostic effect of these genes on COAD. And finally, the associated mechanisms had not been validated in COAD cells. Based on this, our follow-up research will focus on verifying the conclusions of this study in terms of clinical application and molecular mechanisms.

In conclusion, we introduced a 12-gene signature which might be an independent prognostic factor in COAD and a novel nomogram that could predict the survival of COAD patients.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

X-qW, S-wX, WW, S-zP, X-bZ, and YW participated in the design of the study and performed the statistical analysis. X-lM, W-dW, L-pY, and S-wL drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by program of Taizhou Science and Technology Grant (20ywb29 and 1802ky09), Medical Health Science and Technology Project of Zhejiang Province (2021PY083, 2020KY1037, and 2019KY239), Key Technology Research and Development Program of Zhejiang Province (2019C03040), and Major Research Program of Taizhou Enze Medical Center Grant (19EZZDA2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.635863/full#supplementary-material

Supplementary Material | The process and results of stepwise.

Supplementary Figure 1 | The NRI and IDI analysis results for the 1year, 3 year, and 5 years survival prediction using a 12 gene model and 11 gene model. **(A)** 1 year, **(B)** 3 year, and **(C)** 5 year.

Supplementary Table 1 | The information of 1545 genes related to DNA damage and repair.

REFERENCES

Alvi, M. A., Liu, X., O'Donovan, M., Newton, R., Wernisch, L., Shannon, N. B., et al. (2013). DNA methylation as an adjunct to histopathology to detect prevalent, inconspicuous dysplasia and early-stage neoplasia in Barrett's esophagus. *Clin. Cancer Res.* 19, 878–888. doi: 10.1158/1078-0432.CCR-12-2880

Astolfi, A., Fiore, M., Melchionda, F., Indio, V., Bertuccio, S. N., and Pession, A. (2019). BCOR involvement in cancer. *Epigenomics* 11, 835–855. doi: 10.2217/epi-2018-0195

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

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625

Chen, L., Fu, L., Kong, X., Xu, J., Wang, Z., Ma, X., et al. (2014). Jumonji domain-containing protein 2B silencing induces DNA damage response via STAT3 pathway in colorectal cancer. *Br. J. Cancer* 110, 1014–1026. doi: 10.1038/bjc.2013.808

Chen, S., Zhu, B., Yin, C., Liu, W., Han, C., Chen, B., et al. (2017). Palmitoylation-dependent activation of MC1R prevents melanomagenesis. *Nature* 549, 399–403. doi: 10.1038/nature23887

Cox, D. (1972). Regression models and life tables. *J. R. Stat. Soc.* 34, 527–541.

Dekker, E., and Rex, D. K. (2018). Advances in CRC prevention: screening and surveillance. *Gastroenterology* 154, 1970–1984. doi: 10.1053/j.gastro.2018.01.069

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Ganesh, K., Stadler, Z. K., Cercek, A., Mendelsohn, R. B., Shia, J., Segal, N. H., et al. (2019). Immunotherapy in colorectal cancer: rationale, challenges and potential. *Nat. Rev. Gastroenterol. Hepatol.* 16, 361–375. doi: 10.1038/s41575-019-0126-x

Gourley, C., Balmaña, J., Ledermann, J. A., Serra, V., Dent, R., Loibl, S., et al. (2019). Moving from poly (ADP-ribose) polymerase inhibition to targeting DNA repair and DNA damage response in Cancer therapy. *J. Clin. Oncol.* 37, 2257–2269. doi: 10.1200/JCO.18.02050

Hao, Y., Li, D., Xu, Y., Ouyang, J., Wang, Y., Zhang, Y., et al. (2019). Investigation of lipid metabolism dysregulation and the effects on immune microenvironments in pan-cancer using multiple omics data. *BMC Bioinform.* 20(Suppl. 7):195. doi: 10.1186/s12859-019-2734-4

Hoeijmakers, J. H. (2009). DNA damage, aging, and cancer. *N. Engl. J. Med.* 361, 1475–1485. doi: 10.1056/NEJMra0804615

Hu, F. C. (2017). My. stepwise: Stepwise variable selection procedures for regression analysis, version 0.1. 0. Available at: https://cran.r-project.org/web/packages/My.stepwise/index.html

Huang, Z., Han, Z., Wang, T., Shao, W., Xiang, S., Salama, P., et al. (2019). TSUNAMI: translational bioinformatics tool suite for network analysis and mining. bioRxiv [preprint]. doi: 10.1101/787507

Huo, T., Canepa, R., Sura, A., Modave, F., and Gong, Y. (2017). Colorectal cancer stages transcriptome analysis. *PLoS One* 12:e0188697. doi: 10.1371/journal.pone.0188697

Jaiswal, A. S., Williamson, E. A., Srinivasan, G., Kong, K., Lomelino, C. L., McKenna, R., et al. (2020). The splicing component ISY1 regulates APE1 in base excision repair. *DNA Repair (Amst)* 86:102769. doi: 10.1016/j.dnarep.2019.102769

Jeggo, P. A., Pearl, L. H., and Carr, A. M. (2016). DNA repair, genome stability and cancer: a historical perspective. *Nat. Rev. Cancer* 16, 35–42. doi: 10.1038/nrc.2015.4

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Karpov, D. S., Spirin, P. V., Zheltukhin, A. O., Tutyaeva, V. V., Zinovieva, O. L., Grineva, E. N., et al. (2020). LINC00973 induces proliferation arrest of drug-treated cancer cells by preventing p21 degradation. *Int. J. Mol. Sci.* 21:8322. doi: 10.3390/ijms21218322

Khanna, A. (2015). DNA damage in cancer therapeutics: a boon or a curse? *Cancer Res.* 75, 2133–2138. doi: 10.1158/0008-5472.CAN-14-3247

Kleinbaum, D. G. (1998). *Survival analysis, a self-learning text*. Dordrecht, Heidelberg, London, and New York: Springer.

Kobayashi, H., Mochizuki, H., Sugihara, K., Morita, T., Kotake, K., Teramoto, T., et al. (2007). Characteristics of recurrence and surveillance tools after curative resection for colorectal cancer: a multicenter study. *Surgery* 141, 67–75. doi: 10.1016/j.surg.2006.07.020

Konstantinopoulos, P. A., and Papavassiliou, A. G. (2006). The potential of proteasome inhibition in the treatment of colon cancer. *Expert Opin. Investig. Drugs* 15, 1067–1075. doi: 10.1517/13543784.15.9.1067

Lei, L., Zhao, X., Liu, S., Cao, Q., Yan, B., and Yang, J. (2019). MicroRNA-3607 inhibits the tumorigenesis of colorectal cancer by targeting DDI2 and regulating the DNA damage repair pathway. *Apoptosis* 24, 662–672. doi: 10.1007/s10495-019-01549-5

Liu, K., Zheng, M., Lu, R., Du, J., Zhao, Q., Li, Z., et al. (2020). The role of CDC25C in cell cycle regulation and clinical cancer therapy: a systematic review. *Cancer Cell Int.* 20:213. doi: 10.1186/s12935-020-01304-w

Lord, C. J., and Ashworth, A. (2017). PARP inhibitors: synthetic lethality in the clinic. *Science* 355, 1152–1158. doi: 10.1126/science.aam7344

Ma, R., Zhao, Y., He, M., Zhao, H., Zhang, Y., Zhou, S., et al. (2019). Identifying a ten-microRNA signature as a superior prognosis biomarker in colon adenocarcinoma. *Cancer Cell Int.* 19:360. doi: 10.1186/s12935-019-1074-9

Mannini, L., Cucco, F., Quarantotti, V., Amato, C., Tinti, M., Tana, L., et al. (2015). SMC1B is present in mammalian somatic cells and interacts with mitotic cohesin proteins. *Sci. Rep.* 5:18472. doi: 10.1038/srep18472

Mauri, G., Arena, S., Siena, S., Bardelli, A., and Sartore-Bianchi, A. (2020). The DNA damage response pathway as a land of therapeutic opportunities for colorectal cancer. *Ann. Oncol.* 31, 1135–1147. doi: 10.1016/j.annonc.2020.05.027

Miller, T., Williams, K., Johnstone, R. W., and Shilatifard, A. (2000). Identification, cloning, expression, and biochemical characterization of the testis-specific RNA polymerase II elongation factor ELL3. *J. Biol. Chem.* 275, 32052–32056. doi: 10.1074/jbc.M005175200

Mlakar, V., Berginc, G., Volavsek, M., Stor, Z., Rems, M., and Glavac, D. (2009). Presence of activating KRAS mutations correlates significantly with expression of tumour suppressor genes DCN and TPM1 in colorectal cancer. *BMC Cancer* 9:282. doi: 10.1186/1471-2407-9-282

Moody, L., He, H., Pan, Y. X., and Chen, H. (2017). Methods and novel technology for microRNA quantification in colorectal cancer screening. *Clin. Epigenetics* 9:119. doi: 10.1186/s13148-017-0420-9

Mourgues, S., Gautier, V., Lagarou, A., Bordier, C., Mourcet, A., Slingerland, J., et al. (2013). ELL, a novel TFIIH partner, is involved in transcription restart after DNA repair. *Proc. Natl. Acad. Sci. U. S. A.* 110, 17927–17932. doi: 10.1073/pnas.1305009110

Mouw, K. W., Goldberg, M. S., Konstantinopoulos, P. A., and D'Andrea, A. D. (2017). DNA damage and repair biomarkers of immunotherapy response. *Cancer Discov.* 7, 675–693. doi: 10.1158/2159-8290.CD-17-0226

Nakayama, Y., and Yamaguchi, N. (2013). Role of cyclin B1 levels in DNA damage and DNA damage-induced senescence. *Int. Rev. Cell Mol. Biol.* 305, 303–337. doi: 10.1016/B978-0-12-407695-2.00007-X

Nissanka, N., Bacman, S. R., Plastini, M. J., and Moraes, C. T. (2018). The mitochondrial DNA polymerase gamma degrades linear DNA fragments precluding the formation of deletions. *Nat. Commun.* 9:2491. doi: 10.1038/s41467-018-04895-1

O'Connell, J. B., Maggard, M. A., and Ko, C. Y. (2004). Colon cancer survival rates with the new American joint committee on Cancer sixth edition staging. *J. Natl. Cancer Inst.* 96, 1420–1425. doi: 10.1093/jnci/djh275

Pang, B., Xu, X., Lu, Y., Jin, H., Yang, R., Jiang, C., et al. (2019). Prediction of new targets and mechanisms for quercetin in the treatment of pancreatic cancer, colon cancer, and rectal cancer. *Food Funct.* 10, 5339–5349. doi: 10.1039/C9FO01168D

Ranstam, J., and Cook, J. A. (2017). Kaplan-Meier curve. *Br. J. Surg.* 104:442. doi: 10.1002/bjs.10238

Scagliarini, A., Mathey, A., Aires, V., and Delmas, D. (2020). Xanthohumol, a Prenylated flavonoid from hops, induces DNA damages in colorectal Cancer cells and sensitizes SW480 cells to the SN38 chemotherapeutic agent. *Cell* 9:932. doi: 10.3390/cells9040932

Sharma, D., Malik, A., Guy, C. S., Karki, R., Vogel, P., and Kanneganti, T. D. (2018). Pyrin Inflammasome regulates tight junction integrity to restrict colitis and tumorigenesis. *Gastroenterology* 154, 948.e8–964.e8. doi: 10.1053/j.gastro.2017.11.276

Stadler, Z. K. (2015). Diagnosis and management of DNA mismatch repair-deficient colorectal cancer. *Hematol. Oncol. Clin. North Am.* 29, 29–41. doi: 10.1016/j.hoc.2014.09.008

Sun, J., Wang, C., Zhang, Y., Xu, L., Fang, W., Zhu, Y., et al. (2019). Genomic signatures reveal DNA damage response deficiency in colorectal cancer brain metastases. *Nat. Commun.* 10:3190. doi: 10.1038/s41467-019-10987-3

Sveen, A., Kopetz, S., and Lothe, R. A. (2020). Biomarker-guided therapy for colorectal cancer: strength in complexity. *Nat. Rev. Clin. Oncol.* 17, 11–32. doi: 10.1038/s41571-019-0241-1

Syx, D., Malfait, F., Van Laer, L., Hellemans, J., Hermanns-Le, T., Willaert, A., et al. (2010). The RIN2 syndrome: a new autosomal recessive connective tissue disorder caused by deficiency of Ras and Rab interactor 2 (RIN2). *Hum. Genet.* 128, 79–88. doi: 10.1007/s00439-010-0829-0

van der Lelij, P., Lieb, S., Jude, J., Wutz, G., Santos, C. P., Falkenberg, K., et al. (2017). Synthetic lethality between the cohesin subunits STAG1 and STAG2 in diverse cancer contexts. *eLife* 6:e26980. doi: 10.7554/eLife.26980

Vickers, A. J., and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis. Mak.* 26, 565–574. doi: 10.1177/0272989X06295361

Yang, C., Zhang, Y., Xu, X., and Li, W. (2019). Molecular subtypes based on DNA methylation predict prognosis in colon adenocarcinoma patients. *Aging (Albany NY)* 11, 11880–11892. doi: 10.18632/aging.102492

Yu, X., Li, W., Liu, H., Deng, Q., Wang, X., Hu, H., et al. (2020). Ubiquitination of the DNA-damage checkpoint kinase CHK1 by TRAF4 is required for CHK1 activation. *J. Hematol. Oncol.* 13:40. doi: 10.1186/s13045-020-00869-3

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 26 February 2021
doi: 10.3389/fgene.2021.629946

Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression

*Qin Jiang and Min Jin**

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

Exploring the molecular mechanisms of breast cancer is essential for the early prediction, diagnosis, and treatment of cancer patients. The large scale of data obtained from the high-throughput sequencing technology makes it difficult to identify the driver mutations and a minimal optimal set of genes that are critical to the classification of cancer. In this study, we propose a novel method without any prior information to identify mutated genes associated with breast cancer. For the somatic mutation data, it is processed to a mutated matrix, from which the mutation frequency of each gene can be obtained. By setting a reasonable threshold for the mutation frequency, a mutated gene set is filtered from the mutated matrix. For the gene expression data, it is used to generate the gene expression matrix, while the mutated gene set is mapped onto the matrix to construct a co-expression profile. In the stage of feature selection, we propose a staged feature selection algorithm, using fold change, false discovery rate to select differentially expressed genes, mutual information to remove the irrelevant and redundant features, and the embedded method based on gradient boosting decision tree with Bayesian optimization to obtain an optimal model. In the stage of evaluation, we propose a weighted metric to modify the traditional accuracy to solve the sample imbalance problem. We apply the proposed method to The Cancer Genome Atlas breast cancer data and identify a mutated gene set, among which the implicated genes are oncogenes or tumor suppressors previously reported to be associated with carcinogenesis. As a comparison with the integrative network, we also perform the optimal model on the individual gene expression and the gold standard PMA50. The results show that the integrative network outperforms the gene expression and PMA50 in the average of most metrics, which indicate the effectiveness of our proposed method by integrating multiple data sources, and can discover the associated mutated genes in breast cancer.

Keywords: breast cancer, machine learning, classification, feature selection, gradient boosted decision tree

OPEN ACCESS

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

Reviewed by:
Min Chen,
Hunan Institute of Technology, China
Wei Liu,
Xiangtan University, China

***Correspondence:**
Min Jin
jinmin@hnu.edu.cn

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

Received: *16 November 2020*
Accepted: *21 January 2021*
Published: *26 February 2021*

Citation:
Jiang Q and Jin M (2021) Feature Selection for Breast Cancer Classification by Integrating Somatic Mutation and Gene Expression. Front. Genet. 12:629946. doi: 10.3389/fgene.2021.629946

INTRODUCTION

Breast cancer is considered to be the most prevalent cancer among women and the second common cause of death in both developed and undeveloped countries. It is caused by multiple factors including genomic, transcriptomic, and epigenomic involvement in its formation and development. With the development of technology, understanding the pathogenesis of cancer from

the perspective of molecular contributes to effective diagnosis and treatment. The large-scale cancer genomics project, The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015), has produced a large volume of data, providing ways to explore cancer formation and progression.

In general, the cancer transcriptome contains gene expression, including messenger RNA (mRNA), long non-coding RNA (lncRNA), and microRNA (miRNA). Previous studies focused on utilizing the gene expression profile to successfully diagnose individuals based on the differential gene expression (Li et al., 2017) and other clinically relevant phenotypes. Meanwhile, the cancer genome contains many mutations. Among them, one of the most important is somatic mutations, which include single-nucleotide variant (SNVs) and small insertions and deletions (indels). Some mutations that contribute to cancer progression from normal to malignant are called driver mutations, and others that accumulate in cells but do not contribute to cancer development are called passengers (Bozic et al., 2010). Distinguishing driver mutations from the passengers that have no critical effect on cancer cells is a crucial step and challenging task in understanding the molecular mechanisms of cancer, which can guide effective treatment and prognosis for cancer patients and promote the development of targeted drugs. In earlier studies, researchers focused on detecting driver genes that cause tumors (Merid et al., 2014). A common approach is to identify driver genes by detecting positive signals in tumors. Because of the complexity of the cancer genome, driver genes contain not only driver mutations but also passenger mutations. This makes this kind of approach sometimes ineffective.

On the other hand, studies have shown that somatic mutations frequently perturb the expression level of affected genes and thus disrupt the pathways controlling normal growth (Kwong et al., 2020). For example, mRNAs carrying a premature stop codon, which can be introduced by truncation mutations, are typically eliminated by the process called nonsense-mediated mRNA decay, and thus, both the concentration of mRNA transcripts and protein products would be decreased owing to truncation mutations (Jia and Zhao, 2016). Considering the association between the somatic mutation and gene expression, several studies have emphasized the necessity of integrating both types of data to identify candidate driver genes (Masica and Karchin, 2011; Zhang and Wang, 2020). For cancer analysis, many researchers construct a co-expression network by integrating different types of data. He et al. (2017) and Wu et al. (2019) utilized the network by integrating somatic mutation with gene expression to identify the type of cancers and cancer subtypes. Mamidi et al. (2019) integrated germline and somatic mutation to discover biomarkers in triple-negative breast cancer and identified the molecular networks and biological pathways.

As the molecular network has been verified to be effective for the biological discovery of cancers, current studies utilized the network across different types of cancer or cancer subtypes. However, the objective of most researches is the universality of the methods, which makes it difficult to be equally effective in all disease types. In this study, we aim to construct an efficient method of architecture for the diagnosis of breast cancer based on the network of somatic mutation and gene expression. We are focused not only on finding more biomarkers but also on the classification performance of the model. First, the somatic mutation is used to generate a binary mutation network; similarly, an expression network is obtained from the gene expression profiles. Then, for the expression network, we compute both the observed p-value and the adjusted p-value to correct for multiple-hypothesis testing (false discovery rate, FDR) and thus obtain the differential expression network. Meanwhile, an integrative network is constructed by combining the mutation network and the differential expression network. Thirdly, we rank the genes in the integrative network by mutual information (MI) and select the top 50 genes, which are highly correlated with breast cancer. Finally, we use the Bayesian optimization method to optimize the classification model, gradient boosting decision tree (GBDT), which is further applied to assess the features selected from the previous step. In terms of evaluation metrics, the traditional metric of accuracy does not consider the sample imbalance, so we propose a simple and effective metric, balanced accuracy, to reveal the ability of the different model to classify positive and negative samples.

MATERIALS AND METHODS

We used statistical and machine learning methods to develop this novel method for feature selection and classification, including the preprocessing of data, filter method, and embedded method for feature selection, processing of imbalanced data, and the final classification model. **Figure 1** shows the flowchart of the proposed method.

Dataset Construction and Preprocessing

In this research, we use publicly available breast cancer datasets (BRCA) from TCGA, including transcriptome gene expression and somatic mutation. Considering the different structures of these two types of data, we used different methods to preprocess them. **Table 1** shows the numbers of samples and features for the two datasets.

The BRCA gene expression dataset comprises 1222 samples and 57,063 genes. There are 113 normal samples and 1109 tumor samples. We used the edgeR package to filter the genes expressed in small amounts in most samples and normalized the data. The gene expression data was reduced from 57,063 to 34,465 by deleting the genes expressed in small amounts in most samples.

The somatic mutation data comes from the simple nucleotide variation (SNV) in the TCGA-BRCA project. The data file includes SNP, INS, and DEL, three types of mutations. The important fields in the data file are Hugo_Symbol (gene name), Variant_Type, and Tumor_Sample_Barcode (sample name). Statistically, the somatic mutation data contains 18,127 genes and 986 samples. To get the mutation frequency of each gene in all samples, we use a Perl script to process the data file. For example, if gene A is present in sample S, that means sample S has a mutation in gene A, then we code it as "1," otherwise we code

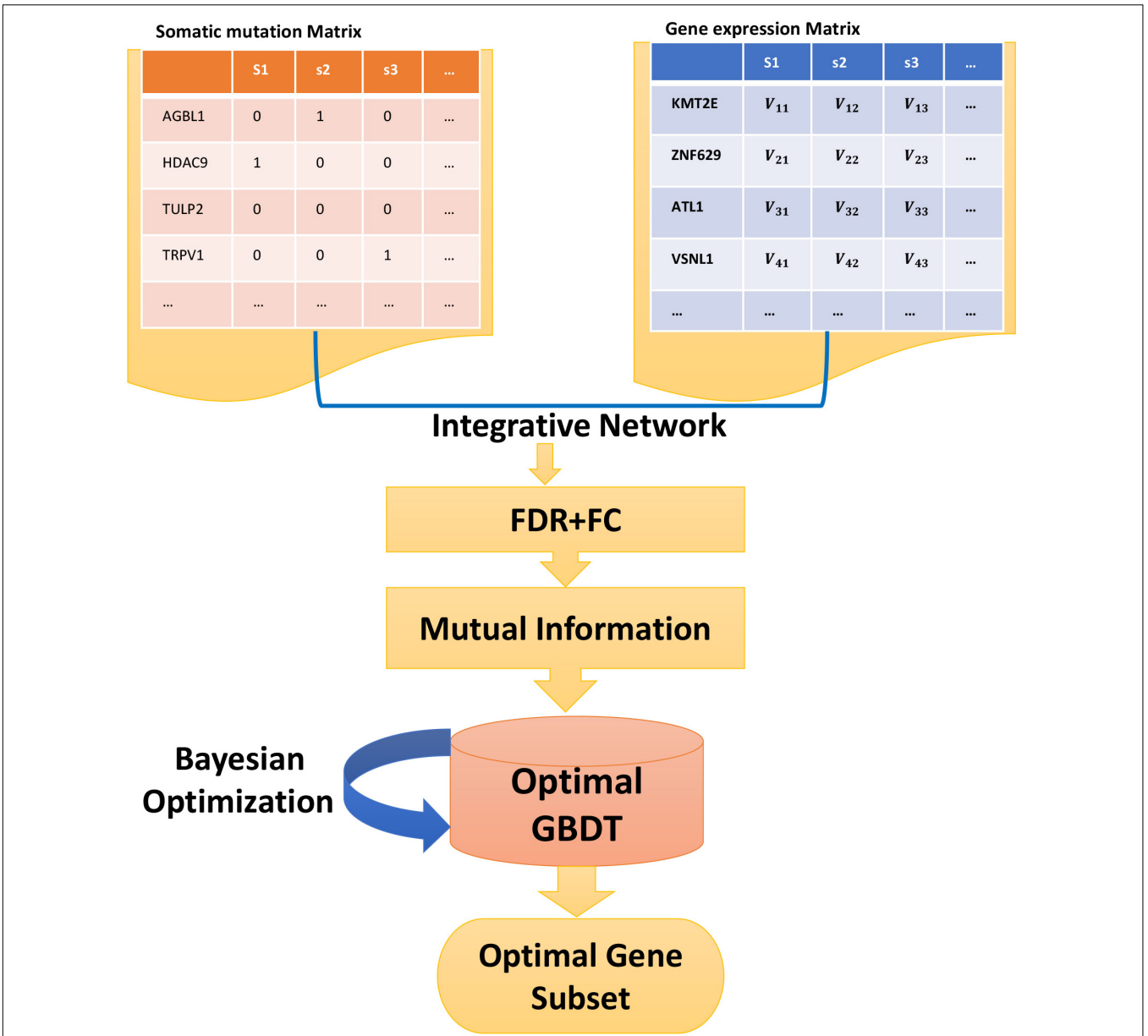

FIGURE 1 | The flowchart of the proposed method. From the somatic mutation matrix, the mutation frequency of each gene is obtained to select the highly mutated genes, which will be integrated by mapping into the gene expression matrix to get the integrative network. After FDR, FC, and mutual information ranking, the feature genes serve as the input to GBDT. Then, the optimal model is obtained by Bayesian optimization. During the training process, the optimal gene subset is obtained simultaneously.

it as "0." **Supplementary Table 1** shows the coding schedule of all genes in samples. Given the sample set S = {s_1, s_2, . . .s_n}, n is the total number of samples, and s_i represents the sample i. Gene set G = {g_1, g_2, . . .g_m}, m is the total number of mutation genes, and g_j represents the gene j. In the set of sample number C = {c_1, c_2, . . .c_m}, c_k represents the number of samples with "1" in gene k. The set C can be calculated by the number of "1" in each row in **Supplementary Table 1**.

According to **Supplementary Table 1** and set C, we can obtain the frequency of mutations across patients to assess the percentage of patients carrying a particular mutation in each mutated gene. To further reduce the interference of genes with low mutation rates, we set the threshold p as the percentage of the total samples to select the genes with high mutation frequency. The selected gene set constitutes the mutation network. In the experiment, we compare the effects of different p on classification accuracy by the proposed model, and the result is shown in **Supplementary Table 2**. Due to the highest accuracy 97.31% obtained by setting the threshold p as 1%, we apply this value in the proposed method.

The Way to Combine Somatic Mutation and Gene Expression

Somatic mutations in cancer genomes frequently perturb the expression level of affected genes. Then, the pathways controlling normal growth are disrupted (Zhang et al., 2013). Similarly, the research by Ding et al. (2015) assessed the impact of mutations on gene expression as a means of quantifying potential phenotypic effects and for novel cancer gene discovery. Fleck et al. (2016) addressed the issue of cancer heterogeneity by using both somatic mutation and gene expression data and proposed a formulation to model the molecular progression of cancer. They discovered that the progression of the disease was reflected in both the accumulation of mutations and changes in gene expression levels. Further study (Jia and Zhao, 2016) focused on the functional footprints of somatic mutations in 12 cancer types and grouped the mutations by mutation type, cluster, and status. This study unraveled the effects of somatic mutation features on mRNA and protein expression.

Our study is based on the assumption that mutations may cause changes in the cell's state, such as underexpression or overexpression of different genes. Then, we combine the somatic mutation network with the gene expression network to obtain an integrative network. In the integrative process of the two types of networks, we refer to the gene expression network to obtain the expression value of the somatic mutation genes in the mutation network. It is important to note that in the subsequent classification task, the normal samples in the expression network are added as the control group.

Fold Change and False Discovery Rate

Fold change (FC) is used to calculate the differential multiples of gene expression values between cancer samples and normal samples, which is the basic method for detecting differential genes, and represents the expression values of feature i and sample j in cancer samples and normal samples; FC is defined as:

$$FC_i = \frac{\bar{X}_i}{\bar{Y}_i}. \tag{1}$$

When FC exceeds the initial set threshold, it can be considered that the feature is different, and it is generally considered that there is a significant difference when the difference multiple is more than 2. FC can directly obtain the differentially expressed values, but in the absence of false-positive control, the rate of false-positive results is relatively high.

According to statistical theory, in multiple-hypothesis testing, it is important to control the probability of making mistakes in multiple statistical inferences, called FDR. FDR can be used to analyze deferentially expressed genes to control the proportion of false positives (Reiner-Benaim, 2010). **Table 2** shows the confusion matrix for the statistical test. FDR can be defined as follows:

$$FDR = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right)(R > 0). \tag{2}$$

The number of false positives in multiple-hypothesis tests can be controlled by controlling that FDR is below the threshold q. In general, keep FDR below 0.01, or ensure that there is at most one false positive for every 100 positive hypotheses. Feature genes with significant differences can be identified by FC and FDR, but these two methods do not evaluate the classification performance of these features.

Fold change and FDR are applied to integrative data to select the differentially expressed genes. By comparing the classification balanced accuracy under different FC and FDR thresholds shown in **Supplementary Tables 3, 4**, the optimal value of FC and FDR thresholds is obtained: log (FC) > 1.0, FDR < 0.05.

Mutual Information

Mutual information (Bonev et al., 2008) is a useful measure of information in information theory and is a kind of filter method. It refers to the correlation between two events set. The datasets consist of tens of thousands of gene columns and one label column. The gene column is defined as G_i, and the label column is defined as L. $MI(G_i, L)$ is represented as the MI between the gene G_i and the label L. The calculation equation is Eq. 3.

$$MI(G_i, L) = H(G_i) + H(L) - H(G_i, L) \tag{3}$$

$H(G_i)$ is the information entropy of the gene column G_i, $H(L)$ is the information entropy of the label L, and $H(G_i, L)$ is the joint information entropy of G_i and L. According to information theory, the information entropy is a measure of the uncertainty of a random variable. Suppose X is a random variable, and the range of possible values is S_x, $x \in S_x$ and the probability is $p(x)$; the information entropy of X is defined as:

$$H(X) \quad -\sum_{x \in S_x} p(x) \log p(x) \tag{4}$$

TABLE 1 | Confusion matrix for statistical tests.

	H_0 is true	H_1 is true	Total
Significant	V	S	R
Not significant	U	T	m-R
Total	m_0	m-m_0	m

H_0 is the null hypothesis, H_1 is the alternative hypothesis or reject null hypothesis. m is the number of hypothesis tests. m_0 is the number of null hypotheses that are true. m-m_0 is the number of alternative hypothesis that are true. V is the number of false-positive cases. S is the number of the true positive cases. U is the number of true negative cases. T is the number of false negative cases. R = V + S is the number of rejected hypotheses. FDR = E(V/R).

TABLE 2 | The optimal parameters for each step in the proposed method.

Parameter	p	FDR	| log(FC)|	*M*
Threshold	1%	0.05	1	50

p is the percentage of the total samples, which represents the mutation frequency of a certain gene. FDR is the false discovery rate, and FC is the fold change. M is the number of genes that top ranking in mutual information.

$H(X, Y)$ is the joint information entropy, defined as:

$$H(X, Y) \; -\sum_{x\in S_x}\sum_{y\in S_y} p(x, y)\, logp(x, y) \tag{5}$$

$p(x, y)$ is the joint probability density function. $MI(G_i, L)$ can be calculated according to Eqs 4 and 5. In our study, MI is used to measure the dependency between a feature and the classification type. In general, the greater value of MI indicates that the feature contains more information for classification. Therefore, we rank the MI values of each feature and selected the top M features from the integrative data, respectively. The final objective of this method is to remove irrelevant features to reduce the dimension of integrative data. We set different values of M to compare the classification-balanced accuracy and obtain the best value of M. The result in **Supplementary Table 5** shows that the optimal M is 50. **Table 3** shows the main parameters applied in the proposed method.

GBDT With Bayesian Optimization

The filter methods obtain a feature subset for which the discriminative capability is limited for classification purposes. Embedded methods can be used to search the optimal feature subset by a given classifier. In the training procedure, the features with high importance can be selected by ranking and the classification algorithm is optimized simultaneously. It is helpful to build a strong link between the feature subset and the classifier. The GBDT is an ensemble learning algorithm based on GBM, which is proposed by Friedman (Friedman, 2001). During the training process, multiple iterations are used to build multiple trees to make joint decisions. When the square error loss function is adopted, each regression tree learns the conclusions and residuals of all previous trees, and a current residual regression tree is obtained by the fitting. The meaning of residuals is as follows:

$$residuals = true\ value - predict\ value$$

The boosting tree (Galicia et al., 2018) is an accumulation of regression trees generated during the entire iteration process. The optimization process of learning is realized by using an additive model and a forward step algorithm. The GBDT was used in our study because of its flexibility for different types of data, excellent classification performance, and robustness for abnormal values.

TABLE 3 | Classification accuracy and balanced accuracy of proposed method.

Case	Testing accuracy	Testing balanced accuracy	Running time
1	0.9796	0.8547	65.2642
2	0.9878	0.9111	20.7672
3	0.9878	0.9255	0.2187
4	**0.9951**	**0.9731**	**0.1925**

The method in case 1 without using any feature selection and the accuracy is the lowest and is time-consuming. In case 2, using FC + FDR to select differentially expressed genes, the results are improved by 0.84 and 6.6%. In case 3, using FC + FDR and MI to select the key 50 features, 1.58% improvement in the balanced accuracy and a significant reduction in running time are obtained. The proposed method shown in case 4, the best performance in the three metrics is obtained. The bold values are the best results.

However, it is tedious and important work to tune the hyperparameters when conducting the GBDT, because it greatly affects the performance of the algorithm. Manual tuning is time-consuming; grid and random searches (Bhat et al., 2018) require no human effort but a long-running time. Therefore, in this research, Bayesian optimization is adopted to find the optimal hyperparameters, which is first proposed by Snoek et al. (2012). Bayesian optimization seeks to minimize the value of the objective function by establishing an alternative function based on the objective function's past evaluation results. The Bayesian method is different from random or grid searches as they consider previous estimates when testing the next set of hyperparameters, thus saving a lot of effort.

Suppose hyperparameters set (represents a hyperparameter's value), the relationship between this set, and the loss function that need to be optimized, defined as $f(X)$. However, machine learning just likes a black box, which means we only know the input and output; f is hard to be sure. So we should turn our attention to a function that can be solved. Assume function, we need to find in:

$$x^* = arg\ min_{x\in X} f(x) \tag{6}$$

Here, we chose Hyperopt in Python library, which adopted Tree Parzen Estimator (TPE), which used the Gaussian Mixture Model (Oh et al., 2019) to learn hyperparameters. First, we split the integrative dataset into 80% learning set and 20% test set then divided the learning set into 60% training set and 40% validation set. The performance of hyperparameters was evaluated on the validation set. The Bayesian optimization assigned a greater probability to the value of the hyperparameters set with a lower loss in the cross-validation. Finally, the best hyperparameters set was output.

A Weighted Metric for Imbalanced Dataset

Class imbalance is a situation in which the number of training samples of different categories varies greatly in the classification task. There are many strategies to deal with the imbalance problem, such as undersampling and oversampling. EasyEnsemble is a method of undersampling, proposed by Li and Liu (2014). Multiple different training sets are generated by putting back the samples several times, and then multiple different classifiers are trained. The final result is obtained by combining the results of multiple classifiers. Another method is BalanceCascade (Liu et al., 2009), which adopts the idea of Boosting. It also uses undersampling to generate a training set, but those correctly classified samples are not put back. Undersampling is easy to lose information, and the way the final result is integrated also has an impact. The most common strategy for oversampling is SMOTE (Synthetic Minority Oversampling Technique) (Blagus and Lusa, 2013). In this method, the new samples are synthesized according to the nearest neighbor in the minority samples and then added into the dataset. However, there two main problems in this algorithm: there is some blindness in the selection of the nearest neighbor and the problem of distribution marginalization is easy to occur. Additionally, undersampling and oversampling may change the distribution

of data. For the task of cancer classification, the size of sample is small, more than a thousand at most, and these strategies do not seem appropriate. Therefore, in this study, we propose a weighted metric to modify the traditional accuracy metric instead of changing the distribution of the dataset. There are far more cancer samples than normal samples, which will lead to the high accuracy of the learning method if it returns a learning model that always predicts the new sample as a cancer category. To solve this problem, we separated the total sample set into a normal set and tumor set. The classification accuracy of the model in the two-sample space embodies the model's ability to correctly classify the positive and negative samples, named the weight for the two-sample spaces. On the final test stage, we multiply this weight with the accuracy of two sample spaces on the test set.

Let N and T denote the sample set of normal class and that of tumor class, respectively. $\overrightarrow{w_n}$ and $\overrightarrow{w_t}$ are the accuracy of normal samples and tumor samples of classifier *clf* in the validation set, respectively. These two weights represent the different capacities of the given classifier for different types of samples. In the final testing stage, the optimized GBDT is conducted as the classifier to predict the independent test set; $\overrightarrow{w_n}$ and $\overrightarrow{w_t}$ will be considered in the final decision. As we split the dataset into 10 equal-sized datasets, $\overrightarrow{w_n}$ and $\overrightarrow{w_t}$ are the average accuracy of the 10 validation sets. Here, the average accuracy of normal samples and tumor samples on the 10 test sets are represented by acc_n and acc_t. So the final balanced accuracy is defined as:

$$balanced\ acc = acc_n \cdot \overrightarrow{w}_n + acc_t \cdot \overrightarrow{w}_t \tag{7}$$

The core procedure of calculating the weighted metric for balanced accuracy is described in **Figure 2**. The weighted metric for the imbalanced dataset is easy to operate. It considers the classification ability of the classifier on samples of different categories and further revises the final test results by multiply weights, thus reducing the impact of class imbalance.

Evaluation Criteria

The following metrics are used to evaluate the performance of the classification model in this study:

Accuracy: $ACC = \frac{TP+TN}{TP+FP+TN+FN}$

Sensitivity: $SES = \frac{TP}{TP+FN}$

Specificity: $SPC = \frac{TN}{TN+FP}$

Precision: $PRC = \frac{TP}{TP+FP}$

F_1 score: $F_1 = \frac{2TP}{2TP+FP+FN}$

In this study, the tumor sample is positive, and the normal sample is the negative sample, where TP (true positive) is the number of tumor samples predicted as tumor, FP (false positive) is the number of tumor samples predicted as normal, TN (true negative) is the number of tumor samples normal and predicted as normal, and FN (false negative) is the number of normal samples and predicted as tumor. Meanwhile, the AUC is obtained.

Due to that the number of samples is much smaller than that of the features, in this study, first, we split the dataset into 10 equal-sized datasets. Then, we divide the datasets into 80% learning set and 20% test set and ensure that the test set does not participate in any training process (Meng et al., 2020). Finally, the independent test set is used to calculate the above evaluation metrics. This procedure is repeated on the 10 datasets. The average of the results generated on the 10 datasets is used as the final performance of the proposed model on the test set.

RESULTS

Classification Results of Proposed Method SFS

In our experiments, the training set is used to train the classifier. The obtained parameters are verified on the validation set. In addition, we calculate $\overrightarrow{w_n}$ and $\overrightarrow{w_t}$ (normal samples' accuracy and tumor samples' accuracy in the validation set). Moreover, balanced accuracy was calculated by Eq. 6. The proposed method adopts FC, FDR, MI, and GBDT with Bayesian optimization. The parameters are applied as follows:

(1) FC: $|log(FC)| > 1.0$

(2) FDR: FDR 0.05

(3) MI: select the top 50 features of MI value ranking

(4) Bayesian optimization: tuning the parameters of GBDT with Bayesian optimization using the 50 features to get the optimal model.

These methods are combined in the ways shown in **Table 4**.

Case 1: None of the above methods are used.

Case 2: FC and FDR are used to obtain the differentially expressed genes.

Case 3: FC + FDR, MI are used to select informative features.

Case 4: FC + FDR, MI, and Bayesian optimization are adopted to optimize GBDT, and this case is the proposed method.

The testing accuracy is obtained by the classifier GBDT on the independent test set. The results shown in case 1 are the classification accuracy using GBDT without any feature selection. It can be observed that the GBDT without any feature selection obtains a testing accuracy of about 97.96%, but the testing balanced accuracy is only about 85.47%, which implied the learning efficiency of the GBDT without feature selection is not much high. In case 2, although FC and FDR effectively reduce the running time, it does not improve the accuracy significantly, because they ignore the correlation between features. In case 3, we add MI to further select key features, and the results show that there is an improvement (1.58%) in balance accuracy and a significant reduction in running time. In case 4, we use Bayesian optimization to optimize GBDT to obtain the optimal model. According to the results, we conclude that the accuracy and balanced accuracy are improved by 0.74 and 5.14%, which were compared with case 3. Particularly, the proposed method shown in case 4 obtains the highest testing accuracy and balanced accuracy. The performance of testing balanced accuracy is

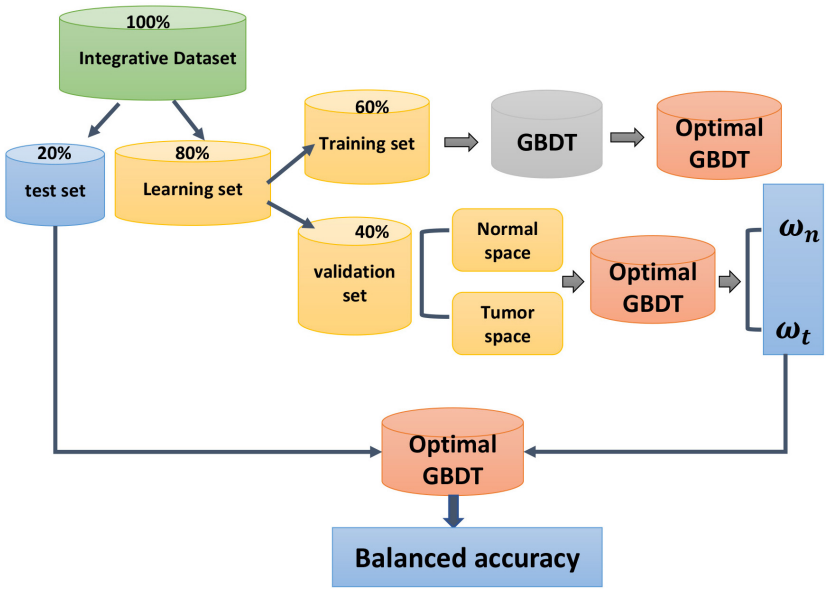

FIGURE 2 | The calculation procedure of balanced accuracy. The raw dataset is split into 10 equal datasets. The diagram shows the procedure on one of the 10 datasets. First, the integrative dataset is derived into a learning set and an independent test set. The learning set is derived into a training set and a validation set. The training set is used to train the GBDT model and the validation set is used to obtain the weight for normal and tumor space (ω_n and ω_t), which is represented by the accuracy of normal and tumor space. Finally, when the optimal model is tested on the test set, ω_n and ω_t will be used to modify the final accuracy to obtain the balanced accuracy.

TABLE 4 | The mean values of seven evaluation metrics obtained from four methods on integrative dataset.

Classifier	B_ACC	ACC	SES	SPC	PRC	F1	AUC
SVM	0.9413	0.9865	0.9910	0.9435	0.9941	0.9926	0.9672
RF	0.9208	0.9902	**0.9968**	0.9261	0.9924	0.9946	0.9615
KNN	0.9480	0.9914	0.9955	0.9522	0.9950	0.9953	0.9738
Proposed	**0.9731**	**0.9951**	0.9964	**0.9826**	**0.9982**	**0.9973**	**0.9895**

In the experiments, we randomly split the dataset into 10 equal-sized datasets. The mean values of the seven metrics are obtained on the 10 test sets. The proposed method outperforms other methods in balanced accuracy, accuracy, specificity, precision, F1 score, and AUC. The bold values are the best results.

improved by 13.85%, compared with the method in case 1. From the perspective of vertical comparison, the features selected by the proposed method have better classification performance. From the perspective of horizontal comparison, balanced accuracy improves more than traditional accuracy, which indicates that the proposed model shows greater advantages when the sample balance is considered.

The Hyperparameters of GBDT Adjusted by Bayesian Optimization

Bayesian optimization aims to find the minimum value of the objective function by establishing a proxy function (probabilistic model). The proxy function is easier to optimize than the objective function (Victoria and Maragatham, 2020), so the next input value to be evaluated is selected by applying some criterion. For hyperparameter optimization, the objective function is the validation error of the machine learning model using a set of hyperparameters. Its goal is to find the hyperparameters that produce the minimum error on the validation set and to generalize these results to the test set. The cost of evaluating an objective function is significant because it requires the training of a machine learning model with a specific set of hyperparameters. Bayesian hyperparameter tuning uses a constantly updated probabilistic model to "focus" the search process on the hyperparameters that are likely to be optimal by reasoning from past results. In this study, for the objective function, the input was a set of hyperparameters, and the output was the fivefold cross-validation loss with classifier GBDT. We chose Tree Parzen Estimation (TPE) as the optimization algorithm. **Figure 3** shows the best sets of hyperparameters obtained by Bayesian optimization and random search with 300 iterations. The balanced accuracy gained on the test set by using the best two sets of hyperparameters in GBDT was 97.31 and 96.8%, respectively. The results indicated that Bayesian optimization outperforms random search in the respect of hyperparameter tuning.

In the comparative experiments, we select three other classifiers, SVM, KNN, and RF. **Supplementary Tables 6, 7** and **Supplementary Figure 2** show the procedure of tuning parameters for the three classifiers. According to the balanced accuracy obtained in those models, the optimal parameters are as follows:

(1) SVM: C = 1, kernel = "linear"
(2) KNN: n_neighbor = 7, metric = "manhattan"
(3) RF: max_depth: 46, min_sample_leaf: 2, min_sample_split: 94, n_estimators: 75

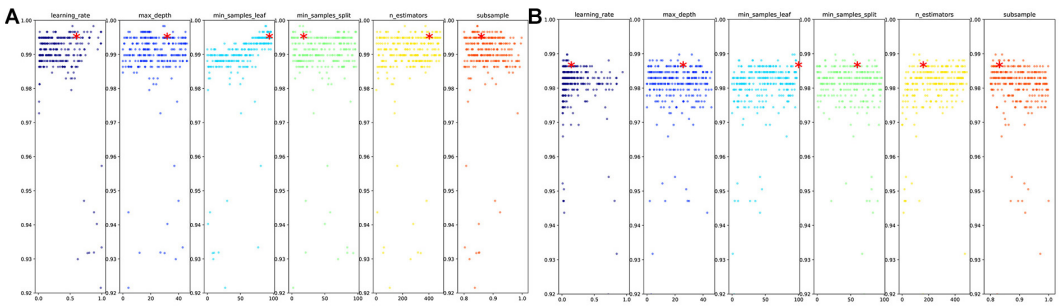

FIGURE 3 | (A) Bayesian optimization for hyperparameters of GBDT. The best hyperparameters set: {"learning_rate": 0.53732209, "max_depth": 29, "min_samples_leaf": 88, "min_samples_split": 12, "n_estimators": 374, "subsample": 0.84620375}, testing accuracy: 0.995102041, testing balanced accuracy: 0.973135976. The best hyperparameter set was obtained by comparing the average metrics on 10 test sets. The detailed results obtained by every test are shown in **Supplementary Datasheet 1**. **(B)** Random search for hyperparameters of GBDT. The best hyperparameters set: {"learning_rate": 0.0829095, "max_depth": 23, "min_samples_leaf": 94, "min_samples_split": 54, "n_estimators": 130, "subsample": 0.817617081}, testing accuracy: 0.994693878, testing balanced accuracy: 0.968032706. The best hyperparameter set was obtained by comparing the average metrics on 10 test sets. The detailed results obtained by every test are shown in **Supplementary Datasheet 1**.

Table 5 shows the mean values of seven evaluation metrics obtained from four methods on the integrative dataset. The results indicate that the proposed method outperforms SVM, KNN, and RF by 3.4, 5.7, and 2.6% with balanced accuracy. Particularly, the AUC obtained by the proposed method is 2.3, 2.9, and 1.6% higher than the above three classifiers, respectively. We can conclude that the proposed method achieves the best performance on the integrative dataset in terms of balanced accuracy (97.31%), accuracy (99.51%), specificity (98.26%), precision (99.82%), F1 score (99.73%), and AUC (98.95%). **Supplementary Datasheet 2** shows the average and variance of each metric, and the proposed method gets the smallest variance in accuracy, balanced accuracy, and F1 score in TCGA-BRCA. Other metrics are the second smallest. It can be seen from the variance table that the proposed method has certain robustness.

The Effect of Integrative Dataset

To explore the effect of the integrative dataset, we apply the proposed method to individual gene expression and integrative dataset, respectively. Besides, we choose PMA50 as the control model. PMA50 refers to a set of 50 genes selected by Parker et al. (2009), which are with a good diagnostic performance that are regarded to be highly related to breast cancer. In **Table 6**, for the gene expression and PMA50, the proposed method achieves the best testing accuracy. The blue and orange bars in **Figures 4A,B** intuitively reflect the results. However, for the integrative dataset, the proposed method obtains 99.51% testing accuracy and 97.31% balanced accuracy, which outperforms the gene expression model and PMA50 model. This fact indicates that the features selected by the proposed model have better classification performance.

The results in **Table 6** and **Figure 4** also show the results obtained by the other classifiers. The SVM classifier gives the accuracy of 98.78% on the gene expression dataset, which is higher than that on the integrative dataset. However, the balanced accuracy is higher on the integrative dataset (94.93%). On the other hand, RF and KNN give a higher testing accuracy on the integrative dataset than that on the gene expression dataset, which is illustrated by the blue bars in **Figure 4A**. However, in **Figure 4A**, the proposed model obtains the highest three bars, which reveals that the proposed method performs better than other classifiers in all three types of datasets. For a balanced accuracy in **Figure 4B**, SVM and the proposed model obtain the best results on the integrative dataset, and RF and KNN obtain the best ones on gene expression and PMA50, respectively. The reason for this difference lies in the sensitivity of different classifiers to data distribution. The feature genes in the PMA50 model and the integrative model obtain higher balanced accuracy 97.4% (KNN) and 97.3% (proposed method) than that in the gene expression model, which illustrates that KNN and the proposed method provide the better capability to classify the minority sample class.

Biomarkers and GO/Pathway Analysis

The 50 genes (listed in the **Supplementary Table 8**) discovered by the proposed model include 16 genes, *IQGAP3* (Hu et al., 2019), *KIF4A* (Xue et al., 2018), *TSHZ2* (Yamamoto et al., 2011), *MKI67* (Schmidt et al., 2007), *TNXB* (Hu et al., 2009), *KIFC1* (Ogden et al., 2017), *KDM5B* (Catchpole et al., 2011), *PPEF1* (Ye et al., 2020), *RYR3* (Shrestha et al., 2012), *TMEM132C* (Zhang et al., 2020), *FANCD2* (Barroso et al., 2006), *ATAD2* (Kalashnikova et al., 2010), *KIF26B* (Wang et al., 2013), *BRCA2* (Wooster et al., 1995), *BLM* (Arora et al., 2015), and *ARFGEF* (Kim et al., 2011), which are reported to be directly associated with breast cancer by previous researches. Although the other 14 genes have not been verified by biological experiments, we further analyze the Gene Ontology and pathway enrichment to explore their impact on the tumor formation and progression.

Gene Ontology and pathway analysis produces biological function and pathway enriched for mutation genes. The result reveals that *BRCA2*, *KDM5B*, and *IQGAP3* are associated with mammary gland epithelial cell proliferation and gland

TABLE 5 | Comparison of related works.

Work	Method	Dataset resource	Evaluation metric	Performance
Mavaddat et al., 2019	Polygenic risk scores (PRSs)	Breast Cancer Association Consortium (BCAC)	AUC	0.63
Chaurasia et al., 2018	Naive Bayes	Breast Cancer Wisconsin dataset	Accuracy	97.36%
Ai et al., 2020	Pearson correlation coefficient (PCC) + SVM	GEO	Accuracy	96.92%
Huang et al., 2017	SVM ensembles	UCI and ACM SIGKDD Cup 2008	Accuracy AUC F-measure	96.85% 0.967 0.988

TABLE 6 | Comparison between the results of different datasets on four classifiers.

Data category	Testing accuracy				Testing balanced accuracy			
	SVM	RF	KNN	Proposed	SVM	RF	KNN	Proposed
Gene	0.9878	0.9918	0.9878	**0.9918**	0.8995	0.9707	0.9619	0.9481
PMA50	0.9743	0.9869	0.9824	**0.9910**	0.8831	0.8980	0.9736	0.9342
Integrative dataset	0.9865	0.9902	0.9914	**0.9951**	0.9413	0.9208	0.9408	**0.9731**

For the gene expression, the proposed method obtains the highest accuracy, but the balanced accuracy is highest in RF. For the PMA50, the proposed method obtains the best accuracy. For the integrative dataset, the proposed method obtains the highest accuracy and balanced accuracy, which illustrates that the integrative dataset contains more useful information after feature selection. The bold values are the best results.

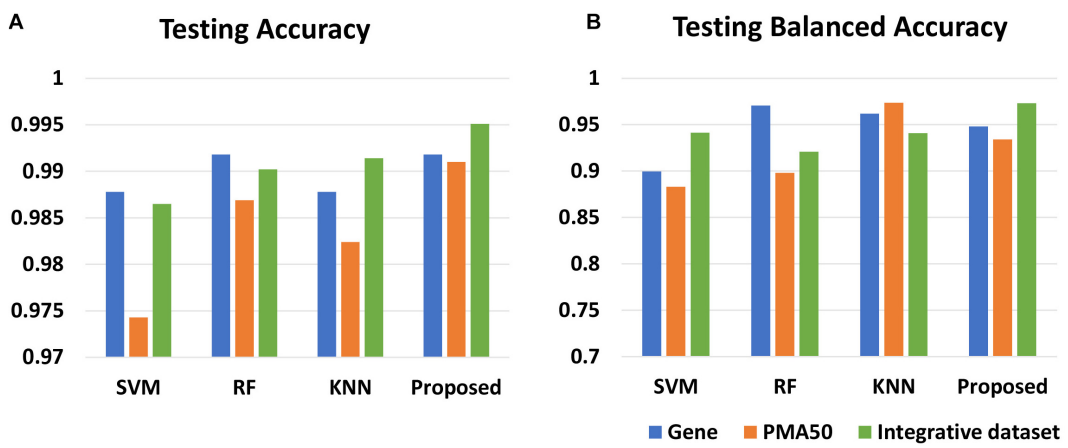

FIGURE 4 | Comparison of the results of different datasets on four classifiers. **(A)** The average testing accuracy obtained by the four methods. **(B)** The average balanced testing accuracy obtained by the four methods. The left data is the accuracy; the height of the blue and red bars represent the performance of each method on gene data and combine data. The red bar obtained by the proposed method in **A** is the highest. The red bar obtained by the proposed method in **B** is also the highest. That means the proposed method performs best in the accuracy and balanced accuracy on the integrative data.

development; *BLM*, *BRCA2*, *CENPE*, *CENPF*, *KIFC1*, *CKAP*, *CIT*, *TTC28*, *KIF4A*, and *ASPM* are associated with cell division; *BRCA*, *CENPE*, *CENPF*, *FANCD2*, *KIFC1*, *MKI67*, *KIF4A*, and *ASPM* are associated with organelle fission; *BLM*, *BRCA2*, *CENPE*, *CENPF*, *EGFR*, *FANCD2*, *MKI67*, *CKAP5*, and *TTC28* are associated with regulation of the mitotic cell cycle; *ABCA10*, *ABCA9*, *ABCA8*, and *ABCA6* enrich in the pathway of ABC transporters; and *EGFR*, *FN1*, *RELN*, and *TNXB* enrich in the pathway of human papillomavirus infection. The main GO and pathway are shown in **Figure 5**. The comprehensive analysis of the whole 50 genes is shown in **Supplementary Datasheet 3**. Overall, the investigation reveals oncogenic interactions and cooperation among mutation genes.

DISCUSSION

This research presents a Staged Feature Selection method for breast cancer classification based on gene expression and somatic mutation datasets. In the proposed method, FC and FDR were used to select differentially expressed genes, MI was adopted to remove the irrelevant and redundant features, and an embedded method based on GBDT with Bayesian optimization was presented to obtain the informative features. Besides, the weighted metric was proposed to evaluate the classification accuracy, which could avoid the impact of sample imbalance on classification. The experiment results showed that the proposed method selected 50 feature genes and achieved the accuracy of

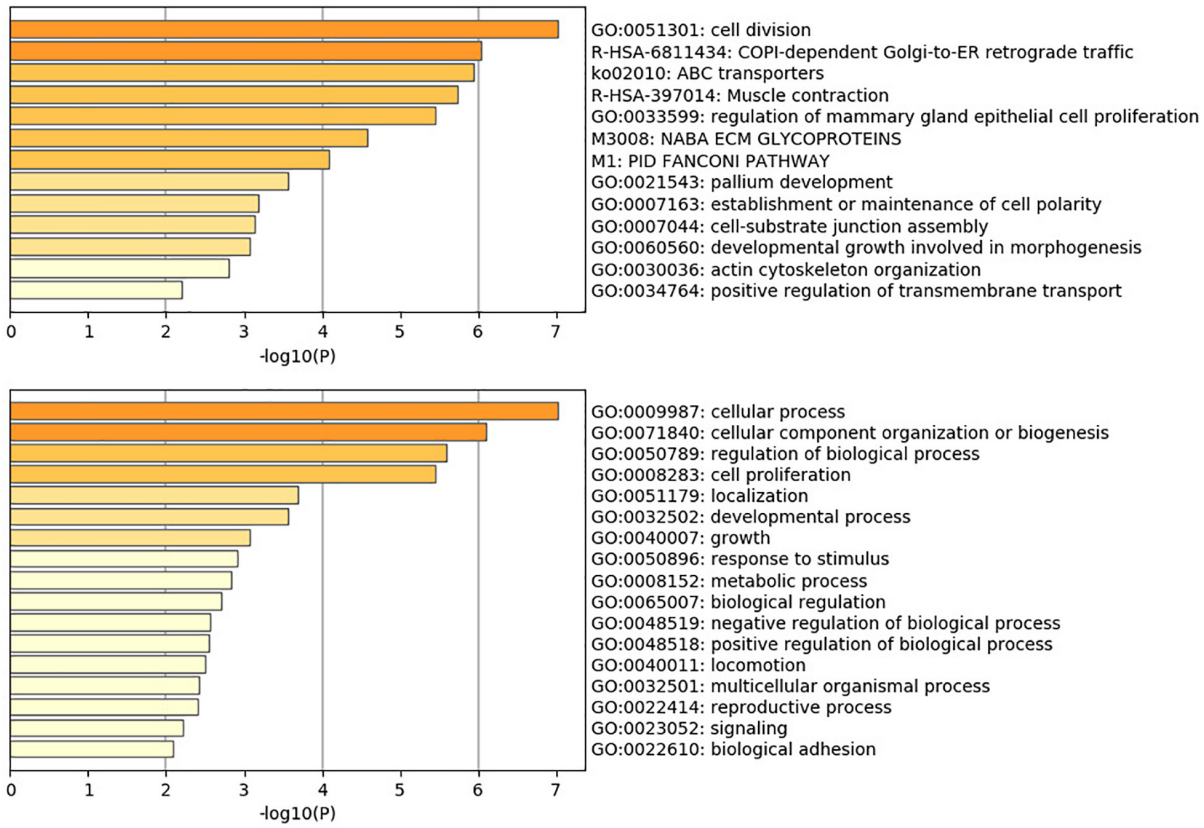

FIGURE 5 | Heatmap of selected Gene Ontology.

99.51%, the balanced accuracy of 97.31% and the sensitivity of 99.64%, the specificity of 98.26%, the precision of 99.82%, the F1 score of above 99.73%, and the AUC of 98.95%, which was superior to the other three classifiers. It was verified that the proposed method was an efficient tool for feature selection in breast cancer classification.

The results presented the effectiveness of integration with gene expression and somatic mutation data for breast cancer classification, which indicated that it could provide more useful information for cancer classification by integrating multiple information. However, this study only focused on breast cancer, and the scalability of the proposed method on other types of cancers remained to be further explored, which will provide helpful information for cancer prevention and treatment. Therefore, in future work, we will apply the approach to classify other types of cancer, explore ways to incorporate more relevant data, and introduce other techniques to boost our method. Besides, the pathogenesis of some biomarkers discovered by the proposed model still has to be verified by biological experiments.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov/repository.

AUTHOR CONTRIBUTIONS

QJ processed the data, designed the algorithm and the programming codes, and wrote the manuscript. MJ supervised the project and revised the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by the National Natural Science Foundation of China under Grant 61773157.

ACKNOWLEDGMENTS

We thank the members for their support and guidance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.629946/full#supplementary-material

REFERENCES

Ai, D., Wang, Y., Li, X., and Pan, H. (2020). Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder. *Biomolecules* 10:1207. doi: 10.3390/biom10091207

Arora, A., Abdel-Fatah, T. M. A., Agarwal, D., Doherty, R., Moseley, P. M., Aleskandarany, M. A., et al. (2015). Transcriptomic and protein expression analysis reveals clinicopathological significance of bloom syndrome helicase (BLM) in breast cancer. *Mol. Cancer Ther.* 14, 1057–1065. doi: 10.1158/1535-7163.mct-14-0939

Barroso, E., Milne, R. L., Fernández, L. P., Zamora, P., Arias, J. I., Benítez, J., et al. (2006). FANCD2 associated with sporadic breast cancer risk. *Carcinogenesis* 27, 1930–1937. doi: 10.1093/carcin/bgl062

Bhat, P. C., Prosper, H. B., Sekmen, S., and Stewart, C. (2018). Optimizing event selection with the random grid search. *Comp. Phys. Commun.* 228, 245–257. doi: 10.1016/j.cpc.2018.02.018

Blagus, R., and Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* 14:106.

Bonev, B., Escolano, F., and Cazorla, M. (2008). Feature selection, mutual information, and the classification of high-dimensional patterns. *Pattern Anal. Applic.* 11, 309–319. doi: 10.1007/s10044-008-0107-0

Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., et al. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18545–18550. doi: 10.1073/pnas.1010978107

Catchpole, S., Spencer-Dene, B., Hall, D., Santangelo, S., Rosewell, I., Guenatri, M., et al. (2011). PLU-1/JARID1B/KDM5B is required for embryonic survival and contributes to cell proliferation in the mammary gland and in ER+ breast cancer cells. *Int. J. Oncol.* 38, 1267–1277.

Chaurasia, V., Pal, S., and Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *J. Algorithms Comp. Technol.* 12, 119–126. doi: 10.1177/1748301818756225

Ding, J., Mcconechy, M. K., Horlings, H. M., Ha, G., Chan, F. C., Funnell, T., et al. (2015). Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nat. Commun.* 6:8554.

Fleck, J. L., Pavel, A. B., and Cassandras, C. G. (2016). Integrating mutation and gene expression cross-sectional data to infer cancer progression. *BMC Syst. Biol.* 10:12.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

Galicia, A., Torres, J. F., Martínez-Álvarez, F., and Troncoso, A. (2018). A novel Spark-based multi-step forecasting algorithm for big data time series. *Inform. Sci.* 467, 800–818. doi: 10.1016/j.ins.2018.06.010

He, Z., Zhang, J., Yuan, X., Liu, Z., Liu, B., and Tuo, S. (2017). Network based stratification of major cancers by integrating somatic mutation and gene expression data. *PLoS One* 12:e0177662. doi: 10.1371/journal.pone.0177662

Hu, G., Liu, H., Wang, M., and Peng, W. (2019). IQ motif containing GTPase-activating protein 3 (IQGAP3) inhibits kaempferol-induced apoptosis in breast cancer cells by extracellular signal-regulated kinases 1/2 (ERK1/2) signaling activation. *Med. Sci. Monit.* 25:7666. doi: 10.12659/msm.915642

Hu, X., Zhang, Y., Zhang, A., Li, Y., Zhu, Z., Shao, Z., et al. (2009). Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls via label-free semiquantitative shotgun technology. *OMICS* 13, 291–300. doi: 10.1089/omi.2009.0016

Huang, M.-W., Chen, C.-W., Lin, W.-C., Ke, S.-W., and Tsai, C.-F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS One* 12:e0161501. doi: 10.1371/journal.pone.0161501

Jia, P., and Zhao, Z. (2016). Impacts of somatic mutations on gene expression: an association perspective. *Brief Bioinform.* 18, 413–425.

Kalashnikova, E. V., Revenko, A. S., Gemo, A. T., Andrews, N. P., Tepper, C. G., Zou, J. X., et al. (2010). ANCCA/ATAD2 overexpression identifies breast cancer patients with poor prognosis, acting to drive proliferation and survival of triple-negative cells through control of B-Myb and EZH2. *Cancer Res.* 70, 9402–9412. doi: 10.1158/0008-5472.can-10-1199

Kim, J. H., Kim, T. W., and Kim, S. J. (2011). Downregulation of ARFGEF1 and CAMK2B by promoter hypermethylation in breast cancer cells. *BMB Rep.* 44, 523–528. doi: 10.5483/bmbrep.2011.44.8.523

Kwong, A., Cheuk, I. W., Shin, V. Y., Ho, C. Y., Au, C.-H., Ho, D. N., et al. (2020). Somatic mutation profiling in BRCA-negative breast and ovarian cancer patients by multigene panel sequencing. *Am. J. Cancer Res.* 10, 2919–2932.

Li, Q. Q., and Liu, X. Y. (2014). EasyEnsemble.M for multiclass imbalance problem. *Moshi Shibie yu Rengong Zhineng* 27, 187–192.

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., et al. (2017). A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics* 18:508.

Liu, X. Y., Wu, J., and Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern B (Cybernetics)* 39, 539–550. doi: 10.1109/tsmcb.2008.2007853

Mamidi, T. K. K., Wu, J., and Hicks, C. (2019). Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. *BMC Cancer* 19:229.

Masica, D. L., and Karchin, R. (2011). Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* 71, 4550–4561. doi: 10.1158/0008-5472.can-11-0180

Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hu. Genet.* 104, 21–34.

Meng, Y., Jin, M., Tang, X., and Xu, J. (2020). Degree-based similarity indexes for identifying potential miRNA-disease associations. *IEEE Access* 8, 133170–133179. doi: 10.1109/access.2020.3006998

Merid, S. K., Goranskaya, D., and Alexeyenko, A. (2014). Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinfrom.* 15:308. doi: 10.1186/1471-2105-15-308

Ogden, A., Garlapati, C., Li, X. B., Turaga, R. C., Oprea-Ilies, G., Wright, N., et al. (2017). Multi-institutional study of nuclear KIFC1 as a biomarker of poor prognosis in African American women with triple-negative breast cancer. *Sci. Rep.* 7:42289.

Oh, C., Tomczak, J. M., Gavves, E., and Welling, M. (2019). "Combinatorial bayesian optimization using graph representations," in *Proceedings of the ICML Workshop on Learning and Reasoning With Graph-Structured Data 2019*, Vancouver, BC.

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

Reiner-Benaim, A. (2010). FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biom J.* 49, 107–126. doi: 10.1002/bimj.200510313

Schmidt, M., Boehm, D., Von Toerne, C., Lehr, H. A., Hengstler, J. G., Koelbl, H., et al. (2007). Prognostic impact of MKI67 and MMP1 in node-negative invasive ductal and invasive lobular carcinoma of the breast. *J. Clin. Oncol.* 38, 239–255.

Shrestha, S., Yan, Q., Joseph, G., Arnett, D. K., Martinson, J. J., and Kingsley, L. A. (2012). Replication of RYR3 gene polymorphism association with cIMT among HIV-infected whites. *AIDS* 26, 1571–1573. doi: 10.1097/qad.0b013e328355359f

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Adv. Neural Inform. Process. Syst.* 25, 2960–2968.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.* 19, A68–A77.

Victoria, A. H., and Maragatham, G. (2020). Automatic tuning of hyperparameters using Bayesian optimization. *Evol. Syst.* 1–7.

Wang, Q., Zhao, Z.-B., Wang, G., Hui, Z., Wang, M.-H., Pan, J.-F., et al. (2013). High expression of KIF26B in breast cancer associates with poor prognosis. *PLoS One* 8:e61640. doi: 10.1371/journal.pone.0061640

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., et al. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* 378, 789–792.

Wu, J., Mamidi, T. K. K., Zhang, L., and Hicks, C. (2019). Integrating germline and somatic mutation information for the discovery of biomarkers in triple-negative breast cancer. *Int. J. Environ. Res. Public Health* 16:1055. doi: 10.3390/ijerph16061055

Xue, D., Cheng, P. U., Han, M., Liu, X., Xue, L., Ye, C., et al. (2018). An integrated bioinformatical analysis to evaluate the role of KIF4A as a prognostic biomarker for breast cancer. *Onco Targets Ther.* 11, 4755–4768. doi: 10.2147/ott.s164730

Yamamoto, M., Cid, E., Bru, S., and Yamamoto, F. (2011). Rare and frequent promoter methylation, respectively, of TSHZ2 and 3 genes that are both downregulated in expression in breast and prostate cancers. *PLoS One* 6:e17149. doi: 10.1371/journal.pone.0017149

Ye, T., Wan, X., Li, J., Feng, J., Guo, J., Li, G., et al. (2020). The clinical significance of PPEF1 as a promising biomarker and its potential mechanism in breast cancer. *Onco Targets Ther.* 13, 199–214. doi: 10.2147/ott.s229432

Zhang, J., Zhang, S., Wang, Y., and Zhang, X.-S. (2013). Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst. Biol.* 7(Suppl. 2):S4.

Zhang, W., and Wang, S. L. (2020). A novel method for identifying the potential cancer driver genes based on molecular data integration. *Biochem. Genet* 58, 16–39. doi: 10.1007/s10528-019-09924-2

Zhang, X., Kang, X., Jin, L., Bai, J., Zhang, H., Liu, W., et al. (2020). ABCC9, NKAPL, and TMEM132C are potential diagnostic and prognostic markers in triple−negative breast cancer. *Cell Biol. Int.* 44, 2002–2010. doi: 10.1002/cbin.11406

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 March 2021
doi: 10.3389/fgene.2021.648898

A Novel Method to Identify the Differences Between Two Single Cell Groups at Single Gene, Gene Pair, and Gene Module Levels

Lingyu Cui[1], *Bo Wang*[2], *Changjing Ren*[1], *Ailan Wang*[2], *Hong An*[3]* *and Wei Liang*[4]*

[1] School of Science, Dalian Maritime University, Dalian, China, [2] Geneis (Beijing) Co., Ltd., Beijing, China, [3] Guangzhou Anjie Biomedical Technology Co., Ltd., Guangzhou, China, [4] Medical Clinical Laboratory, The Second People's Hospital of Lianyungang, Lianyungang, China

Single-cell sequencing technology can not only view the heterogeneity of cells from a molecular perspective, but also discover new cell types. Although there are many effective methods on dropout imputation, cell clustering, and lineage reconstruction based on single cell RNA sequencing (RNA-seq) data, there is no systemic pipeline on how to compare two single cell clusters at the molecular level. In the study, we present a novel pipeline on comparing two single cell clusters, including calling differential gene expression, coexpression network modules, and so on. The pipeline could reveal mechanisms behind the biological difference between cell clusters and cell types, and identify cell type specific molecular mechanisms. We applied the pipeline to two famous single-cell databases, Usoskin from mouse brain and Xin from human pancreas, which contained 622 and 1,600 cells, respectively, both of which were composed of four types of cells. As a result, we identified many significant differential genes, differential gene coexpression and network modules among the cell clusters, which confirmed that different cell clusters might perform different functions.

Keywords: scRNA-seq, differential gene expression analysis, differential correlation analysis, network analysis, differential network analysis

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Liuyi Hao,
University of North Carolina at Greensboro, United States
Qianqian Song,
Wake Forest Baptist Medical Center, United States
Feng Wang,
Emory University, United States

***Correspondence:**
Wei Liang
hslwys@163.com
Hong An
149308555@qq.com

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

Received: *02 January 2021*
Accepted: *15 February 2021*
Published: *15 March 2021*

Citation:
Cui L, Wang B, Ren C, Wang A, An H and Liang W (2021) A Novel Method to Identify the Differences Between Two Single Cell Groups at Single Gene, Gene Pair, and Gene Module Levels. Front. Genet. 12:648898. doi: 10.3389/fgene.2021.648898

INTRODUCTION

The fundamental unit of an organism is the cell. Coordinated gene expression in each cell is essential to biological functions, and aberrations often cause illness. Consequently, the genome-wide quantification of RNA experiments help to understand the growth and development of organism as well as pathogenesis of disease. One traditional technology of mRNA abundance measured at cell line or tissue level averaged over thousands or millions of cells, which is also called bulk RNA-seq (Stark et al., 2019). The bulk RNA-seq experiments has been successfully applied to a multitude of studies, and improved our biology knowledge. However, the disadvantage of bulk RNA-seq is that cell-specific mRNA abundance could not been provided, and some important gene expression signals might be unobserved. Our current knowledge related with cell types and there dynamic changes in biological system remains highly incomplete. Owing to resolution in sequencing technology, single-cell RNA-seq (scRNA-seq) at genome-wide level was first invented by Tang et al. (2009), and has been under rapidly booming development. The scRNA-seq technology makes some very important and challenging scientific research possible.

For instance, unknown cell types were identified (Trombetta et al., 2014; Buettner et al., 2015). How to dissect gene expression changes during dynamic development (Tang et al., 2010; Xue et al., 2013; Yan et al., 2013). Study uncovered how tumorgenesis and cancer cell immune escape and tumor cell heterogeneity (Chung et al., 2017; Zhao et al., 2020). scRNA-seq was also used to predict therapeutic response in patients and understanding drug resistance mechanism (Lee et al., 2014; Liang et al., 2020), and clarify the pathophysiology of complex diseases and guide the successful treatment and intervention of patients with intractable diseases (Shalek and Benson, 2017; Kim et al., 2020). Collectively, the scRNA-seq technology has significantly promoted basic biological research and clinical personalized medicine. At the same time, the analysis of scRNA-seq data is challenging due to a number of problems such as sparsity caused by technical dropout, bimodal and multi-modal expression distributions (Korthauer et al., 2016), and highly biological and technical cell-to-cell variability (Vallejos et al., 2017; Hicks et al., 2018) giving rise to cellular heterogeneity. One very important step of scRNA-seq data analysis is to identify gene-specific expression pattern and/or a gene-gene interacting network within a population of cells or a biological condition in studies. Although numerous computational methods have been developed and applied during the past few years, most of them focused on difference in single gene-level differentiation (Finak et al., 2015; Korthauer et al., 2016; Butler et al., 2018; Miao et al., 2018; Stuart et al., 2019). In the present study, we integrated a variety of computational methods into a variance analysis workflow.

A fundamental question raised of expression data is what genes differentially expressed across conditions and circumstances. Despite technological revolution for scRNA-seq in recent years, technical stability of RNA quantification by scRNA-seq is still worse than that in bulk RNA-seq. Thus, the numerous variation computational tool established for bulk RNA do not work well for single-cell RNA-seq. During the past few years, a couple of computational methods have been designed particularly for single-cell RNA-seq data (Soneson and Robinson, 2018). For example, MAST based on Generalized linear model (Finak et al., 2015); DEsingle based on Zero inflated negative binomial (Miao et al., 2018); D3E based on Cramér-von Mises test, Kolmogorov-Smirnov test, likelihood ratio test (Delmans and Hemberg, 2016); SCDE based on Poisson and negative binomial model (Kharchenko et al., 2014); SigEMD based on Non-parametric earth mover's distance (Wang and Nabavi, 2018) and so on. Marker genes found by differential expression analysis play important role in cell type identification and discovery. It is also essential for downstream drug targets prediction and thus to prevent or treat disease. In addition to analyzing single gene, analyzing the relationship between genes is also crucial for construction of biological networks. For instance, the R package DGCA offers a suite of tools for computing and analyzing differential correlations between genes across multiple conditions (McKenzie et al., 2016).

If some genes always have similar expression patterns in a physiological process or metabolic process, then we can consider these genes to be functionally dependency, so they can be defined as a functional module. If a gene module is identified, then numerous researches would be done based of which, such as screening the core genes of relevant trait modules, modeling metabolic pathways, and establishing gene interaction networks. Weighted correlation network analysis (WGCNA) is a typical analysis tool at the network co-expression level (Langfelder and Horvath, 2008). Since WGCNA is an analysis tool designed for bulk sequencing data, almost no one uses it to analyze scRNA-seq data. Based the correlation between the analyzed module and the sample characteristics we can quickly extract gene co-expression modules related to the sample characteristics from the complex data for subsequent analysis. WGCNA builds a bridge between sample characteristics and gene expression changes (Iancu et al., 2012; Xue et al., 2013).

In the present study, we performed differential expression genes (DEGs) analysis for each two categories in the scRNA-seq data from a single gene level. Based the level of gene pairs, differential correlation analysis for each two categories were analyzed for the purpose of digging deeper biological information. The gene pair with the most significant difference in each category pair was obtained. The results of this analysis provide theoretical support for medical staff. Based the level of gene network, we used WGCNA to perform network analysis on scRNA-seq data, and in order to explore the difference in gene expression of each module, we used DiffCoEx (Tesson et al., 2010) to analyze the difference network module. The results of analysis from different levels of single cells, cell pairs, and cell networks showed that such a complete system is more capable of mining the underlying information contained in the scRNA-seq data. The study provided a comprehensive analysis approach for scRNA-seq researches in future.

MATERIALS AND METHODS

In recent years, with the microfluidic technology that can separate individual cells from a piece of tissue, researchers have made it more accurate to predict the diversity of biological tissues and target drugs for related diseases. Compared with bulk sequencing technology, the resolution of scRNA-seq technology is very accurate for single-cell level analysis, so scRNA-seq technology has developed rapidly. Although a series of work on single-cell sequencing technology has been developed in recent years, most of them are tested and verified in a single field, and there is no complete system to mine the potentially valuable information in single-cell data. Ignore some algorithms that have been developed in bulk sequencing technology, such as WGCNA. In this work, we have established a set of procedures for analyzing scRNA-seq data, including differential gene expression analysis (DEsingle, SigEMD), differential correlation analysis (DGCA), network analysis (WGCNA), differential network analysis (DNA). The specific flow chart is shown in **Figure 1**. These processes are described in detail below.

Data Information

In this work, we used two single-cell data sets. One of them is Usoskin [622 (cells) * 25335 (genes)], which comes from the GEO database (GSE59739) (Usoskin et al., 2015). This data is mainly divided into 4 categories: NF, NP, PEP, and TH. We performed

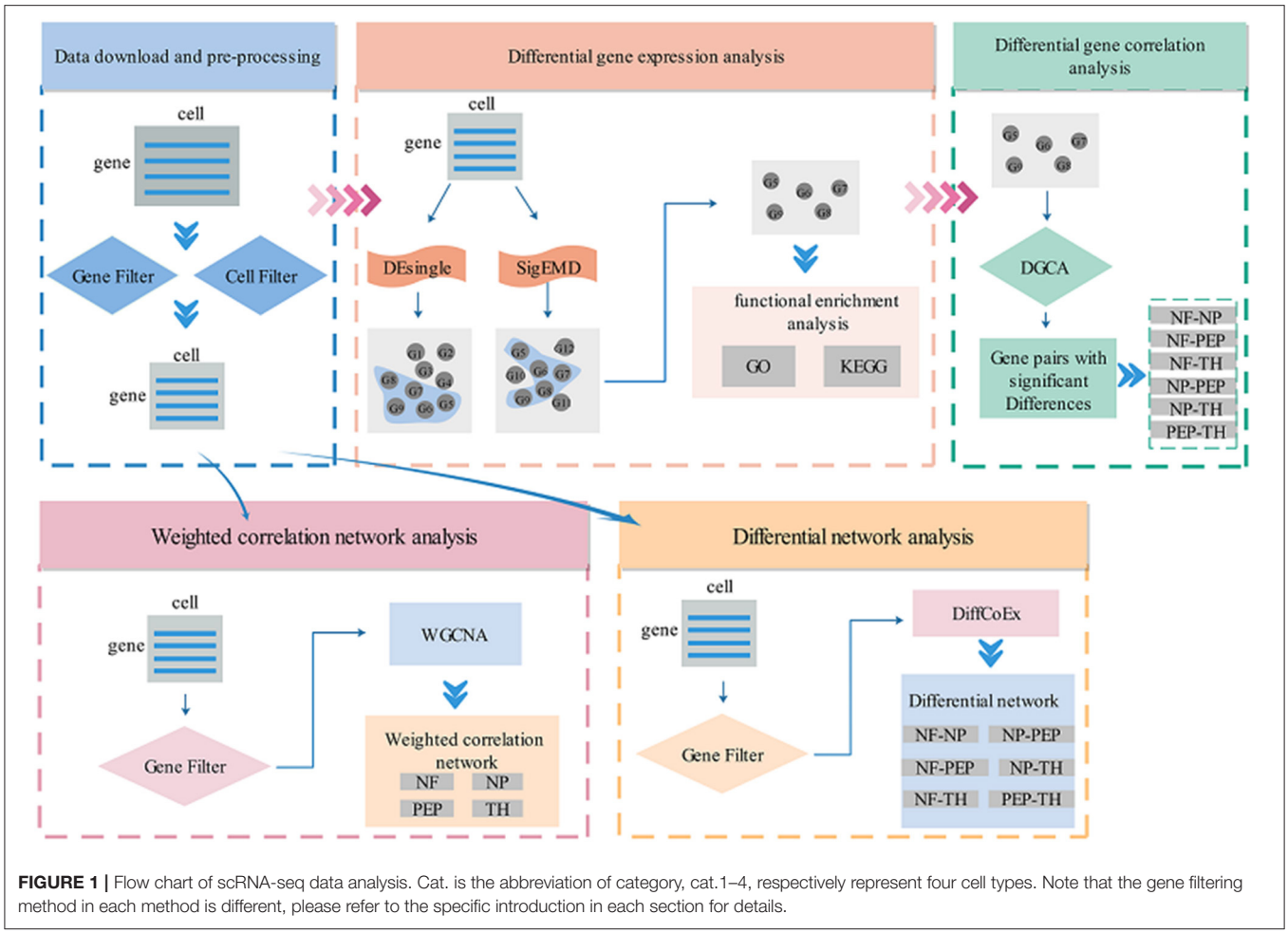

FIGURE 1 | Flow chart of scRNA-seq data analysis. Cat. is the abbreviation of category, cat.1–4, respectively represent four cell types. Note that the gene filtering method in each method is different, please refer to the specific introduction in each section for details.

TABLE 1 | Brief information about Usoskin data.

Usoskin (622)	Num. of cells	Num. of genes	Description of cell groups
NF	139	25333	Neurofilament containing
NP	169		Non-peptidergic nociceptors
PEP	81		Peptidergic nociceptors
TH	233		Tyrosine hydroxylase containing

TABLE 2 | Brief information about Xin data.

Xin (1492)	Num. of cells	Num. of genes	Description of cell groups
α cells	886	28403	Produce glucagon
β cells	472		Insulin
δ cells	49		Somatostatin
PP cells	85		Pancreatic polypeptide

a concise preprocessing of the data, the gene filter removes genes/transcripts that are expressed in <3 cells, and the cell filter removes cells that are expressed in <500 genes, the number of remaining samples is 622, and the gene dimension is 25333. **Table 1** summarizes the basic information of Usoskin data.

The other data used in this article is from human pancreas, named Xin [1600 (cells) * 39851 (genes)], which comes from the GEO database (GSE81608) (Xin et al., 2016). Xin is also divided into four categories: α, β, δ, and PP. The data uses the same preprocessing method as Usoskin data, the number of remaining samples is 1492, and the gene dimension is 28403. **Table 2** summarizes the basic information of Xin data.

Differential Gene Expression Analysis

We performed pairwise difference expression genes (DEG) analysis on the four types of cells in these two data sets. The methods used for DEGs are DEsingle (Miao et al., 2018) and SigEMD (Wang and Nabavi, 2018), both of which are methods for scRNA-seq data.

One of the biggest features of scRNA-seq data is that it contains a high proportion of 0 values, which is mainly due to two reasons: on the one hand, these "true" 0 values are the natural expression values of genes; on the other hand, due to the reverse transcription and sequencing process, there are too many "false" 0 values caused by the technical noise of

the company, we call the latter "dropout." In response to this phenomenon, most of the current differential analysis methods cannot separate the two situations, so DEsingle was developed to solve the differential analysis that contains the dropout problem data. DEsingle employed Zero-Inflated Negative Binomial model to estimate the proportion of real and dropout zeros and detect three types of DEGs in scRNA-seq data with higher accuracy.

The SigEMD method also takes into account the "dropout" problem. Using a logistic regression model and a non-parametric method based on the distance of the earth mover can accurately and effectively identify the DEGs in the scRNA-seq data. Regression models and data imputation are used to reduce the impact of a large number of zero counts, and non-parametric methods are used to improve the sensitivity of detecting DEGs from multimodal scRNAseq data. And used simulated data sets and real data sets to verify the accuracy of this method.

Differential Gene Correlation Analysis

The key step to establish a biological system prediction model is to analyze the regulatory relationship between genes, so an effective solution is to study the difference in correlation between gene pairs. Differential Gene Correlation Analysis (DGCA) is proposed to solve such problems (McKenzie et al., 2016). In order to minimize parameter assumptions, DGCA calculates empirical *p*-values through permutation tests. In order to understand the differential correlation at the system level, DGCA conducted a higher-level analysis through simulation research. The simple method based on Z score adopted by DGCA is significantly better than the existing alternative methods of calculating differential correlation.

Network Analysis

Weighted correlation network analysis (WGCNA) is a systems biology method used to describe gene association patterns between different samples (Liu et al., 2017). WGCNA can be used to identify highly coordinated gene sets, and identify candidate biomarker genes or therapeutic targets based on the interconnectivity of gene sets and the association between gene sets and phenotypes. Compared with only focusing on differentially expressed genes, WGCNA uses the information of thousands or tens of thousands of genes with the greatest changes or all genes to identify the gene set of interest, and conducts significant association analysis with the phenotype. It not only makes full use of information, but also converts the associations between thousands of genes and phenotypes into associations between multiple genomes and phenotypes, eliminating the problems of multiple hypothesis testing and correction.

Differential Network Analysis

In scRNA-seq data, if certain genes always have similar expression changes in a physiological process or in different tissues, then we have reason to believe that these genes are functionally related and can be defined as a module. When the gene module is defined, we can use these results to do a lot of further work. For example, we use DiffCoEx for differential network analysis (Tesson et al., 2010), which is a method for identifying changes in association patterns. This method is based on the commonly used WGCNA framework for co-expression analysis. Prove its usefulness by identifying biologically relevant, differentially co-expressed modules in the mouse dataset.

SOFTWARE AVAILABILITY

The codes for the two methods of differential gene expression analysis are freely available (DEsingle: https://bioconductor.org/packages/DEsingle, SigEMD: https://github.com/NabaviLab/SigEMD); This article uses the DAVID website for feature enrichment analysis. The website is available for free in https://david.ncifcrf.gov/ Difference correlation analysis is freely available in https://github.com/andymckenzie/DGCA WGCNA is freely available in https://cran.r-project.org/web/packages/WGCNA/index.html DiffCoEx is freely available in https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-497.

RESULTS

DEGs Between Two Categories

With the development of high-throughput technology, the field of biomedical related research has entered the omics era, and the research of a single gene can no longer meet the needs of researchers. However, such a large amount of data brings new challenges to the effective extraction and analysis of information. Taking sequencing data as an example, the analysis of sequencing results often results in a list of differentially expressed genes or proteins. But for many researchers, it is difficult to associate this long list of genes or proteins with a biological phenomenon to be studied and its underlying mechanism. Functional enrichment analysis is to divide a gene or protein list into multiple parts, that is, to classify a bunch of genes, and the classification criteria here are often limited according to the function of the gene. In other words, it is to put together genes with similar functions in a gene list and associate them with biological phenotypes.

We use DEsingle and SigEMD two methods to analyze the four types of data contained in Usoskin, overlap the differential genes obtained by the two methods, and select the differential genes with $p < 0.05$ for functional enrichment analysis. In this work, we used DAVID to perform two enrichment analyses of GO and KEGG on overlapping differential genes obtained from two NF-NP data, and correlated them with biological phenotypes. Among them, GO (Gene Ontology) enrichment analysis is mainly divided into three parts: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC), as shown in **Figure 2A**, we have selected the top 20 representative Go terms for BP, CC, and MF. The x-axis represents the first 20 terms selected for each part, the y-axis represents the change of *p*value, and the color represents z-score. The KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database that systematically analyzes the metabolic pathways of gene products in cells and the functions of these gene products. KEGG integrates

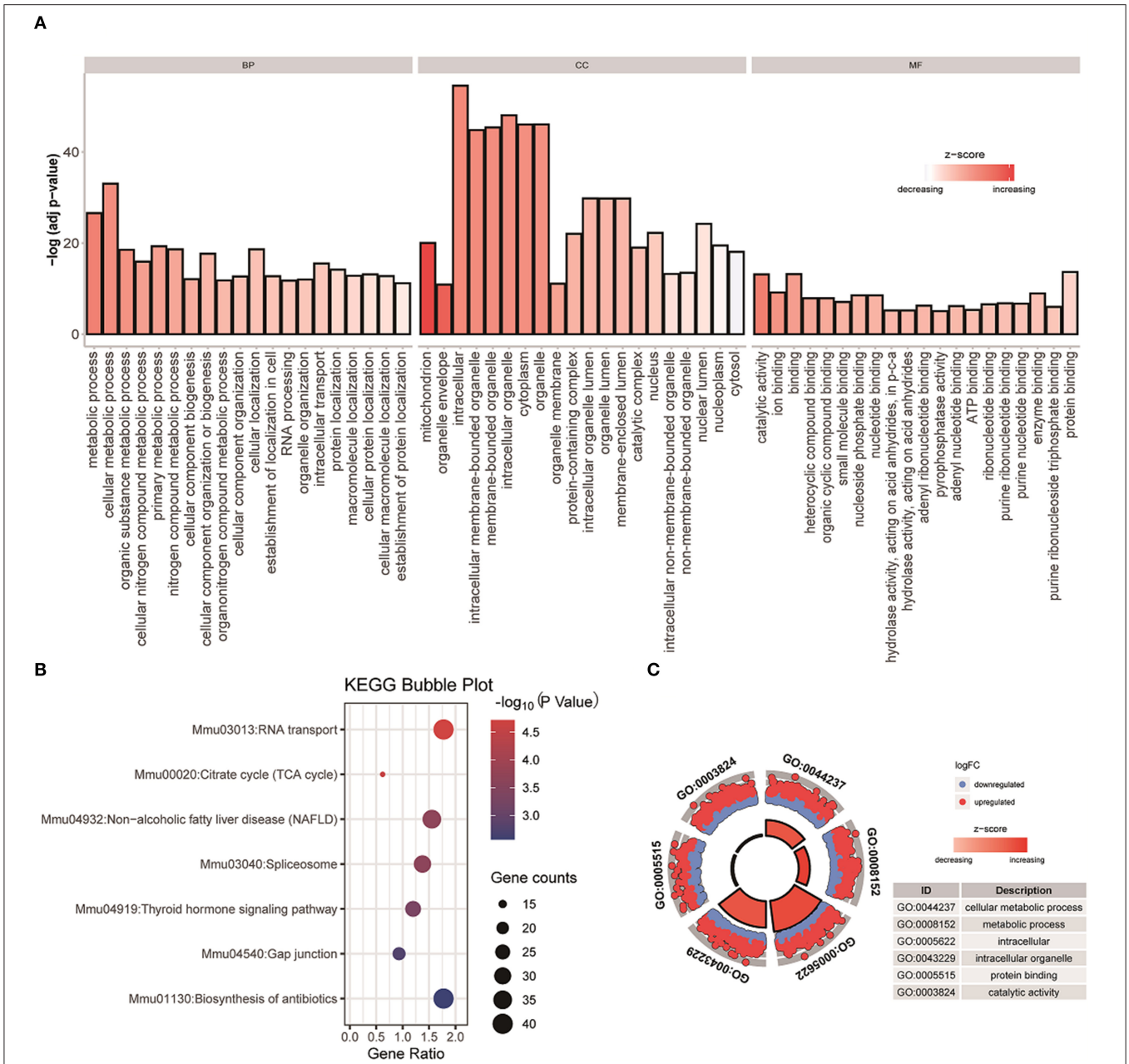

FIGURE 2 | GO and KEGG analysis were performed on the differential genes with overlapping NF-NP data. **(A)** Perform enrichment analysis on the differential genes with overlapping datasets, and display the top 20 most significant terms in BP, CC, and MF. **(B)** Perform KEGG enrichment analysis on the differential genes with overlapping NF-NP data. **(C)** Basic information of six specified terms, among them, blue means down-regulated genes, red means up-regulated genes.

data on the genome, chemical molecules, and biochemical systems, including metabolic pathways (PATHWAY), etc. As shown in **Figure 2B**, we can observe that seven pathways are obtained in the two sets of NF-NP data, and the number of genes expressed in the pathway Mmu030133: RNA transport pathway is large, and the *p*-value lower, indicating that the enrichment of this pathway is the most significant. In addition, we selected two terms with the most significant enrichment among the three indicators of BP, CC, and MF, and analyzed the up-regulated and down-regulated genes of these six terms, as well as their z-score changes, as shown in **Figure 2C**. The Xin data set and the other five analysis results are shown in **Supplementary Material 1**.

Gene Pairs With Significant Differences Between Two Categories

Analyzing the regulatory relationship between genes is a key step in establishing an accurate prediction model of biological

systems. To achieve this goal, a powerful method is to systematically study the correlation differences between gene pairs in more than one situation. In our work, we will perform pairwise analysis on the four data types contained in Usoskin and Xin, and consider the difference and correlation between gene pairs in different types of datasets. We used overlapping differentially expressed genes as the input of DGCA, and listed the most different gene pairs in six different situations, as shown in **Tables 3**, **4**.

The first column in **Table 3** shows the matching analysis pairs of six different data subtypes, corresponding to class1 and class2, respectively in columns four and five, and the sixth column shows the change value of Z-score, indicating the change of correlation between gene pairs. **Table 4** is the same. NF_NP, NF_PEP, NF_TH, NP_PEP, PEP_TH these five pairs of data from class1 to class2 gene pair correlation completely lost, on the contrary, NP_TH this pair of data is completely irrelevant from class1 to class2 correlation has been significantly improved. Please refer to **Table 5** for basic information about the two genes *Il17rd* and *Pde1b*. For detailed information about these two genes, please refer to the database MGI (Mouse Genome Informatics, http://www.informatics.jax.org/).

Co-expression Networks Generated With WGCNA

WGCNA is mainly divided into two steps. In the first step, WGCNA analysis uses the weighted value of the correlation coefficient, that is, the gene correlation coefficient is taken to the power of β, so that the connection between the genes in the network obeys the scale-free network distribution (scale-free networks), determine the β parameter by the square of the correlation coefficient of $\log(k)$ and $logp[(k)]$. In general, the higher the square of the correlation coefficient, the closer the network is to the distribution without network scale. This algorithm has more biological significance. The second step is to construct a hierarchical clustering tree through the correlation coefficients between genes. Different branches of the clustering tree represent different gene modules, and different colors represent different modules. Based on the weighted correlation coefficients of genes, genes are classified according to their expression patterns, and genes with similar patterns are grouped into one module. In this way, tens of thousands of genes can be divided into dozens of modules through gene expression patterns, which is a process of extracting general information.

TABLE 3 | The six gene pairs in Usoskin data have the largest differences in different situations.

	Gene1	Gene2	class1_cor	class2_cor	zScoreDiff	empPVals	Classes
NF_NP	Robo1	Grid1	−0.0797	0.9913	23.5578	3.11102E-08	0/+
NF_PEP	Tomm22	Zc3h13	−0.1886	0.9901	20.0313	3.96294E-08	−/+
NF_TH	Fam84a	Omg	−0.0701	0.9953	25.0786	1.57764E-08	0/+
NP_PEP	Itga3	Synj2	−0.0263	0.9968	19.5367	4.84097E-06	0/+
NP_TH	Il17rd	Pde1b	0.9865	−0.0098	−24.5847	1.25299E-06	+/0
PEP_TH	H2.M11	Pde8b	−0.0716	0.9945	20.6581	2.66991E-07	0/+

TABLE 4 | The six gene pairs in Xin data have the largest differences in different situations.

	Gene1	Gene2	class1_cor	class2_cor	zScoreDiff	empPVals	Classes
α_β	DAPL1	HMOX1	0.9962	−0.0402	−47.0484	9.33E-09	+/0
α_δ	GCG	G6PC2	−0.1662	0.9967	18.8006	3.79E-07	−/+
α_pp	SLC25A53	RPL518	−0.0630	0.9901	23.6008	6.44E-08	0/+
β_δ	INS	IAPP	−0.1771	0.1000	18.4667	1.33E-06	−/+
β_pp	INS	IAPP	−0.1771	0.1000	23.7250	7.50E-08	−/+
δ_pp	RBP4	SST	−0.0414	0.9979	14.5204	1.53E-05	0/+

TABLE 5 | Basic information of genes *Il17rd* and *Pde1b*.

	Il17rd	Pde1b
Name	Interleukin 17 receptor D	Phosphodiesterase 1B, Ca2+-calmodulin dependent
Feature type	Protein coding gene	Protein coding gene
Human ortholog	IL17RD, interleukin 17 receptor D	PDE1B, phosphodiesterase 1B
Chr location	3p14.3; chr3:57089982-57170317 (−) GRCh38.p7	12q13.2; chr12:54549393-54579239 (+) GRCh38.p7
HomoloGene	Vertebrate Homology Class 9717	Vertebrate Homology Class 37370
HCOP	Human homology predictions: IL17RD	Human homology predictions: PDE1B

TABLE 6 | The Hub gene of the NF data subset.

Module	Hub gene	Module	Hub gene
Bisque4	Prr14	Lightyellow	Pml
Black	Tmem8b	Magenta	Acvr2a
Blue	BC052040	Mediumpurple3	Myo1b
Brown	Cntn2	Midnightblue	Atrx
Brown4	BC021891	Orange	Mybl1
Cyan	Tmem130	Orangered4	Acbd3
Darkgreen	Robo3	Paleturquoise	Taf1c
Darkgrey	Mboat1	Pink	Gm13375
Darkmagenta	Fam70b	Plum1	Pkn2
Darkolivegreen	Slc7a8	Plum2	Mmadhc
Darkorange	Slc25a47	Purple	Hhex
Darkorange2	Chd8	Red	Apc2
Darkred	Gpx2.ps1	Royalblue	Med26
Darkslateblue	Catsper2	Saddlebrown	Slc25a44
Darkturquoise	Ebf4	Salmon	Orm3
Floralwhite	Klhl28	Sienna3	Scube2
Green	Grem2	Skyblue	Nanp
Greenyellow	Usp18	Skyblue3	Nek11
Grey60	Zcchc12	Steelblue	Crx
Ivory	Ep300	Tan	Zfp651
Lightcyan	Ptgds	Turquoise	Ap3m2
Lightcyan1	mt.Rnr2	Violet	Qk
Lightgreen	Gstm2	White	B230217O12Rik
Lightsteelblue1	Disp1	Yellow	Inpp4a
		Yellowgreen	Zswim1

TABLE 7 | Functional enrichment analysis of Hub genes in NF data subsets.

Category	Term	Count	*P*-value	Fold enrichment	FDR
BP	GO:0045944	9	9.13E-04	4.19373792	0.233863747
BP	GO:0006351	12	0.001284966	2.951560906	0.233863747
BP	GO:0007275	8	0.005066806	3.604594952	0.422224333
BP	GO:0030154	7	0.005358539	4.160880999	0.422224333
BP	GO:0006355	12	0.005799785	2.441286664	0.422224333
BP	GO:0032206	2	0.020822877	92.72820513	1
BP	GO:0006810	9	0.032934936	2.290213628	1
BP	GO:0042771	2	0.063185914	29.91232423	1
BP	GO:0016055	3	0.07364446	6.530155291	1
CC	GO:0005634	18	0.059352202	1.469995016	1
CC	GO:0032993	2	0.059738529	31.71290323	1
CC	GO:0005654	8	0.083476947	2.032248062	1
MF	GO:0003677	10	0.010936964	2.55286147	0.677053596
MF	GO:0008013	3	0.012979407	16.83976834	0.677053596
MF	GO:0003682	5	0.015065842	5.05915787	0.677053596
MF	GO:0032183	2	0.018423227	104.7807808	0.677053596
MF	GO:0000978	4	0.037404986	5.253632463	1
MF	GO:0035257	2	0.048397769	39.29279279	1
MF	GO:0003713	3	0.053050255	7.858558559	1
MF	GO:0004674	4	0.057674865	4.406668351	1
MF	GO:0000977	3	0.07599868	6.40063593	1
MF	GO:0016740	7	0.078508897	2.242251763	1
MF	GO:0005524	7	0.085817386	2.190175577	1
MF	GO:0035064	2	0.092637771	20.06440483	1
MF	GO:0004672	4	0.09568683	3.551890874	1

In this work, in order to reduce the running time of WGCNA, we calculated the standard deviation of the genes in each data set, and then left the genes with the largest standard deviation of the first 5000. The reason is that data with large variance contains the main biological information in the data, and it can also reduce the complexity of calculation. We first analyze two data subsets of NF and PEP, as shown in **Figure 3**. The Xin data set and the other two analysis results are shown in **Supplementary Material 2**.

Figure 3 shows the heat map of the module. Both the abscissa and the ordinate are genes, and the entire module represents the relationship between genes. On the left and top is the hierarchical clustering tree and module allocation. Red represents higher similarity, and yellow represents lower similarity. Since the module is composed of genes with high similarity, corresponding to the red area of the diagonal line in the figure, the target gene analysis and the correlation between the module and the trait can be performed for the module of interest.

Hub gene is a gene that plays a vital role in biological processes. In related pathways, the regulation of other genes is often affected by this gene. Therefore, hub gene is often an important target and research hot spot. We use *chooseTopHubInEachModule* in the WGCNA package to find the Hub genes in each module, and predict the gene function of the module through functional enrichment analysis. Here we show the Hub genes of the NF data type in **Table 6** and the results of the functional enrichment analysis in **Table 7**.

Differential Network Analysis With DiffCoEx

When we use DiffCoEx to analyze the difference network of each two types of Usoskin and Xin data, using the default parameters will lead to too many modules. In order to reduce the number of modules as much as possible, the gene is sampled, in other words, only 1/2 of the genes were randomly selected as the input to DiffCoEx, and the "cutHeight" parameter of the "mergeCloseModules" function was adjusted to 0.5 (default 0.2). Here, we only show the results of the two data types of NF-NP in the Usoskin dataset, as shown in **Figure 4**. The Xin data set and the other five analysis results are show in **Supplementary Material 3**.

The upper half of the main matrix in the figure shows the relationship between genes and genes in the NF subset, and the lower half shows the relationship between genes and genes in the NP subset. There are a total of 21 modules in the figure. Some modules have a higher expression level in NF and some have higher expression levels in NP. The color difference between the two sides is more obvious, indicating that this module has a large difference between NF and NP, which can be targeted at the difference. Analysis of the more obvious modules plays a vital role in the exploration of downstream target genes and drug prediction.

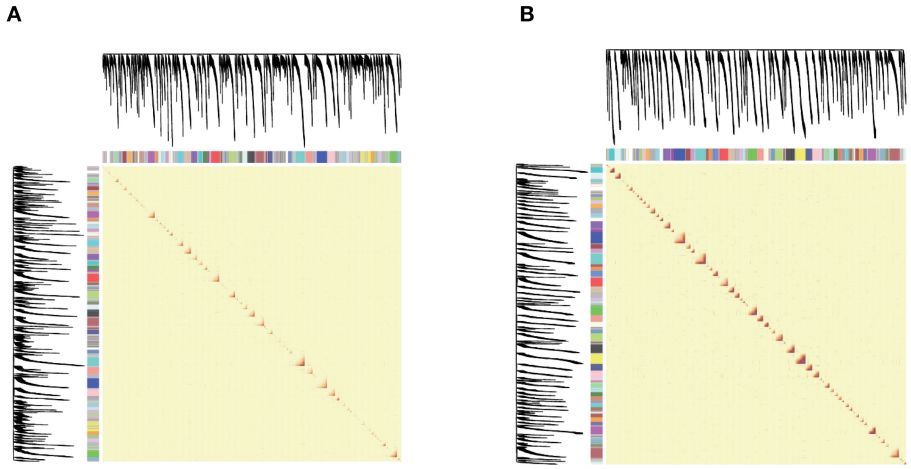

FIGURE 3 | (A,B) are network heat maps of NF and PEP, respectively. On the left side and top are the hierarchical clustering trees and modules of genes. In the figure, red represents higher similarity and yellow represents lower similarity. As the module is composed of genes with high similarity, it corresponds to the diagonal red in the figure.

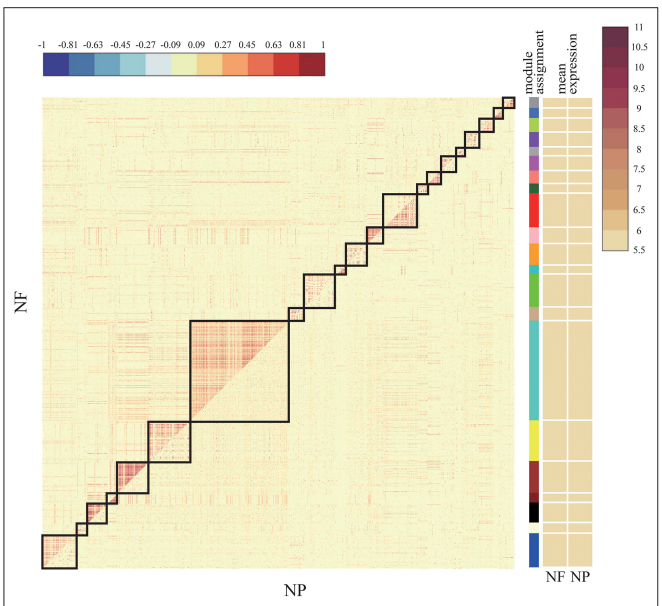

FIGURE 4 | Comparative correlation heat map of NF and NP. The upper diagonal of the main matrix shows a correlation between pairs of genes among the NF (the red color corresponds to positive correlations, blue to negative correlations). The lower diagonal of the heat map shows a correlation between the same gene pairs in the NP controls. Modules are identified in the heat map by black squares and on the right side of the heat map by a color bar. The brown bands on the right side indicate the mean expression of the modules in the NF (first column) and the NP (second column); darker colors indicate higher mean expression levels.

DISCUSSION

In recent years, the rapid development of single-cell sequencing technology can simultaneously measure the expression levels of tens of thousands of cells in a single experiment. Because of this, single-cell sequencing technology has developed rapidly in recent years. Although a large number of research methods have been developed for single-cell sequencing technology, there is no systematic framework on how to compare two single-cell clusters at the molecular level. Due to the difference in gene expression levels, different cells have different biological meanings and different physiological functions. Each gene is involved in a different biological process. It is not feasible to analyze all genes blindly to predict drugs and treat diseases. Therefore, analyzing data from the perspective of genetics plays an important role in clinical trials and scientific research. In this work, we performed a complete process analysis of scRNA-seq data at the molecular level. For example, through DEGs, we can know whether there are differences between different groups, and which genes are different. Furthermore, the functional enrichment analysis (GO, KEGG) of these differential genes was performed to explore the relevant signal pathways and the biological processes mediated by the differences in the expression of these genes. By constructing a gene regulatory network (WGCNA), it is helpful to understand the function of different genes and the interaction between genes as a whole, to better understand the gene expression mechanism inside cells, and to promote the research of disease pathology. By analyzing the difference modules in the entire gene regulatory network, exploring modules that contain more biological information provides effective guidance for the prediction of targeted genes and subsequent analysis.

This work mainly focuses on the analysis of the gene level in single-cell data, including the analysis of differential genes, the analysis of differential correlation, the construction of gene regulatory networks and the analysis of differential networks, without considering the internal dynamics between cells. How to effectively express the biological information contained in genes and cells in words is one of our future research directions. And due to the lack of relevant biological background knowledge, the

analysis and description of the analysis results and the regulatory relationship between genes are insufficient. At the same time, more algorithm models can be considered for constructing the relationship between genes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

WL, HA, and LC conceived, designed, and managed the study. WL and LC performed the experiments and drafted the manuscript. BW, CR, and AW provided computational support and technical assistance. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61803065, 11971347), and the Fundamental Research Funds for the Central Universities of China.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.648898/full#supplementary-material

REFERENCES

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., et al. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* 33, 155–160. doi: 10.1038/nbt.3102

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

Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., et al. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8, 1–12. doi: 10.1038/ncomms15081

Delmans, M., and Hemberg, M. (2016). Discrete distributional differential expression (D3E)–a tool for gene expression analysis of single-cell RNA-seq data. *BMC Bioinformatics* 17:110. doi: 10.1186/s12859-016-0944-6

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A. K., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* 16, 1–13. doi: 10.1186/s13059-015-0844-5

Hicks, S. C., Townes, F. W., Teng, M., and Irizarry, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19, 562–578. doi: 10.1093/biostatistics/kxx053

Iancu, O. D., Kawane, S., Bottomly, D., Searles, R., Hitzemann, R., and McWeeney, S. (2012). Utilizing RNA-Seq data for de novo coexpression network inference. *Bioinformatics* 28, 1592–1597. doi: 10.1093/bioinformatics/bts245

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* 11, 740–742. doi: 10.1038/nmeth.2967

Kim, D., Kobayashi, T., Voisin, B., Jo, J.-H., Sakamoto, K., Jin, S.-P., et al. (2020). Targeted therapy guided by single-cell transcriptomic analysis in drug-induced hypersensitivity syndrome: a case report. *Nat. Med.* 26, 236–243. doi: 10.1038/s41591-019-0733-7

Korthauer, K. D., Chu, L.-F., Newton, M. A., Li, Y., Thomson, J., Stewart, R., et al. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biol.* 17:222. doi: 10.1186/s13059-016-1077-y

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

Lee, M. C. W., Lopezdiaz, F. J., Khan, S. Y., Tariq, M. A., Dayn, Y., Vaske, C. J., et al. (2014). Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4726–E4735. doi: 10.1073/pnas.1404656111

Liang, Wang, S., Mo, X., Li, Y., He, J., Biology, Y. J. C., et al. (2020). Single-cell RNA-seq reveals the immune escape and drug resistance mechanisms of mantle cell lymphoma. *Cancer Biol. Med.* 17, 726–739. doi: 10.20892/j.issn.2095-3941.2020.0073

Liu, W., Li, L., Ye, H., and Tu, W. (2017). [Weighted gene co-expression network analysis in biomedicine research]. *Sheng Wu Gong Cheng Xue Bao* 33, 1791–1801. doi: 10.13345/j.cjb.170006

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., and Zhang, B. (2016). DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst. Biol.* 10:106. doi: 10.1186/s12918-016-0349-1

Miao, Z., Deng, K., Wang, X., and Zhang, X. (2018). DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics* 34, 3223–3224. doi: 10.1093/bioinformatics/bty332

Shalek, A. K., and Benson, M. (2017). Single-cell analyses to tailor treatments. *Sci. Trans. Med.* 9:eaan4730. doi: 10.1126/scitranslmed.aan4730

Soneson, C., and Robinson, M. D. (2018). Bias, robustness, and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi: 10.1038/nmeth.4612

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nat. Rev. Genet.* 20, 631–656. doi: 10.1038/s41576-019-0150-2

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

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., et al. (2010). Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* 6, 468–478. doi: 10.1016/j.stem.2010.03.015

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6, 377–382. doi: 10.1038/nmeth.1315

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics* 11:497. doi: 10.1186/1471-2105-11-497

Trombetta, J. J., Gennert, D., Lu, D., Satija, R., Shalek, A. K., and Regev, A. J. C. P. M. B. (2014). Preparation of single–cell RNA–seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* 107, 4.22.1–4.22.17. doi: 10.1002/0471142727.mb0422s107

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., et al. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.* 18, 145–153. doi: 10.1038/nn.3881

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nat. Methods* 14:565. doi: 10.1038/nmeth.4292

Wang, T., and Nabavi, S. (2018). SigEMD: a powerful method for differential gene expression analysis in single-cell RNA sequencing data. *Methods* 145, 25–32. doi: 10.1016/j.ymeth.2018.04.017

Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., et al. (2016). RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell *Metabol*. 24, 608–615 doi: 10.1016/j.cmet.2016.08.018

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500, 593–597. doi: 10.1038/nature12364

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20:1131. doi: 10.1038/nsmb.2660

Zhao, J., Guo, C., Xiong, F., Yu, J., and Zeng, Z. J. C. L. (2020). Single cell RNA-seq reveals the landscape of tumor and infiltrating immune cells in nasopharyngeal cancer. *Cancer Lett.* 477, 131–143. doi: 10.1016/j.canlet.2020.02.010

Conflict of Interest: BW and AW were employed by the company Geneis Beijing Co., Ltd. HA was employed by the company Guangzhou Anjie Biomedical Technology 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: 26 April 2021
doi: 10.3389/fgene.2021.665173

Identification and Validation of a Novel RNA-Binding Protein-Related Gene-Based Prognostic Model for Multiple Myeloma

Wei Wang[1], Shi-wen Xu[1], Xia-yin Zhu[2,3], Qun-yi Guo[2,3], Min Zhu[1], Xin-li Mao[3,4], Ya-Hong Chen[5], Shao-wei Li[3,4] and Wen-da Luo[1,2,3]**

[1] Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China, [2] Department of Hematology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China, [3] Key Laboratory of Minimally Invasive Techniques & Rapid Rehabilitation of Digestive System Tumor of Zhejiang Province, Taizhou, China, [4] Department of Gastroenterology, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China, [5] Health Management Center, Taizhou Hospital of Zhejiang Province Affiliated to Wenzhou Medical University, Linhai, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Feng Wang,
Emory University, United States
Xiuting Liu,
Washington University in St. Louis, United States

***Correspondence:**
Shao-wei Li
li_shaowei81@hotmail.com
Wen-da Luo
luowd@enzemd.com

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

Received: *07 February 2021*
Accepted: *29 March 2021*
Published: *26 April 2021*

Citation:
Wang W, Xu S-w, Zhu X-y, Guo Q-y, Zhu M, Mao X-l, Chen Y-H, Li S-w and Luo W-d (2021) Identification and Validation of a Novel RNA-Binding Protein-Related Gene-Based Prognostic Model for Multiple Myeloma. Front. Genet. 12:665173. doi: 10.3389/fgene.2021.665173

Background: Multiple myeloma (MM) is a malignant hematopoietic disease that is usually incurable. RNA-binding proteins (RBPs) are involved in the development of many tumors, but their prognostic significance has not been systematically described in MM. Here, we developed a prognostic signature based on eight RBP-related genes to distinguish MM cohorts with different prognoses.

Method: After screening the differentially expressed RBPs, univariate Cox regression was performed to evaluate the prognostic relevance of each gene using The Cancer Genome Atlas (TCGA)-Multiple Myeloma Research Foundation (MMRF) dataset. Lasso and stepwise Cox regressions were used to establish a risk prediction model through the training set, and they were validated in three Gene Expression Omnibus (GEO) datasets. We developed a signature based on eight RBP-related genes, which could classify MM patients into high- and low-score groups. The predictive ability was evaluated using bioinformatics methods. Gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment, and gene set enrichment analyses were performed to identify potentially significant biological processes (BPs) in MM.

Result: The prognostic signature performed well in the TCGA-MMRF dataset. The signature includes eight hub genes: *HNRNPC*, *RPLP2*, *SNRPB*, *EXOSC8*, *RARS2*, *MRPS31*, *ZC3H6*, and *DROSHA*. Kaplan–Meier survival curves showed that the prognosis of the risk status showed significant differences. A nomogram was constructed with age; *B2M*, *LDH*, and *ALB* levels; and risk status as prognostic parameters. Receiver operating characteristic (ROC) curve, C-index, calibration analysis, and decision curve analysis (DCA) showed that the risk module and nomogram performed well in 1, 3, 5, and 7-year overall survival (OS). Functional analysis suggested that the spliceosome pathway may be a major pathway by which RBPs are involved in myeloma development. Moreover, our signature can improve on the R-International

Staging System (ISS)/ISS scoring system (especially for stage II), which may have guiding significance for the future.

Conclusion: We constructed and verified the 8-RBP signature, which can effectively predict the prognosis of myeloma patients, and suggested that RBPs are promising biomarkers for MM.

Keywords: RBP, prediction, prognosis, multiple myeloma, model

INTRODUCTION

Multiple myeloma (MM) is a malignant clonal plasma cell disease of the bone marrow. The main clinical manifestations are monoclonal proteins in the blood or urine and related organ dysfunction (Palumbo and Anderson, 2011). Improved understanding of myeloma and the application of new treatment methods and drugs have greatly improved the survival of patients with myeloma. However, MM is a highly heterogeneous disease, both in response to treatment and in survival, for which the overall survival (OS) of patients ranges from less than 2 years to more than 10 years (Palumbo and Anderson, 2011; Sonneveld et al., 2016). This stark difference may be related to the heterogeneity of myeloma cell biology and multiple host factors (Greipp et al., 2005). Therefore, it is essential to identify disease-related biomarkers and use them to distinguish patients with different prognoses, which will be beneficial for formulating individualized treatments to cope with tumor heterogeneity, thereby improving patients' final prognosis.

Post-transcriptional gene regulation (PTGR) is a crucial biological process (BP). It is involved in maintaining cellular metabolism, coordinating the maturation, transport, stability, and degradation of all classes of RNAs (Gerstberger et al., 2014). RNA-binding proteins (RBPs) are involved in nearly all steps of PTGR, determining the fate and function of each transcript in the cell, and ensuring cellular homeostasis (Pereira et al., 2017). Gerstberger et al. identified 1542 RBP-associated genes, accounting for 7.5% of all protein-coding genes in humans, and half of these genes are involved in mRNA metabolic pathways. Eleven percent of the RBPs constitute ribosomal proteins, and the rest are involved in multiple non-coding RNA metabolic processes (Gerstberger et al., 2014). RBPs constitute a complex network with cancer-associated RNA targets, and these interactions maintain tumor growth, allowing them to escape death and become more invasive (Tu et al., 2015; Pereira et al., 2017). Overexpression of the *LIN28* paralog was shown to synergize with the Wnt pathway to promote aggressive intestinal adenocarcinoma development in mouse models; it has also been detected in a variety of other solid tumors and hematological malignancies. *LIN28*/*LIN28B* blocks let-7 microRNA (miRNA) biogenesis and, in turn, downregulates the expression of let-7 miRNA target genes, which play an important role in tumor progression and metastasis.

The International Staging System (ISS) distinguishes myeloma patients into stages I, II, and III by serum β_2 microglobulin and albumin (Greipp et al., 2005). However, this staging only considers the biochemical factors. The R-ISS staging groups patients into stages I, II, and III based on ISS staging, which integrates chromosomal abnormalities (CA) and serum lactate dehydrogenase (LDH) (Palumbo et al., 2015). Although R-ISS distinguishes patients with a good prognosis (stage I) from those with a poor prognosis (stage III), this staging classifies the larger cohort patients into stage II, which is composed of those who still show significant survival heterogeneity (Gonsalves et al., 2020). RBP-associated genes such as *DIS3* have been shown to be associated with myeloma prognosis (Boyle et al., 2020). Here, we identified several prognostically relevant differentially expressed genes (DEGs) for RBP by analyzing public databases and found that these molecular biomarkers can enrich the understanding of myeloma. We also performed Cox regression to construct an 8-gene prognostic model and nomogram that could effectively predict the survival of MM patients and found that this model could improve on the ISS and R-ISS staging ability.

MATERIALS AND METHODS

Data Processing and DEG Identification

All analyses in this study were conducted using R version 4.03. A list of 1542 RBP-related genes was obtained from a previous study (Gerstberger et al., 2014). Gene expression profiles GSE47552, GSE136337, GSE24080, and GSE57317 were downloaded from the Gene Expression Omnibus (GEO) database[1]. The data for MMRF-CoMMpass were obtained from The Cancer Genome Atlas (TCGA[2]). The array data of GSE47552 were obtained using the GPL6244 platform (HuGene-1.0-st Affymetrix Human Gene 1.0 ST Array). GSE136337 was obtained using the GPL27143 platform (HG-U133 Plus 2) Affymetrix Human Genome U133 Plus 2.0 Array; GSE24080 and GSE57317 were obtained using the GPL570 platform (HG-U133 Plus 2) Affymetrix Human Genome U133 Plus 2.0 Array. The data of GSE47552 included bone marrow samples from five healthy donors and 41 newly diagnosed patients with MM. DEGs between MM patients and healthy donors were identified using the R package "limma." Genes with $P < 0.05$, and [$\log_2$FoldChange ($\log_2$FC)] > 1 were considered as DEGs. Volcanic maps and heat maps were drawn using the R package "ggplot2" and "pheatmap" to visualize DEGs.

The Cancer Genome Atlas-MMRF was used as a training set to develop a prognostic signature, while GSE136337, GSE24080, and GSE57317 were used for validation. To meet the needs of this

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

[2]https://tcga-data.nci.nih.gov/

analysis, we set the following conditions to control data quality: (1) Samples must have complete survival information, including survival status and OS time, where death had to be tumor-related and OS time had to be greater than 30 days (2). Samples must have complete R-ISS or ISS information. Finally, 709 cases of MMRF, 559 cases of GSE24080, 559 cases of GSE136337, and 55 cases of GSE57317 were selected for subsequent analysis.

Gene Ontology and KEGG Enrichment Analysis of DEGs

Gene ontology (GO) term analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis were performed using the R package "clusterProfiler" to identify the functional roles of the upregulated and downregulated DEGs, respectively. GO enrichment was described from three sub-ontologies: BP, molecular function (MF), and cellular component (CC).

Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) version 4.1.0 was used to explore significant BPs between patients in different risk groups. KEGG gene sets as Gene Symbols[3] were chosen as the gene set database and the cut-off values for the significance of outcomes were FDR < 0.25, NOM $P < 0.05$, and $|\text{NES}| > 1$.

RNA-Binding Protein-Related Gene Signature Construction

Screening for Hub Genes in the Training Dataset

The TCGA dataset was used as the training cohort, and three datasets (GSE136337, GSE24080, and GSE57317) were used for validation. Univariate Cox regression analysis and multivariate regression analysis (Cox, 1972) were chosen to screen for RBP-related genes that were closely related to the OS of patients. In the univariate Cox regression analysis, $P < 0.05$ was the criterion to screen candidate genes. Next, the least absolute shrinkage and selection operator (Lasso) (Friedman et al., 2010) regression model was applied to minimize overfitting and identify the most significant survival-associated DEGs of RBP-related genes in myeloma. Stepwise multivariate Cox regression analysis was then applied to further establish the RBP-related risk signature. Finally, the hazard ratios (HRs) and regression coefficients of every gene were calculated, and the satisfactory ones were chosen.

Construction of the Gene-Related Prognostic Signature in the Training Dataset

The prognostic risk-score signature for prognosis prediction of MM patients was to multiply the expression level of each selected prognostic gene by its corresponding relative regression coefficient weight as follows:

Risk score $= \sum_{i=1}^{N} \beta i \times Ei$: ($N$ represents the total number of signature genes, and βi and Ei represent the coefficient index and the gene expression value of each gene, respectively).

The risk score of each patient and the median risk score were calculated using the training dataset. Those with a higher risk score than the median were classified into the high-score group, while those with a lower risk score were classified into the low-score group. Kaplan–Meier survival curves (Ranstam and Cook, 2017) and receiver operating characteristic (ROC) curves (Kamarudin et al., 2017) of the two groups were plotted to evaluate the sensitivity and specificity of the signature we established.

Validation of the Gene-Related Prognostic Signature's Efficacy in the Validation Datasets

As in the training set, the patients in the validation datasets were classified into the high- and low-score groups by comparing the risk score of each patient with the calculated median risk score from each dataset. The time-dependent prognostic values of the gene signature were investigated using the Kaplan–Meier curve and log-rank test (Kleinbaum, 1998) was used to compare the survival difference between the above-mentioned high- and low-score groups.

Construction of the Nomogram

In the GSE24080 dataset, we used the lasso regression analysis to analyze all clinical factors and finally selected the clinical prognostic factors together with risk status as the prognostic parameters, ensuring that the nomogram model will not overfit. Then, through "rms" and "regplot" R packages, a prognostic nomogram was established to evaluate the probability of OS in MM patients at 1/3/5/7 years with the regression coefficients based on the lasso analysis. Calibration plots were used to evaluate the discriminative ability of the nomogram. Harrell's concordance index (C-index) was used to verify the nomogram performance. The ROC curve and calibration curve varying with time were also drawn to estimate the accuracy of the actual observed rate with the predicted survival for 1/3/5/7-year OS of the nomogram. In addition, the clinical application prospects of the eight-gene prognostic signature were determined through decision curve analysis (DCA) (Vickers and Elkin, 2006).

RESULTS

Identification of DEGs

We set a $P < 0.05$, and [$\log_2$FoldChange ($\log_2$FC)] > 1 as the cut-off criterion. Based on this standard, we identified 866 DEGs in MM cases compared with healthy donors, among which 202 were considered significantly upregulated, and 664 were considered significantly downregulated. The volcano plot of DEGs and the heat map of the top 200 DEGs are shown in **Figures 1A,B**. As shown in **Figure 1C**, we obtained 96 differentially expressed RBPs by taking the intersection of DEGs and 1,542 RBPs.

Functional Analysis of Differential RBP Genes

For exploring the potential function of these differentially expressed RBPs, we performed GO and KEGG enrichment analysis using the R package "clusterProfiler." The results of the GO enrichment analysis are presented in three parts. For BP, differentially expressed RNA-binding proteins

[3]http://www.gsea-msigdb.org/gsea/downloads.jsp

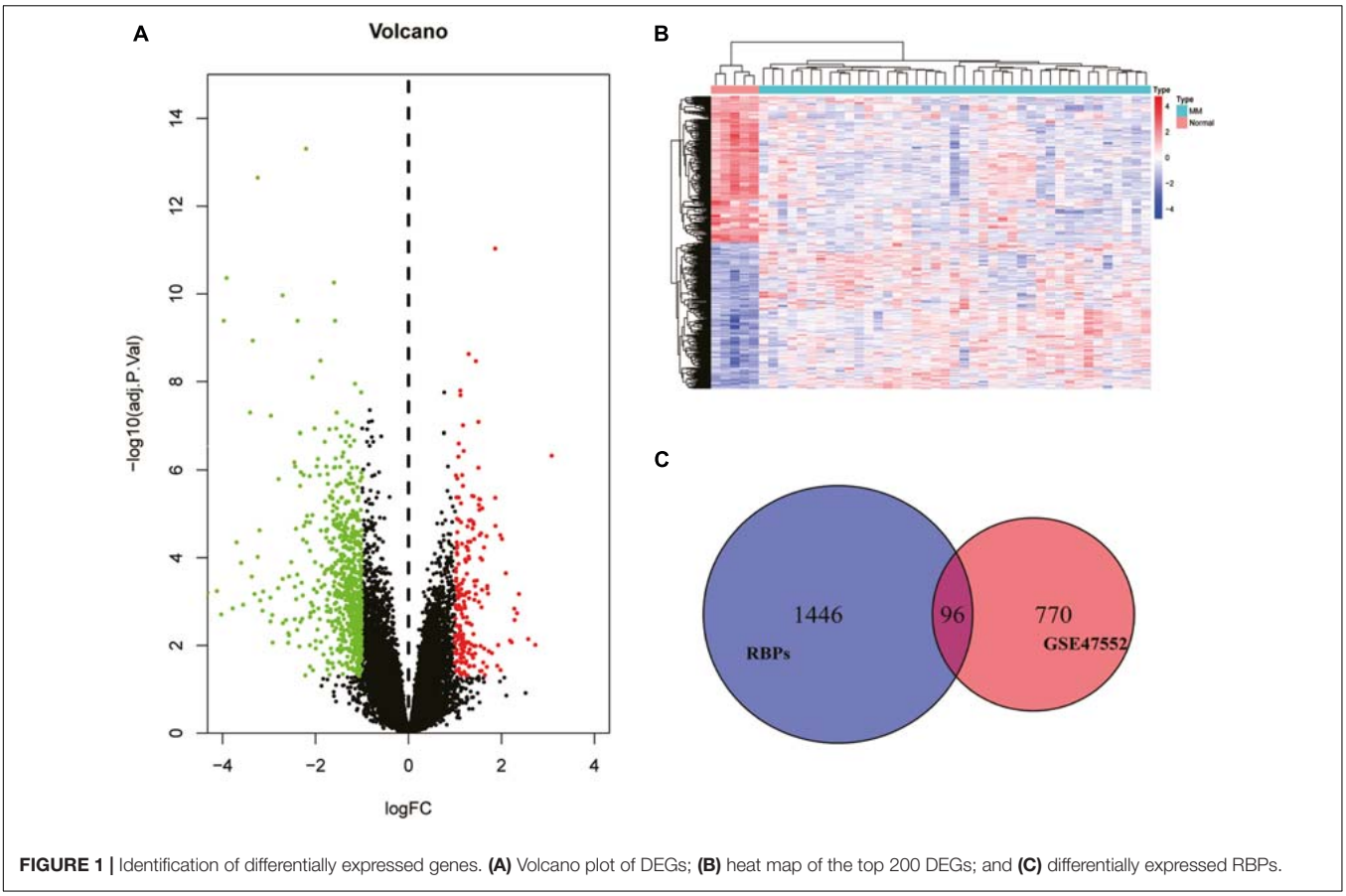

FIGURE 1 | Identification of differentially expressed genes. **(A)** Volcano plot of DEGs; **(B)** heat map of the top 200 DEGs; and **(C)** differentially expressed RBPs.

(DERBPs) were significantly associated with the following terms: RNA catabolic process, mRNA catabolic process, nuclear-transcribed mRNA catabolic process, translational initiation, nuclear-transcribed mRNA catabolic process, nonsense-mediated decay, other important BPs, SRP-dependent cotranslational protein targeting to membrane, cotranslational protein targeting to membrane, protein targeting to ER, the establishment of protein localization to the endoplasmic reticulum, and protein localization to the endoplasmic reticulum (**Figure 2A**). Four of the top five BP terms were related to various RNA catabolic meaning that these processes may be involved with MM disease progression. The CCs analysis indicated that DERBPs were mostly involved in the following terms: ribosome, ribosomal subunit, cytosolic ribosome, large ribosomal subunit, cytosolic large ribosomal subunit, small ribosomal subunit, cytoplasmic ribonucleoprotein granule, cytosolic small ribosomal subunit, polysome, and the polysomal ribosome (**Figure 2B**). MF terms were mainly enriched for the structural constituent of ribosome, catalytic activity (acting on RNA), mRNA 3'-UTR binding, rRNA binding, ribonuclease activity, ribonucleoprotein complex binding, translation regulator activity, telomerase RNA binding, nucleocytoplasmic carrier activity, and Ran GTPase binding (**Figure 2C**). The ribosome, coronavirus disease – COVID-19, RNA degradation, spliceosome, ribosome biogenesis in eukaryotes, and RNA transport pathway terms were significantly enriched in DERBPs, as shown by KEGG enrichment analysis (**Figure 2D**).

Exploration of the Prognostic RBPs in MM

We enrolled 709 patients with a follow-up time of more than 30 days from TCGA as the training dataset for the construction of the signature. Although 96 differentially expressed RBPs were screened before (**Figure 1C**), only 94 of them were included in the TCGA dataset. The prognostic significance of the 94 genes was investigated using univariate Cox regression. As a result, 34 prognostic-associated candidate RBPs were obtained ($P < 0.05$) (**Table 1**). LASSO regression was then performed to identify 34 candidate genes closely related to the prognosis of MM patients, including the following 19 genes: *HNRNPC*, *RPLP2*, *SNRPB*, *SNRPE*, *SF3B3*, *KPNB1*, *GAPDH*, *RPS12*, *NFX1*, *MTIF3*, *CIRBP*, *EXOSC8*, *RARS2*, *MRPS31*, *ZC3H6*, *DROSHA*, *NAT10*, *LSM5*, and *PRIM*1 (**Supplementary Figure 1**). To further screen out the RBPs with the greatest prognostic value, a multiple stepwise Cox regression was conducted to investigate their impact, and eight hub RBPs, *HNRNPC*, *RPLP2*, *SNRPB*, *EXOSC8*, *RARS2*, *MRPS31*, *ZC3H6*, and *DROSHA* were selected to construct the risk model in MM patients (**Figure 3A**). All of the above genes showed an independent prognostic effect ($P < 0.05$). Among them, *HNRNPC*, *SNRPB*, *EXOSC8*, and *DROSHA* may be regarded as

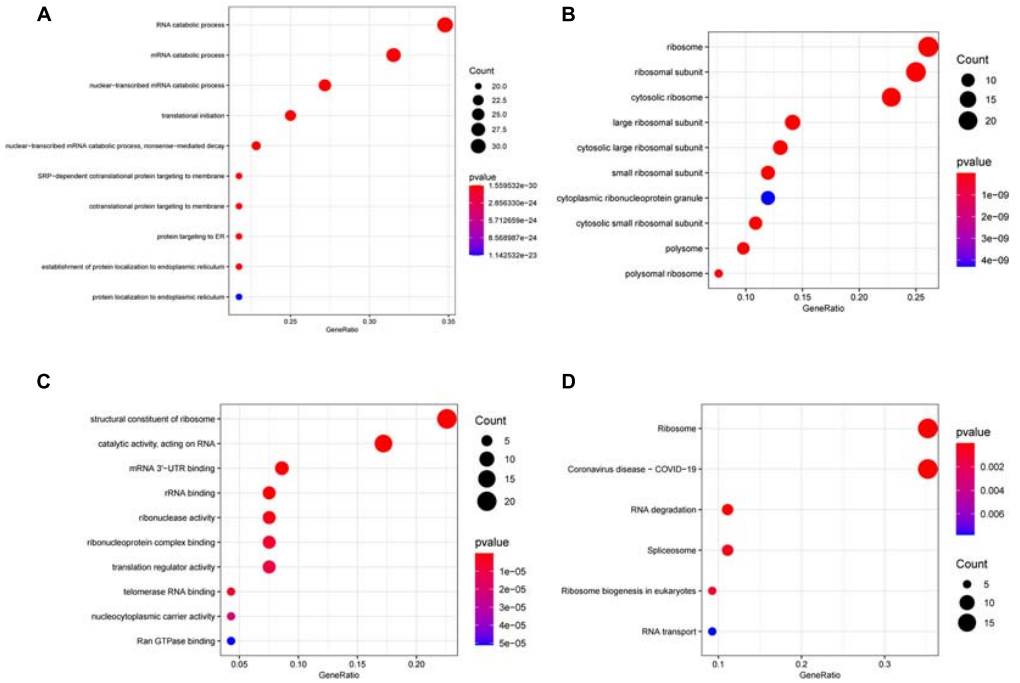

FIGURE 2 | Functional enrichment analysis of DEGs showed by bubble plots. **(A–C)** Three sub ontologies of GO enrichment analysis. **(A)** The biological process (BP) enrichment analysis. **(B)** The cellular component (CC) enrichment analysis. **(C)** The molecular function (MF) enrichment analysis. **(D)** The KEGG enrichment analysis.

oncogenes, whereas *RPLP2*, *RARS2*, *MRPS31*, and *ZC3H6* may be tumor suppressor genes. The coefficients of these genes indicated their impact on survival prediction.

Construction and Validation of the RBP Prognostic Signature

We used the eight hub RBPs selected by multiple Cox regression to establish the eight-gene predictive signature in the TCGA dataset. The risk score for each patient was calculated based on the expression level and the corresponding beta value using the following formula:

$$\text{Risk score} = (-0.6071) \times \text{Exp}ZC3H6 + (0.9575) \times \text{Exp}SNRPB + (-0.4821) \times \text{Exp}RPLP2 + (-0.5116) \times \text{Exp}RARS2 + (-0.4890) \times \text{Exp}MRPS31 + (0.7192) \times \text{Exp}HNRNPC + (0.5315) \times \text{Exp}EXOSC8 + (0.9987) \times \text{Exp}DROSHA$$

We then divided MM patients into the low-score group (n = 355) and high-score group (n = 354) based on the median risk score as the cut-off point. The patients' gene expression levels, status, and survival time are shown in **Figures 3B–D**. The K-M results showed that the OS rate of patients in the high-score group was significantly lower than that in the low-score group ($P < 0.001$, **Figure 3E**). In addition, the time-dependent ROC curve showed that the area under the ROC curve (AUC) of this risk score signature at 1, 2, 3, 4, and 5 years were 0.78, 0.74, 0.77, 0.77, and 0.81, respectively (**Figure 3F**), indicating that this signature has moderate performance.

To verify the predictive value of the 8-gene signature in other MM cohorts, we performed a similar analysis in three datasets: GSE136377, GSE24080, and GSE57317, which all included the risk-related genes selected. The risk score formula described above was validated for the three datasets. We only compared the OS differences of 1–5 years in the TCGA dataset. But when the risk model was applied to the GSE24080 and GSE136337 datasets, comparing for up to 10 years, the results showed that the OS of patients in the high-score group was worse than that of patients in the low-score group (all $P < 0.01$) (**Figures 4A,C**). The AUC of this risk score signature was >0.6, proving the performance of this scoring system (**Figures 4B,D**). More interestingly, in the GSE57317 dataset, the OS difference between the two groups was very significant, with AUC values of 0.84,0.88, and 0.96 at 1, 2, and 3 years, respectively, proving the prognostic value in this dataset (**Figures 4E,F**). In conclusion, this scoring model exhibited acceptable performance for all three datasets.

Establishment and Validation of Nomogram Survival Model

Univariate and Multivariate COX Regression Analysis of the Model

Univariate and multivariate Cox regression analyses were performed using clinical data from the GSE24080 dataset. Using univariate Cox regression analysis, age, *B2M*, *CRP*, *LDH*, *ALB*, *HGB*, and risk score status were selected to assess the independent prognostic factors in the MM sample (**Figure 5A**). Multivariate Cox regression analysis confirmed that age (HR = 1.02, 95% CI [1.00−1.03]; P = 0.042), *B2M* (HR = 1.41, 95% CI [1.17−1.69]; P = 0.000354), *LDH* (HR = 1.00, 95% CI [1.00−1.01]; $P = 6.89 \times 10^{-8}$), *ALB* (HR = 0.60, 95% CI [0.42−0.86]; P = 0.005), and multigene

TABLE 1 | Unicox results of differential RNA-binding proteins.

Gene	Hazard ratios	CI95	*P*-value
SUPT4H1	1.76	1.02–3.06	0.043
HNRNPC	3.67	2.1–6.43	0
RPLP2	0.66	0.5–0.89	0.007
SNRPB	5.09	3.39–7.64	0
EIF3K	0.62	0.43–0.9	0.012
GEMIN5	1.95	1.31–2.89	0.001
SNRPE	2.54	1.79–3.61	0
UTP6	2.7	1.59–4.61	0
SF3B3	2.24	1.39–3.62	0.001
KPNB1	3.97	2.46–6.39	0
GAPDH	2.14	1.51–3.03	0
CNOT1	1.65	1.06–2.58	0.027
DDX17	0.77	0.6–0.98	0.033
NFX1	0.54	0.35–0.83	0.005
MTIF3	0.52	0.38–0.72	0
CPSF2	1.75	1.09–2.81	0.02
NOL11	1.88	1.19–2.98	0.007
ESF1	2.57	1.57–4.2	0
CIRBP	0.5	0.35–0.72	0
EXOSC8	1.9	1.28–2.84	0.002
DDX21	1.89	1.31–2.73	0.001
INTS2	2	1.32–3.01	0.001
RARS2	0.58	0.38–0.88	0.01
MRPS31	0.59	0.46–0.77	0
ZC3H6	0.45	0.29–0.71	0.001
RPF2	2.08	1.45–2.99	0
DROSHA	2.16	1.3–3.6	0.003
NAT10	1.84	1.18–2.89	0.007
XPO1	2.14	1.38–3.3	0.001
LSM5	1.7	1.05–2.74	0.032
PRIM1	2.66	1.97–3.58	0
CPEB2	0.7	0.52–0.93	0.014
SLIRP	1.69	1.04–2.77	0.035
DARS2	1.78	1.35–2.33	0

CI95: 95% confidence interval.

risk status (HR = 1.78; 95% CI [1.29−2.47]; P = 0.000438) were significant independent risk factors (**Figure 5B**). Based on the results shown in **Figure 3C**, the risk score can be used as an independent prognostic factor without being affected by clinicopathological features. The HR of the high-risk group was 1.78 (95% CI: 1.29−2.47) times higher than that of the low-risk group.

Nomogram Construction

To establish a clinical method to predict the survival probability of MM patients, we created a nomogram using lasso regression analysis to estimate the probability of, 1-, 3-, 5-, and 7-year OS with age, *B2M*, *LDH*, *ALB*, and risk score status (**Figures 6A,B**). The AUC of 1-, 3-, 5-, and 7-year OS predictions were 0.78, 0.75, 0.70, and 0.77, respectively (**Figure 6E**). The calibration curve was used to describe the prediction value of the nomogram, and the 45° line indicates the actual survival outcomes. The results for predicting 1-, 3-, 5-, and 7-year OS showed that the nomogram-predicted survival closely matched the best prediction performance (**Figure 6D**), indicating that the nomogram has a significant predictive value for predicting 1-, 3-, 5-, and 7-year OS in patients with MM. The concordance index (C-index) was calculated to evaluate the prognostic capability of the model. The C-index of the nomogram was 0.71 (95% CI [0.69–0.73]), which proved that the nomogram's value a good predictive tool for MM prognosis. We used DCA analysis to confirm the range of the threshold probabilities for a prediction model. As shown in **Figure 6C**, the nomogram threshold probability based on 8-gene combinations was significantly better than the default strategies of treating all or none at a threshold probability > 0.05.

Validation of Classification Capabilities of the Eight-Genes Prognostic Signature for R-ISS and ISS Stage II Patients

To assess whether our model could improve the heterogeneity of patients with R-ISS stage II, we reclassified R-ISS stage II patients in GSE136337 based on the model. Finally, 122 of 267 patients were redefined as II-High, while 145 were defined as II-Low, and the survival curves were subsequently plotted. To highlight the discriminatory effect, we defined the categorized patients as R-ISS II co-plotted in graphs. As shown in **Figure 7**, stage II patients were clearly divided into two groups with different survival, and patients defined as II-High had a worse prognosis. Meanwhile, we found that the classifier also optimized for ISS stage II in GSE136337, and it was validated in two other independent datasets (**Figures 7B–D**). We treated R-ISS stage I and stage III in the same way, but the discrimination was not ideal (**Supplementary Figures 2A,B**). To further evaluate whether a similar effect would be apparent on ISS, we performed the same reclassification for the TCGA, GSE24080, and GSE136337 datasets. In GSE136337, the results were not significant for either ISS stage I or stage III ($P > 0.05$). For GSE24080, the difference in survival after grouping was significant only for ISS stage I ($P = 0.033$), but the effect of differentiation was not good enough. Surprisingly, the TCGA dataset performed the best in the three datasets. Although each dataset performed differently, the overall results were not as good as those of stage II (**Supplementary Figures 2C–H**).

Signaling Pathways Analysis of High-Risk Group

In our study, patients in the high-risk group exhibited worse survival. We used GSEA to investigate the potentially important pathways causing different prognoses in the two groups. A KEGG functional enrichment analysis showed that the base excision repair, nucleotide excision repair, spliceosome, cell cycle, and p53 signaling pathways may be involved in cancer development (**Figure 8**). The spliceosome pathway also appeared in the KEGG enrichment results of DERBPs, further demonstrating the importance of this pathway.

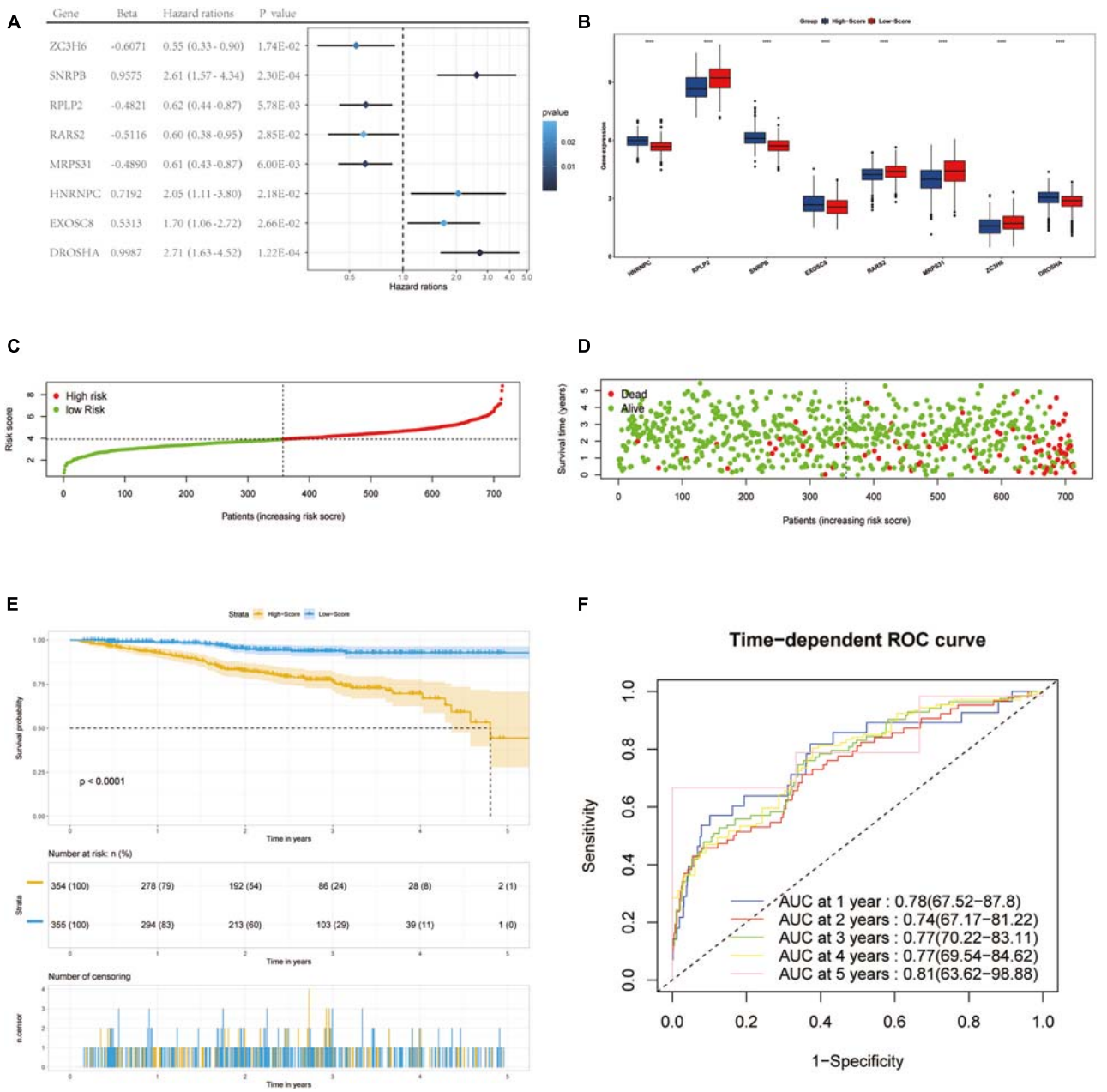

FIGURE 3 | Forest plots of the multivariate Cox regression analysis, the boxplot of eight RBP expression levels, the distribution of risk score, the living status, Kaplan–Meier analysis, and ROC analysis of the eight-gene signature of MM patients in the TCGA cohort. **(A)** Forest plot of the multivariate Cox regression analysis of OS of eight genes. Beta values represent the coefficient index β for each gene. **(B)** The boxplot of eight RBP expression levels in the training set (blue: low-score group. Red: high-score group). **(C)** The distribution of risk scores in the TCGA training set. **(D)** The living status of MM patients in the TCGA training set. **(E)** Kaplan–Meier survival analysis of the low-score and high-score group patients. **(F)** ROC curve analysis according to the 1-to-5-year survival of the area under the AUC value in the training TGCA cohort.

DISCUSSION

With the development of novel diagnostic approaches and treatment strategies, the survival of patients with MM has improved. However, MM remains an incurable disease for the vast majority of patients (Rajkumar, 2020). To ensure the predictive value of RBP-associated genes, we first screened for RBPs with significant differences in expression between newly diagnosed myeloma patients and normal human bone marrow. Subsequently, an eight-gene prognostic signature was established based on the expression levels of RBP-associated genes. By calculating the risk scores, we divided all patients into high- and low-score groups in the training dataset and three validation datasets, respectively. The predictive ability of this scoring model

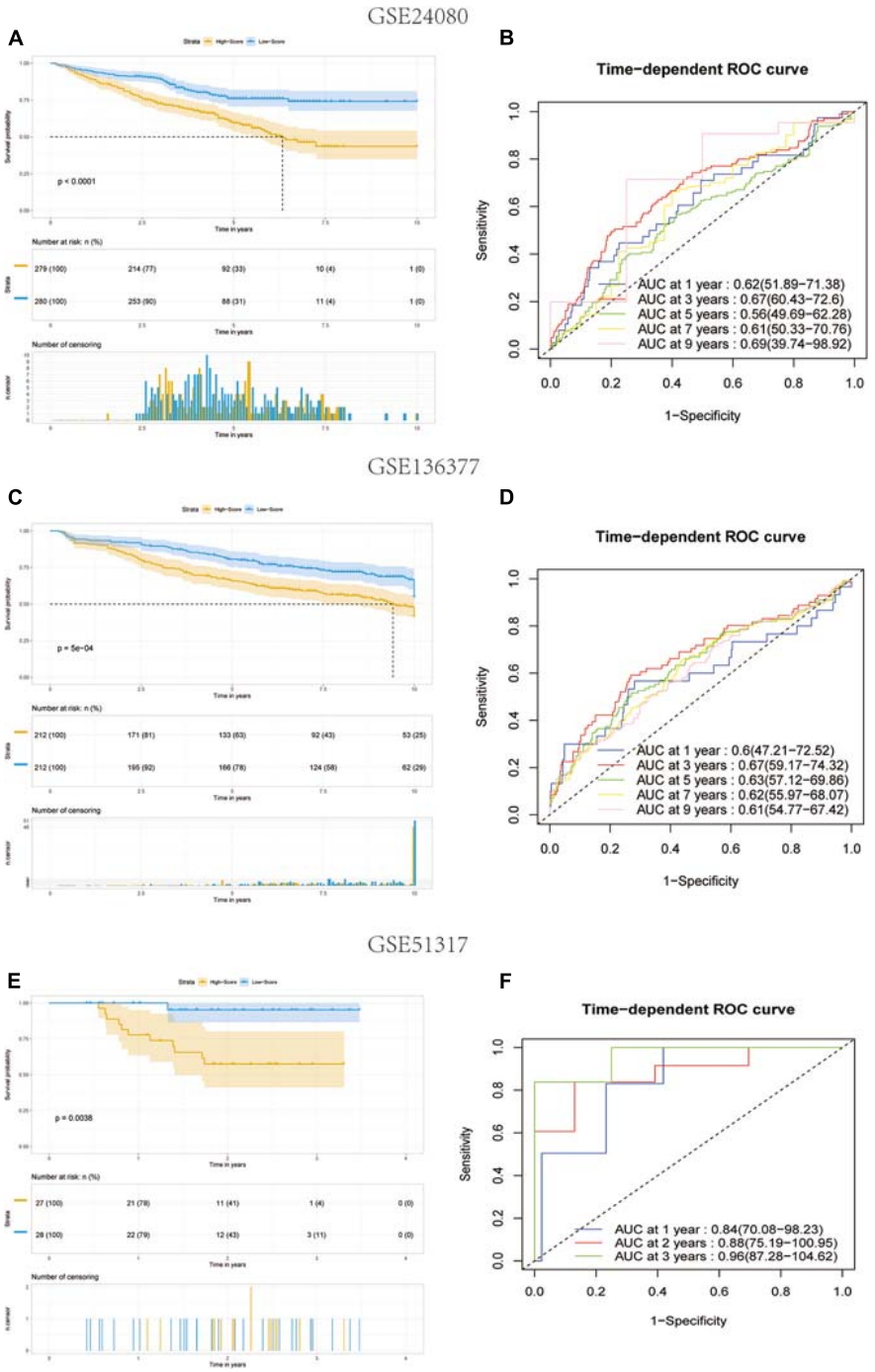

FIGURE 4 | Kaplan–Meier analysis and ROC analysis of 8-gene signature in three validation datasets. **(A,B)** Kaplan–Meier survival analysis of the low-score and high-score group patients and ROC curve analysis according to the 1-, 3-, 5-, 7-, and 9-year survival of the AUC value in the GSE24080 cohort. **(C,D)** Kaplan–Meier survival analysis of the low-score and high-score group patients and ROC curve analysis according to the 1-, 3-, 5-, 7-, and 9-year survival of the area under the AUC value in the GSE136337 cohort. **(E,F)** Kaplan–Meier survival analysis of the low-score and high-score group patients and ROC curve analysis according to the 1-, 2-, and 3-year survival of the AUC value in the GSE51317 cohort.

was evaluated and verified in the training set and the three validation datasets. Meanwhile, we built a nomogram survival model to predict the 1/3/5/7-year survival rate by combining age, *B2M*, *LDH*, *ALB*, and risk score status.

The role of RBPs in promoting cancer has been confirmed, and *DROSHA*, *EXOSC8*, *HNRNPC*, *MRPS31*, *RPLP2*, and *SNRPB* have also been reported to be related to the occurrence and development of a variety of tumors. *DROSHA* and DICER

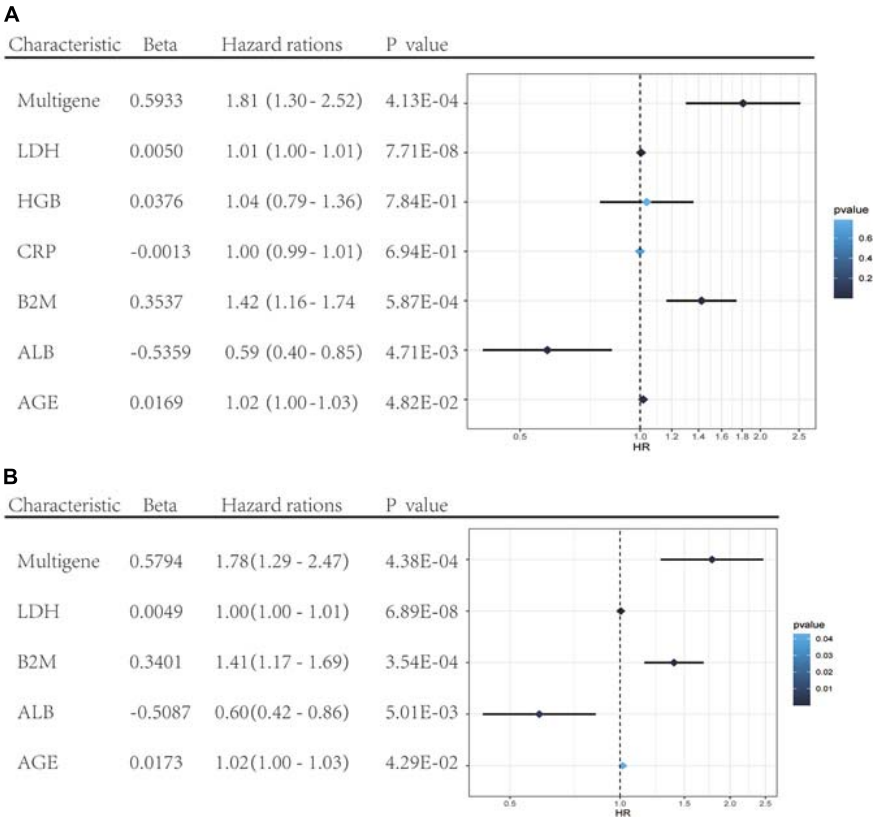

FIGURE 5 | Forest plots of the multivariate and univariate Cox regression analysis in GSE24080 cohorts. **(A)** Forest plot of the univariate Cox regression analysis OS of the clinical factors and risk score. **(B)** Forest plot of the multivariate Cox regression analysis OS of clinical factors screened by univariate Cox analysis and risk score. Beta values represent the coefficient index β for each clinical factor.

are important factors involved in miRNA processing. For neuroblastoma, the expression level of *DROSHA* decreased in advanced-stage patients and was associated with poor prognosis (Lin et al., 2010). *EXOSC8* is an essential component of the exosome complex and is involved in RNA surveillance and epigenetic regulation. Cui et al. found that the expression of *EXOSC8* in colorectal cancer was higher than that in normal tissues in a public database, indicating a poor prognosis. They confirmed that the expression of *EXOSC8* in colorectal cancer was higher than that in matched normal tissues in clinical samples, and verified the cancer-promoting effect of the gene in cell and animal experiments (Cui et al., 2020). As an RBP, *HNRNPC* was reported to be aberrantly expressed at elevated levels in a variety of tumors, besides being involved in some well-established BPs, such as RNA splicing. Further, it was found to control endogenous dsRNA and downstream interferon response functions and is indispensable to a subset of breast cancer cell lines, and partial suppression of this gene can affect cell line activity (Wu et al., 2018). Xu et al. (2014) found that mitochondrial ribosomal protein S31 (*MRPS31*) was associated with thyroid cancer disease progression using a machine-learning method. Ribosomal P2 (*RPLP2*) is an ancient ribosomal stalk protein. It has been shown that *RPLP2* can alleviate ribosome pausing in the DENV envelope coding sequence, thus enhancing protein stability. This effect is achieved by improving the efficiency of co-translational folding. *RPLP2* also influences multipass transmembrane protein biogenesis, making it important in protein synthesis. Moreover, it is associated with DNA repair, proliferation, apoptosis, and tumorigenesis, and is significantly associated with malignancies such as gynecological tumors, digestive system tumors, and lung adenocarcinoma (Campos et al., 2020). The *SNRPB* of SMB/B', the core member of the spliceosome mechanism, promotes cell proliferation and inhibits cell apoptosis. Changes in the core splice protein encoded by *SNRPB* may interrupt RNA processing, resulting in specific changes in the splice of variable exons, thus affecting the entire transcription process (Correa et al., 2016). Besides its important role in splicing, *SNRPB* mutations also have significant effects on cell division and DNA repair (Kittler et al., 2004). *SNRPB* is associated with poor prognosis in a variety of cancers, including glioblastoma, non-small cell lung cancer, and metastatic prostate cancer (Yi et al., 2009). Although the above-mentioned genes have been reported in a variety of cancers, their precise roles in myeloma remain unknown; thus, our study may provide direction for further exploration. *RARS2* encodes mitochondrial arginine tRNA synthetase, a protein essential for the translation of all mitochondrially synthesized proteins (Edvardson et al., 2007). Mutation of the *RARS2* gene causes destructive effects

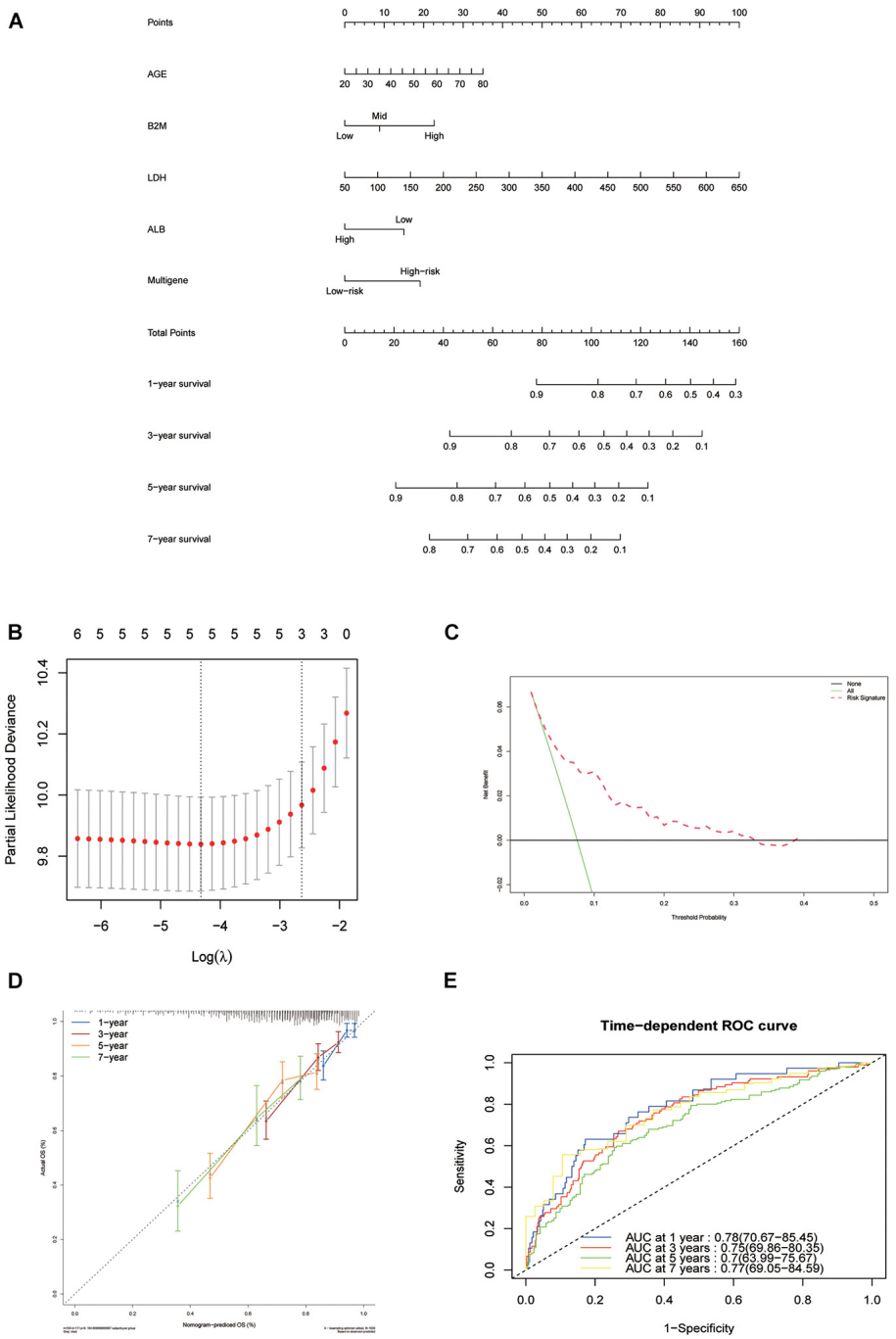

FIGURE 6 | Nomogram construction based on the eight-gene signature and prognostic value of genes. **(A)** The nomogram for predicting the proportion of patients with 1-, 3-, 5-, and 7-year OS of MM. **(B)** LASSO regression analysis used tenfold cross-validation via the maximum criteria. **(C)** Decision curve analysis of nomogram predicting 1-, 3-, 5-, and 7-year OS of MM. **(D)** Calibration plots of the nomogram. **(E)** Time-dependent ROC analysis of nomogram predicting 1-, 3-, 5-, and 7-year OS of MM.

on the cerebellum and cerebellum-associated nuclei (inferior olivary nuclei, pontine base, and dentate nuclei), leading to degenerative changes in the brain. However, the exact mechanism of this effect remains to be elucidated (Joseph et al., 2014). *ZC3H6* is a zinc finger transcription factor, but little is known about its function or expression. However, we found that *ZC3H6* may be closely related to the prognosis of patients with MM. This finding has not been mentioned in previous literature, so it may be a potential research direction in the future. In previous studies, *RARS2* and *ZC3H6* have not been reported

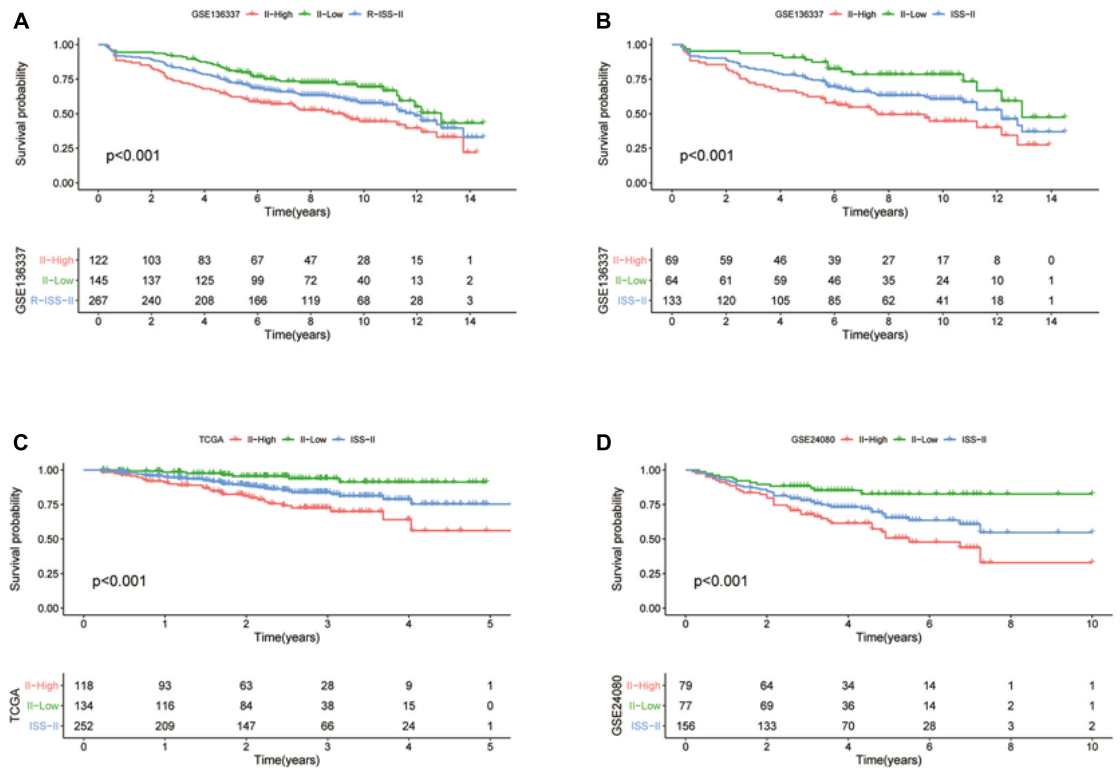

FIGURE 7 | The eight-gene model can enhance the predictive power of R-ISS and ISS for their respective stage II cohorts. **(A)** R-ISS stage II in GSE136337. **(B–D)** ISS stage II in GSE136337, TCGA-MMRF, and GSE24080. (Red: a group that was reclassified as high risk. Green: a group that was reclassified as low risk. Blue: total group before reclassification.)

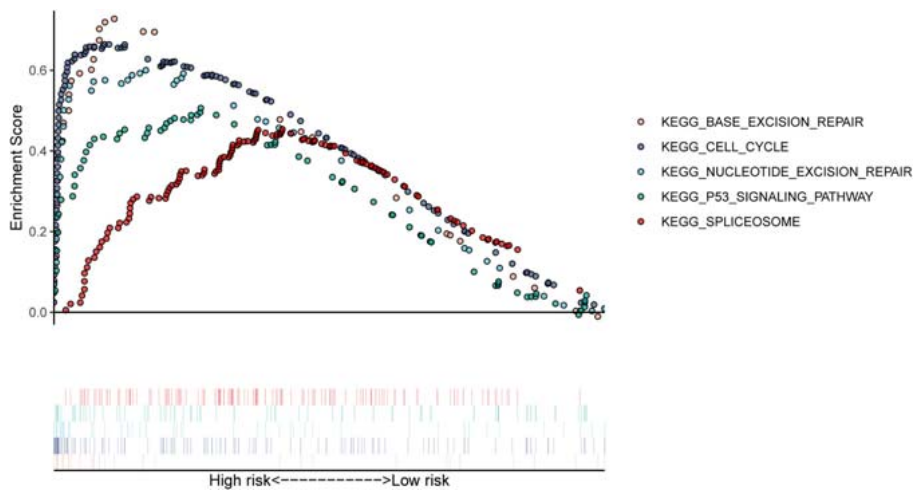

FIGURE 8 | The KEGG pathways were enriched in the high-risk group by performing the GSEA analysis.

to be associated with tumors. In our study, these two genes were differentially expressed in myeloma and correlated with patient survival, suggesting that these two genes are potential tumor-related genes that require further investigation. The vast majority of these eight RBP genes were first reported to be associated with myeloma, and in the future, we intend to establish a real-world cohort of MM patients to validate the value of these genes again.

To explore how RBPs are involved in the development and progression of MM, we performed GO and KEGG enrichment analyses of 96 DERBPs. In the GO enrichment analysis section, the results of enrichment from BP, CC, and MF are described.

The BP results show that the RNA catabolic process was the most significantly enriched result. The CC results suggest that RBPs are mainly localized in the ribosome and its associated locations. MF then reflects the involvement of RBPs in the structural conformation of the ribosome, RNA catalytic activity, and other important MPs. KEGG indicated that RBPs affected the disease by participating in the ribosome, RNA degradation, spliceosome, and RNA transport pathways. The results of conducting enrichment analysis only on DEGs may miss the contribution of genes that are relevant but less biologically significant to disease, so we further analyzed the differences in BPs between high-risk and low-risk groups by GSEA. In the GSEA-KEGG results, as with the results of the KEGG enrichment analysis of the DERBPs, the "Spliceosome pathway" was suggested to be significant. The RNA splicing pathway is associated with a variety of human tumors (Wang and Aifantis, 2020). In MM, aberrant RNA splicing patterns were found to exist, and patients with a large number of novel splice loci tended to have worse survival outcomes, which could be used to distinguish extremely high-risk groups (Bauer et al., 2020). These findings fit our enrichment results, demonstrating the value of the spliceosome pathway in myeloma, but there are currently few relevant studies, and its role in myeloma remains to be comprehensively uncovered. One study showed that spliceosome interference was an unreported mechanism of action of proteasome inhibitors; inhibition of the spliceosome could synergize with carfilzomib to potentiate antitumor effects, suggesting that targeted spliceosome therapy could serve as a future research direction for the treatment of myeloma (Huang et al., 2020).

R-ISS staging had the advantage of distinguishing patients with a very good prognosis (stage I) from those with a very poor prognosis (stage III); however, more patients were classified as stage II. Although stage II patients were intermediate in terms of overall prognosis, the issue of significant heterogeneity within stage II patients has not been addressed. In this study, we constructed a model that we intended to be a powerful predictor of patient survival. Therefore, we wanted to evaluate whether the model could enhance R-ISS prediction. The results showed that the model could further discriminate patients with R-ISS stage II, but performed poorly in stage I and III patients. This result not only further suggests intra-patient heterogeneity at stage II, but also illustrates that our model can optimize R-ISS to some extent. Besides this, we also applied this model for ISS staging in the three databases. The results were similar between the three databases, the optimization effect of the model on the ISS stage II phase was the most obvious, and it had a smaller optimization effect on stages I and III, although the effect was weaker than R-ISS. The distinct results for stages I and III affirm the ability of R-ISS to discriminate between patients with stages I and III diseases, as well as the significant heterogeneity within patients with stage II disease, while also demonstrating the ability of our model to optimize both R-ISS and ISS.

Collectively, we suggest that our 8-RBP-related gene signature and nomogram could be practical and reliable prognostic tools for MM. Although the signature and nomogram showed excellent performance in the training and validation sets, they inevitably had some limitations. First, although it performed well in predicting the survival of patients with MM, it lacked verification of large-scale prospective trials. Second, the R-ISS data were only obtained from the GSE136337 database, and further confirmation is needed to conclude that our model can enhance the predictive power of R-ISS. Third, the associated mechanisms have not been validated in MM cells. Based on this, our follow-up research will focus on verifying the conclusions of this study in terms of clinical applications and molecular mechanisms.

In conclusion, we introduced a prognostic signature based on eight RBP genes that might be independent prognostic factors in MM and a novel nomogram that could predict the survival of patients with MM.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

WW, SX, XZ, QG, and MZ participated in the design of the study and performed the statistical analysis. XM, YC, SL, and WL drafted the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (LY16H070008), the Zhejiang Province Medical and Health Science and Technology Program Project (2017KY166 and 2017KY709), the Program of Taizhou Science and Technology Grant (20ywb29, 1701KY22, and 1701KY23), the Medical Health Science and Technology Project of Zhejiang Province (2021PY083 and 2019KY239), the Key Technology Research and Development Program of Zhejiang Province (2019C03040), and the Major Research Program of Taizhou Enze Medical Center Grant (19EZZDA2).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665173/full#supplementary-material

Supplementary Figure 1 | Lasso cox regression of the MM-related genes in the TCGA dataset.

Supplementary Figure 2 | The optimization ability of the 8-gene model for stages I and III is not as good as that for stage II (either R-ISS or ISS). **(A,B)** R-ISS in GSE136337. **(A)** R-ISS stage I; **(B)** R-ISS stage III. **(C–H)** ISS in GSE24080, GSE136337, and TCGA-MMRF. **(C)** ISS stage I in GSE24080. **(D)** ISS stage III in GSE24080. **(E)** ISS stage I in GSE136337. **(F)** ISS stage III in GSE136337. **(G)** ISS stage I in TCGA-MMRF. **(H)** ISS stage III in TCGA-MMRF. (Red: a group that was reclassified as high-risk. Blue: a group that was reclassified as low risk.)

REFERENCES

Bauer, M. A., Ashby, C., Wardell, C., Boyle, E. M., Ortiz, M., Flynt, E., et al. (2020). Differential RNA splicing as a potentially important driver mechanism in multiple myeloma. *Haematologica* 2020:235424. doi: 10.3324/haematol.2019.235424

Boyle, E. M., Ashby, C., Tytarenko, R. G., Deshpande, S., Wang, H., Wang, Y., et al. (2020). BRAF and DIS3 Mutations Associate with Adverse Outcome in a Long-term Follow-up of Patients with Multiple Myeloma. *Clin. Cancer Res.* 26, 2422–2432. doi: 10.1158/1078-0432.Ccr-19-1507

Campos, R. K., Wijeratne, H. R. S., Shah, P., Garcia-Blanco, M. A., and Bradrick, S. S. (2020). Ribosomal stalk proteins RPLP1 and RPLP2 promote biogenesis of flaviviral and cellular multi-pass transmembrane proteins. *Nucleic Acids Res.* 48, 9872–9885. doi: 10.1093/nar/gkaa717

Correa, B. R., de Araujo, P. R., Qiao, M., Burns, S. C., Chen, C., Schlegel, R., et al. (2016). Functional genomics analyses of RNA-binding proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma. *Genome Biol.* 17:125. doi: 10.1186/s13059-016-0990-4

Cox, D. (1972). Regression models and life tables. *J. Roy. Statist. Assoc.* 1972:34.

Cui, K., Liu, C., Li, X., Zhang, Q., and Li, Y. (2020). Comprehensive characterization of the rRNA metabolism-related genes in human cancer. *Oncogene* 39, 786–800. doi: 10.1038/s41388-019-1026-9

Kleinbaum, D. G. (1998). Survival Analysis, a Self-Learning Text. *Biometric. J.* 1998:4036.

Edvardson, S., Shaag, A., Kolesnikova, O., Gomori, J. M., Tarassov, I., Einbinder, T., et al. (2007). Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. *Am. J. Hum. Genet.* 81, 857–862. doi: 10.1086/521227

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* 33, 1–22.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat. Rev. Genet.* 15, 829–845. doi: 10.1038/nrg3813

Gonsalves, W. I., Jevremovic, D., Nandakumar, B., Dispenzieri, A., Buadi, F. K., Dingli, D., et al. (2020). Enhancing the R-ISS classification of newly diagnosed multiple myeloma by quantifying circulating clonal plasma cells. *Am. J. Hematol.* 95, 310–315. doi: 10.1002/ajh.25709

Greipp, P. R., San Miguel, J., Durie, B. G., Crowley, J. J., Barlogie, B., Bladé, J., et al. (2005). International staging system for multiple myeloma. *J. Clin. Oncol.* 23, 3412–3420. doi: 10.1200/jco.2005.04.242

Huang, H. H., Ferguson, I. D., Thornton, A. M., Bastola, P., Lam, C., Lin, Y. T., et al. (2020). Proteasome inhibitor-induced modulation reveals the spliceosome as a specific therapeutic vulnerability in multiple myeloma. *Nat. Commun.* 11:1931. doi: 10.1038/s41467-020-15521-4

Joseph, J. T., Innes, A. M., Smith, A. C., Vanstone, M. R., Schwartzentruber, J. A., Bulman, D. E., et al. (2014). Neuropathologic features of pontocerebellar hypoplasia type 6. *J. Neuropathol. Exp. Neurol.* 73, 1009–1025. doi: 10.1097/nen.0000000000000123

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Med. Res. Methodol.* 17:53. doi: 10.1186/s12874-017-0332-6

Kittler, R., Putz, G., Pelletier, L., Poser, I., Heninger, A. K., Drechsel, D., et al. (2004). An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature* 432, 1036–1040. doi: 10.1038/nature03159

Lin, R. J., Lin, Y. C., Chen, J., Kuo, H. H., Chen, Y. Y., Diccianni, M. B., et al. (2010). microRNA signature and expression of Dicer and Drosha can predict prognosis and delineate risk groups in neuroblastoma. *Cancer Res.* 70, 7841–7850. doi: 10.1158/0008-5472.Can-10-0970

Palumbo, A., and Anderson, K. (2011). Multiple myeloma. *N. Engl. J. Med.* 364, 1046–1060. doi: 10.1056/NEJMra1011442

Palumbo, A., Avet-Loiseau, H., Oliva, S., Lokhorst, H. M., Goldschmidt, H., Rosinol, L., et al. (2015). Revised International Staging System for Multiple Myeloma: A Report From International Myeloma Working Group. *J. Clin. Oncol.* 33, 2863–2869. doi: 10.1200/jco.2015.61.2267

Pereira, B., Billaud, M., and Almeida, R. (2017). RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer* 3, 506–528. doi: 10.1016/j.trecan.2017.05.003

Rajkumar, S. V. (2020). Multiple myeloma: 2020 update on diagnosis, risk-stratification and management. *Am. J. Hematol.* 95, 548–567. doi: 10.1002/ajh.25791

Ranstam, J., and Cook, J. A. (2017). Kaplan-Meier curve. *Br. J. Surg.* 104:442. doi: 10.1002/bjs.10238

Sonneveld, P., Avet-Loiseau, H., Lonial, S., Usmani, S., Siegel, D., Anderson, K. C., et al. (2016). Treatment of multiple myeloma with high-risk cytogenetics: a consensus of the International Myeloma Working Group. *Blood* 127, 2955–2962. doi: 10.1182/blood-2016-01-631200

Tu, H. C., Schwitalla, S., Qian, Z., LaPier, G. S., Yermalovich, A., Ku, Y. C., et al. (2015). LIN28 cooperates with WNT signaling to drive invasive intestinal and colorectal adenocarcinoma in mice and humans. *Genes Dev.* 29, 1074–1086. doi: 10.1101/gad.256693.114

Vickers, A. J., and Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med. Decis Making* 26, 565–574. doi: 10.1177/0272989x06295361

Wang, E., and Aifantis, I. (2020). RNA Splicing and Cancer. *Trends Cancer* 6, 631–644. doi: 10.1016/j.trecan.2020.04.011

Wu, Y., Zhao, W., Liu, Y., Tan, X., Li, X., Zou, Q., et al. (2018). Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response. *Embo* J. 37:201899017. doi: 10.15252/embj.201899017

Xu, Y., Deng, Y., Ji, Z., Liu, H., Liu, Y., Peng, H., et al. (2014). Identification of thyroid carcinoma related genes with mRMR and shortest path approaches. *PLoS One* 9:e94022. doi: 10.1371/journal.pone.0094022

Yi, Y., Nandana, S., Case, T., Nelson, C., Radmilovic, T., Matusik, R. J., et al. (2009). Candidate metastasis suppressor genes uncovered by array comparative genomic hybridization in a mouse allograft model of prostate cancer. *Mol. Cytogenet.* 2:18. doi: 10.1186/1755-8166-2-18

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: 29 April 2021
doi: 10.3389/fgene.2021.616110

CDK4 Amplification in Esophageal Squamous Cell Carcinoma Associated With Better Patient Outcome

Jie Huang[1†], Xiang Wang[1†], Xue Zhang[1], Weijie Chen[1], Lijuan Luan[1], Qi Song[1], Hao Wang[2], Jia Liu[1], Lei Xu[1], Yifan Xu[1], Licheng Shen[1], Lijie Tan[2], Dongxian Jiang[1], Jieakesu Su[1] and Yingyong Hou[1,3]**

[1] Department of Pathology, Zhongshan Hospital, Fudan University, Shanghai, China, [2] Department of Thoracic Surgery, Zhongshan Hospital, Fudan University, Shanghai, China, [3] Department of Pathology, Zhongshan Hospital, School of Basic Medical Sciences, Fudan University, Shanghai, China

OPEN ACCESS

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

Reviewed by:
Yulin Zhang,
Shandong University, China
Haoyun Lei,
Carnegie Mellon University,
United States
Liuyi Hao,
University of North Carolina
at Greensboro, United States

***Correspondence:**
Yingyong Hou
houyingyong@aliyun.com
Jieakesu Su
13816327136@126.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: 11 October 2020
Accepted: 26 February 2021
Published: 29 April 2021

Citation:
Huang J, Wang X, Zhang X, Chen W, Luan L, Song Q, Wang H, Liu J, Xu L, Xu Y, Shen L, Tan L, Jiang D, Su J and Hou Y (2021) CDK4 Amplification in Esophageal Squamous Cell Carcinoma Associated With Better Patient Outcome. Front. Genet. 12:616110. doi: 10.3389/fgene.2021.616110

In the present study, we aimed to investigate the clinical and prognostic values of *CDK4* amplification and improve the risk stratification in patients with esophageal squamous cell carcinoma. *CDK4* amplification was analyzed by fluorescence *in situ* hybridization using tissue microarray consisting of representative tissues of 520 patients with esophageal squamous cell carcinoma, and its correlation with clinicopathological features and clinical outcomes were evaluated. *CDK4* amplification was found in 8.5% (44/520) of patients with esophageal squamous cell carcinoma. *CDK4* amplification was negatively correlated with disease progression ($P = 0.003$) and death ($P = 0.006$). Patients with *CDK4* amplification showed a significantly better disease-free survival ($P = 0.016$) and overall survival ($P = 0.023$) compared with those patients without *CDK4* amplification. When patients were further stratified into I–II stage groups and III–IV stage groups, *CDK4* amplification was significantly associated with both better disease-free survival ($P = 0.023$) and overall survival ($P = 0.025$) in the I–II stage group rather than the III–IV stage group. On univariate and multivariate analysis, invasive depth and *CDK4* amplification were associated with disease-free survival and overall survival. Taken together, *CDK4* amplification was identified as an independent prognostic factor for survival, which could be incorporated into the tumor–node–metastasis staging system to refine risk stratification of patients with esophageal squamous cell carcinoma.

Keywords: esophageal squamous cell carcinoma, *CDK4* amplification, clinical stage, prognostic value, fluorescence *in situ* hybridization

INTRODUCTION

Esophageal cancer (EC) is a lethal digestive tract malignancy with a poor prognosis, and an increasing incidence and mortality rate worldwide (Malhotra et al., 2017). There are two main histological types of EC: esophageal squamous cell cancer (ESCC) and esophageal adenocarcinoma (EAC), which have significant differences in pathogenesis, epidemiology, and risk factors (Rustgi and El-Serag, 2014; Arnold et al., 2015). ESCC usually occurs in flat cells lining the upper

two thirds of the esophagus, predominantly in Africa and eastern Asia (especially in China), and smoking is the main risk factor (Lin et al., 2013; Rustgi and El-Serag, 2014), while EC mostly originates from the Barrett mucosa in the lower third of the esophagus and is prevalent in many developed countries (Edgren et al., 2013; Rustgi and El-Serag, 2014; Arnold et al., 2016). In addition to environmental and external factors, genetic factors may also contribute to the development of a specific type of EC. Recently, whole-genome sequencing and genome-wide association studies have been undertaken to identify EC-related genetic alterations (Gao et al., 2014; Lin et al., 2014; Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, Bc Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al., 2017).

Cell cycle dysregulation induced by abnormal genetic alterations (mutations, deletions, or amplifications) occur frequently in human malignancies (Hanahan and Weinberg, 2011; Gampenrieder et al., 2016). The deregulation of the cyclin D1–CDK4/6–Rb pathway, which will trigger loss of cell cycle control, is one of the hallmarks of carcinogenesis (Asghar et al., 2015). The cyclin-dependent kinase 4 (*CDK4*) gene located in the chromosomal region 12q14.1 might have oncogenic potential similar to other G1 regulatory genes (Haas et al., 1997). CDK4 is initially identified as a catalytic subunit present in the CDK/cyclin D complex in the G1 phase of the cell cycle (Matsushime et al., 1992). CDK4 coupled with cyclin D1 (CCND1) phosphorylates the retinoblastoma protein 1 (RB1), which leads to the release of the transcription factor EF2 and subsequently enables the cell cycle progress from G1 to S phase (Harbour et al., 1999). Alterations of these key components have been implicated in the pathogenesis of multiple tumor types (An et al., 1999). Overexpression of CDK4 could induce uncontrolled cell growth and eventually lead to tumorigenesis; moreover, amplification of the *CDK4* gene, have been found in various cancers (Lee et al., 2014).

The perturbed cell cycle regulation pathway in ESCC mainly exhibited genetic alterations in the G1/S transition control, including mutations or deletions of *TP53*, *RB1*, *CDKN2A*, *CHEK1*, and *CHEK2*, and amplifications of *CDK4*, *CCND1*, *CDK6*, and *MDM2* (Song et al., 2014). Alterations of these genes, such as inactivation of *RB1* and *CDKN2A* and amplification of *CCND1*, *CDK6*, and *MDM2* have been well documented in ESCC (Huang et al., 2007; Baba et al., 2014; Jiang et al., 2020). To date, the prognostic significance of *CDK4* amplification in ESCC has not been described before. In this article, we describe *CDK4* amplification in ESCC by fluorescence *in situ* hybridization (FISH) and meticulously investigated the clinical and prognostic values of *CDK4* amplification in patients with ESCC to improve the risk stratification.

MATERIALS AND METHODS

Patients and Tissues

This study retrospectively enrolled 520 ESCC patients who had undergone surgical resection in the Department of Thorax Surgery, Zhongshan Hospital, Fudan University (Shanghai, China), between January 2007 and November 2010. Patients who received preoperative antitumor therapy, including neoadjuvant therapy, chemotherapy, and radiotherapy or died within 3 months were excluded from the current study. Ethical approval was granted by the Human Research Ethics Committee of Zhongshan Hospital, Fudan University. Signed informed consent for the acquisition and use of patient tissue specimens and clinical data was obtained from each patient.

All specimens were reassessed independently by two pathologists using hematoxylin and eosin (HE)-stained sections to determine the tumor grade, differentiation, invasion depth, lymph node metastasis, vessel and nerve involvement, and disease stage, according to the American Joint Committee on Cancer guidelines for tumor–node–metastasis (TNM) classification (eighth edition). Patients' clinicopathological characteristics such as gender, age, smoking, tumor location, and clinical stage were collected from medical records. After surgery, patients were followed up with endoscopy and computed tomographic scan of the thorax and abdomen every 3 months for the first year, every 6 months for the second year, and every 6–12 months thereafter. Follow-up data of those patients who did not have themselves examined in our hospital were obtained by telephone.

Tissue Microarray

Tissue microarrays (TMAs) containing tumor tissues of the 520 patients under study were constructed as previously described (Shi et al., 2013). Briefly, the representative areas of 2 mm wide and 6 mm long with rich tumor cells were selected by two experienced pathologists according to HE-stained slides. The corresponding regions on archived formalin-fixed, paraffin-embedded (FFPE) tissue blocks were extracted, vertically planted into the recipient TMA blocks and then aggregated on the instrument.

Fluorescence *in situ* Hybridization and Assessment

Dual-color FISH assay was conducted on the TMA sections of 5 μm thickness using *CDK4*-specific probe (Spectrum orange) together with a centromere-specific probe (Spectrum green) for chromosome 12 (*CEP12*) (Empire Genomics, Buffalo, NY) for assessment of *CDK4* amplification according to established laboratory protocol, as previously described (Zhang et al., 2014). FISH copy number evaluation was performed by two experienced pathologists blinded to patients' clinicopathologic characteristics under a fluorescence microscope (BX43; Olympus, Tokyo, Japan) equipped with a DAPI/green/orange triple band pass filter and a Microscope Digital Camera (DP73; Olympus). At least 100 tumor cell nuclei of each ESCC sample were analyzed by counting orange signals for *CDK4* and green signals for *CEP12* under an oil microscope with a magnification of 1,000 times. Overlapping tumor nuclei were excluded from evaluation to avoid false-positive scoring. Then the average number of *CDK4* and *CEP12* signals and the ratio of *CDK4/CEP12* were calculated for each case. Amplification of *CDK4* was defined as a *CDK4/CEP12* ratio $\geq$2.0 or an average copy number of *CDK4* signals/tumor

cell nucleus ≥5.0 or percentage of tumor cells containing large clusters of *CDK4* signal ≥10%, respectively, based on previously reported modified scoring algorithms for *HER2* and *c-MYC* (Wolff et al., 2007; Huang et al., 2019).

Statistical Analysis

All the statistical analyses were carried out using SPSS 20.0 (SPSS Inc., Chicago, IL, United States). All *P*-values were two sided, and differences were considered statistically significant values of $P < 0.05$. Disease-free survival (DFS) was defined as the interval between surgical resection and recurrence, metastasis, or death from any cause. Overall survival (OS) was defined as the interval from date of curative surgery until death or last follow-up date. Correlations between *CDK4* amplification and clinicopathologic variables were analyzed using the Fisher exact test or Pearson χ^2 test. The Kaplan–Meier method with log-rank test was applied to calculate the cumulative survival proportion for OS and DFS by *CDK4* amplification level and to determine if there were any significant differences between the survival curves. The Cox proportional hazard regression model was used to carry out the univariate and multivariate regression analyses, and the hazard ratio (HR) and 95% confidence intervals (CI) were determined.

RESULTS

Patient Characteristics

Detailed clinicopathological characteristics of the study cohort including 520 ESCC specimens obtained for this study are summarized in **Table 1**. The median age of this cohort was 61 years (range, 34–83 years), of which 81.7% were men and 38.7% were smokers. By anatomic site, 44.0% of tumors were in the middle esophagus, whereas 51.2% of the tumors were in the upper and lower esophagus with a median tumor size of 3 cm (range, 0.3–10 cm). The tumor differentiation was defined as grade I in 20 (3.8%) patients, II in 292 (56.2%) patients, and III in 208 (40.0%) patients. Vessel and nerve invasions were presented in 111 (21.3%) and 177 (34.0%) tumors, respectively. Meanwhile, lymph node metastasis was observed in 238 (45.8%) of the patients. The depth of invasion was also evaluated. 15 (2.9%) cases were confined to the mucosa, 38 (7.3%) were in the submucosa, 115 (22.1%) were in the muscular layer, and 352 (67.7%) were beyond the muscular layer. Among these patients with ESCC, clinical stage was classified as I to II and III to IVb in 290 (55.8%) and 230 (44.2%) cases, respectively, according to the American Joint Committee on Cancer Staging Manual (eighth edition).

Association Between *CDK4* Amplification and Clinicopathological Features

All the patients were classified into two groups by using prespecified criteria for *CDK4* amplification based on previous studies (Wolff et al., 2007; Huang et al., 2019). *CDK4* amplification (a *CDK4/CEP12* ratio ≥2.0 or an average copy number of *CDK4* signals/tumor cell nucleus ≥5.0 or percentage of tumor cells containing large clusters of *CDK4* signal ≥10%) was found in 8.5% (44 of 520) of patients (**Figures 1A,B**), and other patients (91.5%, 476 of 520) showed non-amplification (low polysomy or disomy) (**Figures 1C,D**). The correlations between *CDK4* amplification and clinicopathological features are shown in **Table 1**. *CDK4* amplification status significantly correlated with disease progression ($P = 0.003$) and death ($P = 0.006$). There was no significant difference between *CDK4* amplification and *CDK4* non-amplification group regarding sex ($P = 0.987$),

TABLE 1 | Correlation between *CDK4* amplification and clinicopathological features in full cohort of patients with ESCC.

Clinicopathologic feature	No.	*CDK4* amplification		
		No	Yes	*P*-value
Sex				0.987
Female	95	87	8	
Male	425	389	36	
Age (years)				0.588
<60	221	204	17	
≥60	299	272	27	
Grade				0.403
I + II	312	283	29	
III	208	193	15	
Invasive depth				0.791
I–II	168	153	15	
III	352	323	29	
Vessel invasion				0.592
No	409	373	36	
Yes	111	103	8	
Nerve invasion				0.511
No	343	312	31	
Yes	177	164	13	
Lymph node metastasis				0.556
No	282	260	22	
Yes	238	216	22	
Site				0.768
Up	25	22	3	
Middle	229	211	18	
Down	241	220	21	
Smoking				0.748
No	319	293	26	
Yes	201	183	18	
Clinical stage				0.625
I–II	290	267	23	
III–IVb	230	209	21	
Disease progression				**0.003**
No	242	212	30	
Yes	278	264	14	
Death				**0.006**
No	251	221	30	
Yes	269	255	14	

Grade: I, well differentiated; II, moderately differentiated; III, poorly differentiated. Invasive depth: I, confined to submucosal layer; II, invasion of muscular layer; III, beyond the muscularis.
The features/variables that make significant contribution are in bold.

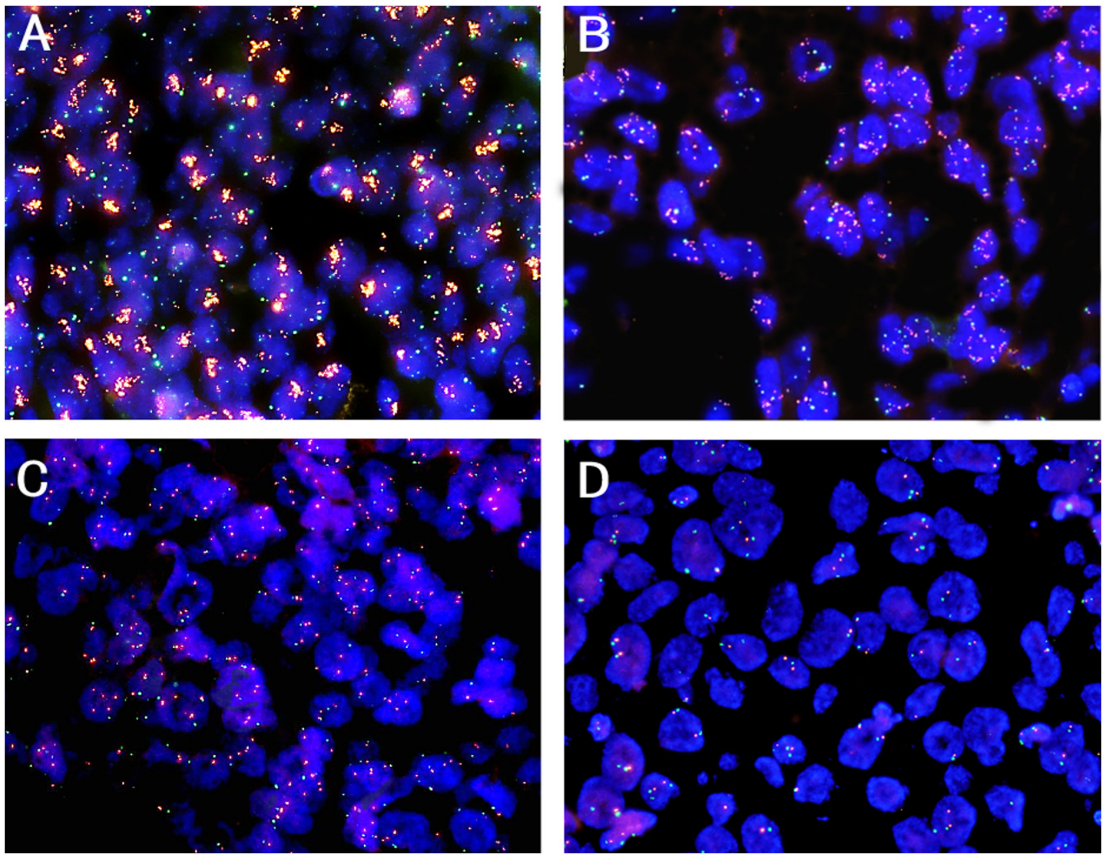

FIGURE 1 | Representative patterns of *CDK4* gene (orange color) and CEP12 (green color) copy number status by FISH (original magnification ×1,000). **(A)** *CDK4* amplification, a *CDK4/CEP12* ratio ≥2.0; **(B)** *CDK4* amplification, an average copy number of *CDK4* signals/tumor cell nucleus ≥5.0; **(C)** *CDK4* non-amplification, low polysomy; **(D)** *CDK4* non-amplification, disomy.

age (P = 0.588), grade (P = 0.403), invasive depth (P = 0.791), vessel (P = 0.592) and nerve invasions (P = 0.511), lymph node metastasis (P = 0.556), anatomic site (P = 0.768), smoking (P = 0.748), and clinical stage (P = 0.625).

Survival Analyses

The 5-year DFS and OS rates for all patients were 32.1% and 32.9%, respectively, with a median follow-up period of 35.5 months (range, 3–102 months). Mean and median times to DFS were 41.6 and 31.0 months, while to OS were 44.9 and 35.5 months, respectively. Instances of disease progression (278), including 106 local recurrences and 172 lymph node or distant metastasis, were documented, and 277 patients (53.3%) died during the follow-up, in which 269 patients (51.7%) died of EC. To further explore the prognostic significance of *CDK4* amplification and clinical outcomes, Kaplan–Meier analysis with log-rank test was used to compare differences between subgroups. The Kaplan–Meier curves revealed that the *CDK4* amplification group with a median DFS and OS of 42.5 and 46.0 months, respectively, gained significant survival benefit compared with the group without *CDK4* amplification (median DFS, 30.0 months, P = 0.016; median OS, 35.0 months, P = 0.023) (**Figure 2**). Univariate analysis of prognostic significance revealed that grade, invasive depth, vessel invasion, nerve invasion, lymph node metastasis, clinical stage, and *CDK4* amplification were significantly associated with DFS and OS. In the multivariate analysis, invasive depth (P = 0.006, HR: 1.560, 95% CI: 1.133–2.149 for DFS; P = 0.008, HR: 1.542, 95% CI: 1.119–2.125 for OS) and *CDK4* amplification (P = 0.015, HR: 0.512, 95% CI: 0.299–0.877 for DFS; P = 0.021, HR: 0.530, 95% CI: 0.309–0.908 for OS) were associated with DFS and OS (**Table 2**).

Survival Analyses Based on Clinical Stage

In stages I–II patients (n = 290, **Figures 3A,B**), *CDK4* amplification was significantly associated with better DFS (P = 0.023) and OS (P = 0.025). Among the 23 patients with *CDK4* amplification, a better prognosis was observed, with a median DFS and OS being both 73.0 months compared with 45.0 and 47.0 months for 267 patients without *CDK4* amplification. As to the stages III–IV patients (n = 230, **Figures 3C,D**), *CDK4* amplification did not play the prognostic role whether in DFS (P = 0.144) or in OS (P = 0.211), since the median DFS and OS

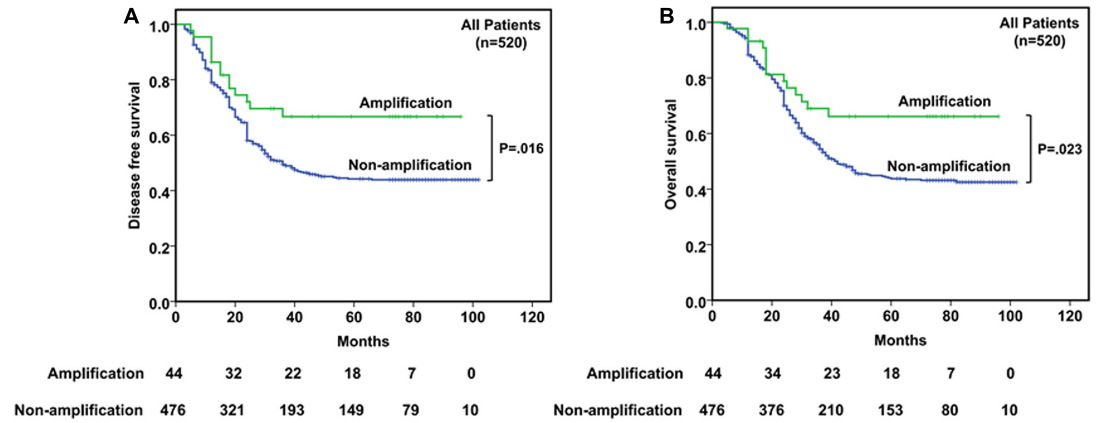

FIGURE 2 | Kaplan–Meier curves of disease-free survival (DFS) **(A)** and overall survival (OS) **(B)** according to *CDK4* amplification status in 520 esophageal squamous cell cancer (ESCC) patients.

were 18.0 and 25.0 months, respectively, in 21 patients with *CDK4* amplification, whereas it was 20.0 and 25.0 months, respectively, for 209 patients without *CDK4* amplification.

TABLE 2 | Univariate and multivariate survival analyses for DFS and OS in full cohort of patients with ESCC.

Variable	DFS		OS	
	P-value	Hazard ratio (CI 95%)	*P*-value	Hazard ratio (CI 95%)
Univariate analysis				
Sex	0.109	1.299 (0.943–1.790)	0.128	1.283 (0.931–1.767)
Age	0.848	0.977 (0.768–1.243)	0.888	1.017 (0.800–1.295)
Grade	**0.032**	1.269 (1.021–1.577)	**0.045**	1.250 (1.005–1.554)
Invasive depth	**<0.001**	1.921 (1.446–2.553)	**<0.001**	1.952 (1.469–2.593)
Vessel invasion	**0.002**	1.536 (1.175–2.007)	**0.001**	1.562 (1.195–2.042)
Nerve invasion	**0.008**	1.394 (1.091–1.782)	**0.002**	1.460 (1.142–1.865)
Lymph node metastasis	**<0.001**	2.809 (2.192–3.600)	**<0.001**	2.854 (2.227–3.658)
Clinical stage	**<0.001**	2.844 (2.222–3.639)	**<0.001**	2.882 (2.252–3.687)
Site	0.980	0.997 (0.810–1.228)	0.813	1.026 (0.832–1.265)
Smoking	0.199	1.173 (0.919–1.495)	0.192	1.176 (0.922–1.499)
CDK4 amplification	**0.020**	0.529 (0.309–0.906)	**0.027**	0.546 (0.319–0.934)
Multivariate analysis				
Grade	0.438	1.093 (0.873–1.367)	0.594	1.063 (0.849–1.331)
Invasive depth	**0.006**	1.560 (1.133–2.149)	**0.008**	1.542 (1.119–2.125)
Vessel invasion	0.977	0.996 (0.749–1.324)	0.907	1.017 (0.766–1.351)
Nerve invasion	0.964	1.006 (0.771–1.313)	0.677	1.058 (0.810–1.382)
Lymph node metastasis	0.192	1.980 (0.709–5.524)	0.158	2.095 (0.750–5.849)
Clinical stage	0.583	1.333 (0.478–3.721)	0.656	1.263 (0.452–3.528)
CDK4 amplification	**0.015**	0.512 (0.299–0.877)	**0.021**	0.530 (0.309–0.908)

ESCC, esophageal squamous cell carcinoma; CI, confidence interval; DFS, disease-free survival; OS, overall survival.
The features/variables that make significant contribution are in bold.

DISCUSSION

Prognosis prediction and treatment guidance for ESCC are currently based on the TNM staging system, which provides prognostic information, and it will continue to be the most commonly applied approach for a fairly long time (Rustgi and El-Serag, 2014). However, patients with the same TNM stage may display different molecular phenotypes and prognoses. Many non-anatomic prognostic factors, especially genetic and molecular markers critical in carcinogenesis and cancer progression, are also found to have great significance in patient prognosis (Cao et al., 2014; Lin et al., 2017; Mei et al., 2017; Wang et al., 2017; Bi et al., 2020). Therefore, it is of great importance to identify accurate biological markers for the prognosis of ESCC, which may help subdivide patients at the same stage into different groups according to their prognosis. A better understanding of patient prognosis would help guide more personalized treatment for ESCC patients after curative surgery.

Aberrant *CDK4* amplification in malignant tissues has been reported to be involved in the development and progression of various cancers including liposarcoma (Creytens et al., 2015), glioblastomas (Schmidt et al., 1994), breast cancer (Piezzo et al., 2020), ovarian cancer (Masciullo et al., 1997), and melanoma (Muthusamy et al., 2006) through the cyclin D1–CDK4/6–Rb pathway. Ricciotti et al. (2017) performed a cut point analysis of the prognostic significance of *CDK4* amplification in patients with dedifferentiated liposarcoma by comparison of Kaplan–Meier survival curves using log rank tests. The study showed that *CDK4* amplification was associated with decreased DFS (P = 0.0169) and disease-specific survival (DSS) (P = 0.0140). Saada-Bouzid et al. (2015) also demonstrated that *CDK4* amplification was significantly associated with shorter recurrence-free survival, and overall survival in dedifferentiated liposarcoma patients. Altogether, the amplification of *CDK4* appears to be a negative event in liposarcoma. In glioblastoma patients, Fischer et al. (2010) reported that lack of amplification of *CDK4* was recognized to be associated with a significant longer survival time.

In the present study, we investigated *CDK4* amplification and its value in the prediction of survival in patients with ESCC. The correlation between *CDK4* amplification and the clinicopathological parameters of ESCC patients was also analyzed. Different from a singular criterion only using *CDK4/CEP12* ratio or *CDK4* copy numbers, we applied a more sophisticated *CDK4* FISH criterion considering percentage of *CDK4* clusters at the same time. Patients with *CDK4* amplification and non-amplification account for 8.5% (n = 44) and 91.5% (n = 476) of all the 520 ESCC patients, respectively. *CDK4* amplification rate (8.5%) determined by FISH analysis in our study is comparative with that of a previous study obtained by high-throughput sequencing methods (Song et al., 2014). There was no significant difference between *CDK4* amplification and *CDK4* non-amplification group regarding sex, age, grade, invasive depth, vessel and nerve invasions, lymph node metastasis, anatomic site, smoking, and clinical stage, which is in line with the conclusion that no significant associations were found between *CDK4* gene amplification and patient's age, tumor size, and lymph node status in breast cancer (An et al., 1999).

Although there was no statistical significance, *CDK4* gene amplification was less common in tumors with higher histological grade. Moreover, it is worth noting that *CDK4* amplification had a significant negative correlation with disease progression (P = 0.003) and death (P = 0.006) (**Table 1**). *CDK4* seems to be negatively correlated with some indicators indicating poor prognosis in ESCC. Interestingly, different from the prognosis value of *CDK4* amplification in dedifferentiated liposarcoma and glioblastoma patients, we demonstrated that *CDK4* amplification was associated with a better DFS (P = 0.016) and OS (P = 0.023) (**Figure 2**). Furthermore, *CDK4* amplification was not a common genetic alteration but proved to be an independent prognostic marker in patients with ESCC (**Table 2**). The results of this study seem to be opposite to the prognosis of other tumor types. This may be as a result of the complexity of the gene regulation process in ESCC. The occurrence and development of ESCC is a multistage and multifactor process, which involves the interaction of multiple oncogenes and tumor suppressor genes (Gao et al., 2014; Lin et al., 2014; Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, Bc Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al., 2017). In addition, we speculate that it may be due to the influence of cancer species, and geographical and environmental factors; the causes of different tumors are not the same, leading to the differences in research results. To the best of our knowledge, this study is the first to evaluate the value of

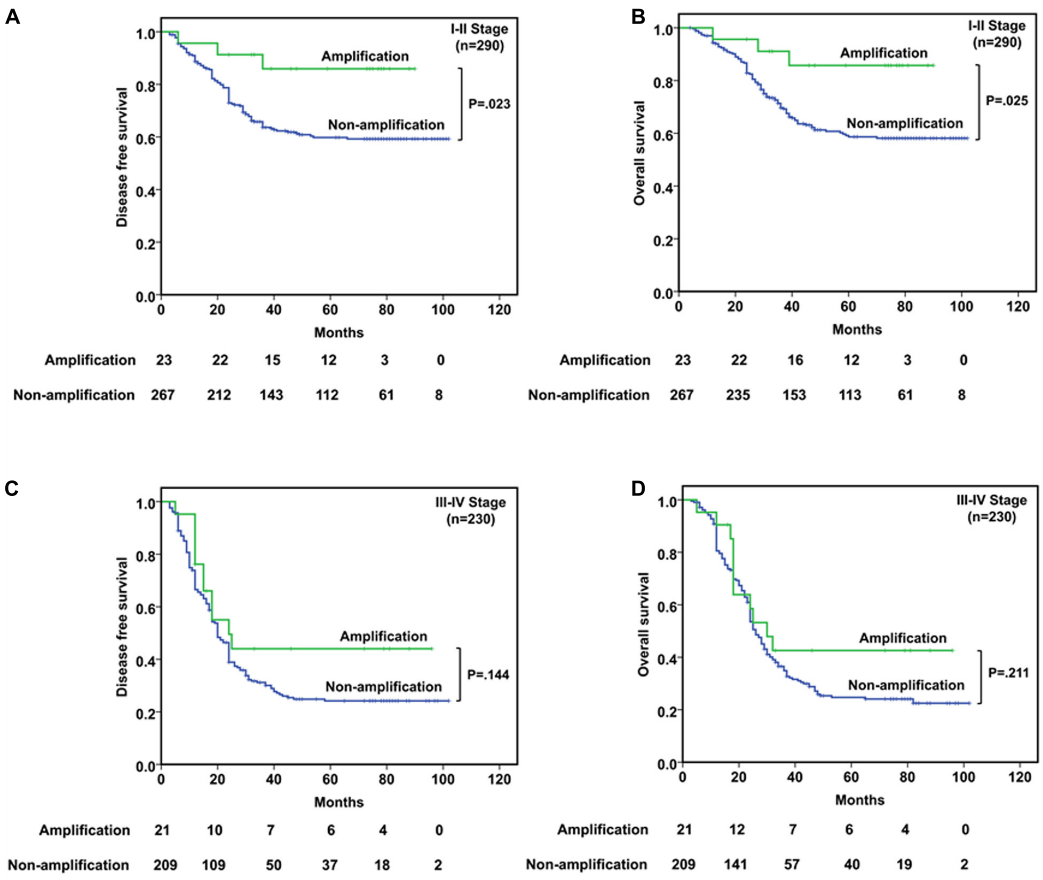

FIGURE 3 | Survival analyses based on clinical stage of ESCC patients. **(A,B)** In stages I–II patients, *CDK4* amplification was significantly associated with better DFS (P = 0.023) and OS (P = 0.025). **(C,D)** In stages III–IV patients, *CDK4* amplification could not predict the prognosis in DFS (P = 0.144) or OS (P = 0.211).

CDK4 amplification as a novel candidate prognostic biomarker in patients with ESCC, so it is necessary to further investigate the upstream and downstream genes of *CDK4* to clarify its role and elucidate the prognostic utility in ESCC.

Given that clinical stage is an important clinicopathological factor, the prognosis usually varies between patients with different stages. Therefore, we categorized the patients into the I–II stage group and III–IV stage group. In the I–II stage group, *CDK4* amplification was significantly associated with both better DFS and OS compared with the non-amplification group. However, this significant correlation was not found in the III–IV stage patients implying that prognostic value of *CDK4* amplification is relying on clinical stage (**Figure 3**). It is suggested that *CDK4* may change in the early stage of ESCC and play an important role in the occurrence and development of the disease. With the increase in clinical stage, more and more genes in ESCC are changed (Sudo et al., 2019), and the interaction between genes becomes complex, which affects the role of *CDK4*.

In summary, we have first proved the prognostic significance of *CDK4* amplification as a favorably independent prognostic factor for DFS and OS in Chinese patients with ESCC. Combining *CDK4* amplification with the TNM staging system might add more information to better predict the prognosis of ESCC patients.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Research Ethics Committee of Zhongshan Hospital, Fudan University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH wrote the draft of the manuscript. XW and YH evaluated CDK4 copy numbers. XZ, WC, and LL analyzed the results of the experiment. QS, HW, and JL constructed Tissue microarrays (TMA). LX, YX, and LS were involved in the picture editing. LT and DJ put forward suggestions for improvement. YH and JS revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Shanghai Natural Science Foundation of China (No. 18ZR1406800), Xiamen Science and Technology Project of Fujian Province, China (No. 3502Z20184003), National Natural Science Foundation of China (No. 81702372), Shanghai Municipal Commission of Science and Technology (No. 19441904000), Shanghai Municipal Key Clinical Specialty (No. shslczdzk01302), and Shanghai Science and Technology Development Fund (No. 19MC1911000).

REFERENCES

An, H., Beckmann, M., Reifenberger, G., Bender, H., and Niederacher, D. (1999). Gene amplification and overexpression of *CDK4* in sporadic breast carcinomas is associated with high tumor cell proliferation. *Am. J. Pathol.* 154, 113–118. doi: 10.1016/s0002-9440(10)65257-1

Arnold, M., Colquhoun, A., Cook, M. B., Ferlay, J., Forman, D., and Soerjomataram, I. (2016). Obesity and the incidence of upper gastrointestinal cancers: an ecological approach to examine differences across age and sex. *Cancer Epidemiol. Biomarkers Prev.* 25, 90–97. doi: 10.1158/1055-9965.epi-15-0753

Arnold, M., Soerjomataram, I., Ferlay, J., and Forman, D. (2015). Global incidence of oesophageal cancer by histological subtype in 2012. *Gut* 64, 381–387. doi: 10.1136/gutjnl-2014-308124

Asghar, U., Witkiewicz, A. K., Turner, N. C., and Knudsen, E. S. (2015). The history and future of targeting cyclin-dependent kinases in cancer therapy. *Nat. Rev. Drug Discov.* 14, 130–146.

Baba, Y., Watanabe, M., Murata, A., Shigaki, H., Miyake, K., Ishimoto, T., et al. (2014). LINE-1 hypomethylation, DNA copy number alterations, and CDK6 amplification in esophageal squamous cell carcinoma. *Clin. Cancer Res.* 20, 1114–1124. doi: 10.1158/1078-0432.ccr-13-1645

Bi, Y., Guo, S., Xu, X., Kong, P., Cui, H., Yan, T., et al. (2020). Decreased ZNF750 promotes angiogenesis in a paracrine manner via activating DANCR/miR-4707-3p/FOXC2 axis in esophageal squamous cell carcinoma. *Cell Death Dis.* 11:296.

Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, Bc Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, et al. (2017). Integrated genomic characterization of oesophageal carcinoma. *Nature* 541, 169–175. doi: 10.1038/nature20805

Cao, F., Han, H., Zhang, F., Wang, B., Ma, W., Wang, Y., et al. (2014). HPV infection in esophageal squamous cell carcinoma and its relationship to the prognosis of patients in northern China. *Sci. World J.* 2014:804738.

Creytens, D., van Gorp, J., Ferdinande, L., Speel, E. J., and Libbrecht, L. (2015). Detection of MDM2/CDK4 amplification in lipomatous soft tissue tumors from formalin-fixed, paraffin-embedded tissue: comparison of multiplex ligation-dependent probe amplification (MLPA) and fluorescence in situ hybridization (FISH). *Appl. Immunohistochem. Mol. Morphol.* 23, 126–133. doi: 10.1097/pdm.0000000000000041

Edgren, G., Adami, H. O., Weiderpass, E., and Nyren, O. (2013). A global assessment of the oesophageal adenocarcinoma epidemic. *Gut* 62, 1406–1414. doi: 10.1136/gutjnl-2012-302412

Fischer, U., Leidinger, P., Keller, A., Folarin, A., Ketter, R., Graf, N., et al. (2010). Amplicons on chromosome 12q13-21 in glioblastoma recurrences. *Int. J. Cancer* 126, 2594–2602.

Gampenrieder, S. P., Rinnerthaler, G., and Greil, R. (2016). CDK4/6 inhibition in luminal breast cancer. *Memo* 9, 76–81. doi: 10.1007/s12254-016-0268-2

Gao, Y. B., Chen, Z. L., Li, J. G., Hu, X. D., Shi, X. J., Sun, Z. M., et al. (2014). Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.* 46, 1097–1102.

Haas, K., Staller, P., Geisen, C., Bartek, J., Eilers, M., and Moroy, T. (1997). Mutual requirement of CDK4 and Myc in malignant transformation: evidence for cyclin D1/CDK4 and p16INK4A as upstream regulators of Myc. *Oncogene* 15, 179–192. doi: 10.1038/sj.onc.1201171

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

Harbour, J. W., Luo, R. X., Dei Santi, A., Postigo, A. A., and Dean, D. C. (1999). Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1. *Cell* 98, 859–869. doi: 10.1016/s0092-8674(00)81519-6

Huang, C., Yang, L., Li, Z., Yang, J., Zhao, J., Dehui, X., et al. (2007). Detection of CCND1 amplification using laser capture microdissection coupled with real-time polymerase chain reaction in human esophageal squamous cell carcinoma. *Cancer Genet. Cytogenet.* 175, 19–25. doi: 10.1016/j.cancergencyto.2007.01.003

Huang, J., Jiang, D., Zhu, T., Wang, Y., Wang, H., Wang, Q., et al. (2019). Prognostic significance of c-MYC amplification in esophageal squamous cell carcinoma. *Ann. Thorac. Surg.* 107, 436–443. doi: 10.1016/j.athoracsur.2018.07.077

Jiang, D., Chen, L., Huang, J., Wang, H., Song, Q., Shi, P., et al. (2020). Mouse double minute 2 amplification in oesophageal squamous cell carcinoma is associated with better outcome. *Histopathology* 77, 963–973. doi: 10.1111/his.14208

Lee, S., Park, H., Ha, S. Y., Paik, K. Y., Lee, S. E., Kim, J. M., et al. (2014). *CDK4* amplification predicts recurrence of well-differentiated liposarcoma of the abdomen. *PLoS One* 9:e99452. doi: 10.1371/journal.pone.0099452

Lin, D. C., Hao, J. J., Nagata, Y., Xu, L., Shang, L., Meng, X., et al. (2014). Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat. Genet.* 46, 467–473.

Lin, Y., Shen, L. Y., Fu, H., Dong, B., Yang, H. L., Yan, W. P., et al. (2017). P21, COX-2, and E-cadherin are potential prognostic factors for esophageal squamous cell carcinoma. *Dis. Esophagus* 30, 1–10. doi: 10.1007/978-981-15-4190-2_1

Lin, Y., Totsuka, Y., He, Y., Kikuchi, S., Qiao, Y., Ueda, J., et al. (2013). Epidemiology of esophageal cancer in Japan and China. *J. Epidemiol.* 23, 233–242. doi: 10.2188/jea.je20120162

Malhotra, G. K., Yanala, U., Ravipati, A., Follet, M., Vijayakumar, M., and Are, C. (2017). Global trends in esophageal cancer. *J. Surg. Oncol.* 115, 564–579. doi: 10.1002/jso.24592

Masciullo, V., Scambia, G., Marone, M., Giannitelli, C., Ferrandina, G., Bellacosa, A., et al. (1997). Altered expression of cyclin D1 and CDK4 genes in ovarian carcinomas. *Int. J. Cancer* 74, 390–395. doi: 10.1002/(sici)1097-0215(19970822)74:4<390::aid-ijc5>3.0.co;2-q

Matsushime, H., Ewen, M. E., Strom, D. K., Kato, J. Y., Hanks, S. K., Roussel, M. F., et al. (1992). Identification and properties of an atypical catalytic subunit (p34PSK-J3/cdk4) for mammalian D type G1 cyclins. *Cell* 71, 323–334. doi: 10.1016/0092-8674(92)90360-o

Mei, L. L., Qiu, Y. T., Zhang, B., and Shi, Z. Z. (2017). MicroRNAs in esophageal squamous cell carcinoma: potential biomarkers and therapeutic targets. *Cancer Biomark.* 19, 1–9. doi: 10.3233/cbm-160240

Muthusamy, V., Hobbs, C., Nogueira, C., Cordon-Cardo, C., McKee, P. H., Chin, L., et al. (2006). Amplification of CDK4 and MDM2 in malignant melanoma. *Genes Chromosomes Cancer* 45, 447–454.

Piezzo, M., Cocco, S., Caputo, R., Cianniello, D., Gioia, G. D., Lauro, V. D., et al. (2020). Targeting cell cycle in breast cancer: CDK4/6 inhibitors. *Int. J. Mol. Sci.* 21:6479.

Ricciotti, R. W., Baraff, A. J., Jour, G., Kyriss, M., Wu, Y., Liu, Y., et al. (2017). High amplification levels of MDM2 and CDK4 correlate with poor outcome in patients with dedifferentiated liposarcoma: a cytogenomic microarray analysis of 47 cases. *Cancer Genet.* 218-219, 69–80. doi: 10.1016/j.cancergen.2017.09.005

Rustgi, A. K., and El-Serag, H. B. (2014). Esophageal carcinoma. *N. Engl. J. Med.* 371, 2499–2509.

Saada-Bouzid, E., Burel-Vandenbos, F., Ranchere-Vince, D., Birtwisle-Peyrottes, I., Chetaille, B., Bouvier, C., et al. (2015). Prognostic value of HMGA2, CDK4, and JUN amplification in well-differentiated and dedifferentiated liposarcomas. *Mod. Pathol.* 28, 1404–1414. doi: 10.1038/modpathol.2015.96

Schmidt, E. E., Ichimura, K., Reifenberger, G., and Collins, V. P. (1994). CDKN2 (p16/MTS1) gene deletion or CDK4 amplification occurs in the majority of glioblastomas. *Cancer Res.* 54, 6321–6324.

Shi, Y., He, D., Hou, Y., Hu, Q., Xu, C., Liu, Y., et al. (2013). An alternative high output tissue microarray technique. *Diagn. Pathol.* 8:9.

Song, Y., Li, L., Ou, Y., Gao, Z., Li, E., Li, X., et al. (2014). Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* 509, 91–95.

Sudo, K., Kato, K., Matsuzaki, J., Boku, N., Abe, S., Saito, Y., et al. (2019). Development and validation of an esophageal squamous cell carcinoma detection model by large-scale microRNA profiling. *JAMA Netw. Open* 2:e194573. doi: 10.1001/jamanetworkopen.2019.4573

Wang, C., Wang, J., Chen, Z., Gao, Y., and He, J. (2017). Immunohistochemical prognostic markers of esophageal squamous cell carcinoma: a systematic review. *Chin. J. Cancer* 36:65.

Wolff, A. C., Hammond, M. E., Schwartz, J. N., Hagerty, K. L., Allred, D. C., Cote, R. J., et al. (2007). American society of clinical oncology/college of American pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J. Clin. Oncol.* 25, 118–145.

Zhang, J., Jiang, D., Li, X., Lv, J., Xie, L., Zheng, L., et al. (2014). Establishment and characterization of esophageal squamous cell carcinoma patient-derived xenograft mouse models for preclinical drug discovery. *Lab. Invest.* 94, 917–926. doi: 10.1038/labinvest.2014.77

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 27 May 2021
doi: 10.3389/fgene.2021.665065

AutoEncoder-Based Computational Framework for Tumor Microenvironment Decomposition and Biomarker Identification in Metastatic Melanoma

Yanding Zhao[1,2†], Yadong Dong[1,2†], Yongqi Sun[3] and Chao Cheng[1,2]**

[1] Department of Medicine, Baylor College of Medicine, Houston, TX, United States, [2] Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, United States, [3] Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

OPEN ACCESS

Edited by:
Min Tang,
Jiangsu University, China

Reviewed by:
Hong Zheng,
Stanford University, United States
Pritam Mukherjee,
Stanford University, United States
Yulin Zhang,
Shandong University, China

***Correspondence:**
Chao Cheng
chao.cheng@bcm.edu
Yongqi Sun
yqsun@bjtu.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 Genetics

Received: *07 February 2021*
Accepted: *12 April 2021*
Published: *27 May 2021*

Citation:
Zhao Y, Dong Y, Sun Y and Cheng C (2021) AutoEncoder-Based Computational Framework for Tumor Microenvironment Decomposition and Biomarker Identification in Metastatic Melanoma. Front. Genet. 12:665065. doi: 10.3389/fgene.2021.665065

Melanoma is one of the most aggressive cancer types whose prognosis is determined by both the tumor cell-intrinsic and -extrinsic features as well as their interactions. In this study, we performed systematic and unbiased analysis using The Cancer Genome Atlas (TCGA) melanoma RNA-seq data and identified two gene signatures that captured the intrinsic and extrinsic features, respectively. Specifically, we selected genes that best reflected the expression signals from tumor cells and immune infiltrate cells. Then, we applied an AutoEncoder-based method to decompose the expression of these genes into a small number of representative nodes. Many of these nodes were found to be significantly associated with patient prognosis. From them, we selected two most prognostic nodes and defined a tumor-intrinsic (TI) signature and a tumor-extrinsic (TE) signature. Pathway analysis confirmed that the TE signature recapitulated cytotoxic immune cell related pathways while the TI signature reflected MYC pathway activity. We leveraged these two signatures to investigate six independent melanoma microarray datasets and found that they were able to predict the prognosis of patients under standard care. Furthermore, we showed that the TE signature was also positively associated with patients' response to immunotherapies, including tumor vaccine therapy and checkpoint blockade immunotherapy. This study developed a novel computational framework to capture the tumor-intrinsic and -extrinsic features and identified robust prognostic and predictive biomarkers in melanoma.

Keywords: biomarker, gene expression profile, SKCM, tumor microenvironment, immunotherapy

INTRODUCTION

Melanoma is one of the most aggressive tumors, with about 160,000 newly diagnosed cases worldwide each year (Schadendorf et al., 2015; Torre et al., 2015). Although the 5-year overall survival of metastatic melanoma patients has increased up to over 50% with checkpoint blockade immunotherapy (CBI) (Larkin et al., 2019), there are still about half of the patients who do not respond to current immunotherapy whose prognosis remain poor (Khair et al., 2019). Thus,

identifying comprehensive gene signatures that predict the responses to immunotherapy and melanoma patients' overall survival would facilitate the clinical practices of melanoma patients.

Both the tumor cell-intrinsic and cell-extrinsic factors influence the progression and regression of cancer. Extrinsically, immune cell infiltration is a hallmark of melanoma (Li et al., 2016; Thorsson et al., 2018). Four molecular subtypes of metastatic melanoma patients based on the gene expression have been identified and the immune subtype patients had significantly prolonged overall survival (Jönsson et al., 2010). This tumor immune microenvironment can be largely affected by tumor intrinsic features (L. Yang et al., 2019). Several studies reported the positive association between the number of non-synonymous somatic mutations and the abundance of tumor-infiltrating immune cells (Li et al., 2016; Varn et al., 2017). On the contrary, copy number variation (CNV) presented a negative association with immune cell infiltration in the tumor microenvironment across multiple cancer types (Davoli et al., 2017; Zhao et al., 2019). In addition to the genomic features, the tumor oncogenic pathways play a profound role in regulating the immunosuppressive tumor microenvironment and immune evasion (Hanahan and Weinberg, 2011). MYC, as an important transcription factor, has been reported to cooperate with Ras to exclude the infiltration of immune cells (L. Yang et al., 2019). In line with these findings, it has been shown that melanoma patients with high somatic mutation burden, low CNV, or low oncogenic activation are more likely to benefit from immunotherapy (Snyder et al., 2014; Van Allen et al., 2015; Davoli et al., 2017; Lauss et al., 2017).

In order to comprehensively characterize these cell-extrinsic and cell-intrinsic factors in patients, linear regression-based models have been widely used to identify gene signatures in patients. Zhao et al. identified 25 immune-associated genes to depict the abundance of tumor-infiltrating immune cells (Zhao et al., 2019), and Liao et al. combined the expression of two immune genes, CCL8 and DEFB1, for prognosis prediction (Liao et al., 2020). However, the algorithms based on linear regression ignored the complicated nonlinear relationships and correlations among genes. Currently, only few methods designed nonlinear models to capture the tumor-infiltrating immune cells in the microenvironment but mostly focused on the function of specific immune cell populations (Yoshihara et al., 2013; Varn et al., 2016). Thus, in this study, we proposed an Autoencoder-based computational framework to extract both the tumor-intrinsic and -extrinsic features from gene expression of melanoma samples. By applying this framework to the TCGA metastatic melanoma RNA-seq dataset, we identified a number of interrelated nodes. Many of these nodes are found to be significantly associated with patients' prognosis. We selected two most prognostic nodes and defined a tumor-intrinsic (TI) signature and a tumor-extrinsic (TE) signature. Using benchmarked experimental data, we validated that the TE signature reflected the immune cell cytotoxicity pathway while the TI signature captured the MYC oncogenic pathway activity. Both signatures were strong predictors for metastatic melanoma patients' overall survival, even after adjusting for several clinical factors. Moreover, the TE signature could predict the patients' response to MAGE-A3 and anti-CTLA4 immunotherapy. Our results provided a generic computational framework for tumor-intrinsic and -extrinsic feature extraction and identified potential biomarkers for predicting clinical outcome in melanoma.

RESULTS

Overview of the Study

We extracted the tumor-intrinsic and -extrinsic signals from the gene expression data of metastatic melanoma patients in TCGA and identified a number of interrelated modules (**Figure 1**). Among these modules, we identified two representatives associated with tumor-extrinsic (TE) and -intrinsic (TI) features, respectively. We further validated that the TE signature reflected the immune cell cytotoxicity pathway while the TI signature indicated the MYC oncogenic pathway activity. Subsequently, we systematically investigated the function of the extrinsic and intrinsic features in melanoma patients' prognosis and response to immunotherapy, which could be summarized as (1) illustrating the prognostic value of the TE signature and TI signature in metastatic and stage III melanoma patients; (2) developing an integrative model to predict patients' overall survival; (3) examining the prediction power of the TE signature in immunotherapy; and (4) identifying the association between the TI signature and anticancer drugs.

Association of the TI and TE Signatures With Molecular and Immunological Features

In total, 40 nodes were acquired (20 nodes from TE-associated modules and 20 from TI-associated modules). An additional feature selection process was performed to select the most clinically relevant nodes. We first examined the prognostic value of each node in the training data (metastatic TCGA SKCM) and chose the TE-signature (H17) and TI-signature (L7) nodes as the representatives for tumor-extrinsic and -intrinsic features given their performances in predicting prognosis (Methods, **Figure 2A**).

As mentioned in **Figure 1**, we only chose the genes that were correlated with lymphocyte abundance as the input for training. Therefore, we further validated that the TE signature and the TI signature are associated with lymphocyte abundance ($p < 2e\text{-}16$, **Figure 2B**; $p = 9e\text{-}08$, **Figure 2C**). Additional correlation analyses with immune-stimulatory and inhibitory genes confirmed that the TE signature and TI signature were correlated with the immune microenvironment in the tumor with TE signature presenting a positive correlation and TI signature presenting a negative correlation (**Figure 2D**). Those evidences showed that the TE signature and TI signature maintained the original correlation structure with the lymphocyte score.

Next, we aimed to explore the pathways that the TE signature and TI signature represent to unravel their biological indications. Based on the pre-ranked GSEA results of the TE signature (**Supplementary Table 2**), we hypothesized that the TE signature

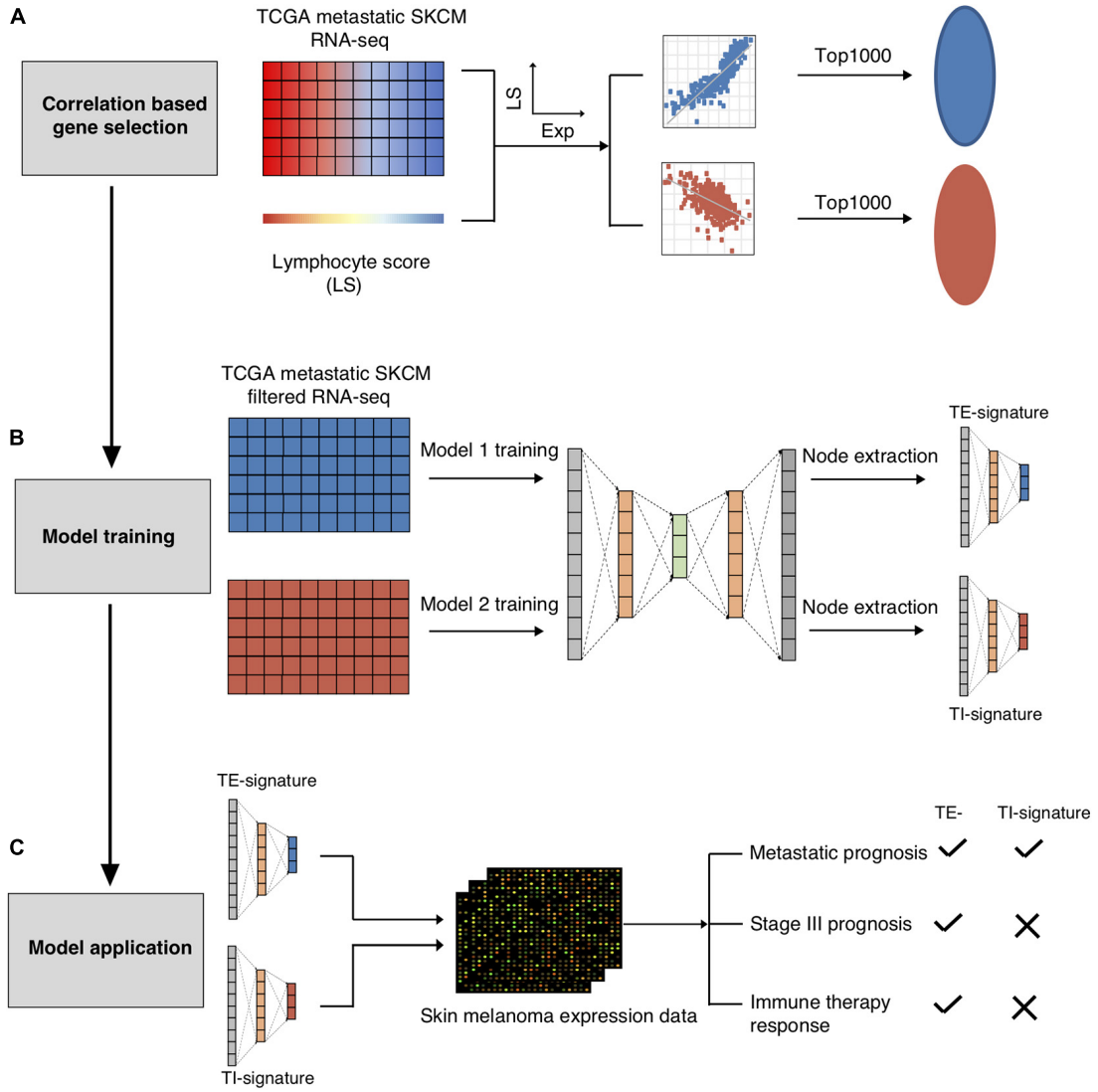

FIGURE 1 | Schematic overview of present study. The TCGA-SKCM metastatic RNA-seq dataset was used to screen out the immune cell and tumor cell related genes. **(A)** The RNA-seq dataset was further split into immune cell related genes expression dataset and tumor cell related gene expression dataset for AutoEncoder decomposition models training. **(B)** Node TE-signature and TI-signature were chosen as the representatives of the immune cell and tumor cell gene expression datasets. **(C)** The trained models were further applied into the independent melanoma gene expression dataset for decomposition. Node TE-signature and TI-signature were then examined for predicting prognosis and immune therapy response.

was associated with immune cell cytotoxicity-related pathways. To test this, the pathway activity for each patient was identified using the TCGA metastatic SKCM patients' expression data of the genes in each pathway of the MsigDB C2 pathway database. The pathway activity of all the pathways in the MsigDB C2 database was then correlated with the TE-signature score for each patient. Shown in **Figure 2E**, the TE-signature score was correlated with the pathway activity of Graft Versus Host Disease (GVHD), mediated by pro-inflammatory immune components (Henden and Hill, 2015; Kuba and Raida, 2018). The hypothesis was further supported by a strong correlation between the TE-signature score and the cytolytic activity (CYT) index in TCGA metastatic melanoma patients (Rho = 0.91, **Figure 2E**). To gain insights on the immune cell subtype contributing to this cytolytic activity, the infiltration levels of six major immune subtypes (NK cell, naive B cell, memory B cell, $CD8^+$ T cell, $CD4^+$ T cell, and monocytes) were correlated with the TE-signature score, which showed that the NK cells having the highest correlation (**Figure 2G**).

We also explored if the TI-signature score captured similar immune profiles. We found strong negative correlations between the TI signature with the CYT index as well as the infiltration of the six immune cell subtypes (Rho = −0.54, **Figures 2E,G**), indicating that the TI signature could rather associate with the tumor-intrinsic but not -extrinsic pathways in the TME. Interestingly, the TI-signature score presented a

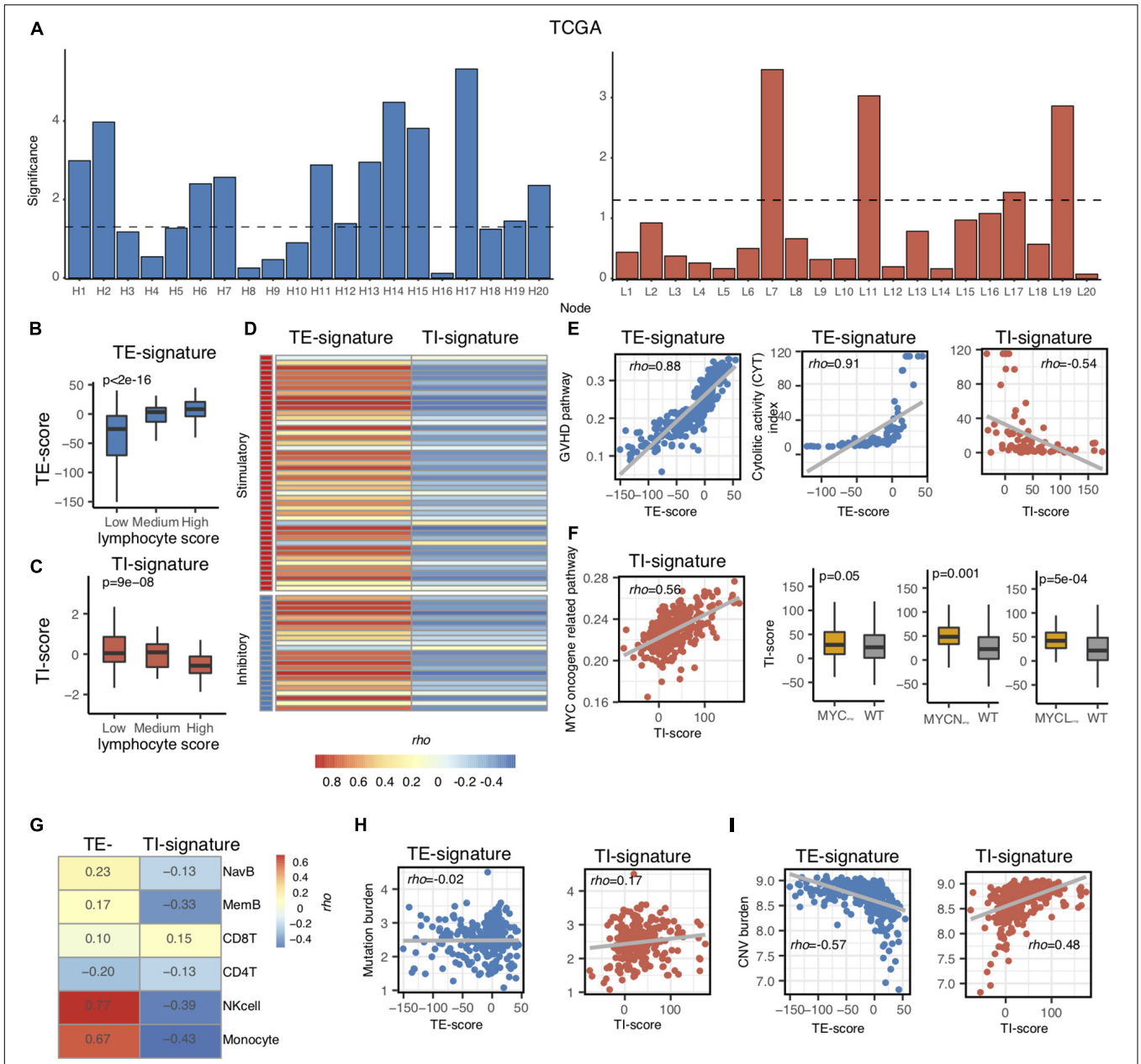

FIGURE 2 | Association of TI and TE signatures with molecular and immunological features. **(A)** Bar plot showing the –log10 (*p*-value) of each node in the TCGA metastatic melanoma dataset. **(B,C)** Boxplot showing the association between TE-signature score and lymphocyte score in **(B)** and between TI-signature score and lymphocyte score in **(C)**. *P*-value was calculated by ANOVA. **(D)** Heat map showing the correlation between immune stimulatory or immune inhibitory gene expression and TE-signature or TI-signature scores. **(E)** Scatterplot showing the correlation between GVHD pathway activity and TE-signature score (left panel) or showing the correlation between CYT index and TE-signature or TI-signature scores (right panel). **(F)** Scatterplot showing the correlation between MYC oncogene pathway activity and TI-signature score. Boxplot indicating the TI-signature scores difference between MYC_{amp} or $MYCN_{amp}$ or $MYCL_{amp}$ and WT. *P*-values were calculated by Wilcoxon rank-sum test. **(G)** Heat map showing the correlation between TE-signature or TI-signature scores and immune cell abundance. **(H)** Scatterplot showing the correlation between Mutation burden and TE-signature (left panel) or TI-signature (right panel) scores. **(I)** Scatterplot showing the correlation between CNV burden and TE-signature (left panel) or TI-signature (right panel) scores. In all scatterplots, the rho was calculated by Spearman correlation.

consistent positive correlation with multiple MYC oncogene-related pathways (**Figure 2F** and **Supplementary Table 3**). MYC, MYCL, or MYCN amplification-induced MYC pathway activation was reported through many studies (Schaub et al., 2018). Thus, the association between the TI-signature and MYC/MYCL/MYCN amplification status were examined and the results indicated that the TI-signature score represented the MYC pathway in the tumor cells.

Evidences above suggested that the TE signature was associated with immune cell cytotoxicity while the TI signature

was associated with MYC pathway activation. These tumor cell-intrinsic and -extrinsic features were largely affected by tumor mutation burden and copy number variation burden (Hanahan and Weinberg, 2011; Chalmers et al., 2017; Taylor et al., 2018). Thus, we further correlated tumor mutation burden and copy number variation burden with both signatures and found that the tumor mutation burden only correlated with the TI-signature score with Rho = 0.17 while the tumor copy number variation burden correlated with both the TE-signature and the TI-signature scores with Rho = −0.57 and Rho = 0.48, respectively (**Figures 2H,I**).

TE and TI Signatures Were Predictive of Prognosis in Metastatic Melanoma

Aforementioned, the TE and TI signatures were chosen based on their prognostic values for metastatic melanoma patients from TCGA, where the TE-signature score associated with better prognosis, yet the TI-signature associated with poor prognosis. The prognosis values of both signatures were further expanded to four other independent metastatic melanoma datasets (GSE8401, GSE65904, GSE19234, and GSE22155). Consistent with the results in the TCGA dataset, patients with higher TE-signature scores had significantly better survival outcomes, while the patients with higher TI-signature scores had worse overall survival (**Figures 3A,B**). Importantly, the distinctive prognostic values of the TE and TI signature were stable across all the datasets, although each dataset had different patient numbers and collection criteria. To further investigate whether the TE signature and TI signature added additional prognostic values to well-established clinical factors, we applied a multivariate Cox regression model and found that both signatures maintained as predictors for patients' overall survival even after adjusting for clinical covariates (e.g., tumor pathological stage at diagnosis, patients age and gender) (**Figures 3C,D**).

TE Signature Predicted Prognosis in Stage III Melanoma Patients

Metastatic melanoma includes distant (stage IV) and regional lymph node metastasis (stage III). After validating that the TE and TI signatures were predictors for stage IV melanoma patients as above, we investigated their prognostic values in stage III melanoma patients. We isolated the stage III SKCM samples in TCGA based on the metastatic regions. We found that the distribution of TE-signature and TI-signature scores are highly different. The stage III samples got the highest TE-signature score while the distal metastatic samples got the highest TI-signature score (**Figures 4A,B**). Then, we calculated the TE-signature and TI-signature scores of samples in two stage III datasets—GSE53118 and GSE54467—and examined their prognostic roles. We found a significant protective association of the TE-signature score with survival (HR = 0.46, P = 0.002, **Figure 4C**) in GSE53118. Adjusting for clinical covariates, including pathological stage at diagnosis, age, and sex, did not substantially change the significant prognostic value of the TE signature we observed (P = 0.02, **Figure 4D**). We were able to repeat this finding in the GSE54467 dataset with the TE signature (HR = 0.38, P = 0.003, **Figure 4E**, P = 0.003, **Figure 4F**). On the contrary, the predictive performance of TI signature was not significant. Therefore, only the TE signature can be used to predict the prognosis of patients with stage III melanoma.

TE and TI Signatures Provided Additional Prognostic Values Than Clinical Factors

Taking into consideration the distinctive associations of the TE signature and TI signature with patients' prognosis, we proposed that the integration of TE signature and TI signature could separate patients much better in terms of overall survival. As a result, we examined the predictive performance of TE signature and TI signature and clinical information on the survival outcome of metastatic melanoma patients. First, we separated the samples in the TCGA SKCM datasets into four groups including TE-signature score-Low and TI-signature score-High, TI-signature score-Low and TE-signature score-High, TE-signature score-Low and TI-signature score-Low, and TE-signature score-High and TI-signature score-High. We found that the survival probability of the four groups of samples was significantly different as shown in **Figure 5A**. As we expected, the group with high TE-signature and low TI-signature scores had the best survival outcome, and the group with low TE-signature and high TI-signature score shaved the worst survival outcome (P = 2E-5, **Figure 5A**). This pattern could still be observed after adjusting for important clinical factors (**Figure 5B**), highlighting the potential of developing clinical applicable model.

Driven by this, we further conducted a multivariate Cox regression analysis on the TCGA cohort to explore the prediction power differences among TE signature, TI signature, and clinical factors and subsequently developed a prognostic prediction model. Shown in **Figure 5C**, the model combined all clinical information with TE signature and TI signature achieving the highest prediction performance, measured by C-index. We further quantified the model's performance on another five independent stage III and stage IV melanoma datasets. The combined model outperformed other models in each independent dataset with the highest C-index = 0.84 being observed in GSE8401 (**Figure 5D**). As expected, the combined model could significantly improve the prediction of patient's survival outcome (P = 0.05, **Figure 5E**).

The TE-Signature Predicted Patients' Response to Immunotherapy

Various immunotherapy strategies have been developed to save metastatic melanoma patients' lives, yet many patients do not respond to current immunotherapies. Precisely predicting that the patient cohort may potentially respond to a certain immunotherapy could maximize the benefit of the therapy to the responding patients while minimizing the risks of severe side effects of immunotherapy for the nonresponding patients. MAGE-A3 anti-gen-specific cancer immunotherapy is a tumor vaccine therapy that has been tested in multiple clinical trials (Daud, 2018; Pol et al., 2019). Therefore, we first investigated whether the TE signature can predict the response

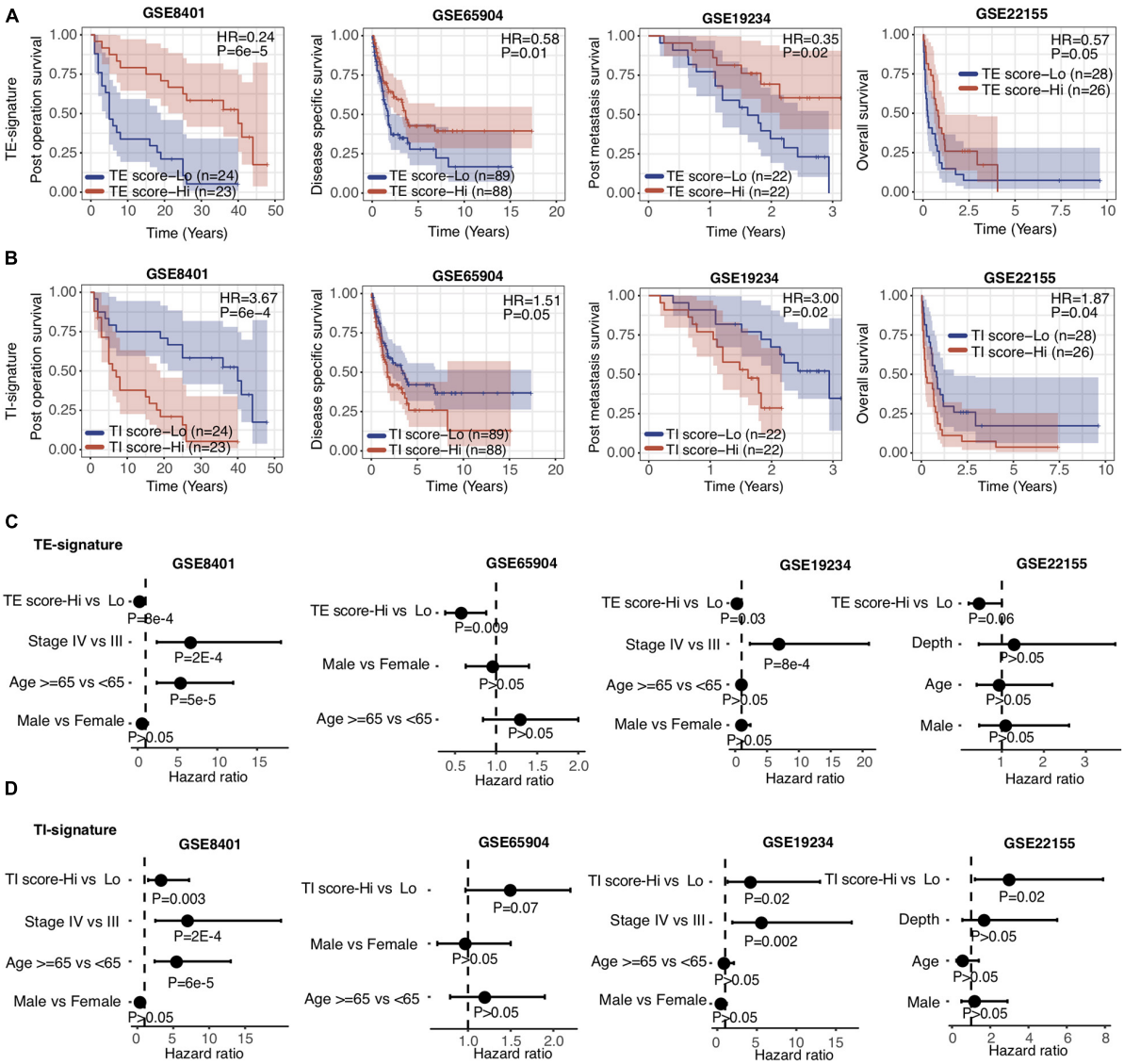

FIGURE 3 | TE signature and TI signature are prognostic in metastatic melanoma. **(A,B)** Kaplan–Meier plots depicting the survival distribution for patients with high (red) and low (blue) TE-signature or TI-signature scores. In Kaplan–Meier plots, *p*-values were calculated using the log-rank test and vertical hash marks indicate censored data. **(C,D)** Forest plot showing hazard ratios and *p*-values of TE-signature score **(C)** or TI-signature score **(D)** and several clinical variables estimated by a multivariate Cox regression model. In all forest plots, HR was presented as the 95% confidence interval, the dotted lines indicate the null association, and the Wald's test was used to determine statistical significance.

of patients with metastatic melanoma to this tumor antigen vaccine therapy. We calculated the TE-signature score and compared its difference between the patients who responded or did not respond to the MAGE-A3 immunotherapy. As shown in **Figure 6A**, there was a significant difference in TE-signature score between two groups of patients (P = 7E-4, **Figure 6A**). Patients who benefited from the MAGE-A3 immunotherapy had significantly higher TE-signature score. An AUC = 0.76 was observed by using the TE-signature score as the predictor (**Figure 6B**).

In addition to antigen-specific immunotherapy, CBI has achieved great success in treating metastatic melanoma patients (Li et al., 2016; Larkin et al., 2019). We additionally analyzed the association between the TE signature and response to anti-CTLA4 therapy. Using the RECIST criteria, patients were labeled as no response (NR), long survival (LS), and complete response (CR). Shown in **Figure 6C**, both CR and LS patients had significantly higher TE-signature scores compared to no response patients (P = 0.01, CR vs. NR; P = 0.01, LS vs. NR). Furthermore, it is not surprising that the TE signature predicted the overall survival in patients treated with anti-CTLA4 therapy and the prediction power remained significant after controlling for clinical factors (P = 0.004, HR = 0.53, **Figure 6D**; P = 0.009, **Figure 6E**).

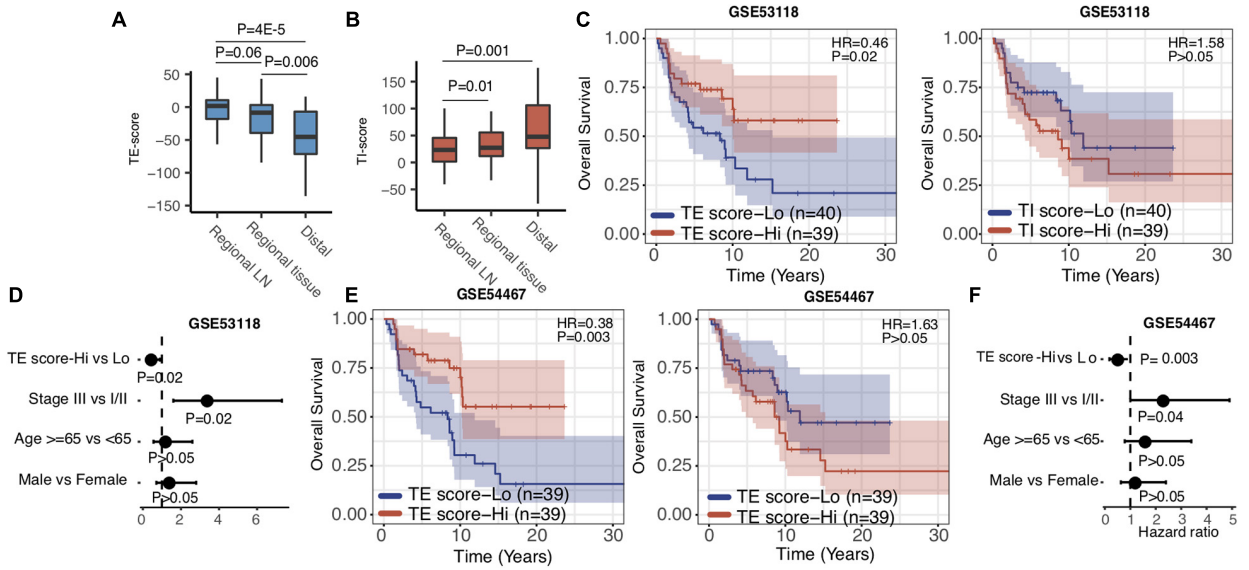

FIGURE 4 | TE signature predicts prognosis in stage III melanoma patients. **(A,B)**. Boxplots indicating the difference of TE-signature or TI-signature scores across different metastatic regions. *P*-values were calculated by the Wilcoxon rank-sum test. **(C)** Kaplan–Meier plots depicting the survival distribution for patients with high (red) and low (blue) TE-signature or TI-signature scores. **(D)** Forest plot showing hazard ratios and *p*-values of TE-signature scores and several clinical variables estimated by a multivariate Cox regression model. **(E)** Kaplan–Meier plots depicting the survival distribution for patients with high (red) and low (blue) TE-signature or TI-signature scores. **(F)** Forest plot showing hazard ratios and *p*-values of TI-signature score and several clinical variables estimated by a multivariate Cox regression model. In Kaplan–Meier plots, *p*-values were calculated using the log-rank test and vertical hash marks indicate censored data. In all forest plots, HR was presented as the 95% confidence interval, the dotted lines indicate the null association, and the Wald's test was used to determine statistical significance.

The TI Signature Was Associated With Cancer Cell Line Sensitivity to Inhibitors of the MYC Pathway

Given that the TI signature reflected poor clinical outcomes of metastatic melanoma patients (**Figures 3**, **6**), we sought for potential drugs that could inhibit the function of the genes in the TI signature which was annotated as the MYC-related pathway (**Figure 2**). Using the GDSC database, we examined the association between anticancer drugs and the TI-signature score (**Supplementary Table 4**). The top three highly correlated anticancer drugs are presented in **Figure 6F**. Interestingly, all those drugs are reported to be kinase inhibitors and have a certain degree of inhibition on the signaling pathway activated by MYC. Erlotinib and Midostaurin were both FDA-approved tyrosine kinase inhibitors and found to inhibit MYC activity (Suenaga et al., 2013; Basit et al., 2018; Allen-Petersen and Sears, 2019). GSK650394 is a novel serum and glucocorticoid-inducible kinase (SGK) inhibitor and has been reported in treating melanoma cancer in some preclinical studies (Scortegagna et al., 2015).

DISCUSSION

In this study, we have built a deep-learning-based computational framework to extract tumor-intrinsic features and extrinsic features from the melanoma gene expression data and define a tumor-intrinsic (TI) signature and a tumor-extrinsic (TE) signature. Then, we systematically investigated how TI and TE signatures affect melanoma patients' prognosis and response to different therapies. To interpret the two signatures, we determined the relative contribution of each gene (bottom node) to them (see Methods). Following that, pathway analyses were performed to identify the underlying pathways. Our results first indicated that the TE signature captured the cytotoxic infiltrating immune cell abundance while the TI signature captured MYC oncogenic pathway activity (**Figures 2B–F**). Next, we examined the prognostic role of the TE signature and TI signature in metastatic melanoma patients and stage III melanoma patients, respectively (**Figures 2–4**). Patients with high TE-signature scores would present a better survival outcome in metastatic and stage III melanoma while patients with high TI-signature scores would present a worse survival outcome in metastatic melanoma (**Figures 3**, **4**). Driven by this, we further constructed different prediction models to quantify the prognostic power of the TE signature, TI signature, and clinical factors. As a result, we found the integrative model using the TE signature; the TI signature with a clinical factor achieved a significantly better performance compared with clinical factor-only model (**Figure 5**). In addition, we showed that the TE signature was predictive of immunotherapy while the TI signature was associated with tyrosine and Ser/Thr kinase inhibitor sensitivity (**Figure 6**).

While many computational methods have been published to capture the immune cell-associated features in the tumor microenvironment, most of them utilized the linear regression-formulized model to characterize the relationship of immune cell-related genes. Given the complicated gene–gene interactions

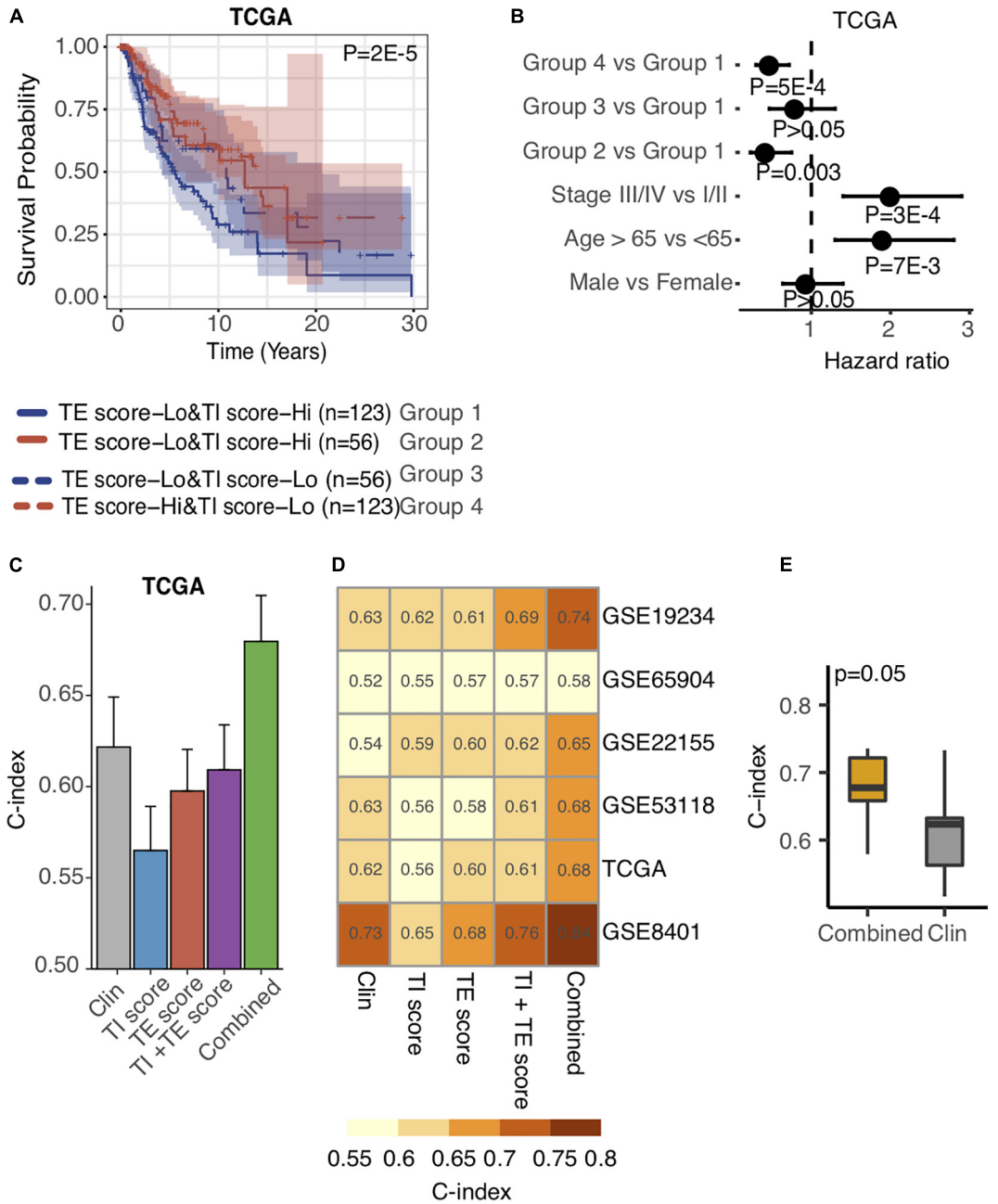

FIGURE 5 | Integration of TE signature and TI signature outperforms prognosis prediction than clinical factors. **(A)** Kaplan–Meier plots depicting the survival distribution for patients in each group. In Kaplan–Meier plots, *p*-values were calculated using the log-rank test, and vertical hash marks indicate censored data. **(B)** Forest plot showing hazard ratios and *p*-values of TE-signature score and several clinical variables estimated by a multivariate Cox regression model. In all forest plots, HR was presented as the 95% confidence interval, the dotted lines indicate the null association, and the Wald's test was used to determine statistical significance. **(C)** Barplot showing the C-index distribution of using Clinical factors, TI-signature scores, TE-signature scores, combination of TI-signature and TE-signature scores, and combination of all features in predicting prognosis in TCGA data. **(D)** Heat map showing the C-index distribution of features listed in **(C)** across different datasets. **(E)** Boxplot showing the C-index difference between combined prognostic model and clinical factor-derived prognostic model. *P*-value was calculated by the Wilcoxon rank-sum test.

in the tumors, our method utilized deep learning, integrating both the linear and nonlinear associations between genes, to capture the function of the tumor-extrinsic features (**Figures 1**, **2**). By choosing IHC-measured lymphocyte score positively associated genes, we decomposed the immune microenvironment into 20 nodes which covered different states or types of immune cells. In our analyses, we only chose the most prognostic node, defined as TE signature, to perform the

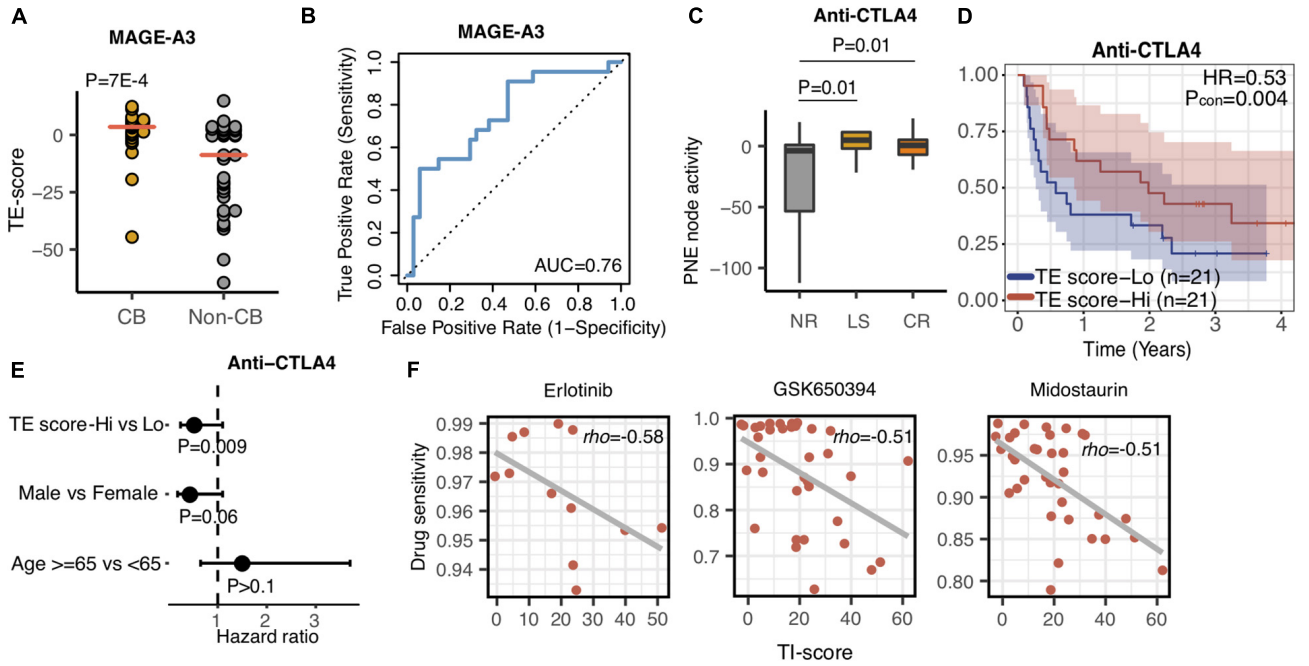

FIGURE 6 | The TE and TI signatures are predictive of drug response. **(A)** Dot plot indicating the TE-signature score difference between responder and non-responder. *P*-value was calculated by Wilcoxon rank-sum test. **(B)** Receiver operating characteristic (ROC) curves for MAGE-A3 therapy response prediction in melanoma patients using the TE-signature score as the predictor. **(C)** Boxplot depicting the TE-signature score difference between different response groups treated with Anti-CTLA4 therapy. *P*-value was calculated by Wilcoxon rank-sum test. **(D)** Kaplan–Meier plots depicting the survival distribution for patients with high (red) and low (blue) TE-signature scores. In Kaplan–Meier plots, *p*-values were calculated using the log-rank test and vertical hash marks indicate censored data. **(E)** Forest plot showing hazard ratios and *p*-values of TE-signature scores and several clinical variables estimated by a multivariate Cox regression model. In all forest plots, HR was presented as the 95% confidence interval, the dotted lines indicate the null association, and the Wald's test was used to determine statistical significance. **(F)** Scatterplot showing the correlation between TI-signature score and Erlotinib, GSK650394, or Midostaurin drug sensitivity. In all scatterplots, the rho was calculated by spearman correlation.

downstream analyses due to its clinical potential (**Figure 2**). However, the more comprehensive analysis of characterizing other nodes will be interesting in the future.

We performed a similar analysis to capture the tumor-intrinsic feature by using IHC-measured lymphocyte score negatively associated genes. It is interesting to observe that the TI-signature score, which reflects MYC oncogene pathway activity, is strongly associated with prognosis. MYC, known as an important oncogenic regulator, has a high fraction of amplification events in melanoma samples, contributing to the overactivation of the MYC oncogenic pathway (Schaub et al., 2018; Schaafsma et al., 2020). As a result, high MYC activity induces melanoma tumor growth, further leading to metastasis. More importantly, MYC also regulates the immune cell function in the tumor microenvironment. MYC could either directly or cooperate with other oncogenes to regulate the expression of PD-L1 to inhibit the function of immune cells or remodel the tumor microenvironment by recruiting macrophages that promote angiogenesis and reduce T cell infiltration (Casey et al., 2018). It is not surprising that MYC activity is negatively associated with the infiltration level of different immune cells (**Figure 2G**). Our study highlighted the significance of MYC in melanoma progression from both tumor-intrinsic and -extrinsic perspectives.

The prognostic value of immune cells in metastatic melanoma has been reported many times, and several-immune-cell-based prognostic biomarkers have been proposed. In this work, we selected genes that best reflected the expression of tumor cells and infiltrating immune cells, respectively. These genes were input into autoencoders to extract tumor-intrinsic and -extrinsic features in the form of bottleneck nodes. From them, we selected two representative nodes and defined a TE signature and a TI signature for prognostic prediction. We first validated the prognostic role of TE signature. Surprisingly, our results indicated that the integration of the TE signature and TI signature could further stratify patients into different risk groups. Patients with high TI-signature and low TE-signature scores had the best survival outcome while patients with high TI-signature and low TE-signature scores had the worst survival outcome. The combination prognostic model, which integrates the TE signature, TI signature, and clinical factors, significantly improved the prediction power of clinical factors derived model (**Figure 5**). These results validated the capability of Autoencoders in denoising and reducing dimensionality for defining prognostic signatures.

Our current model utilized the median score as the cutoff for predicting prognosis because the gene expression profiles from the preclinical cohorts have different scales. To facilitate

the clinical application in the future, we could rescale the expression profiles from those preclinical cohorts to build a cohort-independent threshold for clinical practice. One thing to be noted is that the model prediction power was limited by the clinical information that was provided in the public data. In addition to patients' stage, gender, and Breslow Depth, the surgery information and other treatment information also impact the prognosis in melanoma patients (Bhatia et al., 2015). In the future, with more patient information available, we would like to integrate different clinical information to further improve the prediction accuracy of the combined model.

Targeted immunotherapies have been increasingly used in clinical practice of treating metastatic melanoma patients. MAGE-A3 therapy, a tumor vaccine-based immunotherapy, is still undergoing different clinical trials (Pol et al., 2019). However, several previous clinical trials revealed that MAGE-A3 did not reach the endpoint criteria (Kruit et al., 2005; Dreno et al., 2018). Our results indicated that the TE signature was predictive of MAGE-A3 clinical benefits, which could be further used to guide the design of future clinical trials (**Figures 6A,B**). In addition to tumor vaccine therapy, immune checkpoint blockade therapy has revolutionarily changed immunotherapy and significantly improved overall survival (Larkin et al., 2019). In our results, TE signature could predict anti-CTLA4 response (**Figure 6C**). Patients with high TE-signature scores were more likely to be responders and had a better survival outcome (**Figure 6D**). This result raised the potential of using the TE-signature score as a biomarker for anti-CTLA4 response prediction. In our current analysis, only regular clinical information, including patients' age, gender, and stage, was provided. The efficacy of immunotherapy was also affected by other treatment strategies. For example, chemotherapy administered after immunotherapy might improve the immunotherapy response (Fridlender et al., 2010; Peng et al., 2015). In the future, with such treatment information being released, the prediction accuracy of using the TE signature could be further enhanced.

In the previous section, we mentioned the importance of MYC from both tumor-extrinsic and -intrinsic sides. Inhibiting MYC in melanoma will bring a reduction in tumor proliferation and potentially remodel the tumor microenvironment into immune hot, leading to the increased sensitivity of immunotherapy. Using the GDSC database, we identified that Erlotinib and Midostaurin have inhibitory roles for MYC pathway activity (**Figure 6F**). Erlotinib and Midostaurin were both FDA-approved tyrosine kinase inhibitors and found to repress MYC activity (Suenaga et al., 2013; Basit et al., 2018; Allen-Petersen and Sears, 2019). Interestingly, several clinical trials are ongoing for testing the efficacy of Erlotinib combined with immune-checkpoint blockade therapy (Liang et al., 2018). Our analysis highlighted the potential clinical usage of MYC inhibitors in treating metastatic melanoma patients (Singleton et al., 2017).

In summary, we developed a computational framework to capture the tumor-extrinsic and -intrinsic features in melanoma patients. The two TE- and TI-signature scores we calculated as the representatives of tumor cell feature and immune cell feature are powerful in predicting patient prognosis and response to different treatments. The computational framework could be readily extended to other cancer types.

MATERIALS AND METHODS

Dataset Collection

The TCGA melanoma RNA-seq data were downloaded from Firehose[1] (**Supplementary Table 1**), containing gene expression profiles of 358 metastatic patients. Gene expression values were calculated and normalized by using the RNA-Seq by Expectation-Maximization (RSEM) Algorithm (Li and Dewey, 2011). The clinical information of TCGA melanoma samples was also retrieved from Firehose (see text footnote 1). The information included the patients' age, gender, pathological stage at diagnosis, location of the metastatic tumor, Breslow thickness, lymph node stage, and metastatic stage.

Six additional microarray data sets were used for metastatic melanoma and stage III melanoma prognosis analysis. These data were downloaded from the Gene Expression Omnibus (GEO) database with accession numbers GSE65904 (n = 214), GSE54467 (n = 79), GSE53118 (n = 79), GSE22155 (n = 54), GSE8401 (n = 47), and GSE19234 (n = 44) (Xu et al., 2008; Bogunovic et al., 2009; Jönsson et al., 2010; Mann et al., 2013; Cirenajwis et al., 2015; Jayawardana et al., 2015). GSE65904 and GSE19234 contained disease-specific survival time (DSS) and survival time information after recurrence, respectively, while TCGA-SKCM, GSE54467, GSE53118, GSE22155, and GSE8401 data sets contained overall survival time (OS) information. GSE53118 and GSE54467 provided the survival information for patients with stage III melanoma.

Two datasets were used for immunotherapy response analysis. The treatment information of MAGE-A3 immunotherapy is included in the GSE35640 dataset. It provided the gene expression profiles of a total of 56 patients, among which 34 had no responses and 22 had clinical benefits (Ulloa-Montoya et al., 2013). The anti-CTLA4 immune checkpoint blockade therapy dataset was downloaded from the Database of Genotypes and Phenotypes (dbGaP) under accession number phs000452 (Van Allen et al., 2015). Raw read files were aligned to the GRCh37 human genome assembly using the TopHat v2.1.0 (Kim et al., 2013), and the gene expression was calculated using the Cufflinks v2.2.1 (Trapnell et al., 2012). In total, 42 treatment-naive tumor sample patients were sequenced.

The Genomics of Drug Sensitivity in Cancer (GDSC) dataset was downloaded from the GDSC database[2] for anticancer drug sensitivity testing (W. Yang et al., 2013). It provided a baseline gene expression for a total of 987 cell lines, including with 38 melanoma cell lines, with the corresponding sensitivity to 251 drugs. Drug sensitivity was represented as Area Under the Curve for the fitted model (AUC), with lower values indicating higher sensitivity to a drug (i.e., lower IC50 values).

The genomic characteristics of TCGA melanoma samples were calculated based on the MAF file and DNA sequencing

[1] http://gdac.broadinstitute.org/

[2] https://www.cancerrxgene.org

map downloaded from Firehose (see text footnote 1). Specifically, tumor mutation burden (TMB) was represented as the total number of non-silent somatic mutations in a given TCGA melanoma sample. The copy number variation burden (CNV burden) was calculated using the following equation:

$$CNV - burden = \frac{\sum_{j=1}^{m} |log_2\left(\frac{c_j}{2}\right) * f_j|}{N} \quad (1)$$

where C_j and f_j represent the copy number and the size of the DNA fragment j in the sample; m is the total number of abnormal fragments in the genome, and N is the size of the human genome. For a normal diploid genome, the CNV burden is zero. A higher CNV burden indicates a higher level of copy number variation of the genome.

Gene Expression Decomposition Based on Autoencoder

We applied an autoencoder model to decompose gene expression data for metastatic melanoma samples using the RNA-seq from TCGA. An autoencoder is a type of artificial neural network consisting of two components: an encoder that gradually reduces the input gene expression data into a small number of representative nodes and a decoder that reconstructs the original input (Chen et al., 2018; Way and Greene, 2018; **Supplementary Figure 1**). The configuration of the Autoencoder is shown in **Supplementary Figure 1**; we used two layers for Encoder and Decoder with each layer containing 400 and 100 nodes, respectively. By minimizing the deviation between the reconstructured and the input data, Autoencoder achieves dimensionality reduction using the 20 representative nodes while filtering out noises (**Supplementary Figure 1**). As shown in **Figure 1**, the main steps are elaborated below.

First, TCGA metastatic melanoma RNA-seq data were log transformed and converted into z-scores by subtracting the mean and then dividing the standard deviations of genes across all samples. In order to capture both tumor cell-intrinsic and -extrinsic signals, we selected the top 1000 genes that had the highest positive correlations with lymphocyte infiltration scores ($G_H^{'}$) and the top 1000 genes that had the highest negative correlations ($G_L^{'}$). Lymphocyte infiltration scores were calculated based on IHC staining results from TCGA (Cancer Genome Atlas Network, 2015).

Second, for both of the two gene expression sub-matrices ($G_H^{'}$ and $G_L^{'}$), an Autoencoder model was used to identify 20 informative "hidden" nodes that best capture the whole expression sub-matrices. Autoencoder could integrate both linear and nonlinear structures in the gene expression data and therefore more correctly capture complex gene–gene interactions. Specifically, the configuration of the AutoEncoder model is shown in **Supplementary Figure 1**. There were 1000 nodes of the input layer, corresponding to the gene expression after screening, and then compressed to 400, 100, and 20 nodes in the following layers, and then gradually reconstructed. Each layer of the model is fully connected, and each hidden layer is followed by a rectified linear unit (ReLU) activation function, which is defined as follows.

$$ReLU\,(x) = \begin{cases} x & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \quad (2)$$

In order to train the model, we chose the regularized square loss as the objective function, as shown in equation 5.

$$L = \sum_{i=1}^{n} \varepsilon\,(i) + ||w||^2 = \frac{1}{2}\sum_{i=1}^{n} ||x - D_\theta(E_\theta(x))||^2 + \lambda||w||^2, \quad (3)$$

where n denotes the number of samples and E_θ and D_θ represent the encode and decode functions, respectively. w represents the learnable weight of the AutoEncoder model. λ is the hyperparameter controlling the proportion of the regularization term. We chose a stochastic gradient descent (SGD) optimization method to train the model and to obtain the optimal weight $\boldsymbol{w}$. The compressed features F_H and F_L corresponding to $G_H^{'}$ and $G_L^{'}$ can be obtained by the two well-trained AutoEncoder models, as shown in equations 6 and 7.

$$F_H = E_{\phi_1}(G_H^{'}) \quad (4)$$

$$F_L = E_{\phi_2}(G_L^{'}) \quad (5)$$

where F_H and F_L are two matrices with 20 columns; each row represents a sample, and each column represents a feature compressed by the AutoEncoder model. The performance of the autoencoder model was measured by the R square between the fitted gene expression and the real gene expression. We also tried different numbers of nodes in the bottleneck layer and found the comparable performance.

Finally, from the compressed features F_H and F_L, we selected a feature that best correlated with patient prognosis in TCGA metastatic melanoma samples. Since the two selected features, respectively, capture tumor cell-intrinsic and -extrinsic features, we denoted them as tumor-intrinsic (TI) and tumor-extrinsic (TE) signatures.

Calculation of TE- and TI-Signature Scores in Tumor Samples

For a given melanoma gene expression dataset, we first utilized a Z-score transformation to convert the expression profile to a relative expression profile. We then separated the relative expression profile into two profiles, containing $G_H^{'}$ and $G_L^{'}$ genes, respectively. For each patient in the relative expression profile, we applied the Autoencoder models trained in the TCGA-SKCM metastatic dataset and acquired the corresponding TE- and TI-signature scores according to equations 4 and 5.

Survival Analysis

Cox proportional hazard models were used to investigate the association between signature scores (calculated based on the TE signature or TI) and patient prognosis. Patient samples were dichotomized into two groups by using the median score as

the cutoff value. Univariate Cox regression models were used to determine the association between the dichotomized scores and patient survival. To compare survival between the two groups, Kaplan–Meier plots were used for visualization. The difference between the survival times of different groups was compared by a log-rank test. The multivariate Cox regression model was used to estimate the association between signature scores and patient survival while considering important clinical variables such as age, sex, Breslow score, and tumor stages.

The Kaplan–Meier estimator was implemented in the survival R package. Specifically, the "coxph" function was used to construct Cox proportional hazard models. The "survfit" function was used to generate Kaplan–Meier survival curves. The "survdiff" function was used to statistically compare the difference between survival curves.

Gene Weight Calculation

After model training, we obtained the weights of each layer in TE and TI signature-associated Autoencoder models. The genes with more contributions to the signature tend to have higher weights. The weighted sum of all the possible combinations between each gene and the corresponding signature node (the TE signature-17th node in the F_H and the TI signature-7th node in F_L) can be viewed as the contribution score. The score is defined as follows.

$$\mathrm{GWH\,(i)} = \sum_{\substack{j=1:400 \\ k=1:100}} w_{i,j}^{(1)} \cdot w_{jk}^{(2)} \cdot w_{k,17}^{(3)} \tag{6}$$

$$\mathrm{GWL\,(i)} = \sum_{\substack{j=1:400 \\ k=1:100}} w_{i,j}^{(1)} \cdot w_{jk}^{(2)} \cdot w_{k,7}^{(3)} \tag{7}$$

where $w_{a,b}^{(c)}$ represents the weight between the bth node of the cth hidden layer and the ath node of the prior layer. So GWH (i) and GWL (i) represent the importance score of the ith gene in the TE and TI signature, respectively.

Pathway Analysis

Based on the weight profile that each gene contributes to the node, we performed pre-rank Gene Set Enrichment Analysis using the fgsea R package (Korotkevich et al., 2019). For calculating the specific pathway activity in melanoma patients, Gene Set Variation Analysis was used for integrating the expression profile with the MsigDB C2 pathway database (Subramanian et al., 2005) through GSVA R package (Hänzelmann et al., 2013).

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

CC contributed to conception and design. CC, YS, YD, and YZ contributed to the development of methodology and analysis and interpretation of the data. CC, YD, and YZ contributed to the writing–review, and/or revision of the manuscript. All authors contributed to manuscript revision, read, and approved the final manuscript.

FUNDING

This study is supported by the Cancer Prevention Research Institute of Texas (CPRIT) (RR180061 to CC) and the National Cancer Institute of the National Institutes of Health (1R21CA227996 to CC). CC is a CPRIT scholar in Cancer Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.665065/full#supplementary-material

REFERENCES

Allen-Petersen, B. L., and Sears, R. C. (2019). Mission possible: advances in MYC therapeutic targeting in cancer. *BioDrugs* 33, 539–553. doi: 10.1007/s40259-019-00370-5

Basit, F., Andersson, M., and Hultquist, A. (2018). The Myc/Max/Mxd network is a target of mutated Flt3 signaling in hematopoietic stem cells in Flt3-ITD-induced myeloproliferative disease. *Stem Cells Int.* 2018:3286949. doi: 10.1155/2018/3286949

Bhatia, S., Tykodi, S. S., Lee, S. M., and Thompson, J. A. (2015). Systemic therapy of metastatic melanoma: on the road to cure. *Oncology* 29, 126–135.

Bogunovic, D., O'Neill, D. W., Belitskaya-Levy, I., Vacic, V., Yu, Y.-L., Adams, S., et al. (2009). Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival. *Proc. Natl. Acad. Sci. U.S.A.* 106, 20429–20434. doi: 10.1073/pnas.0905139106

Cancer Genome Atlas Network (2015). Genomic classification of cutaneous melanoma. *Cell* 161, 1681–1696. doi: 10.1016/j.cell.2015.05.044

Casey, S. C., Baylot, V., and Felsher, D. W. (2018). The MYC oncogene is a global regulator of the immune response. *Blood* 131, 2007–2015. doi: 10.1182/blood-2017-11-742577

Chalmers, Z. R., Connelly, C. F., Fabrizio, D., Gay, L., Ali, S. M., Ennis, R., et al. (2017). Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* 9:34. doi: 10.1186/s13073-017-0424-2

Chen, H.-I. H., Chiu, Y.-C., Zhang, T., Zhang, S., Huang, Y., and Chen, Y. (2018). GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization. *BMC Syst. Biol.* 12 (Suppl 8):142. doi: 10.1186/s12918-018-0642-2

Cirenajwis, H., Ekedahl, H., Lauss, M., Harbst, K., Carneiro, A., Enoksson, J., et al. (2015). Molecular stratification of metastatic melanoma using gene expression profiling: prediction of survival outcome and benefit from molecular

targeted therapy. *Oncotarget* 6, 12297–12309. doi: 10.18632/oncotarget.3655

Daud, A. I. (2018). Negative but not futile: MAGE-A3 Immunotherapeutic for melanoma. *Lancet Oncol.* 19, 852–853. doi: 10.1016/S1470-2045(18)30353-X

Davoli, T., Uno, H., Wooten, E. C., and Elledge, S. J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* 355:eaaf8399. doi: 10.1126/science.aaf8399

Dreno, B., Thompson, J. F., Smithers, B. M., Santinami, M., Jouary, T., Gutzmer, R., et al. (2018). MAGE-A3 immunotherapeutic as adjuvant therapy for patients with resected, MAGE-A3-positive, stage III melanoma (DERMA): a double-blind, randomised, placebo-controlled, phase 3 trial. *Lancet Oncol.* 19, 916–929. doi: 10.1016/S1470-2045(18)30254-7

Fridlender, Z. G., Sun, J., Singhal, S., Kapoor, V., Cheng, G., Suzuki, E., et al. (2010). Chemotherapy delivered after viral immunogene therapy augments antitumor efficacy via multiple immune-mediated mechanisms. *Mol. Ther.* 18, 1947–1959. doi: 10.1038/mt.2010.159

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

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

Henden, A. S., and Hill, G. R. (2015). Cytokines in graft-versus-host disease. *J. Immunol.* 194, 4604–4612. doi: 10.4049/jimmunol.1500117

Jayawardana, K., Schramm, S.-J., Haydu, L., Thompson, J. F., Scolyer, R. A., Mann, G. J., et al. (2015). Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, MRNA, MicroRNA, and protein information. *Int. J. Cancer* 136, 863–874. doi: 10.1002/ijc.29047

Jönsson, G., Busch, C., Knappskog, S., Geisler, J., Miletic, H., Ringnér, M., et al. (2010). Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome. *Clin. Cancer Res.* 16, 3356–3367. doi: 10.1158/1078-0432.CCR-09-2509

Khair, D. O., Bax, H. J., Mele, S., Crescioli, S., Pellizzari, G., Khiabany, A., et al. (2019). Combining immune checkpoint inhibitors: established and emerging targets and strategies to improve outcomes in melanoma. *Front. Immunol.* 10:453. doi: 10.3389/fimmu.2019.00453

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36. doi: 10.1186/gb-2013-14-4-r36

Korotkevich, G., Sukhov, V., and Sergushichev, A. (2019). Fast gene set enrichment analysis. *BioRxiv* [Preprint] doi: 10.1101/060012

Kruit, W. H. J., van Ojik, H. H., Brichard, V. G., Escudier, B., Dorval, T., Dréno, B., et al. (2005). Phase 1/2 study of subcutaneous and intradermal immunization with a recombinant MAGE-3 protein in patients with detectable metastatic melanoma. *Int. J. Cancer* 117, 596–604. doi: 10.1002/ijc.21264

Kuba, A., and Raida, L. (2018). Graft versus host disease: from basic pathogenic principles to DNA damage response and cellular senescence. *Mediators Inflamm.* 2018:9451950. doi: 10.1155/2018/9451950

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J.-J., Rutkowski, P., Lao, C. D., et al. (2019). Five-year survival with combined nivolumab and ipilimumab in advanced melanoma. *N. Engl. J. Med.* 381, 1535–1546. doi: 10.1056/NEJMoa1910836

Lauss, M., Donia, M., Harbst, K., Andersen, R., Mitra, S., Rosengren, F., et al. (2017). Mutational and Putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. *Nat. Commun.* 8:1738. doi: 10.1038/s41467-017-01460-0

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* 17:174. doi: 10.1186/s13059-016-1028-7

Liang, H., Liu, X., and Wang, M. (2018). Immunotherapy combined with epidermal growth factor receptor-tyrosine kinase inhibitors in non-small-cell lung cancer treatment. *OncoTargets Ther.* 11:6189–6196. doi: 10.2147/OTT.S178497

Liao, M., Zeng, F., Li, Y., Gao, Q., Yin, M., Deng, G., et al. (2020). A Novel predictive model incorporating immune-related gene signatures for overall survival in melanoma patients. *Sci. Rep.* 10:12462. doi: 10.1038/s41598-020-69330-2

Mann, G. J., Pupo, G. M., Campain, A. E., Carter, C. D., Schramm, S.-J., Pianova, S., et al. (2013). BRAF mutation, NRAS mutation, and the absence of an immune-related expressed gene profile predict poor outcome in patients with stage III melanoma. *J. Invest. Dermatol.* 133, 509–517. doi: 10.1038/jid.2012.283

Peng, J., Hamanishi, J., Matsumura, N., Abiko, K., Murat, K., Baba, T., et al. (2015). Chemotherapy induces programmed cell death-ligand 1 overexpression via the nuclear factor-KB to foster an immunosuppressive tumor microenvironment in ovarian cancer. *Cancer Res.* 75, 5034–5045. doi: 10.1158/0008-5472.CAN-14-3098

Pol, J. G., Acuna, S. A., Yadollahi, B., Tang, N., Stephenson, K. B., Atherton, M. J., et al. (2019). Preclinical evaluation of a MAGE-A3 vaccination utilizing the oncolytic maraba virus currently in first-in-human trials. *Oncoimmunology* 8:e1512329. doi: 10.1080/2162402X.2018.1512329

Schaafsma, E., Zhao, Y., Zhang, L., Li, Y., and Cheng, C. (2020). MYC activity inference captures diverse mechanisms of aberrant MYC pathway activation in human cancers. *Mol. Cancer Res.* 19, 414–428. doi: 10.1158/1541-7786.MCR-20-0526

Schadendorf, D., Fisher, D. E., Garbe, C., Gershenwald, J. E., Grob, J.-J., Halpern, A., et al. (2015). Melanoma. *Nat. Rev. Dis. Primers* 1:15003. doi: 10.1038/nrdp.2015.3

Schaub, F. X., Dhankani, V., Berger, A. C., Trivedi, M., Richardson, A. B., Shaw, R., et al. (2018). Pan-cancer alterations of the MYC oncogene and its proximal network across the cancer genome atlas. *Cell Systems* 6, 282–300.e2. doi: 10.1016/j.cels.2018.03.003

Scortegagna, M., Lau, E., Zhang, T., Feng, Y., Sereduk, C., Yin, H., et al. (2015). PDK1 and SGK3 contribute to the growth of BRAF-mutant melanomas and are potential therapeutic targets. *Cancer Res.* 75, 1399–1412. doi: 10.1158/0008-5472.CAN-14-2785

Singleton, K. R., Crawford, L., Tsui, E., Manchester, H. E., Maertens, O., Liu, X., et al. (2017). Melanoma therapeutic strategies that select against resistance by exploiting MYC-driven evolutionary convergence. *Cell Rep.* 21, 2796–2812. doi: 10.1016/j.celrep.2017.11.022

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371, 2189–2199. doi: 10.1056/NEJMoa1406498

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

Suenaga, M., Yamamoto, M., Tabata, S., Itakura, S., Miyata, M., Hamasaki, S., et al. (2013). Influence of gefitinib and erlotinib on apoptosis and C-MYC expression in H23 lung cancer cells. *Anticancer Res.* 33, 1547–1554.

Taylor, A. M., Shih, J., Ha, G., Gao, G. F., Zhang, X., Berger, A. C., et al. (2018). Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* 33, 676–689.e3. doi: 10.1016/j.ccell.2018.03.007

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T. H., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812–830.e14. doi: 10.1016/j.immuni.2018.03.023

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

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Ulloa-Montoya, F., Louahed, J., Dizier, B., Gruselle, O., Spiessens, B., Lehmann, F. F., et al. (2013). Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J. Clin. Oncol.* 31, 2388–2395. doi: 10.1200/JCO.2012.44.3762

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 350, 207–211. doi: 10.1126/science.aad0095

Varn, F. S., Andrews, E. H., Mullins, D. W., and Cheng, C. (2016). Integrative analysis of breast cancer reveals prognostic haematopoietic activity and patient-specific immune response profiles. *Nat. Commun.* 7:10248. doi: 10.1038/ncomms10248

Varn, F. S., Wang, Y., Mullins, D. W., Fiering, S., and Cheng, C. (2017). Systematic pan-cancer analysis reveals immune cell interactions in the tumor microenvironment. *Cancer Res.* 77, 1271–1282. doi: 10.1158/0008-5472.CAN-16-2490

Way, G. P., and Greene, C. S. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* 23, 80–91.

Xu, L., Shen, S. S., Hoshida, Y., Subramanian, A., Ross, K., Brunet, J.-P., et al. (2008). Gene expression changes in an animal melanoma model correlate with aggressiveness of human melanoma metastases. *Mol. Cancer Res. MCR* 6, 760–769. doi: 10.1158/1541-7786.MCR-07-0344

Yang, L., Li, A., Lei, Q., and Zhang, Y. (2019). Tumor-intrinsic signaling pathways: key roles in the regulation of the immunosuppressive tumor microenvironment. *J. Hematol. Oncol.* 12:125. doi: 10.1186/s13045-019-0804-8

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2013). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi: 10.1093/nar/gks1111

Yoshihara, K., Shahmoradgoli, M., Martínez, 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

Zhao, Y., Schaafsma, E., Gorlov, I. P., Hernando, E., Thomas, N. E., Shen, R., et al. (2019). A leukocyte infiltration score defined by a gene signature predicts melanoma patient prognosis. *Mol. Cancer Res. MCR* 17, 109–119. doi: 10.1158/1541-7786.MCR-18-0173

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